Characterization of Failures in an Operational IP Backbone Network

Athina Markopoulou, *Member, IEEE*, Gianluca Iannaccone, *Member, IEEE*, Supratik Bhattacharyya, Chen-Nee Chuah, *Member, IEEE*, Yashar Ganjali, *Member, IEEE*, and Christophe Diot

***Abstract*—As the Internet evolves into a ubiquitous communication infrastructure and supports increasingly important services, its dependability in the presence of various failures becomes critical. In this paper, we analyze IS-IS routing updates from the Sprint IP backbone network to characterize failures that affect IP connectivity. Failures are first classified based on patterns observed at the IP-layer; in some cases, it is possible to further infer their 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 happen during a period of scheduled maintenance activities. Of the unplanned failures, almost 30% are shared by multiple links and are most likely due to router-related and optical equipment-related problems, respectively, while 70% affect a single link at a time. Our classification of failures reveals the nature and extent of failures in the Sprint IP backbone. Furthermore, our characterization of the different classes provides a probabilistic failure model, which can be used to generate realistic failure scenarios, as input to various network design and traffic engineering problems.**

***Index Terms*—Failure analysis, intermediate system to intermediate system (IS-IS) protocol, link failures, modeling, routing.**

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 queueing 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]. Routing reconvergence may also impose burden on router processors. Furthermore, if failures happen frequently, route-flapping may lead to network instability. For all these reasons, failures are a major concern for an operational network.

In this work, we collect IS-IS routing updates from the Sprint network using a passive listener, located at the New York point-of-presence (PoP). 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 U.S. 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. Our classification proceeds 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;among shared 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

Manuscript received July 15, 2004; revised October 10, 2005, and September 13, 2006; first published February 25, 2008; last published August 15, 2008 (projected); approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor M. Roughan. This work was conducted when the authors were affiliated (or in collaboration) with Sprint Advanced Technologies Lab, Burlingame, CA.

A. Markopoulou is with the Department of Electrical Engineering and Computer Science, University of California at Irvine, Irvine, CA 92697-2625 USA (e-mail: athina@uci.edu).

G. Iannaccone is with Intel Research, Berkeley, CA 94704 USA (e-mail: gianluca.iannaccone@intel.com).

S. Bhattacharyya is with Snaptell Inc., Mountain View, CA 94041 USA (e-mail: supratik@gmail.com).

C.-N. Chuah is with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA 95616-5294 USA (e-mail: chuah@ece.ucdavis.edu).

Y. Ganjali is with the Department of Computer Science, University of Toronto, Bahen Center for Information Technology, Toronto, ON, M5S 2E4 Canada (e-mail: yganjali@cs.toronto.edu).

C. Diot is with Thomson R&D, Paris, France (e-mail: christophe.diot@thomson.net).

Digital Object Identifier 10.1109/TNET.2007.902727

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.

The contributions of this work are as follows. First, 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. Second, we classify failures into classes according to the behavior they exhibit at the IP-layer, such as their time synchronization/overlap and their occurrence on particular links or routers. In some cases, these IP-layer characteristics, combined with supplementary information, can assist the operator to infer, or at least to narrow-down, the possible cause of failure. Finally, we provide the statistical characteristics of each class of failures and, when appropriate, we approximate them with well-known distributions. This provides a probabilistic failure model: one can generate failures according to the statistics for each class and then superimpose. This model can be used to generate realistic failure scenarios, as input to network design and traffic engineering problems that take failures into account.

An earlier version of this work appeared in [7]. This journal paper improves that work and extends it with additional materials, including: 1) the analysis of shorter (one month) time periods in order to understand how the statistical characteristics vary with time and 2) an analysis of SONET alarms in conjunction with the IS-IS data.

The paper is organized as follows. Section II presents 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. Section V describes the results of our classification of failures and the characteristics of each identified class. Section VI applies the classification and characterization methodologies to shorter (one-month) time periods and discusses how the statistical characteristics of each failure class vary with time. Section VII discusses how our characterization can be used to build a failure model, and identifies open issues for further investigation. Section VIII 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 quality-of-service (QoS) objectives such as low loss, latency, and jitter. Recent results show that the Sprint network provides almost no queueing delays [8], [2], negligible jitter [2] and is capable of supporting toll-quality voice service [9].

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 [10], [11], [4]. A significant amount of effort has been made to achieve sub-second convergence in IS-IS [12]. 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 [13], [14], deflection routing techniques to alleviate temporary link overloads due to failures [6], network availability-based service differentiation [15], and failure insensitive routing [16].

All the above techniques react to failures; therefore, their performance depends on the understanding of the characteristics of the underlying 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 [17], stationarity of Internet path properties [18], and evaluation of routing-based and caching techniques to deal with failures [19]. Other times, routing updates are used for failure analysis. For example, in [20], OSPF routing updates were used, although the primary focus was on studying stability of inter-domain paths. The authors continued their work in several directions, including an experimental study of OSPF, in [21]: they studied routing instability, and large scale anomalies caused mainly by external routing protocols. In [22], there is a case study on the characteristics and dynamics of OSPF link state advertisements (LSAs).

In [3], we used IS-IS routing updates to do a preliminary analysis of link failures on the Sprint backbone. In [7], we studied a larger and more recent data set, and characterized the statistical behavior of failures. Here, we improve [7] and extend it by additional materials, namely, the SONET alarms (Section IV.F) and the per-month analysis (Section VI).

References [23] and [24] are some recent developments in failure diagnosis in IP networks, with emphasis on shared risk link groups (SRLG) and cross IP and optical layers fault localization. In [23], the MinSetCover technique was developed to deal with the underdeterminedness of mapping an IP fault to the underlying physical cause. In [25], the SRLG-IP mapping problem is modeled as a Bayesian network and the fact that different SRLGs have different probabilities of failure is exploited to better infer the SRLG responsible for an IP-level failure. In [24], Shrink is proposed to extend the Bayesian approach in the presence of inaccurate SRLG-IP mappings and noisy measurements.

III. Failure Measurements

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

A. Design of the Sprint IP Backbone

The Sprint IP topology in the continental US consists of a collection of PoPs, in various cities, connected via high speed links. Each PoP consists itself of a number of backbone and access routers: the first connect to other PoPs and the latter connect to clients and to backbone routers. We refer to all the links between those routers as 'logical' links, or IP links or just links. This logical IP network is layered over a dense wavelength-division multiplexing (DWDM) optical infrastructure with SONET framing.[1]

B. Failures With an Impact on IP Connectivity

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], [11].

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.

C. Collecting and Processing ISIS Updates

We use the Python Routing Toolkit (PyRT)[2] to collect IS-IS Link State PDUs (LSPs) from our backbone. PyRT includes an IS-IS "listener" that collects LSPs from an IS-IS enabled router over an Ethernet link. The router treats the listener in the same way as adjacent routers: 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, provided that there are no partitions of the network due to failures. 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 were collected from a listener at a Sprint backbone 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 conforming 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.

The authors also note that IS-IS LSPs may contain weight changes that might give additional information; however, this has not been considered in this work.

D. 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 1 and October 21, 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 (which were 3.7% of all failures) are due to links being decommissioned rather than to accidental failures, and therefore we exclude them from the failures data set. Indeed, the usual time-to-repair for links in active use 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. The scale

[1]It is worth mentioning that the Sprint network is designed with sufficient redundancy and careful engineering to be robust in case that some links fail. For example, there are multiple parallel links connecting a pair of PoPs; furthermore these links are terminated on different router within the same PoP. Inside a single POP, backbone routers are connected in a fully meshed topology; each access router connects customers to two different backbone routers. Among all PoPs, a full mesh would be prohibitively expensive; however there is still a large degree of connectivity and the logical links are carefully chosen to use disjoint physical links. Finally, sufficient capacity is provisioned to carry the re-routed traffic in the case of failures.

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

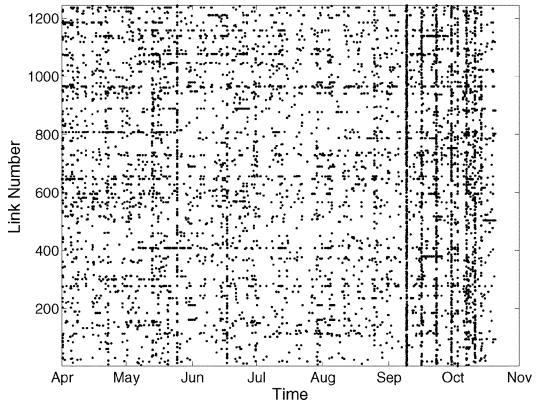


Fig. 1. Data set under study: failures in continental U.S. between April 1 and October 21, 2002.

of the figure is chosen to emphasize the horizontal and vertical patterns. We should mention that the 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.

Although we cannot report the absolute number of failures, for proprietary reasons, we can give a sense of the size and representativeness of the data set, by mentioning that it involved thousands of different links and hundreds of different routers, *i.e.*, on the order of the actual number of links and routers in the continental U.S. backbone. Failures were split roughly equally between inter- and intra-POP links, although there are ten times more intra- than inter-POP links in the Sprint backbone.

E. Additional Data Sets

The IS-IS logs is the main data set used in this paper to identify and characterize failures at the ISIS level. However, in our methodology, we also use two auxiliary data sets to confirm that some failures are optical-related. In Section IV-D, we use the IP-to-optical mapping and in Section IV-F we use SONET alarm logs; see the respective sections for a detailed description. These sets are not as complete (over the seven months period under study) as the main ISIS data set, and are used only for supplementary confirmation.

IV. Classification Methodology

This section describes our methodology for classifying failures according to the patterns observed in the IP-layer data. In some cases, we also attempt to infer their possible causes, using heuristic observations and evidence from supplementary data. However, the main purpose of this classification is *not* to accurate infer the cause of each failure. The main purpose is to partition the entire data set into smaller classes with common patterns at the IP layer. Once the classes are identified, and the statistics characterized, the interested user will be able to generate failures for each class separately and then superimpose.

A. Overview

Our approach is to use several hints obtained from the IS-IS failure data to identify groups of failures, and try to infer the different causes behind them when possible. A visual inspection of Fig. 1 provides insights into how to perform this classification. We observe that the failures are not uniformly scattered and there are vertical and horizontal lines. The vertical lines correspond to links that fail at the same time or to links that fail close in time 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 Window* 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: when several links fail at the same time, we call this group of failures an "event" of shared failures. Such events indicate that the involved links share a network component that fails. The shared 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: 1) how we decide whether a failure belongs to this specific class and 2) how we obtain partial confirmation for the inferred cause.

B. Weekly Maintenance Window

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 U.S. Sprint backbone network is Mondays 5 am-2 pm, UTC time. It turns out that failures during this window are responsible for the most prominent vertical lines in Fig. 1.

We would like to note that it is possible that some unplanned failures accidentally happened during the maintenance window. However, several facts confirm our maintenance hypothesis. First, we observed that the overwhelming majority of failures during the maintenance window were grouped into events on a few routers and happening inside the same POP(s), clearly indicating maintenance activities in these POPs. Second, the durations of these failures were in the order of tens of minutes, even hours, which could be due to router reboots and/or updates and human intervention. Third, we were able to confirm the scheduled operation activities for the most prominent vertical lines in September – October, which all happened during the weekly maintenance windows. Finally, the few unplanned failures that were possibly misclassified during the maintenance, have very low impact on users in the U.S., thanks to the choice of the maintenance window to be off peak-hours. Conversely,

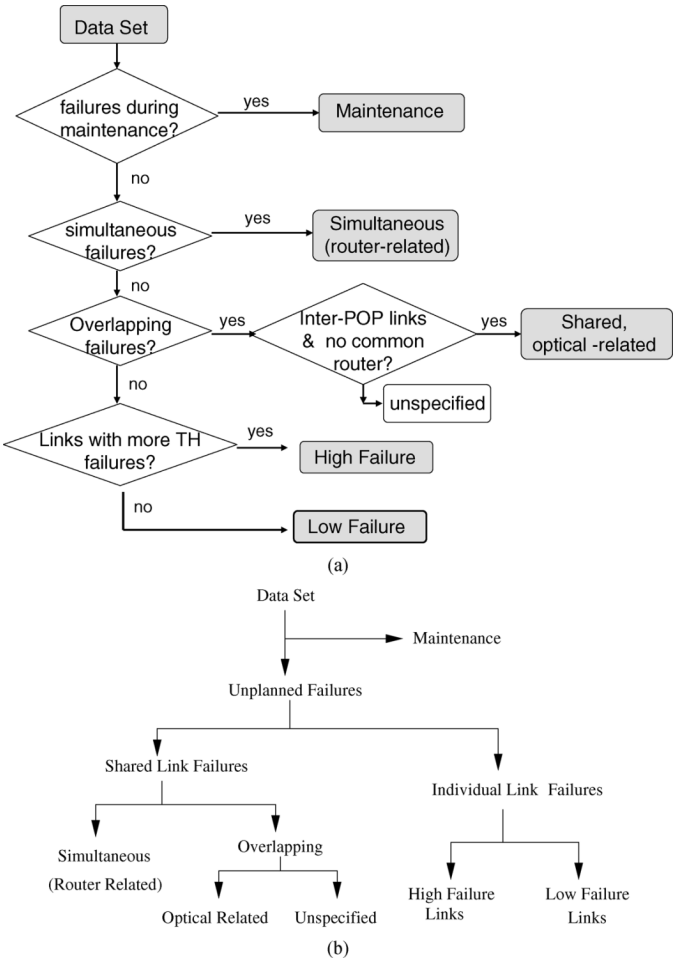


Fig. 2. Classification of failures. (a) Classification methodology. (b) Classification results.

it is possible that some maintenance events take place outside this window, and will be studied within the other categories.

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 same MRTD timestamp 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 MRTD 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 $T_{\rm start}$ (and 3 out of 4 come up at the exactly same time $T_{\rm end}$). We refer to such failures as simultaneous failures and we group them into events. Simultaneous failures start and/or finish at the exact same time, with an accuracy of microseconds. We conjectured that they are more likely to be due to a common cause (e.g., they may share a common component which fails

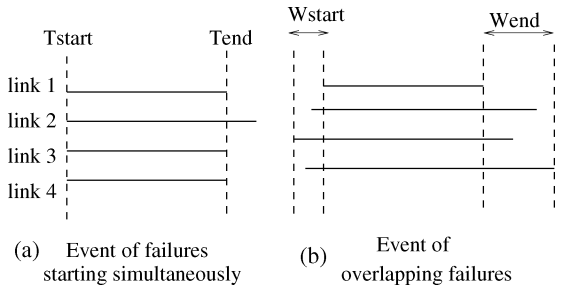


Fig. 3. Example events of simultaneous and overlapping failures. The time is according to the MRTD prepended at the IS–IS listener.

and causes all links to go down together) rather than to a coincidence. Furthermore, a router must have reported in the same LSP that these links go down together; our listener receives this LSP and prepends the same MRTD timestamp on all failures – that is why they appear to fail simultaneously.

We then checked this conjecture. 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 conversely, there is no simultaneous failure event that does not involve a common router, which confirms our intuition. Therefore, these events (simultaneous failures on links connected to the same router) are due either to problems on the common router (such as a router crash or reboot, a linecard failure or reset, a CPU overload, software or hardware error or human misconfiguration) or due to problems at the optical layer outside the router (if the failing links are carried over the same fiber). However, they are definitely *related* to this router for all modeling purposes. Therefore, in the rest of the paper, we refer to simultaneous failures as *Shared Router-Related*. Unfortunately, without any router logs it is not possible to determine for sure whether simultaneous failures are also *caused* by a failure in the common router.

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 min 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 four links start within $W_{\rm start}$ and finish within $W_{\rm end}$ 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 $W_{\rm start}$ and $W_{\rm end}$. 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 milliseconds up to a few seconds. The dominant ones are

TABLE I
SUMMARY OF OVERLAPPING EVENTS

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

the IS-IS carrier delay timer [3] with default 2 s to report a link going down and 12 s 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} matters for a meaningful definition of overlapping failures. If they 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 s to match the default timers used to report a link down or up respectively. We also varied W_{start} from 0.5 to 10 s and W_{end} from 0.5 to 20 s 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.

1) 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 check 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 "overlapping" event found in the database, we check whether *all (not just a subset of) links in the event share some optical components*. 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

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

are most likely due to the failure of some optical component shared by multiple IP links.

2) Unspecified: 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, because the cause of the problem is ambiguous: it could be a problem at the router or at the optical infrastructure. They also include overlapping failures of links in different PoPs, that clearly have no components in common and are due to coincidence. For all these events, we are not able to satisfactorily infer their causes and we call them *Unspecified*. We do not attempt to further analyze them as 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 for link l, where $l = 1, \ldots, L$. Let the maximum number of failures in a single link be $\max n = \max_l(n(l))$. For proprietary reasons, we show the normalized value $nn(l) = 100 \cdot n(l)/\max n$, 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 have the same slope; therefore, the interested reader is able to use the normalized value to simulate this behavior. The dashed lines intersect approximately at a point that corresponds to 2.5% of the links and to a normalized number of failures $nn(l) = 15.2$. We use this as the threshold ($THR = 15.2$) to distinguish between two sub-classes: the *High Failure Links* ($nn(l) \geq THR$) and the *Low Failure Links* ($1 \leq nn(l) \leq THR$).

We would like to emphasize that the distinction between high and low failure links is based on the number of failures, and not on the total downtime or time-to-repair, which is a different aspect and will be addressed in the Results section. A link failing for a long period of time, and another link failing repeatedly over a very short time period can impact the performance of the network in very different ways. In [26], we have followed-up on that and defined different metrics to measure the impact of such failures.

High failure links represent only 2.5% of all links but account for more than half of individual failures. All of them

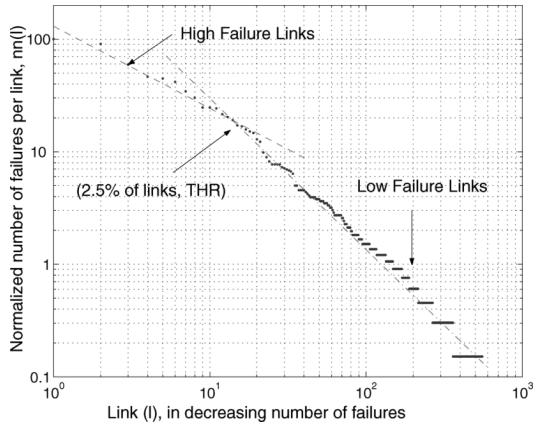


Fig. 4. Number of individual failures per link.

were backbone links; most of them were connected to different POPs; half of them had a router in common with at least another high-failure link. 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. The remaining half of individual failures happen on low failure links. Unlike all previous failure classes, low failure links do not have a prominent pattern either in time or across links.

It is difficult to infer the cause of individual failures based merely on the IS–IS data; unlike the shared failures, there is no timing correlation with other links to exploit. They could be due to optical or any other of the possible reasons. That is partially why we turned to additional data sets.

F. SONET Alarms

In order to further confirm our classification methodology (and, in particular, to confirm the shared-optical related failures and explore the individual failures), we tried to match the IS-IS failures with failures reported in signals by the SONET layer, which we will refer to as the SONET alarms. We had SONET alarm logs available for the four later months of the measurement period, from July to October 2002.

A number of alarms from the SONET layer can be recorded and timestamped by routers in the Sprint network. Of these alarms, the "Section Loss of Signal" (SLOS) is the most critical one and is triggered when a failure occurs in the optical layer. When a failure occurs at the optical layer, a SLOS alarm is generated. A router then waits for a small period of time, $T_{\rm dn}$, before reporting this failure to the IS-IS layer. This is done to damp out very short failures that disappear so as to avoid triggering route re-computation at the IP level. Similarly, when an optical layer fault is restored, a router receives a "SLOS cleared" signal. After this, the router waits for a period of time, $T_{\rm up}$, before reporting this to the IS-IS protocol. This again is done to damp out "flaps", when an optical layer fault appears and disappears several times and unnecessarily triggers route re-computation at the IP level. Typical values of for T_{dn} and $T_{\rm up}$ are 20 ms and 10 s, respectively; following the principle that bad news should travel fast and good news should travel slow.

In view of this damping mechanism, the correlation of IS-IS failures and SLOS alarms, *for the same link*, is done as follows. If an IS-IS link is reported to be down at time t, then we search for a SLOS alarm message at any time between $t-20$ ms and t (we make slight adjustments to compensate for marginal errors in timing). Similarly, when a IS-IS link is reported to be up, we search for a "SLOS cleared" message at any time between $t-10$ and t seconds.

TABLE III
% OF IS–IS FAILURES MATCHED WITH SLOS ALARMS FOR FOUR MONTHS

Failure Class	Jul	Aug	Sep	Oct
All failures	53%	54%	24%	35%
Shared Optical-related	*87%*	*70%*	*60%*	*69%*
Shared Router-related	42%	40%	34%	27%
Individual Low-Failure Links	47%	52%	30%	49%
All Individual (High+Low Failure)	41%	52%	20%	41%

Table III shows the percentage of IS-IS failures that were matched with SLOS signals for the failure classes defined by our classification methodology. Note that the percentages shown in this table should be interpreted as qualitative rather than as quantitative arguments, as the focus of this paper is on the ISIS failures and not the SONET alarms. for various reasons: first because of the ambiguities in inferring the cause, mentioned below; second because there are various optical-layer logs in addition to the SONET alarms; third, there are many SONET alarms, in addition to the critical SLOS signals.

We can make the following observations from Table III. The most important observation is that the large majority (up to 87%) of the shared failures classified by our methodology as 'shared optical-related' matches the SLOS alarms, which is a confirmation of their cause. This percentage is clearly higher than in any of the other classes. We believe that the small remaining percentage that is not matched to SLOS, could be matched to other SONET alarms or different optical-layer logs or could indicate some false positives in the classification. However, even this first step of matching ISIS failures to SLOS signals, sufficiently confirmed the cause of the large majority of shared optical-related failures. A second observation is that the shared failures classified as shared router-related have the lowest percentage matching the SLOS signals, as expected. There is however a nonzero percentage of them matching SLOS, because the optical layer could still affect two or more links of the same router sharing some optical component. Finally, it is not possible to infer the causes of Individual Link failures based solely on the ISIS data. They can be due to any of the possible reasons, including router or optical-layer problems. Therefore it is expected that some of the individual failures (up to 52%) match the SLOS signals. The percentages are similar for low-failure links and for the remaining high-failure links, which do not seem to have unusually high correlation with SLOS alarms.

V. FAILURE ANALYSIS

We now consider each class of failures separately and we study its characteristics that are useful for re-producing its behavior. First, we apply the classification methodology of the previous section and count how many failures fall in each class. Table IV shows the contribution of each class to the total number of failures. However, for proprietary reasons, we cannot report absolute numbers of failures, and we report normalized values (typically percentages) instead. Throughout the section, we note what it is a percentage of, as appropriate in different places.

Then, we study the properties of each one of the four classes and in particular: 1) the time-between-failures; 2) the distribution of failures across components (links or routers); 3) the

TABLE IV
PARTITIONING FAILURES INTO CLASSES

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

TABLE V
STATISTICAL CHARACTERIZATION OF EACH FAILURE CLASS

	Time between Failures	Number of failures per component	Time-to-repair	Number of links failing
Low-Failure Links	Weibull (network-wide)	Power-Law	Empirical	N/A
High-Failure Links	Empirical (per-link)	Power-Law	Empirical	N/A
Shared Router-Related	Weibull (network-wide)	Power-Law	Empirical	Empirical
Shared Optical-Related	Weibull (network-wide)	N/A	Empirical	Empirical

time-to-repair, and (iv) the number of links that fail together in a shared (router- or optical-related) event. Table V summarizes these properties for each class. Rows correspond to classes and columns correspond to properties. In the rest of this section, we discuss each class separately: we provide empirical distributions for the above three properties and, when possible, we also fit them to well-known distributions. In each subsection, we focus on one class and characterize how failures happen in time and across components (routers, links, etc). Fig. 5 also provides the empirical cumulative distribution function of time-to-repair for all classes. Notice that the frequency of failure has a larger effect on IP connectivity, because the network reacts and reconfigures itself shortly after a failure occurs; e.g., frequent but short failures may be more disruptive than a single long failure.

A failure model is specified by this information (% of failures per class and statistics for each class) and allows the interested reader to reproduce a realistic failure scenario. For example, one can generate failures according to the statistics of each class (how frequently a failure happens, on what router/link it happens and how long it lasts) separately and then superimpose failures from different classes.

A. Weekly Maintenance Window

20% of all failures happen during the window of 9-h 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.

More than half of the 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 cumulative distribution function (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

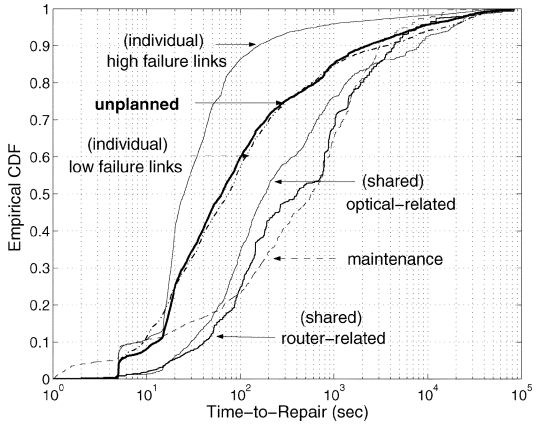


Fig. 5. CDF of the time-to-repair for each class of unplanned failures.

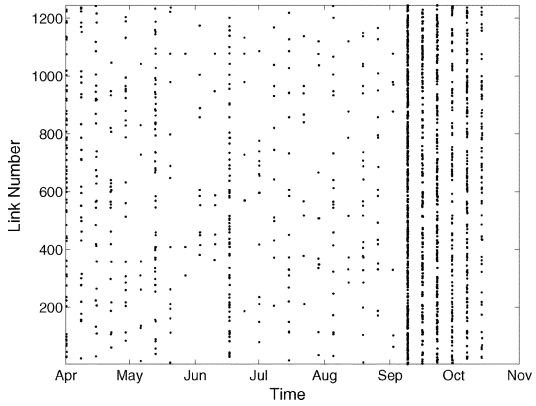


Fig. 6. Failures during weekly maintenance windows.

are router-related. A typical maintenance window for a given interface/router is one hour, although it can take less than that.

B. Router-Related Failures

Router-related events are responsible for 16.5% of the 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 connecting to 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 $nn(r) = 100 \cdot n(r)/\max n$ be its normalized value with respect to its maximum value $\max n = \max_r(n(r))$. Fig. 7 shows the normalized number of events per router, for routers ranked in decreasing number of events. 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. The mean time between events varies from five days up to several months, for different routers.

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 2–20 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 events are due to failures of the two links of access routers.

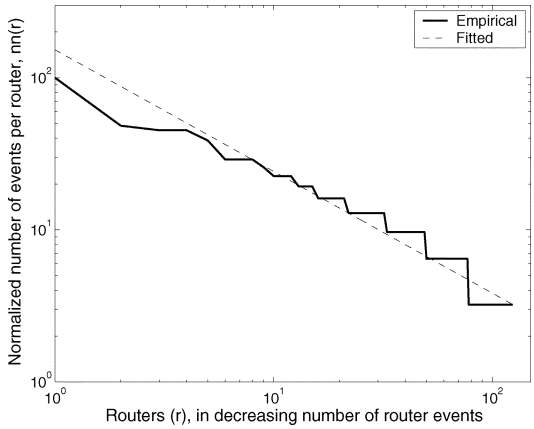


Fig. 7. Normalized number of events per router, in decreasing order.

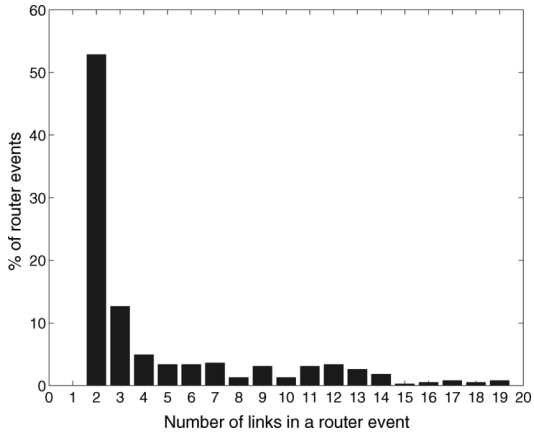


Fig. 8. Empirical PDF for the number of links in a router event.

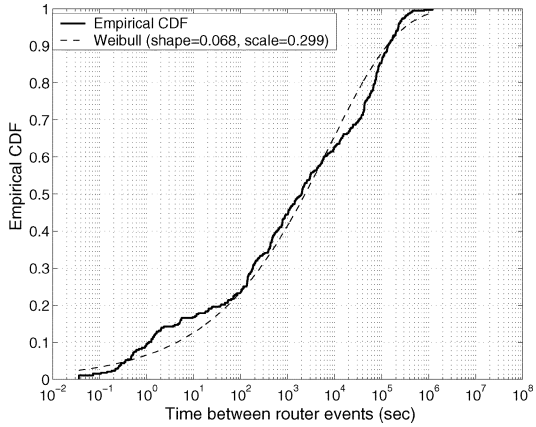


Fig. 9. CDF for the network-wide time between router events.

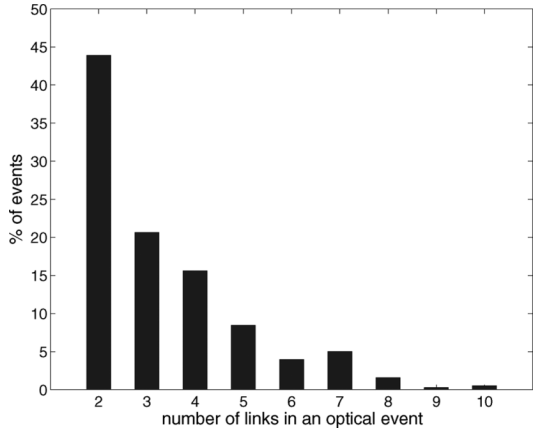


Fig. 10. Empirical pdf for the number of links in an optical event.

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 a separate figure, omitted here for lack of space, we noticed that the autocorrelation function decreases 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 VII.

C. Shared 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 (in order to overlap by definition) and at most 10 links in the same event. This is in agreement with sharing information derived from the IP-to-Optical mapping. E.g., 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).

The CDF of the *time-to-repair* for optical-related failures is shown in Fig. 5. Short time-to-repair is more likely due to faults in the optical components, while longer values correspond to fiber cuts or other failures that require human intervention to be repaired. Similarly to the previous classes of shared failures, the CDF is shifted towards larger values. 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.

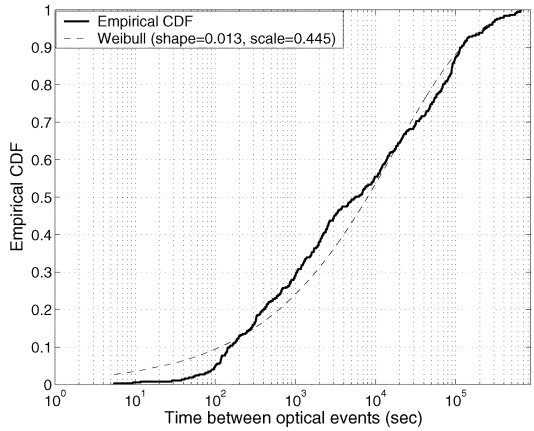


Fig. 11. CDF for the network-wide time between optical events.

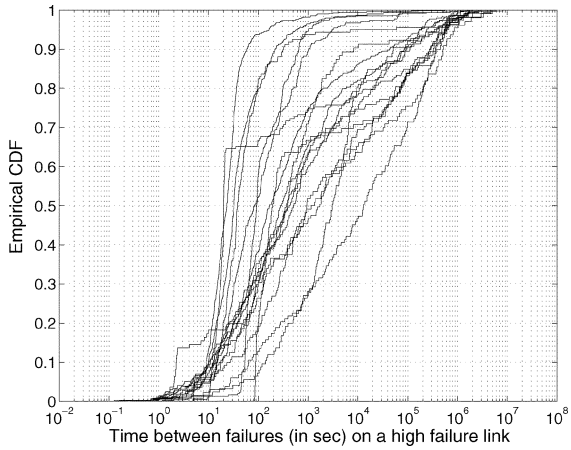


Fig. 12. Time between failures on each high failure link.

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). 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 IV.

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, as opposed to the previous classes that allowed only for a network-wide characterization.

The empirical CDF of the *time between failures* on each of the high failure links is shown in Fig. 12. There is a large heterogeneity among the behavior of the high-failure links. 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. These 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 maybe due to the high failure links being in an advanced stage of their life and their components being subject to intermittent and recurring faults.

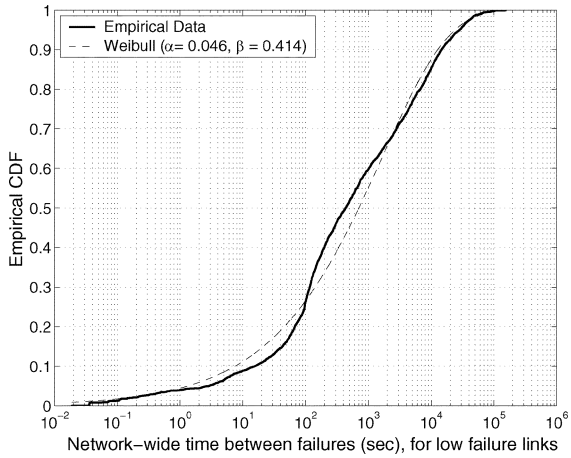


Fig. 13. Network-wide time between failures on low-failure links.

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*, i.e., the two failures may happen anywhere in the network and not necessarily on the same link. Fig. 13 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. We also looked at the autocorrelation function for the time between failures at the 90% confidence interval, omitted here for lack of space. We notice that correlation in the time between failures drops fast after a small lag. Therefore, as a first approximation, we can use independent and identically distributed (i.i.d.) Weibull r.v.

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. Failure Analysis Over Shorter Time Periods

The classification and characterization, so far, was based on the entire measurements period. It is also important to understand how properties remain constant or change with time. For this purpose, we now consider shorter, in particular one-month[3], time periods and we investigate to what extent the failure characteristics identified, based on the entire measurement period, still hold or change with time. For every one-month period, we now apply the same classification and characterization as we did for the entire time period. As far as the classification is concerned, we show how the total number of failures and their breakdown to the four failure classes (router-related, optical-related, high-failure links, low-failure links) varies with time; we find that most of the variability is due to the high-failure links. As far as the statistical characterization is concerned, we find that the same type of distributions (Weibull for the time-between-failures, power-laws for the heterogeneity of failures per network element, as well as the empirical distributions for the failure durations) still apply for each month. Furthermore, in the case of the low-failure links class, these distributions are estimated to have roughly the same parameters.

A. Number of Failures and Classification

First, we break the entire period into seven months and look at the number of failures (reported as the% of the total number of failures in the entire measurement period) and their causes for each month separately, see Fig. 14. The top bold line shows the number of failures in each month, including all causes. The four bottom thin lines show the number of failures per month for each class of failures.

We can make the following observations from this figure. First, the number of failures in the class of low-failure links does not vary much across months. Second, the number of failures in the high-failure links varies significantly from month to month. For example, in July and August there are no such failures, while in June and September they are the majority of failures. This is expected, as this class of failures corresponds to exceptionally rare events. Third, the shared link failures (router-related and optical-related) have consistently a smaller and less variable contribution. Finally, the top line is the sum of the four bottom lines and corresponds to all failures per month, including all causes. Its variability is mainly due to the month-to-month variability of high-failure links.

[3]As a first cut, we chose periods one-month long, because they are significantly shorter than the entire seven-month period to reveal time variability but they are still long enough for a sufficient number of failures to occur and allow for a statistical characterization. In general, the entire measurement period might need to be partitioned into time intervals of variable length, in order to better capture bursty behavior. This is a more general and difficult problem and is not addressed here.

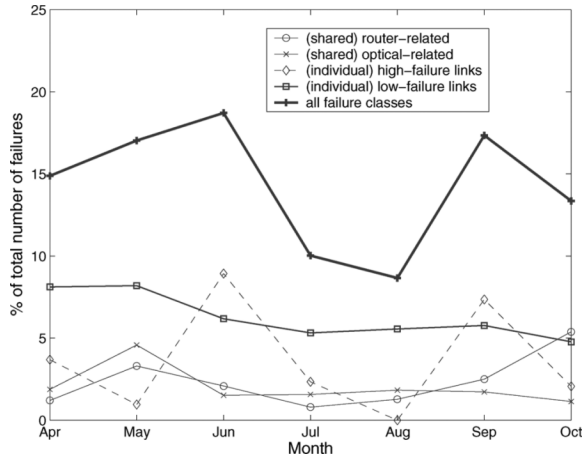


Fig. 14. Number of failures per month, per class and aggregate, (reported as the percentage of the total number of failures in the entire measurement period).

B. Statistical Characterization

We now look at the statistical characteristics of three classes: low-failure links, high-failure links, and router-related. We exclude the high-failure links, because both the small number of high failures per month and the variability from month-to-month do not allow for a sufficient statistical characterization. For the remaining classes, we find that the distributions describing their characteristics (time-between-failures, heterogeneity of network elements, time-to-repair) also hold for each month, with some small variation in the parameters.

Let us first consider the low-failure links class. Fig. 15 shows the empirical distributions for each month separately (a thin line for each month) and for the entire measurement period (shown in a bold line). Fig. 15(a) shows the empirical distributions for the time-between-failures. We can observe visually that the curves for each month have the same shape and are close to the curve for the entire period. A Weibull distribution can be fitted to each one of them, using maximum likelihood estimation; the estimated parameters with 95% confidence interval are shown in Table VI. These parameters are close to each other and to those estimated for the entire measurement period. Fig. 15(b) shows the number of failures in each link (with at least one failure in a month) in decreasing number of failures, for each month separately. We see that the curve for each month can be approximated by a straight line with roughly the same slope as the slope for the entire period. Finally, Fig. 15(c) shows the empirical CDF for the time-to-repair a failure for each month. All CDFs (with the exception of August which seems to be an outlier) have the same shape and are close to the empirical CDF for the entire period.

Applying the same steps to the router-related and optical-related classes, we observe that the same distributions also apply in the per-month analysis. First, Weibull distributions describe well both the time-between-router-related and the time-between-optical-related events (network-wide) for each month. Second, power-laws describe well the router events per router, for each month. Finally, the empirical distributions of time-to-repair for each month have the usual consistent shape.

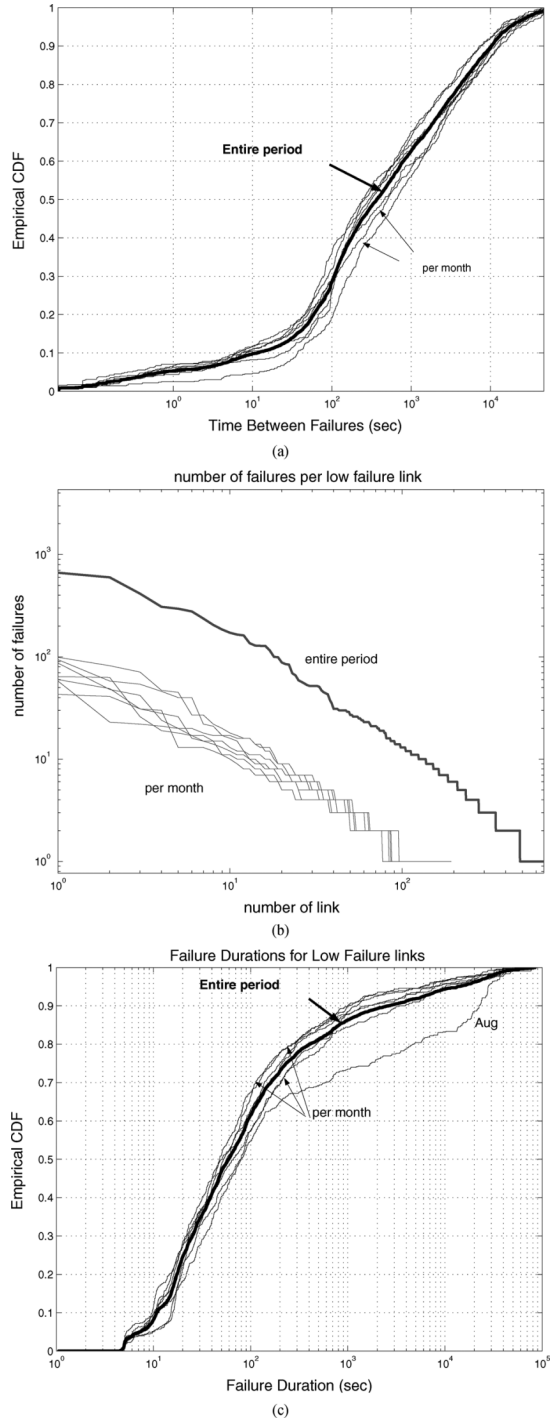


Fig. 15. Characterizing low-failure links, per month. (a) Time-between-failures. (b) Number of failures per low-failure link. (c) Failure durations.

However, and contrary to the low-failure class, the estimated parameters vary significantly from month-to-month. This can be explained by the small number of failures per month in the router- and optical-related failure classes.

VII. Discussion

This work offers a detailed characterization of link failures and 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. Furthermore, it provides a failure model, in terms of the statistical properties of each class. Such a model would be useful as an input to various engineering problems that need to account for failures.

TABLE VI
Time-Between-Failures on Low Failure Links and Over One-Month Periods. Weibull Parameters Estimated With 95% Confidence Interval

Month	Apr	May	Jun	Jul	Aug	Sep	Oct
α	0.044	0.056	0.026	0.056	0.063	0.046	0.043
β	0.450	0.421	0.490	0.403	0.388	0.423	0.438

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. 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. Our backbone failure model consists of the superposition of the models for each class. 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 IS–IS data set, discussed in detail in Section IV. In the same section, we used the IP-to-Optical database and confirmed to a 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 I.

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 shared router-related (Fig. 9, with parameters $\alpha = 0.068, \beta = 0.299$), the shared optical-related (Fig. 11, with parameters $\alpha = 0.013, \beta = 0.445$) and the low- failure links (Fig. 13, with parameters $\alpha = 0.046, \beta = 0.414$). The Weibull distribution has been found widely applicable in reliability engineering to describe the lifetime of components, primarily due to its versatile shape [27]. In addition, the Weibull distribution is derived as an extreme value distribution: for a large number of identical and independent components, the time to the first failure follows a Weibull, e.g., see [27]. This provides a theoretical justification for 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.

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, with slope -0.8) as well as for the number of individual failures per high or low failure link (see Fig. 4, with slope -1.35). 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 [28]. The Internet has also been found to display a number of power-law distributions [29], [30].

As an example of using the failure model, consider the class of low-failure links in Section V-E and let us try to regenerate failures with similar statistics as the measured ones. To decide when the next failure happens, one can pick a random number from the Weibull distribution for the network-wide time between failures. To decide on which link this failure happens, one could pick a link using the power-law distribution. We can use the power-law to assign a failure to a link as follows. If we observe a link l over a long period of time ($T = 7$ months) and find that it suffers n_l failures, the failure probability p_l is proportional to n_l/T. If the links are independent and n_1, $n_2, \ldots n_L$ are given (here follow the power-law), so do p_1, $p_2, \ldots p_L$. Given that a failure happens, it happens on link l with probability $p_1/(p_1 + \cdots + p_L) = n_1/(n_1 + \cdots + n_L)$ which can now be easily calculated. Similar steps can be followed to reproduce the router events using the network-wide time between events and the distribution of events across routers.

Different networks may vary in their topology, design, maintenance, technology, age and other specific traits. However, our two main contributions of this paper, namely the classification methodology and the resulting failure model, are useful in a general context. First, our classification methodology can be applied to any network using only IP-level failure logs. Indeed, the paper describes the data set that would be needed (just IS–IS logs) and the heuristics to derive a failure model from any network that is using OSPF or IS–IS. Operators could use our method to derive a failure model from their own network (i.e., to classify failures at the IP layer and estimate the distribution parameters for each class). Second, we applied our methodology to the Sprint's backbone network and estimated the parameters of the distributions based on the failures measured on this network. Our analysis provides a realistic data point that can be used as input to simulation/analytical studies, e.g., to evaluate a new protocol/architecture targeting the IP core.

VIII. Conclusion

A variety of failures or at below the IP layer, can potentially result in loss of IP connectivity. In this paper, we analyze seven months of IS-IS routing updates from the Sprint's IP backbone to characterize link failures. We classify failures according to their characteristics and we infer their probable causes. Our findings indicate that failures are part of the everyday operation: 20% of them are happen during a period of scheduled maintenance, 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 provides a better understanding of the nature and the extent of link failures, and a statistical failure model that can be used as input to many network design and traffic engineering problems.

Acknowledgment

The authors would like to thank R. Mortier for contributing his PyRT listener software; R. Gass and E. Kress for their help in the routing data collection; and their colleagues at SprintLink Operations for allowing them to collect data on Sprint's backbone and for providing their invaluable feedback.

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 Mag.*, vol. 17, no. 6, pp. 6–16, Nov.-Dec. 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 J. Sel. Areas Commun.*, vol. 21, no. 6, pp. 908–921, Aug. 2003.

[3] G. Iannaccone, C.-N. Chuah, R. Mortier, S. Bhattacharyya, and C. Diot, "Analysis of link failures in an IP backbone," in *Proc. ACM Internet Measurement Workshop*, Marseilles, France, Nov. 2002, pp. 237–242.

[4] G. Iannaccone, C.-N. Chuah, S. Bhattacharyya, and C. Diot, "Feasibility of IP restoration in a tier-1 backbone," *IEEE Netw.*, vol. 18, no. 2, pp. 13–19, Mar. 2004.

[5] D. Oran, "OSI IS-IS Intra-Domain Routing Protocol," RFC 1142, 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. IEEE INFOCOM*, San Francisco, CA, Mar. 2003, vol. 1, pp. 406–416.

[7] A. Markopoulou, G. Iannaccone, S. Bhatacharyya, C.-N. Chuah, and C. Diot, "Characterization of failures in an IP backbone," in *Proc. IEEE INFOCOM*, Hong Kong, Mar. 2004, vol. 4, pp. 2307–2317.

[8] C. Fraleigh, F. Tobagi, and C. Diot, "Provisioning IP backbone networks to support latency sensitive traffic," in *Proc. IEEE INFOCOM*, San Francisco, CA, Mar.-Apr. 2003, vol. 1, pp. 375–385.

[9] C. Boutremans, G. Iannaccone, and C. Diot, "Impact of link failures on VoIP performance," in *Proc. ACM NOSSDAV*, Miami Beach, FL, May 2002, pp. 63–71.

[10] 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.

[11] L. Sahasrabuddhe, S. Ramamurthy, and B. Mukherjee, "Fault management in IP-over-WDM networks: WDM protection versus IP restoration," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 1, pp. 21–33, Jan. 2002.

[12] A. Alaettinoglou and S. Casner, "Detailed analysis of ISIS Routing Protocol on the Qwest backbone," NANOG [Online]. Available: http://www.nanog.org/mtg-0202/ppt/cengiz.pdf

[13] A. Nucci, B. Schroeder, S. Bhattacharyya, N. Taft, and C. Diot, "IGP link weight assignment for transient link failures," in *Proc. 18th Int. Teletraffic Congr.*, Berlin, Germany, Sep. 2003.

[14] B. Fortz and M. Thorup, "Optimizing OSPF/IS-IS weights in a changing world," *IEEE J Sel. Areas Commun.*, vol. 20, no. 4, pp. 756–767, Apr. 2002.

[15] M. Durvy, C. Diot, N. Taft, and P. Thiran, "Network availability based service differentiation," in *Proc. IWQoS*, Monterey, CA, Jun. 2003.

[16] S. Nelakuditi, S. Lee, Y. Yu, Z.-L. Zhang, and C.-N. Chuah, "Fast local rerouting for handling transient link failures," *IEEE/ACM Trans. Netw.*, vol. 15, no. 2, pp. 359–372, Apr. 2007.

[17] V. Paxson, "End-to-end routing behavior in the Internet," *IEEE/ACM Trans. Netw.*, vol. 5, no. 5, pp. 601–615, Oct. 1997.

[18] Y. Zhang, V. Paxson, and S. Shenker, "The stationarity of Internet path properties: Routing, loss and throughput," Tech. Rep. ICIR, 2000 [Online]. Available: http://www.icir.org/

[19] M. Dahlin, B. Chandra, L. Gao, and A. Nayate, "End-to-end WAN service availability," *IEEE/ACM Trans. Netw.*, vol. 11, no. 2, pp. 300–313, Apr. 2003.

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

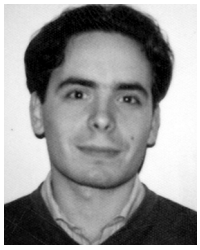
[21] D. Watson, F. Jahanian, and C. Labovitz, "Experiences with monitoring OSPF on a regional service provider network," in *Proc. IEEE ICDCS*, May 2003.

[22] A. Shaikh, C. Isett, A. Greenberg, M. Roughan, and J. Gottlieb, "A case study of OSPF behavior in a large enterprise network," in *Proc. ACM IMW*, Marseille, France, Nov. 2002, pp. 217–230.
[23] R. R. Kompella, J. Yates, and A. Greenberg, "IP fault localization via risk modeling," in *Proc. ACM/USENIX NSDI*, Apr. 2005.
[24] S. Kandula, D. Katabi, and J.-P. Vasseur, "Shrink: A tool for failure diagnosis in IP networks," in *ACM SIGCOMM Workshop on Mining Network Data (MineNet-05)*, Philadelphia, PA, Aug. 2005, pp. 173–178.
[25] M. Steinder and A. Sethi, "Increasing robustness of fault localization through analysis of lost, spurious and positive symptoms," in *Proc. IEEE INFOCOM*, New York, NY, Jun. 2002, vol. 1, pp. 322–331.
[26] Y. Ganjali, S. Bhattacharyya, and C. Diot, "Limiting the impact of failures on network performance," Sprint ATL Tech. Res. Rep. RR04-ATL-020666, 2003.
[27] P. Tobias and D. Trindade, *Applied Reliability*, 2nd ed. London, U.K.: Chapman Hall/CRC, 1995.
[28] L. Adamic, "Zipf, power-laws and Pareto: A ranking tutorial," Xerox Palo Alto Research Center, Palo Alto, CA [Online]. Available: http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html
[29] G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos, "Power-laws and the AS-level Internet topology," *IEEE/ACM Trans. Netw.*, vol. 11, no. 4, pp. 514–524, Aug. 2003.
[30] 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. ACM SIGCOMM*, Cambridge, MA, Sep. 1999, pp. 301–303.

Athina Markopoulou (S'98–M'02) received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Greece, in 1996, and the M.S. and Ph.D. degrees, both in electrical engineering, from Stanford University in 1998 and 2002, respectively.

Prior to joining UCI, she was a postdoctoral research fellow at Stanford University, and a member of the technical staff at Sprint Advanced Tech. Labs and Arastra Inc. She is currently an Assistant Professor with the EECS Department, University of California at Irvine (UCI). Her research interests include voice and video over IP networks, Internet Denial-of-Service, network measurement and control, and applications of network coding.

Dr. Markopoulou received the NSF CAREER award in 2008.



Gianluca Iannaccone (M'98) received the B.S. and M.S. degrees in 1998 and the Ph.D. degree in 2002 from the University of Pisa, Italy, all in computer engineering.

He joined Sprint as a research scientist in October 2001 working on network performance measurements, loss inference methods, and survivability of IP networks. In September 2003, he joined Intel Research, Berkeley, CA. His current interests are in network data mining, monitoring and management and the design of high-speed routers using off-the-shelf components.

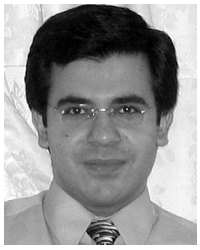
Supratik Bhattacharyya received the Ph.D. degree in computer science from the University of Massachusetts, Amherst, in 1999.

From 1999 to 2006, he was a Distinguished Member of Technical Staff at Sprint Advanced Technology Labs in Burlingame, CA. His research at Sprint covered a broad range of topics such as Internet routing, networking monitoring and fault tolerance, data mining and streaming, and disruption tolerant wireless services. He is currently a co-founder at SnapTell Inc., Mountain View, CA, a Silicon Valley startup specialized in mobile services based on cutting-edge image processing.

Chen-Nee Chuah (S'92–M'01) received the B.S. degree in electrical engineering from Rutgers University, Piscataway, NJ, and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1997 and 2001, respectively.

She is currently an Associate Professor in the Electrical and Computer Engineering Department at the University of California, Davis (UCD). Before joining UCD, she held a visiting researcher position at Sprint Advanced Technology Laboratories. Her research interests include Internet measurements, network management, overlay/peer-to-peer systems, network security, wireless/mobile networking, and opportunistic communications. She has served on the technical program committee of several ACM and IEEE conferences.

Dr. Chuah received the NSF CAREER Award in 2003 and the UCD College of Engineering Outstanding Junior Faculty Award in 2004.



Yashar Ganjali (S'03–M'07) received the B.Sc. degree in computer engineering from Sharif University of Technology, Tehran, Iran, in 1999, and the M.Sc. degree in computer science from the University of Waterloo, Waterloo, Canada, in 2001. He joined the High Performance Networking Group at Stanford University, and received the Ph.D. degree in electrical engineering in 2007.

Since January 2007, he has been a faculty member of the Computer Science Department, University of Toronto, Toronto, Canada. His research interests include packet switching architectures/algorithms, wireless networking and optical networking.

Christophe Diot received the Ph.D. degree from INP Grenoble, France, in 1991.

He was with INRIA Sophia-Antipolis, France, from October 1993 to September 1998, with Sprint, Burlingame, CA, from October 1998 to April 2003, and with Intel Research, Cambridge, U.K., from May 2003 to September 2005. He joined Thomson in October 2005 to start and manage the Paris Research Lab, Paris, France (http://parislab.thomson.net). His research activities focus on communication services and platforms for the future. He is a Fellow of the ACM.