Characterization of Failures in an IP Backbone

Athina Markopoulou[†], Gianluca Iannaccone[‡], Supratik Bhattacharyya[§], Chen-Nee Chuah[¶], Christophe Diot[‡]

[†]EE Department, Stanford Univ.,CA, USA
[§]Sprint ATL, Burlingame, CA, USA
[¶]ECE Department, UC Davis, CA, USA
[‡]Intel Research, Cambridge, UK

***Abstract*— We analyze IS-IS routing updates from Sprint's IP network to characterize failures that affect IP connectivity. Failures are first classified based on probable causes such as maintenance activities, router-related and optical layer problems. Key temporal and spatial characteristics of each class are analyzed and, when appropriate, parameterized using well-known distributions. Our results indicate that 20% of all failures is due to planned maintenance activities. Of the unplanned failures, almost 30% are shared by multiple links and can be attributed to router-related and optical equipment-related problems, while 70% affect a single link at a time. Our classification of failures according to different causes reveals the nature and extent of failures in today's IP backbones. Furthermore, our characterization of the different classes can be used to develop a probabilistic failure model, which is important for various traffic engineering problems.**

I. Introduction

The core of the Internet consists of several large networks (often referred to as backbones) that provide transit services to the rest of the Internet. These backbone networks are usually well-engineered and adequately provisioned, leading to very low packet losses and negligible queuing delays [1], [2]. This robust network design is one of the reasons why the occurrence and impact of failures in these networks have received little attention. The lack of failure data from operational networks has further limited the investigation of failures in IP backbones. However, such failures occur almost everyday [3] and an in-depth understanding of their properties and impact is extremely valuable to Internet Service Providers (ISPs).

In this paper, we address this deficiency by analyzing failure data collected from Sprint's operational IP backbone. The Sprint network uses an IP-level restoration approach for safeguarding against failures with no protection mechanisms in the underlying optical fiber infrastructure [4]. Therefore, problems with any component at or below the IP layer (e.g., router hardware/software failures, fiber cuts, malfunctioning of optical equipment, protocol misconfigurations) manifest themselves as the loss of connectivity between two directly connected routers, which we refer to as an IP link failure.

IS-IS [5] is the protocol used for routing traffic inside the Sprint network. When an IP link fails, IS-IS automatically recomputes alternate routes around the failed link, if such routes exist. The Sprint network has a highly meshed topology to prevent network partitioning even in the event of widespread failures involving multiple links. However, link failures may still adversely affect packet forwarding. While IS-IS recomputes alternate routes around a failure, packets may be dropped (or caught in a routing loop) by routers that lack up-to-date forwarding information. Moreover, when traffic fails over to backup paths, links along that path may get overloaded leading to congestion and eventually to packet loss [6].

In this work, we collect IS-IS routing updates from the Sprint network using passive listeners installed at geographically diverse locations. These updates are then processed to extract the start-time and end-time of each IP link failure. The data set analyzed consists of failure information for all links in the continental US from April to October 2002.

The first step in our analysis is to classify failures into different groups according to their underlying cause, i.e. the network component that is responsible. This is a necessary step for developing a failure model where the faults of each component can be addressed independently. In our classification, we proceed as follows. First, link failures resulting from scheduled maintenance activities are separated from unplanned failures. Then, among the unplanned failures, we identify shared failures, i.e. failures on multiple IP links at the same time due to a common cause. Among shared link failures, we further distinguish those that have IP routers in common and those that have optical equipment in common. The remaining failures represent individual link failures, i.e. faults that affect only one link at a time. For the individual failures, we further differentiate groups of links, based on the number of failures on each link.

The second step in our analysis is to provide the spatial and temporal characteristics for each class separately, e.g., the distributions of the number of failures per link, time between failures, time-to-repair, etc. When possible, we provide parameters for these characteristics using well-known distributions.

Our results indicate that 20% of all failures can be attributed to scheduled network maintenance activities. Of the remaining unplanned failures, 30% can be classified as shared. Half of the shared failures affected links connected to a common router, pointing to a router-related problem; the rest affect links that share optical infrastructure, indicating an optical layer fault. The remaining 70% of the unplanned failures are individual link failures caused by a variety of problems. Interestingly, the failure characteristics of individual links vary widely- less than 3% of the links in this class contribute to 55% of all individual link failures.

This work was conducted when the authors were affiliated (or in collaboration) with Sprint ATL. Email addresses: A. Markopoulou - amarko@stanford.edu, G. Iannaccone - gianluca.iannaccone@intel.com, S. Bhattacharyya - supratik@sprintlabs.com, C. N. Chuah - chuah@ece.udavis.edu, C. Diot - christophe.diot@intel.com

The original contributions of the paper are as follows:

- We perform an in-depth analysis of IS-IS failure data from a large operational backbone. This has not been attempted before, largely due to the lack of availability of such data sets.
- We classify failures based on their causes. This methodology can enable an ISP to isolate failures attributable to a cause and pinpoint areas of improvement. For example, this approach can help determine whether a significant number of failures are related to optical-layer problems and identify optical components that should potentially be updated.
- We provide the statistical characteristics of each class of failures and, when appropriate, we approximate them with well-known distributions. The parameters obtained can be used as input to generate realistic failure scenarios, which is important to various traffic engineering problems that take failures into account. Those include routing protocols, network management systems and network design itself.

The paper is organized as follows. Section II presents some related work in the area of failure analysis and fault management. Section III describes the data collection process in the Sprint backbone and provides an overview of the data set under study. Section IV describes our classification methodology and evaluates its accuracy. Section V describes the results of our classification of failures and the characteristics of each identified class. Section VI discusses how our characterization can be used to build a failure model, and identifies open issues for further investigation. Section VII concludes the paper.

II. Related Work

The availability of spare capacity and sound engineering practices in commercial IP backbones makes it easy to achieve traditional QoS objectives such as low loss, latency and jitter. Recent results show that the Sprint network provides almost no queuing delays [7], [2], negligible jitter [2] and is capable of supporting toll-quality voice service [8].

On the other hand, failures can degrade network performance by reducing available capacity and disrupting IP packet forwarding. Common approaches for ensuring network survivability in the presence of failures include protection and restoration at the optical layer or the IP layer [9], [10], [4]. A significant amount of effort has been investing in achieving sub-second convergence in IS-IS [11]. In addition, a number of new approaches have been proposed to account for backbone failures, including the selection of link weights in the presence of transient link failures [12], [13], deflection routing techniques to alleviate temporary link overloads due to failures [6], network availability-based service differentiation [14], and failure insensitive routing [15].

All of the above approaches require a thorough understanding of the cause and characteristics of backbone failures. However, such an understanding has been limited partly by a lack of measurement data from operational networks, and partly by a focus on traditional QoS objectives such as loss and delay. In some cases, traceroutes were used to study the routing behavior in the Internet. These include studies on routing pathologies, stability and symmetry [16], stationarity of Internet path properties [17], and evaluation of routing-based and caching techniques to deal with failures [18]. To the best of our knowledge, [19] is the only work that uses OSPF routing updates for failure analysis, although its primary focus is on studying stability of inter-domain paths. More recently, [3] has performed a preliminary analysis of backbone link failures based on IS-IS routing updates. Our work builds on [3] and makes the following new contributions. We study a larger and more recent data set, classify failures according to probable causes, and provide characterization for each class that can be used to build a failure model.

To put things in perspective, reliability is an aspect of dependable distributed systems and has been extensively studied in the context of fault-tolerant distributed systems. A landmark paper on failures in Tandem systems and the techniques to prevent them is [20]. In parallel and even earlier, a mathematical framework was developped in the Operations Research world to manage the reliability and risk in systems composed of various components [21]. In general, complex systems that have passed the stage of proof of concept and have matured into industrial grade systems, are expected to provide high reliability/availability to their users. Telephone networks and power grids are examples of such mature networks. Recently, there has been an increasing interest in the reliability of end-to-end services in the Internet. Examples of recent work include [22] where human errors and incorrect configurations are identified as a main source of errors, [23] where giant scale web services are examined, and [24] where fast recovery is compared to high reliability. Within the above context, our work targets failures that affect routing and availability across a single backbone network.

III. Failure Measurements

In this section, we discuss the types of failures that impact IP connectivity, present our methodology for extracting link failure information from IS-IS routing updates and briefly summarize the data set.

A. Failures with an impact on IP connectivity

The Sprint IP network has a layered structure, with an IP layer operating directly above a dense wavelength-division multiplexing (DWDM) optical infrastructure with SONET framing.

There are two main approaches for sustaining end-to-end connectivity in IP networks in the event of failures: protection and restoration. Protection is based on fixed and predetermined failure recovery, with a working path set up for traffic forwarding and an alternate protection path provisioned to carry traffic if the primary path fails. Restoration techniques attempt to find a new path on-demand to restore connectivity when a failure occurs. Protection and restoration mechanisms can be provided either at the optical or at the IP layer, with different cost-benefit tradeoffs [4], [10].

The Sprint IP network relies on IP layer restoration (via IS-IS protocol) for failure recovery. All failures at or below the IP layer that can potentially disrupt packet forwarding manifest themselves as the loss of IP links between routers. The failure or recovery of an IP link leads to changes in the IP-level network topology. When such a change happens, the routers at the two ends of the link notify the rest of the network via IS-IS. Therefore, the IS-IS update messages constitute the most appropriate data set for studying failures that affect connectivity.

Failures can happen at various protocol layers in the network for different reasons. At the physical layer, a fiber cut or a failure of optical equipment may lead to loss of physical connectivity. Hardware failures (e.g. linecard failures), router processor overloads, software errors, protocol implementation and misconfiguration errors may also lead to loss of connectivity between routers. When network components (such as routers, linecards, or optical fibers) are shared by multiple IP links, their failures affect all the links. Finally, failures may be unplanned or due to scheduled network maintenance. Note that at the IS-IS level, we observe the superposition of all the above events. Inferring causes from the observed IS-IS failures is a difficult reverse engineering problem.

B. Collecting and Processing ISIS Updates

We use the Python Routing Toolkit (PyRT)[1] to collect IS-IS Link State PDUs (LSPs) from our backbone. PyRT includes an IS-IS "listener" that collects these LSPs from an IS-IS enabled router over an Ethernet link. The router treats the listener in the same way as other adjacent routers, hence it forwards to the listener all LSPs that it receives from the rest of the network. Since IS-IS broadcasts LSPs through the entire network, our listener is informed of every routing-level change occurring anywhere in the network. However, the listener is passive in the sense that it does not transmit any LSPs to the router. The session between the listener and the router is kept alive via periodic IS-IS keepalive (Hello) messages. Upon receiving an LSP, the listener prepends it with a header in MRTD format (extended to include timestamp of micro-second granularity) and writes it out to a file. The data presented in this paper were collected from a listener at a Sprint backbone Point-of-Presence (POP) in New York.

Whenever IP level connectivity between two directly connected routers is lost, each router independently broadcasts a "link down" LSP through the network. When the connectivity is restored, each router broadcasts a "link up" LSP. We refer to the loss of connectivity between two routers as a *link failure*.

The LSPs from the two ends of a link reporting loss or restoration of IP connectivity may not reach our listener at the same time. The start of a failure is recorded with the MRTD timestamp of the first LSP received at our listener that reports "link down". The end of each failure is recorded with the MRTD timestamp of the second LSP received at our listener that reports "link up". This asymmetry is conformant with how the IS-IS protocol reacts to routing updates. As soon as a router receives the first LSP reporting an "link down", it considers the IP connectivity to be lost without waiting for the second LSP. Hence, the first LSP is sufficient to trigger a route re-computation, which may lead to a disruption in packet forwarding. However, in order to consider the IP connectivity restored, a router waits until it receives LSPs reporting "link up" from both ends of a link. In the rest of the paper, we refer to the time between the start and the end of a failure, as defined above, as the *time-to-repair* for the failure.

[1]The source code is publicly available at http://ipmon.sprint.com/pyrt

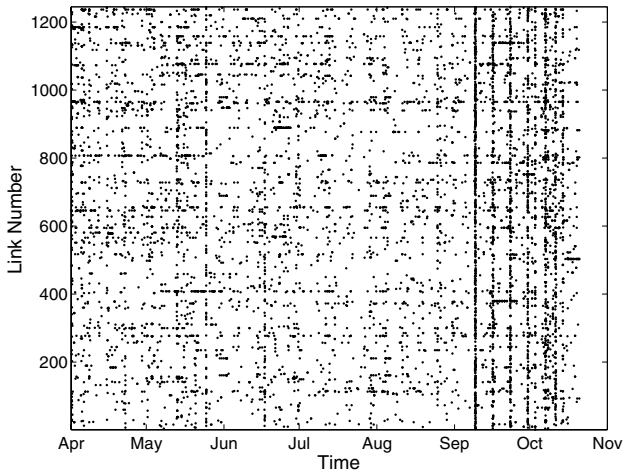


Fig. 1. Data set under study: failures in continental US between April 1 and October 21, 2002.

C. The Failures Data Set

Using the steps described above, we can determine the start and end times for failures on every link in the Sprint backbone. The data are collected for the period between April 1st and October 21st 2002, in the continental US. This data set involves a large number of links, routers and POPs (in the order of thousands, hundreds and tens respectively). We consider that link failures with time-to-repair longer than 24 hours are due to a links being decommissioned rather than to accidental failures, and therefore we exclude them from the failures data set. Indeed, the specified time-to-repair for any failure is in the order of hours and not in the order of days.

Fig. 1 shows the failures in the data set under study, across links and time. A single dot at (t, l) is used to represent a failure that started at time t, on link l. One can see that failures are part of the everyday operation and affect a variety of links. We also observe that the failures occurrence follows patterns, such as (more or less prominent) vertical and horizontal lines of different lengths. In the rest of the paper, we further use these visual patterns as guidance for our failure classification.[2]

[2]The scale of the figure is chosen to emphasize the horizontal and vertical patterns. Times-to-repair are not represented in the figure and the area covered by the dots represents neither the total duration nor the impact of link failures on the Sprint backbone.

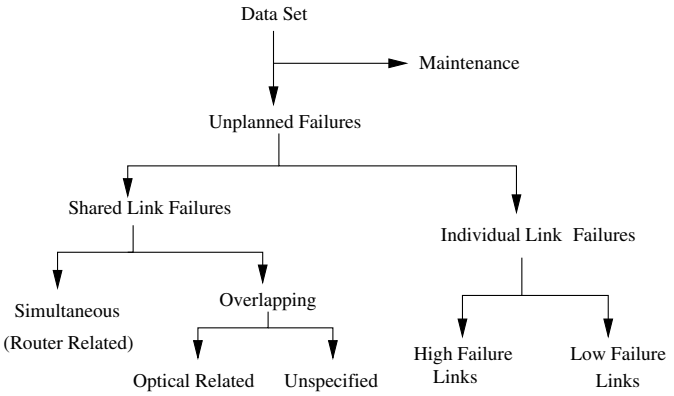


Fig. 2. Classification of failures

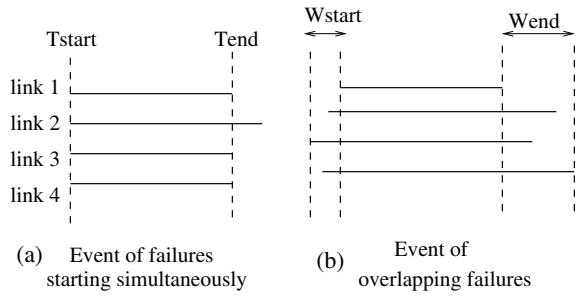


Fig. 3. Example events of simultaneous and overlapping failures.

IV. Classification Methodology

This section describes our methodology for classifying failures according to their causes and properties. We attempt to infer the causes by leveraging patterns observed in the empirical data and by correlating them with the possible causes. We first give an overview of our classification and then we discuss each class in detail.

A. Overview

Our approach is to use several hints obtained from the IS-IS failure data to identify groups of failures due to different causes. A visual inspection of Fig. 1 provides insights into how to perform this classification. We observe that the failures are not uniformly scattered- there are vertical and horizontal lines. The vertical lines correspond to links that fail at the same time (indicating a shared network component that fails) or to links that fail close in time (e.g. due to a maintenance activity) but appear almost aligned in the plot. The horizontal lines correspond to links that fail more frequently than others. Apart from these lines, the remaining plot consists of roughly uniformly scattered points.

Our classification of failures is summarized in Fig. 2 and consists of the following steps. We first separate failures due to scheduled *Maintenance* from *Unplanned* failures. We analyze the unplanned failures in greater depth since these are the ones that an operator seeks to minimize. We distinguish between *Individual Link Failures* and *Shared Link Failures*, depending on whether only one or multiple links fail at the same time. Shared failures indicate that the involved links share a network component that fails. This component can be located either on a common router (e.g. a linecard or route processor in the router) or in the underlying optical infrastructure (a common fiber or optical equipment). Therefore, we further classify shared failures into three categories according to their cause: *Router-Related*, *Optical-Related* and *Unspecified* (for shared failures where the cause cannot be clearly inferred). We divide links with individual failures into *High Failure* and *Low Failure Links* depending on the number of failures per link. In Fig. 2, maintenance and shared failures correspond to the vertical lines, high failure links correspond to the horizontal lines, and low failure links correspond to the roughly uniform plot that remains after excluding all the above classes of failures.

We now consider each class separately and describe *(i)* the rules that we use to decide whether a failure belongs to this specific class and *(ii)* how we obtain partial confirmation for the inferred cause.

B. Maintenance

Failures resulting from scheduled maintenance activities are unavoidable in any network. Maintenance is usually scheduled during periods of low network usage, in order to minimize the impact on performance. The maintenance window in the US Sprint backbone network is Mondays 5am-2pm, UTC time. It turns out that failures during this window are responsible for the most prominent vertical lines in Fig. 1, including the ones in September - October.

C. Simultaneous Failures

In the shared failures class, we first identify failures that happen simultaneously on two or more links. Failures on multiple links can start or finish at exactly the same time, when a router reports them in the same LSP. For example, when a linecard fails, a router may send a single LSP to report that all links connected to this linecard are going down. When our listener receives this LSP, it will use the timestamp of this LSP as the start for all the reported failures. Similarly, when a router reboots, it sends an LSP reporting that many of the links connected to it are going up. When our listener receives this LSP, it will use the same timestamp as the end for all the reported failures. (However, it still needs to receive an LSP from the other end to declare the end of a failure.)

In our data, we identify many such cases. An example is shown in Fig. 3(a): 4 links are going down at exactly the same time T_{start} (and 3 out of 4 come up at the exactly same time T_{end}). We refer to such failures as simultaneous failures and we group them into events.

For every event of simultaneous failures found in the data set, we verified that all involved links are indeed connected to a common router. And reversely, there is no simultaneous failure event that does not involve a common router, which confirms our intuition. Therefore, we attribute these events (simultaneous failures on links connected to the same router) to problems

on the common router and we call them *router events*. Such problems include a router crash or reboot, a linecard failure or reset, a CPU overload. In the rest of the paper, we refer to these failures as *Router-Related*. Unfortunately, based on the available data set, it is impossible to do any finer classification, i.e. we are unable to identify whether the failure is due to high load, software or hardware error or human misconfiguration. Such a root-cause analysis is a direction for future work and requires additional failure logs.

Occasionally, a link in a router event may come up later than the others, as shown in Fig. 3(a). This can happen either because the link comes up later (e.g. router interfaces coming up one-by-one) or because the LSP from the other end of the link reaches our listener later (either delayed or lost). However, in 50% of the router events identified in the data set, all links came up at the exact same time; in 90% of the cases the last link came up no later than 2 minutes after the first link.

D. Overlapping Failures

After excluding the simultaneous failures, we relax the time constraint from "simultaneous" to "overlapping", i.e. we look for events where all failures start and finish within a time window of a few seconds. An example is shown in Fig. 3(b), failures on all 4 links start within W_{start} and finish within W_{end} seconds from each other.

Overlapping failures on multiple links can happen when these links share a network component that fails and our listener records the beginning and the end of the failures with some delays W_{start} and W_{end}. For example, a fiber cut leads to the failure of all IP links over the fiber, but may lead to overlapping failures in our listener for several reasons. First, there are multiple protocol timers involved in the failure notification and in the generation of LSPs by the routers at the ends of the links. Most of these timers are typically on the order of tens of milli-seconds up to a few seconds. The dominant ones are the IS-IS carrier delay timer [3] with default 2 seconds to report a link going down and 12 seconds to report a link going up. The timers can be configured to have different values on different routers. Finally, the LSPs from the two ends of the link can reach our listener through different paths in the network and thus may incur different delays; or an LSP may be lost, leading to an additional retransmission delay.

The choice of windows, W_{start} and W_{end}, becomes important for a meaningful definition of overlapping failures. If the windows are chosen too long, failures that overlap by coincidence may be wrongly interpreted as shared failures. Windows that are too short may fail to detect some shared failures. We choose W_{start} and W_{end} to be 2 and 12 sec to match the default timers used to report a link down or up respectively. We also varied W_{start} from 0.5 to 10 sec and W_{end} from 0.5 to 20 sec and observed that the number of overlapping failures or events is relatively insensitive around the chosen values.

We now focus on identifying the network component that is responsible for the overlapping failures. Links can share components either at a router or in the optical infrastructure.

TABLE I
SUMMARY OF OVERLAPPING EVENTS

Classification of event	% events	% failures
Overlapping	100%	100%
Optical-Related	75%	80%
Unspecified	25%	20%

TABLE II
USING THE IP-TO-OPTICAL MAPPING TO CONFIRM THAT LINKS IN THE SAME OPTICAL EVENT SHARE AN OPTICAL COMPONENT

Optical-Related Events	%
Found in the database	93% of optical events
All links have common site(s)	96% of found events
All links have common segment(s)	98% of found events

Optical-Related. Among all overlapping events, we identify those that involve only inter-POP links and that do not share a common router. It turns out that 75% of all overlapping events and 80% of all overlapping failures are of this type, see Table I. We consider those events to be *Optical-Related* for the following reason. Since the links in the same event have no router in common, an explanation for their overlapping failures is that they share some underlying optical component that fails, such as a fiber or another piece of optical equipment.

To verify this conjecture, we use an additional database: the IP-to-Optical mapping of the Sprint network. This database provides the mapping from the IP logical topology to the underlying optical infrastructure. It provides the list of optical equipment used by every IP link. The optical topology consists of sites (cities where optical facilities are located) and segments (pair of sites connected with an optical fiber). IP links share necessarily some sites or segments.

Table II summarizes our findings in the IP-to-Optical database. Out of all overlapping events that we classify as optical-related (i.e. inter-POP without a common router), we were able to find 93% of them in the database (meaning that all links in the same event were found in the database). Not all links are found in the database due to changes in the mapping. For each event found in the database, we check whether all links in the event share some optical component. We find that 96% of the events found in the database, involve links that all share at least one site; 98% of the found events involve links that all share at least one segment. In fact, links in the same event share even more than just one site or segment. They share from 1 up to 30 sites (8.3 on average) and from 1 up to 27 segments (7.3 on average). These findings validate our conjecture that the events classified as optical-related are most likely due to the failure of some optical component shared by multiple IP links.

Unspecified. All the overlapping failures that are not classified as optical-related fall in this class. These include overlapping failures on inter-POP links connected to the same router. The cause is ambiguous: they could be due to a problem at the router or to an optical problem. They also include overlapping

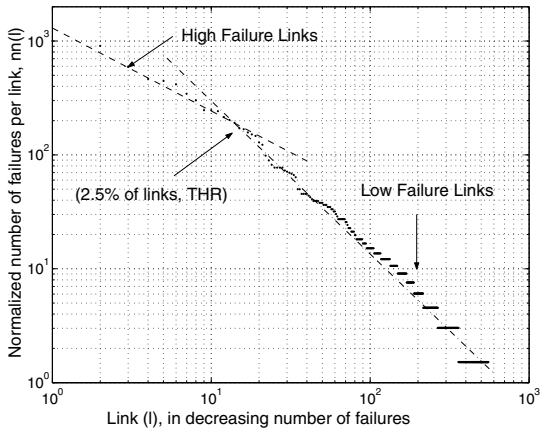


Fig. 4. Number of individual failures per link

failures of links in the same POP that could be due to a problem or operation in the POP. However, since we are not able to satisfactorily infer their cause, we call these events *Unspecified* and we do not attempt to analyze them further. They account for only 20% of the overlapping failures (see Table I), which is less than 3% of all the unplanned failures.

E. Individual Link Failures

After excluding all the above classes of failures from the data set, we refer to the remaining failures collectively as *Individual Failures* because they affect only one link at a time.

Let $n(l)$ be the number of individual failures per link l=1,..,L. Let the maximum number of failures in a single link be $maxn = max_l(n(l))$. For proprietary reasons, we show the normalized value $nn(l) = 1000 \cdot n(l)/maxn$, instead of the absolute number $n(l)$. In Fig. 4, we plot $nn(l)$, for all links in decreasing order of number of failures. There are several interesting observations based on this graph. First, links are highly heterogeneous: some links fail significantly more often than others, which motivates us to study them separately. Second, there are two distinct straight lines in this log-log plot in Fig. 4. We use a least-square fit to approximate each one of them with a power-law: $n(l) \propto l^{-0.73}$ for the left line and $n(l) \propto l^{-1.35}$ for the right line. Notice that both the absolute ($n(l)$) and the normalized ($nn(l)$) values follow a power-law with the same slope; therefore, the interested reader is still able to use the normalized value to simulate this behavior. The dashed lines in the figure, intersect approximately at a point that corresponds to 2.5% of the links and to a normalized number of failures $nn(l) = 152$.

We use this value as the threshold ($THR = 152$) to distinguish between two sub-classes: the *High Failure Links* ($nn(l) \geq THR$) and the *Low Failure Links* ($1 \leq nn(l) \leq THR$). High failure links represent only 2.5% of all links but account for more than half of individual failures. It is difficult to determine the cause of individual failures. High failure links may be in an advanced stage of their lifetime and their components fail frequently; or they may be undergoing an upgrade or testing operation for a period of time. Unlike all previous failure classes, low failure links do not have a prominent pattern either in time or across links.

TABLE III
PARTITIONING FAILURES INTO CLASSES

Failure Class		% of all	% of unplanned
Data Set		100%	
Maintenance		20%	
Unplanned		80%	100%
Shared	Router-Related		16.5%
	Optical-Related		11.4%
	Unspecified		2.9%
Individual	High Failure Links		38.5%
	Low Failure Links		30.7%

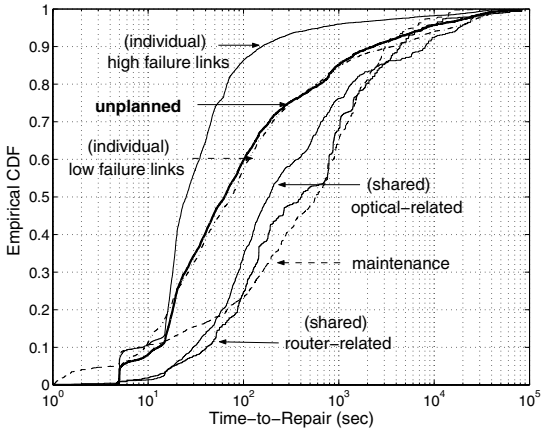


Fig. 5. Cumulative distribution function (CDF) of the time-to-repair for each class of unplanned failures

V. FAILURE ANALYSIS

We now consider each class of failures separately and we study its characteristics that are useful for re-producing its behavior. These characteristics include time between failures, time-to-repair, number of links involved in an (router or optical-related event) event and distribution of failures/events across links/routers. We provide empirical distributions and, when possible, we also fit them to simple distributions. [3]

Table III and Fig. 5 summarize and compare all classes, and will be referenced repeatedly throughout this section. Table III shows the contribution of each class to the total number of failures. Fig. 5 provides the empirical cumulative distribution function of time-to-repair for each class of unplanned failures.

A. Maintenance

20% of all failures happen during the window of 9-hours weekly maintenance, although each such window accounts only for 5% of a week. Fig.6 shows the occurrence of link failures due to scheduled maintenance. It turns out that those account for many of the vertical lines in Fig.1.

[3]For proprietary reasons, we provide a characterization in terms of percentages and statistical properties, rather than in terms of absolute numbers. However, the information provided should be sufficient for the reader to reproduce realistic failure scenarios.

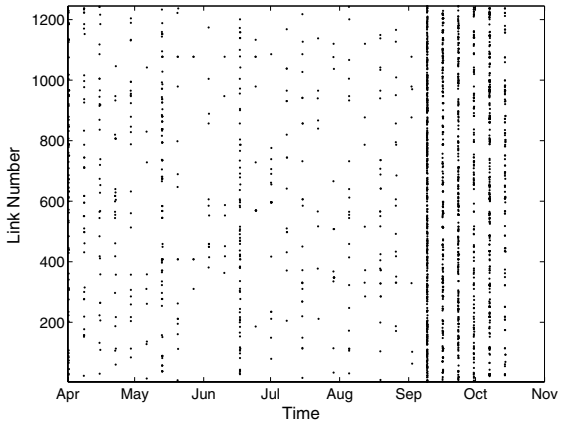


Fig. 6. Failures during weekly maintenance windows

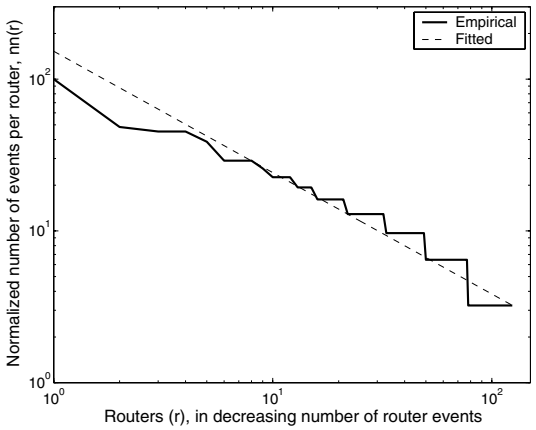


Fig. 7. Normalized number of events per router, in decreasing order.

More than half failures during the maintenance window are also router-related (according to the definition of Section IV-C). This is expected as maintenance operations involve shutting down and (re)starting routers and interfaces. Also, Fig. 5 shows that the CDF of time-to-repair for maintenance-related failures, is close to the CDF for the router-related failures, which further supports the observation that many of the maintenance failures are router-related. A typical maintenance window for a given interface/router is one hour, although it can take less.

B. Router-Related Failures

Router-related events are responsible for 16.5% of unplanned failures. They happen on 21% of all routers. 87% of these router events (or 93% of the involved failures) happen on backbone routers and the remaining 7% happens on access routers. An access router runs IS-IS only on two interfaces facing the backbone but not on the customer side.

Router events are unevenly distributed across routers. Let $n(r)$ be the *number of events* in router r and $nn(r) = 100 \cdot n(r)/maxn$ be its normalized value with respect to its maximum value $maxn = max_r(n(r))$. For proprietary reasons, we present the normalized instead of the absolute value. Fig. 7 shows the normalized number of events per router, for all routers ranked in decreasing number of events. Interestingly, the straight line in the log-log plot indicates that $nn(r)$ follows roughly a power-law. Both $n(r)$ and $nn(r)$ follow a power-law with the same slope. An estimate of the parameters of the power-law using least-square method yields $n(r) \propto r^{-0.8}$, which we plot as a dashed line in Fig. 7. One can use the same figure to calculate the mean time between events for different routers: it varies from 5 days up to several months.

When a router event happens, multiple links of the same router fail together. The distribution of the *number of links in an event* is shown in Fig. 8. Events involve 2 to 20 links. This is related to the number of ports per linecard, which varies typically between 2 and 24. Most events involve two links; 12% of these events are due to failures on the two links of access routers.

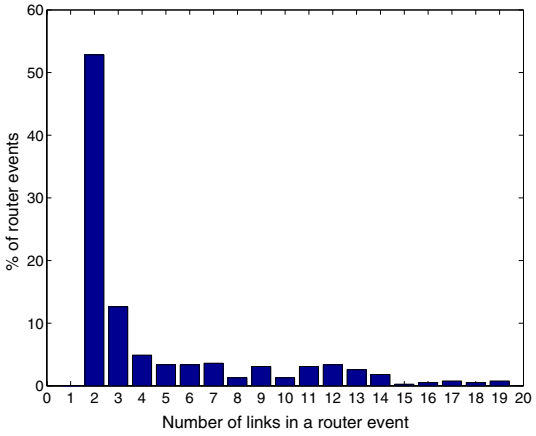


Fig. 8. Empirical PDF for the number of links in a router event.

The empirical CDF of *time-to-repair* for router-related failures is shown in Fig. 5, together with those of the other classes. The CDF for the router and the maintenance-related classes are close to each other, and shifted toward larger values compared to other classes. This could be due to human intervention for repair or due to the rebooting process that takes on the order of several minutes for backbone routers. Repair times for failures belonging to the same event are roughly equal.

Another characteristic of interest is the *frequency* of such events. Because not all routers experience enough events for a statistically significant derivation of per router inter-arrival times, we consider the time between any two router events, anywhere in the network. Fig. 9 shows the empirical cumulative distribution of network-wide time between two router events. We observe that the empirical CDF is well approximated by the CDF of a Weibull distribution: $F(x) = 1-exp(-(x/\alpha)^{\beta}), x \geq 0$. We estimate the Weibull parameters using maximum-likelihood estimation as $\alpha = 0.068$ and $\beta = 0.299$. The fitted CDF is shown in dashed line in Fig. 9. In addition, we notice that the autocorrelation function decreases

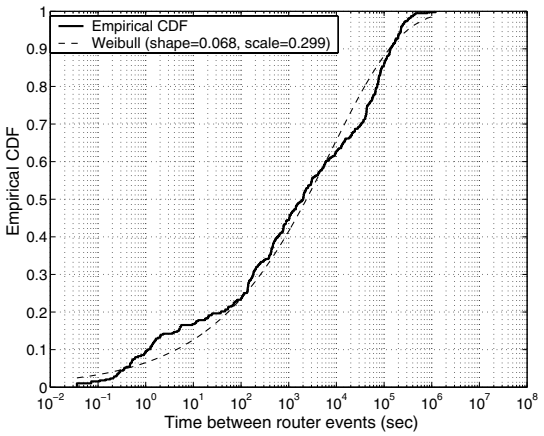


Fig. 9. Cumulative distribution function (CDF) for the network-wide time between router events.

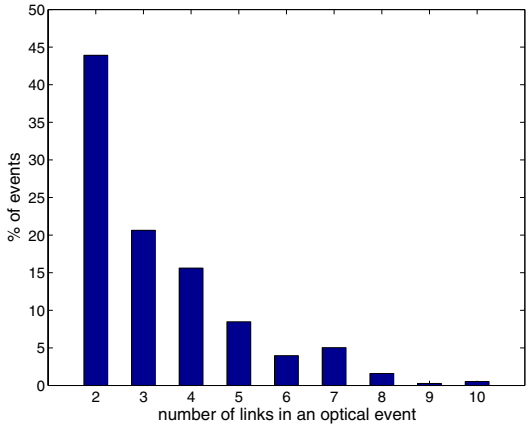


Fig. 10. Number of links in an optical event.

fast beyond small values of the lag. This means that, for practical purposes, one could use i.i.d Weibull random variables to simulate the time between router events. The appropriateness of the Weibull distribution for the time between failures, is discussed in Section VI.

C. Optical-Related Failures

Shared optical failures have an important impact on the network operation, as they affect multiple links and are therefore more difficult to recover from than individual link failures. Shared optical-related failures are responsible for 11.4% of all unplanned failures.

Fig. 10 shows the histogram of the *number of IP links* in the same optical event. There are at least two (as their definition requires an overlap in time) and at most 10 links in the same event. This is in agreement with sharing information derived from the IP-to-Optical mapping. For example, the most frequent number of links sharing a segment according to the mapping is 2 (which is also the case in optical events); the maximum number of links that share a segment according to the mapping is 25 (larger than the maximum number of links in any optical event).

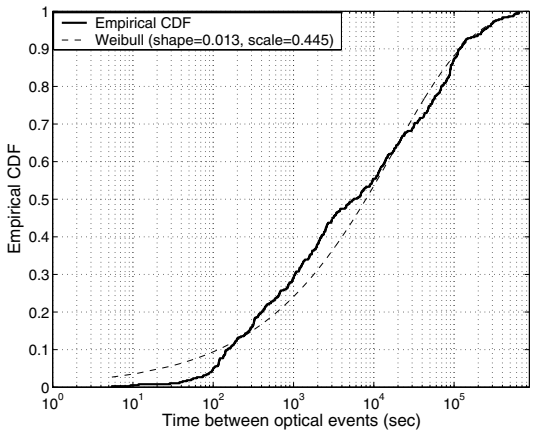


Fig. 11. Cumulative distribution function (CDF) for the network-wide time between optical events.

The CDF of *time-to-repair* for optical-related failures is shown in Fig. 5. Short time-to-repair values are more likely due to faults in the optical switches, while longer times correspond to fiber cuts or other failures that require human intervention to be repaired. Similar to the previous classes of shared failures, the CDF is shifted towards larger values, compared to individual failures. By their definition, failures in the same optical event happen within a few seconds from each other.

Another characteristic of interest is the *frequency* of optical failures in the network. Fig. 11 shows the CDF for the time between two successive optical events, anywhere in the network. The values range from 5.5 sec up to 7.5 days, with a mean of 12 hours. We use maximum likelihood estimation to estimate the parameters of a Weibull distribution from the empirical data and we obtain $\alpha = 0.013$ and $\beta = 0.445$. The resulting CDF, shown in dashed line in Fig. 11, is an approximation of the empirical CDF. However, one can observe that there are more distinct modes in this distribution (e.g. one from 0 up to 100 sec, a second from 100 sec up to 30 hours and a third one above that), hinting to more factors that could be further identified. A closer look in the sequence of events reveals that times between events below 100 sec correspond to many closely spaced events on the same set of links that could be due to a persistent problem in the optical layer. However, the Weibull fit of the aggregate CDF sufficiently characterizes the frequency of optical events network-wide.

D. High Failure Links

High failure links include only 2.5% of all links. However, they are responsible for more than half of the individual failures and for 38.5% of all unplanned failures, which is the largest contribution among all classes, see Table III.

As we discussed earlier in Fig. 4, the *number of failures* $n(l)$ per high failure link l follows a power-law: $n(l) \propto l^{-0.73}$. Each high failure link experiences enough failures to allow for a characterization by itself.

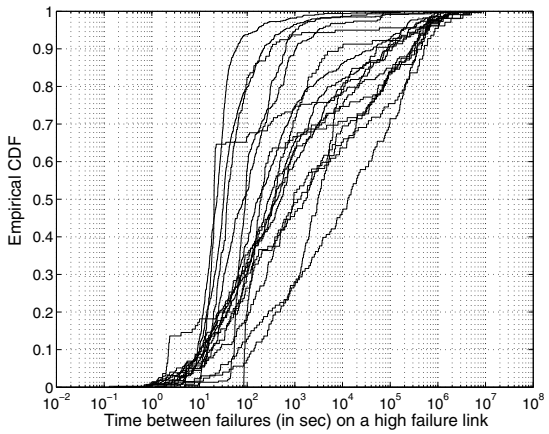


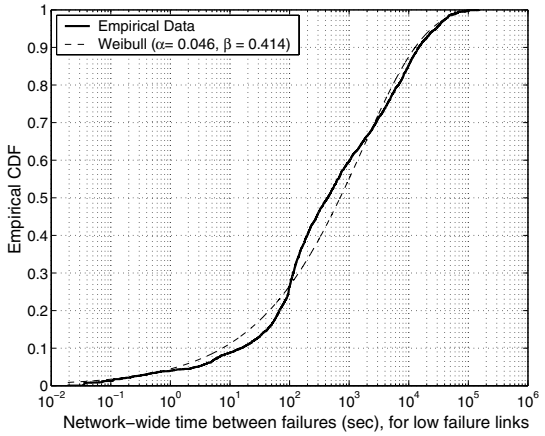
Fig. 12. Time between failures on each high failure link.

The empirical CDF of the *time between failures* on each of the high failure links is shown in Fig. 12. Some of them experience failures well spread across the entire period. They correspond to the long horizontal lines in Fig. 1 and the smooth CDFs in Fig. 12. Some other high failure links are more bursty: a large number of failures happens over a short time period. They correspond to the short horizontal lines in Fig. 1 and to the CDFs with a knee in Fig. 12. The mean time between failures varies from 1 to 40 hours for various links, i.e. a shorter range than for the other classes. Finally, the CDF of the *time-to-repair* for failures on high failure links is shown in Fig. 5. It is clearly distinct from all other classes- failures last significantly shorter (up to 30% difference from the CDF of all unplanned failures and up to 70% from the CDF of the shared failures). The larger number of shorter failures is in accordance with our conjecture that high failure links are in an advanced stage of their lifetime and their components are probably subject to intermittent and recurring faults.

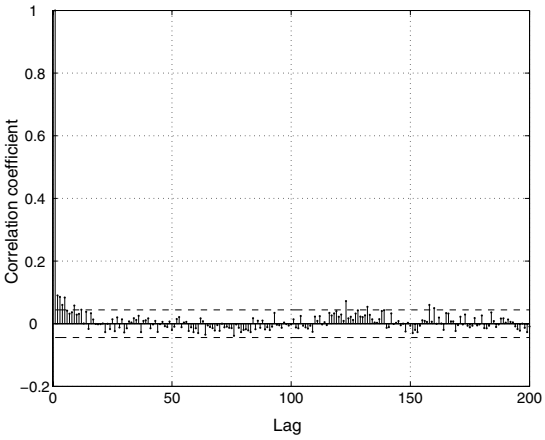
E. Low Failure Links

In Fig. 4, we have already defined low failure links are those with less individual failures than the threshold THR. The number of failures $n(l)$ per link (l) follows roughly a power-law: $n(l) \propto l^{-1.35}$.

A statistically significant characterization is not possible for every low failure link, as many of them do not experience enough failures. We group all low failure links together and study *the time between any two failures*- the two failures may happen anywhere in the network and not necessarily on the same link. Fig. 13(a) shows the empirical CDF for the network-wide time between failures. It turns out that in this case too, the empirical CDF is well approximated by a Weibull distribution with maximum-likelihood estimated parameters $\alpha = 0.046$ and $\beta = 0.414$; the fitted distribution is shown in dashed line in the same figure. In Fig. 13(b), we also show the autocorrelation function for the time between failures at the 90% confidence interval. We notice that correlation in the time between failures drops fast after a small lag. This means



(a) CDF of (network-wide) time between failures



(b) Autocorrelation of (network-wide) time between failures

Fig. 13. Network-wide time between failures on low failure links

that, as a first approximation, we can use i.i.d. Weibull random variables with the fitted parameters to regenerate the network-wide time between individual failures on low failure links.

Finally, the empirical CDF for the *time-to-repair* in this class is shown in Fig. 5, together with the rest of the classes. It is interesting to note that the CDF is very close to the CDF for all unplanned failures. This fact together with the observation that low failure links correspond to the roughly random part of Fig. 1 indicate that, unlike the previous classes, failures in this class have an "average" behavior and are the norm rather than the exception of the entire data set.

VI. Discussion

This work offers a detailed characterization of link failures and is useful in two ways. First, it reveals the nature and extent of failures in today's IP backbones. Our methodology can be used to identify failing network components and pinpoint

areas for improvement. Second, it is the first step toward building a failure model, as we provide information about the cause of failures as well as their statistical properties. Such a model would be useful as an input to various engineering problems that need to account for failures. In this section, we discuss how our classification and analysis can be used toward building a failure model, as well as open issues and future directions.

IP link failures occur due to several causally unrelated events at or below the IP layer. Accordingly, we have divided failures into a number of classes such that their underlying causes are unrelated. Therefore a backbone failure model can be obtained by developing a model to characterize each class independently, and then combining them. For each class, we have identified a few key properties (such as the time between failures, the time-to-repair and the distribution of failures across links and routers), provided their statistics and, when possible, fitted them using well-known distributions with a small number of parameters.

Let us first discuss the validity of our *classification* and then the modeling of *each class* separately.

Our *classification* is based on hints from the ISIS data set, discussed in detail in Section IV. In the same section, we used the IP-to-Optical database and confirmed to a very satisfactory degree the validity of our optical-related class of failures. The fact that all simultaneous failures involved a common router was also a confirmation for the router-related class. When we applied our classification methodology to the measurements, the statistics of the identified classes turned out to be quite different from each other, which provides further assurance about our classification. For example, the CDF of time-to-repair in Fig. 5 are well separated from each other: the shared failures "pull" the CDF toward larger values, the high failure links "pull" it toward smaller values, while the low failure links are in the middle. A similar separation happens in the initial Fig. 1: the maintenance and shared failures capture the vertical lines, the high failure capture the horizontal lines, the low failure links capture the remaining "random" plot. However, inferring the failures causes based solely on IS-IS logs is a difficult reverse engineering problem and results to a coarse classification. In future work, we plan to correlate the IS-IS data set with additional logs, such as SONET alarm logs, router logs, maintenance activity schedules. Even with such cross examinations, it will still be difficult to uniquely identify the causes for all failures.

The characterization in Section V provides the basis for modeling *each class* separately. There are two interesting observations from parameterizing the properties of various classes. First, we observe that the empirical CDF for the network-wide time between failures (or events) for three classes of failures was well approximated by a *Weibull distribution*. These three classes are the router-related (Fig. 9), the optical-related (Fig. 11) and the low failure links (Fig. 13(a)). The Weibull distribution has been found widely applicable in reliability engineering to describe the lifetime of components, primarily due to its versatile shape [25]. Interestingly enough, the Weibull distribution is derived as a smallest extreme value distribution: for a large number of identical and independent components, the time to the first failure follows a Weibull [25]. One could say that this explains the good fit in our case: there is a large number of components in each class and the network-wide time between failures can be interpreted as the time to the first failure, assuming a renewal process. However, there are implicit assumptions in such a claim, which have not been validated in our data: e.g. independence and similarity of components and the renewal assumption.

Our second finding is that *power-laws* describe well the distribution of failures (or events) across components in the same class. Indeed, power-laws fitted well the number of router events per router (see Fig. 7) as well as for the number of individual failures per high or low failure link (see Fig. 4). Power-laws are often found to describe phenomena in which, small occurrences are extremely common, whereas large instances are extremely rare. Examples include man-made or naturally occurring phenomena, such as word frequencies, income distribution, city sizes and earthquake magnitudes, [26]. Recently, the Internet has been found to display quite a number of power-law distributions, such as the node outdegrees and other properties of Internet topologies [27], the sizes and other attributes of objects in a web page [28] (Pareto distributions are another expression for power-laws).

Care must be taken if the characterization of Section V is used as the basis for a failure model. For example, consider the low failure links in Section V-E and let us try to re-generate failures that have similar statistics with the measured ones. In order to decide when the next failure happens, one can pick a random number from the Weibull distribution for the network-wide time between failures. In order to decide on which link this failure happened, one could pick a link using the power-law distribution. (Similar steps can be followed to reproduce the router events using the network-wide time between events and the distribution of events across routers.) However, using these distributions incorporates the assumption that the arrival of successive failures is independent of the distribution across links. Clustering of failures per link would break this assumption. Therefore, the dependence among the two dimensions, i.e. "time of occurrence" and "link/router of occurrence", needs to be studied further. Furthermore, the independence among components in the same class is important. We did find small correlation among low failure links; however this is only a necessary and not a sufficient condition for the independence among components.

The characterization presented in this paper is specific to Sprint's IP network and one should be careful before blindly applying it to any kind of network. (For example, the network-wide time between failures is tied to the specific topology. In order to develop a model that is applicable to an IP network regardless of its topology, size and underlying infrastructure, we need a further characterization of each component and graph statistics that describe the spatial occurrence of failures.) In general, the failure behaviour of a network is very related to its design, maintenance, technology, age and other specific

traits. These traits may vary between networks and also for the same network in time, which is make them inherently difficult to model in great precision.

One specific aspect we plan to address in future work, is the variation in time. Either the stationarity of the failure classification and characteristics needs to be established or the parameters of the model need to vary with time. This is an entire problem by itself and we plan to address it leveraging on the continuous collection of IS-IS updates from the Sprint network, and applying our methodology in shorter intervals of a long total period of time.

It also needs to be explored whether there is a dependence between the time-to-repair and the failure arrival process, in which case a marked Markov process (with the mark of each point being the time-to-repair) could be a more appropriate modeling approach.

Finally, an important direction for future work is to understand how the failures actually affect the network and service availability. Factors that determine the impact of a failure to the service, as perceived by the user, include (i) the characteristics of the failure (e.g. a shared failure affects multiple links and is more difficult to recover from) (ii) the network topology and the actual traffic carried over the network (note that with appropriate network design, the effect of failures on the traffic may be minimized) and (iii) the routing protocol (which determines the forwarding disruption associated with each failure and thus the effect perceived by the user) or other protection/restoration mechanisms used at lower layers (that are able to hide a failure from higher layers).

VII. Conclusions

In this paper, we analyze seven months of ISIS routing updates from the Sprint's IP backbone to characterize failures that affect IP connectivity. We classify failures according to their cause and describe the key characteristics of each class. Our findings indicate that failures are part of the everyday operation: 20% of them are due to scheduled maintenance operation, while 16% and 11% of the unplanned failures are shared among multiple links and can be attributed to router-related and optical-related problems respectively. Our study not only provides a better understanding of the nature and the extent of link failures, but is also the first step towards developing a failure model. Directions for future work include (i) the modeling aspects discussed in the previous section (ii) more root-cause analysis, using correlation with different failure logs (iii) a better understanding of the impact of failures on network availability.

Acknowledgments

Athina Markopoulou is grateful to professors Peter Glynn and Balaji Prabhakar, from Stanford University, for insightful discussions on the data analysis and on failure problems.

References

[1] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, R. Rockell, D. Moll, T. Seely, and C. Diot, "Packet-level traffic measurements from the Sprint IP backbone," *IEEE Network Magazine*, vol. 17, no. 6, Nov. 2003.

[2] K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, and C. Diot, "Measurement and analysis of single-hop delay on an IP backbone network," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 6, Aug. 2003.

[3] 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*, Nov. 2002.

[4] G. Iannaccone, C. Chuah, S. Bhattacharyya, and C. Diot, "Feasibility of IP restoration in a tier-1 backbone," in *IEEE Network, Special Issue on Protection, Restoration and Disaster Recovery*, Mar. 2004.

[5] D. Oran, "OSI IS-IS intra-domain routing protocol," RFC 1142, Feb. 1990.

[6] S. Iyer, S. Bhattacharyya, N. Taft, and C. Diot, "An approach to alleviate link overload as observed on an IP backbone," in *Proc. of IEEE Infocom*, Mar. 2003.

[7] C. Fraleigh, F. Tobagi, and C. Diot, "Provisioning IP backbone networks to support latency sensitive traffic," in *Proc. of IEEE Infocom*, Mar. 2003.

[8] C. Boutremans, G. Iannaccone, and C. Diot, "Impact of link failures on VoIP performance," in *Proc. of NOSSDAV*, May 2002.

[9] A. Fumagalli and L. Valcarenghi, "IP restoration versus WDM protection: Is there an optimal choice?" *IEEE Network Magazine*, vol. 14, no. 6, pp. 34–41, Nov. 2000.

[10] L. Sahasrabuddhe, S. Ramamurthy, and B. Mukherjee, "Fault management in IP-over-WDM networks: WDM protection versus IP restoration," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 1, Jan. 2002.

[11] S. C. A. Alaettinoglou, "Detailed analysis of isis routing protocol on the qwest backbone," Nanog presentation, Feb. 2002, http://www.nanog.org/mtg-0202/ppt/cengiz.pdf.

[12] A. Nucci, B. Schroeder, S. Bhattacharyya, N. Taft, and C. Diot, "IGP link weight assignment for transient link failures," in *Proc. of IEEE ITCI 8*, Berlin, Germany, Sept. 2003.

[13] B. Fortz and M. Thorup, "Optimizing OSPF/IS-IS weights in a changing world," *IEEE Journal on Selected Areas in Communications*, Feb. 2002.

[14] M. Duvry, C. Diot, N. Taft, and P. Thiran, "Network availability based service differentiation," in *Proc. of IWQoS*, June 2003.

[15] S. Nelakuditi, S. Lee, Y. Yu, and Z.-L. Zhang, "Failure insenstive routing for ensuring service availability," in *Proc. of IWQoS*, 2003.

[16] V. Paxson, "End-to-end routing behavior in the internet," *IEEE/ACM Transactions on Networking*, vol. 5, no. 5, pp. 601–615, Oct. 1997.

[17] Y. Zhang, V. Paxson, and S. Shenker, "The stationarity of internet path properties: Routing, loss and throughput," available at http://www.icir.org/, ACIRI, Tech. Rep., May 2000.

[18] M. Dahlin, B. Chandra, L. Gao, and A. Nayate, "End-to-end WAN service availability," *IEEE/ACM Transactions on Networking*, vol. 11, no. 2, Apr. 2003.

[19] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental study of internet stability and wide-area network failures," in *Proc. of FTCS*, June 1999.

[20] J. Gary, "Why do computers stop and what can be done about it?" in *Symposium on Reliability in Distributed Software and Database Systems*, 1986.

[21] R. Barlow and F. Proschan, *Statistical Analysis of Reliability and Life Testing Models*. New York: Holt, Rinehart and Winston, 1975.

[22] D. Oppenheimer, A. Ganapathi, and D. Patterson, "Why do internet services fail, and what can be done about it?" in *4th USENIX Symposium on Internet Technologies and Systems (USITS'03)*, 2003.

[23] E. Brewer, "Lessons from giant-scale services," in *IEEE Internet Computing*, vol. 5, no. 4, 2001.

[24] A. Fox and D. Patterson, "When does fast recovery trump high reliability?" in *Proc. 2nd Workshop on Evaluating and Architecting System Dependability*, San Jose, CA, 2002.

[25] P. Tobias and D. Trindade, *Applied Reliability*, 2nd ed. Chapman Hall/CRC, 1995.

[26] L. Adamic, "Zipf, power-laws and pareto: a ranking tutorial," Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html.

[27] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *Proc. of ACM Sigcomm*, Sept. 1999.

[28] A.Feldmann, A.Gilbert, P.Huang, and W.Willinger, "Dynamics of IP traffic; a study of the role of variability and the impact of control," in *Proc. of ACM Sigcomm*, 1998.