Network Traffic Characteristics of Data Centers in the Wild

Theophilus Benson*, Aditya Akella* and David A. Maltz†
*University of Wisconsin–Madison
†Microsoft Research–Redmond

ABSTRACT

Although there is tremendous interest in designing improved networks for data centers, very little is known about the network-level traffic characteristics of current data centers. In this paper, we conduct an empirical study of the network traffic in 10 data centers belonging to three different types of organizations, including university, enterprise, and cloud data centers. Our definition of cloud data centers includes not only data centers employed by large online service providers offering Internet-facing applications, but also data centers used to host data-intensive (MapReduce style) applications. We collect and analyze SNMP statistics, topology, and packet-level traces. We examine the range of applications deployed in these data centers and their placement, the flow-level and packet-level transmission properties of these applications, and their impact on network utilization, link utilization, congestion, and packet drops. We describe the implications of the observed traffic patterns for data center internal traffic engineering as well as for recently-proposed architectures for data center networks.

Categories and Subject Descriptors

C.4 [Performance of Systems]: Design studies; Performance attributes

General Terms

Design, Measurement, Performance

Keywords

Data center traffic, characterization

1. INTRODUCTION

A data center (DC) refers to any large, dedicated cluster of computers that is owned and operated by a single organization. Data centers of various sizes are being built and employed for a diverse set of purposes today. On the one hand, large universities and private enterprises are increasingly consolidating their IT services within on-site data centers containing a few hundred to a few thousand servers. On the other hand, large online service providers, such as Google, Microsoft, and Amazon, are rapidly building geographically diverse cloud data centers, often containing more than 10K servers, to offer a variety of cloud-based services such as E-mail, Web servers, storage, search, gaming, and Instant Messaging. These service providers also employ some of their data centers to run large-scale data-intensive tasks, such as indexing Web pages or analyzing large data-sets, often using variations of the MapReduce paradigm [6].

Despite the growing applicability of data centers in a wide variety of scenarios, there are very few systematic measurement studies [19, 3] of data center usage to guide practical issues in data center operations. Crucially, little is known about the key differences between different classes of data centers, specifically university campus data centers, private enterprise data centers, and cloud data centers (both those used for customer-facing applications and those used for large-scale data-intensives tasks).

While several aspects of data centers still need substantial empirical analysis, the specific focus of our work is on issues pertaining to a data center network's operation. We examine the sending/receiving patterns of applications running in data centers and the resulting link-level and network-level performance. A better understanding of these issues can lead to a variety of advancements, including traffic engineering mechanisms tailored to improve available capacity and reduce loss rates within data centers, mechanisms for improved quality-of-service, and even techniques for managing other crucial data center resources, such as energy consumption. Unfortunately, the few recent empirical studies [19, 3] of data center networks are quite limited in their scope, making their observations difficult to generalize and employ in practice.

In this paper, we study data collected from ten data centers to shed light on their network design and usage and to identify properties that can help improve operation of their networking substrate. The data centers we study include three university campus data centers, two private enterprise data centers, and five cloud data centers, three of which run a variety of Internet-facing applications while the remaining two predominantly run MapReduce workloads. Some of the data centers we study have been in operation for over 10 years, while others were commissioned much more recently. Our data includes SNMP link statistics for all data centers, fine-grained packet traces from select switches in four of the data centers, and detailed topology for five data centers. By studying different classes of data centers, we are able to shed light on the question of how similar or different they are in terms of their network usage, whether results taken from one class can be applied to the others, and whether different solutions will be needed for designing and managing the data centers' internal networks.

We perform a top-down analysis of the data centers, starting with

IMC'10, November 1–3, 2010, Melbourne, Australia.

the applications run in each data center and then drilling down to the applications' send and receive patterns and their network-level impact. Using packet traces, we first examine the type of applications running in each data center and their relative contribution to network traffic. We then examine the fine-grained sending patterns as captured by data transmission behavior at the packet and flow levels. We examine these patterns both in aggregate and at a per-application level. Finally, we use SNMP traces to examine the network-level impact in terms of link utilization, congestion, and packet drops, and the dependence of these properties on the location of the links in the network topology and on the time of day.

Our key empirical findings are the following:

- We see a wide variety of applications across the data centers, ranging from customer-facing applications, such as Web services, file stores, authentication services, Line-of-Business applications, and custom enterprise applications to data intensive applications, such as MapReduce and search indexing. We find that application placement is non-uniform across racks.
- Most flows in the data centers are small in size ($\leq$ 10K B), a significant fraction of which last under a few hundreds of milliseconds, and the number of active flows per second is under 10,000 per rack across all data centers.
- Despite the differences in the size and usage of the data centers, traffic originating from a rack in a data center is ON/OFF in nature with properties that fit heavy-tailed distributions.
- In the cloud data centers, a majority of traffic originated by servers (80%) stays within the rack. For the university and private enterprise data centers, most of the traffic (40-90%) leaves the rack and traverses the network's interconnect.
- Irrespective of the type, in most data centers, link utilizations are rather low in all layers but the core. In the core, we find that a subset of the core links often experience high utilization. Furthermore, the exact number of highly utilized core links varies over time, but never exceeds 25% of the core links in any data center.
- Losses occur within the data centers; however, losses are not localized to links with persistently high utilization. Instead, losses occur at links with low average utilization implicating momentary spikes as the primary cause of losses. We observe that the magnitude of losses is greater at the aggregation layer than at the edge or the core layers.
- We observe that link utilizations are subject to time-of-day and day-of-week effects across all data centers. However in many of the cloud data centers, the variations are nearly an order of magnitude more pronounced at core links than at edge and aggregation links.

To highlight the implications of our observations, we conclude the paper with an analysis of two data center network design issues that have received a lot of recent attention, namely, network bisection bandwidth and the use of centralized management techniques.

- Bisection Bandwidth: Recent data center network proposals have argued that data centers need high bisection bandwidth to support demanding applications. Our measurements show that only a fraction of the existing bisection capacity is likely to be utilized within a given time interval in all the data centers, even in the "worst case" where application instances are spread across racks rather than confined within a rack. This is true even for MapReduce data centers that see relatively higher utilization. From this, we conclude that load balancing mechanisms for spreading traffic across the existing links in the network's core can help manage occasional congestion, given the current applications used.
- Centralization Management: A few recent proposals [2, 14] have argued for centrally managing and scheduling network-wide transmissions to more effectively engineer data center traffic. Our measurements show that centralized approaches must employ parallelism and fast route computation heuristics to scale to the size of data centers today while supporting the application traffic patterns we observe in the data centers.

Data Center Study	Type of Data Center	Type of Apps	#of DCs Measured
Fat-tree [1]	Cloud	MapReduce	0
Hedera [2]	Cloud	MapReduce	0
Portland [22]	Cloud	MapReduce	0
BCube [13]	Cloud	MapReduce	0
DCell [16]	Cloud	MapReduce	0
VAL2 [11]	Cloud	MapReduce	1
MicroTE [4]	Cloud	MapReduce	1
Flyways [18]	Cloud	MapReduce	1
Optical switching [29]	Cloud	MapReduce	1
ECMP. study 1 [19]	Cloud	MapReduce	1
ECMP. study 2 [3]	Cloud	MapReduce Web Services	19
Elastic Tree [14]	ANY	Web Services	1
SPAIN [21]	Any	Any	0
Our work	Cloud Private Net Universities	MapReduce Webservices Distributed F'S	10

Table 1: Comparison of prior data center studies, including type of data center and application.

The rest of the paper is structured as follows: we present related work in Section 2 and in Section 3 describe the data centers studied, their high-level design, and typical uses. In Section 4, we describe the applications running in these data centers. In Section 5, we zoom into the microscopic properties of the various data centers. In Section 6, we examine the flow of traffic within data centers and the utilization of links across the various layers. We discuss the implications of our empirical insights in Section 7, and we summarize our findings in Section 8.

2. RELATED WORK

There is tremendous interest in designing improved networks for data centers [1, 2, 22, 13, 16, 11, 4, 18, 29, 14, 21]; however, such work and its evaluation is driven by only a few studies of data center traffic, and those studies are solely of huge (> 10K server) data centers, primarily running data mining, MapReduce jobs, or Web services. Table 1 summarizes the prior studies. From Table 1, we observe that many of the data architectures are evaluated without empirical data from data centers. For the architectures evaluated with empirical data, we find that these evaluations are performed with traces from cloud data centers. These observations imply that the actual performance of these techniques under various types of realistic data centers found in the wild (such as enterprise and university data centers) is unknown and thus we are motivated by this to conduct a broad study on the characteristics of data centers. Such a study will inform the design and evaluation of current and future data center techniques.

This paper analyzes the network traffic of the broadest set of data centers studied to date, including data centers running Web services and MapReduce applications, but also other common enterprise and campus data centers that provide file storage, authentication services, Line-of-Business applications, and other custom-written services. Thus, our work provides the information needed to evaluate data center network architecture proposals under the broad range of data center environments that exist.

Previous studies [19, 3] have focused on traffic patterns at coarse time-scales, reporting flow size distributions, number of concurrent connections, duration of congestion periods, and diurnal patterns. We extend these measures by considering additional issues, such as the applications employed in the different data centers, their transmission patterns at the packet and flow levels, their impact on link and network utilizations, and the prevalence of network hot-spots. This additional information is crucial to evaluating traffic engineering strategies and data center placement/scheduling proposals.

The closest prior works are [19] and [3]; the former focuses on a single MapReduce data centers, while the latter considers cloud data centers that host Web services as well as those running MapReduce. Neither study considers non-cloud data centers, such as enterprise and campus data centers, and neither provides as complete a picture of traffic patterns as this study. The key observations from Benson's study [3] are that utilizations are highest in the core but losses are highest at the edge. In our work, we augment these findings by examining the variations in link utilizations over time, the localization of losses to link, and the magnitude of losses over time. From Kandula's study [19], we learned that while most traffic in the cloud is restricted to within a rack and a significant number of hot-spots exist in the network. Our work supplements these results by quantifying the exact fraction of traffic that stays within a rack for a wide range of data centers. In addition, we quantify the number of hot-spots, show that losses are due to the underlying burstiness of traffic, and examine the flow level properties for university and private enterprise (both are classes of data centers ignored in Kandula's study [19]).

Our work complements prior work on measuring Internet traffic [20, 10, 25, 9, 8, 17] by presenting an equivalent study on the flow characteristics of applications and link utilizations within data centers. We find that data center traffic is statistically different from wide area traffic, and that such behavior has serious implications for the design and implementation of techniques for data center networks.

3. DATASETS AND OVERVIEW OF DATA CENTERS

In this paper, we analyze data-sets from 10 data centers, including 5 commercial cloud data centers, 2 private enterprise data centers, and 3 university campus data centers. For each of these data centers, we examine one or more of the following data-sets: network topology, packet traces from select switches, and SNMP polls from the interfaces of network switches. Table 2 summarizes the data collected from each data center, as well as some key properties.

Table 2 shows that the data centers vary in size, both in terms of the number of devices and the number of servers. Unsurprisingly, the largest data centers are used for commercial computing needs (all owned by a single entity), with the enterprise and university data centers being an order of magnitude smaller in terms of the number of devices.

The data centers also vary in their proximity to their users. The enterprise and university data centers are located in the western/mid-western U.S. and are hosted on the premises of the organizations to serve local users. In contrast, the commercial data centers are distributed around the world in the U.S., Europe, and South America. Their global placement reflects an inherent requirement for geo-diversity (reducing latency to users), geo-redundancy (avoiding strikes, wars, or fiber cuts in one part of the world), and regulatory constraints (some data can not be removed from the E.U. or U.S.).

Data Center Name	Number of Locations
EDU1	1
EDU2	1
EDU3	1
PRV2	4

Table 3: The number of packet trace collection locations for the data centers in which we were able to install packet sniffers.

In what follows, we first describe the data we collect. We then outline similarities and differences in key attributes of the data centers, including their usage profiles, and physical topology. We found that understanding these aspects is required to analyze the properties that we wish to measure in subsequent sections, such as application behavior and its impact on link-level and network-wide utilizations.

3.1 Data Collection

SNMP polls: For all of the data centers that we studied, we were able to poll the switches' SNMP MIBs for bytes-in and bytes-out at granularities ranging from 1 minute to 30 minutes. For the 5 commercial cloud data centers and the 2 private enterprises, we were able to poll for the number of packet discards as well.

For each data center, we collected SNMP data for at least 10 days. In some cases (e.g., EDU1, EDU2, EDU3, PRV1, PRV2, CLD1, CLD4), our SNMP data spans multiple weeks. The long time-span of our SNMP data allows us to observe time-of-day and day-of-week dependencies in network traffic.

Network Topology: For the private enterprises and university data centers, we obtained topology via the Cisco CDP protocol, which gives both the network topology as well as the link capacities. When this data is unavailable, as with the 5 cloud data centers, we analyze device configuration to derive properties of the topology, such as the relative capacities of links facing endhosts versus network-internal links versus WAN-facing links.

Packet traces: Finally, we collected packet traces from a few of the private enterprise and university data centers (Table 2). Our packet trace collection spans 12 hours over multiple days. Since it is difficult to instrument an entire data center, we selected a handful of locations at random per data center and installed sniffers on them. In Table 3, we present the number of sniffers per data center. In the smaller data centers (EDU1, EDU2, EDU3), we installed 1 sniffer. For the larger data center (PRV2), we installed 4 sniffers. All traces were captured using a Cisco port span. To account for delay introduced by the packet duplication mechanism and for endhost clock skew, we binned results from the spans into 10 microsecond bins.

3.2 High-level Usage of the Data Centers

In this section, we outline important high-level similarities and differences among the data centers we studied.

University data centers: These data centers serve the students and administrative staff of the university in question. They provide a variety of services, ranging from system back-ups to hosting distributed file systems, E-mail servers, Web services (administra-

Data Center Role	Data Center Name	Location	Age (Years) (Curr Ver/Total)	SNMP	Packet Traces	Topology	Number Devices	Number Servers	Over Subscription
Universities	EDU1	US-Mid	10	X	X	X	22	500	2:1
	EDU2	US-Mid	(7,20)	X	X	X	36	1093	47:1
	EDU3	US-Mid	N/A	X	X	X	1	147	147:1
Private	PRV1	US-Mid	(5,5)	X	X	X	96	1088	8:3
	PRV2	US-West	> 5	X	X	X	100	2000	48:10
Commercial	CLD1	US-West	> 5	X	X	X	562	10K	20:1
	CLD2	US-West	> 5	X	X	X	763	15K	20:1
	CLD3	US-East	> 5	X	X	X	612	12K	20:1
	CLD4	S. America	(3,3)	X	X	X	427	10K	20:1
	CLD5	S. America	(3,3)	X	X	X	427	10K	20:1

Table 2: Summary of the 10 data centers studied, including devices, types of information collected, and the number of servers.

tive sites and web portals for students and faculty), and multicast video streams. We provide the exact application mix in the next section. In talking to the network operators, we found that these data centers "organically" evolved over time, moving from a collection of devices in a storage closet to a dedicated room for servers and network devices. As the data centers reached capacity, the operators re-evaluated their design and architecture. Many operators chose to move to a more structured, two-layer topology and introduced server virtualization to reduce heating and power requirements while controlling data center size.

Private enterprises: The private enterprise IT data centers serve corporate users, developers, and a small number of customers. Unlike university data centers, the private enterprise data centers support a significant number of custom applications, in addition to hosting traditional services like Email, storage, and Web services. They often act as development testbeds, as well. These data centers are developed in a ground-up fashion, being designed specifically to support the demands of the enterprise. For instance, to satisfy the need to support administrative services and beta testing of database-dependent products, PRV1 commissioned the development of an in-house data center 5 years ago. PRV2 was designed over 5 years ago mostly to support custom Line-of-Business applications and to provide login servers for remote users.

Commercial cloud data centers: Unlike the first two classes of data centers, the commercial data centers cater to external users and offer support for a wide range of Internet-facing services, including: Instant Messaging, Webmail, search, indexing, and video. Additionally, the data centers host large internal systems that support the externally visible services, for example data mining, storage, and relational databases (e.g., for buddy lists). These data centers are often purpose-built to support a specific set of applications (e.g., with a particular topology or over-subscription ratio to some target application patterns), but there is also a tension to make them as general as possible so that the application mix can change over time as the usage evolves. CLD1, CLD2, CLD3 host a variety of applications, ranging from Instant Messaging and Webmail to advertisements and web portals. CLD4 and CLD5 are primarily used for running MapReduce style applications.

3.3 Topology and Composition of the Data Centers

In this section, we examine the differences and similarities in the physical construction of the data centers. Before proceeding to examine the physical topology of the data centers studied, we present a brief overview of the topology of a generic data center. In Figure 1, we present a canonical 3-Tiered data center. The 3 tiers of the data center are the edge tier, which consists of the Top-of-Rack switches that connect the servers to the data center's network fabric; the aggregation tier, which consists of devices that interconnect the

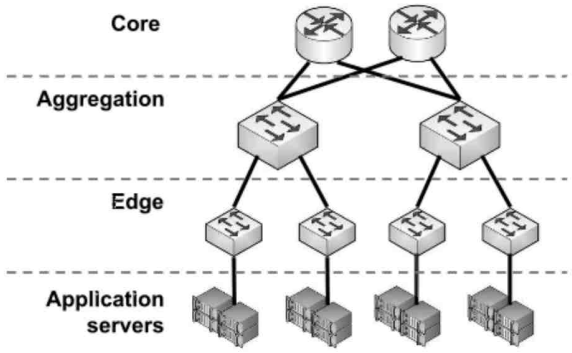


Figure 1: Canonical 3-Tier data center topology.

ToR switches in the edge layer; and the core tier, which consists of devices that connect the data center to the WAN. In smaller data centers, the core tier and the aggregation tier are collapsed into one tier, resulting in a 2-Tiered data center topology.

Now, we focus on topological structure and the key physical properties of the constituent devices and links. We find that the topology of the data center is often an accident of history. Some have regular patterns that could be leveraged for traffic engineering strategies like Valiant Load Balancing [11], while most would require either a significant upgrade or more general strategies.

Topology. Of the three university data centers, we find that two (EDU1, EDU2) have evolved into a structured 2-Tier architecture. The third (EDU3) uses a star-like topology with a high-capacity central switch interconnecting a collection of server racks – a design that has been used since the inception of this data center. As of this writing, the data center was migrating to a more structured set-up similar to the other two.

EDU1 uses a topology that is similar to a canonical 2-Tier architecture, with one key difference: while the canonical 2-Tier data centers use Top-of-Rack switches, where each switch connects to a rack of 20-80 servers or so, these two data centers utilize Middle-of-Rack switches that connect a row of 5 to 6 racks with the potential to connect from 120 to 180 servers. We find that similar conclusions hold for EDU2 (omitted for brevity).

The enterprise data centers do not deviate much from textbook-style constructions. In particular, the PRV1 enterprise data center utilizes a canonical 2-Tier Cisco architecture. The PRV2 data center utilizes a canonical 3-Tier Cisco architecture.

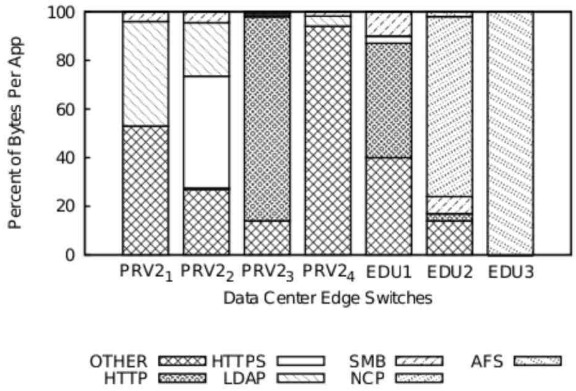


Figure 2: Classification of network traffic to application using Bro-Id. Each of the sniffers sees a very different mix of applications, even though the first 4 sniffers are located on different switches in the same data center.

Note that we do not have the physical topologies from the cloud data centers, although the operators of these data centers tell us that these networks uniformly employ the 3-Tier textbook data center architectures described in [11].

4. APPLICATIONS IN DATA CENTERS

We begin our "top-down" analysis of data centers by first focusing on the applications they run. In particular, we aim to answer the following questions: (1) What type of applications are running within these data centers? and, (2) What fraction of traffic originated by a switch is contributed by each application?

We employ packet trace data in this analysis and use Bro-Id [26] to perform application classification. Recall that we collected packet trace data for 7 switches spanning 4 data centers, namely, the university campus data centers, EDU1, EDU2, and EDU3, and a private enterprise data center, PRV2. To lend further weight to our observations, we spoke to the operators of each data center, including the 6 for which we did not have packet trace data. The operators provided us with additional information about the specific applications running in their data centers.

The type of applications found at each edge switch, along with their relative traffic volumes, are shown in Figure 2. Each bar corresponds to a sniffer in a data center, and the first 4 bars are from the 4 edges switches within the same data center (PRV2). In conversing with the operators, we discovered that this data center hosts a mixture of authentication services (labeled "LDAP"), 3-Tier Line-Of-Business Web applications (captured in "HTTP" and "HTTPS"), and custom home-brewed applications (captured in "Others").

By looking at the composition of the 4 bars for PRV2, we can infer how the services and applications are deployed across racks in the data center. We find that each of the edge switches monitored hosts a portion of the back-end for the custom applications (captured in "Others"). In particular, the rack corresponding to PRV2$_4$ appears to predominantly host custom applications that contribute over 90% of the traffic from this switch. At the other switches, these applications make up 50%, 25%, and 10% of the bytes, respectively.

Further, we find that the secure portions of the Line-of-Business Web services (labeled "HTTPS") are hosted in the rack corresponding to the edge switch PRV2$_2$, but not in the other three racks monitored. Authentication services (labeled "LDAP") are deployed across the racks corresponding to PRV2$_1$ and PRV2$_2$, which makes up a significant fraction of bytes from these switches (40% of the bytes from PRV2$_1$ and 25% of the byes from PRV2$_2$). A small amount of LDAP traffic (2% of all bytes on average) originates from the other two switches, as well, but this is mostly request traffic headed for the authentication services in PRV2$_1$ and PRV2$_2$.

Finally, the unsecured portions of the Line-of-Business (consisting of help pages and basic documentation) are located predominantly on the rack corresponding to the edge switch PRV2$_3$—nearly 85% of the traffic originating from this rack is HTTP.

We also see some amount of file-system traffic (SMB) across all the 4 switches (roughly 4% of the bytes on average).

Clustering of application components within this data center leads us to believe that emerging patterns of virtualization and consolidations have not yet led to applications being spread across the switches.

Next, we focus on the last 3 bars, which correspond to an edge switch each in the 3 university data centers, EDU1, EDU2 and EDU3. While these 3 data centers serve the same types of users we observe variations across the networks. Two of the university data centers, EDU2 and EDU3, seem to primarily utilize the network for distributed file systems traffic, namely AFS and NCP—AFS makes up nearly all the traffic seen at the EDU3 switch, while NCP constitutes nearly 80% of the traffic at the EDU2 switch. The traffic at the last data center, EDU1, is split 60/40 between Web services (both HTTP and HTTPS) and other applications such as file sharing (SMB). The operator of this data center tells us that the data center also hosts payroll and benefits applications, which are captured in "Others."

Note that we find file system traffic to constitute a more significant fraction of the switches in the university data centers we monitored compared to the enterprise data center.

The key take-aways from the above observations are that (1) There is a wide variety of applications observed both within and across data centers, such as "regular" and secure HTTP transactions, authentication services, file-system traffic, and custom applications and (2) We observe a wide variation in the composition of traffic originated by the switches in a given data center (see the 4 switches corresponding to PRV2). This implies that one cannot assume that applications are placed uniformly at random in data centers.

For the remaining data centers (i.e., PRV1, CLD1–5), where we did not have access to packet traces, we used information from operators to understand the application mix. CLD4 and CLD5 are utilized for running MapReduce jobs, with each job, scheduled to pack as many of its nodes as possible into the same rack to reduce demand on the data center's core interconnect. In contrast, CLD1, CLD2, and CLD3 host a variety of applications, ranging from messaging and Webmail to Web portals. Each of these applications is comprised of multiple components with intricate dependencies, deployed across the entire data center. For example, the Web portal requires access to an authentication service for verifying users, and it also requires access to a wide range of Web services from which data is aggregated. Instant Messaging similarly utilizes an authentication service and composes the user's buddy list by aggregating data spread across different data stores. The application mix found in the data centers impacts the traffic results, which we look at next.

5. APPLICATION COMMUNICATION PATTERNS

In the previous section, we described the set of applications running in each of the 10 data centers and observed that a variety of applications run in the data centers and that their placement is non-uniform. In this section, we analyze the aggregate network transmission behavior of the applications, both at the flow-level and at the finer-grained packet-level. Specifically, we aim to answer the following questions: (1) What are the aggregate characteristics of flow arrivals, sizes, and durations? and (2) What are the aggregate characteristics of the packet-level inter-arrival process across all applications in a rack — that is, how bursty are the transmission patterns of these applications? These aspects have important implications for the performance of the network and its links.

As before, we use the packet traces in our analysis.

5.1 Flow-Level Communication Characteristics

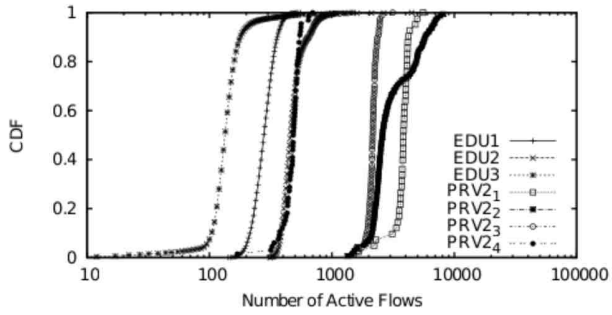
First, we examine the number of active flows across the 4 data centers where we have packet-level data, EDU1, EDU2, EDU3, and PRV2. To identify active flows, we use a long inactivity timeout of 60 seconds (similar to that used in previous measurements studies [19]).

In Figure 3(a), we present the distribution of the number of active flows within a one second bin, as seen at seven different switches within 4 data centers. We find that although the distribution varies across the data centers, the number of active flows at any given interval is less than 10,000. Based on the distributions, we group the 7 monitored switches into two classes. In the first class are all of the university data center switches EDU1, EDU2 and EDU3, and one of the switches from a private enterprise, namely $PRV2_4$, where the number of active flows is between 10 and 500 in 90% of the time intervals. In the second class, are the remaining switches from the enterprise, namely, $PRV2_1$, $PRV2_2$, and $PRV2_3$, where the number of active flows is between 1,000 and 5,000 about 90% of the time.

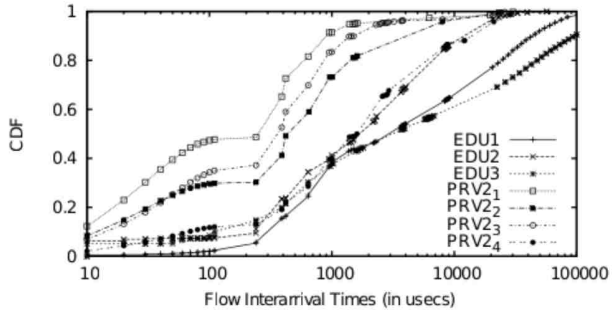
We examine the flow inter-arrival times in Figure 3(b). We find that the time between new flows arriving at the monitored switch is less than 10ms for 2-13% of the flows. For most of the switches in PRV2, 80% of the flows have an inter-arrival time under 1ms. This observation supports the results of a prior study [19] of a cloud data center. However, we found that this observation does not hold for the university data centers, where we see 80% of the flow inter-arrival times were between 4ms and 40ms, suggesting that these data centers have less churn than PRV2 and the previously studied cloud data center [19]. Among other issues, flow inter-arrival time affects what kinds of processing can be done for each new flow and the feasibility of logically centralized controllers for flow placement. We return to these questions in Section 7.

Next, we examine the distributions of flow sizes and and lengths in Figure 4(a) and (b), respectively. From Figure 4(a), we find that flow sizes are roughly similar across all the studied switches and data centers. Across the data centers, we note that 80% of the flows are smaller than 10KB in size. Most of the bytes are in the top 10% of large flows. From Figure 4(b), we find that for most of the data centers 80% of the flows are less than 11 seconds long. These results support the observations made in prior a study [19] of a cloud data center. However, we do note that the flows in EDU2 appear to be generally shorter and smaller than the flows in the other data centers. We believe this is due to the nature of the predominant application that accounts for over 70% of the bytes at the switch.

Finally, in Figure 5, we examine the distribution of packet sizes in the studied data centers. The packet sizes exhibit a bimodal pattern, with most packet sizes clustering around either 200 Bytes and 1400 Bytes. Surprisingly, we found application keep-alive packets as a major reason for the small packets, with TCP acknowledgments, as expected, being the other major contributor. Upon close inspection of the packet traces, we found that certain applications, including MSSQL, HTTP, and SMB, contributed more small packets than large packets. In one extreme case, we found an application producing 5 times as many small packets as large packets. This result speaks to how commonly persistent connections occur as a design feature in data center applications, and the importance of continually maintaining them.



(a)



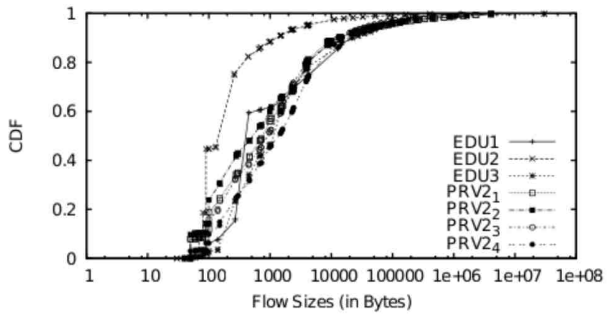
(b)

Figure 3: CDF of the distribution of the number of flows at the edge switch (a) and the arrival rate for flows (b) in EDU1, EDU2, EDU3, and PRV2.

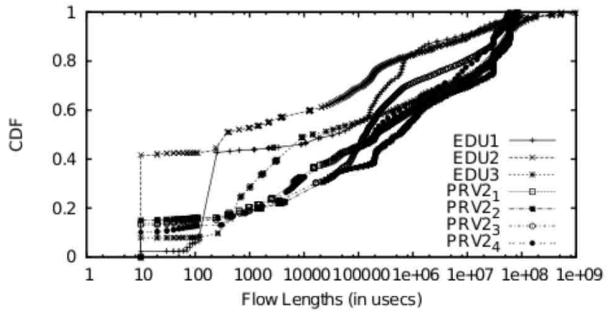
5.2 Packet-Level Communication Characteristics

We first examine the temporal characteristics of the packet traces. Figure 6 shows a time-series of packet arrivals observed at one of the sniffers in PRV2, and the packet arrivals exhibit an ON/OFF pattern at both 15ms and 100ms granularities. We observed similar traffic patterns at the remaining 6 switches as well.

Per-packet arrival process: Leveraging the observation that traffic is ON/OFF, we use a packet inter-arrival time threshold to identify the ON/OFF periods in the traces. Let $arrival_{95}$ be the 95th percentile value in the inter-arrival time distribution at a particular switch. We define a $period_{on}$ as the longest continual period during which all the packet inter-arrival times are smaller than $arrival_{95}$. Accordingly, a $period_{off}$ is a period between two ON periods. To characterize this ON/OFF traffic pattern, we focus on three aspects: (i) the durations of the ON periods, (ii) the durations



(a)



(b)

Figure 4: CDF of the distribution of the flow sizes (a) and of flow lengths (b) in PRV2, EDU1, EDU2, and EDU3.

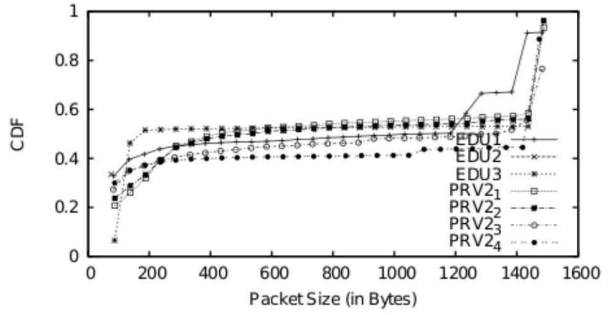
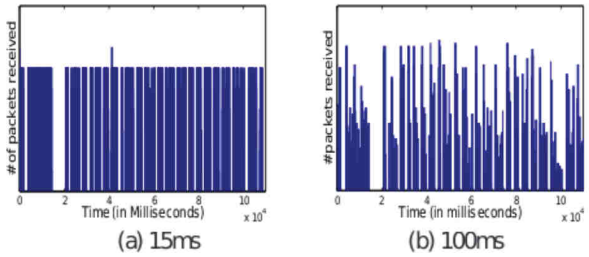


Figure 5: Distribution of packet size in the various networks.

of the OFF periods, and (iii) the packet inter-arrival times within ON periods.

Figure 7(a) shows the distribution of inter-arrival times within ON periods at one of the switches for PRV2. We bin the inter-arrival times according to the clock granularity of 10 μs. Note that the distribution has a positive skew and a heavy tail. We attempted to fit several heavy-tailed distributions and found that the lognormal curve produces the best fit with the least mean error. Figure 7(b)



(a) 15ms (b) 100ms

Figure 6: ON/OFF characteristics: Time series of Data Center traffic (number of packets per time) binned by two different time scales.

Data center	Off period Distribution	ON period Distribution	Interarrival Rate Distribution
$PRV2_1$	Lognormal	Lognormal	Lognormal
$PRV2_2$	Lognormal	Lognormal	Lognormal
$PRV2_3$	Lognormal	Lognormal	Lognormal
$PRV2_4$	Lognormal	Lognormal	Lognormal
EDU1	Lognormal	Weibull	Weibull
EDU2	Lognormal	Weibull	Weibull
EDU3	Lognormal	Weibull	Weibull

Table 4: The distribution for the parameters of each of the arrival processes of the various switches.

shows the distribution of the durations of ON periods. Similar to the inter-arrival time distribution, this ON period distribution also exhibits a positive skew and fits well with a lognormal curve. The same observation can be applied to the OFF period distribution as well, as shown in Figure 7(c).

We found qualitatively similar characteristics at the other 6 switches where packet traces were collected. However, in fitting a distribution to the packet traces (Table 4), we found that only the OFF period at the different switches consistently fit the lognormal distribution. For the ON periods and interarrival rates, we found that best distribution was either Weibull and lognormal, varying by data center.

Our findings indicate that certain positive skewed and heavy-tailed distributions can model data center switch traffic. This highlights a difference between the data center environment and the wide area network, where the long-tailed Pareto distribution typically shows the best fit [27, 24]. The differences between these distributions should be taken into account when attempting to apply models or techniques from wide area networking to data centers.

Per-application arrival process: Recall that the data centers in this analysis, namely, EDU1, EDU2, EDU3, and PRV2, are dominated by Web and distributed file-system traffic (Figure 2). We now examine the arrival processes for these dominant applications to see if they explain the aggregate arrival process at the corresponding switches. In Table 5, we present the distribution that best fits the arrival process for the dominant application. From this table, we notice that the dominant applications in the universities (EDU1, EDU2, EDU3), which account for 70–100% of the bytes at the respective switches, are indeed characterized by identical heavy-tailed distributions as the aggregate traffic. However, in the case of two of the PRV2 switches (#1 and #3), we find that the dominant application differs slightly from the aggregate behavior. Thus, in the general case, we find that simply relying on the characteristics of the most dominant applications is not sufficient to accurately model the aggregate arrival processes at data center edge switches.

Data center	Off period Distribution	Interarrival Rate Distribution	ON period Distribution	Dominant Applications
PRV2$_1$	Lognormal	Weibull	Exponential	Others
PRV2$_2$	Weibull	Lognormal	Lognormal	LDAP
PRV2$_3$	Weibull	Lognormal	Exponential	HTTP
PRV2$_4$	Lognormal	Lognormal	Weibull	Others
EDU1	Lognormal	Lognormal	Weibull	HTTP
EDU2	Lognormal	Weibull	Weibull	NCP
EDU3	Lognormal	Weibull	Weibull	AFS

Table 5: The distribution for the parameters of each of the arrival processes of the dominant applications on each switch.

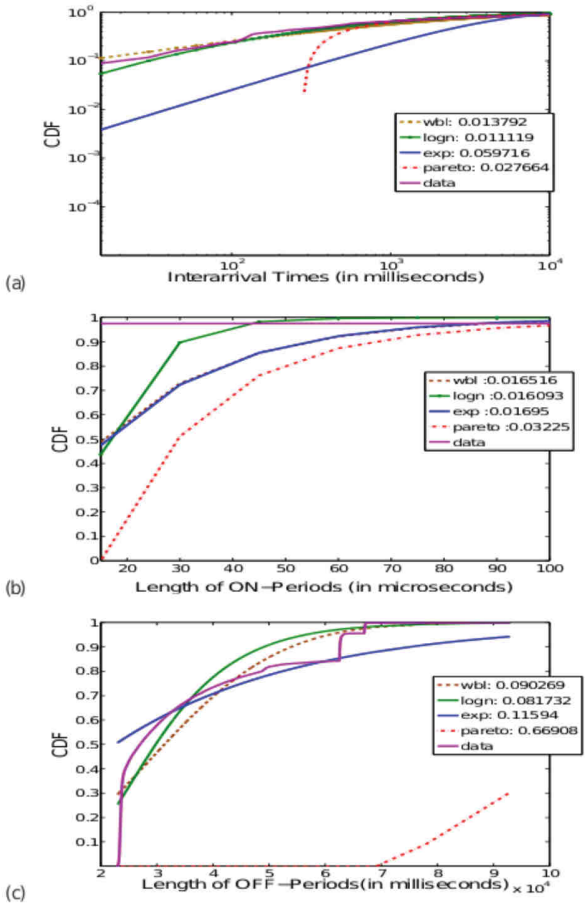


Figure 7: CDF of the distribution of the arrival times of packets at 3 of the switches in PRV2. The figure contains best fit curve for lognormal, Weibull, Pareto, and Exponential distributions, as well as the least mean errors for each.

Finally, we compare the observed distributions for HTTP applications in the data center against HTTP applications in the wide area and find that the distribution of ON periods in the data center does match observations made by others [7] in the WAN.

The take aways from our observations are that: (1) The number of active flows at a switch in any given second is, at most, 10,000 flows. However, new flows can arrive within rapid succession (10μs) of each other, resulting in high instantaneous arrival rates; (2) Most flows in the data centers we examined are small in size ($\leq$ 10K B) and a significant fraction last under a few hundreds of milliseconds; (3) Traffic leaving the edge switches in a data center is bursty in nature and the ON/OFF intervals can be characterized by heavy-tailed distributions; and (4) In some data centers, the predominant application drives the aggregate sending pattern at the edge switch. In the general case, however, simply focusing on dominant applications is insufficient to understand the process driving packet transmission into the data center network.

In the next section, we analyze link utilizations at the various layers within the data center to understand how the bursty nature of traffic impacts the utilization and packet loss of the links at each of the layers.

6. NETWORK COMMUNICATION PATTERNS

In the two previous sections, we examined the applications employed in each of the 10 data centers, their placement, and transmission patterns. In this section, we examine, with the goal of informing data center traffic engineering techniques, how existing data center applications utilize the interconnect. In particular, we aim to answer the following questions: (1) To what extent does the current application traffic utilize the data center's interconnect? For example, is most traffic confined to within a rack or not? (2) What is the utilization of links at different layers in a data center? (3) How often are links heavily utilized and what are the properties of heavily utilized links? For example, how long does heavy utilization persist on these links, and do the highly utilized links experience losses? (4) To what extent do link utilizations vary over time?

6.1 Flow of Traffic

We start by examining the relative proportion of traffic generated by the servers that stays within a rack (Intra-Rack traffic) versus traffic that leaves its rack for either other racks or external destinations (Extra-Rack traffic). Extra-Rack traffic can be directly measured, as it is the amount of traffic on the uplinks of the edge switches (i.e., the "Top-of-Rack" switches). We compute Intra-Rack traffic as the difference between the volume of traffic generated by the servers attached to each edge switch and the traffic exiting edge switches.

In Figure 8, we present a bar graph of the ratio of Extra-Rack to Intra-Rack traffic in the 10 data centers we studied. We note that a predominant portion of server-generated traffic in the cloud data centers CLD1–5—nearly, 75% on average—is confined to within the rack in which it was generated.

Recall from Section 4 that only two of these 5 data centers, CLD4 and CLD5, run MapReduce style applications, while the other three run a mixture of different customer-facing Web services. Despite this key difference in usage, we observe surprisingly little difference in the relative proportions of Intra-Rack and Extra-Rack traffic. This can be explained by revisiting the nature of applications in these data centers: as stated in Section 4, the services running in CLD1–3 have dependencies spread across many servers in the data center. The administrators of these networks try to colocate applications and dependent components into the same racks to avoid sharing a rack with other applications/services. Low Extra-Rack traffic is a side-effect of this artifact. In the case of CLD4 and CLD5, the operators assign MapReduce jobs to co-located servers for similar reasons. However, fault tolerance requires placing redundant components of the application and data storage into different racks, which increases the Extra-Rack communication. Our findings of high Intra-Rack traffic within data centers supports observations made by others [19], where the focus was on cloud data centers running MapReduce.

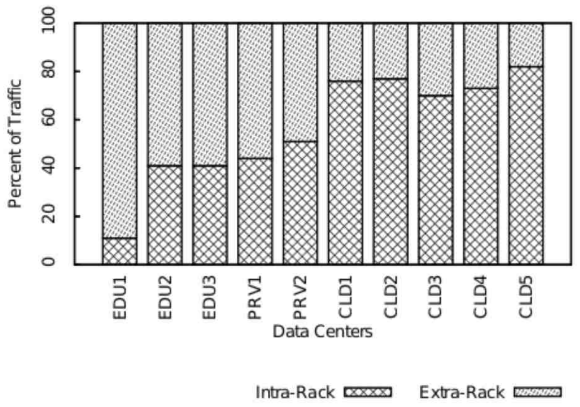


Figure 8: The ratio of Extra-Rack to Intra-Rack traffic in the data centers.

Next, we focus on the enterprise and university data centers. With the exception of EDU1, these appear to be both very different from the cloud data centers and qualitatively similar to each other: at least 50% of the server-originated traffic in the data centers leaves the racks, compared with under 25% for the cloud data centers. These data centers run user-facing applications, such as Web services and file servers. While this application mix is similar to CLD1–3 discussed above, the Intra/Extra rack usage patterns are quite different. A possible reason for the difference is that the placement of dependent services in enterprise and campus data centers may not be as optimized as the cloud data centers.

6.2 Link Utilizations vs Layer

Next, we examine the impact of the Extra-Rack traffic on the links within the interconnect of the various data centers. We examine link utilization as a function of location in the data center topology. Recall that all 10 data centers employed 2-Tiered or 3-Tiered tree-like networks.

In performing this study, we studied several hundred 5-minute intervals at random for each data center and examined the link utilizations as reported by SNMP. In Figure 9, we present the utilization for links across different layers in the data centers for one such representative interval.

In general, we find that utilizations within the core/aggregation layers are higher than those at the edge; this observation holds across all classes of data centers. These findings support observations made by others [3], where the focus was on cloud data centers.

A key point to note, not raised by prior work [3], is that across the various data centers, there are differences in the tail of the distributions for all layers–in some data centers, such as CLD4, there is a greater prevalence of high utilization links (i.e., utilization 70% or greater) especially in the core layer, while in others there are no high utilization links in any layer (e.g., EDU1). Next, we examine these high utilization links in greater depth.

6.3 Hot-spot Links

In this section, we study the hot-spot links—those with 70% or higher utilization—unearthed in various data centers, focusing on the persistence and prevalence of hot-spots. More specifically, we aim to answer the following questions: (1) Do some links frequently appear as hot-spots? How does this result vary across layers and data centers? (2) How does the set of hot-spot links in a layer change over time? (3) Do hot-spot links experience high packet loss?

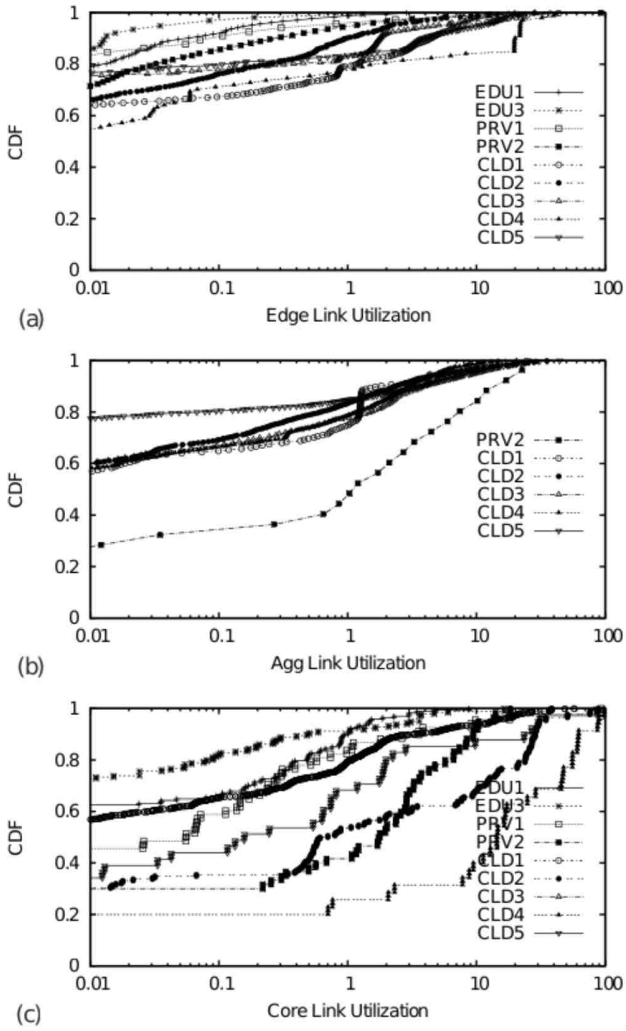


Figure 9: CDF of link utilizations (percentage) in each layer.

6.3.1 Persistence and Prevalence

In Figure 10, we present the distribution of the percentage of time intervals that a link is a hot-spot. We note from Figures 10(a) and (b) that very few links in either the edge or aggregation layers are hot-spots, and this observations holds across all data centers and data center types. Specifically, only 3% of the links in these two layers appear as a hot-spot for more than 0:1% of time intervals. When edge links are congested, they tend to be congested continuously, as in CLD2, where a very small fraction of the edge links appear as hot-spots in 90% of the time intervals.

In contrast, we find that the data centers differ significantly in their core layers (Figure 10(c)). Our data centers cluster into 3 hot-spot classes: (1) Low Persistence-Low Prevalence: This class of data centers comprises those where the hot-spots are not localized to any set of links. This includes PRV2, EDU1, EDU2, EDU3, CLD1, and CLD3, where any given core link is a hot-spot for no more than 10% of the time intervals; (2) High Persistence-Low Prevalence: The second group of data centers is characterized by hot-spots being localized to a small number of core links. This includes PRV1 and CLD2 where 3% and 8% of the core links, respectively, each appear as hot-spots in > 50% of the time intervals; and (3) High Persistence-High Prevalence: Finally, in the last group containing CLD4 and CLD5, a significant fraction of the core links

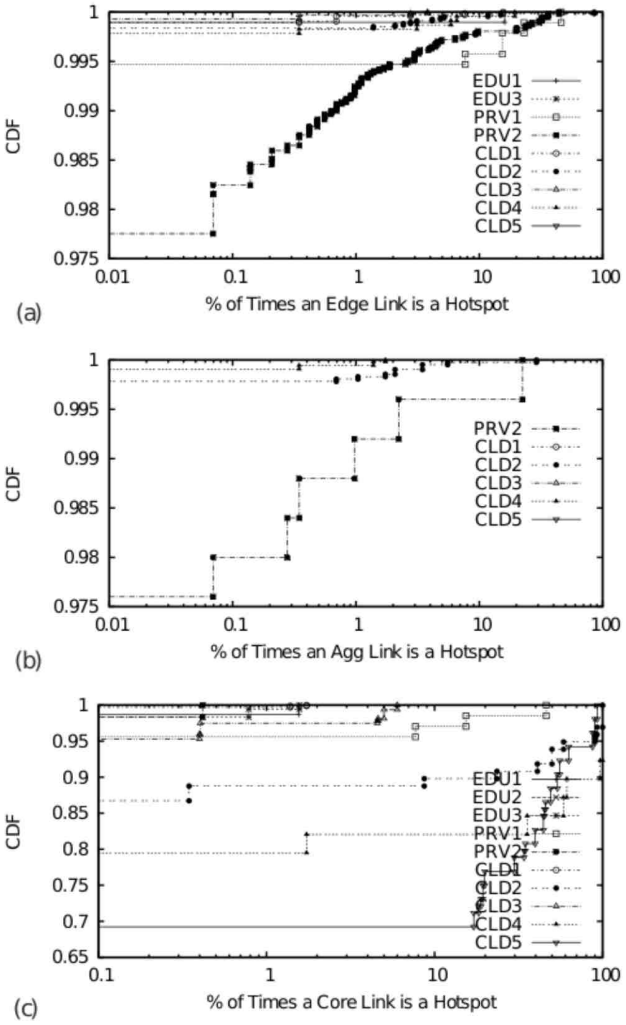


Figure 10: A CDF of the fraction of times that links in the various layers are hot-spots.

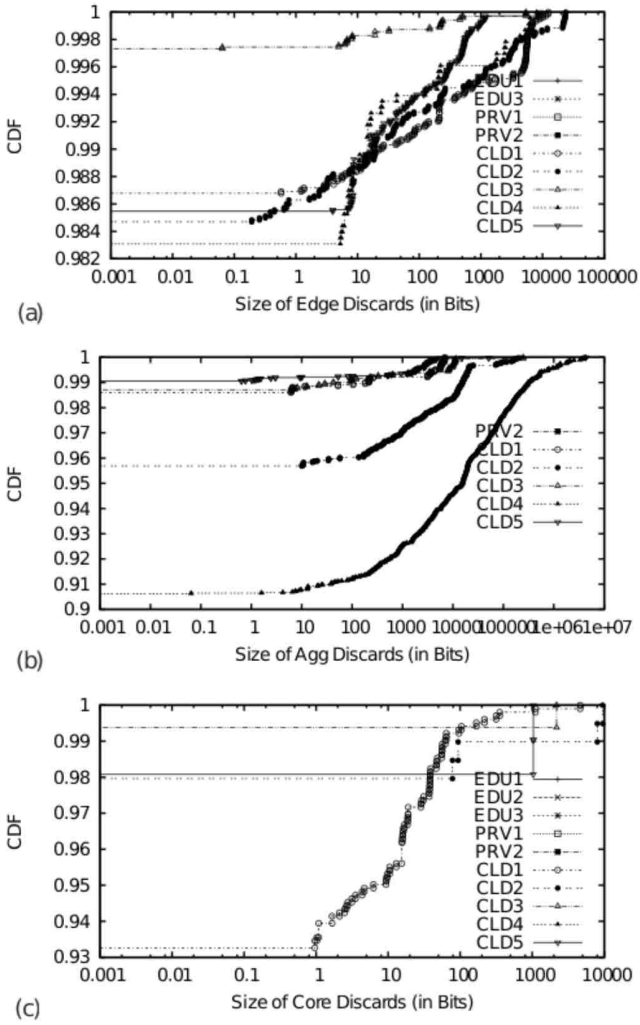


Figure 11: A CDF of the number of bits lost across the various layers.

appear persistently as hot-spots. Specifically, roughly 20% of the core links are hot-spots at least 50% of the time each. Note that both CLD4 and CLD5 run MapReduce applications.

Next, we examine the variation in the fraction of the core links that are hot-spots versus time. In Figure 13, we show our observations for one data center in each of the 3 hot-spot classes just described. From this figure, we observe that each class has a different pattern. In the low persistence-low prevalence data center, CLD1, we find that very few hot-spots occur over the course of the day, and when they do occur, only a small fraction of the core links emerge as hot-spots (less than 0.002%). However, in the high persistence classes, we observe that hot-spots occur throughout the day. Interestingly, with the high persistence-high prevalence data center, CLD5, we observe that the fraction of links that are hot-spots is affected by the time of day. Equally important is that only 25% of the core links in CLD5 are ever hot-spots. This suggests that, depending on the traffic matrix, the remaining 75% of the core links can be utilized to offload some traffic from the hot-spot links.

6.3.2 Hot-spots and Discards

Finally, we study loss rates across links in the data centers. In particular, we start by examining the discards for the set of hot-spot links. Surprisingly, we find that none of the hot-spot links experience loss. This implies that in the data centers studied, loss does not correlate with high utilization.

To understand where losses are prevalent, we examine Figures 11 and 12 that display the loss rates and link utilization for the links with losses. In the core and aggregation, all the links with losses have less than 30% average utilization, whereas at the edge, the links with losses have nearly 60% utilization. The fact that links with relatively low average utilization contain losses indicates that these links experience momentary bursts that do not persist for a long enough period to increase the average utilization. These momentary bursts can be explained by the bursty nature of the traffic (Section 5).

6.4 Variations in utilization

In this section, we examine if the utilizations vary over time and whether or not link utilizations are stable and predictable.

We examined the link utilization over a one week period and found that diurnal patterns exist in all data centers. As an example, Figure 14 presents the utilization for input and output traffic at a

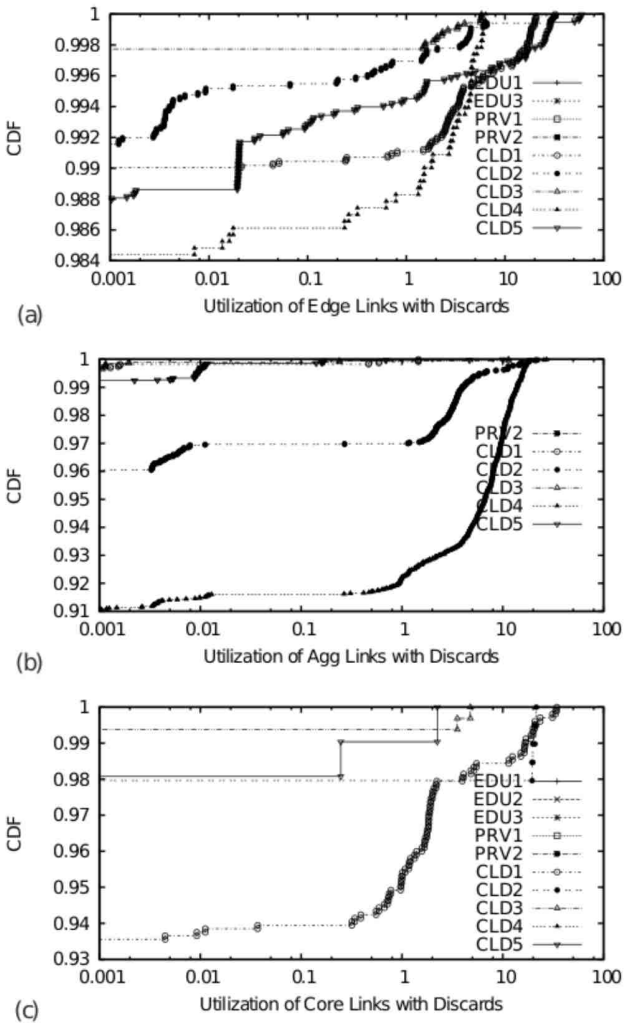


Figure 12: A CDF of the utilization of links with discards.

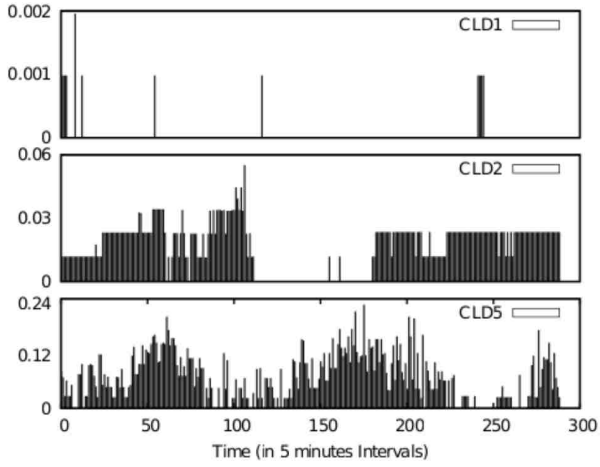


Figure 13: Time series of the fraction of links that are hot-spots in the core layer for CLD1, CLD2, and CLD5.

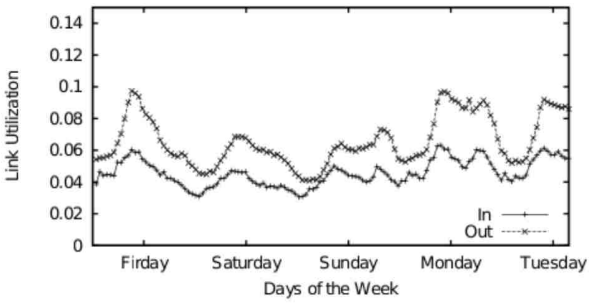


Figure 14: Time-of-Day/Day-of-Week traffic patterns.

router port in one of the cloud data centers. The 5-day trace shows diurnal and pronounced weekend/weekday variations.

To quantify this variation, we examine the difference between peak and trough utilizations for each link across the studied data centers. In Figure 15, we present the distribution of peak versus trough link utilizations across the various data centers. The x-axis is in percentage. We note that edge links in general show very little variation (less than 10% for a least 80% of edge links). The same is true for links in the aggregation layer (where available), although we see slightly greater variability. In particular, links in the aggregation layer of PRV2 show significant variability, whereas those in the other data centers do not (variation is less than 10% for a least 80% of edge links). Note that links with a low degree of variation can be run at a slower speed based on expected traffic volumes. This could result in savings in network energy costs [14].

The variation in link utilizations at the edge/aggregation are similar across the studied data centers. At the core, however, we are able to distinguish between several of the data centers. While most have low variations (less than 1%), we find that two cloud data centers (CLD4 and CLD5) have significant variations. Recall that unlike the other cloud data centers, these two cloud data centers run primarily MapReduce-style jobs. The large variations reflect differences between the periods when data is being reduced from the worker nodes to the master and other periods.

To summarize, the key take-aways from our analysis of network traffic patterns are as follows: (1) In cloud data centers, a significant fraction of traffic stays inside the rack, while the opposite is true for enterprise and campus data centers; (2) On average, the core of the data center is the most utilized layer, while the data center edge is lightly utilized; (3) The core layers in various data centers do contain hot-spot links. In some of the data centers, the hot-spots appear only occasionally. In some of the cloud data centers, a significant fraction of core links appear as hot-spots a large fraction of the time. At the same time, the number of core links that are hot-spots at any given time is less than 25%; (4) Losses are not correlated with links with persistently high utilizations. We observed losses do occur on links with low average utilization indicating that losses are due to momentary bursts; and (5) In general, time-of-day and day-of-week variation exists in many of the data centers. The variation in link utilization is most significant in the core of the data centers and quite moderate in other layers of the data centers.

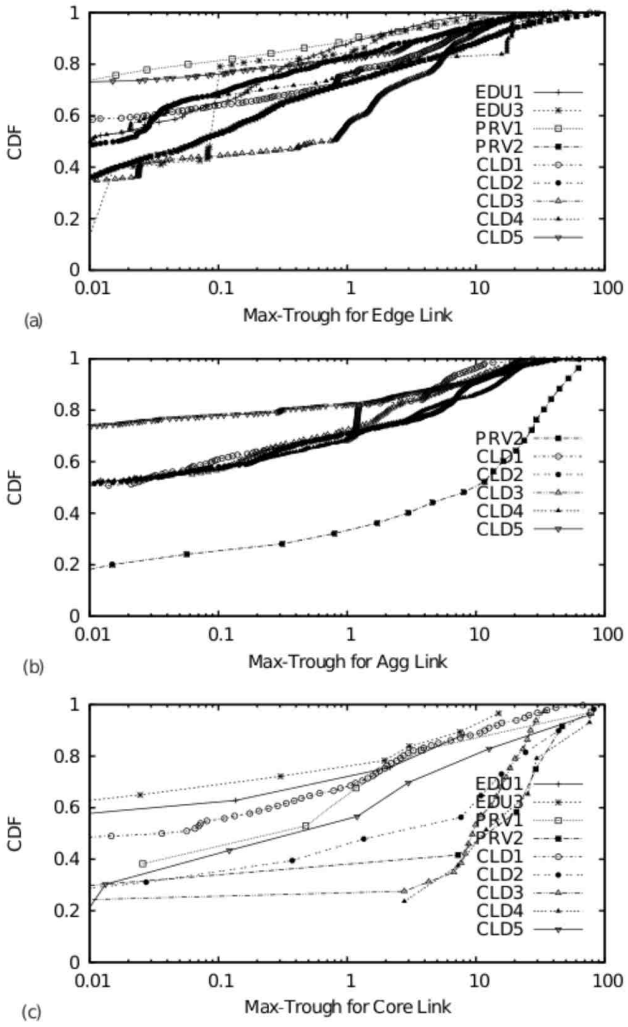


Figure 15: Difference between the peak and trough utilization.

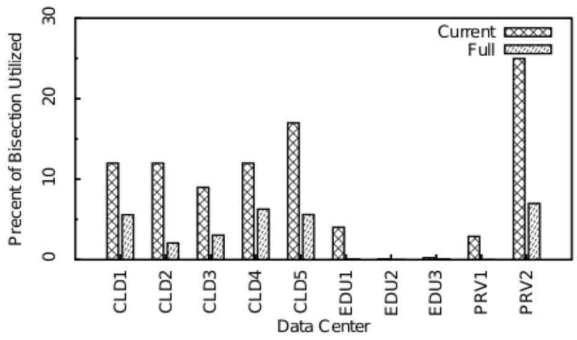


Figure 16: The first bar is the ratio of aggregate server traffic over Bisection BW and the second bar is the ratio of aggregate server traffic over full bisection capacity. The y-axis displays utilization as a percentage.

7. IMPLICATIONS FOR DATA CENTER DESIGN

7.1 Role of Bisection Bandwidth

Several proposals [1, 22, 11, 2] for new data center network architectures attempt to maximize the network bisection bandwidth. These approaches, while well suited for data centers, which run applications that stress the network's fabric with all-to-all traffic, would be unwarranted in data centers where the bisection bandwidth is not taxed by the applications. In this section, we re-evaluate the SNMP and topology data captured from the 10 data centers and examine whether the prevalent traffic patterns are likely to stress the existing bisection bandwidth. We also examine how much of the existing bisection bandwidth is needed at any given time to support the prevalent traffic patterns.

Before explaining how we address these questions, we provide a few definitions. We define the bisection links for a tiered data center to be the set of links at the top-most tier of the data center's tree architecture; in other words, the core links make up the bisection links. The bisection capacity is the aggregate capacity of these links. The full bisection capacity is the capacity that would be required to support servers communicating at full link speeds with arbitrary traffic matrices and no oversubscription. The full bisection capacity can be computed as simply the aggregate capacity of the server NICs.

Returning to the questions posed earlier in this section, we use SNMP data to compute the following: (1) the ratio of the current aggregate server-generated traffic to the current bisection capacity and (2) the ratio of the current traffic to the full bisection capacity. In doing so, we make the assumption that the bisection links can be treated as a single pool of capacity from which all offered traffic can draw. While this may not be true in all current networks, it allows us to determine whether more capacity is needed or rather better use of existing capacity is needed (for example, by improving routing, topology, or the migration of application servers inside the data center).

In Figure 16, we present these two ratios for each of the data centers studied. Recall (from Table 2) that all data centers are oversubscribed, meaning that if all servers sent data as fast as they can and all traffic left the racks, then the bisection links would be fully congested (we would expect to find utilization ratios over 100%). However, we find in Figure 16 that the prevalent traffic patterns are such that, even in the worst case where all server-generated traffic is assumed to leave the rack hosting the server, the aggregate output from servers is smaller than the network's current bisection capacity. This means even if the applications were moved around and the traffic matrix changed, the current bisection would still be more than sufficient and no more than 25% of it would be utilized across all data centers, including the MapReduce data centers. Finally, we note that the aggregate output from servers is a negligible fraction of the ideal bisection capacity in all cases. This implies that should these data centers be equipped with a network that provides full bisection bandwidth, at least 95% of this capacity would go unused and be wasted by today's traffic patterns.

Thus, the prevalent traffic patterns in the data centers can be supported by the existing bisection capacity, even if applications were placed in such a way that there was more inter-rack traffic than exists today. This analysis assumes that the aggregate capacity of the bisection links forms a shared resource pool from which all offered traffic can draw. If the topology prevents some offered traffic from reaching some links, then some links can experience high utilization while others see low utilization. Even in this situation, however, the issue is one of changing the topology and selecting a routing algorithm that allows offered traffic to draw effectively

from the existing capacity, rather than a question of adding more capacity. Centralized routing, discussed next, could help in constructing the requisite network paths.

7.2 Centralized Controllers in Data Centers

The architectures for several proposals [1, 22, 12, 2, 14, 21, 4, 18, 29] rely in some form or another on a centralized controller for configuring routes or for disseminating routing information to endhosts. A centralized controller is only practical if it is able to scale up to meet the demands of the traffic characteristics within the data centers. In this section, we examine this issue in the context of the flow properties that we analyzed in Section 5.

In particular, we focus on the proposals (Hedera [2], MicroTE [4] and ElasticTree [14]) that rely on OpenFlow and NOX [15, 23]. In an OpenFlow architecture, the first packet of a flow, when encountered at a switch, can be forwarded to a central controller that determines the route that the packet should follow in order to meet some network-wide objective. Alternatively, to eliminate the setup delay, the central controller can precompute a set of network paths that meet network-wide objectives and install them into the network at startup time.

Our empirical observations in Section 5, have important implications for such centralized approaches. First, the fact that the number of active flows is small (see Figure 4(a)) implies that switches enabled with OpenFlow can make do with a small flow table, which is a constrained resource on switches today.

Second, flow inter-arrival times have important implications for the scalability of the controller. As we observed in Section 5, a significant number of new flows (2–20%) can arrive at a given switch within 10μs of each other. The switch must forward the first packets of these flows to the controller for processing. Even if the data center has as few as a 100 edge switches, in the worst case, a controller can see 10 new flows per μs or 10 million flows per second. Depending on the complexity of the objective implemented at the controller, computing a route for each of these flows could be expensive. For example, prior work [5] showed a commodity machine computing a simple shortest path for only 50K flow arrivals per second. Thus, to scale the throughput of a centralized control framework while supporting complex routing objectives, we must employ parallelism (i.e., use multiple CPUs per controller and multiple controllers) and/or use faster but less optimal heuristics to compute routes. Prior work [28] has shown, through parallelism, the ability of a central controller to scale to 20 million flows per second.

Finally, the flow duration and size also have implications for the centralized controller. The lengths of flows determine the relative impact of the latency imposed by a controller on a new flow. Recall that we found that most flows last less than 100ms. Prior work [5] showed than it takes reactive controllers, which make decisions at flow start up time, approximately 10ms to install flow entries for new flows. Given our results, this imposes a 10% delay overhead on most flows. Additional processing delay may be acceptable for some traffic, but might be unacceptable for other kinds. For the class of workloads that find such a delay unacceptable, OpenFlow provides a proactive mechanism that allows the controllers, at switch start up time, to install flow entries in the switches. This proactive mechanism eliminates the 10ms delay but limits the controller to proactive algorithms.

In summary, it appears the number and inter-arrival time of data center flows can be handled by a sufficiently parallelized implementation of the centralized controller. However, the overhead of reactively computing flow placements is a reasonable fraction of the length of the typical flow.

8. SUMMARY

In this paper, we conducted an empirical study of the network traffic of 10 data centers spanning three very different categories, namely university campus, private enterprise data centers, and cloud data centers running Web services, customer-facing applications, and intensive Map-Reduce jobs. To the best of our knowledge, this is the broadest-ever large-scale measurement study of data centers.

We started our study by examining the applications run within the various data centers. We found that a variety of applications are deployed and that they are placed non-uniformly across racks. Next, we studied the transmission properties of the applications in terms of the flow and packet arrival processes at the edge switches. We discovered that the arrival process at the edge switches is ON/OFF in nature where the ON/OFF durations can be characterized by heavy-tailed distributions. In analyzing the flows that constitute these arrival process, we observed that flows within the data centers studied are generally small in size and several of these flows last only a few milliseconds.

We studied the implications of the deployed data center applications and their transmission properties on the data center network and its links. We found that most of the server generated traffic in the cloud data centers stays within a rack, while the opposite is true for campus data centers. We found that at the edge and aggregation layers, link utilizations are fairly low and show little variation. In contrast, link utilizations at the core are high with significant variations over the course of a day. In some data centers, a small but significant fraction of core links appear to be persistently congested, but there is enough spare capacity in the core to alleviate congestion. We observed losses on the links that are lightly utilized on average and argued that these losses can be attributed to the bursty nature of the underlying applications run within the data centers.

On the whole, our empirical observations can help inform data center traffic engineering and QoS approaches, as well as recent techniques for managing other resources, such as data center network energy consumption. To further highlight the implications of our study, we re-examined recent data center proposals and architectures in light of our results. In particular, we determined that full bisection bandwidth is not essential for supporting current applications. We also highlighted practical issues in successfully employing centralized routing mechanisms in data centers.

Our empirical study is by no means all-encompassing. We recognize that there may be other data centers in the wild that may or may not share all the properties that we have observed. Our work points out that it is worth closely examining the different design and usage patterns, as there are important differences and commonalities.

9. ACKNOWLEDGMENTS

We would like to thank the operators at the various universities, online services providers, and private enterprises for both the time and data that they provided us. We would also like to thank the anonymous reviewers for their insightful feedback.

This work is supported in part by an NSF FIND grant (CNS-0626889), an NSF CAREER Award (CNS-0746531), an NSF NetSE grant (CNS-0905134), and by grants from the University of Wisconsin-Madison Graduate School. Theophilus Benson is supported by an IBM PhD Fellowship.

10. REFERENCES

[1] M. Al-Fares, A. Loukissas, and A. Vahdat. A scalable, commodity data center network architecture. In SIGCOMM, pages 63–74, 2008.

[2] M. Al-Fares, S. Radhakrishnan, B. Raghavan, W. College, N. Huang, and A. Vahdat. Hedera: Dynamic flow scheduling for data center networks. In Proceedings of NSDI 2010, San Jose, CA, USA, April 2010.

[3] T. Benson, A. Anand, A. Akella, and M. Zhang. Understanding Data Center Traffic Characteristics. In Proceedings of Sigcomm Workshop: Research on Enterprise Networks, 2009.

[4] T. Benson, A. Anand, A. Akella, and M. Zhang. The case for fine-grained traffic engineering in data centers. In Proceedings of INM/WREN '10, San Jose, CA, USA, April 2010.

[5] M. Casado, M. J. Freedman, J. Pettit, J. Luo, N. McKeown, and S. Shenker. Ethane: taking control of the enterprise. In SIGCOMM, 2007.

[6] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. volume 51, pages 107–113, New York, NY, USA, 2008. ACM.

[7] A. B. Downey. Evidence for long-tailed distributions in the internet. In In Proceedings of ACM SIGCOMM Internet Measurment Workshop, pages 229–241. ACM Press, 2001.

[8] M. Fomenkov, K. Keys, D. Moore, and K. Claffy. Longitudinal study of Internet traffic in 1998-2003. In WISICT '04: Proceedings of the Winter International Symposium on Information and Communication Technologies, pages 1–6. Trinity College Dublin, 2004.

[9] H. J. Fowler, W. E. Leland, and B. Bellcore. Local area network traffic characteristics, with implications for broadband network congestion management. IEEE Journal on Selected Areas in Communications, 9:1139–1149, 1991.

[10] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot. Packet-level traffic measurements from the Sprint IP backbone. IEEE Network, 17:6–16, 2003.

[11] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. VL2: a scalable and flexible data center network. In SIGCOMM, 2009.

[12] A. Greenberg, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. Towards a next generation data center architecture: scalability and commoditization. In PRESTO '08: Proceedings of the ACM workshop on Programmable routers for extensible services of tomorrow, pages 57–62, New York, NY, USA, 2008. ACM.

[13] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu. BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers. In Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication, Barcelona, Spain, August 17 - 21 2009.

[14] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown. Elastictree: Saving energy in data center networks. April 2010.

[15] NOX: An OpenFlow Controller. http://noxrepo.org/wp/.

[16] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu. Dcell: a scalable and fault-tolerant network structure for data centers. In SIGCOMM '08: Proceedings of the ACM SIGCOMM 2008 conference on Data communication, pages 75–86, New York, NY, USA, 2008. ACM.

[17] W. John and S. Tafvelin. Analysis of Internet backbone traffic and header anomalies observed. In IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pages 111–116, New York, NY, USA, 2007. ACM.

[18] S. Kandula, J. Padhye, and P. Bahl. Flyways to de-congest data center networks. In Proc. ACM Hotnets-VIII, New York City, NY. USA., Oct. 2009.

[19] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken. The Nature of Data Center Traffic: Measurements and Analysis. In IMC, 2009.

[20] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic. In SIGCOMM '93: Conference proceedings on Communications architectures, protocols and applications, pages 183–193, New York, NY, USA, 1993. ACM.

[21] J. Mudigonda, P. Yalagandula, M. Al-Fares, and J. C. Mogul. Spain: Cots data-center ethernet for multipathing over arbitrary topologies. In Proceedings of NSDI 2010, San Jose, CA, USA, April 2010.

[22] 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, 2009.

[23] The OpenFlow Switch Consortium. http://www.openflowswitch.org/.

[24] V. Paxson. Empirically-Derived Analytic Models of Wide-Area TCP Connections. 2(4):316–336, Aug. 1994.

[25] V. Paxson. Measurements and analysis of end-to-end internet dynamics. Technical report, 1997.

[26] V. Paxson. Bro: a system for detecting network intruders in real-time. In SSYM'98: Proceedings of the 7th conference on USENIX Security Symposium, pages 3–3, Berkeley, CA, USA, 1998. USENIX Association.

[27] V. Paxson and S. Floyd. Wide area traffic: the failure of poisson modeling. IEEE/ACM Trans. Netw., 3(3):226–244, 1995.

[28] A. Tavakoli, M. Casado, T. Koponen, and S. Shenker. Applying nox to the datacenter. In Proc. of workshop on Hot Topics in Networks (HotNets-VIII), 2009.

[29] G. Wang, D. G. Andersen, M. Kaminsky, M. Kozuch, T. S. E. Ng, K. Papagiannaki, M. Glick, and L. Mummert. Your data center is a router: The case for reconfigurable optical circuit switched paths. In Proc. ACM Hotnets-VIII, New York City, NY. USA., Oct. 2009.