Network Traffic Characteristics of Data Centers in the Wild

Proceedings of the 10th annual conference on Internet measurement, ACM

Outline

- Introduction
- Traffic Data Collection
- Applications in Data Centers
- Application Communication Patterns
- Network Communication Patterns
- Implications for Data Center Design
- Conclusions

Introduction

Introduction (1/3)

- This paper conduct an empirical study of the network traffic in 10 data centers, including :
  - 3 University
  - 2 Private Enterprise
  - 5 Cloud

Introduction (2/3)

- Large **universities** and **private enterprises** are increasingly consolidating their IT services within data centers.

Introduction (3/3)

- Large online service providers, such as Google are rapidly building **cloud data centers**.
  - Each data centers often containing more than 10K servers
  - Offer a variety of cloud-based services, such as Email, search...

Traffic Data Collection

Traffic Data Collection

- **SNMP polls**
  - We poll the switches' SNMP MIBs for bytes-in and bytes out at granularities ranging from 1 minute to 30 minutes.
- **Network Topology**
  - We obtained topology via the Cisco CDP protocol, which gives both the network topology as well as the link capacities.
- **Packet traces**
  - packet trace collection spans 12 hours over multiple days.

Data Center Role	Data Center Name	Location	Age (Years) (Curr Ver/Total)	SNMP
Universities	EDU1	US-Mid	10	✓
	EDU2	US-Mid	(7/20)	✓
	EDU3	US-Mid	N/A	✓
Private	PRV1	US-Mid	(5/5)	✓
	PRV2	US-West	> 5	✓
Commercial	CLD1	US-West	> 5	✓
	CLD2	US-West	> 5	✓
	CLD3	US-East	> 5	✓
	CLD4	S. America	(3/3)	✓
	CLD5	S. America	(3/3)	✓

Packet Traces	Topology	Number Devices	Number Servers	Over Subscription
✓	✓	22	500	2:1
✓	✓	36	1093	47:1
✓	✓	1	147	147:1
X	✓	96	1088	8:3
✓	✓	100	2000	48:10
X	X	562	10K	20:1
X	X	763	15K	20:1
X	X	612	12K	20:1
X	X	427	10K	20:1
X	X	427	10K	20:1

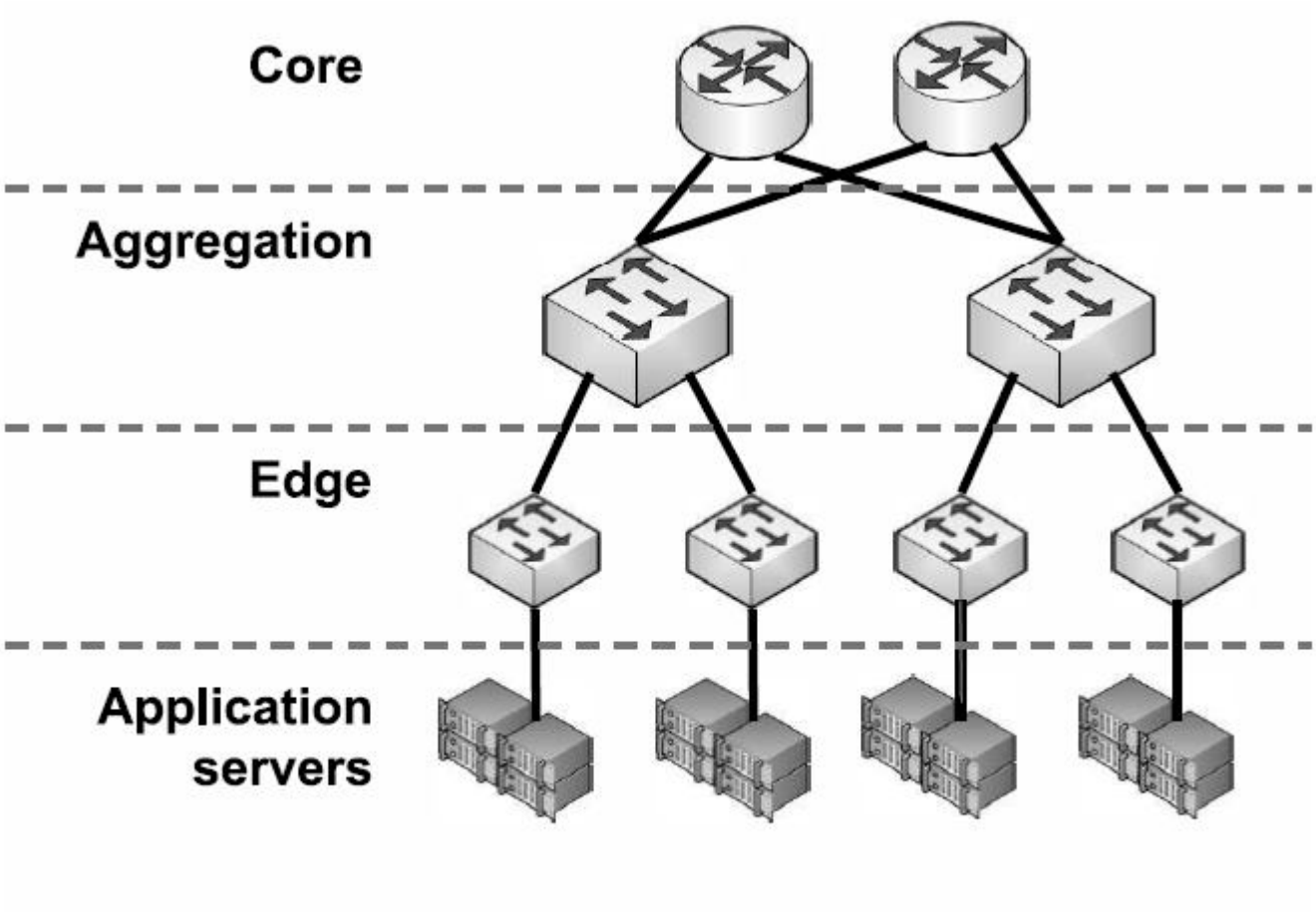


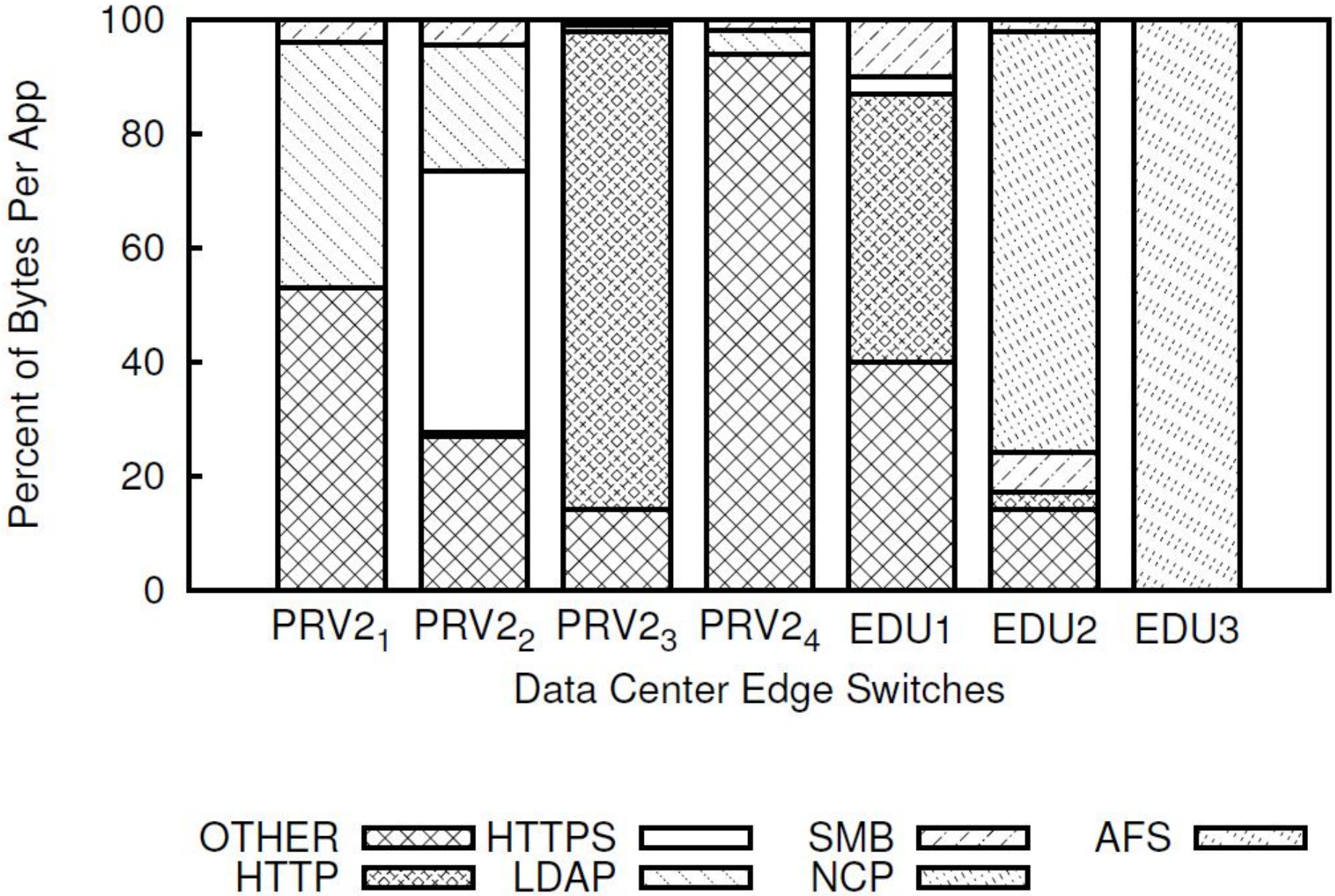
Figure 1: Canonical 3-Tier data center topology.

Applications in Data Centers

Questions

(1) What type of applications are running within these data centers?

(2) What fraction of traffic originated by a switch is contributed by each application?


Percent of Bytes Per App
100
80
60
40
20
0
PRV2$_1$
PRV2$_2$
PRV2$_3$
PRV2$_4$
EDU1
EDU2
EDU3
Data Center Edge Switches
OTHER
HTTP
HTTPS
LDAP
SMB
NCP
AFS

Answers

(1)There is a wide variety of applications observed both within and across data centers.

(2)We observe a wide variation in the composition of traffic originated by the switches in a given data center .

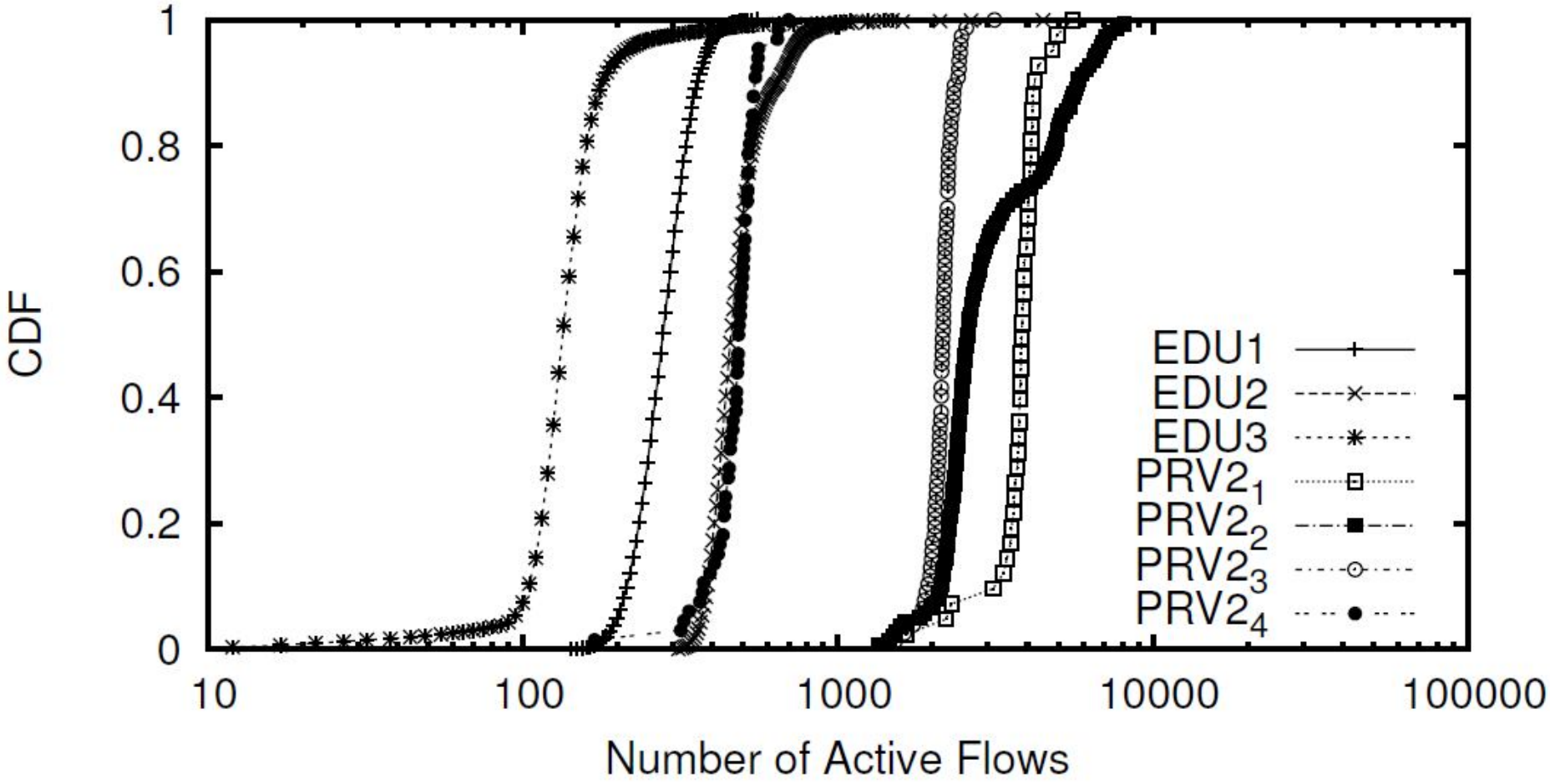
Application Communication Patterns

Flow-Level

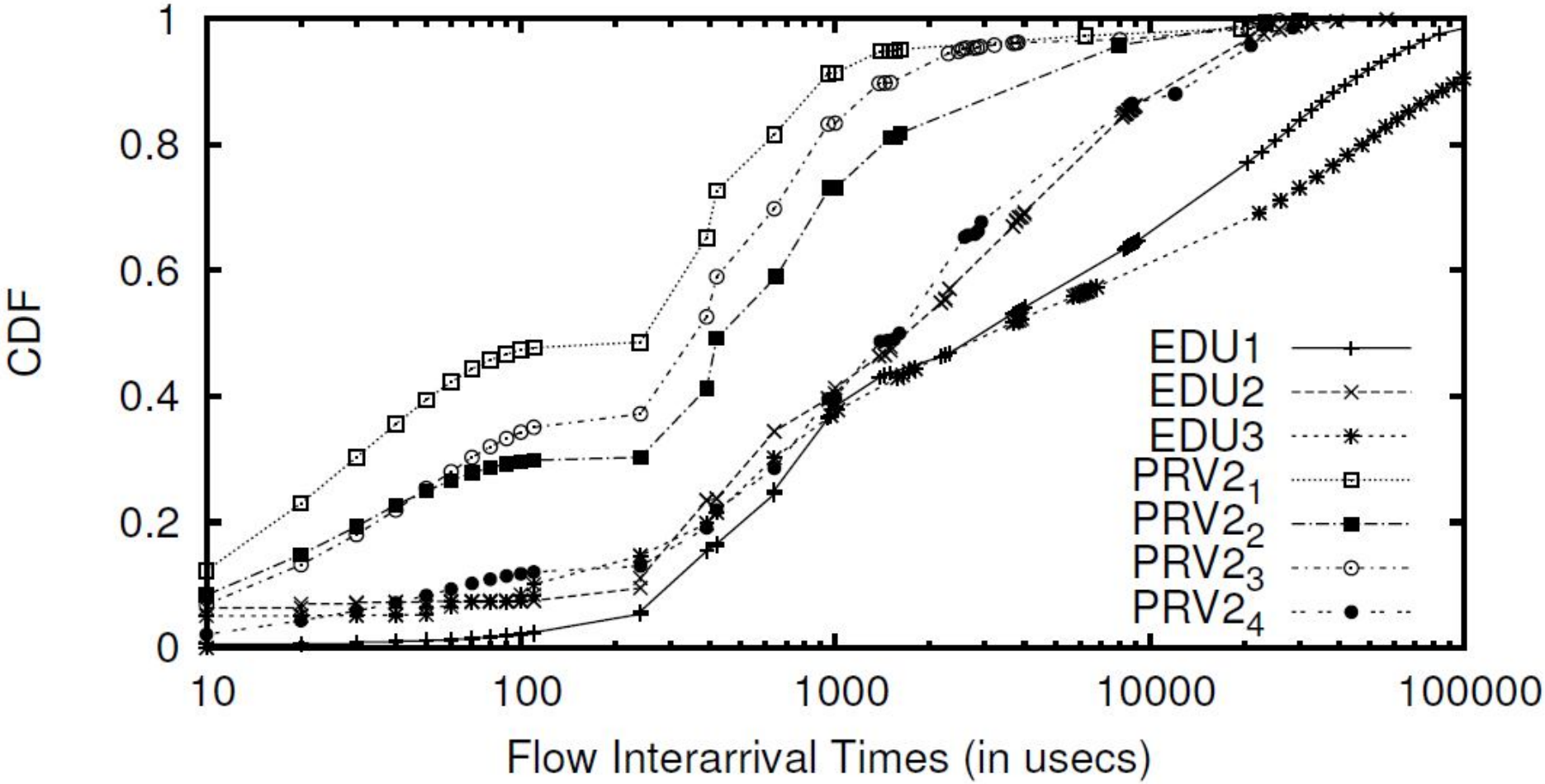
(1) What are the aggregate characteristics of flow arrivals, sizes, and durations?

(2) What are the aggregate characteristics of the packet-level inter-arrival process across all applications in a rack?

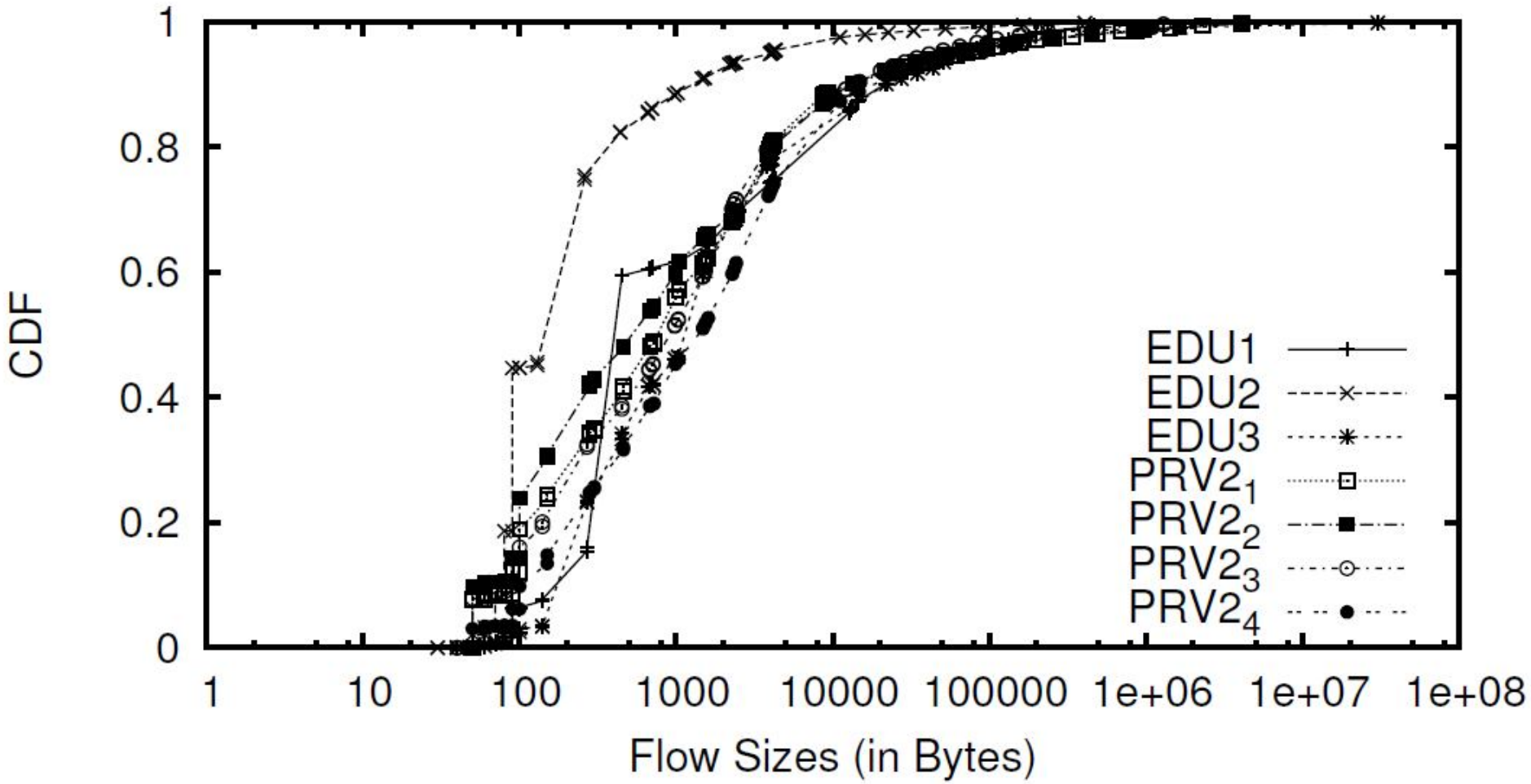
The number of active flows per second



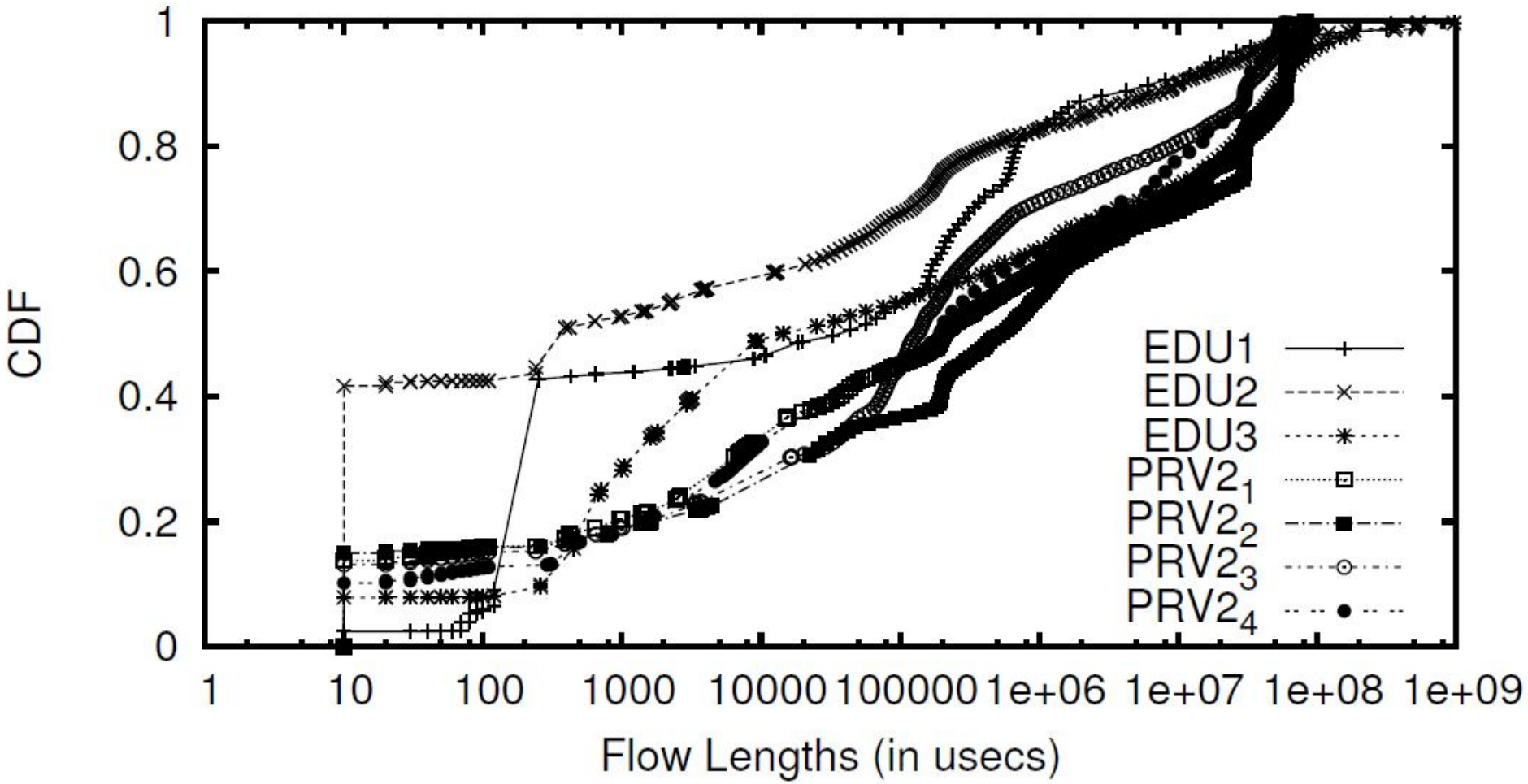
Flow inter-arrival times



Flow sizes


CDF
1
0.8
0.6
0.4
0.2
0
1
10
100
1000
10000
100000
1e+06
1e+07
1e+08
Flow Sizes (in Bytes)
EDU1
EDU2
EDU3
PRV2_1
PRV2_2
PRV2_3
PRV2_4

Flow durations



Packet sizes

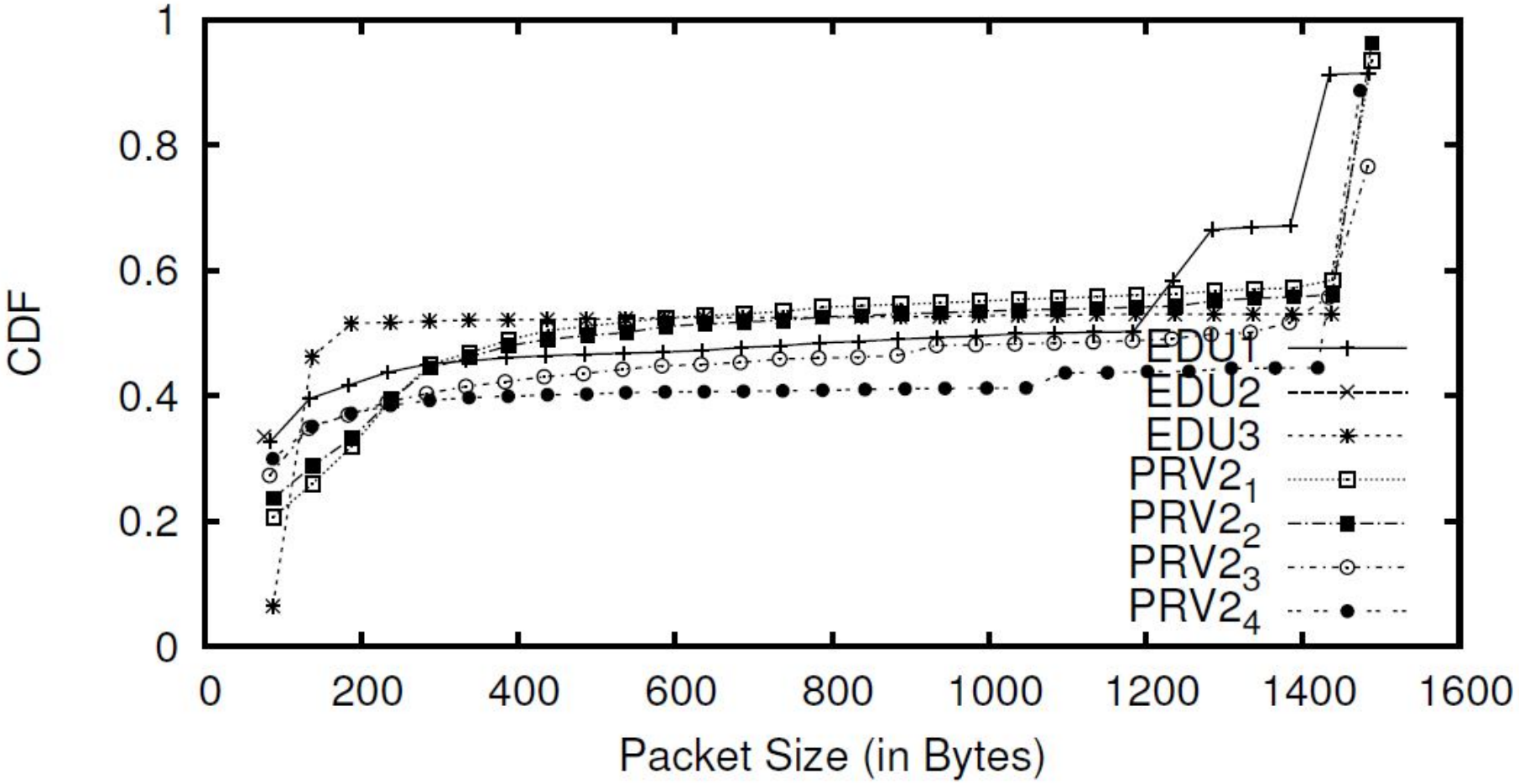


Figure 5: Distribution of packet size in the various networks.

Packet-Level

(1) the durations of the ON periods.

(2) the durations of the OFF periods.

(3) the packet inter-arrival times within ON periods.

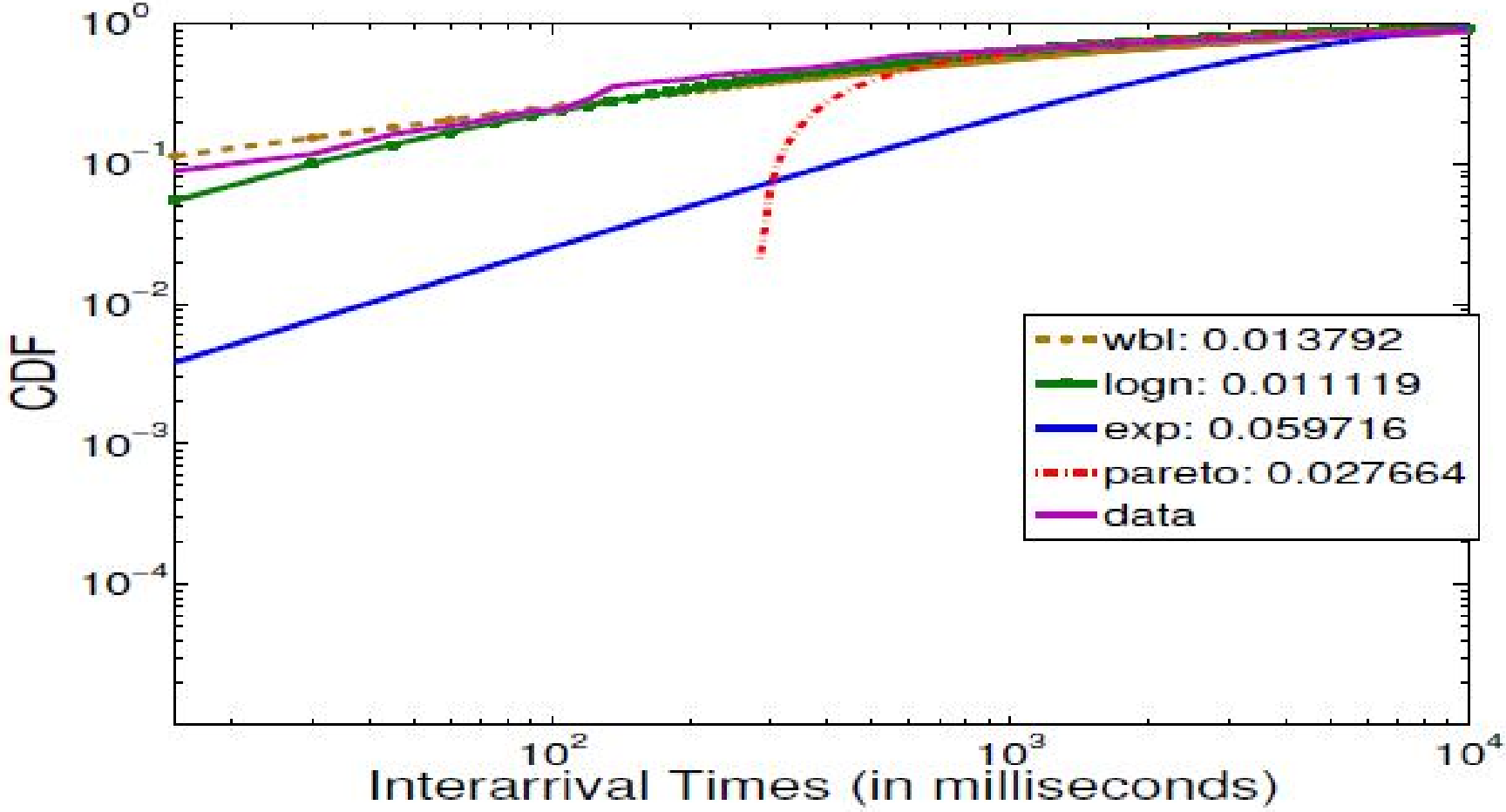
Data center	Off period Distribution	ON period Distribution	Interarrival Rate Distribution
$PRV2_1$	Lognormal	Lognormal	Lognormal
$PRV2_2$	Lognormal	Lognormal	Lognormal
$PRV2_3$	Lognormal	Lognormal	Lognormal
$PRV2_4$	Lognormal	Lognormal	Lognormal
EDU1	Lognormal	Weibull	Weibull
EDU2	Lognormal	Weibull	Weibull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

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

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

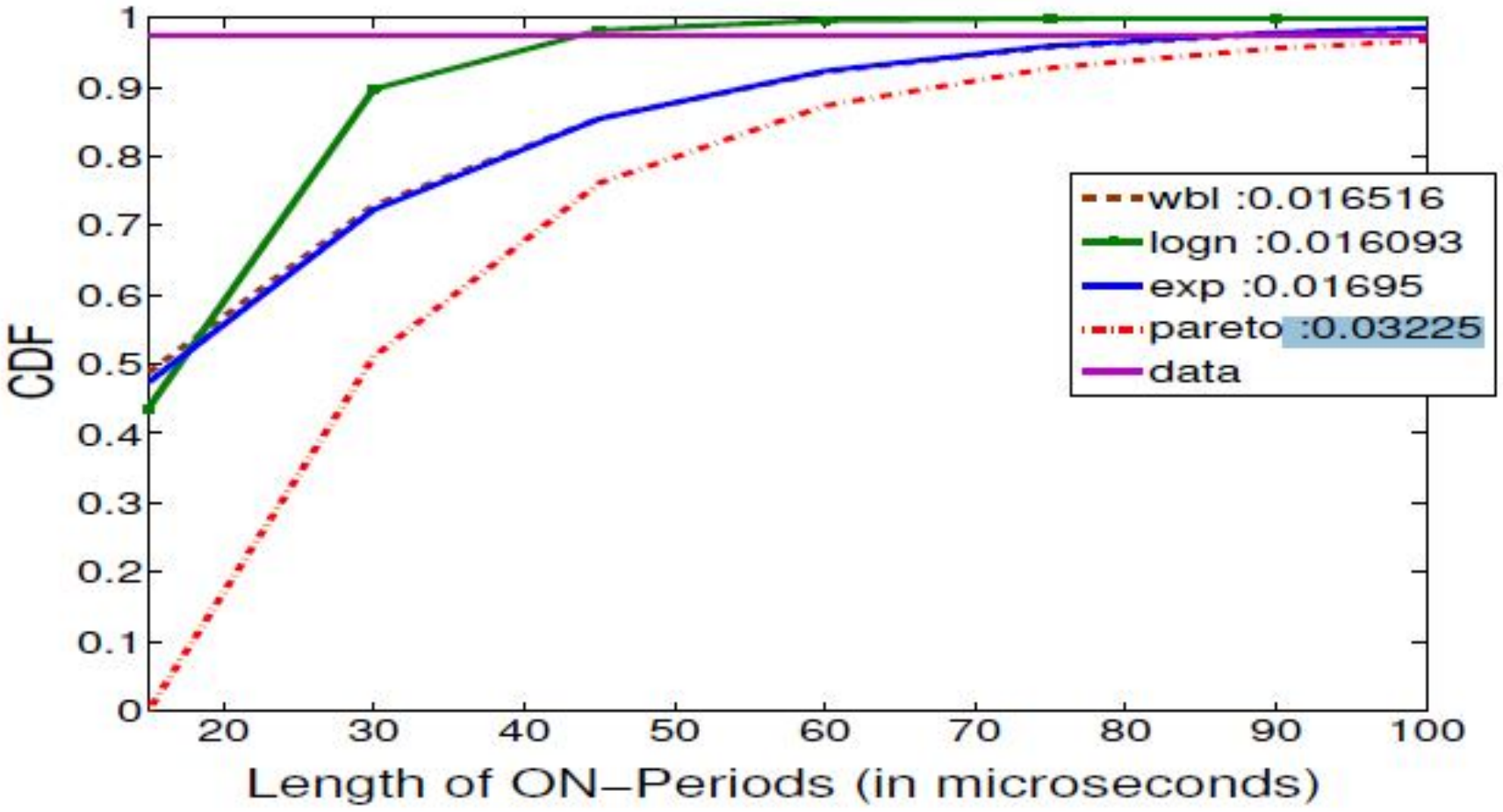
Table 5: The distribution for the parameters of each of the arrival processes of the dominant applications on each switch.

Interarrival Times



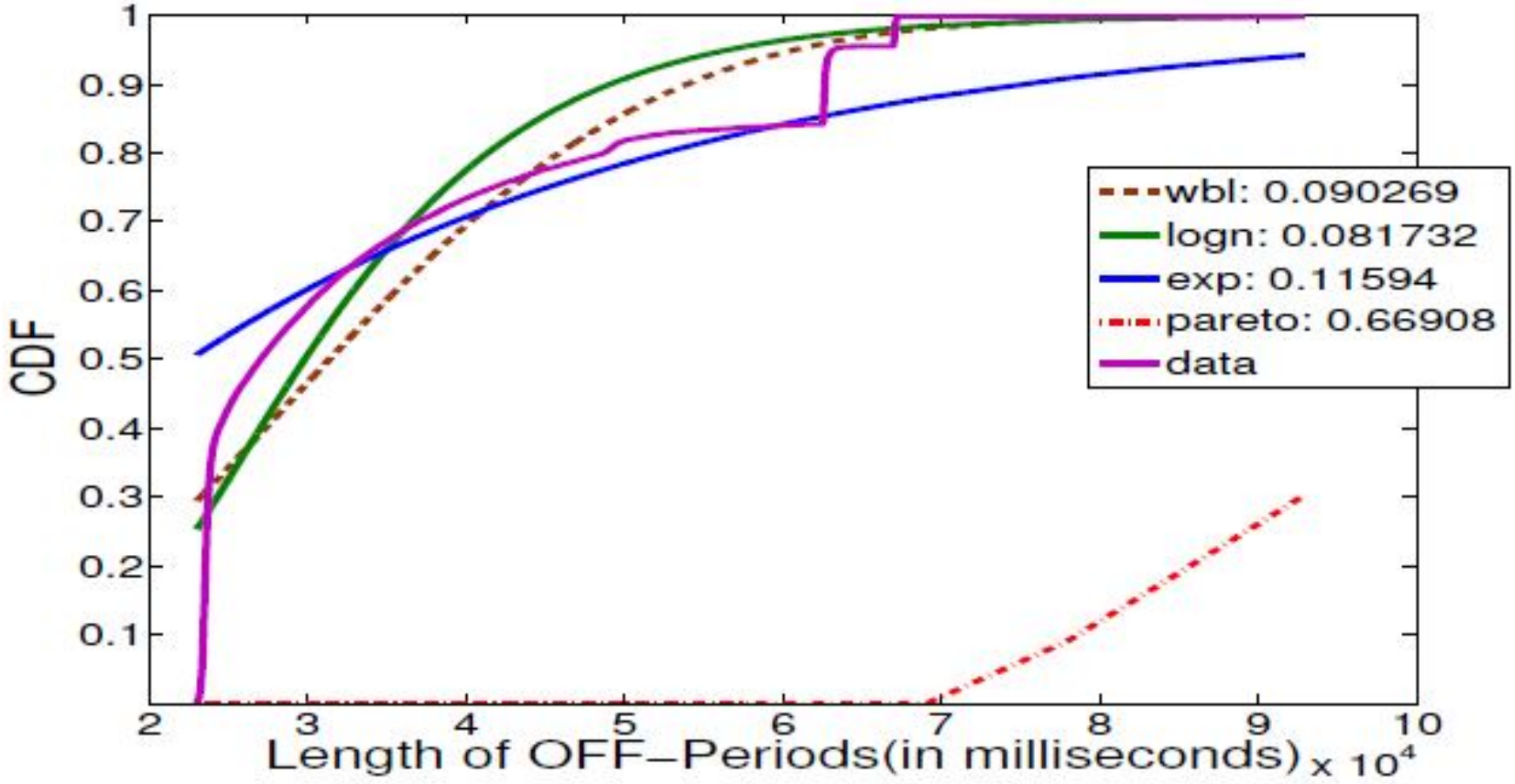
(a)

Length of ON−Periods



(b)

Length of OFF−Periods



(c)

Network Communication Patterns

Questions(1/2)

(1) 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)To what extent do link utilizations vary overtime?

Questions(2/2)

(4)How often are links heavily utilized and what are the properties of heavily utilized links?

- how long does heavy utilization persist on these links?
- do the highly utilized links experience losses?

The ratio of Extra-Rack to Intra-Rack traffic

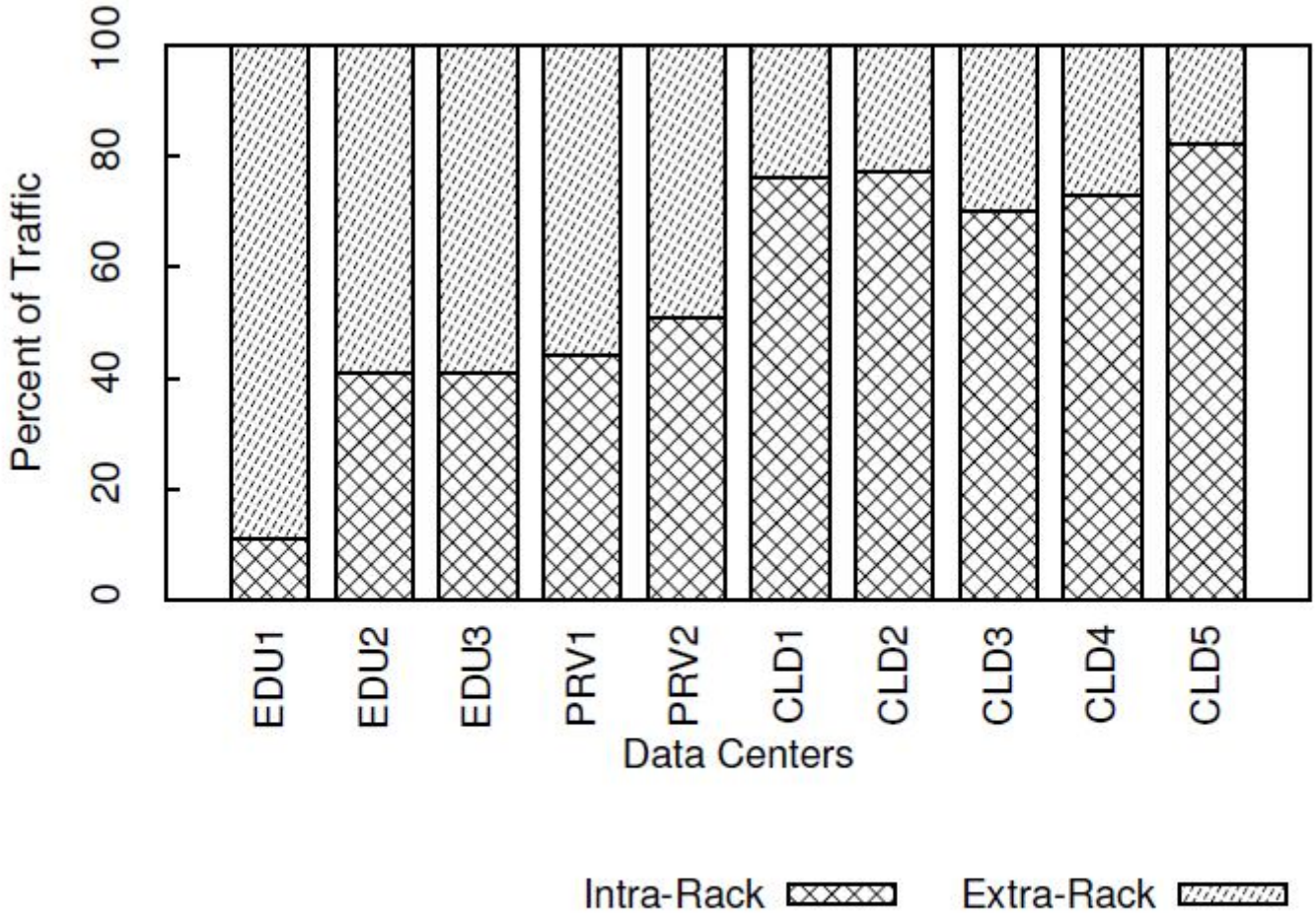
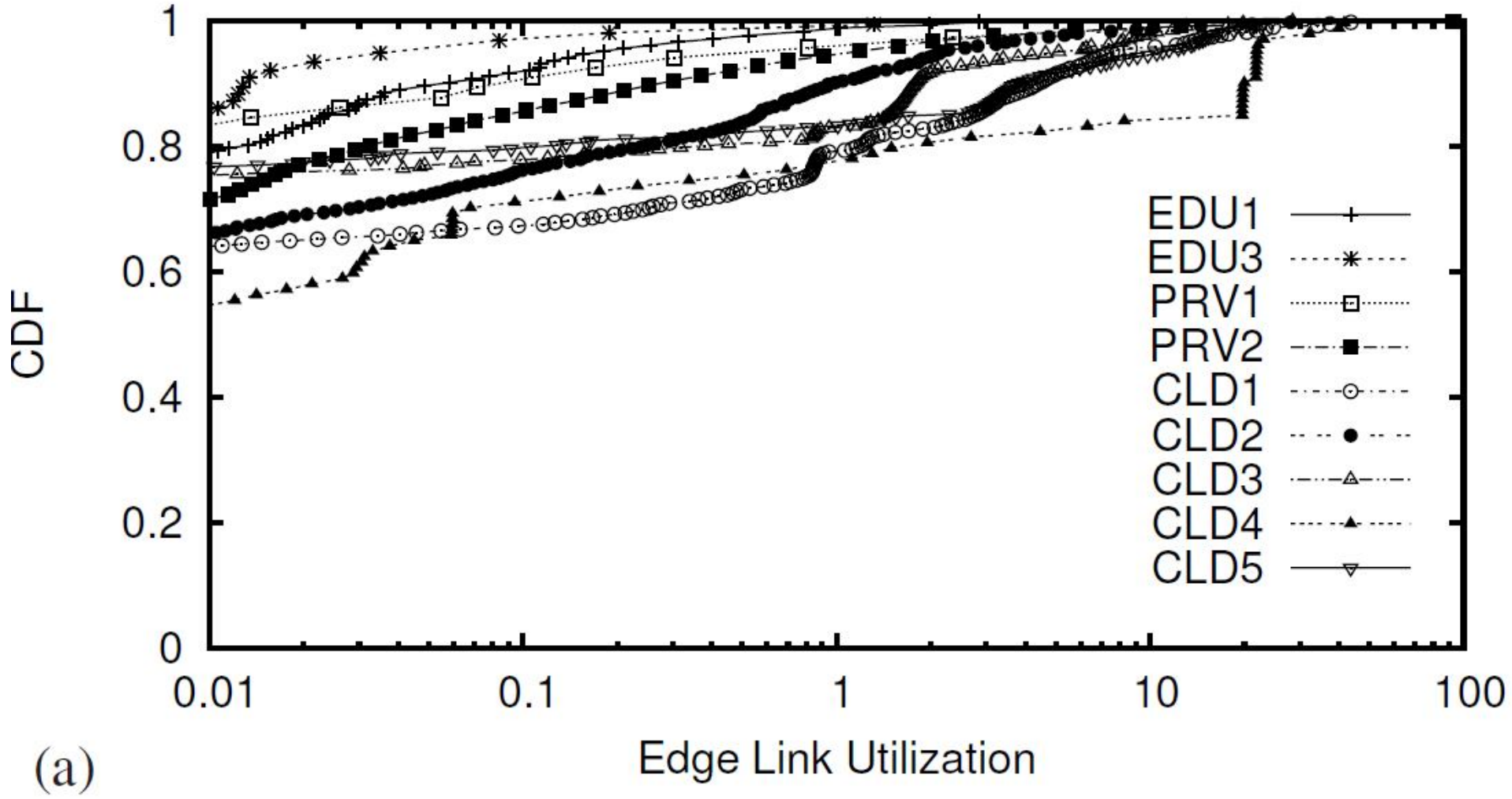


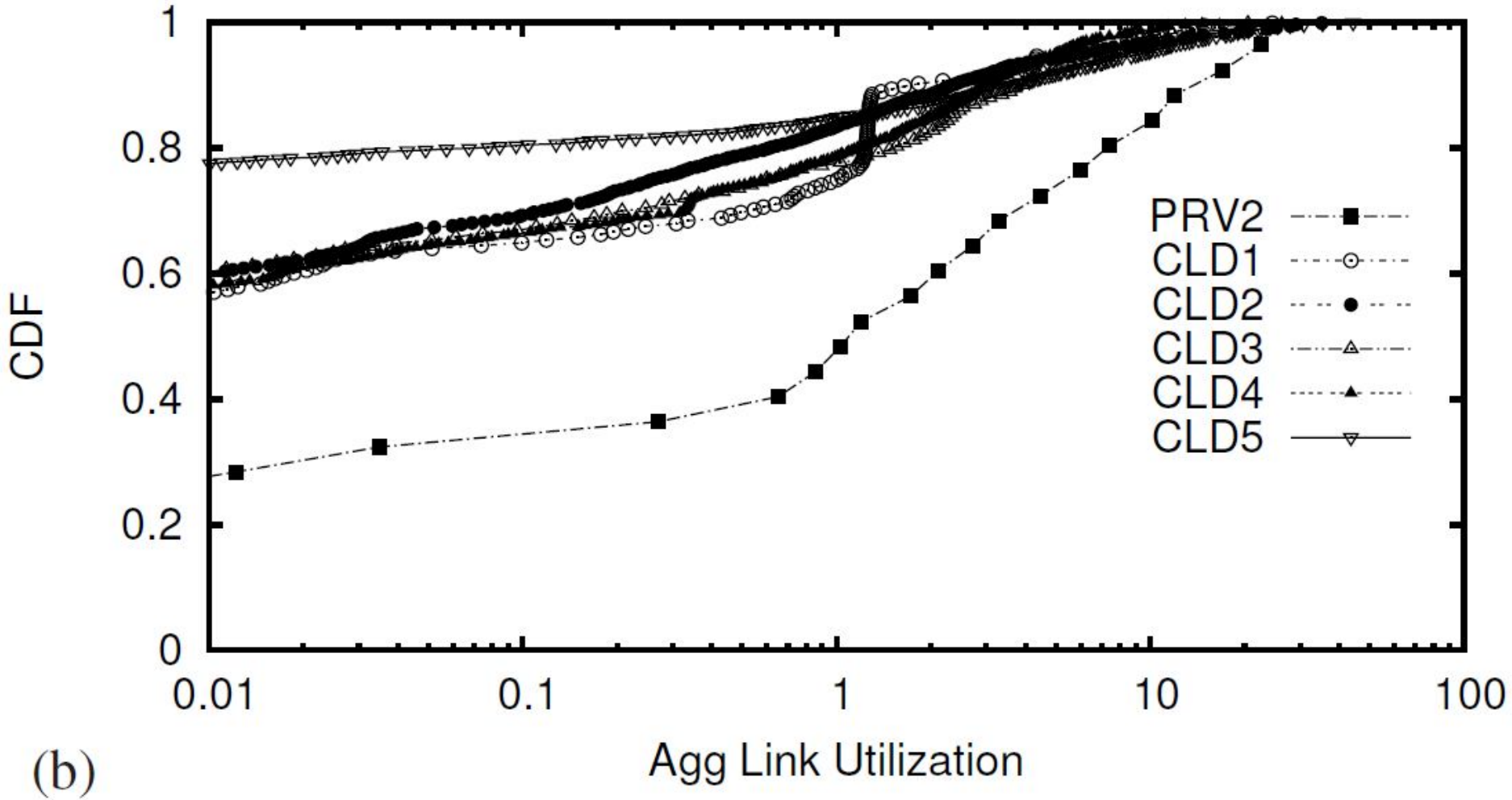
Figure 8: The ratio of Extra-Rack to Intra-Rack traffic in the data centers.

Link utilizations in Edge layer



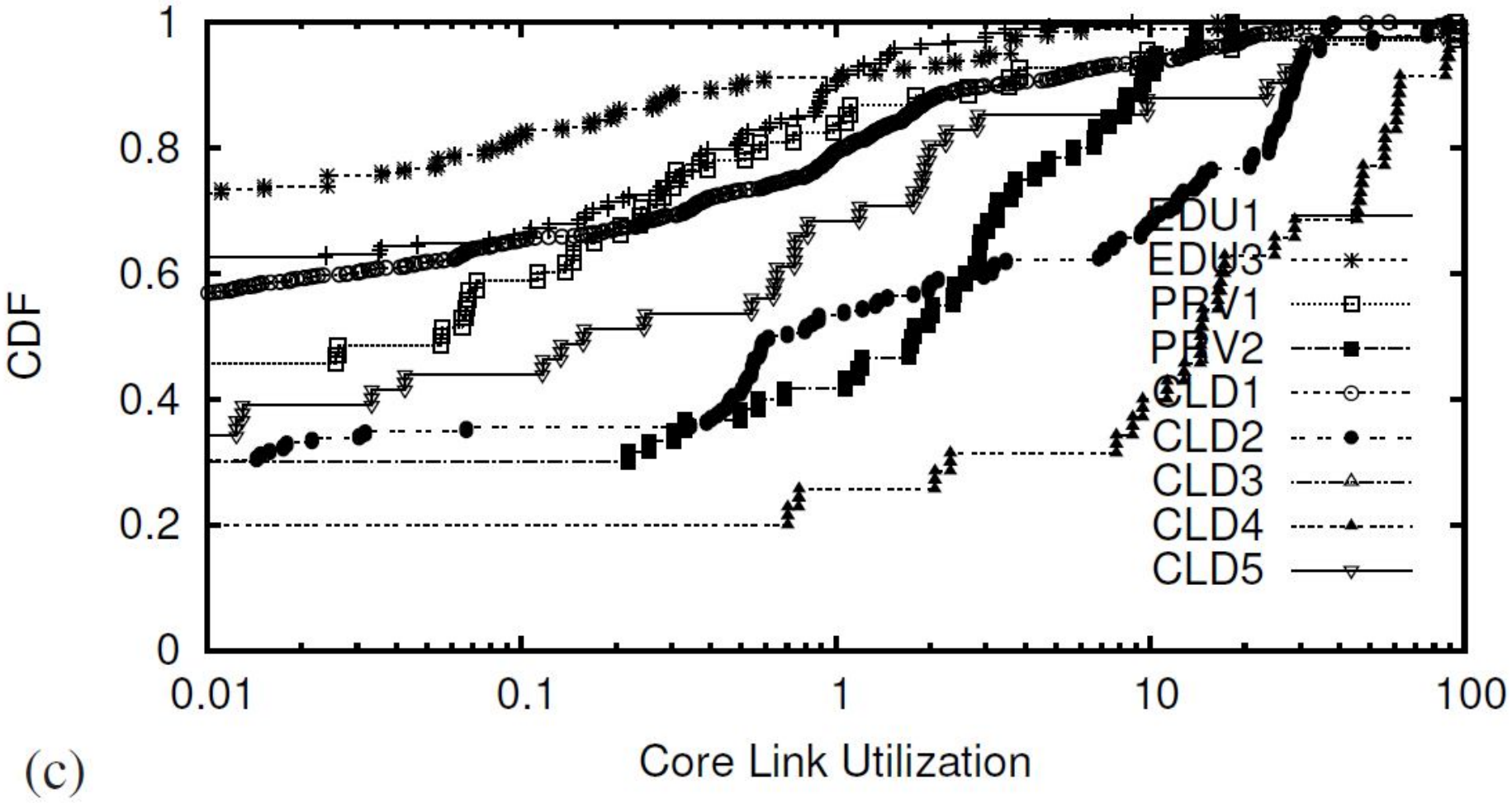
(a)

Link utilizations in Aggregation layer



(b)

Link utilizations in Core layer



(c)

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?

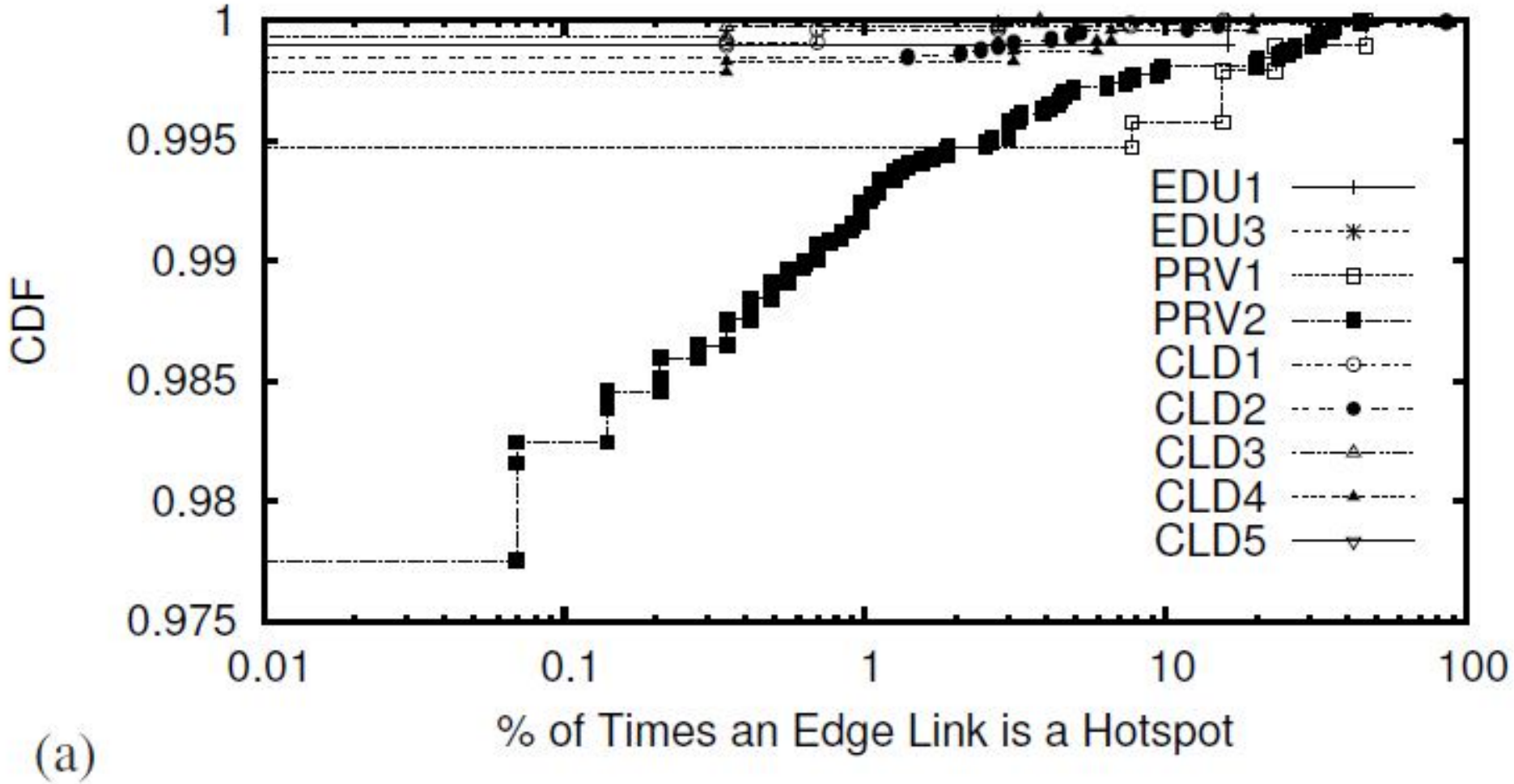
Hot-spots Core link

(1) Low Persistence-Low Prevalence

(2) High Persistence-Low Prevalence

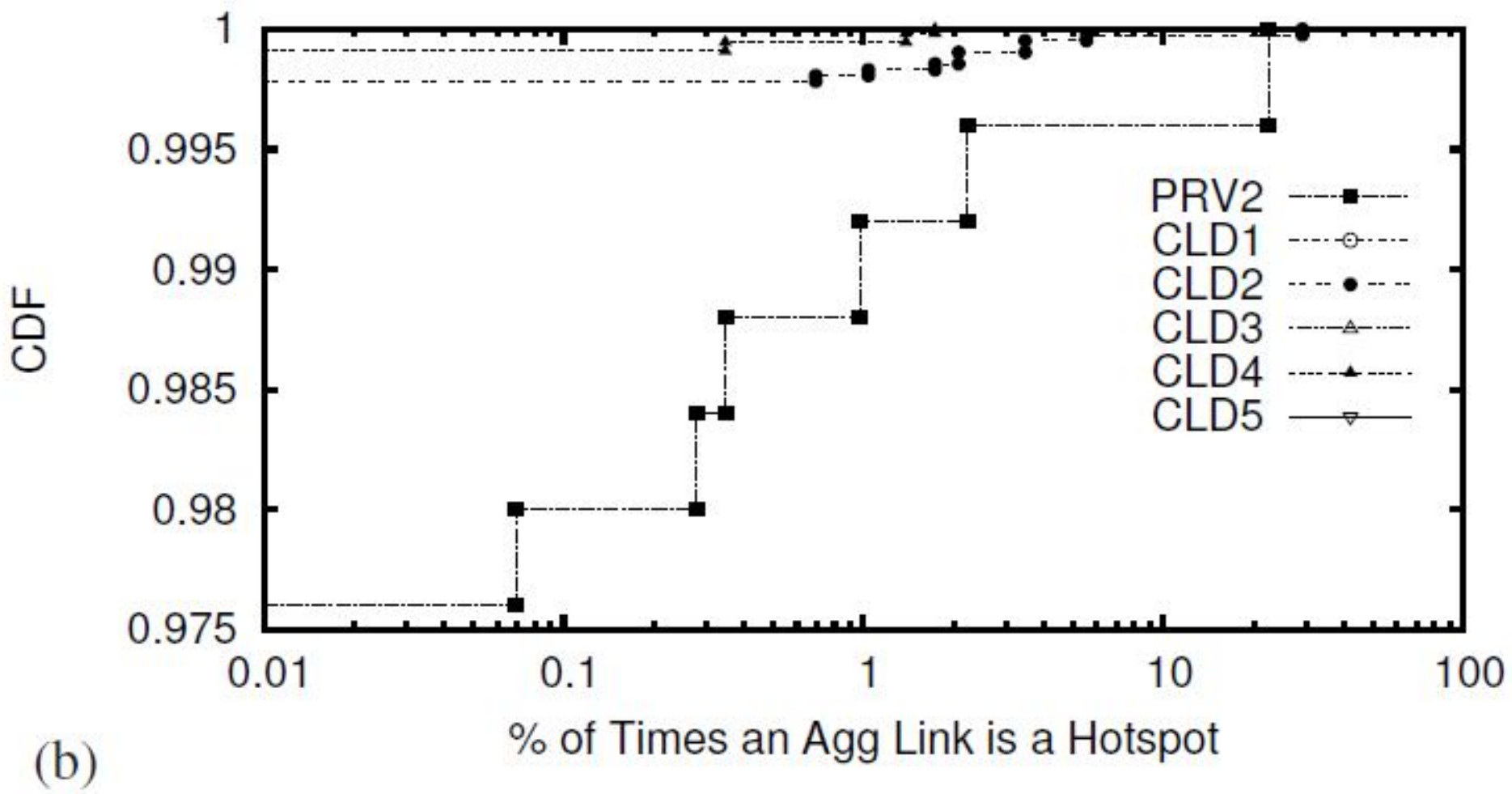
(3) High Persistence-High Prevalence

% of Times a Edge Link is a Hotspot



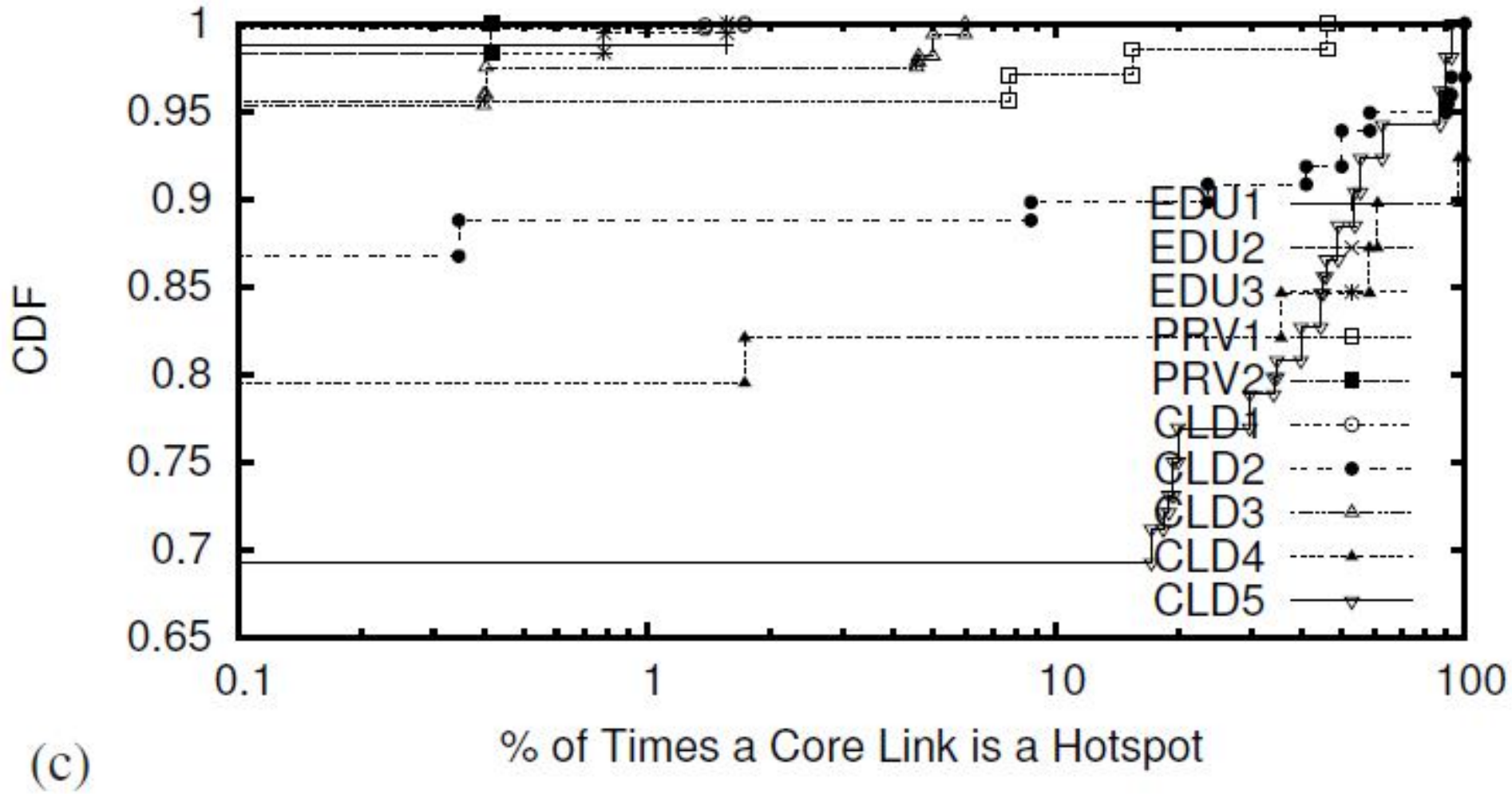
(a)

% of Times a Aggregation Link is a Hotspot



(b)

% of Times a Core Link is a Hotspot



(c)

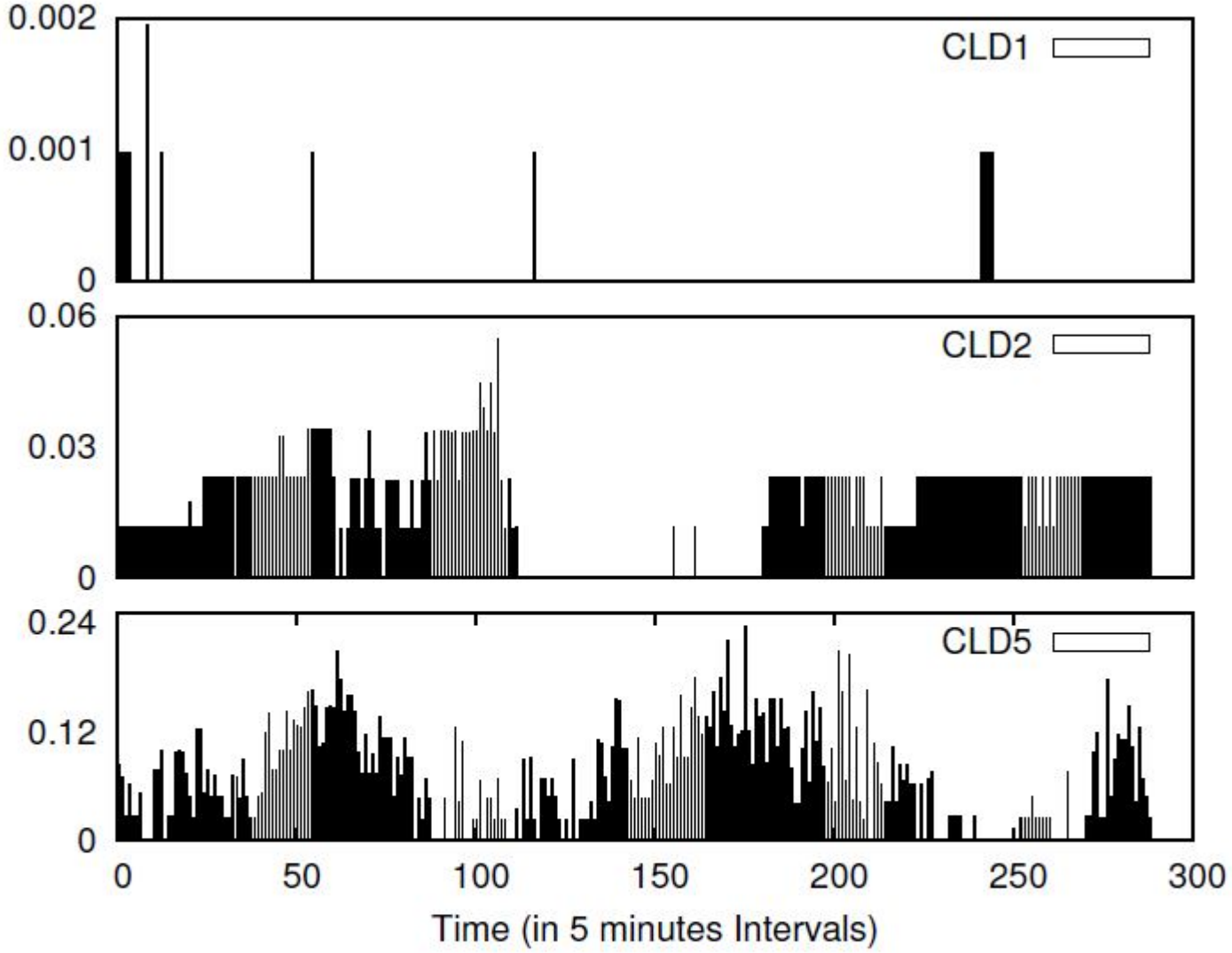
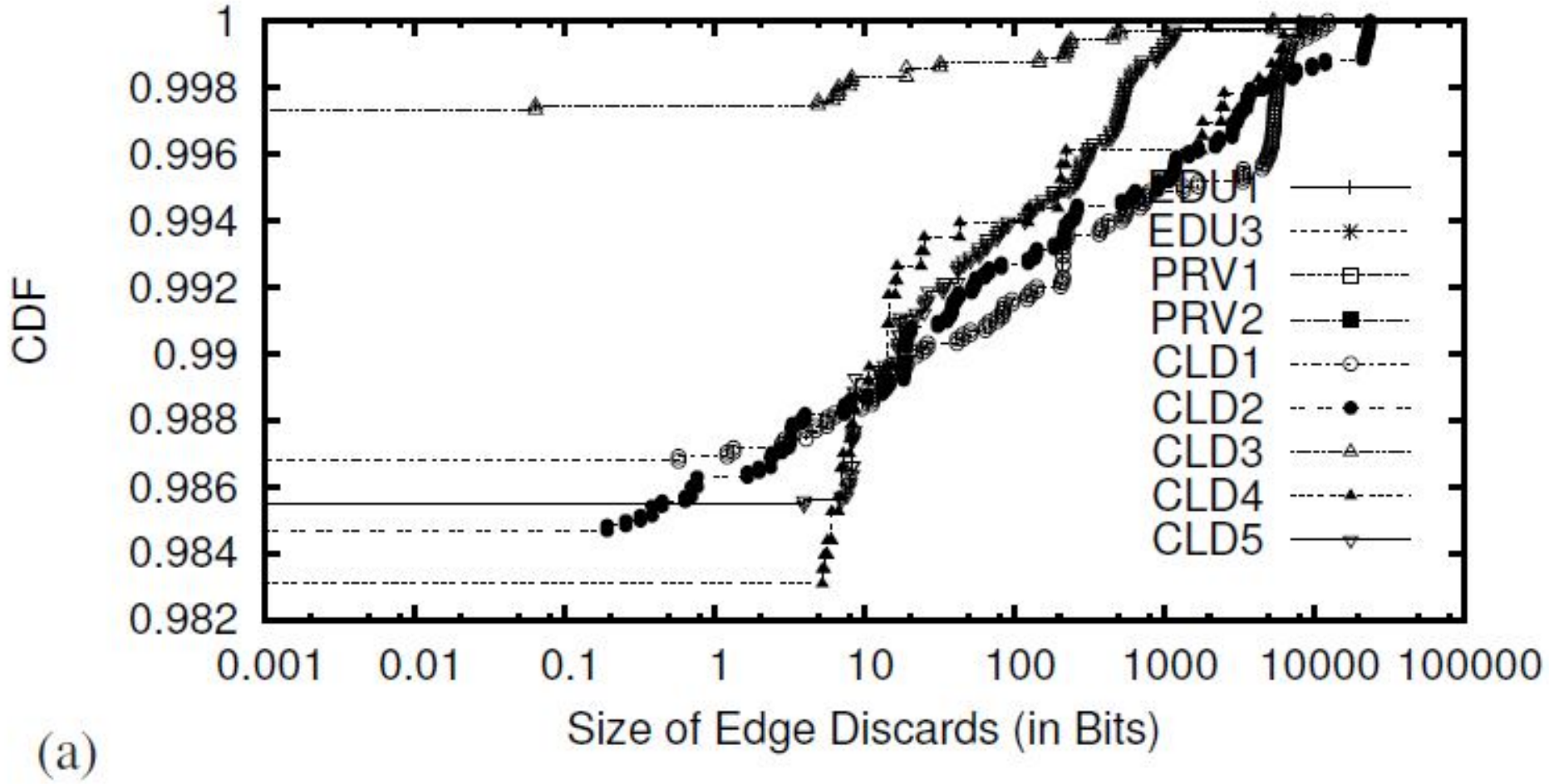


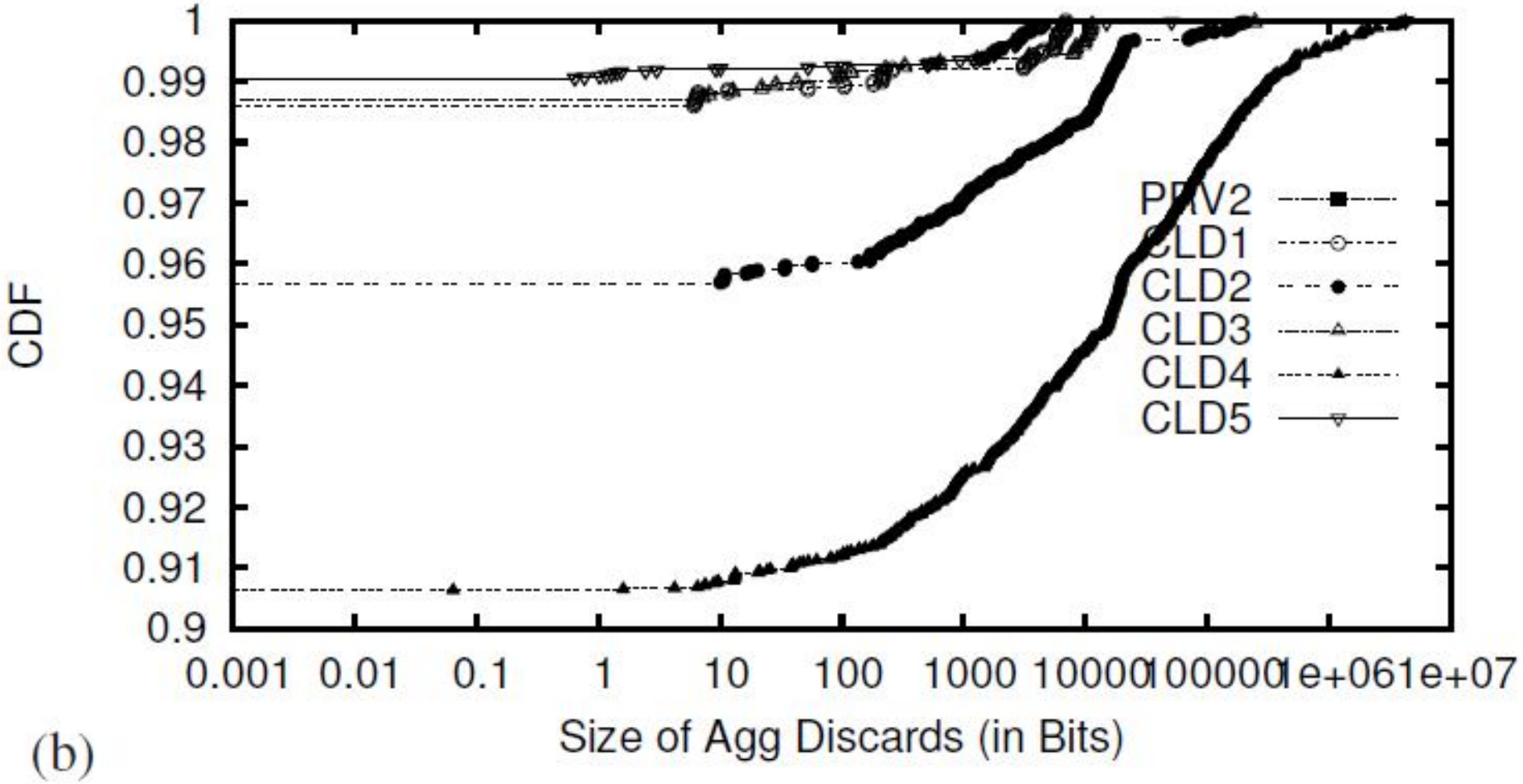
Figure 13: Time series of the fraction of links that are hot-spots in the core layer for CLD1, CLD2, and CLD5.

Size of Edge Discards



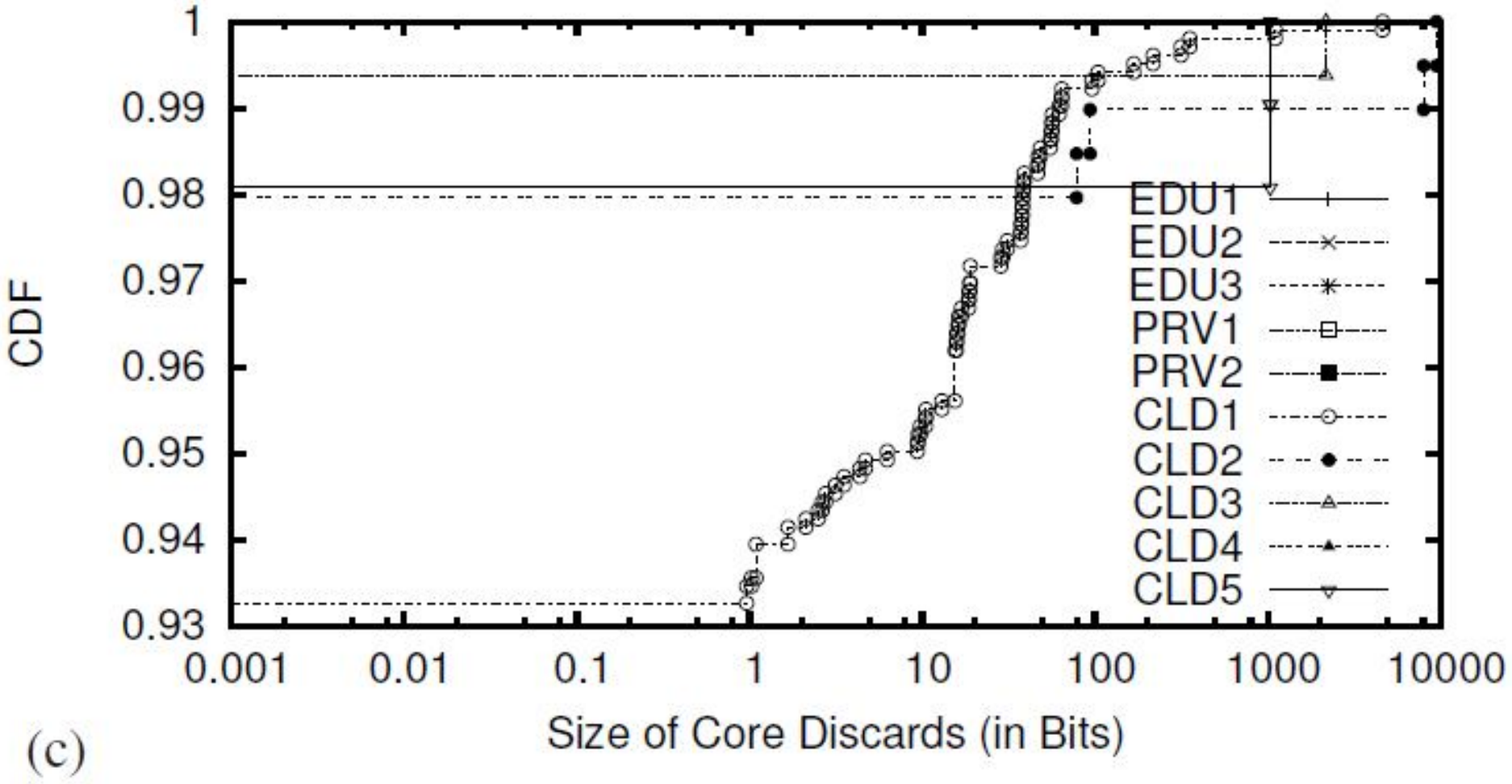
(a)

Size of Agg Discards



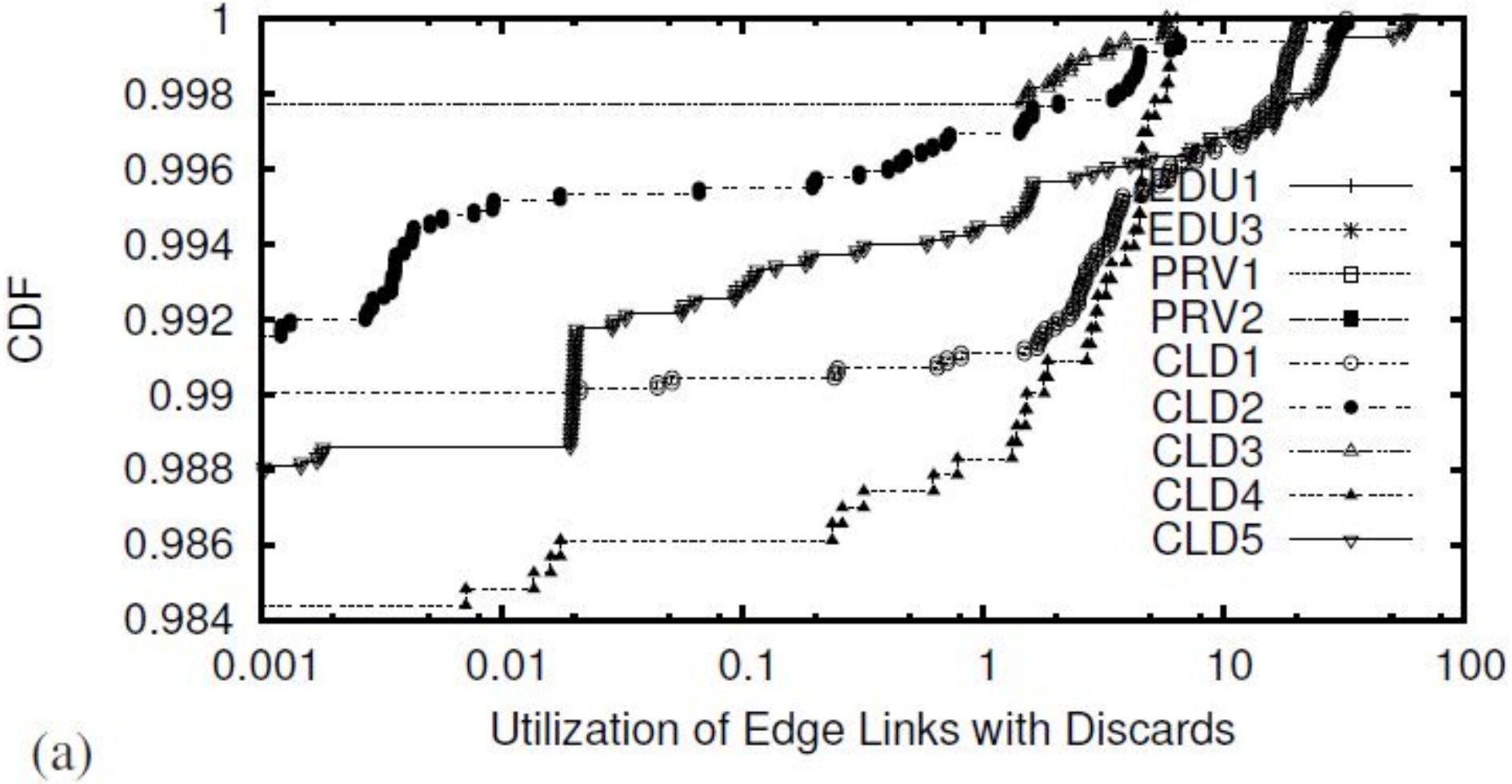
(b)

Size of Core Discards



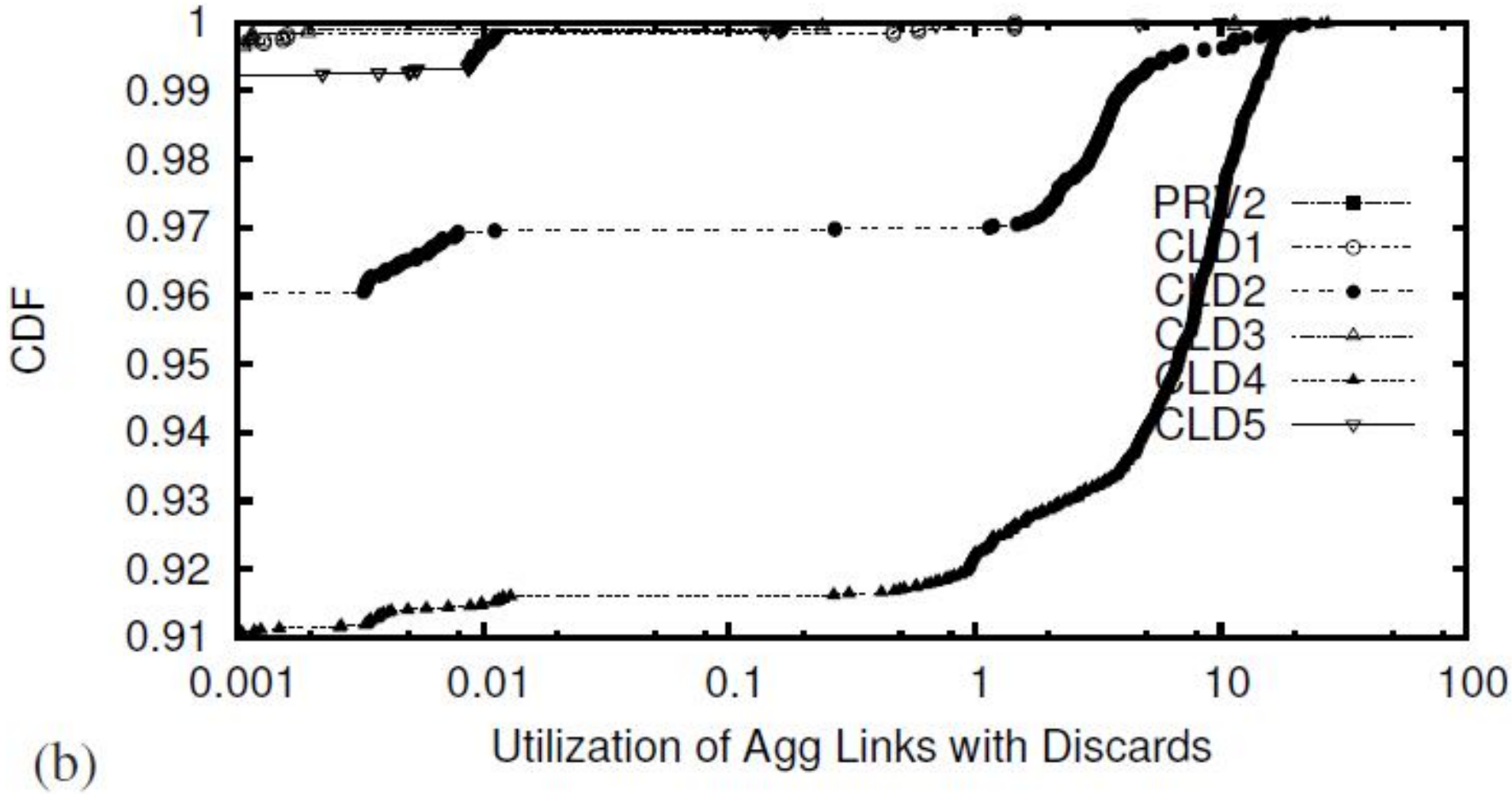
(c)

Utilization of Edge Links with Discards



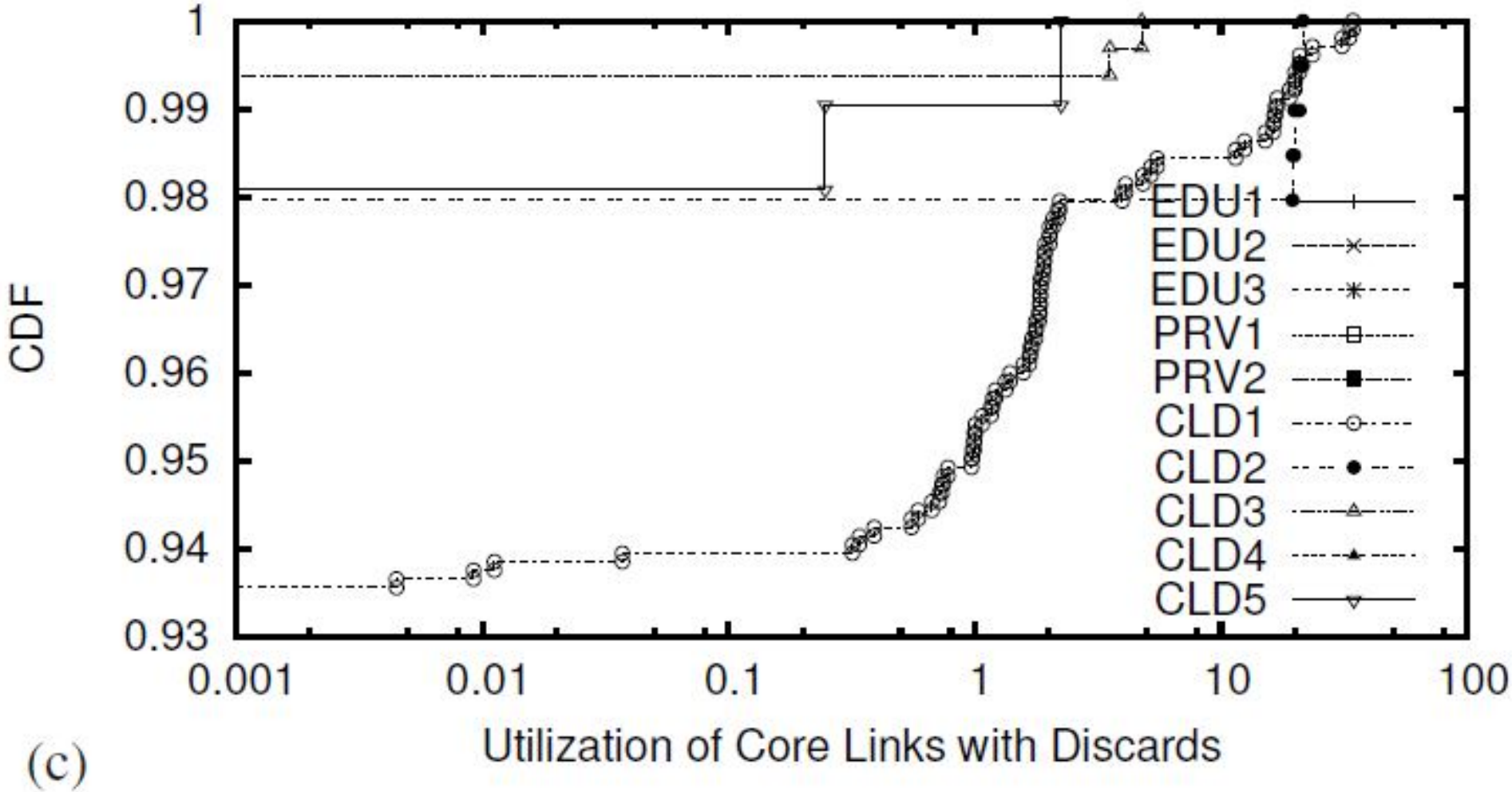
(a)

Utilization of Agg Links with Discards



(b)

Utilization of Core Links with Discards



(c)

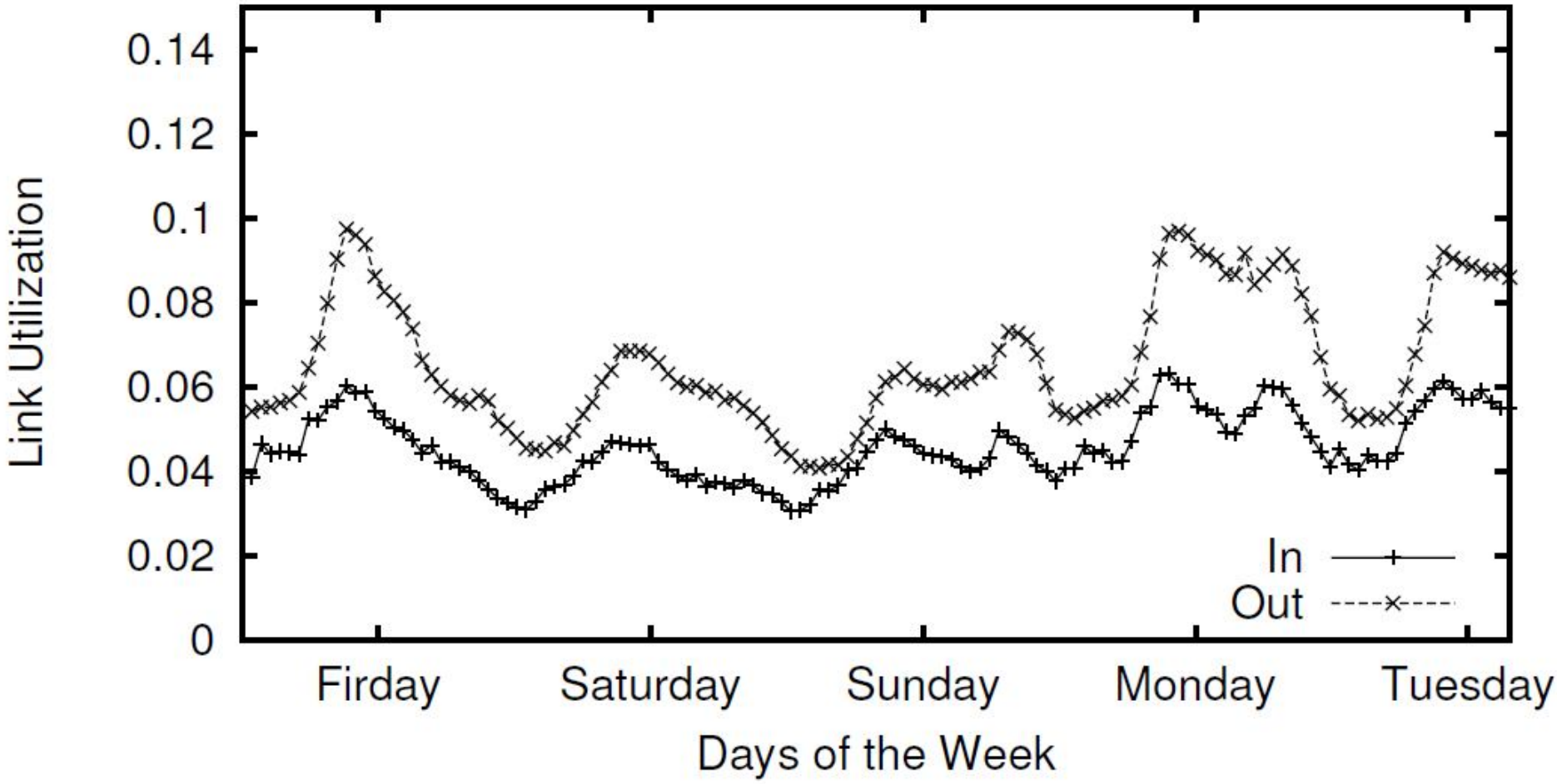


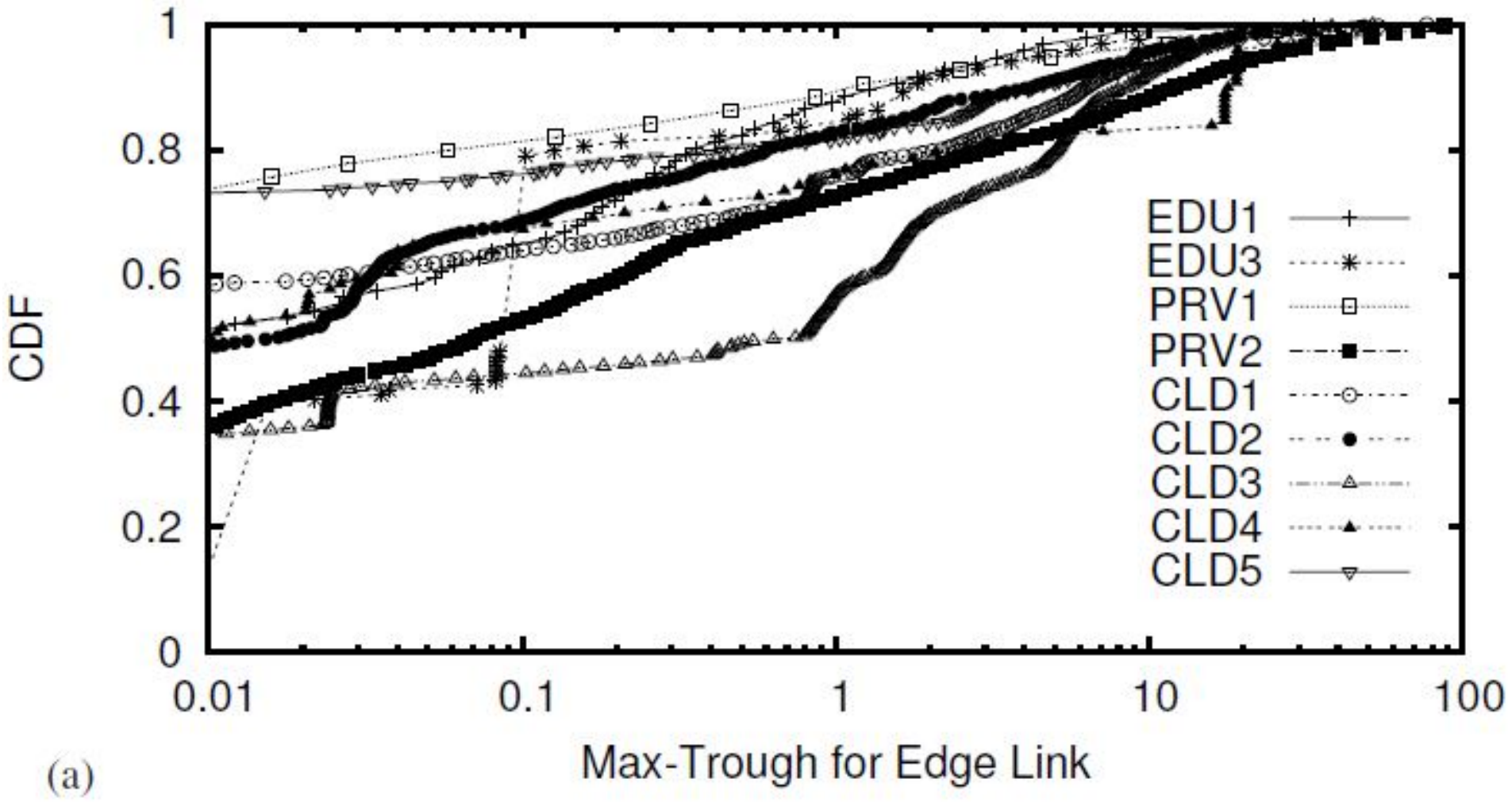
Figure 14: Time-of-Day/Day-of-Week traffic patterns.

Questions

(1)The utilizations vary over time ?

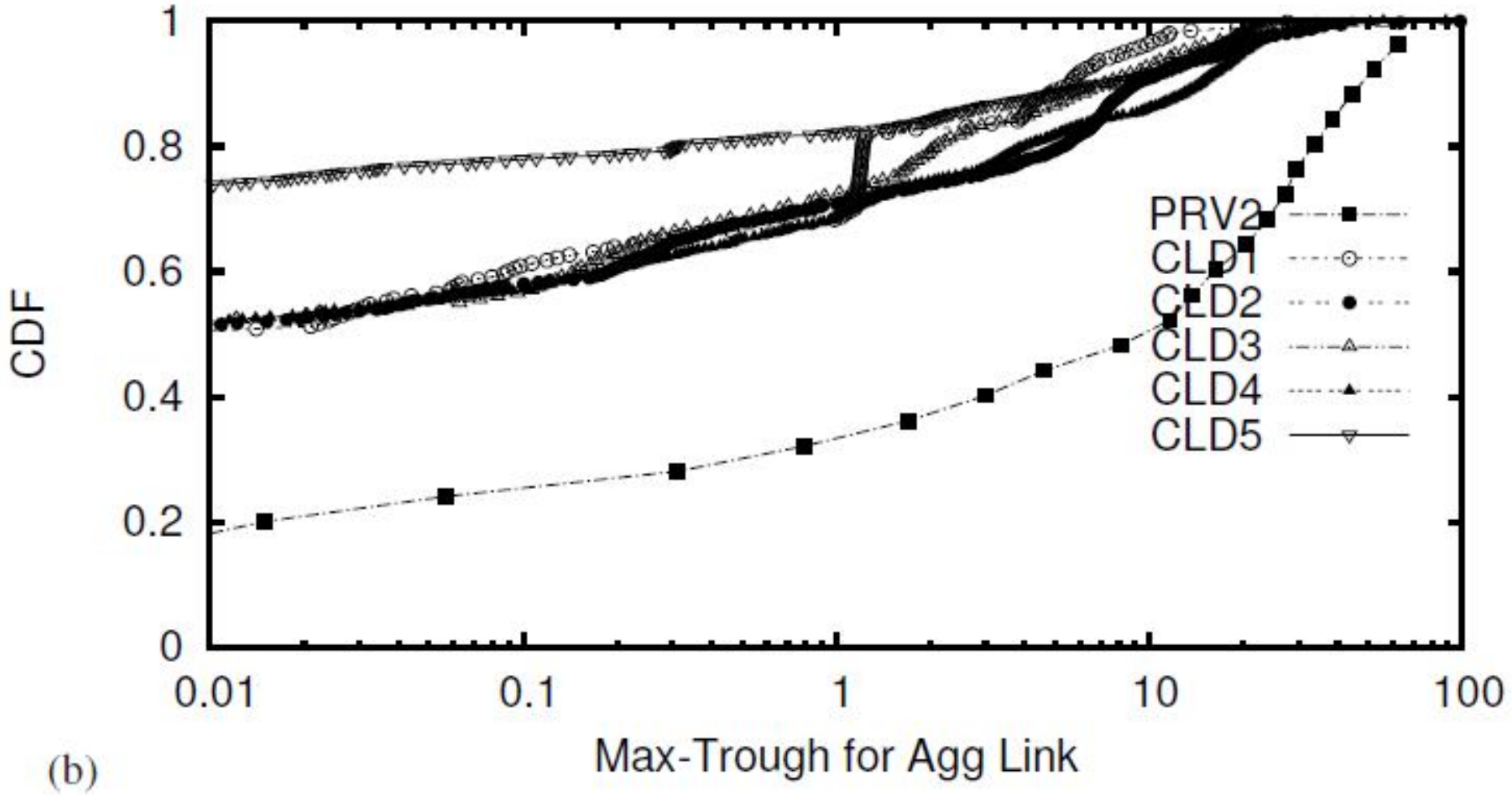
(2)Whether or not link utilizations are stable and predictable ?

Max-Trough for Edge Link

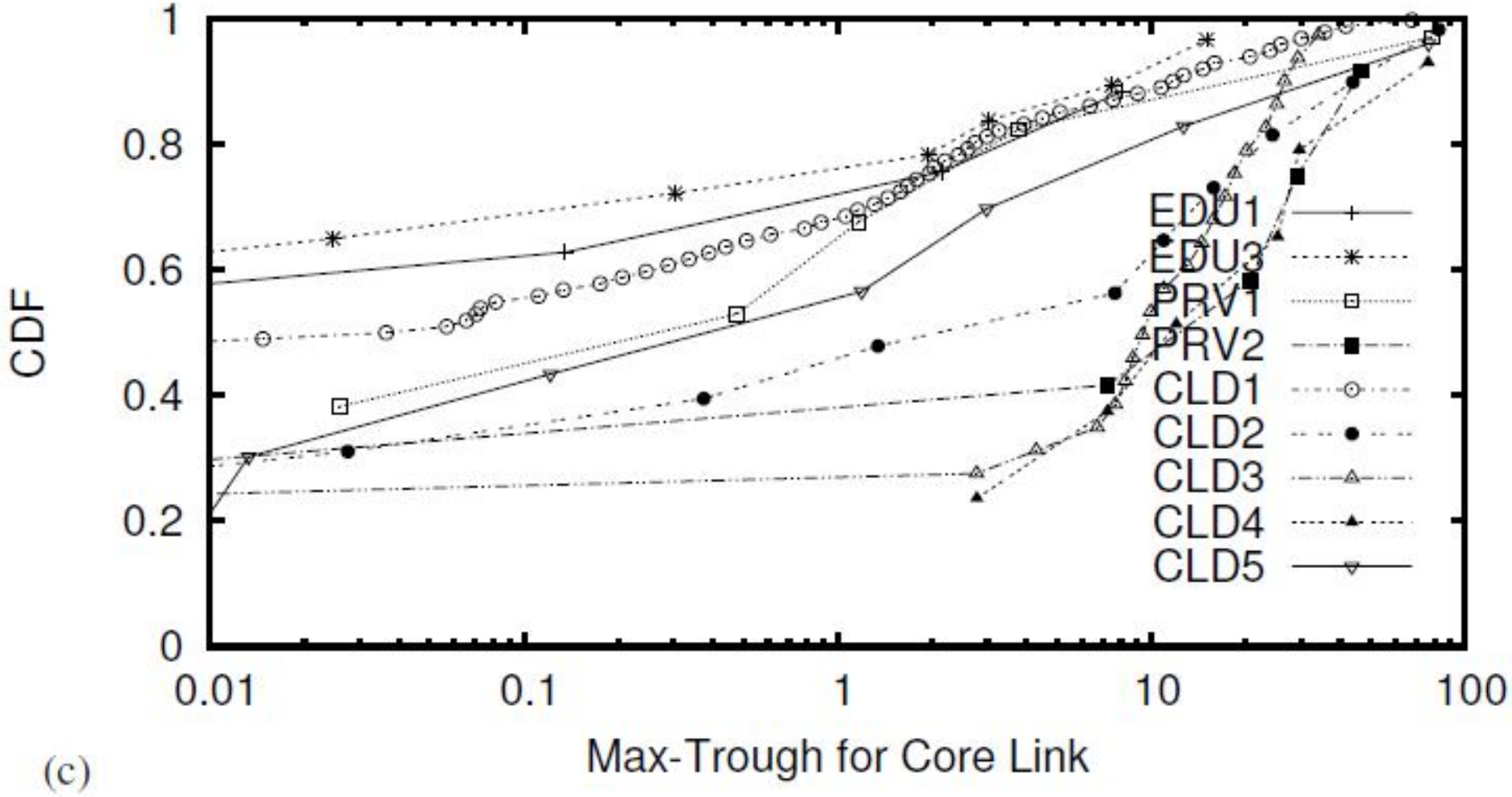


(a)

Max-Trough for Agg Link



Max-Trough for Core Link



(c)

Implications for Data Center Design

Bisection Bandwidth

- **Bisection capacity** : The aggregate capacity of core links.

- **The full bisection capacity** : The capacity that would be required to support servers communicating at full link speeds with arbitrary traffic matrices and no oversubscription.

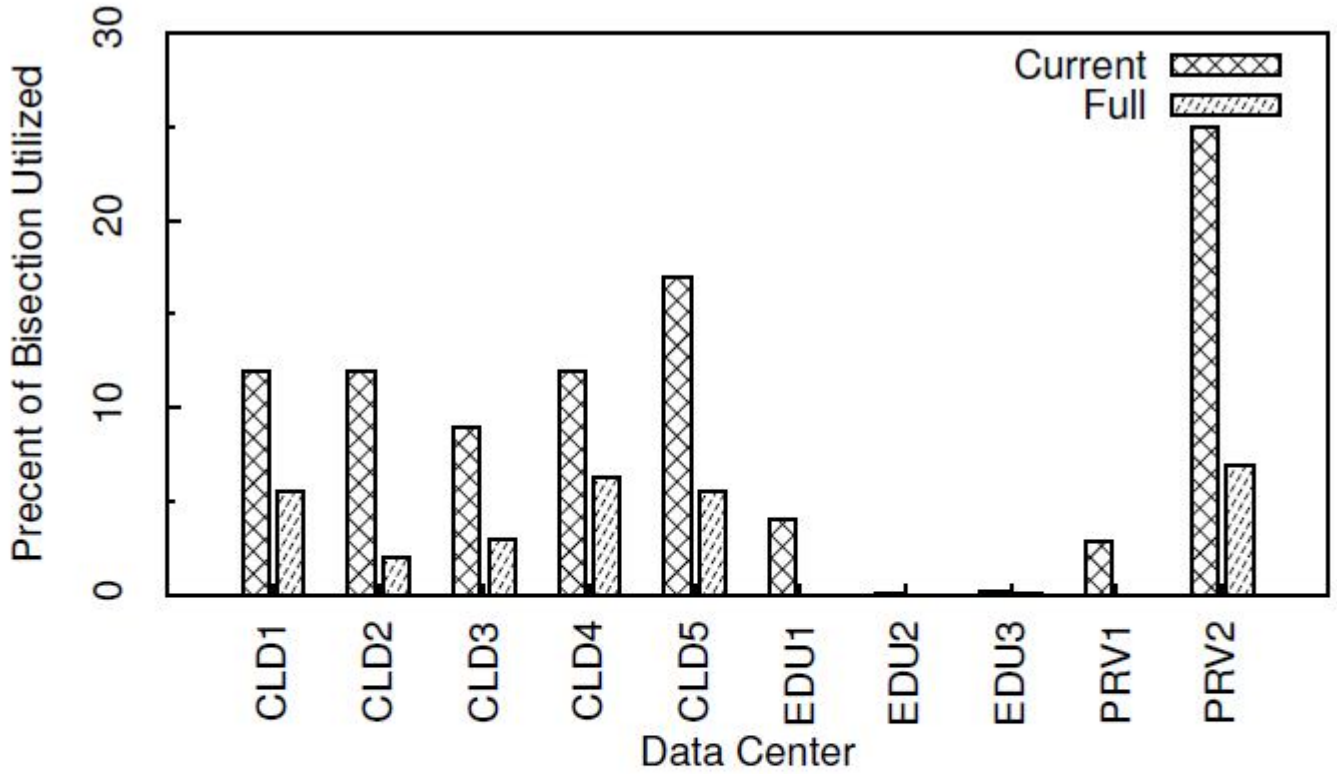


Figure 16: The first bar is the ratio of aggregate server traffic over Bisection BW and the second bar is the ratio of aggregate server traffic over full bisection capacity. The y-axis displays utilization as a percentage.

Conclusions (1/7)

- We see a wide variety of applications across the data centers , and find that application placement is non-uniform across racks.

- Traffic originating from a rack in a data center is ON/OFF in nature that fit heavy-tailed distributions.

Conclusions (2/7)

- In the cloud data centers, a majority of traffic originated by servers stays within the rack.

- For the university and private enterprise data centers, most of the traffic leaves the rack.

Conclusions (3/7)

- Most flows in the data centers are small in size (10KB).
- A significant fraction of which last under a few hundreds of milliseconds.
- The number of active flows per second is under 10,000 per rack.

Conclusions (4/7)

- Link utilizations are rather low in all layers but the core.

- The number of highly utilized core links never exceeds 25% in any data center.

Conclusions (5/7)

- Losses are not localized to links with persistently high utilization.

- Instead, losses occur at links with low average utilization implicating momentary spikes.

Conclusions (6/7)

- 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.

Conclusions (7/7)

- We determined that full bisection bandwidth is not essential for supporting current applications.

Understanding Data Center Traffic Characteristics

ACM SIGCOMM Computer Communication Review
Volume 40 Issue 1, January 2010

Outline

- Introduction
- Data set classification
- Empirical study
  - Macroscopic view
  - Microscopic view
- Generating fine-grained observations from coarse-grained data
- Conclusion

1.Introduction

- There is growing interest both in the research and the operations communities on data center network design.

- But very little is known about traffic characteristic within data center network
  - How do traffic and loss rates vary with time and with the location in a multi-tier data center ?

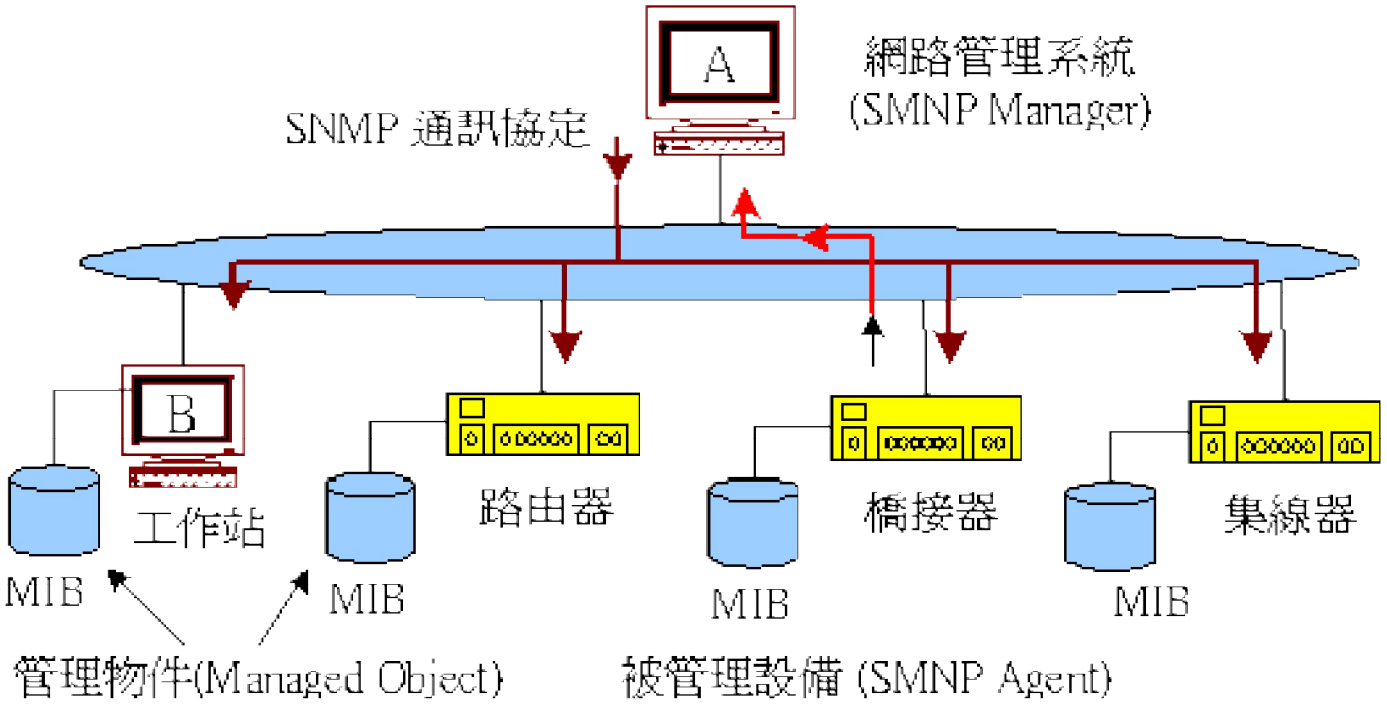
Motivation

- Recent research on data centers utilized toy traffic models or WAN-based models for evaluation (e.g., constant traffic)

- Our observations can be used to create more realistic workloads to evaluate current and future proposals for data center design and management.

Simple Network Management Protocol(SNMP)

- SNMP is used in network management system to monitor status of devices and also spot problem.



2.Data Sets

- Collected two sets of measurement data.

1. Comprised of SNMP data extracted from 19 corporate and enterprise data centers(search, video streaming, instant messaging, map reduce, and web applications)

2. Comprised of packet traces from five switches in one of the data centers.

Device information about the first data set

Data-Center Name	Fraction Core Devices	Frac Aggr Devices	Frac Edge Devices
DC1	0.000	0.000	1.000
DC2	0.667	0.000	0.333
DC3	0.500	0.000	0.500
DC4	0.500	0.000	0.500
DC5	0.500	0.000	0.500
DC6	0.222	0.000	0.778
DC7	0.200	0.000	0.800
DC8	0.200	0.000	0.800
DC9	0.000	0.077	0.923
DC10	0.000	0.043	0.957
DC11	0.038	0.026	0.936
DC12	0.024	0.072	0.904
DC13	0.010	0.168	0.822
DC14	0.031	0.018	0.951
DC15	0.013	0.013	0.973
DC16	0.005	0.089	0.906
DC17	0.016	0.073	0.910
DC18	0.007	0.075	0.918
DC19	0.005	0.026	0.969

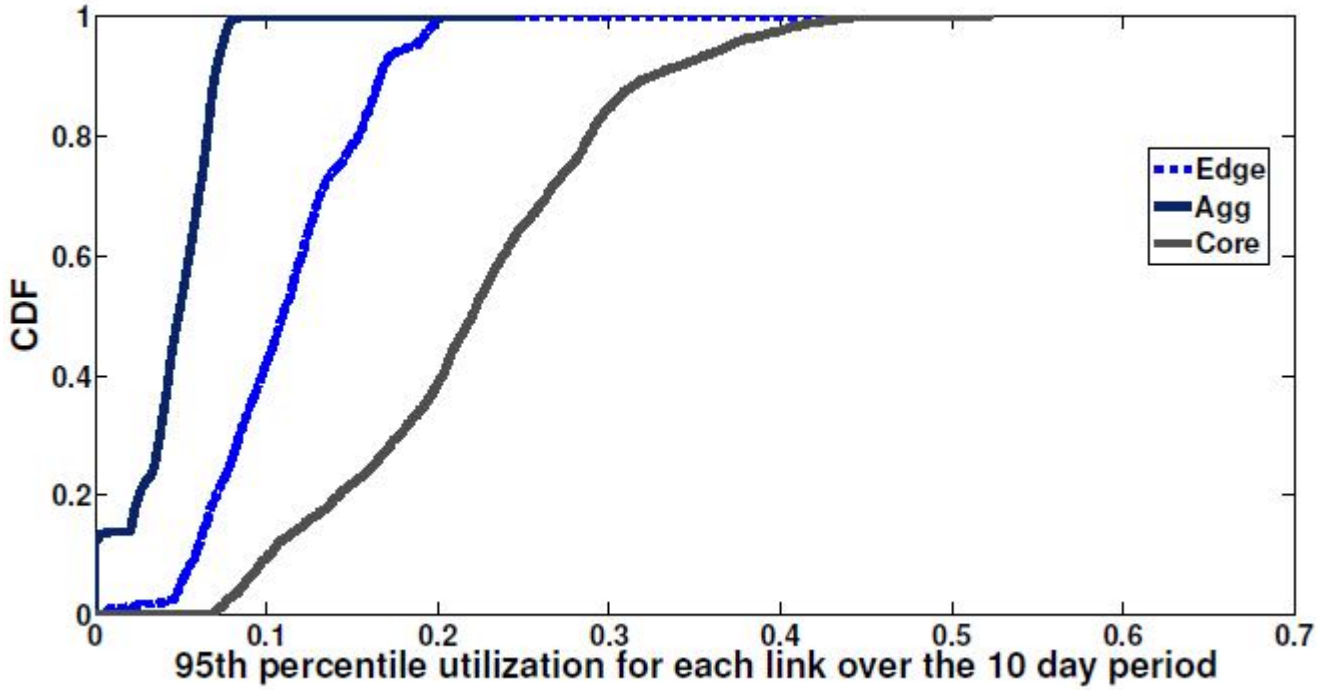
Table 1: We present information about the devices in the 19 data centers studied. For each data center, we present the total number of devices and the fraction of devices in each layer.

3. EMPIRICAL STUDY

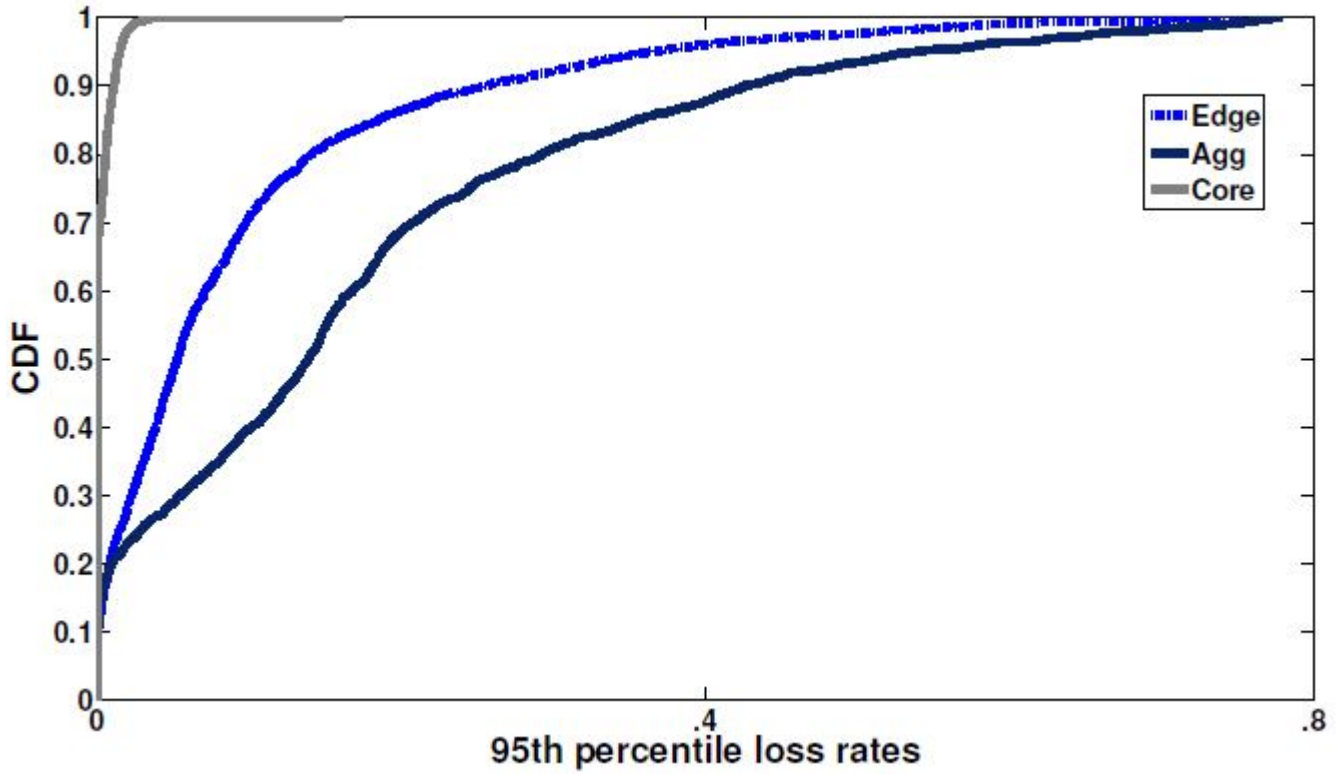
- First examine the SNMP data to study the link utilization and packet loss of core, edge, and aggregation devices.
- Then characterize the temporal patterns of data center traffic using the packet traces.
- The observations we make will help to build a realistic traffic model

3.1 Data Center Traffic: Macroscopic View

- Following shows the CDF of the 95th percentile utilization of those used links (where the 95th percentile is computed over all the 5 minute intervals where the link was utilized).



Data Loss rate

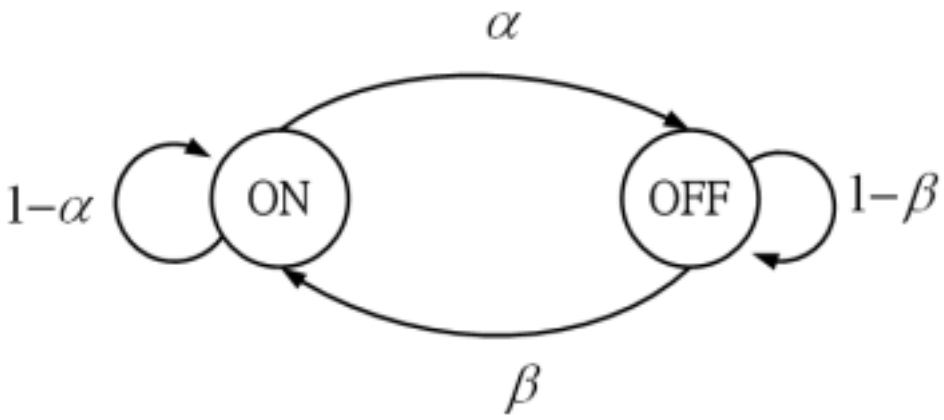


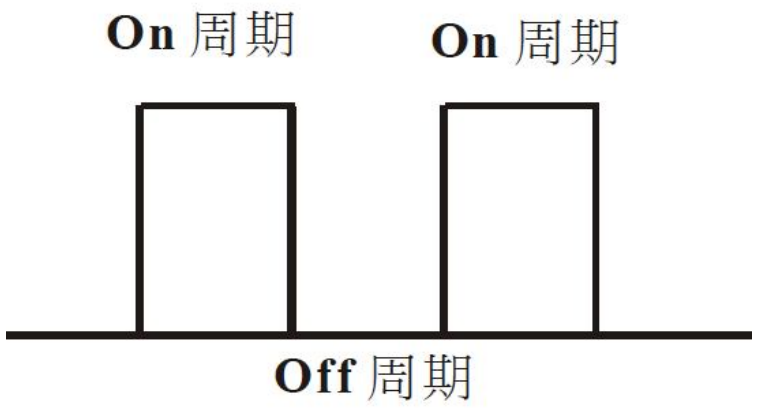
3.2 Data Center Traffic: Microscopic View

- Although aggregate traffic rate may be below the link capacity, a momentary traffic burst can lead to short-lived congestion in the network.

- To understand the properties of such a traffic burst, more detailed information is needed.

On/off traffic





On/off 模型

On/off traffic (cont.)

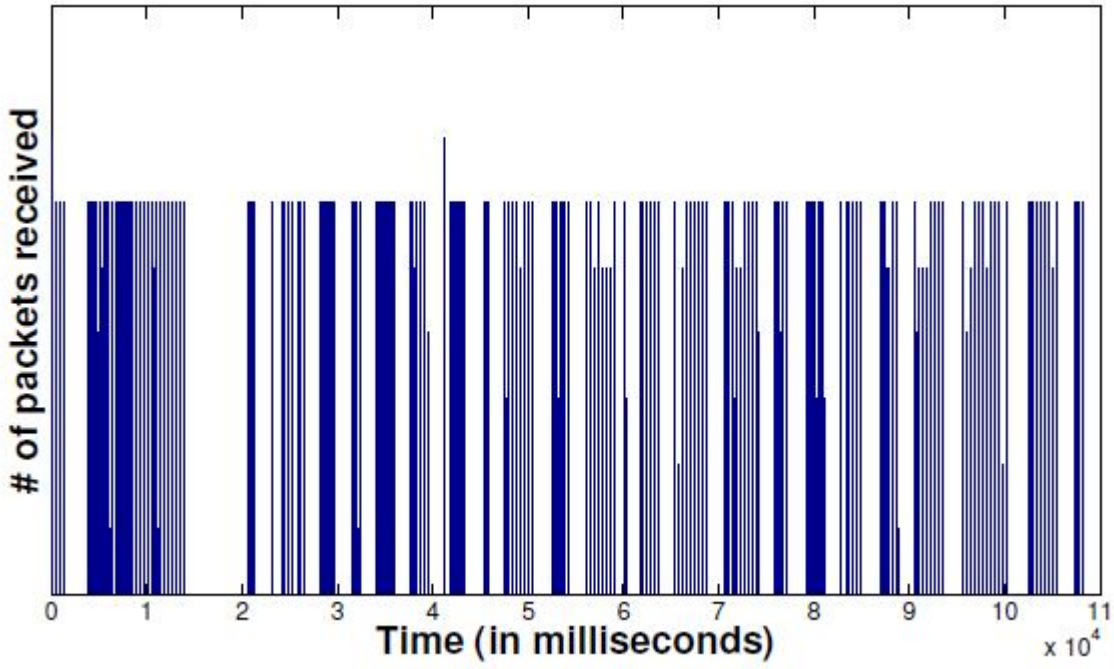
- We define a **period**$_{on}$ as a longest continual period during which all the packet inter-arrival times are smaller than **arrival**$_{95}$.

- **Period**$_{off}$ is a period between two on periods.

- To characterize this ON/OFF traffic pattern, we focus on three aspects:
  - the durations of the ON periods
  - the durations of the OFF periods
  - the packet inter-arrival times

Pictorial proof of ON/OFF characteristics

- Following shows a time-series of the number of packets received during a short time interval at one of the switches.



CDF of the distribution of the arrival times of packets

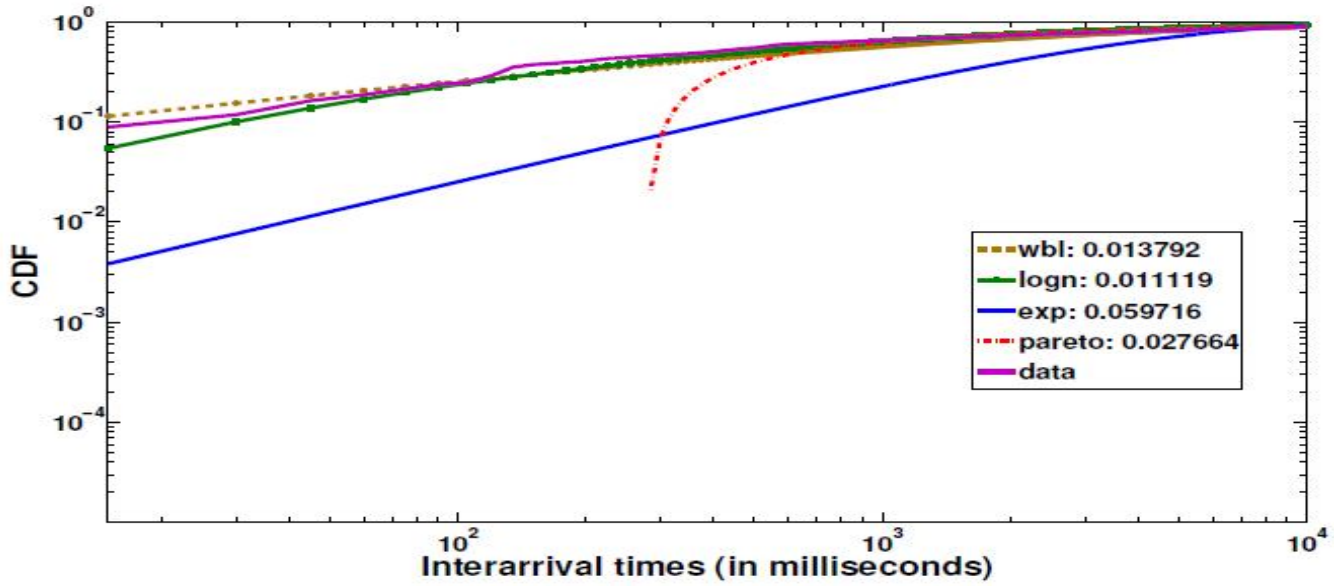


Figure 6: CDF of the distribution of the arrival times of packets at one of the switches in DC10. The figure contains best fit curve for lognormal,weibul, pareto, and exponential as well as the least mean errors for each curve. We notice that the lognormal fit produces the least error

CDF of the distribution of the ON period lengths at one of the switches

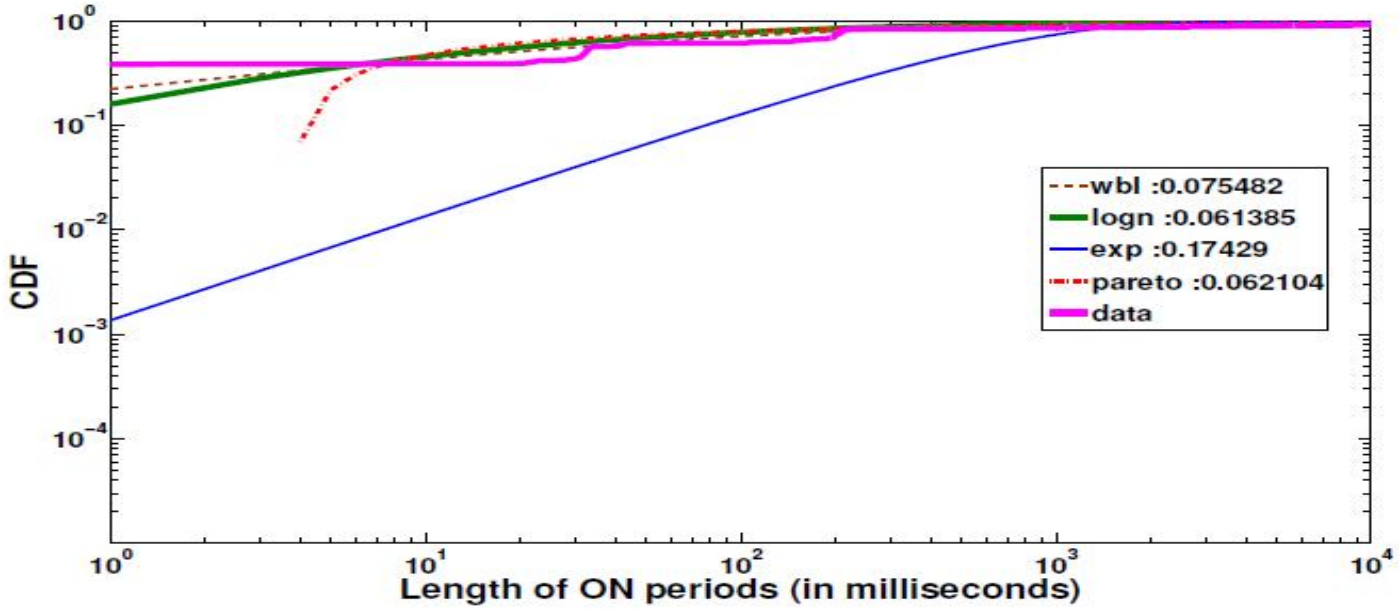


Figure 8: CDF of the distribution of the ON period lengths at one of the switches in DC10. The figure contains best fit curve for lognormal,weibul, pareto, and exponential as well as the least mean errors for each curve. We notice that the lognormal fit produces the least error

CDF of the distribution of OFF period lengths at one of the switches

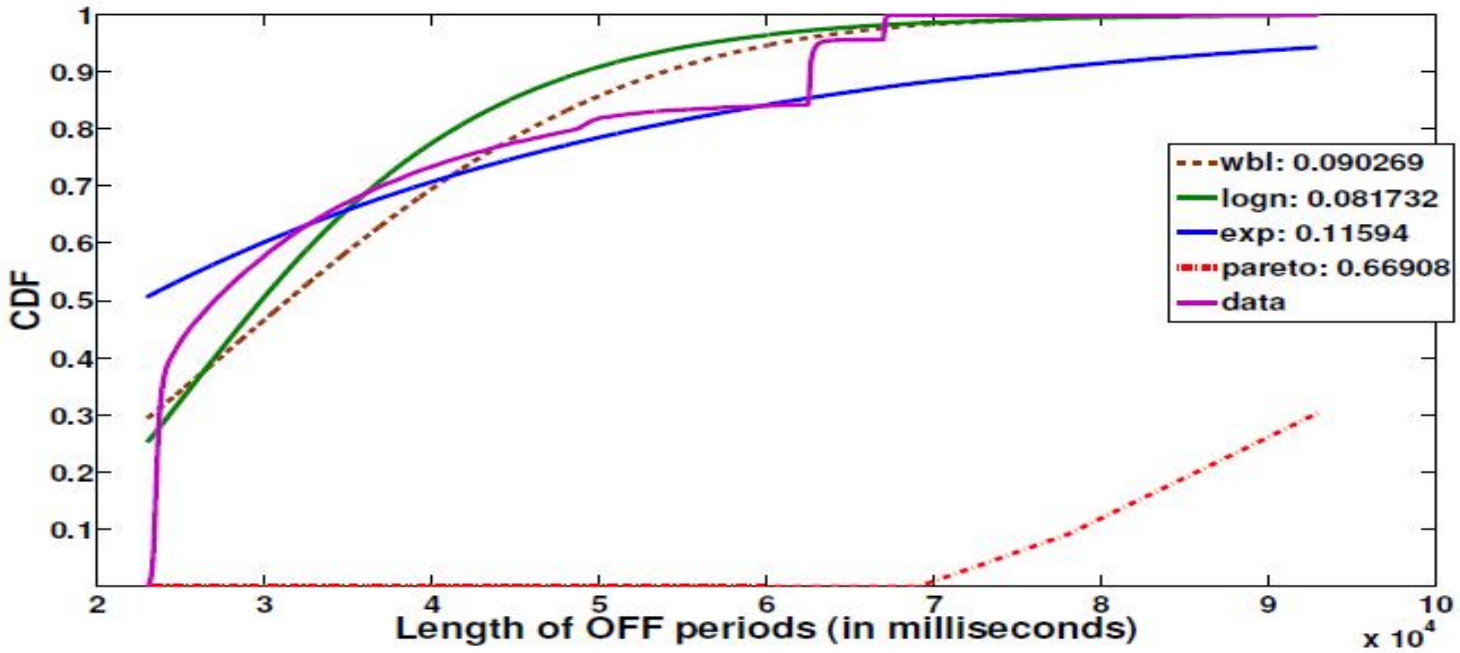


Figure 7: CDF of the distribution of OFF period lengths at one of the switches in DC10. The figure contains best fit curve for lognormal,weibul, pareto, and exponential as well as the least mean errors for each curve. We notice that the lognormal fit produces the least error

3.3 summarize

- The durations of the ON/OFF periods and the packet inter-arrival times within ON periods all follow some lognormal distributions

- Finding the appropriate lognormal random processes that can generate traffic under certain volume and loss rate conditions.

4. GENERATING FINE-GRAINED OBSERVATIONS FROM COARSE-GRAINED DATA

- Access to fine-grained data provides insight into traffic characteristics such as time-scales of congestion which can then be used to inform traffic engineering and switching fabric design.

4.1 Parameter Discovery Algorithm

- There are four challenges in developing such an algorithm:
  1. Developing an accurate scoring function for each point.
  2. Determining a set of terminating conditions.
  3. Defining a heuristic to avoid getting stuck in local maxima and selecting an appropriate starting point.
  4. Defining the neighbors of a point and selecting the next move.

Parameter Discovery Algorithm (cont.)

- Our framework takes as input the distribution of SNMP-derived volumes (***volume***$_{SNMP}$), and loss rates(***lossrate***$_{SNMP}$) for a given link at the edge of the data center.

- The approach returns as output, the parameters for the 3 distributions (**on**$_{times}$, **off**$_{times}$, **arrival**$_{times}$) that provide fine-grained descriptions of the traffic on the edge link.

4.1.1 Scoring Function

- To score the parameters at a point, we utilize two techniques:
  - We use a heuristic algorithm to approximate the distributions of loss and volume(we refer to these as **volume**$_{\text{generated}}$ and **lossrate**$_{\text{generated}}$)
  - we employ a statistical test to score the parameters based on the similarity of the generated distributions to the input distributions **volume**$_{\text{SNMP}}$ and **lossrate**$_{\text{SNMP}}$

Parameter Discovery Algorithm (cont.)

- We use a simple heuristic approach to obtain the loss rate and volume distributions generated by the traffic parameters (μ_{on}, σ_{on}, μ_{off} , σ_{off} , $\mu_{arrival}$, $\sigma_{arrival}$) corresponding to a given poi nt in the search space

- Obtaining $\mathbf{volume}_{generated}$ and $\mathbf{lossrate}_{generated}$.

DERIVEONOFFTRAFFICPARAMS($\mu_{on}, \sigma_{on}, \mu_{off}, \sigma_{off}, \mu_{arrival}, \sigma_{arrival}$)

// Calculate the mean on and OFF period lengths

1 $mean_{on} \leftarrow \exp(\mu_{on}) + \sigma_{on}$

2 $mean_{off} \leftarrow \exp(\mu_{off}) + \sigma_{off}$

// Determine the total on-time in a 5 minute interval

3 $total_{on} = 300 * (mean_{on}/(mean_{off} + mean_{on}))$

// Calculate the average number of ON periods

4 $NumOnPeriods = total_{on}/mean_{on}$.

// Calculate the maximum number of bytes

// that can be sent during the ON period

5 $link_{capacity} = links_{speed} * mean_{on}/8$.

// Determine how much bytes can be absorbed by buffering

// during the OFF period

6 $buf_{capacity} = min(bitsofbuffering, links_{speed} * mean_{off})/8$

// Iterate over ON period to calculate net volume and loss rate

// observed over the 5 minute interval

7 for $i = 0$ to $NumOnPeriods$

a. $a_i \in A\{interarrival\ time\ distribution\}$

b. $vol_{on} = (mean_{on}/a_i) * pktSize$

c. $vol_{total} += min(vol_{on}, link_{capacity} + buf_{capacity})$

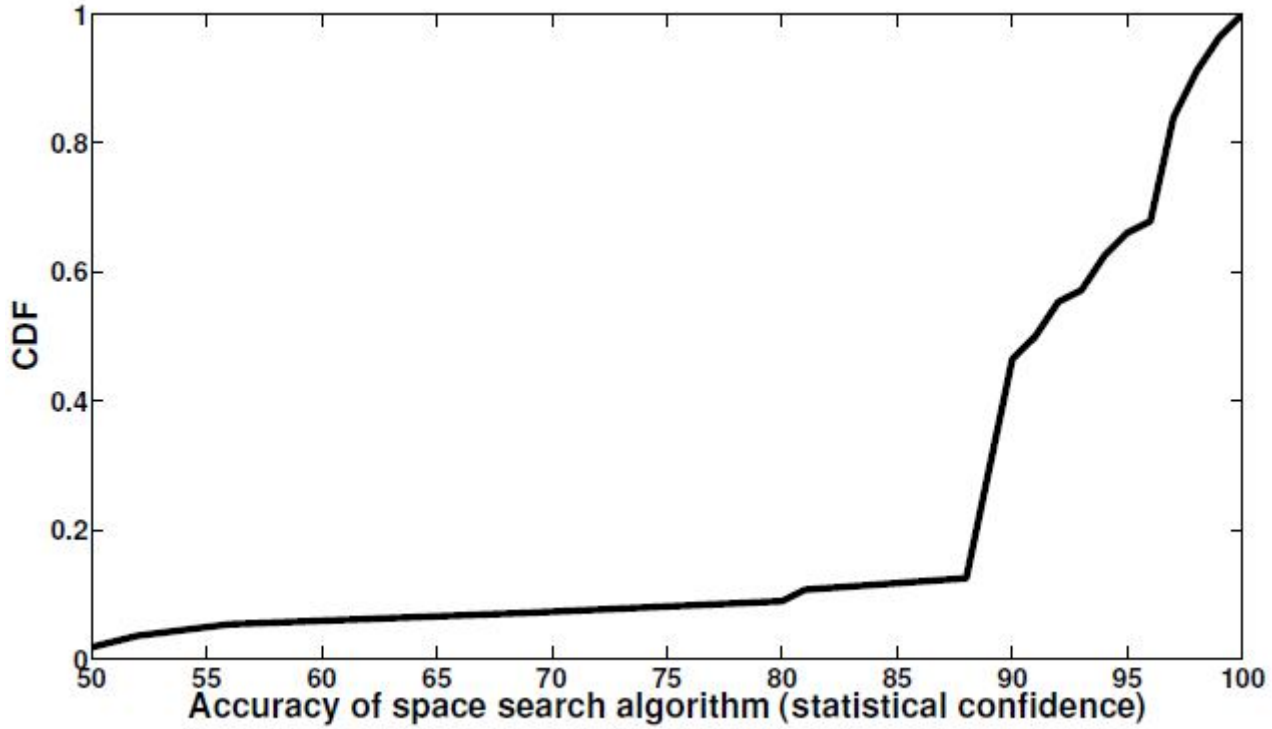
d. $loss_{total} += max(vol_{on} - link_{capacity} - buf_{capacity}, 0)$

Parameter Discovery Algorithm (cont.)

- We run the above subroutine several (100) times to obtain multiple samples for $\mathbf{vol}_{\text{total}}$ and $\mathbf{loss}_{\text{total}}$. From these samples, we derive the distributions $\mathbf{volume}_{\text{generated}}$ and $\mathbf{loss\ rate}_{\text{generated}}$.

Validation by simulation

- Our algorithm can find appropriate arrival processes for over 70% of the devices with at least a 90% confidence according to the **Wilcoxon** test.



5.CONCLUSION AND FUTURE WORK

- We found that links in the core of data centers are more heavily utilized on average, but those closer to the edge observe higher losses on average.

- Using a limited set of packet traces collected at a handful of data center switches we found preliminary evidence of ON-OFF traffic patterns.

- This general framework can be used to examine how to design traffic engineering mechanisms that ideally suit the prevailing traffic patterns in a data center.

Question & Answer

- Q1 : What is the utilization of links at different layers (i.e. Core、Aggregation、Access) in a data center, please write down the orders.

- Ans : Core > Aggregation > Access