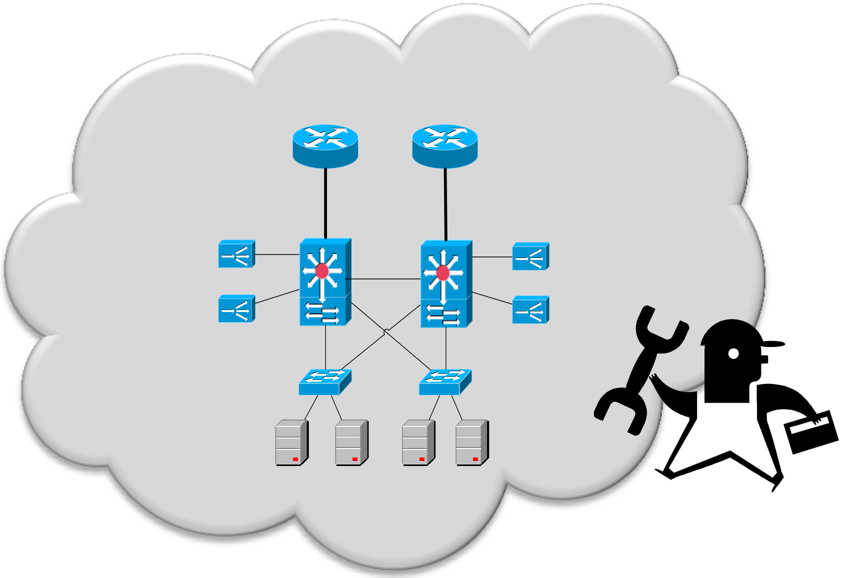
SIGCOMM 2011
Toronto, ON
Aug. 18, 2011

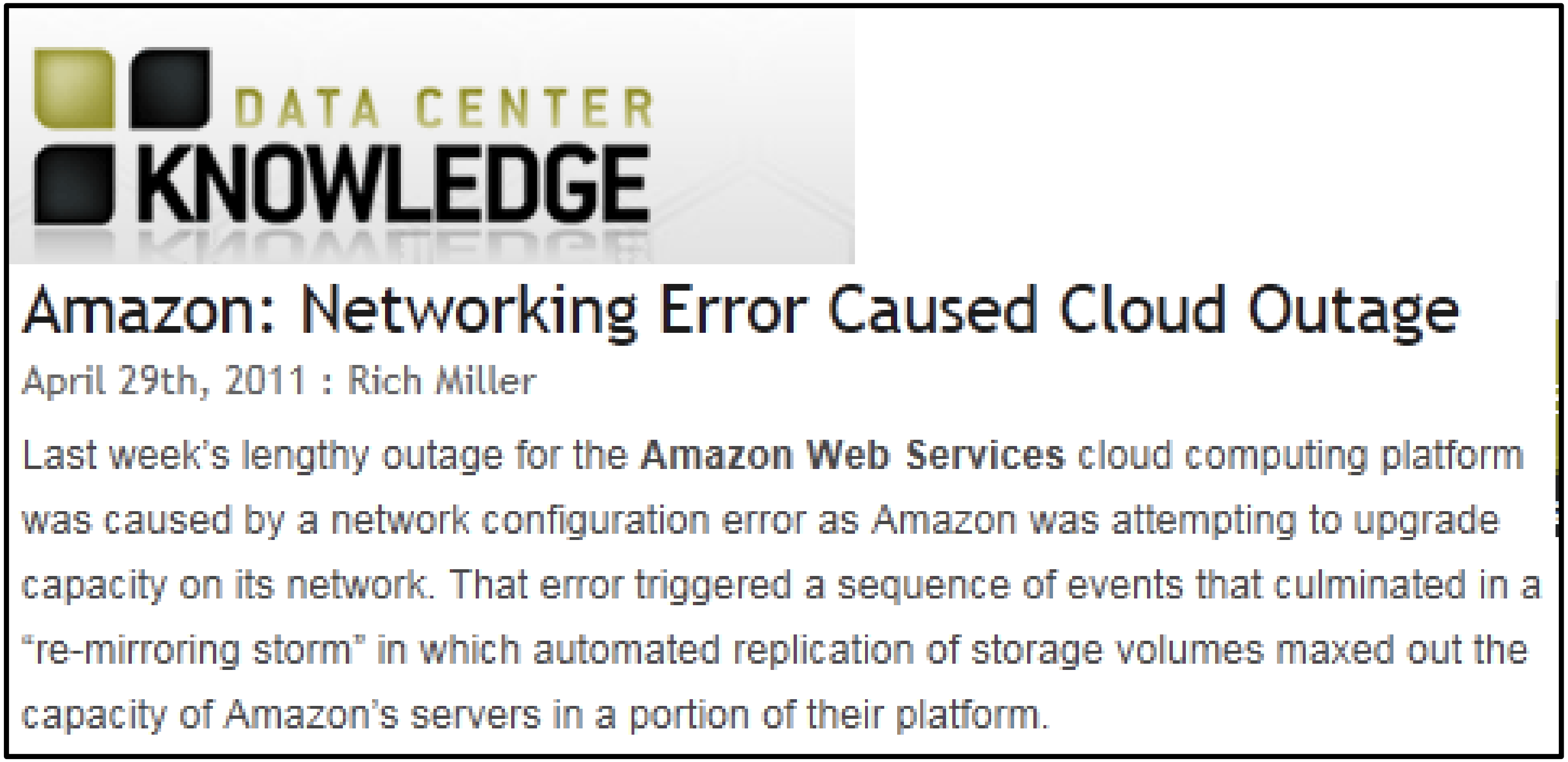


Understanding Network Failures in Data Centers: Measurement, Analysis and Implications

Phillipa Gill
University of Toronto

Navendu Jain & Nachiappan Nagappan
Microsoft Research

Motivation



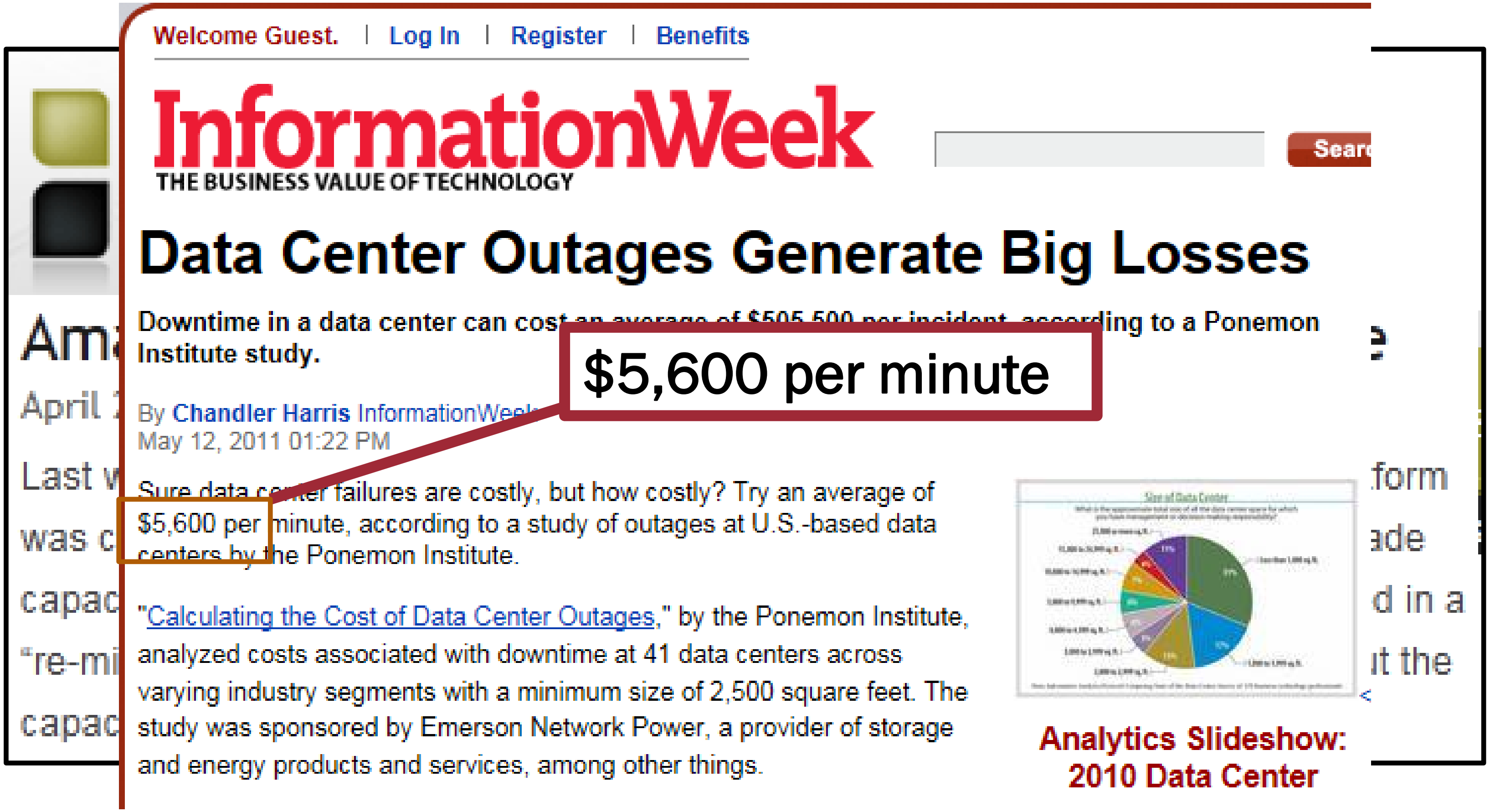
DATA CENTER KNOWLEDGE

Amazon: Networking Error Caused Cloud Outage

April 29th, 2011 : Rich Miller

Last week's lengthy outage for the **Amazon Web Services** cloud computing platform was caused by a network configuration error as Amazon was attempting to upgrade capacity on its network. That error triggered a sequence of events that culminated in a "re-mirroring storm" in which automated replication of storage volumes maxed out the capacity of Amazon's servers in a portion of their platform.

Motivation



Welcome Guest. | Log In | Register | Benefits

InformationWeek
THE BUSINESS VALUE OF TECHNOLOGY

Data Center Outages Generate Big Losses

Downtime in a data center can cost an average of $505,500 per incident, according to a Ponemon Institute study.

By Chandler Harris InformationWeek
May 12, 2011 01:22 PM

Sure data center failures are costly, but how costly? Try an average of $5,600 per minute, according to a study of outages at U.S.-based data centers by the Ponemon Institute.

"Calculating the Cost of Data Center Outages," by the Ponemon Institute, analyzed costs associated with downtime at 41 data centers across varying industry segments with a minimum size of 2,500 square feet. The study was sponsored by Emerson Network Power, a provider of storage and energy products and services, among other things.

Analytics Slideshow: 2010 Data Center

$5,600 per minute

We need to understand failures to prevent and mitigate them!

Overview

Our goal: Improve reliability by understanding network failures

1. Failure **characterization**
   - Most failure prone components
   - Understanding root cause
2. What is the **impact** of failure?
3. Is **redundancy** effective?

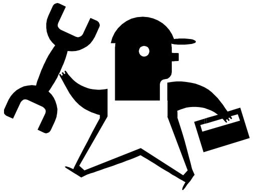
Our contribution: First large-scale empirical study of network failures across multiple DCs

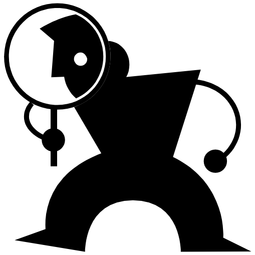
- Methodology to extract failures from noisy data sources.
- Correlate events with network traffic to estimate **impact**
- Analyzing implications for future data center networks

Road Map

Motivation

Background & Methodology



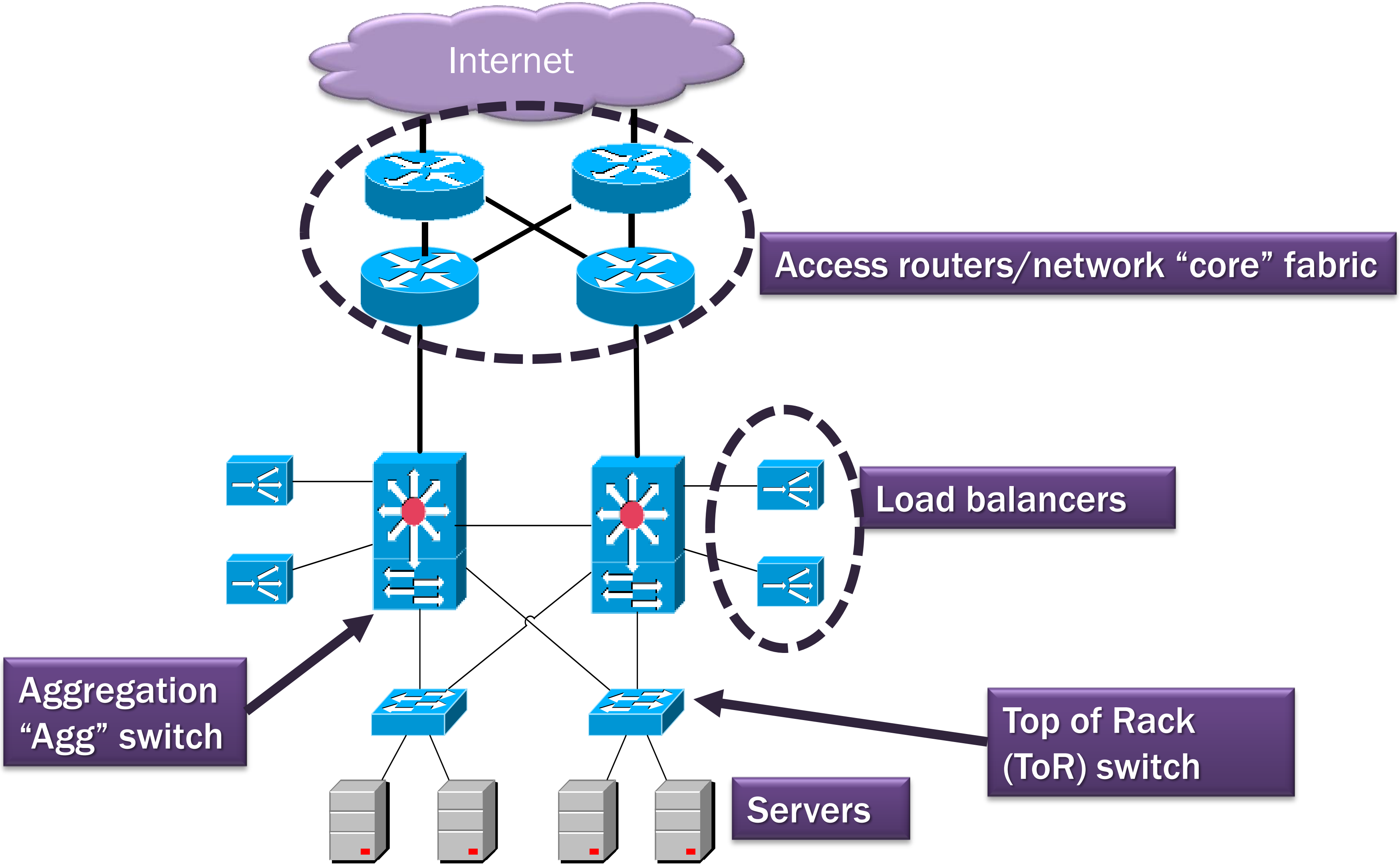


Results

1. Characterizing failures
2. Do current network redundancy strategies help?

Conclusions

Data center networks overview



Access routers/network "core" fabric

Load balancers

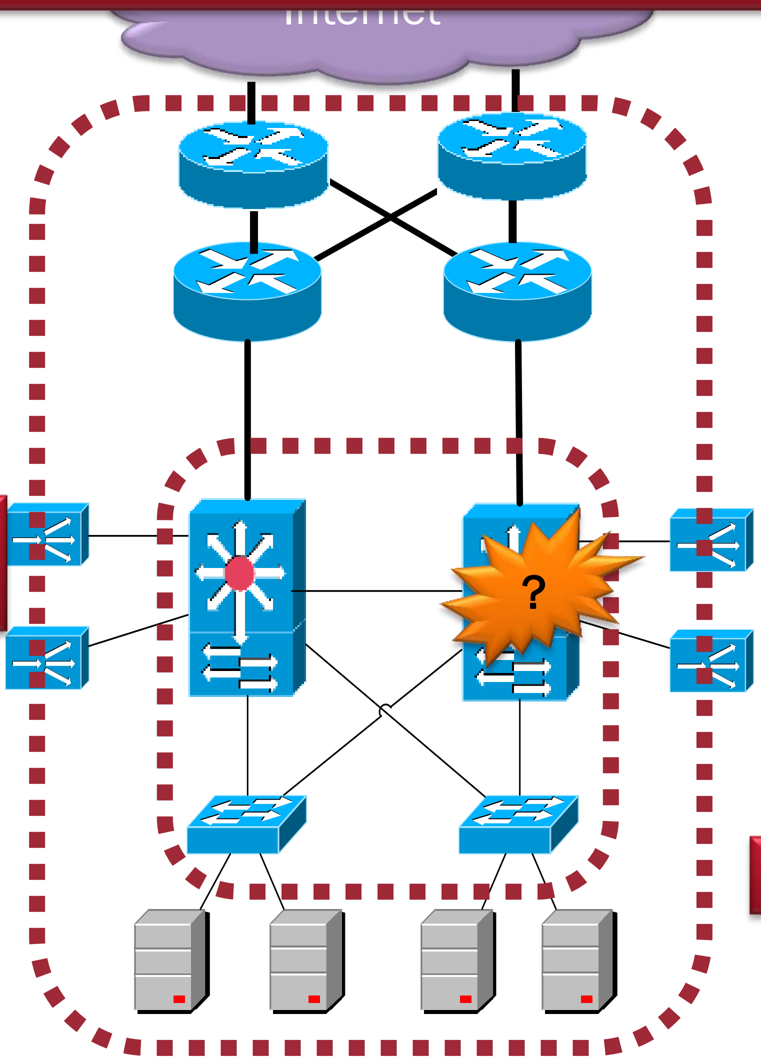
Aggregation "Agg" switch

Top of Rack (ToR) switch

Servers

Data center networks overview

Which components are most failure prone?



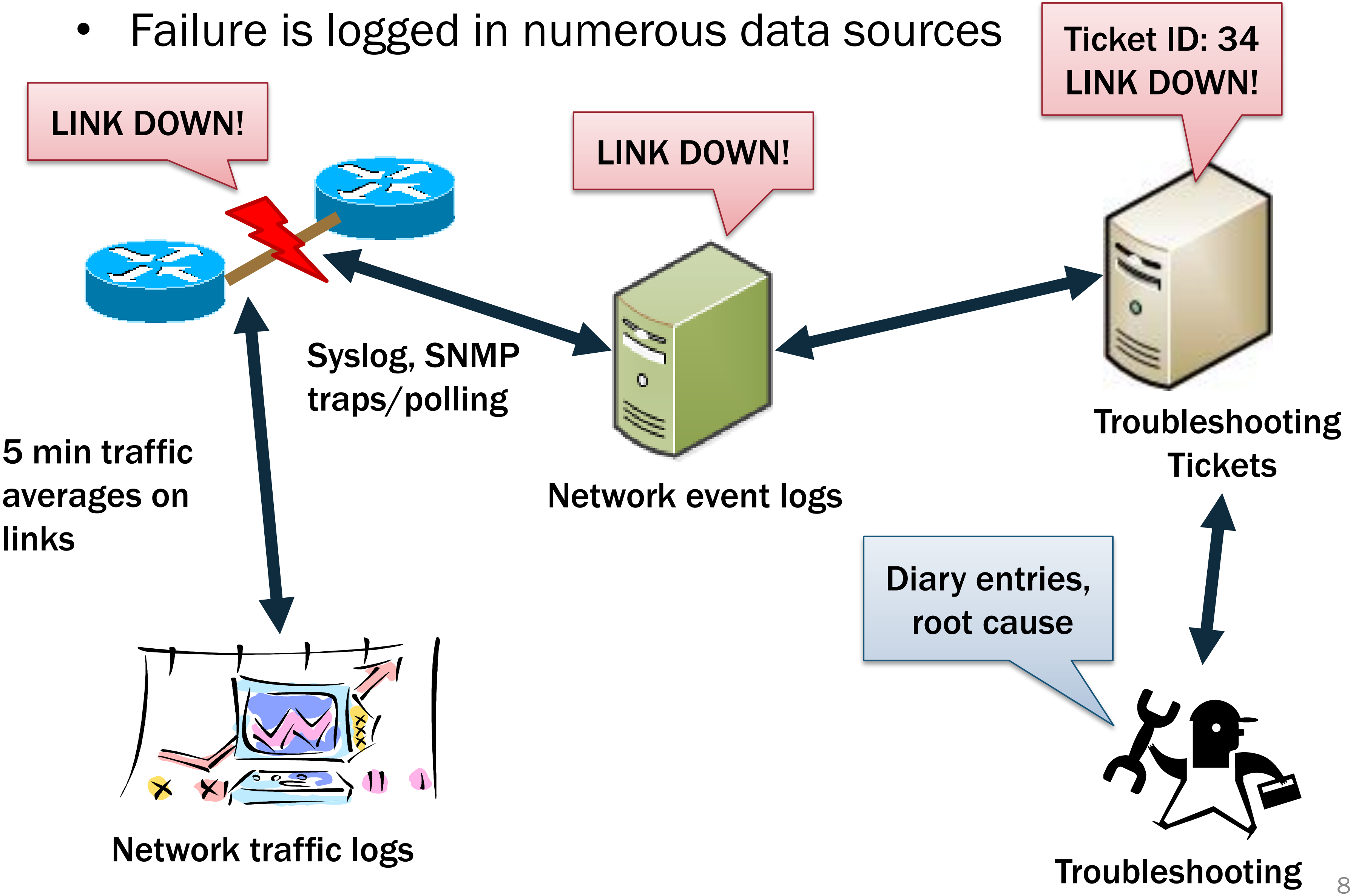
How effective is redundancy?

What is the impact of failure?

What causes failures?

Failure event information flow

- Failure is logged in numerous data sources



Ticket ID: 34
LINK DOWN!

LINK DOWN!

LINK DOWN!

Syslog, SNMP traps/polling

5 min traffic averages on links

Network event logs

Troubleshooting Tickets

Diary entries, root cause

Network traffic logs

Troubleshooting

Data summary

- One year of event logs from Oct. 2009-Sept. 2010
  - Network event logs and troubleshooting tickets

- Network event logs are a combination of Syslog, SNMP traps and polling
  - Caveat: may miss some events e.g., UDP, correlated faults

- Filtered by operators to ***actionable*** events
  - … still many warnings from various software daemons running

Key challenge: How to extract failures of interest?

Extracting failures from event logs

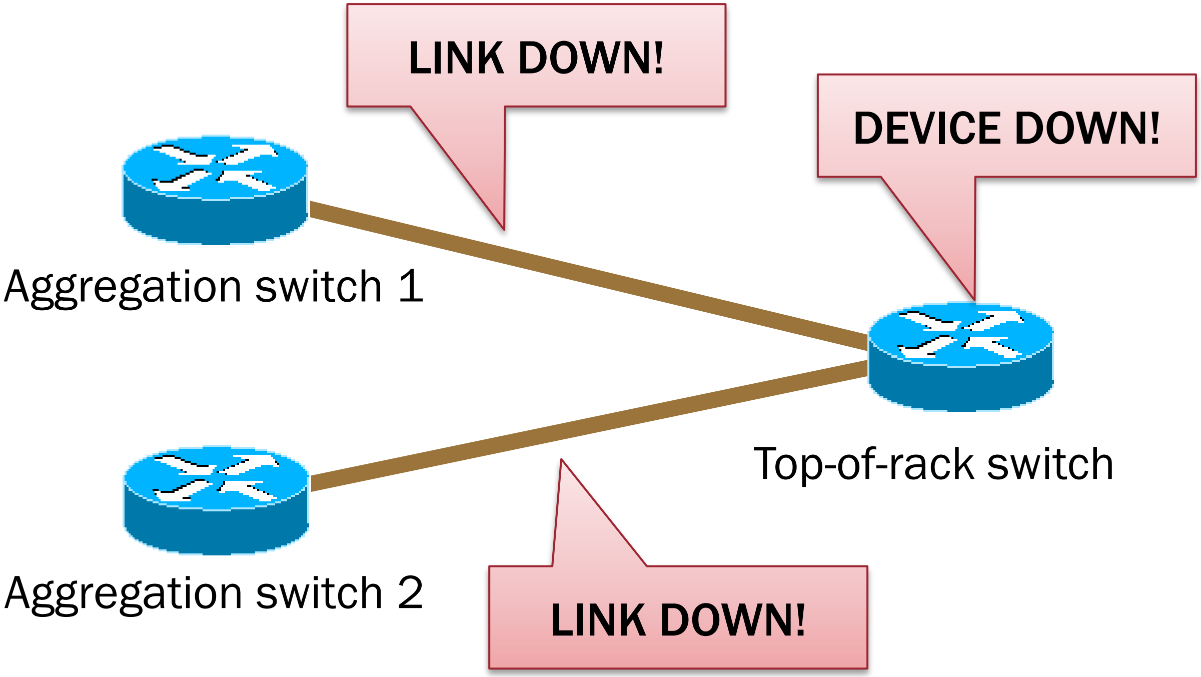
Network event logs

- **Defining failures**
  - **Device failure:** device is no longer forwarding traffic.
  - **Link failure:** connection between two interfaces is down. Detected by monitoring interface state.

- **Dealing with inconsistent data:**
  - **Devices:**
    - Correlate with link failures
  - **Links:**
    - Reconstruct state from logged messages
    - Correlate with network traffic to determine impact

Reconstructing device state

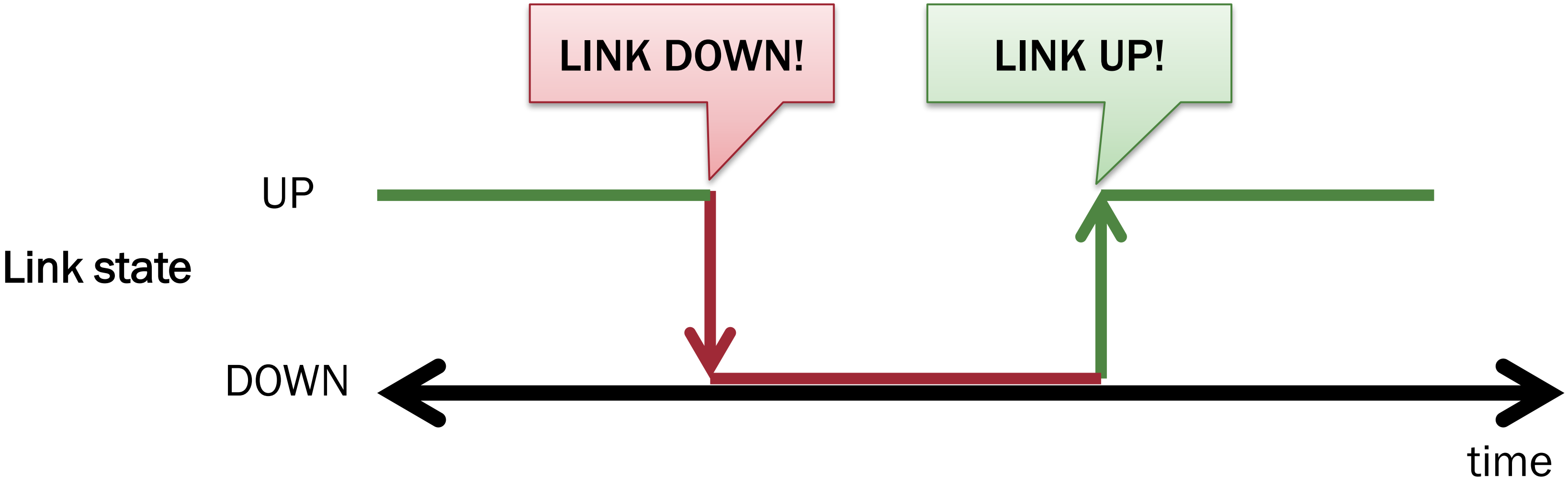
- Devices may send spurious DOWN messages
- Verify **at least one** link on device fails within five minutes
  - Conservative to account for message loss (correlated failures)



This sanity check reduces device failures by 10x

Reconstructing link state

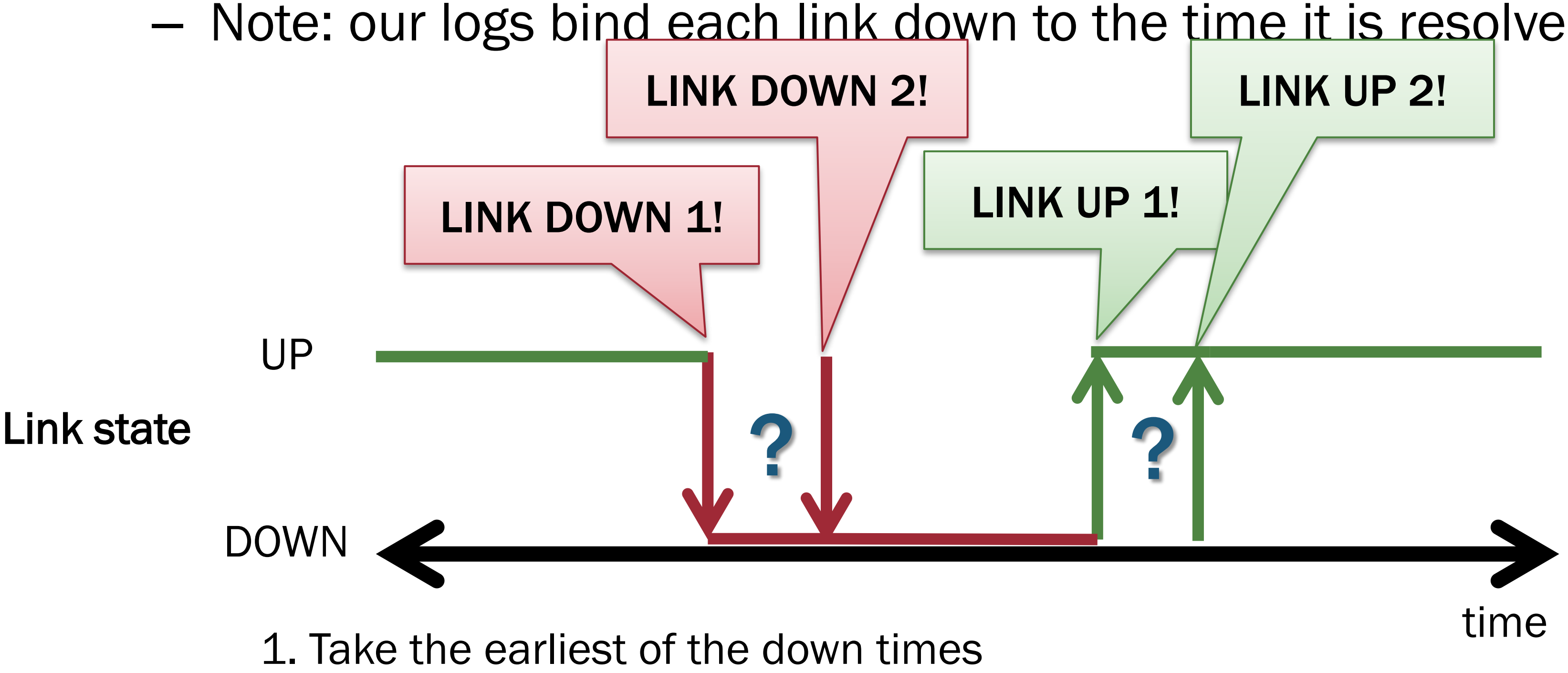
- Inconsistencies in link failure events
  - Note: our logs bind each link down to the time it is resolved



What we expect

Reconstructing link state

- Inconsistencies in link failure events
  - Note: our logs bind each link down to the time it is resolved

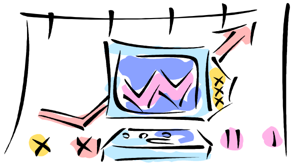


1. Take the earliest of the down times

2. Take the earliest of the up times

How to deal with discrepancies?

Identifying failures with impact

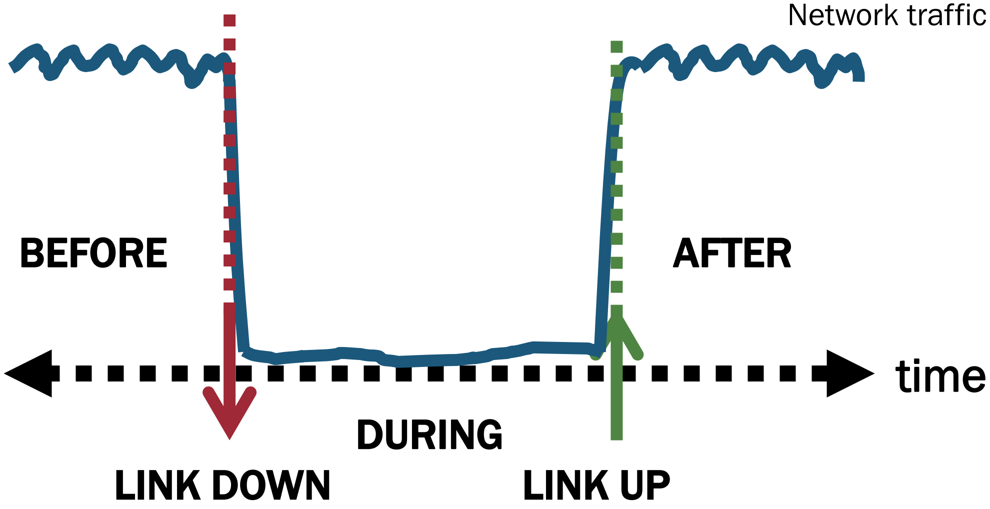


Network traffic logs

Correlate link failures with network traffic

Only consider events where traffic decreases

$$\frac{traffic\ during}{traffic\ before} < 1$$

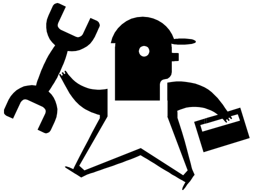


- **Summary of impact:**
  - 28.6% of failures impact network traffic
  - 41.2% of failures were on links carrying **no traffic**
    - E.g., scheduled maintenance activities
- **Caveat:** Impact is only on network traffic **not necessarily applications!**
  - Redundancy: Network, compute, storage mask outages

Road Map

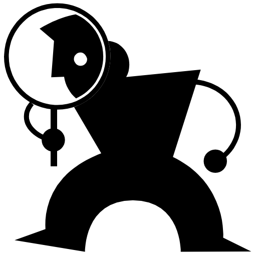
Motivation

Background & Methodology



Results

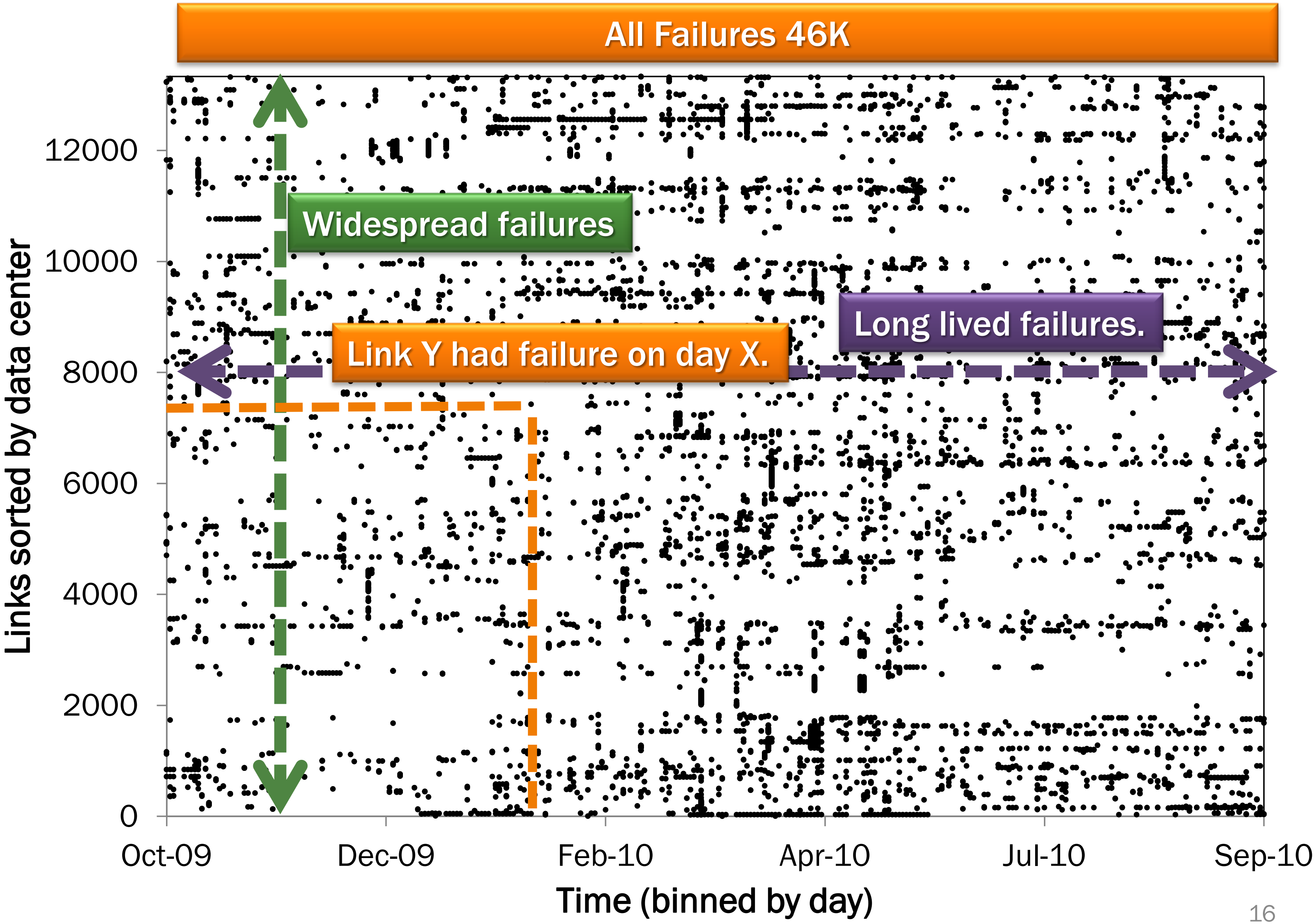
1. **Characterizing failures**
   - Distribution of failures over measurement period.
   - Which components fail most?
   - How long do failures take to mitigate?



2. Do current network redundancy strategies help?

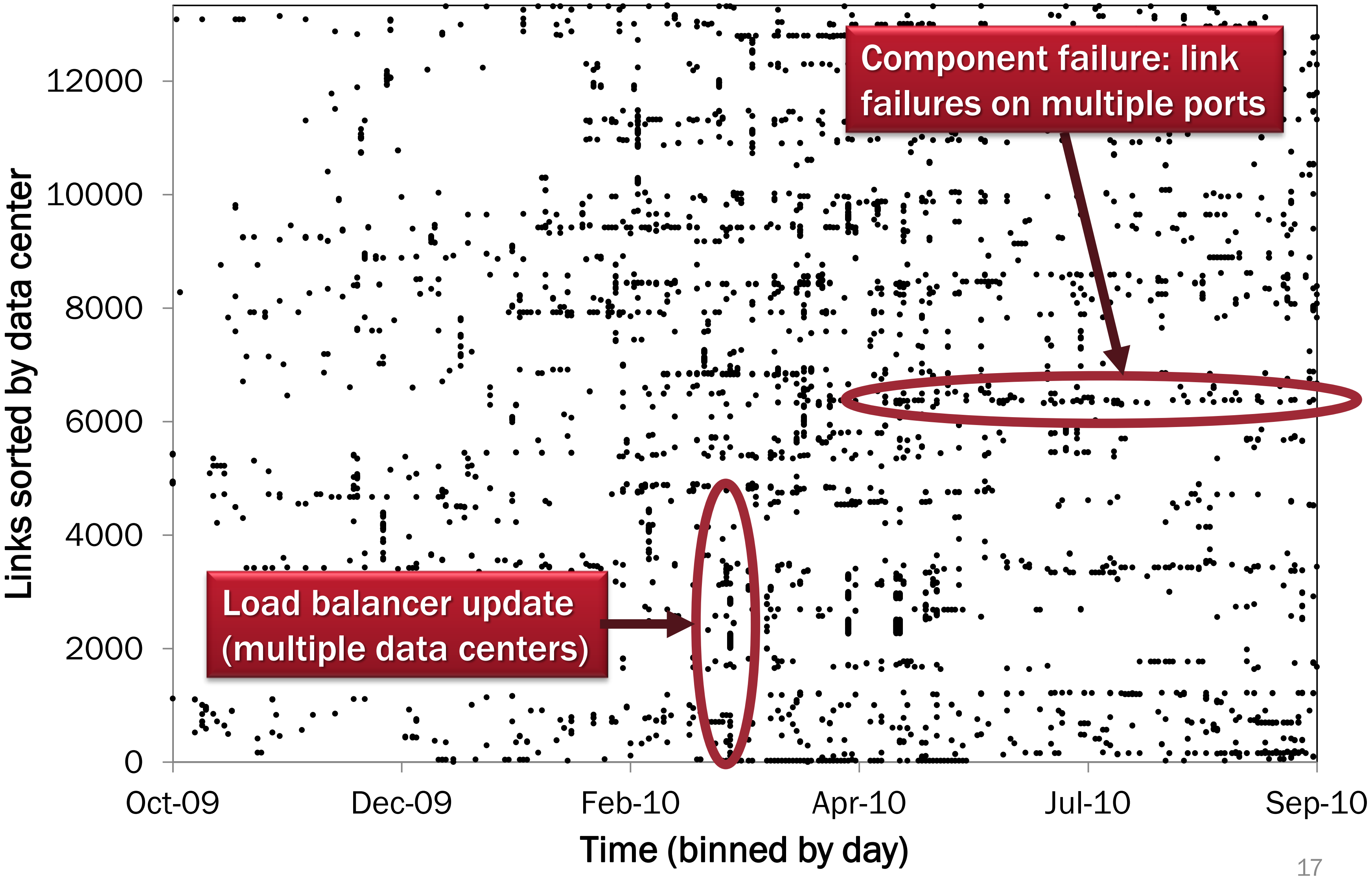
Conclusions

Visualization of failure panorama: Sep'09 to Sep'10



Visualization of failure panorama: Sep’09 to Sep’10

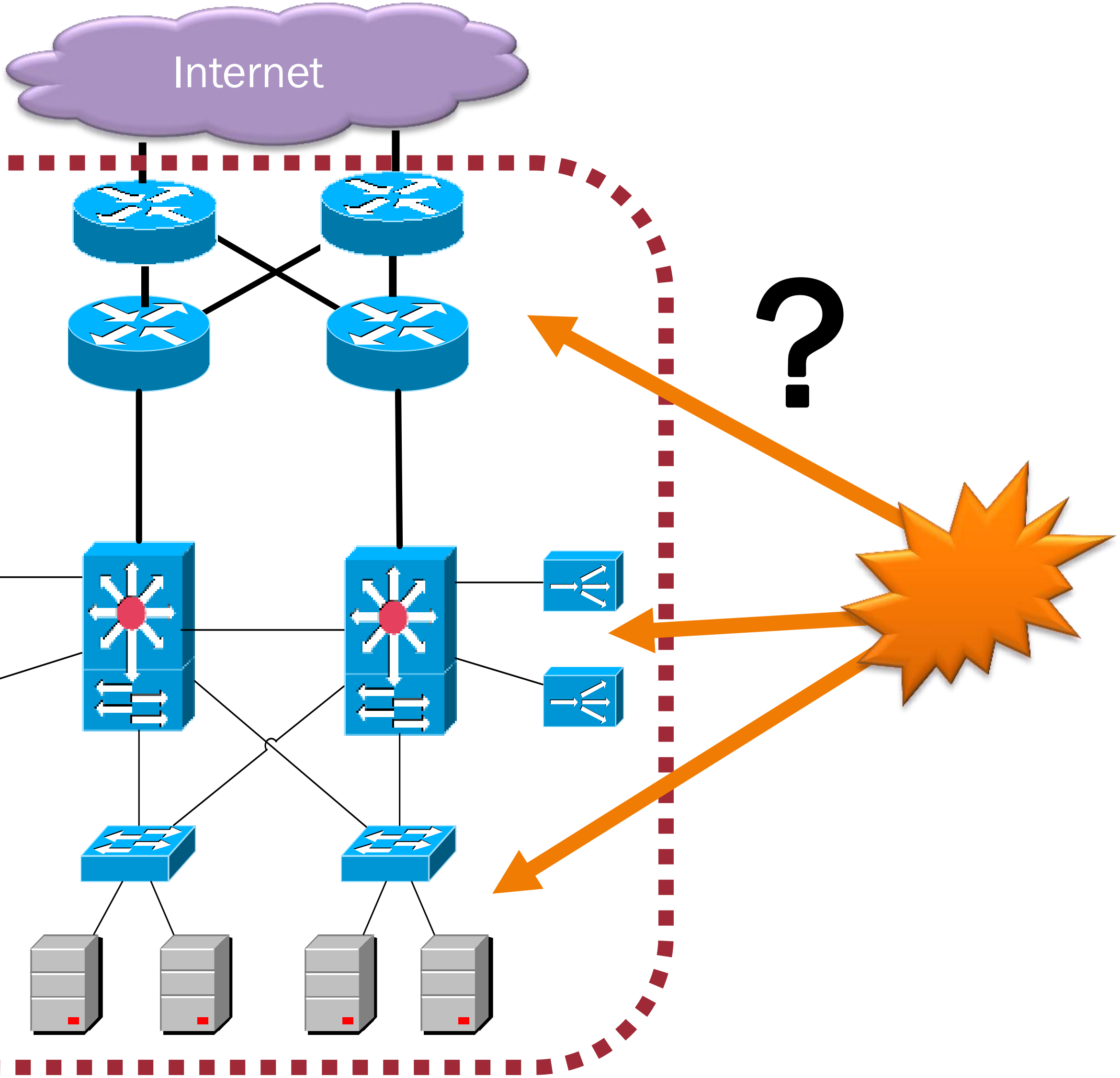
Failures with Impact 28%



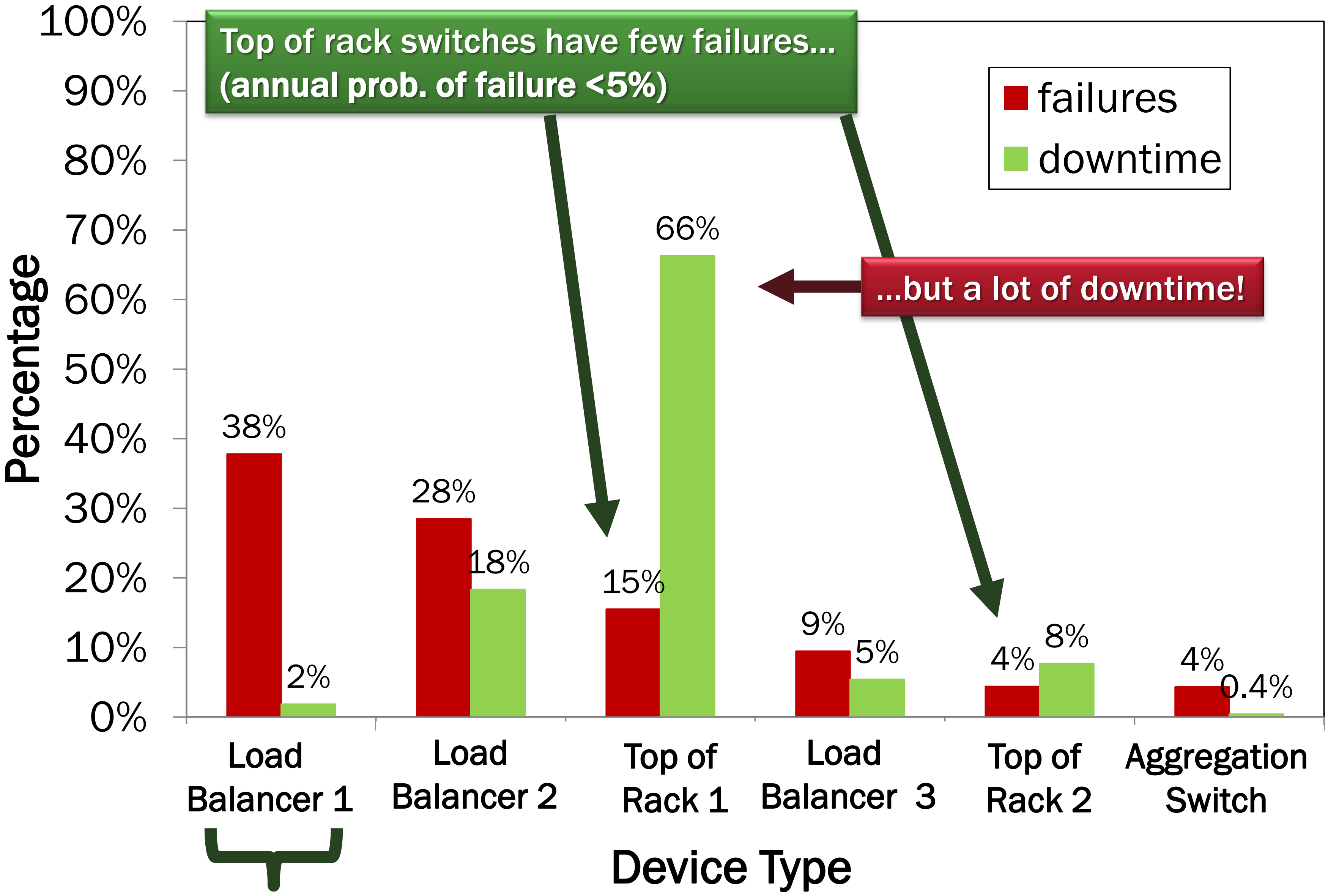
Component failure: link failures on multiple ports

Load balancer update (multiple data centers)

Which devices cause most failures?



Which devices cause most failures?

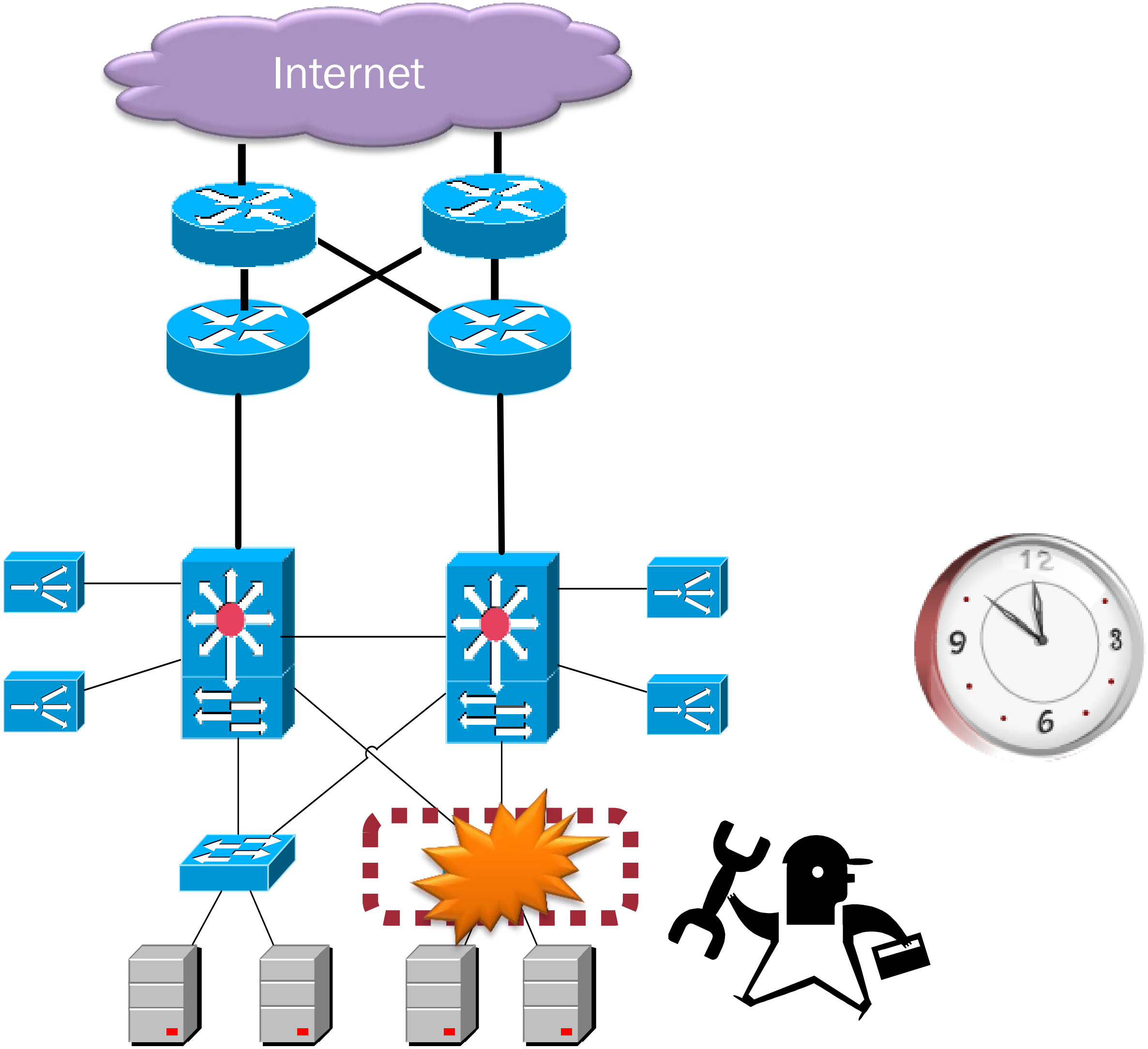


Top of rack switches have few failures… (annual prob. of failure <5%)

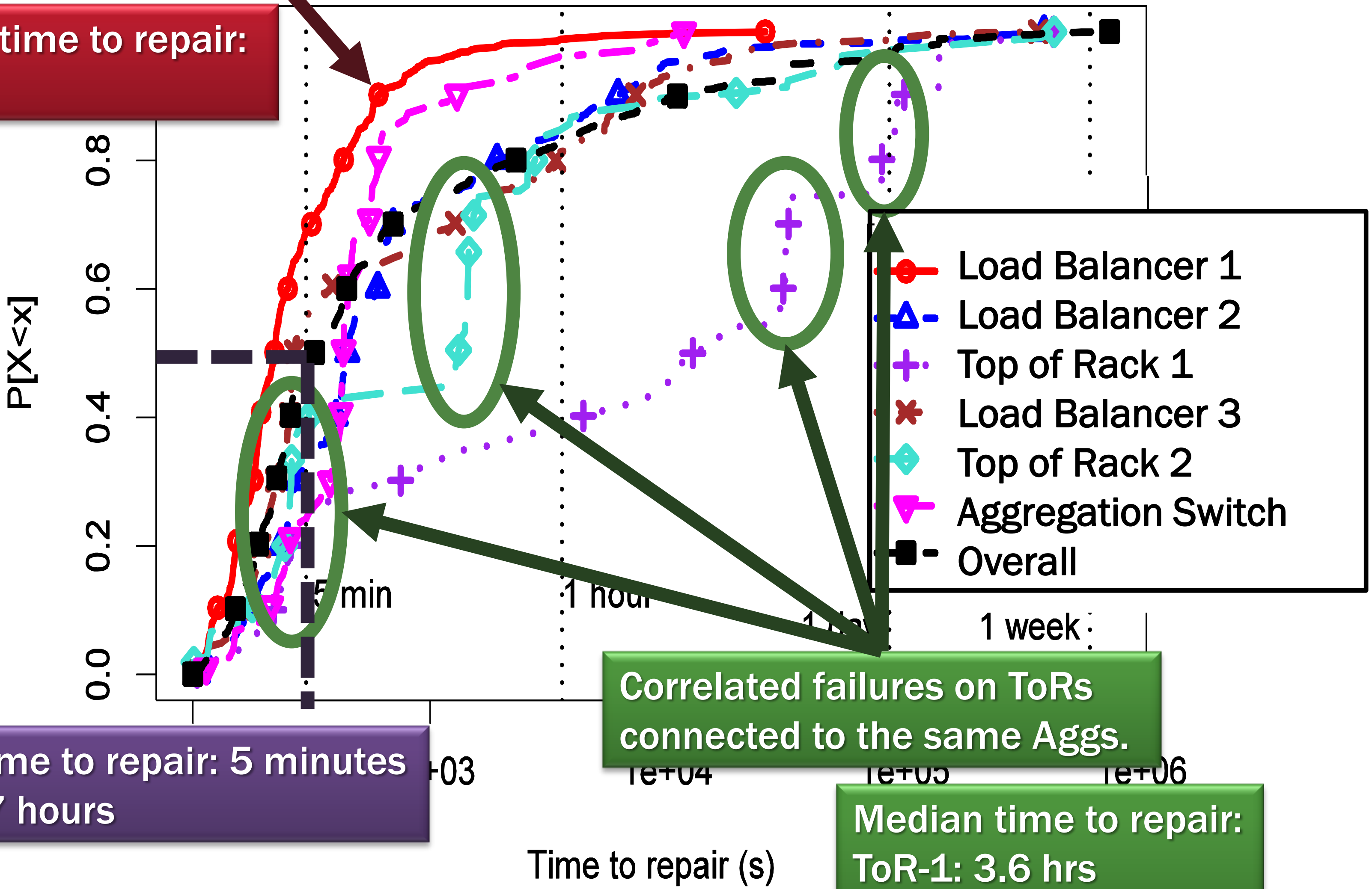
…but a lot of downtime!

Load balancer 1: very little downtime relative to number of failures.

How long do failures take to resolve?



How long do failures take to resolve?



Load balancer 1: short-lived *transient* faults

Median time to repair:
4 mins

Median time to repair: 5 minutes
Mean: 2.7 hours

Correlated failures on ToRs connected to the same Aggs.

Median time to repair:
ToR-1: 3.6 hrs
ToR-2: 22 min

Summary

- Data center networks are highly reliable
  - Majority of components have four 9's of reliability

- Low-cost top of rack switches have highest reliability
  - <5% probability of failure
- ...but most downtime
  - Because they are lower priority component

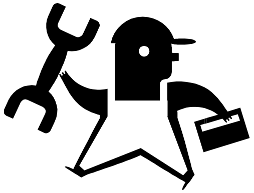
- Load balancers experience many short lived faults
  - Root cause: software bugs, configuration errors and hardware faults

- Software and hardware faults dominate failures
  - ...but hardware faults contribute most downtime

Road Map

Motivation

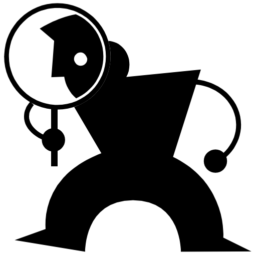
Background & Methodology



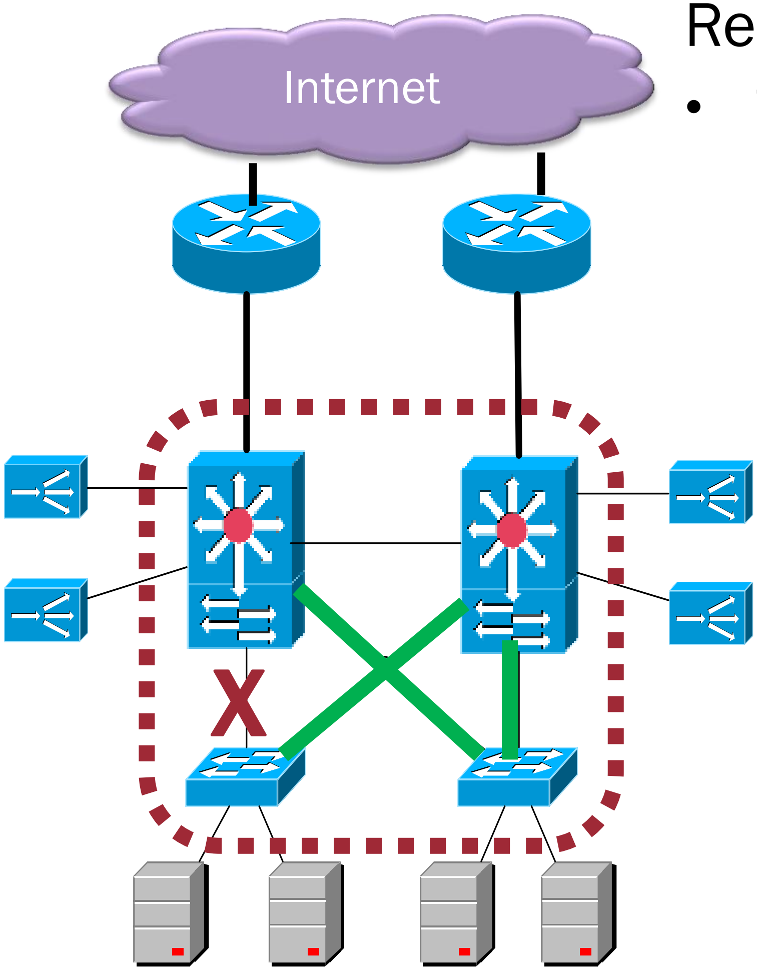
Results

1. Characterizing failures
2. **Do current network redundancy strategies help?**

Conclusions



Is redundancy effective in reducing impact?



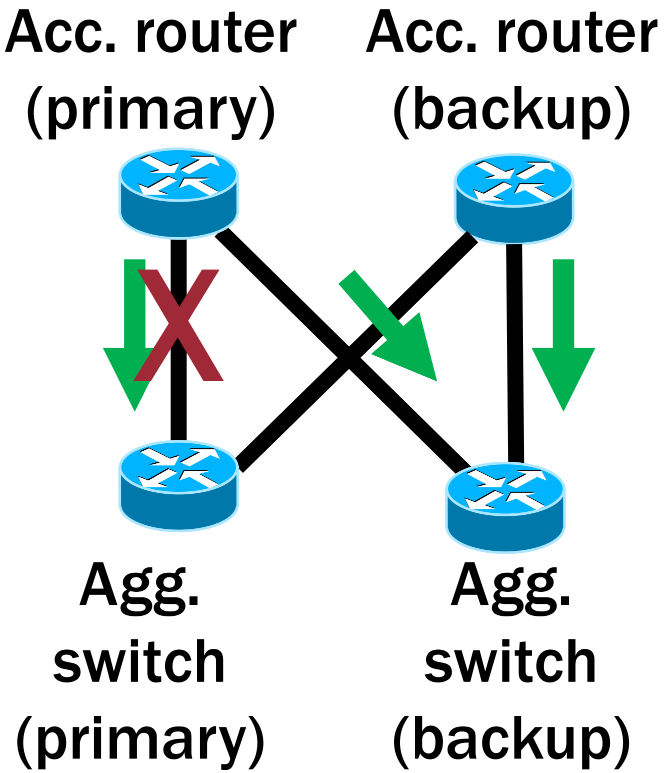
Redundant devices/links to mask failures

- This is expensive! (management overhead + $$$)

Goal: Reroute traffic along available paths

How effective is this in practice?

Measuring the effectiveness of redundancy



Idea: compare traffic before and during failure

Measure traffic on links:

1. Before failure
2. During failure
3. Compute "normalized traffic" ratio:

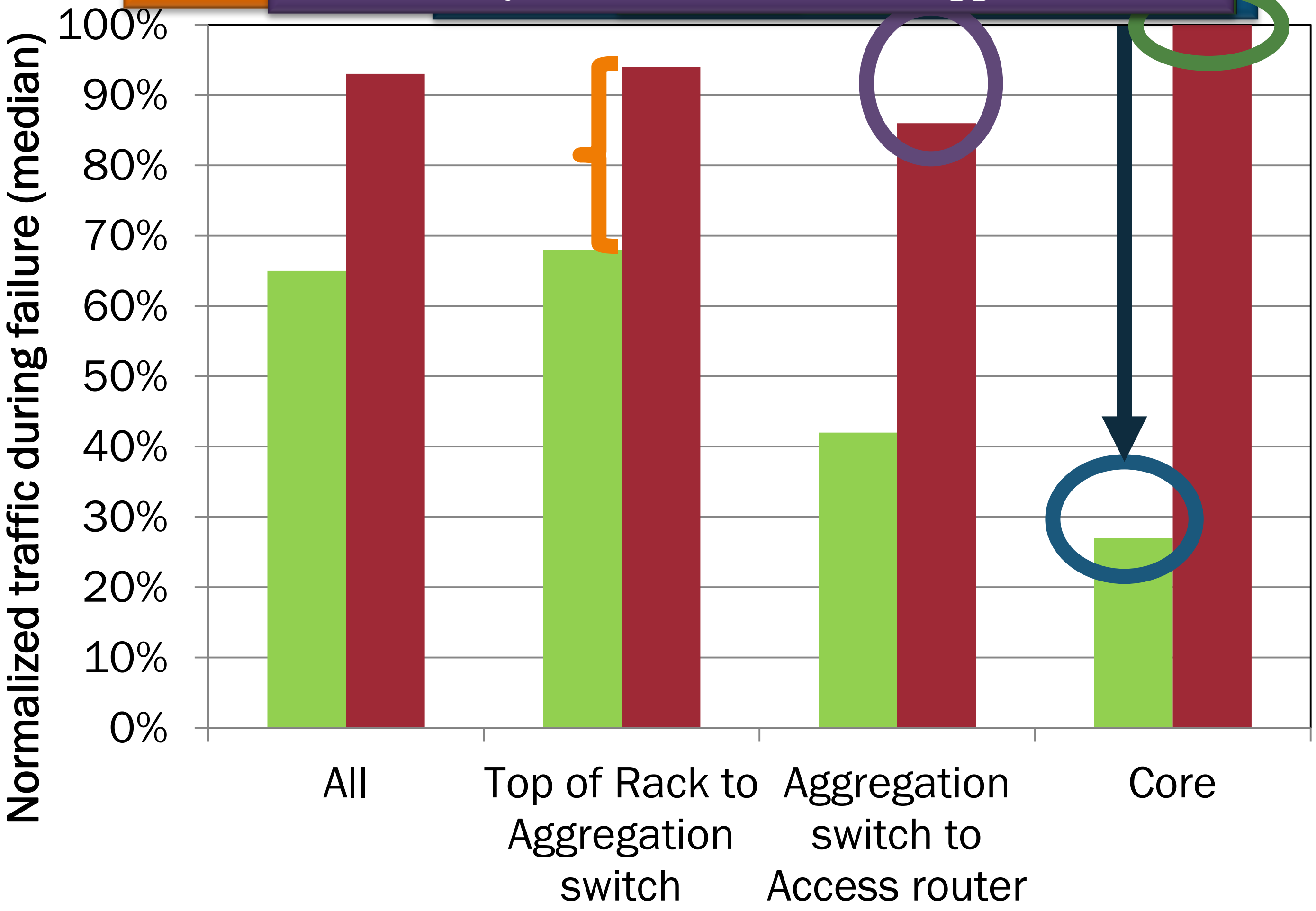
$$\frac{traffic\ during}{traffic\ before} \sim 1$$

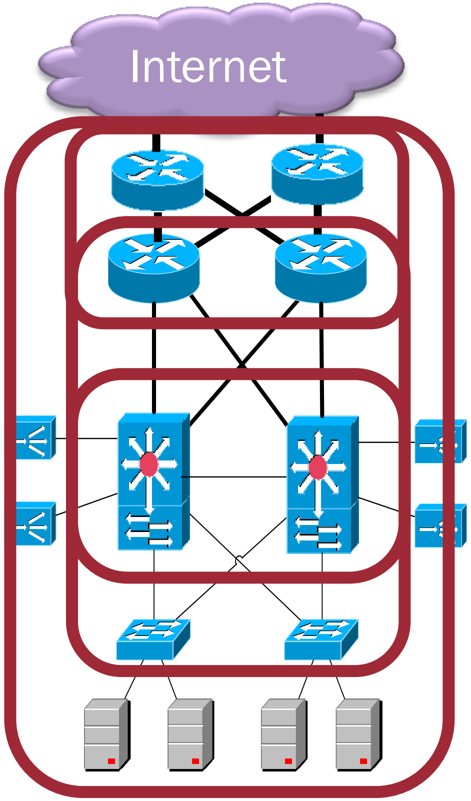
Compare normalized traffic over redundancy groups to normalized traffic on the link that failed

Is redundancy effective in reducing impact?

Less ir

Redundancy is least effective for AggS and AccR



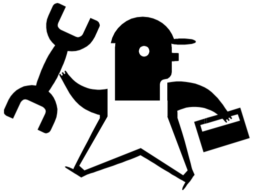


Overall increase of 40% in terms of traffic due to redundancy

Road Map

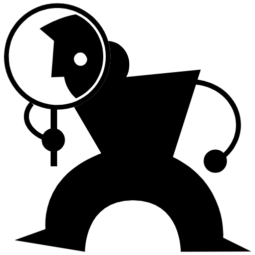
Motivation

Background & Methodology



Results

1. Characterizing failures
2. Do current network redundancy strategies help?



Conclusions

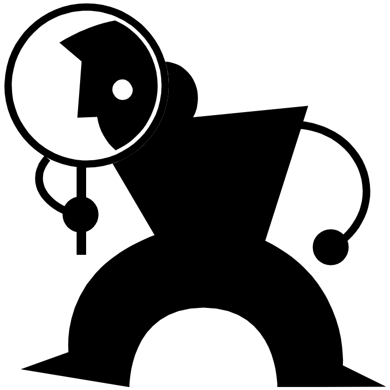
Conclusions

- **Goal: Understand failures in data center networks**
  - Empirical study of data center failures

- **Key observations:**
  - Data center networks have high reliability
  - Low-cost switches exhibit high reliability
  - Load balancers are subject to transient faults
  - Failures may lead to loss of small packets

- **Future directions:**
  - Study application level failures and their causes
  - Further study of redundancy effectiveness

Thanks!



Contact: phillipa@cs.toronto.edu

Project page:

http://research.microsoft.com/~navendu/netwiser