Robustness of Bus Overlays in Optical Networks*

Ari Levon Libarikian and Muriel Médard
Laboratory for Information and Decision Systems (LIDS), MIT
ari@mit.edu, medard@mit.edu

Abstract

We present a bus overlay which uses simple access nodes and is robust to single failures. Our architecture allows the use of existing optical backbone infrastructure. We consider a linear folded bus architecture and introduce a T-shaped folded bus. Although buses are generally not able to recover from failures, we propose a loop-back approach. Our approach allows optical bypass of some routers during normal operation, thus reducing the load on routers, but makes use of routers in case of failures. We analyze the behavior of our linear and T-shaped systems under average conditions and failure conditions. We show that certain simple characteristics of the traffic matrix give meaningful performance characterization. We show that our architecture provides solutions which limit loads on the router.

1 Introduction

In this paper, we present an architecture for LAN overlay networks over an optical Wavelength Division Multiplexing (WDM) backbone. Our architecture is robust to single failures by making use of routers in the optical backbone. Nodes in the network *access* the fiber through the use of access ports, allowing for cheaper data access than architectures involving *only* full-blown optical switches ([6]) and/or routers. The role of these access nodes is to simply place data on, and extract data from a fiber bus. Bus topologies generally are not recoverable. The simplicity of access ports implies that recovery in the backbone needs to be provided by some other source, possibly switches and routers within the network. Hence, this *mixture* of passive access ports, and switching and routing nodes yields a system that meets the recoverability needs of networks, as well as reduce the complexity and cost of the nodes accessing the optical backbone.

The architecture described in this paper makes use of a few powerful nodes, already present in the backbone, that have full switching and routing capabilities. An optical switch is a device that connects a wavelength (or wavelengths) on a fiber to a wavelength (or wavelengths) on another fiber. A router transmits smaller entities such as individual IP packets or MPLS. The entire system is a virtual network topology supported by a real topology. As a result, routers placed on top of switches can help the system support traffic in the case of failure. Figure 1 shows this setup. Traffic going to this switch/router node can either be processed by the router, or optically bypassed. The capacity of an optical switch greatly exceeds that of a router. Therefore, it is beneficial for the backbone to optically bypass a router. Along with recovery and traffic support, the routers also maintain their regular tasks, primarily table look-ups and traffic grooming.

The high data rates carried by optical networks make recovery in the case of failure a crucial feature of these networks. Robustness is the ability of a network to maintain proper service during failures within it, whereas recovery is the ability of the network

*This work was supported by DARPA NGI.

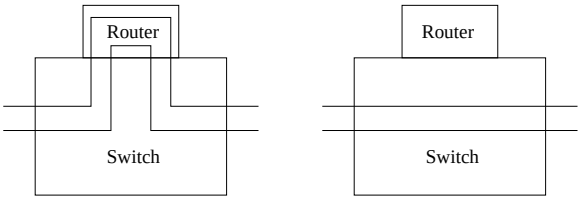

Figure 1: Diagram of a router/switch combination in the overlay. The diagram on the left represents traffic being processed by the router, whereas the diagram on the right represents the router being optically bypassed.

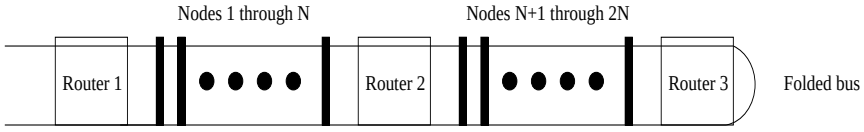

Figure 2: Diagram representing a system with $2N$ nodes and 3 end-routers.

to resume proper operation after a failure has occurred. The traditional router/switch approach would leave access nodes in the fiber unconnected between two switches in the case of a fiber failure.

The implementation of the virtual topology mentioned is done through the use of a folded bus. The folded bus is an overlay requiring a fiber (or fibers), with a wavelength, or group of wavelengths, running through all the nodes (access ports and switches/routers) in each direction. The idea here is that the nodes should be able to both "place" their data onto the bus, as well as "collect" it. Figure 2 represents a $2N$-node, 3-router system employing a folded bus. Folded buses have been used in other systems in previous architectures. Distributed Queue Dual Bus (DQDB) is an example of a folded bus system that is used to interchange data among nodes. DQDB has been analyzed extensively ([1, 3, 4, 5, 8, 11, 12]). HLAN (Helical Local Area Network) is another example of a bus system. HLAN may be constructed as a helical ring structure or a single folded bus and is used mainly to maintain fair utilization of bandwidth among all users. Besides HLAN and DQDB, other dual bus schemes have been studied ([2, 7, 9, 10, 13]).

Drawbacks of these systems are the lack of recoverability, and the limitations of linear systems. The virtual topology introduced by the additional switches and routers in the system ensure recoverability and topology flexibility in the network. At the same time, these routers can access the backbone themselves.

Figure 3 shows an example of a linear system with a folded bus as its backbone. The middle router is optically bypassed, and data is flushed out of one end of the folded bus.

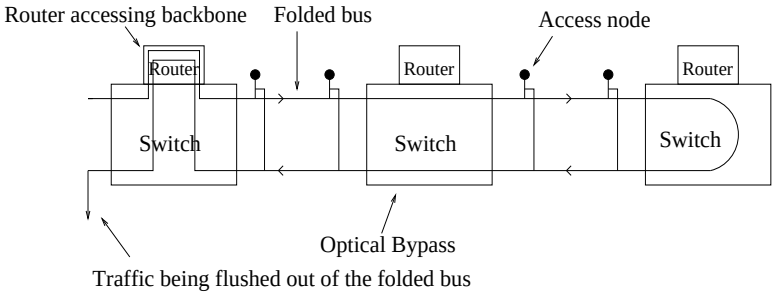

Figure 3: Diagram of a 3-router linear network with a folded bus, and 4 access ports.

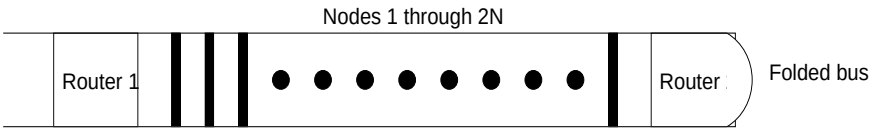

Figure 4: Diagram of a 2-router folded-buy overlay.

Folded buses have advantage that no routing is needed to trasmit the data.

The architecture presented here combines the advantages of folded buses, access ports, optical bypass in switches, and robustness of routing during failures. The main idea behind the virtual topology mentioned earlier is the flexibility introduced by using routers in combination with optical switches. The overlay networks provide increased flexibility to the local-area network. The problem with using routers for recovery purposes, is that the load they support may exceed normal limits. Hence, we evaluate a system's recovery performance not only in the *time-average* sense, but also under *failure* conditions.

This paper compares the performance of various linear systems, as well as that of various T-shaped systems which we describe below. Linear systems are comprised of a set of colinear router-switch pairs and access nodes. The entire system lies on a line and folded buses are adjacent to each other. T-shaped systems are a special type of non-linear systems in which a central node may serve as a hub for up to three folded buses.

The paper is organized as follows. Section 2 describes the system architectures and topologies examined, namely the structure and operating schemes of linear and T-shaped topologies. Section 3 presents the analysis of the performance of these systems as well as several relations and bounds. Section 4 introduces ratios that characterize the network traffic and present the simulation results for both the linear and T-shaped systems. Section 5 presents concluding remarks and directions for further research.

2 System Architecture

The access ports share resources such as a wavelength or a fiber. The nodes represent entities that "access" the fiber both for placing data to be transmitted to another node (distribution portion), and for collecting data transmitted from other nodes (collection portion). This structure is called a folded bus. The upper portion of the folded bus (shown in Figure 4), represents the distribution part, whereas the lower portion represents the collection part. As a result, any traffic meant to be transmitted from node i to node j is simply placed on the distribution portion at node i, and collected by node j on the collection portion.

Traffic is defined in terms of an integer number of units, such as packets or cells. For instance, if u units of traffic or data are to be transmitted from node i to node j in a given interval of time, then the element in the i^{th} row and j^{th} column of the traffic matrix will be u. We define the load on a router as the statistical mean of the number of packets transmitted by it. Routers can handle less traffic than switches, but routers may not need to process all incoming traffic. Routers in the system support traffic emanating from or destined to nodes not included in the overlay (external traffic), and as a result, may have to carry a large burden. Thus, it is desirable that the amount of load placed on them during recovery situations be minimized. The routers of this overlay can and do support all the external traffic and a portion of the internal traffic (during both failure and non-failure states). Figure 1 shows a typical system in which the router associated with a switch is not bypassed, as well as a system in which the router is bypassed. Once the decision has been made to make a router active (i.e. not optically bypassed; processing the traffic) in a local-area network system, the decision of which traffic, and how much

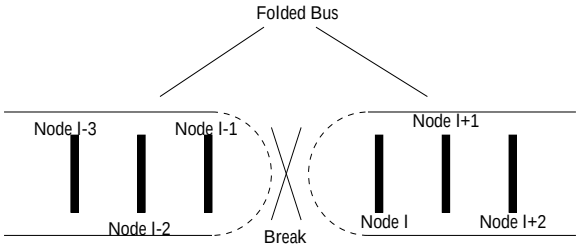

Figure 5: Diagram representing the loopbacks performed on both sides of a folded bus during a failure (e.g. a break in the fiber).

traffic it will handle also has to be made. These are the main issues of the analysis of this paper. Figure 4 shows a system with the number of routers in the system set to $N_R = 3$. It is important to note here that N_R represents the number of routers that are *active* in the system. A router-switch combination may be physically present in the system, but if the switch is configured to bypass the traffic optically, then the router doesn't affect the traffic flow (i.e. the optical bypass would act as an extension of the folded bus). In the case of N_R active routers, there could be up to $N_R - 1$ folded buses (one between every pair of successive routers). It is this value (N_R) whose choice drastically affects system performance in both the *time-average* scenario and the *failure* scenario.

The proposed architecture has a recovery scheme on the folded bus. Let us assume that a failure occurs on link l (the link connecting nodes $l-1$ and l). The system recovers as follows:

- Node $l-1$ "connects" the two terminals of the folded bus so that all the data that was to be transferred across link l is now sent back, in the reverse direction, across link $l-1$. This assures that the router immediately to the left of link l can access the data meant to be transferred across the link, and send it to routers on the other side of the link, via external connections.
- Node l "connects" the two terminals of the folded bus so that all the data meant to cross the link is now looped-back, as in the previous case.

Figure 5 shows this loopback effect.

Using these loop-backs, the routers now become active participants in the recovery-scheme, as all the data that was meant to cross the now-failed link will be sent through at least 2 routers.

Varying the number of routers present in a system causes different changes in system-performance. Firstly, in the *no failure* case, the routers have two purposes:

- To support traffic which does not have both a source and a destination on the same folded bus.
- To support traffic being sent between two nodes not located on the same folded-bus.

Thus, when $N_R = 2$, the only routers in the system are end-routers, and the second responsibility above becomes irrelevant. However, when $N_R > 2$, the routers work to carry traffic from one folded-bus to another. In the *failure* case, the routers have an additional responsibility:

- To provide new recovery paths and to support all re-routed traffic through these new paths.

It is this third purpose that can increase load on the routers.

As the number of routers present in the system increases, the average amount of traffic to be handled *per router* decreases. However, even when there is no link failure in

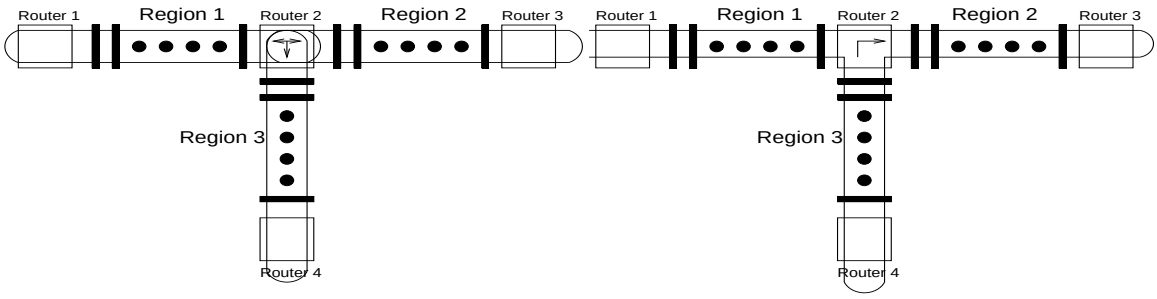

Figure 6: Diagram representing the T-shaped system with 3 folded buses, as well as the general T-shaped system with 1 folded bus.

the system, these additional routers are transferring data from one folded bus to another (something the $N_R = 2$ system doesn't suffer from), and as a result, the amount of average load per router is increased. We explicitly address this trade-off.

We may extend the concept of traditional linear architectures. We introduce T-shaped systems. These systems are formed with 4 routers, one of which is a central router, with the other three each adjacent to it. N nodes separated each pair of routers. Nodes 1 to N in these systems represent the nodes in the first region, nodes $N+1$ to $2N$ represent the nodes in the second region, and nodes $3N$ to $3N+1$ represent the nodes in the third region. Figures 6 and 7 show this system in four different forms. We consider four main variations, which we discuss below.

System A represents the system with 3 folded buses. This means that the central router is connected to every other router through a folded bus. Each folded bus is devoted entirely to the nodes between central router and the corresponding adjacent router. As a result, the central router is a junction where three folded buses meet. It is clear then that any traffic meant to go from one node to another node in another leg of the tree needs to be routed. In System A, router 2, the central router, will carry all of this "cross-over" traffic. The first diagram in Figure 6 shows System A.

System B is another variation of the T-shaped system. In System B, there is one folded bus going through all four routers. However, as evident in the second diagram of Figure 6, this folded bus is not sufficient to transmit all the internal traffic. Specifically, all the traffic meant to go from nodes in the third region to nodes in the second region cannot be transmitted through the single folded bus (unless the folded bus is extended, but this case will be considered in a later system). Hence, the central router is devoted solely to the transmission of data from region 3 to region 2 (but *not* vice-versa).

System C represents a variation of System B in which the central router is not used to transmit any traffic. The single folded bus is extended from the first to the third router, as shown in the first diagram of Figure 7. It is clear now that in the case of no failure, System C utilizes no routers for internal traffic.

System D represents a hybrid of Systems B and C. In effect, the central router is now used to transmit only a fraction of the total traffic meant to go from nodes in region 3 to nodes in region 2. The rest of this traffic will be sent over the extended folded bus. It is clear that in System D, less bandwidth is used on the "extension" portion of the folded bus, yet more load is applied to the central router. The second diagram in Figure 7 shows the architecture of System D.

System A routes all cross-over traffic, whereas System B only routes part of the cross-over traffic. All other traffic (non-crossover, external,...etc.) is routed the same way in both systems. Hence, we see that the average load per router is at least as high

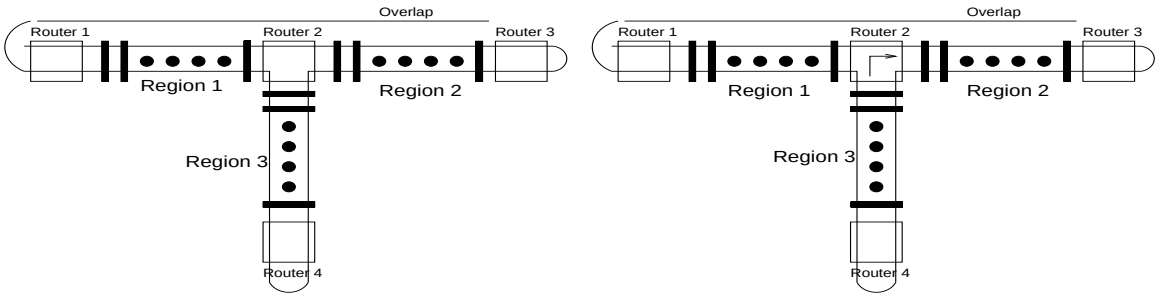

Figure 7: Diagram representing the general T-shaped system with an *extended* folded bus, as well as one with an extended folded bus and usage of the central router.

in System A as in System B. We may also see that Systems B, C, and D only differ in the proportion of cross-over traffic from the lower leg to the right leg routed through the central router (Note: This is related to the share-factor, to be defined later). Since added traffic on the folded bus does not increase the average load per router, we see that the average load per router in System B is greater than that of System D which, in turn, is greater than that of System C.

When considering the performance of these systems, we will assume that the probability of link failure is uniform throughout the system (i.e. there is no bias towards one link failing more than another).

3 Analysis

In this section, we will outline the load equations for two different linear systems ($N_R = 2$ and $N_R = 3$), as well as the T-shaped systems. In the case of the linear systems, we will assume that an $(N+1)$by$(N+1)$ traffic matrix is available, and that there are N nodes per region, and $K = N(N_R - 1)$ total nodes in the system. Let S_i represent the total amount of traffic to be processed by router i (measured in units). **For $N_R = 2$:** In the case of no failure, we have $S_1 = S_2 = \frac{1}{2}(\sum_{i=1}^{K} t_{iK+1})$ where the summation above represents all the traffic destined to, or arriving from the *external* node. It is important to note that, in the case of low or zero external traffic, the routers in this system have almost no load on them. In the case of failure of link l, we have $S_1 = \sum_{i=1}^{l-1} t_{iK+1} + \sum_{i=1}^{l-1}\sum_{j=l}^{K}(t_{ij} + t_{ji})$ and $S_2 = \sum_{i=l}^{K} t_{iK+1} + \sum_{i=1}^{l-1}\sum_{j=l}^{K}(t_{ij} + t_{ji})$.

In the above equations, the first term represents traffic to be carried to and from the external node to the partition of the region accessible by the router. In other words, router 1 deals with the traffic meant for all the nodes to the *left* of link l, and router 2 deals with the traffic meant for all the nodes to the *right* of link l. In addition to this traffic, the routers also have to process the traffic that is to be transmitted from one node on the folded bus on one side of the failure to another node on the folded bus on the other side of the failure. As discussed before, this traffic gets looped-back to the router closest to it. **For $N_R = 3$:** In this scenario there are now 2 regions to deal with. In the case of no failure, we have:

$$S_1 = \sum_{i=1}^{N} t_{iK+1},$$
$$S_2 = \sum_{i=1}^{N}\sum_{j=N+1}^{K}(t_{ij} + t_{ji}) + \sum_{i=N+1}^{K} t_{iK+1}, \text{ and}$$
$$S_3 = 0.$$

In the above equations, S_1 represents all the external traffic to the left-most region. S_2 represents all the external traffic to the right-most region plus the *crossover* traffic (i.e. traffic meant to be transmitted from one region to the other). As a result, the right-most router has no traffic to process. It should be noted that this traffic can be

balanced better, but, due to linearity, this clearly will not affect the average load per router. In the case of a failure in link l we have, for $l \leq N$:

$$S_1 = \sum_{i=1}^{l-1} \sum_{j=l}^{K} (t_{ij} + t_{ji}) + \sum_{i=1}^{l-1} t_{iK+1},$$
$$S_2 = \sum_{i=l}^{N} \sum_{j=N+1}^{K} (t_{ij} + t_{ji}) + \sum_{i=l}^{N} t_{iK+1}, \text{ and}$$
$$S_3 = \sum_{i=N+1}^{K} \sum_{j=l}^{l-1} (t_{ij} + t_{ji}) + \sum_{i=N+1}^{K} t_{iK+1}.$$

Likewise, for $l \geq N$:

$$S_1 = \sum_{i=1}^{N} \sum_{j=l}^{K} (t_{ij} + t_{ji}) + \sum_{i=1}^{N} t_{iK+1},$$
$$S_2 = \sum_{i=l}^{N} \sum_{j=N+1}^{l-1} (t_{ij} + t_{ji}) + \sum_{i=N+1}^{l-1} t_{iK+1}, \text{ and}$$
$$S_3 = \sum_{i=l}^{K} \sum_{j=1}^{l-1} (t_{ij} + t_{ji}) + \sum_{i=l}^{K} t_{iK+1}.$$

In the above equations, we see that, in the case of failure, the load is redistributed. The *crossover* traffic used to be sent over 1 router (the middle one), but now is sent through 2 routers (the two end-routers). At the same time, depending on where the link-failure occurs, different routers have to shoulder different proportions of the load.

In analyzing the T-shaped systems, we show here the equations pertinent to System A. Under non-failure conditions, the following equations hold:

$$S_1 = \sum_{i=1}^{N} t_{i3N+1},$$
$$S_2 = \sum_{i=1}^{N} \sum_{j=N+1}^{3N} (t_{ij} + t_{ji}) + \sum_{i=N+1}^{2N} \sum_{j=2N+1}^{3N} (t_{ij} + t_{ji}),$$
$$S_3 = \sum_{i=N+1}^{2N} t_{i3N+1}, \textit{and}$$
$$S_4 = \sum_{i=2N+1}^{3N} t_{i3N+1}.$$

Now, in the case of failure, note that the system is symmetric, so a failure in any one of the three regions will yield the same equations (subject to a rotation).

Let us consider the case where the failure is between nodes l and $l+1$ in region 1. We have:

$$S_1 = \sum_{i=1}^{l} t_{i3N+1} + \sum_{i=1}^{l} \sum_{j=l+1}^{3N} (t_{ij} + t_{ji}),$$
$$S_2 = \sum_{i=l+1}^{N} \sum_{j=N+1}^{3N} (t_{ij} + t_{ji}) + \sum_{i=N+1}^{2N} \sum_{j=2N+1}^{3N} (t_{ij} + t_{ji}) + \sum_{i=l+1}^{N} t_{i3N+1},$$
$$S_3 = \sum_{i=N+1}^{2N} t_{i3N+1} + \sum_{i=1}^{l} \sum_{j=N+1}^{2N} (t_{ij} + t_{ji}), \text{ and}$$
$$S_4 = \sum_{i=2N+1}^{3N} t_{i3N+1} + \sum_{i=1}^{l} \sum_{j=2N+1}^{3N} (t_{ij} + t_{ji}).$$

We may derive a break-even point between Systems A and B which represents the failure probability at which the two systems exhibit the same average router load. Because System A routers bear some load even under no failure, we would expect that under very small values of p_f, System B would exhibit better average performance (i.e. lower load in System B due to the central router in System A). On the other hand, System A's load is distributed among 3 routers, whereas that of System B is distributed among 2 routers, hence, under certain conditions, the average load *per* router should be lower in System A than in System B.

Let F_C represent the amount of crossover traffic in System A under failure conditions, let F represent the amount of traffic that has it's transmission path changed due to a failure, let C represent the total crossover traffic, under no failure conditions, and let p_f represent the total probability of failure in the system (i.e. the probability that at least one link is down in the system). We have $A_{load} = \frac{1}{3}(C(1-p) + pF + pF_C)$, and $B_{load} = \frac{1}{2}pF$. For $A_{load} > B_{load}$, $p_f < \frac{C}{\frac{1}{2}F + C - F_C} = p^*$. Here, p^* represents the break-even point of p_f (i.e., the point at which the two systems have the same average load per router). For $p_f \geq p^*$, $A_{load} \leq B_{load}$, and for $p_f \leq p^*$, $A_{load} \geq B_{load}$.

Let us assume that $t_{ij} = k \; \forall \; i \neq j$, and that System A has $N_R = 3$. We find that $p^* = \frac{3N^2}{13N^3 + 12N^2 - N}$. Hence, for $N = 2$ (i.e., 2 nodes per region), $p^* = \frac{1}{8}$, whereas for $N = 10$, $p^* = \frac{10}{473}$. As the number of nodes in the system increases, p^* drops, and, as a result, System A becomes more and more attractive (for a given p_f). We see that for any reasonable value of N, the break-even probability is very high - orders of magnitude higher than a typical failure probability.

4 Simulation Results

An $(N+1)$x$(N+1)$ traffic matrix T is assumed to represent all the traffic that is relevant to the system. The element t_{ij} represents the amount of traffic to be sent from node i to node j. The $(N+1)^{st}$ row and column represent all the traffic to be sent to and received from any node outside of the overlay network (i.e. the *external* traffic). The diagonal elements of this matrix represent all the traffic to be sent from a node to itself. We are going to assume that there is internal processing at these nodes, and that the folded buses are not used to process this traffic. Hence, these diagonal elements are set to zero.

Each nonzero element of T is assumed to have a probability distribution. For the purposes of the simulations in this paper, all the distributions are assumed to be of the form $t = \lceil x \rceil - 1$, where $p(x) = \lambda\ e^{-\lambda x}$ for $0 \leq x < \infty$. As a result, the entries in the matrix are integers that have distributions that are *quasi-exponential*.

The parameter λ is chosen to vary across the matrix. We assume that it is more likely that two neighboring nodes send each other traffic than two nodes far from each other, hence the parameter is chosen so that nodes closer together will have a higher traffic rate than nodes farther apart. We therefore choose $\lambda_{ij} = a \mid i - j \mid$ where a represents a positive scale factor.

The *external* traffic is also chosen to be exponentially distributed, but with a constant parameter λ across the nodes, that varies only with the total number of nodes in the system (and a positive scale factor): $\lambda_{ij} = \frac{b}{N}$ where b is another scale factor. $t_{(N+1)(N+1)}$ is chosen to be zero, for obvious reasons. Hence, the overall traffic matrix T is an $(N+1)$x$(N+1)$ matrix with zero diagonal elements, and *quasi-exponentially* distributed non-diagonal elements. For all the elements in the upper-left N x N matrix, the parameter λ_{ij} varies with i and j, whereas for all the elements in the last row and column (excluding $t_{(N+1)(N+1)}$), λ is constant with respect to matrix coordinates.

4.1 Simulation Results of Linear Systems

During normal (*non-failure*) conditions, the system operates as follows:

- All nodes place the data they want to send onto the folded bus.
- All routers on the folded bus process data to be sent within the LAN.
- All routers on the folded bus process data to be sent from external nodes to nodes in the LAN, and *vice-versa*.

Let us consider the case of a link failure. Let N now represent the number of nodes between any two successive routers on the folded bus. We will say that all nodes lying between the same pair of active routers represent a *region*. Hence, there are $N_R - 1$ regions, of N nodes each, yielding a total of $N(N_R - 1)$ nodes in the system. A *link* connects successive nodes, and nodes to neighboring routers. It is clear then, that there are $(N+1)(N_R - 1)$ links in the system.

Let p be the probability that a link between two successive nodes *not separated by a router* fails. Then, the probability that there is a failure between two nodes that *are* separated by a router is $2p$ because this section is comprised of two links of which only one can fail at a time. If p_f represents the probability of any link failure in the system, then $p = \frac{p_f}{(N+1)(N_R-1)}$. Thus, the value of p may be chosen such that the total probability of error equals p_f. When a link failure occurs, it is clear that no traffic can be sent across that link (at least until it recovers).

The first simulation conducted tests the relative performance of two basic linear systems - the first of which is the $N_R = 3$ system (i.e. two end-routers and a router in the middle of the system), and the second of which is the $N_R = 2$ system (i.e. two

end-routers). From now on, System A will represent the $N_R = 3$ system, and System B will represent the $N_R = 2$ system. The *average load* on the routers will serve as the measure of performance (i.e. the lower the average load on the routers, the more efficient and desirable the system). The average load will be defined as the statistical mean of the number of units (or packets) of traffic that is routed in a system per router. Depending on the incident traffic, the systems will perform differently according to this metric.

Since System A has more routers than System B it may appear that the average traffic handled per router is less in System A, than in System B (because the total traffic is constant). However, the middle router in System A has non-zero load even in the case of *no failure*. As a result, the traffic that is simply transported down the folded bus from one node in the first region to another node in the second region will be processed by the middle-router. This load is not present in System B. Hence, we see a tradeoff arising between Systems A and B. As a result, we need to measure how the relative performance of the systems varies with respect to a variable measuring the cross-over traffic. Let $R = \frac{Intra-region\,Traffic}{Inter-region\,Traffic}$. R measures the ratio of traffic that doesn't cross the central router to the the traffic that does. One would expect that, as this ratio increases, the performance of System A relative to system B improves. It should be made clear that R represents a statistic that extracts most of the pertinent information from the traffic matrix T. R can be viewed as a *simplification* of considering the entire traffic matrix. Let $\rho = \frac{A_{load}}{B_{load}}$, where A_{load} represents the average load per router in system A, and B_{load} represents the average load per router in System B. We can use ρ as a measure of the relative performance of the two systems, and see how it varies with R. The main results of the simulation are outlined and described below.

1) Because R does not capture all the information in the matrix, it does not fully determine the value of ρ. In other words, the simulations may, and in fact do produce multiple values of ρ for the same value of R. This is because different matrices with the same value of R may have different distributions of traffic across the matrix, resulting in different recovery mechanisms, and overall load on the routers.

However, for a very special case of $N = 4$ and cross-over traffic restricted to 1 (2 nodes per region, and a total of 1 packet crossing over), the randomness is lost because we have $R = t_{12} + t_{21} + t_{34} + t_{43}$. Two matrices with the same value of R have the same traffic distribution across the matrix.

2) In general, as R increases, ρ decreases. Although R is a simplification of the entire matrix T, it is still a good metric.

Regardless of the value of p_f, the value of ρ drops as R increases. Since the middle router in System A is always processing cross-over traffic, as the relative amount of this traffic is decreased (with respect to the local traffic), the average load in System A with respect to that of System B drops, as expected. The first diagram in Figure 8 shows the ρ vs. R plot for a given system (whose details are outlined on the plot itself) with $p_f = 10^{-3}$. Although more reasonable values of p_f are orders of magnitude smaller, the qualitative results are unaffected. It is interesting to note that the ρ reaches values on the order of 10^2 or 10^3. This can be explained as follows. Since the probability of link-failure is so low, and the external traffic flow has a lower rate than the internal traffic flow, System B's end routers are rarely used. However, the middle-router in System A is always being used to process cross-over traffic. Hence, System B is only processing significant traffic 0.1% of the time, whereas System A is always processing significant traffic. This accounts for the extremely high values of ρ. As R is increased, most of the traffic remains local (in the same region as it was generated), and this deficiency of

System A is not exploited as often.

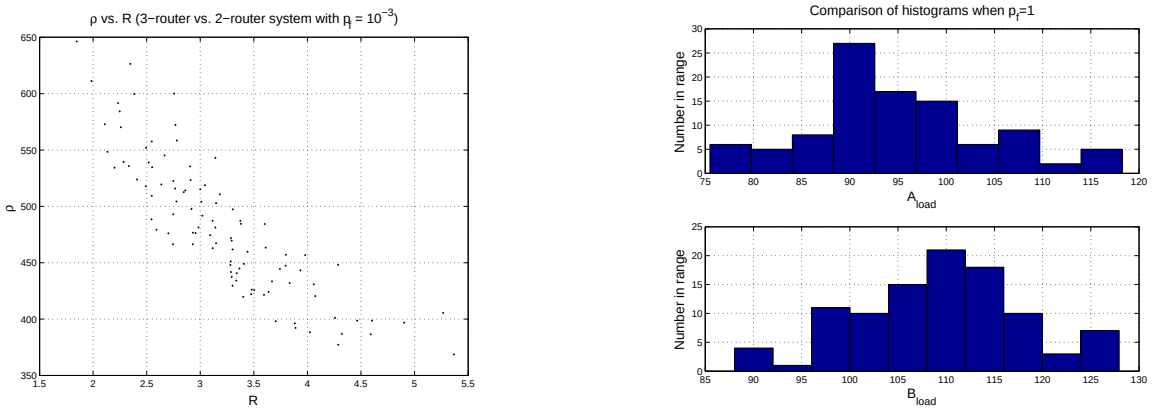

Figure 8: Graphs showing the general monotonically-decreasing relationship between ρ and R, as well as the comparable performance of the two systems in the failure state.

3) If the number of nodes, and external traffic rate are chosen properly, the graph of ρ vs. R becomes decomposed into two curves, one representing the cases with zero external traffic, and the other representing the cases with positive external traffic. Since the external traffic flows to each node are represented by I.I.D. random variables, as the number of nodes is increased, the probability that there is no external traffic decreases. Alternatively, for a fixed number of nodes, as the external traffic rate increases, the probability of zero external traffic decreases. Let p_e represent the probability that there is no external traffic sent to or from the system, and let p_{e1} represent the probability that there is no external traffic sent to a given node in the system. Then, $p_e = (p_{e1})^{2N}$ where N here is the total number of nodes in the LAN.

If the values of N and p_{e1} are chosen properly (p_{e1} can be adjusted by adjusting the incoming traffic rate), situations can occur where different iterations within the same simulation causes traffic matrices with both positive and zero total external traffic.

When the external traffic is zero, the end-routers in System B have no load in the case of no failure. As a result, the only load incurred on those routers is during the time when a failure has occured. When p_f is low, it is clear that B_{load} will be correspondingly low. As a result, two separate curves are present, the lower of which represents those traffic matrices with positive external traffic, and the higher of which represents those traffic matrices with zero total external traffic.

4) Conditioning on having a failure, we see that System A and System B have comparable performance. The second diagram (histogram) of Figure 8 shows this result.

The comparable performance is because of the fact that there are two contending factors. Firstly, the crossover traffic causes System A's performance to degrade with respect to that of System B. However, because 3 routers are shouldering the load in System A, as opposed to 2 routers in System B, we can expect the performances to be comparable. The second diagram of Figure 8 shows the histogram of the proportion of iterations that result in an average value of load within an interval (for both Systems A and B). We see that the histograms for both systems are quite similar.

When the amount of traffic crossing over is kept small during any given iteration, the performance of System A is actually better than that of System B. We see from the second diagram of Figure 8 that System A can possibly outperform System B. Since a large portion of System A's load is due to the crossover traffic during non-failure states, we see that if we limit this to a very low value, the performance of System A dramatically

improves, and eventually surpasses the performance of System B.

Further simulation involved the two comparisons: $N_R > 2$ vs. $2 < N_{R2} < N_R$, and $N_R > 2$ vs. $N_{R2} = 2$. The main results of this simulation are:

1) In general, when $p_f << 1$, the higher the value of N_R, the poorer the performance. As N_R is decreased to 2, the performance improves.

2) A ρ vs. R plot is bounded from below by $\frac{N_{R2}}{N_{R1}}$. In other words, when comparing the performance of the $N_R = 5$ system to the $N_R = 3$ system, the curve will be lower-bounded by $\frac{3}{5}$, and will approach this limit as $R \to \infty$. The reason for this is as follows. Assuming no external traffic, as $R \to \infty$, most of the traffic becomes intra-region traffic, meaning the routers are not used in the case of no failure. Hence, when a failure *does* occur, and each unit passes through exactly 2 routers (because of the recovery scheme), the average load in the system is simply equal to the ratio of the total traffic to be sent through routers to the number of routers, N_R. We see that the ratio of average load values from system to system approaches $\frac{N_{R2}}{N_{R1}}$.

4.2 Simulation Results of T-shaped Systems

As before, random traffic matrices are generated, and the systems are simulated over a large number of iterations. In this case, since there are N nodes in each region, and 3 regions in each system, the T-matrices are $(3N+1)$x$(3N+1)$ matrices, with diagonal entries being zero. Systems A, B, C, and D are all compared.

The ratios of the average load in each system are plotted against new values of R. Let C represent the total crossover traffic, and L represent all the traffic to be sent from the lower leg (region 3) to the right leg (region 2) in the network. Specifically, when comparing Systems A and B, consider the parameter $R_2 = \frac{C}{L}$. R_2 now replaces R as the metric because the difference in the traffic routed through the central router in Systems A and B is highly dependent on the ratio of the crossover traffic to the traffic meant to go from the lower leg of the tree to the right leg. This appears to be a good choice because the dominant factors in determining the average load of these systems when $p_e << 1$ are the traffic going from region 3 to region 2, and all the cross-over traffic.

When comparing Systems B and C, we notice that the only difference in load among the routers when $p_e << 1$ is the amount of traffic to be transferred from region 3 to region 2. Hence, we consider the parameter $R_3 = \frac{C}{L} - 1$. Again, the choice of this value can be attributed to the structures of the two systems. Since System D is a slight variation of Systems B and C, we see that R_3 can be used to measure the relative performance of System D and either System B or System C, and therefore replaces R as the best metric. Once again, all the plots of load-ratios versus values of R are created, and the histograms of the amount of average-load per router of each system are created.

In comparing Systems A and B, we see that System A always performs worse than System B. The reason for this is simple. A certain fraction f of all the cross-over traffic is not routed in System B, whereas all of the cross-over traffic is routed in System A. Moreover, this fraction f is usually greater than $\frac{1}{2}$. Recovery of all other traffic is performed the same way. It is interesting to note that as R_2 increases, the two systems become more comparable in performance. This is because as R_2 increases, the ratio $\frac{L}{C}$ decreases, and more of the crossover traffic is not in the region included in L. Hence, the systems start to behave the same way. As R_2 decreases, System A's performance suffers relative to that of System B. The first graph of Figure 9 shows these results.

In comparing Systems B and C, we see that System C always performs better in terms of average load per router. This can be explained by the fact that System C utilizes more bandwidth in order to save router load. The traffic meant to go from region 3 to region

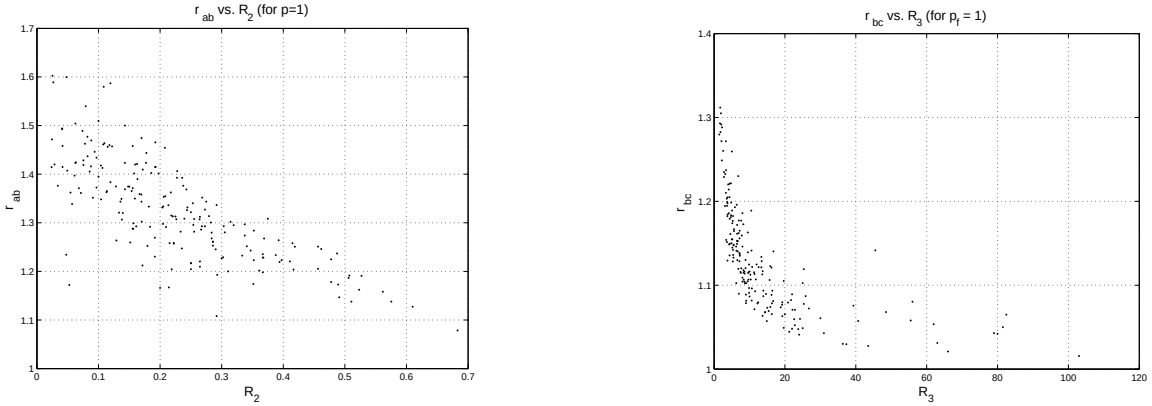

Figure 9: Graphs showing the comparison of performance of Systems A and B, as well as B and C.

2 is passed through the folded bus (in C), thus resulting in savings on load. The obvious tradeoff is that bandwidth on two links is needed for System C. As expected, as R_3 increases, the systems become more and more comparable. This is because most of the cross-over traffic is not included in L, and so neither system carries much internal traffic in the case of no failure. The second graph of Figure 9 shows these results.

Finally, in comparing Systems B and D, a share-factor (α) needs to be selected. This value represents the proportion of the traffic included in L that is to be routed through the central router (in System D), and, together with R_3 provides for a good metric in comparing Systems B and D. The rest of this traffic will be sent down the folded bus. The simulation results show that, as expected, System B's performance is worse than System D's performance (this is because System D is a hybrid of System B and System C), and the ratio of the average load is bounded from above by $\frac{1}{\alpha}$. As $R_3 \to 0$, almost all the traffic is included in L, and so the share-factor affects the average load. On the other hand, as R_3 gets large, the system performances become closer to each other.

In conclusion, we must note that the simulation results showed that the following inequalities always hold, independently of the traffic matrix, as predicted: $A_{load} \geq B_{load} \geq D_{load} \geq C_{load}$. It should be noted that Systems C and D use extra bandwidth, the amount of which depends on the traffic matrix. The ratio of the extra bandwidth needed in C to that needed in D is $r = \frac{1}{1-\alpha}$.

5 Conclusion

We introduce a network bus overlay which uses access nodes in a network relying on a few routers and switches for recoverability and robustness. The architecture uses the existing optical backbone infrastructure. The topologies proposed in this paper are the linear and T-shaped topologies. Recovery is done through the use of loopback - performed by the access nodes on the folded bus. Some routers are optically bypassed under normal operating conditions, and only used in the case of failure. We show that simple values, extracted from the traffic matrix, characterize the performance of the systems.

A direction for further research is the performance of these different systems using different traffic models. The elements of the traffic matrix are independent, identically-distributed (I.I.D.) exponential random variables with different parameters. Interesting paths to follow would be testing the system under uniform, and Pareto (heavy-tailed) traffic. Using uniform traffic would be similar to modelling a system in which all node-pairs have the same amount of demand to be exchanged. Analysis and simulations using heavy-tailed traffic would help us understand how the system performs under more real-

istic traffic in which the probability of a very large deviation is small, but not negligible.

Another avenue to follow would be to consider non-linear bus architectures beyond the T-shaped systems. Examining combinations of linear and T-shaped systems may yield results about general mesh networks. It would be interesting to see if the results pertaining to the simpler linear and T-shaped systems can be extended to account for cascading.

A more interesting result would be measuring the dropoff in performance of a particular system with respect to changing the distribution of the input traffic. For instance, how would the average load of a system change under exponential traffic with a smaller parameter? Given, the performance of a system under a certain traffic model, how would the performance degrade with respect to another distribution? We can measure the "distance" between the distributions of traffic, and in a sense, develop the relationship between the amount a system degrades from one traffic distribution to another, and the distance between the two distributions.

References

[1] C. Bisdikian. Waiting time analysis in a single buffer DQDB (802.6) network. In *Proceedings IEEE INFOCOM*, pages 610–616, 1990.

[2] I. Cidon and Y. Ofek. Metaring - a full duplex ring with fairness and spatial reuse. In *Proceedings IEEE INFOCOM*, pages 969–981, 1990.

[3] M. Conti, E. Gregori, and L. Lenzini. DQDB under heavy load: performance evaluation and fairness analysis. In *Proceedings IEEE INFOCOM*, pages 133–145, 1990.

[4] M. Conti, E. Gregori, and L. Lenzini. A methodological approach to an extensive analysis of DQDB performance and fairness. In *IEEE Journal on Selected Areas in Communications*, volume 9, pages 76–87, January 1991.

[5] A.E. Kamal. Efficient multi-segment message transmission with slot reuse on DQDB. In *Proceedings IEEE INFOCOM*, pages 869–878, 1991.

[6] V.P. Lang, E.A. Varvarigos, and D.J. Blumenthal. The λ-scheduler: A multiwavelength scheduling switch. In *Journal of Lightwave Technology*, volume 18, pages 1049 –1063, August 2000.

[7] J. Limb and C. Flores. Description of Fasnet - a unidirectional local area communication network. In *Bell System Technical Journal*, volume 61, pages 1413–1440, September 1982.

[8] M.A. Rodrigues. Erasure nodes: performance improvements for the IEEE 802.6 MAN. In *Proceedings IEEE INFOCOM*, pages 636–643, 1990.

[9] F.A. Tobagi, F. Borgonovo, and L. Fratta. Expressnet: a high performance integrated-services local area network. In *IEEE Journal on Selected Areas in Communications*, volume SAC-1, pages 898–913, November 1983.

[10] G. Watson and S. Ooi. What *should* a Gbits/s network interface lool like. In North-Holland, editor, *Protocols for High-Speed Networks*, volume II, pages 237–250, Amsterdam, 1990. M.J. Johnson, editor.

[11] G.C. Watson and S. Tohme. S++-anew MAC protocol for Gb/slocal area networks. In *IEEE Journal on Selected Areas in Communications*, volume 11, pages 531–539, may 1993.

[12] J.W. Wong. Throughput of DQDB networks under heavy load. In *Proceedings of EFOC/LAN Conference*, pages 146–151, 1989.

[13] H.-T. Wu, Y. Ofek, and K. Sohraby. Integration of synchronous and asynchronous traffic on the MetaRing architecture and its analysis. In *Proceedings of the IEEE International Conference on Communications*, 1992.