Availability-Oriented Path Selection in Multi-Path Routing

Xin Zhang, Adrian Perrig, and Hui Zhang

August 25, 2007
CMU-CyLab-07-012

CyLab
Carnegie Mellon University
Pittsburgh, PA 15213

Availability-Oriented Path Selection in Multi-Path Routing

Xin Zhang, Adrian Perrig, and Hui Zhang
Carnegie Mellon University

Abstract

Multi-path routing is effective to enhance network availability, by selecting multiple *failure-independent* paths for reaching one destination in the hope to survive individual path failures. Researchers suggest to select IP-layer topologically disjoint paths, assuming that they are failure-independent and can hardly fail simultaneously. Unfortunately, failure correlations lurking behind the IP-layer topology can surreptitiously squash availability gained through multi-path routing, because selected paths can fail simultaneously. Spurred by this observation, we propose a new path metric and selection scheme resilient to failure correlations between topologically disjoint paths, by utilizing path *availability history* to reveal failure correlations. This paper presents a first stride towards the new direction of availability-oriented multi-path selection, with formal and systematic problem definition, modeling, algorithms, and evaluation of our scheme.

1 Introduction

In recent years, *multi-path routing* has been popularly advocated amongst the research community [9, 12, 19, 21, 24–26]. In multi-path routing, each router can use multiple different paths for reaching one destination prefix, with the goal to enhance end-to-end availability: when one of the paths fails, packets can still be delivered via other working paths and thereby maintaining end-to-end availability, *as long as not all paths between the source and destination fail concurrently*. We term such end-to-end availability provided by multiple paths between a source-destination pair as *multi-path availability*. One crucial component in multi-path routing is *multi-path selection*: the decision process of determining which paths to use. Obviously, the selection tactic and resulting path qualities will exercise direct influence on the effectiveness of multi-path routing. While respecting local policies [5] (if any) during the multi-path selection process, in this paper we focus on optimizing multi-path availability, which is the underlying requirement of multi-path routing.

As a new subject, multi-path selection has drawn little attention in the literature so far, despite its importance. Previous proposals on multi-path routing primarily focused on routing infrastructure design. Early investigations on path selection [3, 10, 18] were mostly in the context of conventional single-path routing (where each router is limited to using only a single "best path" for one destination prefix). Wendlandt et al. explicitly recognize multi-path selection as a vital component in multi-path routing in the context of defending against BGP attacks [24]. As a simple approach to handling availability-oriented multi-path selection, it was suggested to choose the most "IP-layer topologically disjoint" paths, in an attempt to minimize the probability that all the paths simultaneously fail [12,26]. While such an approach is simple and intuitive, it relies on the assumption that *IP-layer topologically disjoint paths are indeed failure-independent*. Regretably, this assumption has been suspected repeatedly in the literature [2, 8, 11, 22]. As we investigate in Section 2.2, IP-layer topologically disjoint paths can still be failure-correlated, due to the discrepancies between IP-layer and Physical-layer topologies, i.e., IP-layer disjoint paths can still share the same physical elements. Furthermore, routing-level congestion or common software vulnerabilities can also give rise to failure correlation between even physical-layer disjoint paths.

Motivated by the presence of failure-correlation between topologically disjoint paths, we strive to tackle multi-path selection with resilience to such failure-correlation. The key idea is to base multi-path selection on the knowledge of paths' *availability history*, in light of that failure correlation between paths can be automatically derived from their availability history: if two paths present concurrent failures in the history, we regard them as failure-correlated, otherwise failure-independent. Admitted that two failure-independent paths may also present certain concurrent failures by chance, such *false correlation* has a rather negligible probability to occur, observing that in the real-world link failures are rare [8]. On the other hand, by using the historical failure correlation to predict future ones (Section 7.3), our scheme is most resilient to the types of failures that can repeat in the future, while the effectiveness of our scheme will be hampered in circumstances where failures happen only once. Our underlying rationale for leveraging availability history to exploit failure-correlation lies in that: given the intricate causes to failure-correlation between different paths (Section 2.2), the best we can do is to derive such correlation *after* the failures take place, while it is an open challenge to detect failure-correlation *before* failures happen.

In this paper, we present a first step to study this important and novel topic. Proved by analysis and simulation results, our scheme enables more accurate selection of failure-independent paths compared to using topological disjointness as metric. Our contributions are four-fold:

Problem Definition. This paper raises the awareness of the multi-path selection problem, and formally presents a problem statement.

Metric. We derive a new metric from path availability history to reflect the failure correlation between disjoint paths.

This metric enables more accurate estimation of multi-path availability.

Problem Modeling. We mathematically model multi-path selection as an Integer Programming problem. This precise model facilitates us to better understand the problem, and also enables us to use various well-studied algorithms for Integer Programming.

Algorithm. For the ease of practical deployment, we also suggest a simple heuristic and prove its effectiveness via both theoretical worst-case analysis and simulation.

The remainder of the paper is organized as follows. In Section 2, we introduce the problem statement of multi-path selection and the key observation which motivates our endeavors. Section 3 sketches the high-level picture of our scheme. Then, Sections 4 and 5 elaborate the mathematical modeling and algorithms of the problem, respectively. Section 6 presents performance evaluation results, and Section 7 discusses deployment issues. Finally, Section 8 concludes the paper and presents future work.

2 Problem Statement and Observations

2.1 Problem Statement

For multi-path routing to be effective, it is desired to select paths that can provide the least probability that all the selected paths fail simultaneously, which in turn yields the highest multi-path availability. To this end, we focus on selecting a set of *failure-independent* paths. To sum up, the multi-path selection problem can be stated as follows:

A node Z has n candidate paths to reach destination D. The problem is to find the k most failure-independent paths such that the resulting multi-path availability is maximum, where the value of k is restricted by the communication overhead allowed in the routing infrastructure.

2.2 Why is Disjointness Inaccurate?

In light of the problem statement above, it is clear that an accurate metric for evaluating failure-correlation and multi-path availability forms the prerequisite of a successful multi-path selection scheme. Several proposals suggest to select "disjoint" paths in the IP-layer topology to minimize failure correlation [12,26] between selected paths, based on the intuition that paths overlapping at certain links must suffer from failure correlation introduced by the shared links. While this underlying intuition is correct, the practical effectiveness of such an approach is nevertheless hampered: it has been observed that certain subsets of IP-layer links can be failure-correlated even if they are topologically disjoint, and the popularity of such correlation is not negligible [2, 8, 11]. Therefore, paths can be failure-correlated even when their constitute links are all disjoint in the IP-layer topology.

An important source of causes to such failure correlation has been well recognized as the Shared Risk Link Groups (SRLG) [11, 22]: multiple IP-layer links that are associated with the same physical network element (e.g., fiber span, optical amplifier etc.) will experience failure at the same time when the physical object fails. However, the selection scheme described above intrinsically embeds the inaccuracy of using the *logical* network topology to conclude the *real-world* disjointness of links. More specifically, the IP-layer topology is basically a set of nodes representing routers, interconnected via point-to-point links. Such a logical abstraction is far insufficient to reveal the actual correlation of links at the physical layer. For example, a set of IP-layer disjoint routers may be supplied by the same power source and be "fate sharing", and a set of IP-layer disjoint links can share the same physical fiber paths, span, etc. Even if the links use disparate physical fibers, they are prone to be correlated if the physical fibers are placed in close geographical locations in which case they can be affected by the same exogenous accident. A notorious anecdote is the earthquake off the coast of Taiwan in Dec. 2006, which destroyed all the IP-layer "independent" fibers lying nearby.

In addition, we can also list multiple routing-level or software-related factors that may trigger failure correlation even between physical-layer topologically disjoint links/paths (*behavioral-level fate-sharing*). For example, some routers' software may exhibit vulnerabilities to the same attacks; system administrators may load the same faulty configuration onto different devices; and furthermore, different links can get congested together as well. Figure 1 provides an example of correlated congestion, where Z_1 directs its traffic via both A and B to D. When Z_1 sends large amounts of bursty traffic to D, the links $A \rightarrow D$ and $B \rightarrow D$ will get congested together and become prone to be failure-correlated, although they are topologically disjoint.

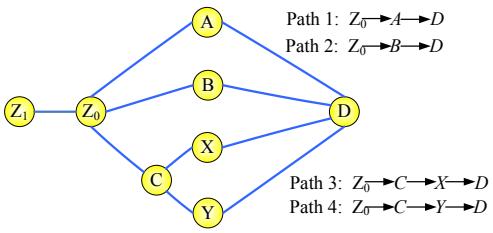


Figure 1: Example topology

In summary, the discrepancies between (a) IP-layer and physical-layer topologies as well as (b) physical-level and behavioral-level fate-sharing give rise to the considerable inaccuracy of using topological disjointness as a gauge of failure independence. This gap motives us to seek better solutions as presented below.

3 Overview of Our Approach

The success of multi-path selection necessitates two components, namely, (a) a *metric* that can accurately reflect failure correlation between different paths, and (b) a *selection algorithm* that can effectively leverage the metric to rule out failure-correlated paths from being selected together. In Section 3.1 we present a new mechanism which can not only evaluate individual path availability, but can also help deriving an accurate multi-path availability metric even in the presence of failure correlation between different paths. Then in Section 3.2, we sketch how the new mechanism helps to derive

a precise multi-path availability metric, and how we base our multi-path selection scheme on the derived metric. Since our scheme relies on the knowledge of link failure history, we assume that *the link failure history is recorded, and each node has access to such data*. In Section 7, we discuss the realization of this assumption in realistic routing infrastructures.

3.1 Availability History Window

Given potentially Byzantine dynamics and interactions in Internet routing, as well as the tremendous complexity of the hardware and software which an IP network relies on, it is difficult (if not impossible) to precisely predict or analyze the correlation between different paths. To bypass such complexity while still exploring the failure correlation between different paths, we propose a new mechanism called *availability history window* (AHW), to record path availability histories, from which the failure correlation between different paths can be learned. In the following, we first define AHW on a per-link basis, from which path (multi-path) availability can be then easily derived.

A straightforward interpretation of an AHW is a 0-1 time history window, where '1' corresponds to the time instant when the link is *available* (working), while '0' corresponds to the time instant when the link is *unavailable* (failed). We define a *stable interval* in an AHW as a *continuous* time interval with duration *longer than* a threshold L, where the link is always available. Conversely, a *failure interval* in an AHW is defined as a *continuous* time interval, within which there is no stable interval.

The top two time series in Figure 2 present the example AHWs of links $A \rightarrow D$ and $Z_0 \rightarrow A$ in Figure 1. According to the definition, note that a short time interval during which a link is available is nevertheless attributed to failure interval, as long as its duration is shorter than L. This makes an unstable path less appealing in the selection algorithm (Section 3.2), and reduces implementation overhead (Section 7). We can also infer that an AHW is *exclusively* composed of a set of *disjoint* stable and failure intervals, i.e., any time epoch in an AHW must be *either* in a stable *or* a failure interval.

So far, AHW is used to characterize individual links. Now we present how to derive an AHW for an entire path consisting of concatenating links or sub-paths, using the following *series combination* operation.

Series Combination. The AHW of a complete path is computed as the logical AND operation of all 0-1 AHWs of the constitute links or sub-paths. For example, the AHW of path 1 in Figure 1 is computed as the AND operation of the AHWs of links $A \rightarrow D$ and $Z_0 \rightarrow A$, as depicted as the third AHW in Figure 2.

3.2 AHW-based Multi-Path Selection

From the availability history carried by AHWs, we can infer that two paths are highly correlated if they tend to fail at the same time in their AHWs, and vice versa. Several algorithms have been proposed to leverage such history records to cluster correlated links into groups [22, 23]. For our purpose of selecting failure-independent routing paths, we derive a multi-path availability metric from AHWs and employ a selection scheme that can *automatically* preclude failure-correlated paths from being selected together, without additional clustering efforts.

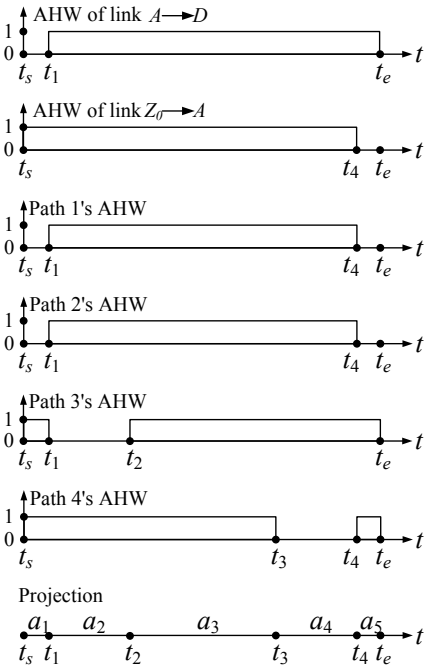


Figure 2: Example AHWs. t_s and t_e represent the start and end times, respectively

Recall that in multi-path routing, we aim at selecting multiple paths that provide the highest multi-path availability, thus we first derive the AHW of a given set of k paths using the following *parallel combination* operation, from which we can eventually compute the *multi-path availability metric*.

Parallel Combination. The AHW of multiple paths between a source-destination pair is computed as the logical OR operation of all AHWs of the paths. For instance in Figure 1, if Z_0 propagates all four paths whose AHWs are given in Figure 2, the resulting AHW is an entire stable interval.

Multi-Path Availability Metric (MAM). It is computed as the duration of all stable intervals in the AHW of multiple paths between a source-destination pair.

Accordingly, our algorithm selects the k AHWs that can produce the largest MAM, by which it can ensure that failure-correlated paths are bound to be less likely chosen together. To illustrate the intuition, consider the example topology in Figure 1 with the AHW of each path given in Figure 2. Suppose path 1 has already been selected, if we further select path 2 which is failure-correlated with path 1, the resulting combined AHW gains no increase in the duration of the stable interval. In contrast, if we parallel-combine path 3 which is failure-independent with path 1, the resulting combined AHW benefits from a significant increase in MAM, thus possessing precedence over path 2 in our selection mechanism.

4 Mathematical Modeling

Following the high-level description above, now we formally present our mathematical modeling of the multi-path selection problem, and ultimately characterize the problem using an Integer Programming formulation. We use the following standard notations: (a) "$\bigcup$" is the UNION operation of sets; (b) "$\vee$" stands for the logical OR operation; and (c) "$|\mathcal{X}|$" operation returns the cardinality of set $\mathcal{X}$. In our context, $|\mathcal{X}|$ refers to the duration time of interval $\mathcal{X}$.

4.1 Mathematical Problem Formulation

First, we formalize AHW, multi-path availability and multi-path selection problem using mathematical notation. Then we prove that the problem is NP-Complete.

Definition 1 An AHW of path i, denoted by $\mathcal{A}_i$, is a 0-1 time series defined as:

$$\mathcal{A}_i\colon t \rightarrow r_i(t) \in \{0,1\},\ \ t \in [t_s, t_e]$$

where t is a historical time epoch in the window range $[t_s, t_e]$; and $r_i(t)$ is the availability record of path i at time t, i.e.:

$$r_i(t) = \begin{cases} 1 & \text{if } t \text{ is in a stable interval of } \mathcal{A}_i, \\ 0 & \text{if } t \text{ is in a failure interval of } \mathcal{A}_i. \end{cases}$$

Recall that $\mathcal{A}_i$ is exclusively composed of interleaving stable and failure intervals. Since the stable intervals are of particular interest for our subsequent modeling purpose, we further have:

Definition 2 Let d_i be the number of stable intervals in $\mathcal{A}_i$, and S_i be the union of all d_i stable intervals, i.e.:

$$S_i = \bigcup s_i^j = \sum s_i^j, \quad j = 1, 2, \ldots, d_i$$

where s_i^j is a single stable interval in $\mathcal{A}_i$.

Definition 3 Let $U = \bigcup S_{i=1,2,\ldots,n}$ be the 1-dimensional universe. Path i's availability is defined as $\theta(i) = |S_i|$. Let M be a set of paths between the same source-destination pair. The multi-path availability of M, denoted by $\theta(M)$, is given by:

$$\theta(M) = |\bigcup S_i|,$$

Definition 4 The multi-path selection problem is defined as a triple of $\langle input, output, objective \rangle$, as follows:

1. An $input$ consists of a family $N = \{S_1, S_2, \ldots, S_n\}$ and an integer k ($\leq n$).
2. An $output$ is a subfamily $M \subseteq N$, with $|M| \leq k$.
3. The $objective$ is to find a subfamily M^* such that:

$$\forall M \neq M^*,\ \ \theta(M^*) \geq \theta(M). \quad (|M| \leq k, |M^*| \leq k)$$

Theorem 1 *The multi-path selection problem in Definition 4 is NP-complete.*

Proof: We prove this theorem by reducing the well-know NP-complete problem "set-covering decision" [6] to our problem. In the set-covering decision problem, we are given a universe U and a family N of subsets of U. A *cover* is a subfamily $M \subseteq F$ whose union is U. The input to the problem is a pair $\langle U, N \rangle$ and an integer k; the question is whether there is a set-covering of size k or less.

In our problem, the universe U is defined as $\bigcup S_i,\ \ i \in N$. A family N is composed of $S_{i=1,2,\ldots,n}$, each of which is a subset of U. The integer k corresponds to the constrained number of paths that we can select. Given the same input as the set-covering decision problem, by solving our problem we can derive the maximum multi-path availability $\theta(M^*)$. If $\theta(M^*) = |U|$, according to Definition 3, we have:

$$\bigcup S_i = U,\ \ i \in M^*$$

Then there definitely exists a cover of U of size k, and the corresponding subfamily is given by M^*. If $\theta(M^*) < |U|$ on the other hand, we can ascertain that there is no set-covering of size k or less, since $\theta(M^*)$ by definition is the maximum value from all possible selections under constraint k. We can thus correctly answer the set-covering problem by solving the problem given in Definition 4. ■

4.2 Optimization Modeling

Based on preceding mathematical definitions, in this section we frame the multi-path selection problem given in Definition 4 as an optimization model and finally evolve it into a '0-1' Integer Programming formulation [4] in Theorem 2.

4.2.1 General Optimization Model

Define $x_i \in \{0,1\}$ as the *indicator variable* for path i, such that:

$$x_i = \begin{cases} 1 & \text{if } i \in M, \\ 0 & \text{if } i \notin M. \end{cases} \tag{1}$$

Then we can formulate the problem as the *optimization model* shown in Equation 2:

$$\begin{aligned} &\textit{Maximize}: \quad \theta(M) = |S_1 x_1 \cup S_2 x_2 \ldots \cup S_n x_n| \\ &\textit{Subject to}: \quad \sum_{i=1}^{n} x_i \leq k,\ \ x_i \in \{0,1\} \end{aligned} \tag{2}$$

4.2.2 Optimization Model with Linear Objective

We first linearize the objective function shown in Equation 2, by the following two steps.

Step 1. Recall that all $S_{i=1,2,\ldots,n}$ constitute a 1-dimensional universe U (Definition 3), which can be interpreted as a time interval. And each $\mathcal{A}_i$ consists of d_i disjoint stable intervals $s_i^1, s_i^2, \ldots, s_i^{d_i}$ (Definition 2). Each stable interval can be identified by its two endpoints in the time interval of U. Now we project all end points of all stable intervals in $\mathcal{A}_{i=1,2,\ldots,n}$ into the time interval of U, and assume there are z end points in total after the projection. Let every two adjacent end points on the number line form a new *atomic interval*, then we obtain $z-1$ atomic intervals denoted by $a_1, a_2, \ldots, a_{z-1}$. Figure 2 gives an example, where the end points t_s, t_1, t_2, t_3, t_4 and t_e

produces five new atomic intervals after projection onto the bottom line in Figure 2.

Before we delve into the second step, we first introduce a definition and highlight an important property of atomic intervals. The property facilitates the transformation towards an Integer Programming model in our subsequent discussion.

Definition 5 An atomic interval a_j is *covered* (*not covered*) by $\mathcal{A}_i$ if a_j is *completely within* a stable (failure) interval of $\mathcal{A}_i$. a_j *intersects* with $\mathcal{A}_i$ when only *part* of a_j is within a stable or failure interval of $\mathcal{A}_i$.

Property 1 *The atomic intervals are all* disjoint. *An atomic interval a_j must be either covered or not covered by path i. It cannot intersect with $\mathcal{A}_i$.*

Step 2. Now we introduce a new indicator variable $y_j \in \{0,1\}$ for atomic interval a_j, such that:

$$y_j = \begin{cases} 1 & \text{if } \exists\, i \in M,\ a_j \text{ is covered by } \mathcal{A}_i, \\ 0 & \text{if } \forall\, i \in M,\ a_j \text{ is not covered by } \mathcal{A}_i. \end{cases} \tag{3}$$

Then according to Property 1, we can transform the objective function of Equation 2 into the linear one below:

$$\textit{Maximize:} \quad \theta(M) = \sum_{j=1}^{z-1} |a_j| \cdot y_j \tag{4}$$

Be aware of that, the import of y_j is also accompanied by new constraints, i.e., the restricting relationship between $x_{i=1,2,\ldots,n}$ (Equation 1) and $y_{j=1,2,\ldots,z-1}$. We first introduce the concept of a *covering set* of an atomic interval as follows.

Definition 6 For an atomic interval a_j, its *covering set* $C(j)$ is the set of all $\mathcal{A}_i$ each of which covers a_j. Denote the indices in $C(j)$ as $p_1, p_2, \ldots, p_{|C(j)|}$. Then we have, $\forall p_i \in C(j)$, a_j is covered by $\mathcal{A}_{p_i}$.

Now also according to Property 1, we can express the relationship between $x_{i=1,2,\ldots,n}$ and $y_{j=1,2,\ldots,z-1}$ as follows:

$$y_i = x_{p_1} \vee x_{p_2} \vee \ldots \vee x_{p_{|C(j)|}}, \quad p_i \in C(j) \tag{5}$$

Take Figure 2 for instance (assume $k = 2$), we have:

$$\begin{aligned} y_1 &= x_3 \vee x_4, \\ y_2 &= x_1 \vee x_2 \vee x_4, \\ y_3 &= x_1 \vee x_2 \vee x_3 \vee x_4, \\ y_4 &= x_1 \vee x_2 \vee x_3, \\ y_5 &= x_3 \vee x_4. \end{aligned} \tag{6}$$

4.2.3 Integer Programming Model

Now we remove the logical OR operation in the constraints in Equation 5 to finally linearize the problem into an Integer Programming formulation.

Lemma 1 *The constraint $y_j = x_{p_1} \vee x_{p_2} \ldots \vee x_{p_{|C(j)|}}$ in Equation 5 is equivalent to $y_j \leq x_{p_1} + x_{p_2} .. + x_{p_{|C(j)|}}$.*

Proof: We prove this theorem by showing that in any case, the values of y_j yielded by both expressions are equal.

1. When $\forall p_i \in C(j),\ x_{p_i} = 0$, both expressions yield the same value, i.e., $y_j = x_{p_1} \vee x_{p_2} \ldots \vee x_{p_{|C(j)|}} = 0$, and $y_j \leq x_{p_1} + x_{p_2} .. + x_{p_{|C(j)|}} = 0 \Rightarrow y_j = 0$.

2. When $\exists p_i \in C(j),\ x_{p_i} = 1$, let w be the number of p_i that $x_{p_i} = 1$. Then $y_j = x_{p_1} \vee x_{p_2} \ldots \vee x_{p_{|C(j)|}} = 1$. On the other hand, $y_j \leq x_{p_1} + x_{p_2} .. + x_{p_{|C(j)|}} = w$. Note that $y_j \in \{0,1\}$, now y_j can take either the value '0' or '1' under this inequality constraint ($y_j \leq w$). Yet, observe that the coefficient of y_j in the objective function (Equation 4) is positive, and we are maximizing the value of objective function. So given the constraint $y_j \leq l$, in the optimal solution there must be $y_j = w$.

Therefore in any case, both expressions render the same value to y_j. This proves the theorem. ■

Theorem 2 *The multi-path selection problem given in Definition 4 can be modeled as the following 0-1 integer programming formulation.*

$$\begin{aligned} \textit{Maximize:} \quad & \theta(M) = \sum_{j=1}^{z-1} |a_j| \cdot y_j \\ \textit{Subject to:} \quad & \sum_{i=1}^{n} x_i \leq k,\ \ x_i \in \{0,1\},\ \ y_j \in \{0,1\}, \\ & y_j \leq \sum_{j=p_1}^{p_{|C(j)|}} x_i,\ \ j = 1, 2, \ldots, z-1 \end{aligned} \tag{7}$$

The proof of this theorem is straightforwardly given by Lemma 1. Till now, we eventually formulate the problem as a standard 0-1 Integer Programming.

5 Algorithm

In this section, we start with the solutions directly following the Integer Programming model, and then propose and analyze a simple yet efficient heuristic for practical deployment.

5.1 Integer Programming Solution

The model we gave in Section 4 is a precise mathematical formulation which helps us to thoroughly understand the problem. This model also enables us to leverage various well-established algorithms to solve the Integer Programming problem, such as branch-and-bound and cutting plane methods [4], or some efficient approximation solutions [1,14]. The entire algorithm is described in Table 1.

5.2 Simple Heuristic

The algorithm given above may be computationally complex, depending on what Integer Programming algorithm is adopted in Line 6, Table 1. Also note that in the current Internet, a router can already have hundreds of neighbors, thus possibly dozens of different paths for one destination. In multi-path routing where each neighbor can announce more than one path for a destination, this number will be even

Table 1: Multi-path selection algorithm using I.P. model

line	action
1	**for** each destination D in the network
2	*get* all the n candidate paths associated with AHWs
3	*get* z end points and $z-1$ atomic intervals
4	**for** each atomic interval a_j
5	*compute* $\|a_j\|$ and $C(j)$
6	*call* I.P. procedure with input $\|a_j\|$ and $C(j)$

larger. Therefore, a simple brute-force method for this NP-Complete problem (Theorem 1) with such an input size described above can be computationally prohibitive as well.

In this subsection we present a simple greedy algorithm which can efficiently produce an approximate result. Basically, in each iteration the algorithm greedily selects the path that can maximize the multi-path availability accumulated so far. Table 2 describes the algorithm. Suppose the number of stable intervals in each AHW is bounded by a small constant, then the complexity for selecting k out of n paths for one destination is at most $O(nk)$.

Table 2: Greedy Algorithm for multi-path selection.

line	action
1	**for** each destination D in the network
2	*get* all the n candidate paths associated with AHWs
3	*initialize* $M=\emptyset, \theta(M)=\emptyset$)
4	//begin loops for greedy selection **while** $\|M\| \le k$ and $\theta(M) \neq \|U\|$
5	*select* a path p that maximizes $\|\theta(M) \cup \theta(p)\|$
6	*add* p to M, *update* $\theta(M) = \theta(M) \cup \theta(p)$

Theorem 3 *Let $\theta(M^*)$ and $\theta(M)$ be the multi-path availabilities yielded by the optimal solution and the greedy heuristic given in Table 2, respectively. Then $\theta(M)$ has the lower-bound*

$$\theta(M) \ge \theta(M^*) - \left(1-\frac{1}{k}\right)^{k-1}(\alpha_1-\beta_1)$$

where k is the number of paths that we can select, α_1 is the union of time intervals that is not covered by the first greedily selected path but covered in $\theta(M^*)$, and β_1 is the union of time intervals that is covered by the first greedily selected path but not covered in $\theta(M^*)$.

Proof: Let $\theta(M)_i$ be the multi-path availability produced by the first i selected paths in the greedy algorithm ($\theta(M)_k = \theta(M)$), and Δ_i be the difference between $\theta(M^*)$ and $\theta(M)_i$ (i.e., $\Delta_i = \theta(M^*) - \theta(M)_i$). In the following we prove the theorem by deriving the upper bound of Δ_k (i.e., $\theta(M^*) - \theta(M)$).

Let α_i be the union of time intervals in the universe U that is covered by $\theta(M^*)$ but not covered by $\theta(M)_i$; and conversely β_i be the union of time intervals in the universe U that is covered by $\theta(M)_i$ but not covered by $\theta(M^*)$. Accordingly, we have:

$$\Delta_i = \alpha_i - \beta_i \ \ (\ge 0) \tag{8}$$

Now we start from the beginning. After greedily selecting the first path (the one with the highest individual availability), we have $\Delta_1 = \alpha_1 - \beta_1$. Since by definition, α_1 is covered in $\theta(M^*)$ (by the k paths in K^*). This means that there must exist a single path (say p) of the totally n paths that can cover $\frac{1}{k}\alpha_1$ (otherwise any k paths cannot cover α_1). Thus if p is selected as the second path by the greedy algorithm, $\theta(M)_2$ can be increased by at least $\frac{1}{k}\alpha_1$. And according to the greedy nature of the heuristic, by selecting the second path the increase in $\theta(M)_2$ must be at least $\frac{1}{k}\alpha_1$, i.e.:

$$\theta(M)_2 \ge \theta(M)_1 + \frac{1}{k}\alpha_1$$

Then Δ_2 must be reduced by at least $\frac{1}{k}\alpha_1$, i.e.:

$$\Delta_2 \le \Delta_1 - \frac{1}{k}\alpha_1$$

Analogously, after greedily selecting the i^{th} path, we have:

$$\Delta_i \le \Delta_{i-1} - \frac{1}{k}\alpha_{i-1} \tag{9}$$

From Equation 8, we have $\Delta_i \le \alpha_i$, yielding:

$$\Delta_i \le \Delta_{i-1} - \frac{1}{k}\alpha_{i-1} \le \Delta_{i-1} - \frac{1}{k}\Delta_{i-1} = \left(1-\frac{1}{k}\right)\Delta_{i-1} \tag{10}$$

By resolving this iteration finally we get:

$$\Delta_k \le \left(1-\frac{1}{k}\right)^{k-1}\Delta_1 = \left(1-\frac{1}{k}\right)^{k-1}(\alpha_1-\beta_1) \tag{11}$$

Since $\theta(M) = \theta(M^*) - \Delta_k$, this proves the theorem. ■

Corollary 2 *As k increases, the lower-bound of $\theta(M)$ is also increased, with the limit $\lim_{k\to\infty}\theta(M) = \theta(M^*) - \frac{1}{e}(\alpha_1-\beta_1)$.*

Proof: Let δ_k be the upper-bound of Δ_k. First we prove $\theta(M)$ is increasing with k by showing that δ_k is decreasing with k. According to Equation 11, we have:

$$\begin{aligned}\frac{\delta_k}{\delta_{k+1}} &= \frac{(1-\frac{1}{k})^{k-1}}{(1-\frac{1}{k+1})^k} = \frac{(\frac{k-1}{k})^{k-1}}{(\frac{k}{k+1})^k} = \left(\frac{k^2-1}{k^2}\right)^{k-1}\frac{k+1}{k} \\ &= \left(1-\frac{1}{k^2}\right)^{k-1}\left(1+\frac{1}{k}\right)\end{aligned} \tag{12}$$

Since $(1-x)^n \ge 1-nx$, we have:

$$\frac{\delta_k}{\delta_{k+1}} = \left(1-\frac{1}{k^2}\right)^{k-1}\left(1+\frac{1}{k}\right) \ge 1+\frac{1}{k^3} \ge 1 \tag{13}$$

This proves δ_k is decreasing. Then we prove the limit. Leveraging $\lim_{m\to\infty}(1+\frac{1}{m})^m = e$, we have:

$$\begin{aligned}\lim_{k\to\infty}\delta_k &= \lim_{k\to\infty}\left(1-\frac{1}{k}\right)^k \cdot \frac{\alpha_1-\beta_1}{1-\frac{1}{k}} \quad \text{(Equation 11)} \\ &= (\alpha_1-\beta_1)\cdot\lim_{k\to\infty}(1+\frac{1}{-k})^{-(-k)} \\ &= (\alpha_1-\beta_1)\cdot\frac{1}{\lim_{k\to\infty}(1+\frac{1}{-k})^{-k}} \\ &= (\alpha_1-\beta_1)\cdot\frac{1}{e}\end{aligned} \tag{14}$$

Since $\theta(M) = \theta(M^*) - \Delta_k$, this proves the theorem. ■

6 Performance Evaluation

As the first study in this research direction, performing a real-world evaluation of our scheme is out of scope for this paper, due to the absence of precise real-world measurement data of individual link failures[1]. The exact failure correlation degree between disjoint links is also left unstudied in the literature thus far (to the best of our knowledge). While a thorough and realistic evaluation is deferred as future work, in this paper we conduct simplified simulations based on realistic topologies and synthetic failure data. By using different types of real-world topologies and varying key parameters that affect the effectiveness of our scheme, the results can provide useful insight into the strengthens and limitations of our scheme in various circumstances. We also summarize and analyze some important characteristics of our mechanism from the results.

6.1 Methodology

In the absence of real-world per-link failure data, we synthesize link failures and correlations, vary the correlation degrees over a wide range, and apply them to real network topologies. The simulation setup and workflow are described below.

Synthesized Failure and Correlation. Suppose between a source-destination pair there are q links in total, denoted by $l_1, l_2, \ldots, l_q$. We introduce three parameters to feature link failures and correlation between them, i.e.,

1. the *individual link failure* $\lambda_i \in [0, 1]$ to specify the failure percentage in l_i's AHW;
2. a *correlation matrix*, denoted by $[\rho_{i,j}]_{q\times q}$, to specify the failure correlation between links, where $\rho_{i,j} = 1$ if l_i and l_j are failure-correlated, otherwise $\rho_{i,j} = 0$; and
3. a *correlation degree*, denoted by $\mu \in [0, 1]$, to specify the percentage of failure-correlated link pairs among all the link pairs.

Note that by restricting $\rho_{i,j} \in \{0, 1\}$, we essentially use a *strict correlation model* where two links have either completely the same failure intervals (strict correlation) or no concurrent failure at all (strict independence). Admitted that two independent links may also have a fraction of simultaneous failures, nevertheless our simplification can be justified by the observation that the real-world link failures are rare, therefore the probability of concurrent failures of independent links is negligible compared to the values of μ used in our simulation. Accordingly, if $\rho_{i,j} = 1$, then we set $\lambda_i = \lambda_j$.

Real-World Topologies and Multi-Path. We use multiple inter- and intra-domain topologies for our simulation. We acquire AS-level inter-domain topologies from Routeviews [7], and collect intra-domain topologies from Rocketfuel [20], over two large ISPs, i.e., ISP 1239 and ISP 1221. From these topologies, we select various source-destination pairs to run the simulation. To inspect the effectiveness and sensitivity of our scheme in different circumstances, here we intentionally present the results from four different types of topologies, as summarized in Table 3. (The paths in ISP-Deep are of 5 or 6 hops, and in ISP-Wide are of 2 or 3 hops.)

Table 3: Topology Characteristics

Label	Type	# of Links	# of Paths
AS-Small	Inter-domain	26	13
AS-Large	Inter-domain	129	79
ISP-Deep	Intra-ISP 1221	268	698
ISP-Wide	Intra-ISP 1239	322	296

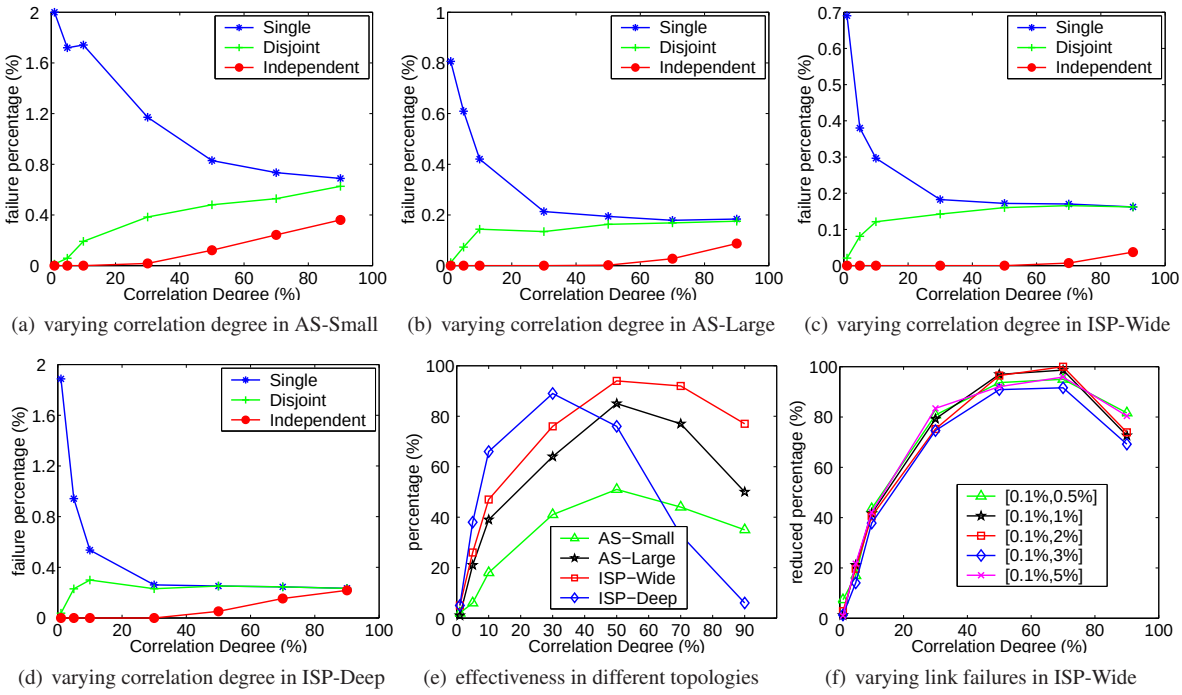
Simulation Workflow. Given the four source-destination pairs selected above, we first extract all the paths between them, randomly set failure-correlated link pairs $\rho_{i,j}$, and assign individual link failures λ_i onto the links. Since the performance of our mechanism is dominated by the failure correlation degree, in our simulation, we intentionally vary the value of μ over a wide range from 1% to 90%, to study the trend of resulting availabilities as μ increases. We first set the value of λ_i uniformly distributed in [0.1%, 3%] for most of the simulations, and then show the sensitivity of resulting availabilities to λ_i. In our simulation, we set $k = 2$ so that only 2 paths can be selected for one destination. We first select the path with the highest individual availability as a default path. Subsequently, we select the second paths using disjointness and AHW respectively. Finally we compute and compare the resulting path availabilities (in our simulation we directly compute and compare the failure percentage for ease of presentation). For each source-destination pair, we run the simulation for 100 rounds for each value of μ.

6.2 Results and Traits

Figures 3(a) to 3(d) plot the results from four different topologies, where labels "Single", "Disjoint" and "Independent" correspond to the results produced by the best single path alone, with the most disjoint path, and with the most independent path according to AHWs, respectively. Figure 3(e) depicts the percentage of simulation rounds when "Independent" achieves a smaller failure percentage than "Disjoint" in different topologies. Figure 3(f) compares the effectiveness of "Independent" with different individual link failures in ISP-Wide topology, where we compute how much *more* failure percentage of "Single" can be reduced by using "Independent" compared to using "Disjoint ".

Form Figures 3(a) to 3(e), we can see that with a low correlation degree (e.g., $\phi \leq 1\%$), the disjointness method can produce a good improvement over the single best path. However, our scheme presents an advantage in any case. Particularly, within a practical range of correlation degree (say, less than 10% or 20% in the real world), our scheme presents remarkably increasing benefits. As the correlation degree approaches 1, the failure percentage of "Disjoint" is close to that of "Single", because in this extreme case disjoint links are all

[1]Recent measurement studies mostly focus on path delay, which is fundamentally different from path failure (availability). Furthermore, current de facto routing protocols (such as BGP or OSPF) only support using a single path for reaching one destination at one time (single-path routing), thus the previously measured path qualities do not provide *concurrent* states of multiple paths between the same source-destination pair which are required to infer failure correlations.



(a) varying correlation degree in AS-Small (b) varying correlation degree in AS-Large (c) varying correlation degree in ISP-Wide

(d) varying correlation degree in ISP-Deep (e) effectiveness in different topologies (f) varying link failures in ISP-Wide

Figure 3: Experimental Results. Labels "Single", "Disjoint" and "Independent" correspond to the results produced by the best single path alone, with the most disjoint path, and with the most independent path according to AHWs, respectively.

failure-correlated thus produce no gain in multi-path availability. From Figure 3(f), we validate that individual link failures have minor impact on the effectiveness of our scheme. In addition, we further summarize the following interesting traits from the results, and give explanations.

Trait 1. As the correlation degree increases, the failure percentage of "Single" decreases. This is because in our model the single path's availability is computed using AND operation (Section 3.1), thus its failure percentage is computed as the *sum* of all *distinct* failure intervals of constitute links; when the consisting links are heavily failure-correlated (with similar failure time distribution), the number of distinct failure intervals decreases, so does the resulting failure percentage of the entire path.

Trait 2. As the correlation degree increases, the failure percentages of both "Disjoint" and "Independent" increase, which is intuitive. However when the correlation degree is high, at some point the failure percentages of "Disjoint" and "Independent" may decrease, which is because the failure percentage of a single path will decrease as shown in Trait 1. This also serves as the reason why when the correlation degree is high, "Independent" performs better in ISP-Wide than ISP-Deep, since in ISP-Deep there introduces more AND operation due to the longer path lengths.

7 Case Study and Discussion

Thus far, we have presented the problem statement, modeling and algorithms of multi-path selection mostly from a conceptual and mathematical perspective. In this section we deal with a set of questions that arise in a deployment of our mechanism, through two case studies in the context of intra- and inter-domain routing. We also discuss several requirements, challenges and tentative countermeasures during the course of a real-world deployment of our scheme.

7.1 Link-State Intra-domain Routing

AHW Acquisition and Propagation. In a link-state routing protocol such as OSPF, each node can monitor its adjacent links and generate corresponding AHWs. An AHW of a link can be encapsulated in each Link State Announcement (LSA) and propagated along with LSAs. In this manner, every node has the knowledge of complete network topology with each link annotated with its AHW, from which the paths' AHWs can be locally computed by each node via the operations defined in Section 3. Considering that AHWs may not need to be as frequently updated as LSAs, AHWs can also be propagated out-of-band: the network administrator can deploy a publicly accessible *availability database* to gather and store AHWs from the network. This alternative can relieve each node of piggybacking an AHW onto every LSA. In intra-domain routing within a network, privacy is not a common concern and thus the required link failure history can be acquired.

It may be expensive to straightforwardly propagate the naive time series of an AHW. Since an AHW is composed of a 0-1 binary sequence, and failure intervals are relatively rare and usually condensed into a short time range [8], the time series can be efficiently compressed (e.g., using run-length encoding). Since failures are likely to be bursty within a short time which can produce short interleaving failure and stable intervals, we regard such an entire instable interval as one sin-

gle failure interval (Section 3.1) to eliminate encoding overhead. This also provides more reasonable estimation because instable intervals can cause routing oscillation which may be even worse than link failures.

Global Selection. By using multi-path availability as the optimization objective, upstream and downstream nodes along the same path can have different "best" selections. To illustrate, consider the example topology in Figure 1 and suppose link $Z_1 \rightarrow Z_0$'s AHW is exactly the same as path 1's shown in Figure 2. For Z_0, paths 3 and 4 produce the optimal multi-path availability. However when combined with link $Z_1 \rightarrow Z_0$ using series combination, Z_1 may find paths 1 and 2 are the "best" choice. In this scenario, Z_1 needs to "notify" its next hop Z_0 about Z_1's preference[2], or simply selects from the path set that will be used by Z_0.

7.2 Path-Vector Inter-domain Routing

AHW Propagation. In a path-vector routing protocol such as BGP, each Autonomous System (AS) prepends its own AS number into the ASPath received downstream, and the new path is propagated upstream. Analogously, the AHW for an entire path can also be aggregated and propagated in such a manner along with the ASPath announcement. More specifically, each AS originally records and stores the AHWs for its adjacent links. Note that for reaching different destination prefixes, an AS composed of multiple routers can use different links (border routers) through a neighbor, and can use different intra-AS paths from its incoming point to outgoing point, thus presenting different availability states. Therefore, AHW needs to be recorded on a per-prefix basis. After receiving from the downstream neighbor the ASPath announcement coupled with the downstream path's AHW, the AS combines the local link's AHW with the downstream path's AHW using series combination. Only the newly aggregated AHW is further propagated to the upstream neighbor. For example in Figure 1, after receiving the downstream AHW of link $X \rightarrow D$ from X, C combines it with its local link $C \rightarrow X$'s AHW using the series combination. Then only this newly aggregated AHW of $C \rightarrow X \rightarrow D$ is reported to the upstream node Z_0, while the AHW of link $X \rightarrow D$ is hidden. In this way, each ASPath is associated with exactly one AHW in the announcement; and each AHW can be efficiently compressed as described in Section 7.1.

The inter-domain environment also brings forth incentive issues. It is apparent that a link/path with "superior" AHW is advantageous to attract traffic. In the AHW propagation mode discussed above, each node is supposed to *self-report* its local links' AHWs. In the competing inter-domain setting, however, an AS with bad performance may not have the incentive to honestly report its poor links. This is an *open challenge* to stimulate ASes to deploy the new metric with incentives [13] and to ensure the metric integrity. As alternatives of the self-reporting mode above to mitigate such concerns, (a) the availability history can be recorded by various third-party network monitoring mechanisms [16, 17, 27], and be accessed from central public databases [2] or be disseminated via an information plane [15], at the expense of setting up those infrastructures and of guaranteeing the involved entities are honest.

Global Selection. In traditional path-vector routing protocol, each node only announces to neighbors the paths that it uses by itself. Therefore there also exists a risk that Z_1 cannot even learn the presence of paths 1 and 2. In this case, Z_1 may proactively inform its downstream neighbor Z_0 of the AHW of local link $Z_1 \rightarrow Z_0$, hoping that Z_0 can announce the paths preferred by Z_1 according to link $Z_1 \rightarrow Z_0$'s AHW. This requires to establish a negotiation phase between the upstream and downstream nodes, which can be seamlessly supported by MIRO [25], a recently proposed multi-path routing protocol.

7.3 Requirements and Cautions

Correlation Predictability. Observe that, by utilizing availability *history*, in effect we only directly learn the *historical* failure correlation between different paths, and we essentially use such historical failure correlation to steer the *future* multi-path selections. In a realistic environment, we posit that the failure correlation between paths can remain invariant within a reasonably long time range, if the update frequency and time scale of AHW are appropriately chosen. Our scheme is most effective in environments where failures repeat in the future.

Time Synchronization. Recall that in our scheme, the failure correlation between paths is indicated by the *time correlation* between failure intervals in the AHWs; and our selection algorithm needs to take the end points (time epochs) of stable and failure intervals as input. Hence, to accurately exploit the failure correlation from AHWs, it is required that all AHWs involved in the selection must be time-synchronized.

Isolation for Stability. It is an open challenge to almost all path selection schemes that use metric-based optimization: at a certain time instance, all the nodes direct their traffic via the path with the currently optimal metric; overwhelmed by the traffic, the path's metric may then degrade and become suboptimal, which in turn causes the main volume of traffic switch to another new optimal path. In this way, network traffic can fluctuate between paths and cause instability. Using availability as metric, however, this problem can be alleviated. Firstly, the availability metric is relatively insensitive to the volume of traffic, meaning that it will not immediately degrade as the traffic volume increases, compared to bandwidth and delay. Secondly, due to the construction of history *window*, it takes a suboptimal path considerable time to become optimal.

8 Conclusion and Future Work

In this paper we launch an initial investigation in availability-oriented multi-path selection. We propose a new way to detect and avert failure-correlated paths by recording and utilizing path/link availability history. Through the mathematical modeling, we first realize that the general multi-path selection problem is NP-Complete. Then leveraging the real-world traits of the problem, we propose a simple yet efficient ap-

[2]Luckily, several proposed multi-path routing protocols can already support such "notification" functionality [9, 21, 25, 26].

proximation heuristic with a proven bound on accuracy and experimental evaluation. Through the case studies of real-world deployment, we show that our scheme is particularly beneficial in intra-domain settings, while the effectiveness in inter-domain environments may be limited by incentive issues. However, we also propose countermeasures to improve incentives. We believe availability-oriented multi-path selection is a promising research direction, and AHW mechanism can be potentially used for network planning as well. We hope this paper can motivate future endeavors in this direction. As following work, we plan to perform a dedicated measurement on real-world failure correlation and inspect the real-world effectiveness of the proposed scheme. We can also study how to seamlessly integrate the AHW-based mechanism into multi-path routing protocols.

References

[1] N. Alon and A. Srinivasan. Improved parallel approximation of a class of integer programming programming problems. In *ICALP: Proceedings of International Colloquium on Automata, Languages and Programming*, 1996.

[2] D. Andersen, A. Snoeren, and H. Balakrishnan. Best-path vs. multi-path overlay routing. In *Internet Measurement Conference*, 2003.

[3] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao. Overview and Principles of Internet Traffic Engineering. RFC 3272 (Informational), May 2002.

[4] Bradley. *Applied Mathematical Programming, Chapter 9*. Addison-Wesley, 1977.

[5] M. Caesar and J. Rexford. BGP routing policies in ISP networks. *IEEE Network Magazine*, Special issue on interdomain Routing, Dec 2005.

[6] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms (Second Edition), pp.1033*. 2001.

[7] http://www.routeviews.org/. University of oregon route views project.

[8] G. Iannaccone, C. Chuah, R. Mortier, S. Bhattacharyya, and C. Diot. Analysis of link failures in an ip backbone. In *Proc. of ACM SIGCOMM Internet Measurement Workshop*, 2002.

[9] H. T. Kaur, S. Kalyanaraman, A. Weiss, S. Kanwar, and A. Gandhi. Bananas: An evolutionary framework for explicit and multipath routing in the internet. In *ACM SIGCOMM Future Directions in Network Architecture*, 2003.

[10] A. Khanna and J. Zinky. The revised ARPANET routing metric. *SIGCOMM Computer Communincation Review*, 1989.

[11] R. Kompella, J. Yates, A. Greenberg, and A. Snoeren. IP fault localization via risk modeling, May 2005.

[12] N. Kushman, S. Kandula, D. Katabi, and B. M. Maggs. R-BGP: Staying Connected In a Connected World. In *Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, April 2007.

[13] P. Laskowski and J. Chuang. Network monitors and contracting systems: competition and innovation. In *ACM SIGCOMM*, 2006.

[14] C.-J. Lu. A deterministic approximation algorithm for a min-max integer programming problem. In *SODA: Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, 1999.

[15] H. Madhyastha, T. Isdal, M. Piatek, C. Dixon, T. Anderson, A. Krishnamurthy, and A. Venkataramani. iplane: an information plane for distributed services. In *Proceedings of USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2006.

[16] V. Padmanabhan, L. Qiu, and H. Wang. Server-based inference of internet performance, 2002.

[17] V. Paxson. End-to-end routing behavior in the Internet. In *Proceedings of ACM SIGCOMM*, 1996.

[18] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. E. Anderson. The end-to-end effects of internet path selection. In *Proc. of ACM SIGCOMM*, 1999.

[19] A. Snoeren, K. Conley, and D. Gifford. Meshbased content routing using XML, 2001.

[20] N. Spring, R. Mahajan, and D. Wetherall. Measuring isp topologies with rocketfuel, 2002.

[21] I. Stocia and H. Zhang. Lira: An approach for service differentiation in the internet. In *Proceedings of Nossdav*, 1998.

[22] J. Strand, A. L. Chiu, and R. Tkach. Issues for routing in the optical layer. *IEEE Communications Magazine*, 39(2):81–87, Feb. 2001.

[23] A. Tachibana, S. Ano, T. Hasegawa, M. Tsuru, and Y. Oie. Locating congested segments on the Internet by multiple paths' delay performance clustering. *IEEE International Conference on Communications*, 2007.

[24] D. Wendlandt, I. Avramopoulos, D. Andersen, and J. Rexford. Don't secure routing protocols, secure data delivery. In *Proc. of ACM Workshop on Hot Topics in Networks (Hotnets-V)*, Nov. 2006.

[25] W. Xu and J. Rexford. MIRO: Multi-path Interdomain ROuting. In *ACM SIGCOMM*, 2006.

[26] X. Yang and D. Wetherall. Source Selectable Path Diversity via Routing Deflections. In *ACM SIGCOMM*, 2006.

[27] Y. Zhao, Y. Chen, and D. Bindel. Towards unbiased end-to-end network diagnosis. *SIGCOMM Comput. Commun. Rev.*, 36(4):219–230, 2006.