An Empirical Analysis of Intra- and Inter-Datacenter Network Failures for Geo-Distributed Services

[Extended Abstract]

Rahul Potharaju
Purdue University
rpothara@purdue.edu

Navendu Jain
Microsoft Research
navendu@microsoft.com

ABSTRACT

As cloud services continue to grow, a key requirement is delivering an 'always-on' experience to end users. Of the several factors affecting service availability, network failures in the hosting datacenters have received little attention. This paper presents a preliminary analysis of *intra-datacenter* and *inter-datacenter* network failures from a service perspective. We describe an empirical study analyzing and correlating network failure events over an year across multiple datacenters in a service provider. Our broader goal is to outline steps leveraging existing network mechanisms to improve end-to-end service availability.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Network**]: Network Operations

General Terms

Network Management, Availability, Reliability

Keywords

Data center networks, Geo-distributed services

1. INTRODUCTION

Cloud services are growing rapidly to provide a fast-response and an always-on experience to end users. Reliability is critically important for these services as failures not only hurt site availability and revenue, but also risk data loss. For example, the hurricane Sandy led to flooding of many data centers in NYC, taking down several major services such as The Huffington Post and Gawker [2]. Further, it caused failures of a large number of trans-atlantic fiber links peering from NYC significantly degrading capacity [10]. Last year, the entire US East region of Amazon became un-available due to a faulty fail-over during maintenance [1].

To increase service availability in a cost-effective manner, service providers are deploying their services across geo-distributed datacenters [5], and building their networks based on a scale-out design using inexpensive commodity hardware [7, 4]. However, as the number of devices and links in a datacenter grows, failures become the norm rather than the exception. Further, the network infrastructure also comprise long- haul links between datacenters whose failures can lead to loss of service traffic. Unfortunately, recent studies (e.g., [9, 3]) do not analyze network failures from a cloud service perspective nor do they examine inter- datacenter network failures.

In this extended abstract, we present a preliminary study of intra- and inter-datacenter network failures from a service perspective. Our study using real-world field data focuses on understanding the failure characteristics for a network stamp of a service comprising Top-of-Rack (ToR) switches connecting servers, aggregation switches (AGGs) that aggregate traffic from ToRs, and access routers (ARs) that in-turn connect to multiple AGGs. Specifically, we present two key aspects of failure characteristics: (a) problem root causes and (b) annualized downtime. Our broader goal is to use this study to provide new insights towards improving service availability.

2. METHODOLOGY

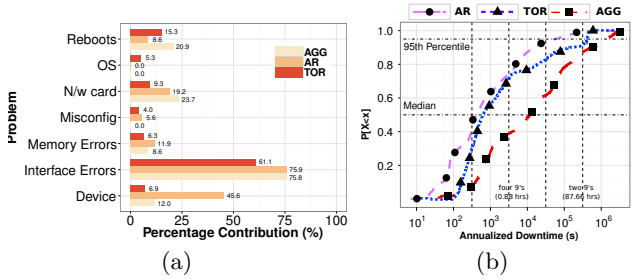
Our methodology is based on analyzing an year's worth (July 24, 2010-11) of network event logs across thousands of devices spanning multiple datacenters. Our data covers a wide range of network data sources, including syslog and SNMP alerts, network trouble tickets, maintenance tracking and revision control system, and traffic carried by links.

There are several challenges in using these network data sources to extract failures for our study such as (i) syslog messages can be spurious with devices logging 'down' notifications even when they are operational, (ii) redundant events resulting from two devices (e.g., neighbors) sending notifications for the same event, and (iii) multiple down and up messages getting logged as different events due to a flapping event.

To address these challenges, we build on our prior work [3, 8] in analyzing network failures. We define a *failure* of a network device or link as an event that causes it to be unavailable to carry traffic. To extract meaningful failures from network logs, we apply two stages of event filtering to analyze and correlate network event data sources.

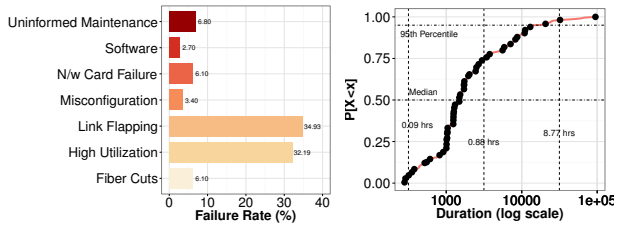
In the first stage, a *time-based filter* removes duplicate events by grouping all events with the same start and end time originating on the same interface. Further, for overlapping events on the same interface, it picks the earliest of start and end times. In the second stage, the *impact filter* identifies events impacting application performance in terms of throughput loss, number of failed connections or increased latency. Since we did not have access to application-level logs, we estimate failure impact by leveraging network traffic logs and computing the ratio of the median traffic on a

SIGMETRICS'13, June 17-21, 2013, Pittsburgh, PA, USA.
ACM 978-1-4503-1900-3/13/06.



(a) (b)

Figure 1: (a) Problems root-causes across device types, and (b) Comparing annualized downtime across ARs, ToRs and AGGs



(a) (b)

Figure 2: (a) Problem root-causes for long-haul links, and (b) Annualized downtime for long-haul links

failed device/link during a failure and its value in the recent past (e.g., preceding 8-hour window): a failure has impact if this ratio is less than one [3]. Finally, we apply NetSieve to do automated problem inference on network trouble tickets [8] for determining the problem root causes.

3. RESULTS

We study the annualized downtime and root causes of failures of network elements that significantly impact service availability.

3.1 Intra-datacenter failures

Figure 1(a) shows the histogram of the top-k problems observed obtained from network trouble tickets associated with intra- datacenter failures. Observe that many problems are due to both hardware issues and misconfigurations (e.g., ARP conflict). Interface-level errors, network card problems, and unexpected reloads were prominent amongst all three types. In addition, ToR failures were also due to OS-related problems and misconfigurations.

Next, we present the annualized downtime (see Figure 1(b)). Surprisingly, AGGs exhibited the highest downtime while ARs the lowest. ToR failures tend to be relatively infrequent compared to ARs and AGGs and hence the lower downtime. A significant contribution to ToR downtime was due to a set of older generation devices that became susceptible to end-of-life (or wear-out) problems, a phenomenon explained by the well-known bathtub curve of hardware failures.

3.2 Inter-datacenter Failures

In this section, we analyze the root causes and annualized downtime on the links connecting datacenters.

We first analyze the network trouble tickets associated with link failures. We found that link flapping (e.g., due to BGP, OSPF protocol issues and convergence) dominates problem root causes (about 35%) in inter-datacenter links. Due to optical protection configured in some areas of the network, a physical layer problem might end up triggering an optical re-route, a technique which is used to reduce the bandwidth loss by shifting existing lightpaths to new wavelengths without changing their route. However, it incurs control overhead (of about 10 seconds), and, more important, the service in the rerouted lightpaths can get disrupted [6]. Depending on the protocol timers, such an event is observed as a "link flap", yet the true underlying issues could be an optical re-route, possibly in response to a fiber cut (e.g., due to construction, hunting, shark attack). Therefore, it is not possible in many cases to attribute the exact cause.

The second major root cause is high link utilization (about 32%). However, note that high utilization does not necessarily imply a physical circuit failure or take- down, but it may be an indicator of packet errors. Fiber cuts (ones that could be observed), configuration errors and unnotified maintenance were observed but they do not constitute a significant fraction (see Figure 2(a)).

Finally, we show the annualized downtime. We observe (see Figure 2(b)) that the average failure duration is about 1.27 hours while the median is 0.41 hours with the 95P value >3 hours, which is 2x-3x higher than the expected values.

References

[1] Amazon. Summary of the Amazon EC2 and Amazon RDS Service Disruption in the US East Region. `http://goo.gl/yUlTJ`, May 2011.

[2] S. G. and I. B. Websites Scramble as Hurricane Sandy Floods Data Centers. `http://goo.gl/zOXDb`, October 31 2012.

[3] P. Gill, N. Jain, and N. Nagappan. Understanding network failures in data centers: measurement, analysis, and implications. In *Proc. of Sigcomm*, 2011.

[4] A. Greenberg, J. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. Maltz, P. Patel, and S. Sengupta. Vl2: a scalable and flexible data center network. *ACM Sigcomm CCR*, 2009.

[5] D. C. Knowledge. Data center global expansion trend. `http://goo.gl/SOvtA`, November 2012.

[6] G. Mohan and C. Murthy. Lightpath restoration in wdm optical networks. *Network, IEEE*, 14(6), 2000.

[7] R. Niranjan Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat. Portland: a scalable fault-tolerant layer 2 data center network fabric. In *Sigcomm CCR*. ACM, 2009.

[8] R. Potharaju, N. Jain, and C. Nita-Rotaru. Juggling the jigsaw: Towards automated problem inference from network trouble tickets. In *Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation*, 2013.

[9] D. Turner, K. Levchenko, A. Snoeren, and S. Savage. California fault lines: understanding the causes and impact of network failures. In *ACM Sigcomm CCR*, 2010.

[10] S. Works. Hurricane Sandy - AC2 Transatlantic Cable Cut. `http://goo.gl/dywVO`, October 2012.