Backup Path Allocation Based On A Correlated Link Failure Probability Model In Overlay Networks

Weidong Cui, Ion Stoica, Randy H. Katz

Department of Electrical Engineering and Computer Science
University of California at Berkeley
{wdc,istoica,randy}@eecs.berkeley.edu

Abstract

Communication reliability is a desired property in computer networks. One key technology to increase the reliability of a communication path is to provision a disjoint backup path. One of the main challenges in implementing this technique is that two paths that are disjoint at the IP or overlay layer may share the same physical links. As a result, although we may select a disjoint backup path at the overlay layer, one physical link failure may cause the failure of both the primary and the backup paths.

In this paper, we propose a solution to address this problem. The main idea is to take into account the correlated link failure at the overlay layer. More precisely, our goal is to find a route for the backup path to minimize the joint path failure probability between the primary and the backup paths. To demonstrate the feasibility of our approach, we perform extensive evaluations under both single and double link failure models. Our results show that, in terms of robustness, our approach is near optimal and is up to 60% better than no backup path reservation and is up to 30% better than using the traditional shortest disjoint path algorithm to select the backup path.

1. Introduction

With the rapid development of the Internet and the emergence of new services, such as Internet Telephony [15], video conferencing, and Virtual Private Networks (VPNs) [7], communication reliability is becoming more and more important. An approach to increase the reliability of a communication path is to provision a disjoint backup path. In case the primary path fails, the traffic is routed through a backup path, if available. However, since it is very hard if not impossible to control which physical links belong to an link at the overlay or IP layer, it is possible that multiple overlay paths share the same physical link despite the fact these paths are disjoint at the overlay level. As a result, the failure of a single physical link may cause multiple overlay paths to fail. Our goal then is to find a backup path that is least likely to share any physical link with the primary path. This choice would minimize the probability of backup path failure when the primary path fails.

Network restoration has been studied in a variety of contexts, such as lightpaths in Wavelength Division Multiplexing (WDM) Optical Networks [11], Virtual Paths (VPs) in Asynchronous Transfer Mode (ATM) networks [12], and most recently Label Switched Paths (LSPs) in Multiprotocol Label Switching (MPLS) networks [19] and application-layer paths in application overlay networks [1] [14]. Research has focused on three main issues: *robustness*, *efficiency*, and *fast restoration*. Robustness is a measure of the probability that primary paths cannot be restored. Efficiency is a measure of the capability of accommodating traffic. Fast restoration is a measure of the time taken to detect primary path failures and switch the route to backup paths.

Overlay networks are usually constructed at the application-layer. However, the Internet consists of multiple layers and each layer is actually an overlay network on top of another network in the underlying layer, such as IP over WDM [13] and application-layer overlay over IP [5]. We will use the term *overlay network* for a general overlay network on top of the physical network. An inherent property of overlay networks is correlation among links because overlay links may share links in the physical network. Thus correlations are introduced among seemingly orthogonal overlay links. On the other hand, restoration is necessary at network layers other than the physical layer. For example, node failures within a service layer can only be dealt with by the actions of peer-level network elements. These facts motivate us to focus our attention on path restoration in the context of overlay networks. To tackle this problem, we propose a novel failure model that takes into account the correlation of overlay link failures. We call this model the *correlated overlay link failure probability model*. We assume the overlay and physical network support bandwidth reservation. Therefore the backup path routing and bandwidth allocation algorithms can be applied not only to application-layer overlay but also to overlay networks at other layers like IP or MPLS.

In particular, we formulate the backup path routing problem based on the correlated overlay link failure probability model as an Integer Quadratic Programming (IQP) problem. We refer to this optimal approach as the *OPtimal backup path Routing algorithm* (OPR). To tackle this NP-hard IQP problem [4], we propose a new backup path routing algorithm called the *Failure Probability cost backup path Routing algorithm* (FPR). FPR decouples backup path routing from primary path routing by routing primary paths based on latency and backup paths based on a new metric called *Failure Probability Cost* (FPC). FPC is a measure of the incremental path failure probability caused by using a link in the path. We compare the FPR algorithm to the OPR algorithm and the *Secondary Shortest backup path Routing algorithm* (SSR) which finds a secondary latency-based shortest path (backup) link-disjoint to a given latency-based shortest path (primary).

We also study the tradeoff between robustness and efficiency for backup path bandwidth sharing under not only single link failures but also double link failures. We refer to the first as the *Single backup path Bandwidth Allocation algorithm* (SBA) and the second as the *Double backup path Bandwidth Allocation algorithm* (DBA). We compare these two algorithms to a naive approach called the *Full backup path Bandwidth Allocation algorithm* (FBA) which reserves the same dedicated bandwidth over the backup path as the primary path.

We undertake an extensive performance evaluation of backup path routing and bandwidth allocation algorithms. Simulation results show that (1) in terms of robustness, our new FPR algorithm is close to the optimal and is up to 60% better than no backup path reservation and is up to 30% better than ignoring link failure probabilities; (2) DBA has a better tradeoff between robustness and efficiency than SBA because DBA is 25% more robust and only 10% less efficient than SBA.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Section 3 presents our overlay link failure probability model, its assumptions, and implications. Since our backup path routing algorithms optimize robustness without taking into account efficiency, we describe backup path routing and bandwidth allocation algorithms in Section 4 and Section 5, separately. Section 6 presents simulations and experimental results, and Section 7 summarizes our work and discusses the future work.

2. Related Work

There have been many research efforts studying the problem of backup path routing and bandwidth allocation in different contexts such as optical networks [11], ATM networks [12], MPLS networks [9] [10] [19], IP networks [20], and application-layer overlay networks [1] [14]. The main technical challenge is to find the right tradeoff among robustness, efficiency, and fast restoration in the specific context.

Restoration methods can be classified as reactive or proactive. In a reactive method, backup paths are not identified before failures happen. A search for a new path is initiated when an existing path fails. In a proactive method, at least one backup path is reserved when establishing the primary path. Both reactive and proactive methods can be link-based or path-based. The link-based approach locally reroutes traffic around the failed component, while path-based methods reroute traffic through a backup path between the source and destination node. Moreover, backup path bandwidth can be dedicated or shared in the proactive approach. In this paper, we study backup path routing and bandwidth allocation in path-based proactive restoration. The key novelty of our work is our study of backup path routing based on a correlated overlay link failure probability model.

Our work differs from previous research in three significant ways. First, we do not require the overlay network to be a fully connected mesh, which reflects the true constraints of application-layer overlay routing. For example, data communication between two overlay nodes may go through a specific server for data transcoding. Second, our work is based on a novel failure model that takes into account the correlation of overlay link failures. In contrast, to the best of our knowledge, prior research has only considered independent link failures. In Section 3, we use Internet data measurements [17] to show that overlay link failures are indeed correlated in the Internet. Third, we consider backup bandwidth sharing assuming both single and double link failures.

Since as shown by Kodialam and Lakshman [9] the problem of node failures can be reduced to the problem of link failures, we assume only link failures in this paper.

3. Correlated Overlay Link Failure Probability Model

3.1. Motivation

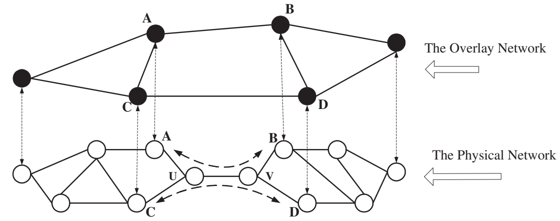


Figure 1. A sample overlay network structure.

Consider an overlay network built on top of the physical network (see Fig. 1). An overlay link is a virtual link directly connecting two overlay nodes in the overlay network. It can be mapped to a physical path. Failures of two overlay links may be correlated because they may share some physical links or nodes. For example, overlay link (A, B) is mapped to a physical path (A, U, V, B), and overlay link (A, B) and (C, D) share physical link (U, V).

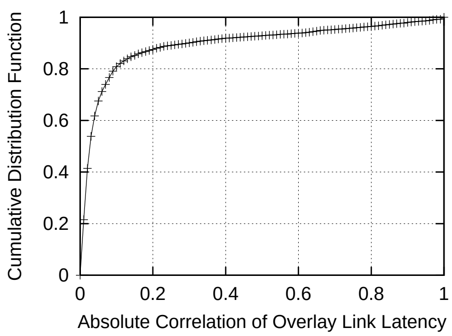


Figure 2. CDF of absolute correlation of overlay link latency.

We define an overlay link failure to occur when the performance degrades to an unacceptable level. In [1], Andersen *et al.* define a virtual application-layer link failure as the length of time τ (on the order of several minutes) over which the packet loss-rate is larger than some threshold p (e.g., 30%). We apply this definition to the link failure in a general overlay network. The choice of τ and p in overlays at different network layer is beyond the scope of this paper. However, our backup path routing algorithm is not dependent on the definition of overlay link failures but on the inherent overlay link correlation. To justify the correlation of overlay link failures, we use the correlation of overlay link latency to prove it indirectly. We analyze end-to-end measurement data (called UW4a) collected by Savage *et al.* in [17]. We refer to the correlation of two random variables X and Y as follows.

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \tag{1}$$

For each pair of measured end hosts u and v, we define $Z_{uv}(t)$ as the latency at time t. We assume the total number of measurements is N. Then we have

$$\begin{aligned} E[Z_{uv}] &= \frac{1}{N}\sum_t Z_{uv}(t) \\ E[Z_{uv}Z_{sd}] &= \frac{1}{N}\sum_t Z_{uv}(t)Z_{sd}(t) \end{aligned}$$

For any two overlay links (u, v) and (s, d), we compute their correlation $\mathrm{Corr}(Z_{uv}, Z_{sd})$ as defined in Eq. 1. Finally we compute the Cumulative Distribution Function (CDF) of the absolute correlation of overlay link latency as shown in Fig. 2. We can see that 20% pairs of overlay links have correlation at least 0.1. This shows that there exist correlations of overlay link latency in today's Internet.

3.2. Overlay Link Failure Probability Model

Our approach to backup path routing is founded on a correlated overlay link failure probability model. In it, we assume that double overlay link failure probabilities $\Pr(\overline{L}_{ij}, \overline{L}_{mn})$ $(\forall (i,j), (m,n) \in E_o)$ are given (see Table 1 for notations). In Section 3.3, we discuss how to obtain these failure probabilities in practice.

In addition to the assumption that overlay link failures are performance-based, we assume that overlay link failures may be transient and persist for periods of time measured in minutes.

By assuming that overlay link failure probabilities are small[1] [3], we have the following approximation: the event that two overlay paths fail at the same time is approximately equivalent to the sum of the small probability events that one overlay link in the first overlay path and another overlay link in the second overlay path fail at the same time. According to this approximation, we compute double overlay path failure probabilities as follows:

$$\Pr(\overline{P}_k^{sd}, \overline{P}_l^{sd}) \approx \sum_{(i,j)\in P_k^{sd}} \sum_{(m,n)\in P_l^{sd}} \Pr(\overline{L}_{ij}, \overline{L}_{mn}) \tag{2}$$

Note that the approximation in Eq. 2 is actually a conservative upper bound[2]. This approximation gives us a way to calculate double overlay path failure probabilities with insufficient information (assuming only single and double overlay link failure probabilities).

3.3. Computing Failure Probabilities

There are two approaches to estimate the single and double overlay link failure probabilities: (1) measurement-based, and (2) based on the knowledge of the physical network topology. In a measurement-based approach, each overlay node periodically probes all its neighbors, and reports statistics of incoming probes to a centralized server. The centralized server processes the data sent by the overlay nodes and computes single and double overlay link failure probabilities periodically. The implication of this approach is that overlay nodes need to be synchronized. We believe this is not a technical challenge because GPS [6] can support accurate synchronization and is increasingly being adopted. It may not be necessary to continuously generate periodical probes because the single and double overlay link failure probabilities will not change frequently when the topology of

[1] This smallness is in the sense of numeric not the quality of service.
[2] de Morgan's Laws.

Table 1. Notations used in this paper

Notation	Comments	Notation	Comments
E_o	the set of all the overlay links	E_u	the set of all the physical paths
D_{ij}	the latency on overlay link (i, j)	d_{ij}	the latency on physical link (i, j)
P^{sd}	the set of all the overlay paths from s to d	p^{ij}	the physical shortest path from i to j
P_k^{sd}	the k-th overlay path from s to d	OPR	the OPtimal backup path Routing algorithm
$\mathcal{F}$	the set of flows set up in the overlay network	FPR	the failure probability cost backup path routing algorithm
b_k	the requested bandwidth of flow k	SSR	the Secondary Shortest backup path Routing algorithm
$\mathcal{P}_{ij}$	the set of flows whose primary paths use link (i, j)	NBR	the No Backup path Routing algorithm
$\mathcal{B}_{ij}$	the set of flows whose backup paths use link (i, j)	FBA	the Full backup path Bandwidth Allocation algorithm
B_{ij}	the bandwidth of overlay link (i, j)	DBA	the Double backup path Bandwidth Allocation algorithm
M_{ij}	the bandwidth reserved for primary paths on link (i, j)	SBA	the Single backup path Bandwidth Allocation algorithm
N_{ij}	the bandwidth reserved for backup paths on link (i, j)	ZBA	the Zero backup path Bandwidth Allocation algorithm

the overlay network is stable. Thus we trade the accuracy of estimating the failure probabilities and the probing overhead. The detail of how the active measurements work is beyond the scope of this paper. Another approach to compute overlay link failure probabilities is to use the knowledge of physical topology and link failure probabilities. In Section 6, we will describe this approach in more detail.

4. Backup Path Routing Algorithms

Backup path routing algorithms need to consider how to route not only backup paths but also primary paths because a primary and backup path pair need to be routed simultaneously to achieve optimal performance in terms of robustness or efficiency. Previous research efforts like [9] [10] [12] have focused on the problem of routing primary paths and backup paths for optimal efficiency, i.e., maximize the amount of traffic admitted, such that primary paths can be restored upon any single link failure. Our goal, however, is to achieve optimal robustness based on the correlated overlay link failure probability model, i.e., minimize the joint failure probability of a primary and backup path pair.

4.1. Optimal Backup Path Routing

Optimal backup path routing seeks to find a primary and backup path pair such that they have minimal joint failure probability. We define a vector $\mathbf{x} = (\cdots, x_{ij}, \cdots)^T$ to represent the flow on the primary path, where x_{ij} is set to 1 if link (i, j) is used on the primary path and is set to 0 otherwise. Similarly, we define a vector $\mathbf{y} = (\cdots, y_{mn}, \cdots)^T$ to represent the flow on the backup path, where y_{mn} is set to 1 if link (m, n) is used on the backup path and is set to 0 otherwise. Based on the approximation made in Eq. 2, we can formulate the optimal backup path routing as the following optimization problem.

Original Optimization Problem

Minimize $\sum_{(i,j)\in E_o} \sum_{(m,n)\in E_o} x_{ij} y_{mn} \Pr(\overline{L}_{ij}, \overline{L}_{mn})$, s.t.,

$$\sum_{j:(i,j)\in E_o} x_{ij} - \sum_{j:(j,i)\in E_o} x_{ji} = \begin{cases} 1 & i = s \\ -1 & i = d \\ 0 & \text{o.w.} \end{cases} \tag{3}$$

$$\sum_{n:(m,n)\in E_o} y_{mn} - \sum_{n:(n,m)\in E_o} y_{nm} = \begin{cases} 1 & m = s \\ -1 & m = d \\ 0 & \text{o.w.} \end{cases} \tag{4}$$

$$\sum_{(i,j)\in E_o} x_{ij} D_{ij} \le \sum_{(m,n)\in E_o} y_{mn} D_{mn} \tag{5}$$

$$x_{ij}, y_{mn} \in \{0, 1\}, \ \forall (i, j), (m, n) \in E_o$$

Eq. 3 and Eq. 4 give the flow balance for the primary path and the backup path, respectively. In the rest of paper, we will refer to Eq. 3 as the primary flow constraint and Eq. 4 as the backup flow constraint. Eq. 5 gives the constraint that the latency of the primary path is no greater than the backup path. We call it the latency constraint. In the literature, people usually add a link-disjoint constraint to make sure that a primary path and a backup path do not share any link. This is implicitly handled in the above optimization problem because link sharing is least likely to happen given that $\Pr(\overline{L}_{ij}) > \Pr(\overline{L}_{ij}, \overline{L}_{mn})$ if $(i, j) \neq (m, n)$.

This optimization problem is actually an Integer Quadratic Programming (IQP) problem. We define a double overlay link failure probability matrix $\mathbf{P}_{|E_o|\times|E_o|}$ having $\Pr(\overline{L}_{ij}, \overline{L}_{mn})$ as entries. We can convert the objective function to the following standard form while keeping all the constraints in Eq. 3 and Eq. 4.

Integer Quadratic Programming Problem

Minimize $\frac{1}{2}\mathbf{z}^T \mathbf{H} \mathbf{z}$ such that the primary and backup flow constraints (Eq. 3 and Eq. 4) and the latency constraint (Eq. 5) are satisfied where we have

$$
\begin{aligned}
\mathbf{z}_{2|E_o|\times 1} &= \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \\
\mathbf{H}_{2|E_o|\times 2|E_o|} &= \begin{pmatrix} 0 & \mathbf{P} \\ \mathbf{P}^T & 0 \end{pmatrix}
\end{aligned}
$$

We call the algorithm which solves the above IQP problem the *OPtimal backup path Routing algorithm* (OPR). It is known that the IQP problem is NP-hard [4]. Currently, we solve OPR by enumerating all possible pairs of paths from a source s to a destination d. This restricts our solution to small overlay networks in our simulations.

4.2. Failure Probability Cost Backup Path Routing

We notice that, in real world networks, primary paths rather than backup paths are used most of the time. Therefore, we can relax the above optimization problem by decoupling backup path routing from primary path routing in such a way that the primary path is routed through a latency-based shortest path for better quality of service and the backup path is routed to minimize the joint double path failure probability. Thus we reduce the IQP problem to two Integer Programming (IP) problems: (1) compute the primary path as the shortest path in terms of latency, and (2) find the backup path that minimizes the joint failure probability given the primary path.

The optimization problem of finding the latency-based shortest path can be formulated as follows. Note that the vector $\mathbf{x}$ represents the flow on the primary path.

Integer Programming Problem 1 (IP1)
Minimize $\sum_{(i,j)\in E_o} x_{ij} D_{ij}$ such that the primary flow constraint (Eq. 3) is satisfied.

Let $\mathbf{x}^*$ denote an optimal solution to IP1. Then the objective function becomes

$$\sum_{(m,n)\in E_o} y_{mn}\Bigg(\sum_{(i,j)\in E_o} x^*_{ij}\,\Pr(\overline{L}_{ij}, \overline{L}_{mn})\Bigg)$$

Note that the term in the parentheses can be regarded as the incremental "cost" on the joint double path failure probability caused by using link (m,n) in the backup path. Thus we define a new metric $C^{\mathbf{x}^*}_{mn}$ called *Failure Probability Cost* (FPC) of using link (m,n) in the backup path where $\mathbf{x}^*$ determines the flow on the primary path.

$$C^{\mathbf{x}^*}_{mn} = \sum_{(i,j)\in E_o} x^*_{ij}\,\Pr(\overline{L}_{ij}, \overline{L}_{mn}) \tag{6}$$

Next, the optimization problem of finding a backup path that minimizes the joint double path failure probability given the primary path $\mathbf{x}^*$ can be formulated as the following IP problem. Note that the vector $\mathbf{y}$ represents the flow on the backup path.

Integer Programming Problem 2 (IP2)
Minimize $\sum_{(m,n)\in E_o} y_{mn} C^{\mathbf{x}^*}_{mn}$ such that the backup flow constraint (Eq. 4) is satisfied.

From the definition of IP1 and IP2, we can easily see that they both are Shortest Path Problems which can be solved in polynomial time[3]. In fact, the primary path is a shortest path based on latency while the backup path is a shortest path based on FPC. Thus we can just use Dijkstra's algorithm [4] to solve these two optimization problems. We call this routing algorithm which solves IP1 for the primary path and then IP2 for the backup path the *Failure Probability cost backup path Routing algorithm* (FPR). Once we obtain the single and double overlay link failure probabilities (refer to Section 3.3), the complexity of the FPR algorithm is comparable to other Link State routing algorithms.

4.3. Secondary Shortest Backup Path Routing

For comparison, we implement a baseline backup path routing algorithm that does not consider overlay link failure probabilities. After a latency-based shortest path is found as the primary path, the backup path should also be routed through a latency-based shortest path which is link-disjoint to the primary path. Here we implicitly assume that the fewer the number of used links is, the smaller the failure probability is. To implicitly guarantee the link-disjoint constraint, we can assume the latency of links in the primary path is infinite when we compute the backup path. Then we define a modified latency $D^{\mathbf{x}^*}_{mn}$ where $\mathbf{x}^*$ determines the flow on the primary path.

$$D^{\mathbf{x}^*}_{mn} = \begin{cases} \infty & x^*_{mn} = 1 \\ D_{mn} & \text{o.w.} \end{cases} \tag{7}$$

Thus we can formulate the problem of routing the backup path based on latency as the following optimization problem.

Integer Programming Problem 3 (IP3)
Minimize $\sum_{(m,n)\in E_o} y_{mn} D^{\mathbf{x}^*}_{mn}$ such that the backup flow constraint (Eq. 4) is satisfied.

The solution to IP3 is a shortest path based on the modified latency $D^{\mathbf{x}^*}_{mn}$. Then we can just use Dijkstra's algorithm to solve it. We call this algorithm the *Secondary Shortest backup path Routing algorithm* (SSR).

We will compare the performance of these backup path routing algorithms, OPR, FPR, SSR, and NBR by simulations in Section 6.

[3]This is because the constraint coefficient matrix is unimodular [18]

5. Backup Path Bandwidth Allocation Algorithms

We assume that both the overlay and physical network provide bandwidth reservation. After backup path routing algorithms find routes for the primary and backup path, the question becomes how much bandwidth should be allocated along the primary and backup path. The goal is to achieve high efficiency with little loss of robustness (see Section 1 for the definition of robustness and efficiency). It is obvious that the requested bandwidth of each flow should be allocated along the primary path. Thus for any link (i, j), the amount of bandwidth reserved for the primary paths is the sum of the requested bandwidth of those flows whose primary paths use that link: (see Table 1 for notations).

$$M_{ij} = \sum_{k \in \mathcal{P}_{ij}} b_k \tag{8}$$

So the goal of the backup path bandwidth allocation algorithms is to determine N_{ij}, the amount of bandwidth reserved on link (i, j) for the backup paths across this link.

Similar to primary path bandwidth allocation, a naive approach for backup path bandwidth allocation is to reserve the requested bandwidth of each flow along the backup path. We define it as the *Full backup path Bandwidth Allocation algorithm* (FBA). In this case, the amount of bandwidth reserved for the backup paths across link (i, j) is the sum of the requested bandwidth of those flows whose backup paths use this link. Formally, we have

$$N_{ij}^{FBA} = \sum_{k \in \mathcal{B}_{ij}} b_k \tag{9}$$

However, since in general the probability that two or more overlay links fail at the same time is much smaller than the probability of a single link failure, reserving the bandwidth for each primary path along the backup path can be inefficient. One way to avoid this inefficiency is to assume only single overlay link failures. This would allow multiple primary paths to statistically share the bandwidth of their backup path.

In particular, for any link (i, j), we allocate the maximum required bandwidth for backup paths on this link by considering all possible single link failures. Formally, we have

$$N_{ij}^{SBA} = \max_{(u,v) \in E_o \backslash (i,j)} \sum_{k \in (\mathcal{P}_{uv} \bigcap \mathcal{B}_{ij})} b_k \tag{10}$$

We call this approach the *Single backup path Bandwidth Allocation algorithm* (SBA).

Previous research efforts have focused on restoration under single link failures. To evaluate the effect on robustness and efficiency by increasing the amount of reserved backup bandwidth, we consider another case where at most two overlay links fail at any time. Similar to SBA, we have

$$N_{ij}^{DBA} = \max_{\substack{(u,v) \in E_o \backslash (i,j) \\ (p,q) \in E_o \backslash (i,j) \\ (u,v) \neq (p,q)}} \sum_{k \in ((\mathcal{P}_{uv} \bigcup \mathcal{P}_{pq}) \bigcap \mathcal{B}_{ij})} b_k \tag{11}$$

We call this approach the *Double backup path Bandwidth Allocation algorithm* (DBA). We will study the tradeoff between robustness and efficiency for SBA and DBA by simulations in Section 6.

To put an upper bound on efficiency, we also consider the baseline approach in which no backup path bandwidth is allocated. We call it the *Zero backup path Bandwidth Allocation algorithm* (ZBA).

$$N_{ij}^{ZBA} = 0 \tag{12}$$

To implement FBA, we can use any signaling protocol that can reserve the requested bandwidth along the primary path to reserve the same amount of bandwidth along the backup path. To implement SBA and DBA, we need to guarantee that every overlay node i has the information of the requested bandwidth b_k and the primary path of flow k whose backup path uses any link (i, j). This kind of information can be distributed by a signaling protocol when it signals every overlay node along the backup path to reserve backup bandwidth. To avoid the expense of maintaining per-flow state, we can take advantage of the algorithm for the *partial information scenario* proposed in [9]. How the signaling protocol works in detail is beyond the scope of this paper.

In Section 6, we will compare the performance of these three backup path bandwidth allocation algorithms in detail.

6. Simulation Experiments

In this section, we present simulation results to evaluate the performance of backup path routing and bandwidth allocation algorithms with respect to robustness, efficiency, and tolerance to inaccurate overlay link failure probability estimates.

6.1. Failure Models

In the simulation design, a key problem is to randomly generate overlay link failures given correlated overlay link failure probabilities. It is hard to directly simulate overlay link failures because failures of any two overlay links may be correlated. We use an indirect approach to solve this problem. First, we randomly assign failure probabilities to physical links uniformly and independently. Then we generate physical link failures at random following the exponential link failure model discussed below. Finally, we use physical link failures to trigger overlay link failures. An overlay link fails whenever at least one physical link in the path fails. Our simulations are based on discrete rather than continuous time.

Due to the limited access to the information of link failure patterns in the the real networks, we use an exponential physical link failure model in our simulations. We assume that link failures are not permanent but can be fixed by some means. With the exponential link failure model, we assume that both up-times and down-times of a physical link follow exponential distributions. To make the failure probability of a physical link be p, the rate of down-times should be $\mu(1-p/p$ if the rate of up-times is μ.

Since we use an indirect approach to generate overlay link failures, we need to compute overlay link failure probabilities based on independent physical link failure probabilities such that they conform to indirectly generated overlay link failures. We assume physical link failure probabilities are small [8]. Then we can make the following approximation to compute single overlay link failure probabilities (see Table 1 for notions).

$$\Pr(\overline{L}_{ij}) \approx \sum_{(u,v)\in p^{ij}} \Pr(\overline{l}_{uv}) \quad (13)$$

Two kinds of physical link failures can cause two overlay links to fail at the same time. The first kind is any failure of a physical link shared by the two overlay links. The second is any simultaneous failures of one physical link used by the first overlay link and another physical link used by the second overlay link. Thus we can approximately compute double overlay link failure probabilities as follows:

$$\Pr(\overline{L}_{ij}, \overline{L}_{mn}) \approx \sum_{(u,v)\in(p^{ij}\bigcap p^{mn})} \Pr(\overline{l}_{uv}) + \sum_{(u,v)\in(p^{ij}\setminus p^{mn})} \sum_{(x,y)\in(p^{mn}\setminus p^{ij})} \Pr(\overline{l}_{uv}) \Pr(\overline{l}_{xy}) \quad (14)$$

Note that the approximation in Eq. 14 is more accurate than that in Eq. 2. We use Eq. 14 instead of Eq. 2 here because we want to have accurate double overlay link failure probabilities for FPC-based routing. When we compute double overlay path failure probabilities in Eq. 2, however, we only need the relative values of different primary and backup routes to make a choice.

6.2. Simulation Setup

We use GT-ITM [2] to generate random network topologies for our simulations. For each physical and overlay network size, we generate 10 random network topologies. We randomly map nodes in the overlay network to nodes in the physical network except transit nodes. This is because transit nodes are core routers or switches which will not correspond to end hosts in the overlay network. All the links in both the physical network and the overlay network are duplex. The source and destination are selected randomly from the set of overlay nodes. Simulation parameters are listed in Table 2.

Table 2. Simulation Parameters

Parameter	Value	Comment
E_o	10 or 30 or 50	using random graph model
E_u	500 $\sim$ 10000	using transit-stub graph model
B_{ij}	500Mb/s	
b_k	[0,10]Mb/s	following uniform distribution
$\Pr(\overline{l}_{ij})$	[10^{-5},10^{-4}]	following uniform distribution [8]

At the beginning of every experiment, we add both primary and backup paths into the overlay network for randomly generated flow requests until a total of 10 requests have been rejected. At this point, we assume the overlay network is saturated. This provides a consistent network state upon which we can compare the performance of different backup path routing and bandwidth allocation algorithms. For every setting, we run our experiment 30 times and compute the average and standard deviation.

6.3. Experimental Results

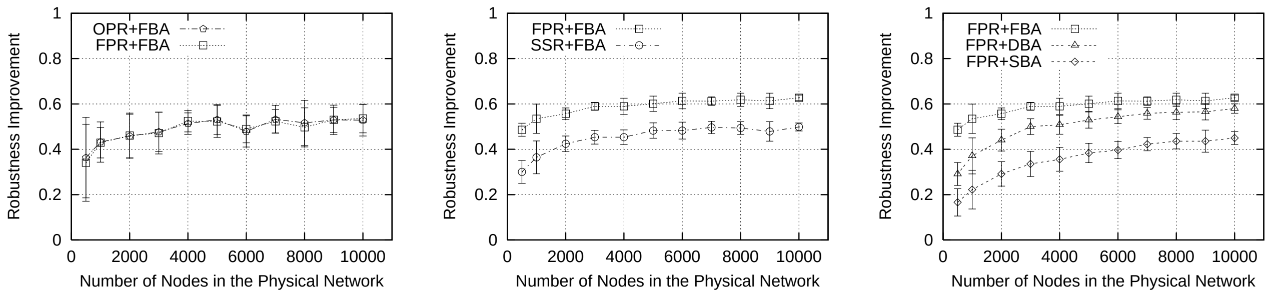
We have discussed four backup path routing algorithms: OPR, FPR, SSR, and NPR; and four backup path bandwidth allocation algorithms: FBA, SBA, DBA, and ZBA. We extensively compare the performance of various combinations of backup path routing and bandwidth allocation algorithms. (see Table 3). Note that in NBR and ZBA there is no backup path at all. Due to space limitations, we will only present experimental results of using overlay networks with 50 nodes (except the robustness experiments of OPR where 10-node networks are used). We observe similar performance on overlay networks with either 10, 30 or 50 nodes.

Table 3. Combinations of Backup Path Routing and Bandwidth Allocation Algorithms Used in Experiments

	FBA	SBA	DBA	ZBA
OPR	$\surd$	$\surd$	$\surd$	-
FPR	$\surd$	$\surd$	$\surd$	-
SSR	$\surd$	$\surd$	$\surd$	-
NBR	-	-	-	$\surd$

6.3.1 Robustness Experiments

Recall that robustness is a measure of the probability that primary paths cannot be restored. To evaluate robustness of different backup path routing and bandwidth allocation algorithms, we define a metric called the *Fatal Path Failure*

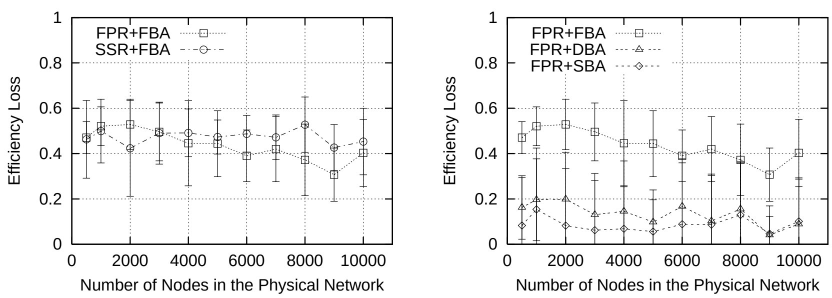


(a) OPR vs. FPR (on 10-Node Overlay Networks)

(b) FPR vs. SSR

(c) SBA vs. DBA vs. FBA

Figure 3. Robustness Experiments



(a) FPR vs. SSR

(b) SBA vs. DBA vs. FBA

Figure 4. Efficiency Experiments

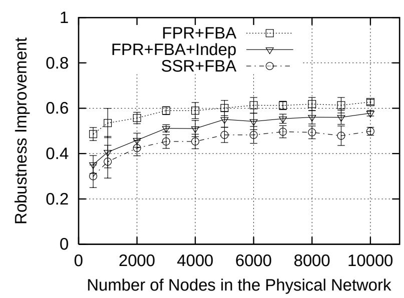


Figure 5. Fault Tolerance Experiments

Probability (FP). It measures the average probability that any primary and backup path pair fail simultaneously. Formally, we have

$$\mathrm{FP} = \frac{1}{|E_o|(|E_o|-1)} \sum_{i \in E_o} \sum_{j \in E_o \setminus \{i\}} \frac{T_{ij}}{T_{total}} \tag{15}$$

where T_{ij} is the total simultaneous failure time of the primary and backup path pair from overlay node i to j and T_{total} is the total simulation time. Since we only care about relative performance, we compare the robustness improvement of different algorithms instead of using FP directly. For example, the robustness improvement achieved by FPR+FBA is defined as follows.

$$\frac{\mathrm{FP}_{\mathrm{NBR+ZBA}} - \mathrm{FP}_{\mathrm{FPR+FBA}}}{\mathrm{FP}_{\mathrm{NBR+ZBA}}} \tag{16}$$

We compare FPR to OPR in Fig. 3(a) and compare FPR to SSR in Fig. 3(b). In these experiments, we use FBA which guarantees that backup path bandwidth is enough for all link failures. Thus robustness is determined only by routing. We compare the robustness of different bandwidth allocation algorithms in Fig. 3(c). We use FPR in these experiments.

6.3.2 Efficiency Experiments

Recall that efficiency is a measure of the capability of accommodating traffic. In the set of efficiency experiments,we use the number of admitted flow requests (defined as AF) to evaluate efficiency of different backup path routing and bandwidth allocation algorithms. The number of admitted flow requests reflects the amount of traffic transferred in the overlay network because each flow request is uniformly generated at random in terms of the source, the destination, and the requested bandwidth. Similar to robustness experiments, we compare the efficiency loss of different algorithms. For example, efficiency loss of FPR+FBA is defined as follows.

$$\frac{\mathrm{AF}_{\mathrm{NBR+ZBA}} - \mathrm{AF}_{\mathrm{FPR+FBA}}}{\mathrm{AF}_{\mathrm{NBR+ZBA}}} \tag{17}$$

We compare the efficiency loss of backup path routing algorithms in Fig. 4(a) and the efficiency loss of different backup path bandwidth allocation algorithms in Fig. 4(b).

6.3.3 Fault Tolerance Experiments

To test the sensitivity of the robustness of FPR to errors in overlay link failure probability estimates, we run a set of experiments with inaccurate overlay link failure probabilities. Specifically, we ignore the correlation of link failures in these experiments. This represents an extreme case of inaccurate overlay link failure estimates. The results are shown in Fig. 5. We also run experiments by adding noise into the double overlay link failure probability matrix $\mathbf{P}$. Noise follows Gaussian or uniform distribution with mean 0. FPR based on the inaccurate $\mathbf{P}$ degrades little on robustness because the noises of link failure probabilities are canceled out along a path and then most backup paths found by FPR are not changed.

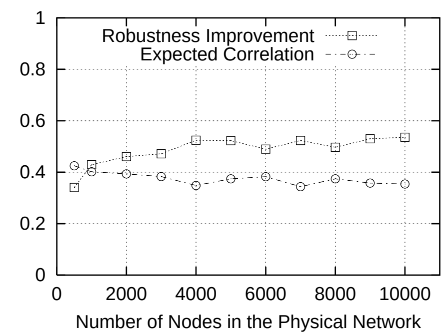


Figure 6. The robustness improvement of FPR+FBA and the expected correlation between primary and backup path failures in overlay networks with 10 nodes.

6.4. Discussion

Simulation results show that FPR achieves high robustness and good efficiency and is tolerant to inaccurate overlay link failure probability estimates. The robustness of FPR is very close to the optimal solution OPR. In particular, FPR's robustness improvement is up to 60% better than using no backup path reservations, and is up to 30% better than SSR. Furthermore, FPR is more efficient than SSR in most cases. While SSR uses fewer links than FPR, these links tend to become bottlenecks and limit the number of backup paths. In contrast, FPR spreads out the backup paths thus reducing the bottlenecks. Furthermore, FPR is robust in the presence of inaccurate overlay link failure estimates. Even if we ignore the correlation of link failures, FPR is still 7% better than SSR in terms of robustness. By considering overlay link failure correlations, the robustness increases by another 8%. This suggests that FPR can significantly improve the performance of overlay networks.

Simulation results also show that we can reduce efficiency loss significantly by using backup path bandwidth sharing. The efficiency losses of SBA or DBA are less than 20%, while the efficiency loss of FBA can be as high as 50% when compared to the case when no backup paths are used. Moreover, we can see that DBA makes a better tradeoff than SBA between robustness and efficiency because DBA is 25% more robust and only 10% less efficient. However, we conjecture that SBA, DBA, and FBA are complementary. Using FPR+FBA, FPR+DBA, FPR+SBA, and NBR+ZBA would make it possible to provide differentiated services to users with different priorities and service requirements.

In our simulations, the best robustness improvement is up to 60% while the upper bound of the robustness improvement is 100% (the maximum happens when FP of the backup path routing and bandwidth allocation algorithms becomes 0). We will use a simple example to explain the reason.

Consider the primary and backup path from node S to D. For simplicity, we assume that the failure probability of the primary path and the backup path are both p, and the joint failure probability of the primary and backup path is q. Thus the failure probability of the connection from node S to D is p when there is no backup path and is q when the backup path is reserved with full backup bandwidth allocation (FBA). Moreover, we define two random variables X and Y. X is 1 if the primary path fails and is 0 otherwise. Y is 1 if the backup path fails and is 0 otherwise. Then we have

$$\begin{aligned} E[X] &= E[Y] = p \\ E[XY] &= q \end{aligned}$$

Let ρ denote the correlation of X and Y. According to Eq. 1, we have

$$\rho = \frac{q - p^2}{\sqrt{p - p^2}\sqrt{p - p^2}} = \frac{q - p^2}{p - p^2} \tag{18}$$

Then we have

$$q = p^2 + \rho(p - p^2) \tag{19}$$

Hence, the robustness improvement (see Eq. 16) is

$$\frac{p - q}{p} = \frac{p - [p^2 + \rho(p - p^2)]}{p} = (1 - \rho)(1 - p) \approx 1 - \rho \tag{20}$$

The last approximation is due to the assumption that link failure probabilities are small (refer to Section 3.2). This demonstrates that the robustness improvement is dependent on the correlation between primary path failures and backup path failures. The less correlated are primary paths and backup paths, the larger the robustness improvement. This confirms the intuition that two "orthogonal" paths are preferred for restoration. In Fig. 6, we show the robustness improvement and the expected correlation between primary and backup path failures in overlay networks with 10 nodes. It is obvious that their relationship follows our deduction made from the simple example above. Thus, the reason we can only achieve up to 60% robustness improvement in our experiments is that the correlation between primary path failures and backup path failures is large in the simulated networks.

7. Conclusions and Future Work

In this paper, we studied the problem of backup path routing and bandwidth allocation in generic overlay networks. The main contributions of this paper are:

- We propose a correlated overlay link failure probability model which reflects the mapping of the overlay network on the physical network topology. In particular, failure probabilities corresponding to two overlay paths that share the same physical link will be highly correlated.

- We use the correlated overlay link failure probability model to formulate the backup path routing problem as an Integer Quadratic Programming (IQP) problem. To efficiently solve this problem we use a new metric (FPC)–which measures the incremental path failure probability caused by using a link in the path–to reduce the IQP problem to a shortest path routing problem.

- We evaluate our solution by using extensive simulations. The results show that, in terms of robustness, our approach is close to the optimal and is up to 60% better than no backup path reservation and is up to 30% better than ignoring link failure probabilities.

In the future, we plan to extend our work in two directions. First, we wish to explore efficient and effective methods for measuring and estimating correlated overlay link failure probabilities. A potential direction is to leverage technologies of detecting shared congestion of overlay paths [16]. Second, we plan to design dynamic backup path routing algorithms that can trade robustness to efficiency rather than simply optimize either robustness or efficiency.

Acknowledgments

We thank the anonymous reviewers for helping us improve the paper. We are very grateful to our colleagues Sharad Agarwal, Matthew Caesar, Adam Costello, Sridhar Machiraju, Bhaskaran Raman, Mukund Seshadri, Lakshminarayanan Subramanian, for their very helpful comments on the paper draft. We would also like to thank Stefan Savage and Andy Collins for sharing their end-to-end measurement traces with us.

References

[1] D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris. Resilient overlay networks. In *Proc. of SOSP 2001*, October 2001.

[2] K. Calvert, M. Doar, and E. W. Zegura. Modeling Internet topology. *IEEE Communications Magazine*, June 1997.

[3] B. Chandra, M. Dahlin, L. Gao, and A. Nayate. End-to-end WAN service availability. In *Proc. of the Third Usenix Symposium on Internet Technologies and System (USITS01)*, March 2001.

[4] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2 edition, 2001.

[5] H. Eriksson. Mbone: The multicast backbone. *Communications of the ACM*, 37(8):54–60, 1994.

[6] C. Fraleigh, S. Moon, C. Diot, B. Lyles, and F. Tobagi. Packet-level traffic measurement from a tier-1 IP backbone. Sprint Technical Report TR-01-110101, 2001.

[7] B. Gleeson, A. Lin, J. Heinanen, G. Armitage, and A. Malis. A framework for IP based virtual private networks. RFC 2764, February 2000.

[8] Genuity. Inc. ATM service level commitments. http://www.genuity.com/services/transport/atm/atm_slc.pdf, 2001.

[9] M. Kodialam and T. V. Lakshman. Dynamic routing of bandwidth guaranteed tunnels with restoration. In *Proc. of INFOCOM 2000*, Tel-Aviv, Israel, March 2000.

[10] M. Kodialam and T. V. Lakshman. Dynamic routing of locally restorable bandwidth guaranteed tunnels using aggregated link usage information. In *Proc. of INFOCOM 2001*, April 2001.

[11] G. Mohan and C. S. R. Murthy. Lightpath restoration in WDM optical networks. *IEEE Network*, 14(6), Nov.-Dec. 2000.

[12] K. Murakami and H. S. Kim. Virtual path routing for survivable ATM networks. *IEEE Trans. on Networking*, 4(1), February 1996.

[13] B. Rajagopalan, J. Luciani, D. Awduche, B. Cain, and B. Jamoussi. IP over optical networks: A framework. Internet Draft draft-ietf-ipo-framework-01.txt, February 2002.

[14] B. Raman and R. H. Katz. Emulation-based evaluation of an architecture for wide-area service composition. In *International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2002)*, July 2002.

[15] J. Rosenberg and H. Schulzrinne. The IETF Internet telephony architecture and protocols. Technical Report, 1999.

[16] D. Rubenstein, J. Kurose, and D. Towsley. Detecting shared congestion of flows via end-to-end measurement. In *Proc. of SIGMETRICS 2000*, June 2000.

[17] S. Savage, A. Colins, E. Hoffman, J. Snell, and T. Anderson. The end-to-end effects of Internet path selection. In *Proc. of SIGCOMM 1999*, August 1999.

[18] A. Schriver. *Theory of Linear and Integer Programming*. Wiley, 1986.

[19] V. Sharma and et al. Framework for MPLS recovery. Internet Draft draft-ietf-mpls-recovery-frmwrk-03.txt, July 2001.

[20] D. Stamatelakis and W. D. Grover. IP layer restoration and network planning based on virtual protection cycles. *IEEE Journal on Selected Areas in Communications*, 18(10), October 2000.