Characterization of failures in n IP backbone

A. Markopolou, G. Iannacconne, S. Bhattacharyya, C-N Chuah and C. Diot

INFOCOM 2004
Failures in a backbone

• Little is known about failures in backbones
  – Typically well-engineered and adequately provisioned
    • Few packet losses, negligible queuing delays
  – Lack of failure data

• Sprint uses an IP-level restoration approach for dealing with failures
  – Using IS-IS for routing traffic
  – Over a highly meshed topology
IS-IS overview

• A link-state interior gateway protocol
• Reliably floods link state information throughout a network of routers
• Uses shortest-path-first to determine routes
• When an IP link fails, IS-IS automatically recomputes alternative routes around it, if such routes exists
  – Still packets drops by routers lacking up-to-date info
  – Backup paths that get overloaded and drop packets
This work

• Analyze IS-IS routing updates from Sprint’s IP network to characterize failures that affect IP connectivity

• A classification of failures types

• Analysis of failures and statistical characteristics of each class
This work

• IS-IS routing updates from Sprint’s IP network
  – Link State PDUs (LSPs)
  – Collected using PyRT from Sprint backbone
  – Six month of data April-October 2002

• Failures
  – When IP level connectivity between two connected routers is lost (“link failure”), each router broadcasts a “link down” LSP through the network
  – LSPs reporting loss or restoration may not reach listener at the same time
    • Link down time, start of failure – time of first arrival
    • Link up time, end of failure – time of last arrival
    • Time to repair – time between start and end of failure
Failure dataset and classification

Not uniformly scattered failures

A problematic link

A problematic time, maintenance?

Our classification of failures is summarized in Fig. 2 and consists of the following steps. We first separate failures due to individual links and then infer the cause of each failure. The main purpose is to partition the entire data set into smaller classes with common causes behind them when possible. A visual inspection of the figure is chosen to emphasize the horizontal and vertical correlation. We observe that the failures are not uniformly scattered and correspond to the roughly uniform plot that remains after excluding all the lines in September – October, which all happened during the maintenance window.

We now consider each class separately and describe: 1) how to perform this classification and 2) how to infer the cause of each failure. We divide links with individual failures into two classes:

- **Individual Link Failures**: a link fails due to an individual cause.
- **Shared Link Failures**: the involved links share a network component that fails.

We decide whether a failure belongs to this specific class and infer the cause of each failure. The main purpose of this classification is to classify failures for each class separately and then superimpose. Therefore, we further classify shared failures into three classes:

- **Router-Related**: 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). This class includes failures where the cause is related to the router.
- **Optical-Related**: failures that were possibly misclassified during the maintenance, in the optical layer. We should mention that the times-to-repair are not represented in the figure and the area covered by the dots represents the representative failures for each class separately and then superimpose. In this way, the interested user will be able to generate statistics characterized, the interested user will be able to generate statistics characterized, the interested user will be able to generate statistics characterized, the interested user will be able to generate statistics characterized, the interested user will be able to generate statistics characterized, the interested user will be able to generate statistics characterized.
Classification of failures

Data set

Unplanned failures

Maintenance

Unavoidable, Mondays 5am-2PM for Sprint

Shared link failures

Multiple links failing at once

Simultaneous (router related)

A router may report them in the same LSP

Optical related

Involve only inter-POP links and don’t share a router (use an IP-to-Optical mapping DB)

Overlapping

There's a delay from layers of protocols

Unspecified

Not sure so better unspecified; small %

Individual link failures

High failure links

Low failure links
Failures during maintenance windows

Figure 6: Failures during weekly maintenance windows.

20% of all failures
# Partition of failure classes

<table>
<thead>
<tr>
<th>Failure class</th>
<th>% of all</th>
<th>% of unplanned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Maintenance</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Unplanned</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>Shared</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Router related</td>
<td></td>
<td>16.5%</td>
</tr>
<tr>
<td>Optical-related</td>
<td></td>
<td>11.4%</td>
</tr>
<tr>
<td>Unspecified</td>
<td></td>
<td>2.9%</td>
</tr>
<tr>
<td>Individual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-failure links</td>
<td></td>
<td>38.5%</td>
</tr>
<tr>
<td>Low-failure links</td>
<td></td>
<td>30.7%</td>
</tr>
</tbody>
</table>
When a failure occurs at the optical layer, a SLOS alarm is generated and is triggered when a failure occurs in the optical layer. Among these alarms, the "Section Loss of Signal" (SLOS) is the most critical and timestamped by routers in the Sprint network. Of these failures, we tried to match the IS-IS failures with failures reported in signals by the SONET layer, and explore the individual failures, we tried to match the IS-IS failures with failures reported in signals by the SONET layer, and explore the individual failures.

We can make the following observations from Table III. The table shows the percentage of IS-IS failures that were classified as high failure (25% of IS–IS failures are high-failure links), low failure (75% of IS–IS failures are low-failure links), shared router-related failures, and shared optical-related failures. Table IV shows the contribution of each class to the total number of failures.

We now consider each class of failures separately and we efficiently confirmed the cause of the large majority of shared failures classified by our methodology as 'shared optical-related' matches the SLOS alarms, which is a confirmation of our classification methodology. Note that the percentages shown in this table should be interpreted as qualitative rather than as 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.

In order to further confirm our classification methodology, we studied the properties of each one of the four classes separately and we found that:

1. 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 their lifetime.
2. Low-failure link. Low failure links are backbone links; most of them were connected to different routers and did not appear to be critical to the network.
3. Shared router-related failures. A second observation is that the shared router-related failures have the lowest percentage matching the SLOS signals, as expected. There is how significant amount of shared router-related failures, including router or optical-layer problems. Therefore it is not possible to have unusually high correlation with SLOS alarms.
4. Shared optical-related failures. A second observation is that the shared optical-related failures account for a significant amount of the failures, typically 20% of all shared failures. They could be due to any of the possible reasons. That is partially explained by the fact that the IS-IS data is not reliable and unnecessarily triggers route re-computation at the IP level.

If an IS-IS link is reported to be down at time $t$, then it was down for a SLOS alarm message at any time between $t - 20$ ms and $t + 10$ s, respec-

Typical values of for power-laws are 20 ms and 10 s, respec-

$\frac{\text{normalized number of failures}}{\text{normalized number of failures per link}} \times 1000 = \frac{n(l)}{\max n} = 1000 \times \frac{n(l)}{\max n}$

The graph shows the normalized number of failures per link, $nn(l)$, as a function of the link index, $l$. The normalization is given by $nn(l) = 1000 \times \frac{n(l)}{\max n}$, where $n(l)$ is the number of failures for link $l$ and $\max n$ is the maximum number of failures among all links. The graph has a log-log scale, with the x-axis representing the link index $l$, and the y-axis representing the normalized number of failures per link, $nn(l)$. The graph shows two different power-laws: one for high failure links and another for low failure links. The high failure links account for 2.5% of all links and are responsible for more than 50% of individual failures. The low failure links account for 97.5% of all links and are responsible for less than 50% of individual failures. The graph also shows that the normalized number of failures per link decreases as the link index increases.
supports the observation that many of the maintenance failures close to the CDF for the router-related failures, which further (CDF) of time-to-repair for maintenance-related failures, is involve shutting down and (re)starting routers and interfaces. Section IV-C). This is expected as maintenance operations window are also router-related (according to the definition of the vertical lines in Fig. 1. Scheduled maintenance. It turns out that those account for many of a week. Fig. 6 shows the occurrence of link failures due to maintenance, although each such window accounts only for 5% A. Weekly Maintenance Window pose failures from different classes. It happens and how long it lasts) separately and then superim- each class (how frequently a failure happens, on what router/link ample, one can generate failures according to the statistics of interes reader to reproduce a realistic failure scenario. For ex- umes per class and statistics for each class) and allows the in- rues may be more disruptive than a single long failure. on IP connectivity, because the network reacts and reconfigures all classes. Notice that the frequency of failure has a larger effect empirical cumulative distribution function of time-to-repair for across components (routers, links, etc). Fig. 5 also provides the on one class and characterize how failures happen in time and them to well-known distributions. In each subsection, we focus discuss each class separately: we provide empirical distributions TABLE IV TABLE V F I E ACH C LASSES F AILURE C LASS

More than half of the failures during the maintenance 20% of all failures happen during the window of 9-h weekly number of events. The straight line in the log-log plot indicates that close to router related since maintenance require shutting down and restarting routers.
Problems are unevenly distributed across routers

Power-laws fit well the distribution of failures across components in the same class

Roughly a power-law

PDF of number of links in a router event (2-24 ports per linecard)
Frequency of events

Weibull has been found generally useful in reliability engineering to describe the lifetime of components.
High-failure links – failure freq

Potentially old equipment with intermittent and recurring faults

![Empirical CDF of time between failures on high failure links](image)
Time to repair

![Graph showing empirical cumulative distribution function of time-to-repair for different classes of failures.]

- **Unplanned** failures:
  - (individual) high failure links
  - (individual) low failure links

- **Maintenance** failures:
  - (shared) optical-related

- **Router-related** failures:
  - (shared) router-related

- **Routing-related** failures:
  - (individual) fast repair

Clearly a different class, fast repair.

Table IV

<table>
<thead>
<tr>
<th>Class</th>
<th>Time-to-Repair, sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintenance</td>
<td>10^0</td>
</tr>
<tr>
<td>Router</td>
<td>10^1</td>
</tr>
<tr>
<td>Optical</td>
<td>10^2</td>
</tr>
</tbody>
</table>

Table V

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintenance</td>
<td>12%</td>
</tr>
<tr>
<td>Router</td>
<td>16.5%</td>
</tr>
<tr>
<td>Optical</td>
<td>2%</td>
</tr>
</tbody>
</table>

More than half of the failures during the maintenance window of a week. Most events involve two links; 12% 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.
Optical-related failures

11.4% of all unplanned failures
Hard to recover from

At least 3 modes
[0, 100sec, 100-1,800sec, >1,800sec]

Many links affected

Persistent problems – closely spaced events over the same set of links

Mean of 12hrs
Range 5.5 secs – 7.5 days
supports the observation that many of the maintenance failures close to the CDF for the router-related failures, which further involve shutting down and (re)starting routers and interfaces. This is expected as maintenance operations window are also router-related (according to the definition of scheduled maintenance). It turns out that those account for many of a week. Fig. 6 shows the occurrence of link failures due to maintenance, although each such window accounts only for 5%.

A. Weekly Maintenance Window

A failure model is specified by this information (% of failures from different classes). Each class (how frequently a failure happens, on what router/link) may be more disruptive than a single long failure. One can generate failures according to the statistics of failures per class and statistics for each class) and allows the in-}

Fig. 5. CDF of the time-to-repair for each class of unplanned failures. Notice that the frequency of failure has a larger effect 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 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 different properties for each class. Rows correspond to classes and columns correspond to properties. In the rest of this section, we discuss each class separately.

TABLE IV

<table>
<thead>
<tr>
<th>Class</th>
<th>Time-to-repair</th>
<th>Number of links that fail together in unplanned failures</th>
<th>Maximum value parameters of the power-law using least-square method yields</th>
</tr>
</thead>
<tbody>
<tr>
<td>Router-related</td>
<td>~ cuts</td>
<td>~ switches</td>
<td>~ switches</td>
</tr>
<tr>
<td>Optical-related</td>
<td>~ cuts</td>
<td>~ switches</td>
<td>~ switches</td>
</tr>
<tr>
<td>Maintenance</td>
<td>~ cuts</td>
<td>~ switches</td>
<td>~ switches</td>
</tr>
</tbody>
</table>

B. Router-Related Failures

Router-related events are responsible for 16.5% of the unplanned failures from different classes. Most events involve two links; 12% of the maximum value parameters of the power-law using least-square method yields.
Low failure links

All the scattered dots – the norm

CDF matches that of unplanned – avg behavior

<table>
<thead>
<tr>
<th>Link Number</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200</td>
<td>May</td>
</tr>
<tr>
<td>1000</td>
<td>Jun</td>
</tr>
<tr>
<td>800</td>
<td>Jul</td>
</tr>
<tr>
<td>600</td>
<td>Aug</td>
</tr>
<tr>
<td>400</td>
<td>Sep</td>
</tr>
<tr>
<td>200</td>
<td>Oct</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time-to-Repair (sec)</th>
<th>10^0</th>
<th>10^1</th>
<th>10^2</th>
<th>10^3</th>
<th>10^4</th>
<th>10^5</th>
</tr>
</thead>
</table>
Conclusions

• A thorough analysis of seven months of IS-IS routing updates from a backbone
• Good/useful failure classification
• A first step toward a failure model
  – Much left undone – e.g., the dependence between time of occurrence and link/router of occurrence needs work
  – Parameters derived from one backbone network …
  – Variation in time, impact on network and service availability, …
• Increasingly relevant with our increase dependency on connectivity …