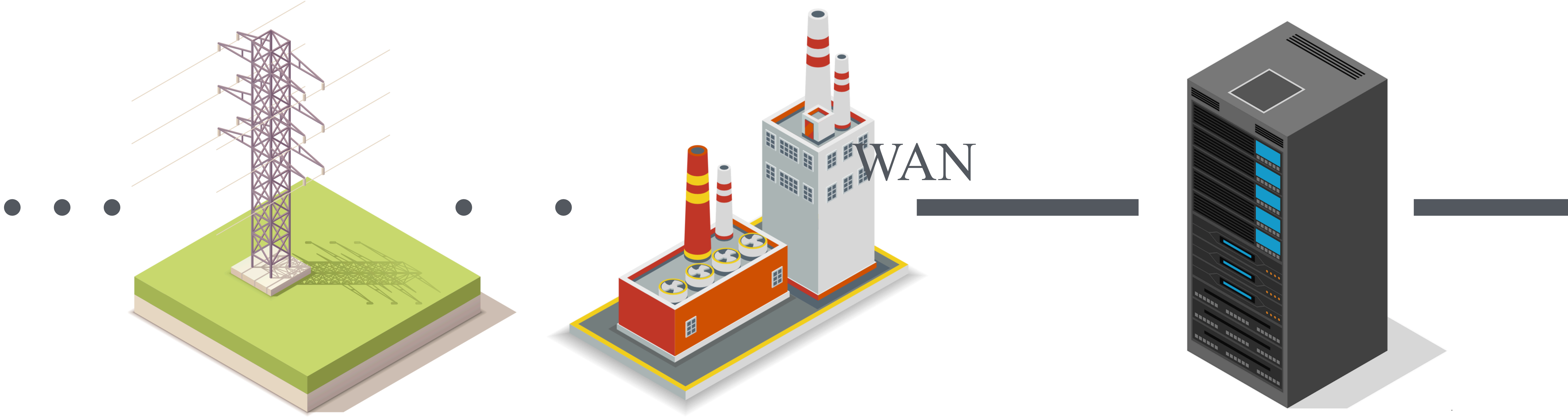
A Large Scale Study of Data Center Network Reliability

Justin Meza
Carnegie Mellon University
Facebook, Inc.
jjm@fb.com

Tianyin Xu
University of Illinois
Urbana-Champaign
Facebook, Inc.
tyxu@illinois.edu

Kaushik Veeraraghavan
Facebook, Inc.
kaushikv@fb.com

Onur Mutlu
ETH Zurich
Carnegie Mellon University
onur.mutlu@inf.ethz.ch

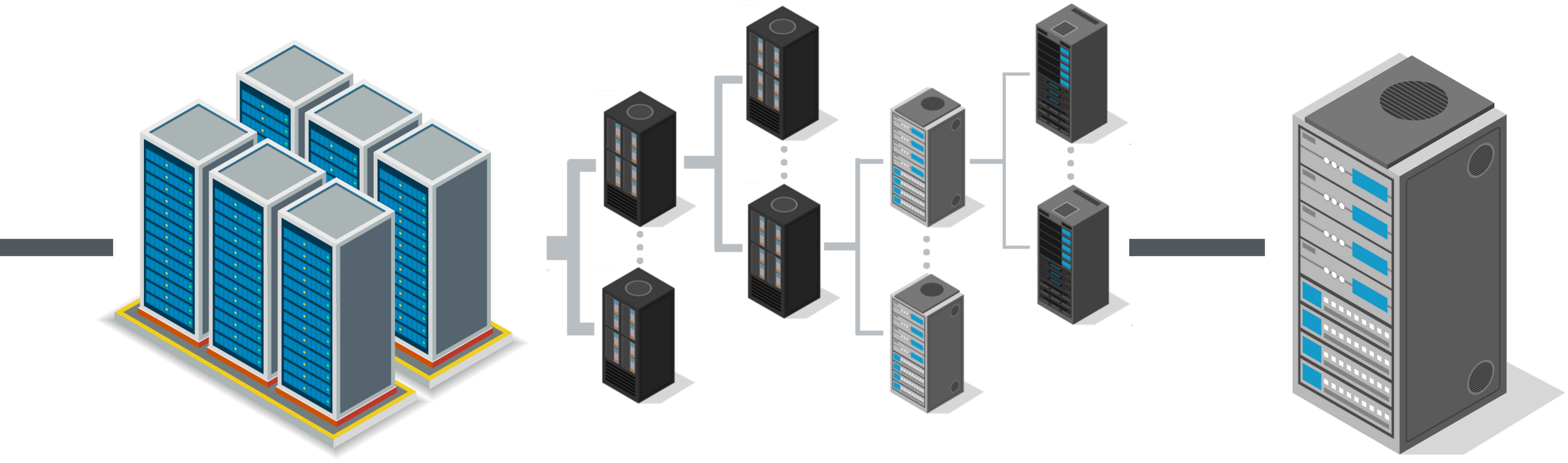


WAN

Internet

ISP

Edge Node

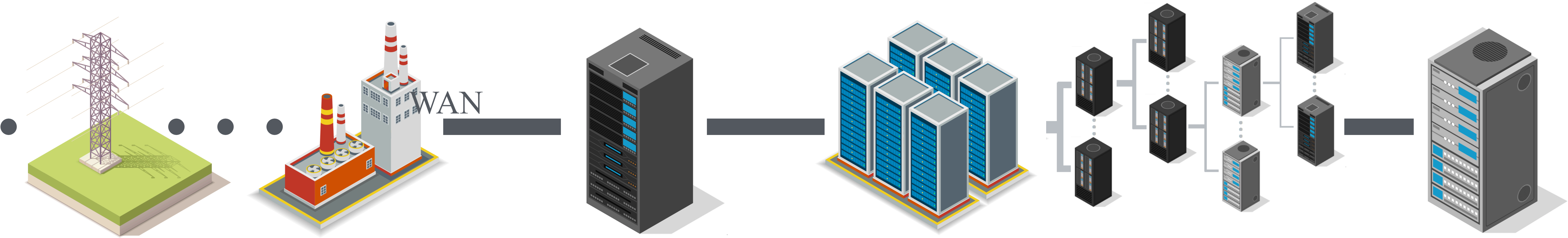


Core Switches

Data Center Fabric

Top of Rack Switch

Software managed



WAN

Internet

ISP

Edge Node

Core Switches

Data Center Fabric

Rack Switch

Problem

- **Network incidents** → *major root* cause of DC **outages**.
- Little research on **reliability** characteristics of **large scale DC** network infrastructure and the *impact* on **software systems**.
- Difficulty lies in **correlating** *device-* and *link-level* failure with software system impact.

Goal

- Cover reliability characteristics of both **intra** and **inter** data center networks.

Key takeaways

- DC → more **software managed**
  - next challenge: make the first and last hop more reliable
- **Backbone network reliability planning**
  - more important than ever for ensuring good overall site reliability.

Outline

- **Introduction to data center networks**
- Intra data center networks
- Inter data center networks
- Concluding thoughts

Terminology

Network Incidents
cause
Software Failures
that result in
Site Events (SEVs)

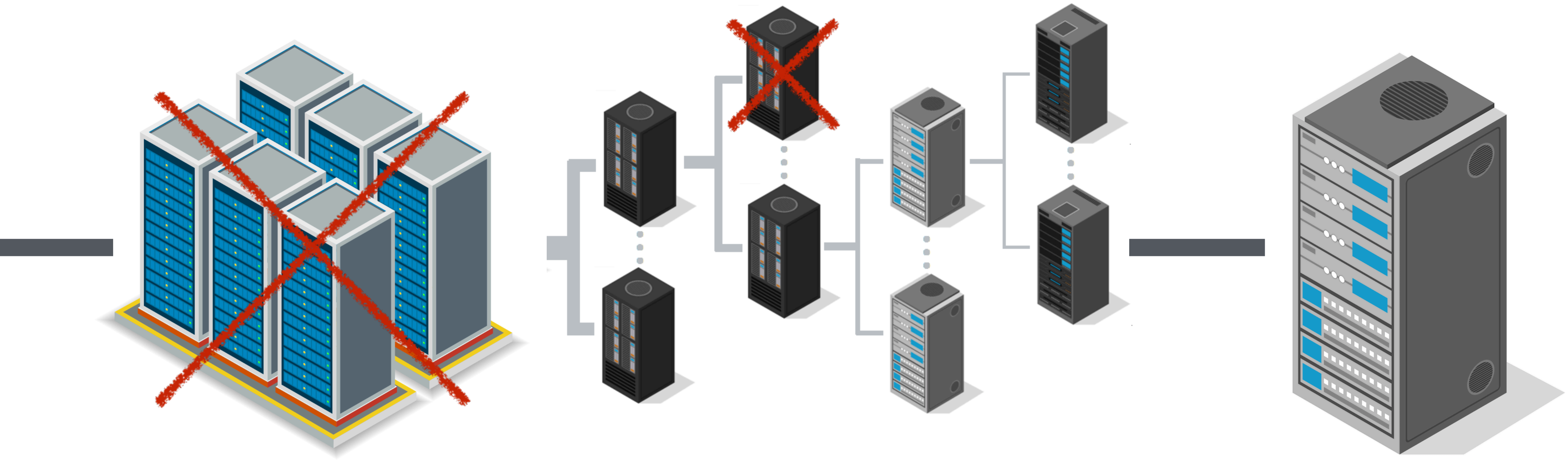
- SEVs classified into 3 severity categories
- Engineers write the reports
- Report contain:
  - Incident’s root cause
  - Root cause’s effect on software systems
  - Steps to prevent the incident from happening again
- Network SEV report contain details about:
  - Network device implicated in the incident
  - Duration of the incident
  - Incident’s effect on software systems

Methodology

Data

Intra data center reliability:
7 years of service level event data collected from SEV database

- Root cause
  - May be undetermined
- Device type
  - Used to classify a network incident by the implicated device's type
- Network design
  - Classify a network incident based on network architecture



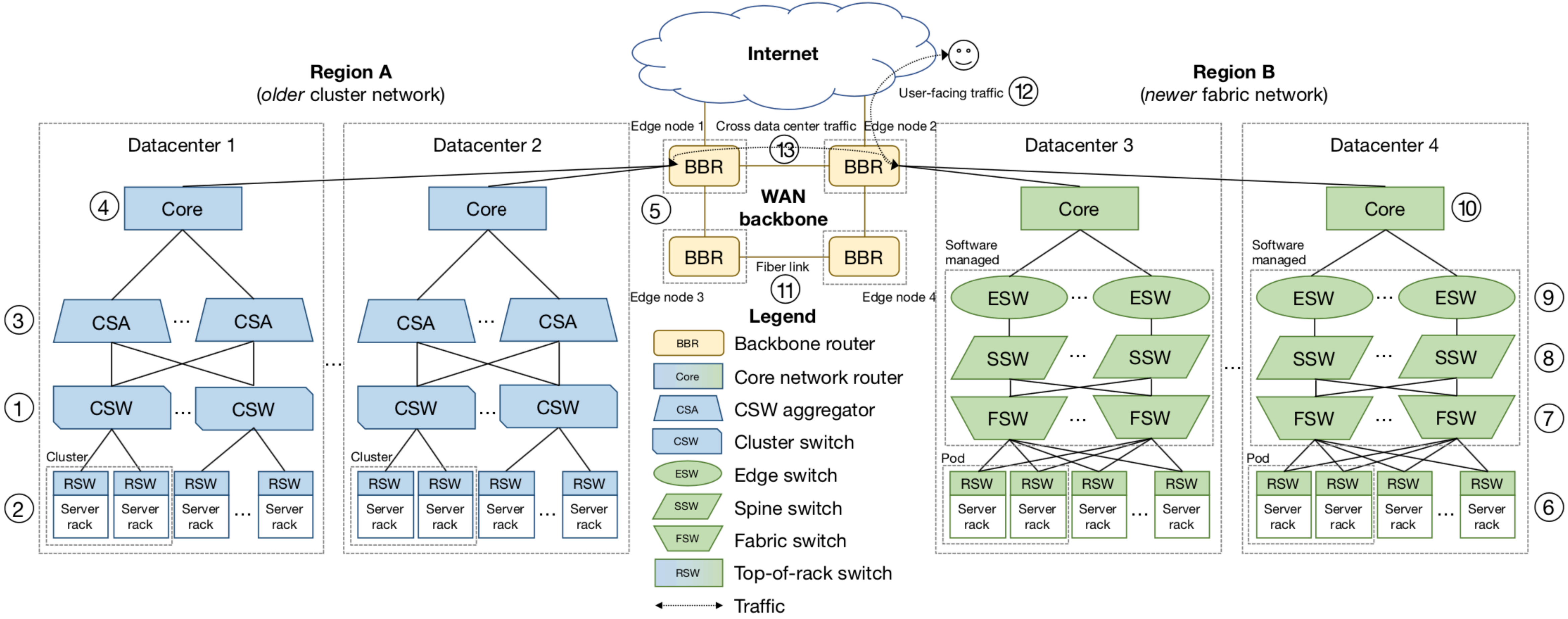
Core Switches

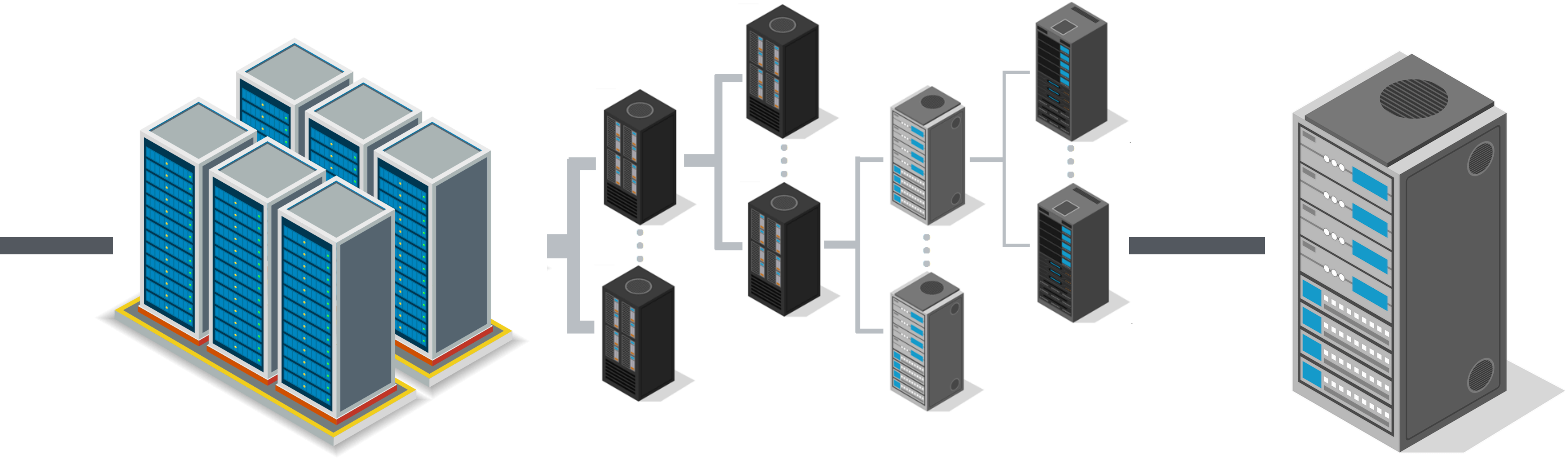
Data Center Fabric

Top of Rack Switch

Background

Facebook’s network architecture





Core Switches

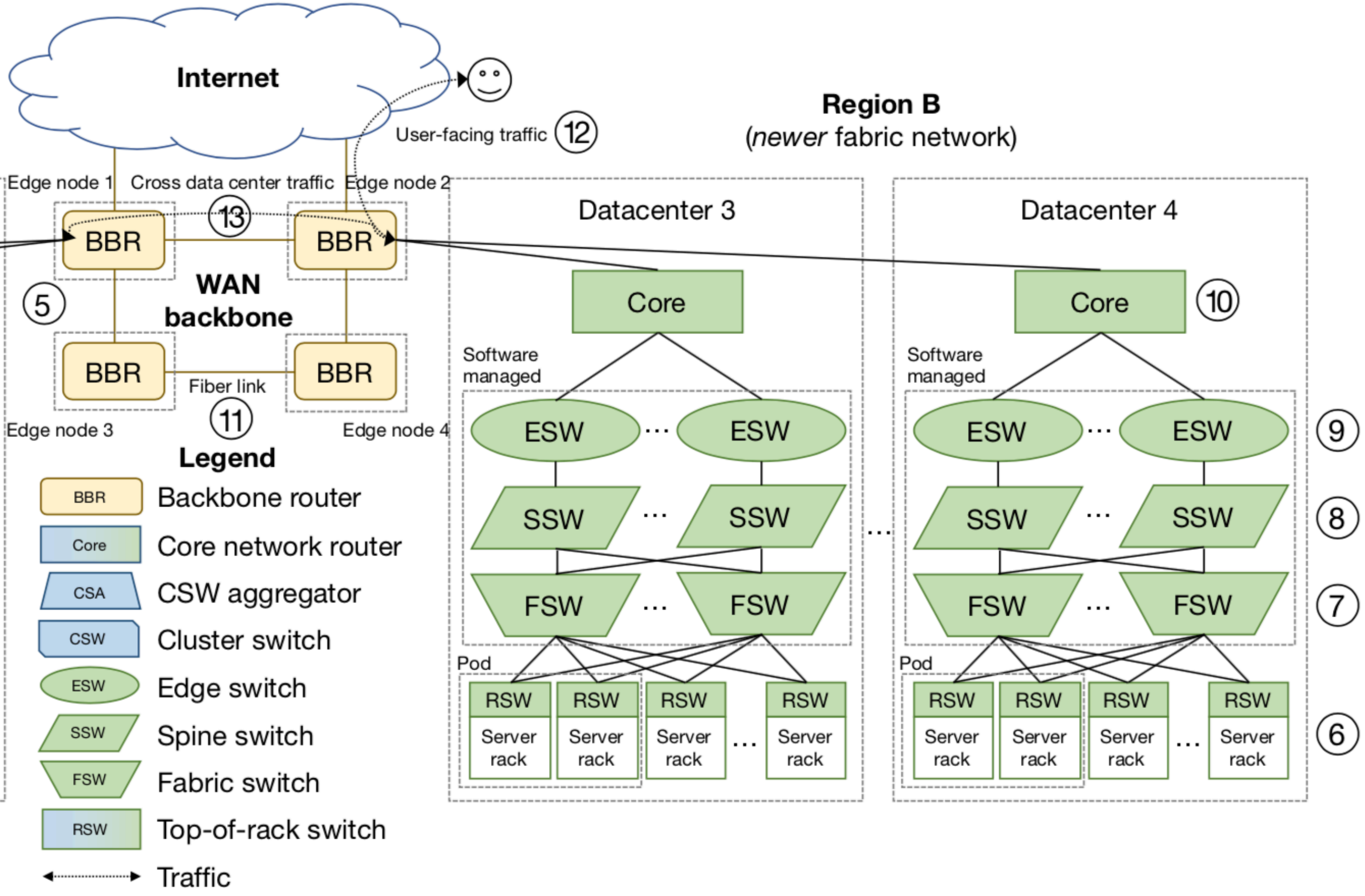
Data Center Fabric

Top of Rack Switch

Software managed

Background

Facebook’s network architecture

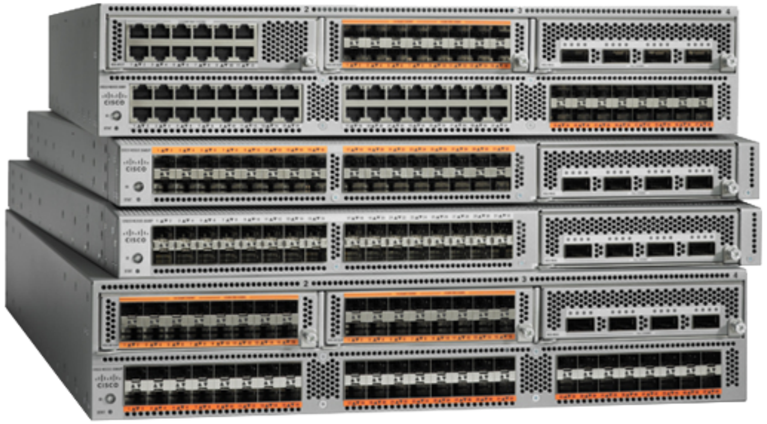


Outline

- Introduction to data center networks
- **Intra data center networks**
- Inter data center networks
- Concluding thoughts

Data center trends

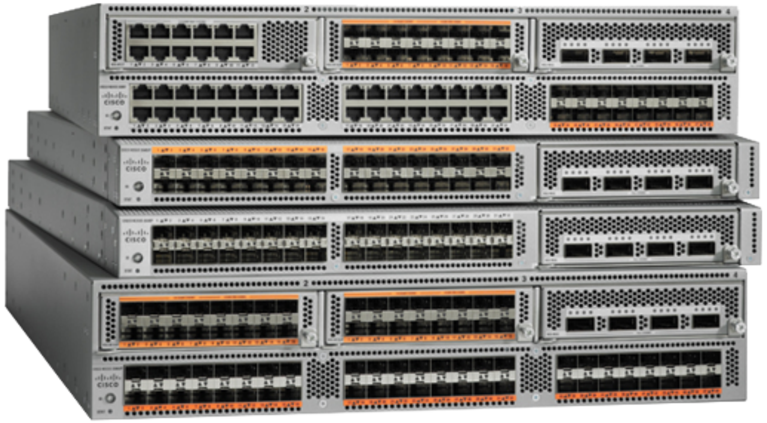
- Simple, custom switches
- Software-based fabric networks
- Automated repairs



Picture: www.cisco.com

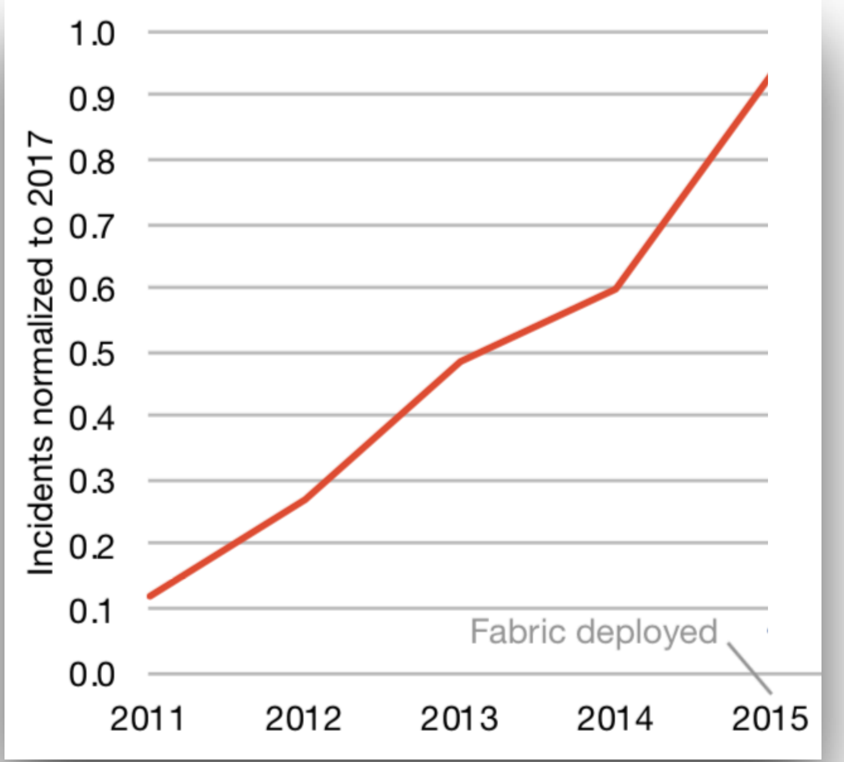
Data center trends

- Simple, custom switches
- Software-based fabric networks
- Automated repairs

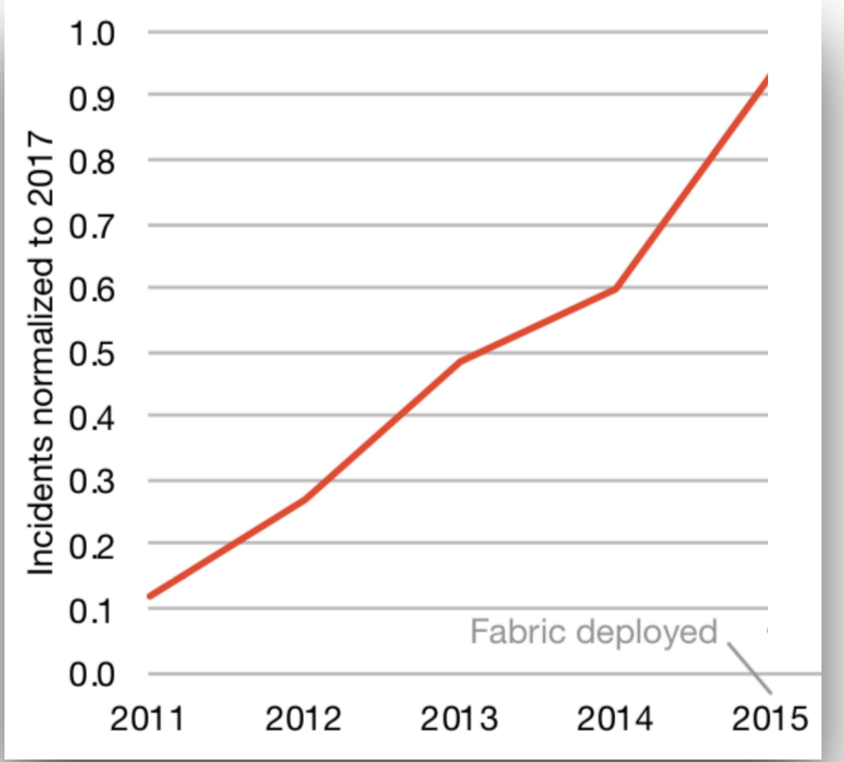


Picture: www.cisco.com

Older cluster based networks

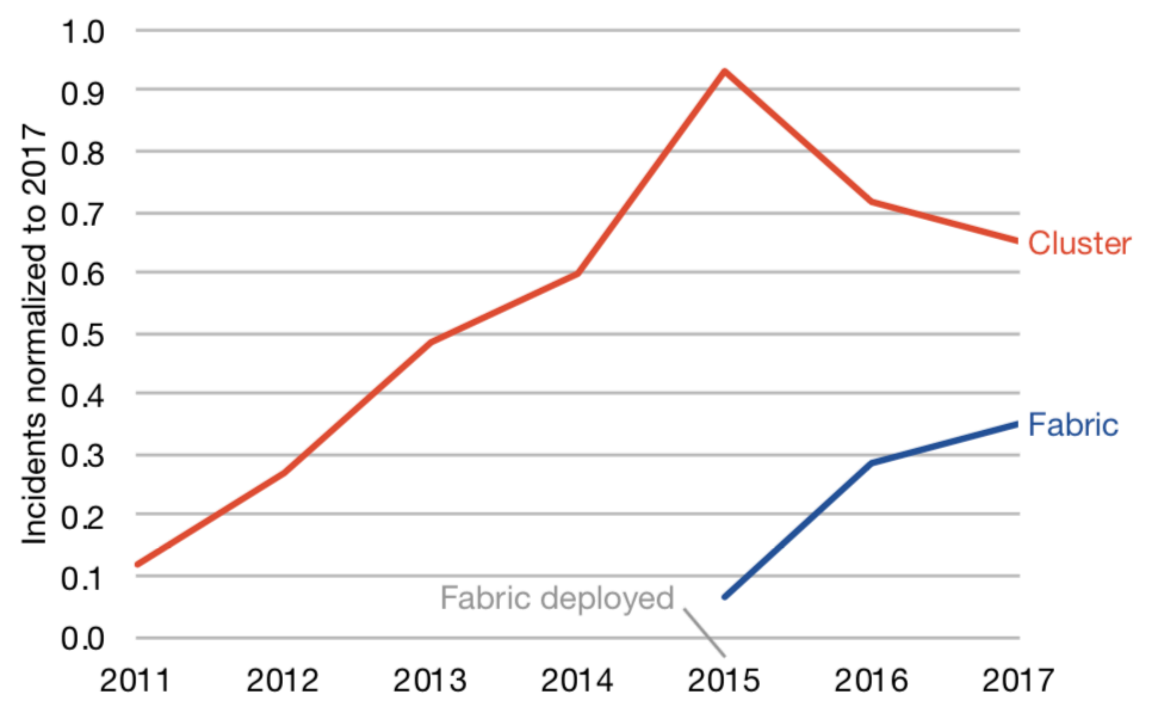


Older cluster based networks

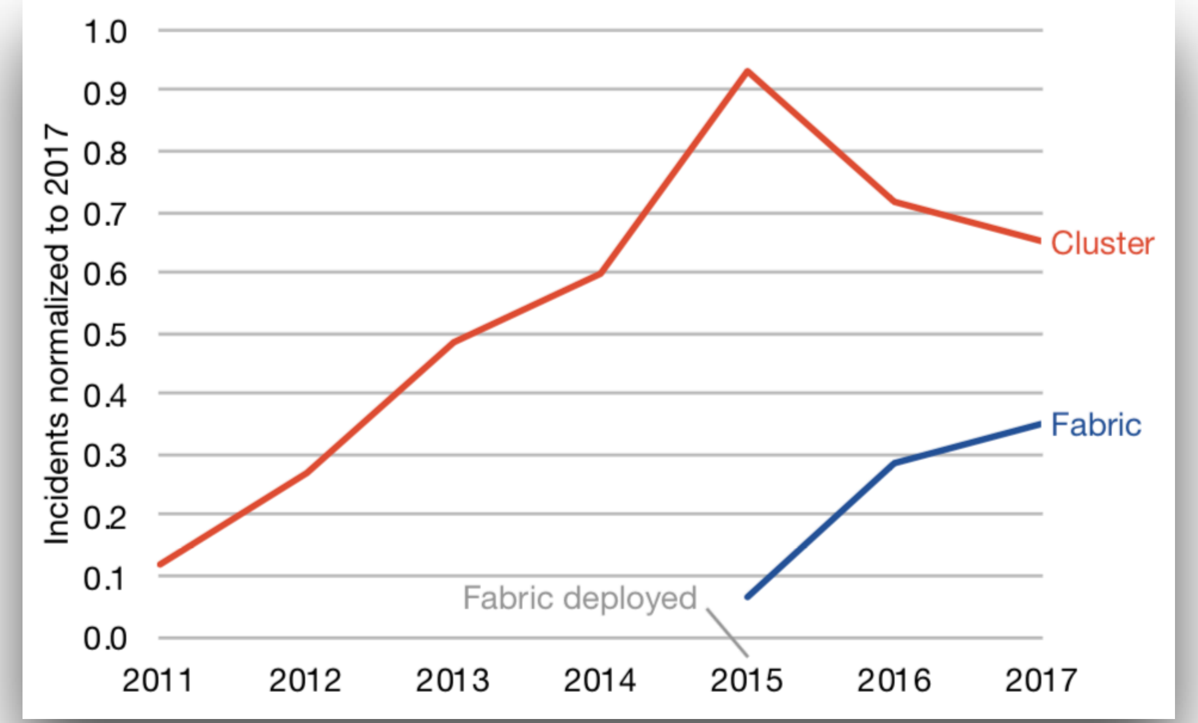


Cluster network incidents increased 9x over 4 years

Cluster versus Fabric Design

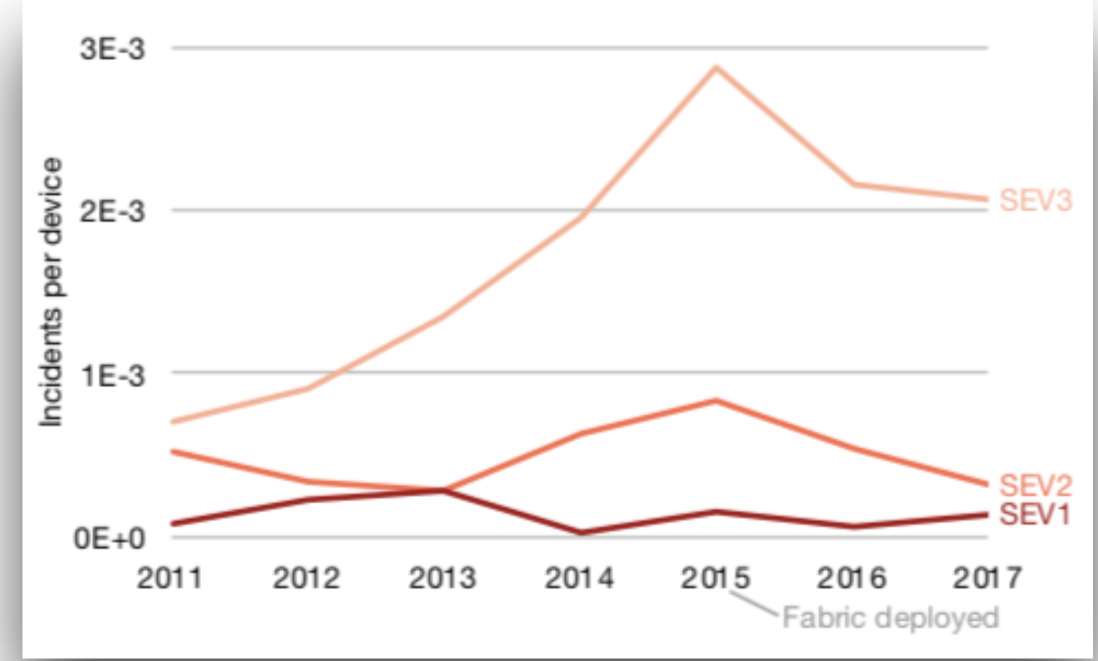


Cluster versus Fabric Design



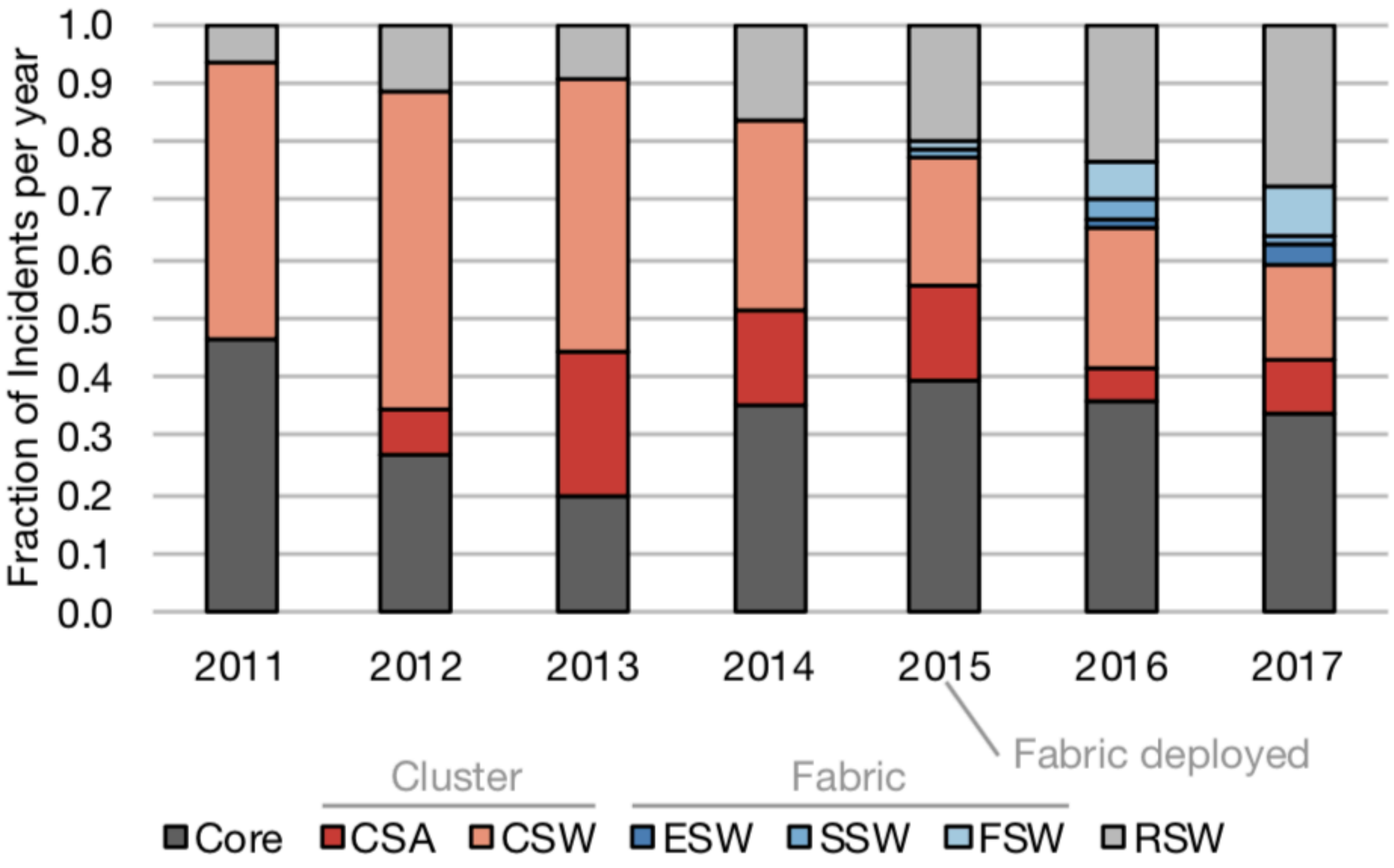
Cluster have 2x total incidents & 2.8x on a per-device level compared to fabric design

Data center fabric design has fewer incidents

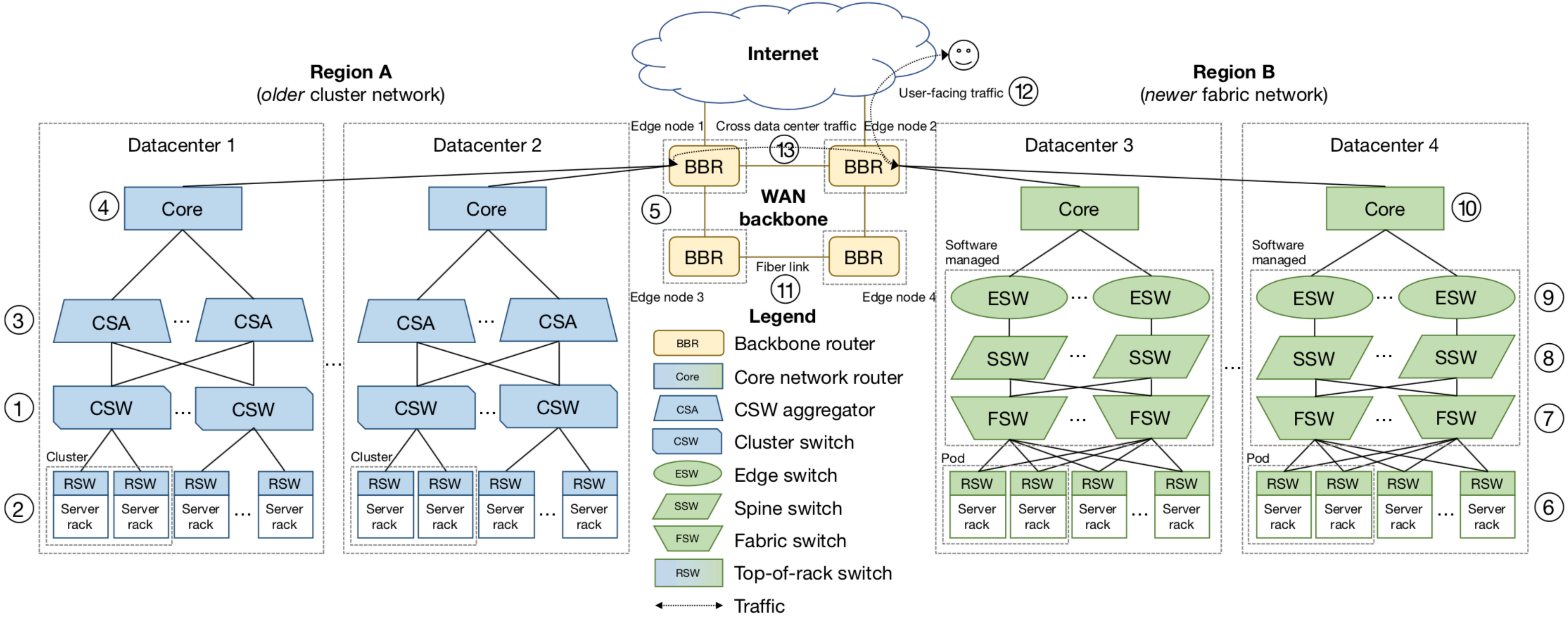


Reversing the negative software-level reliability trend

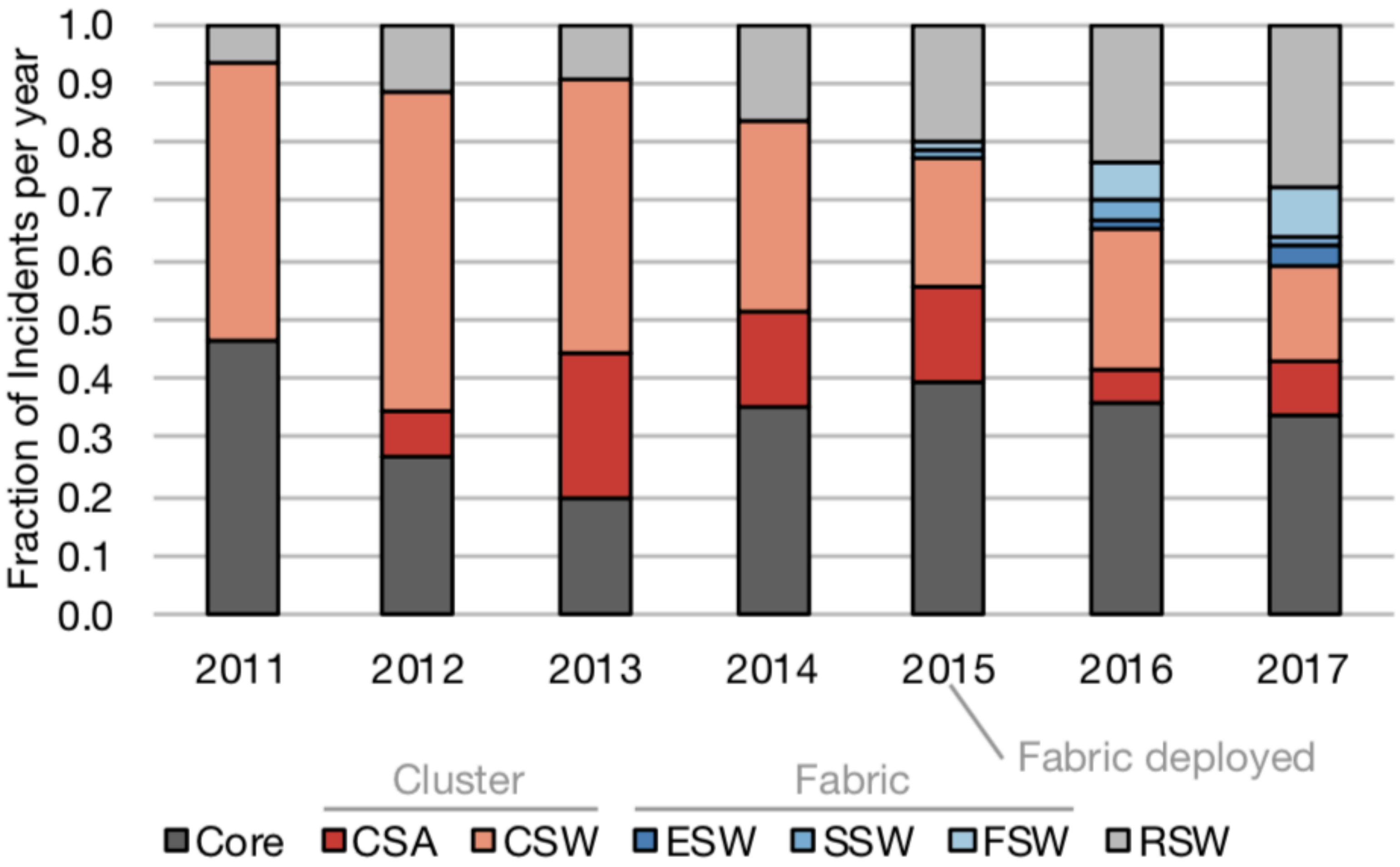
First and last hop reliability



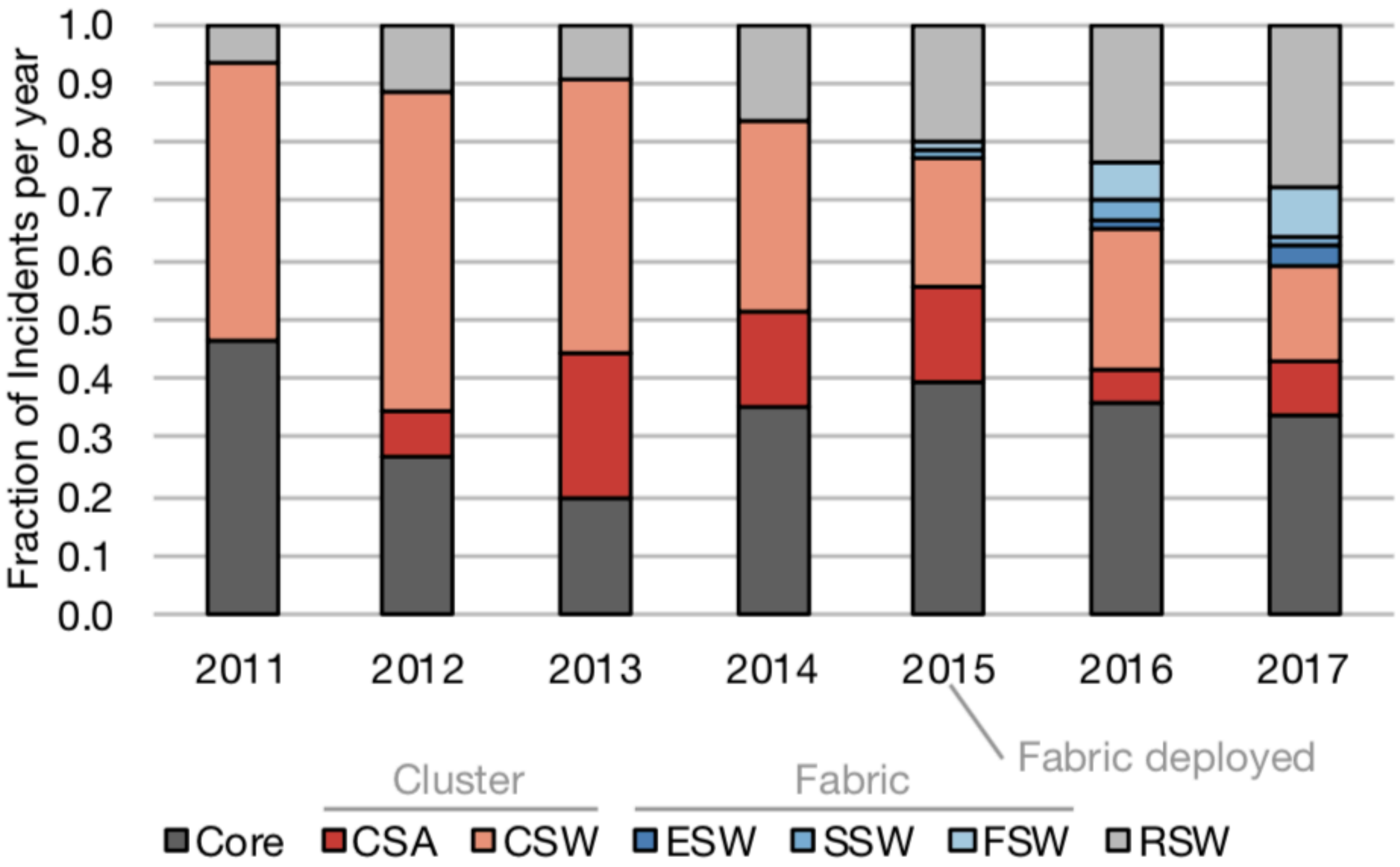
Background
Facebook’s network architecture



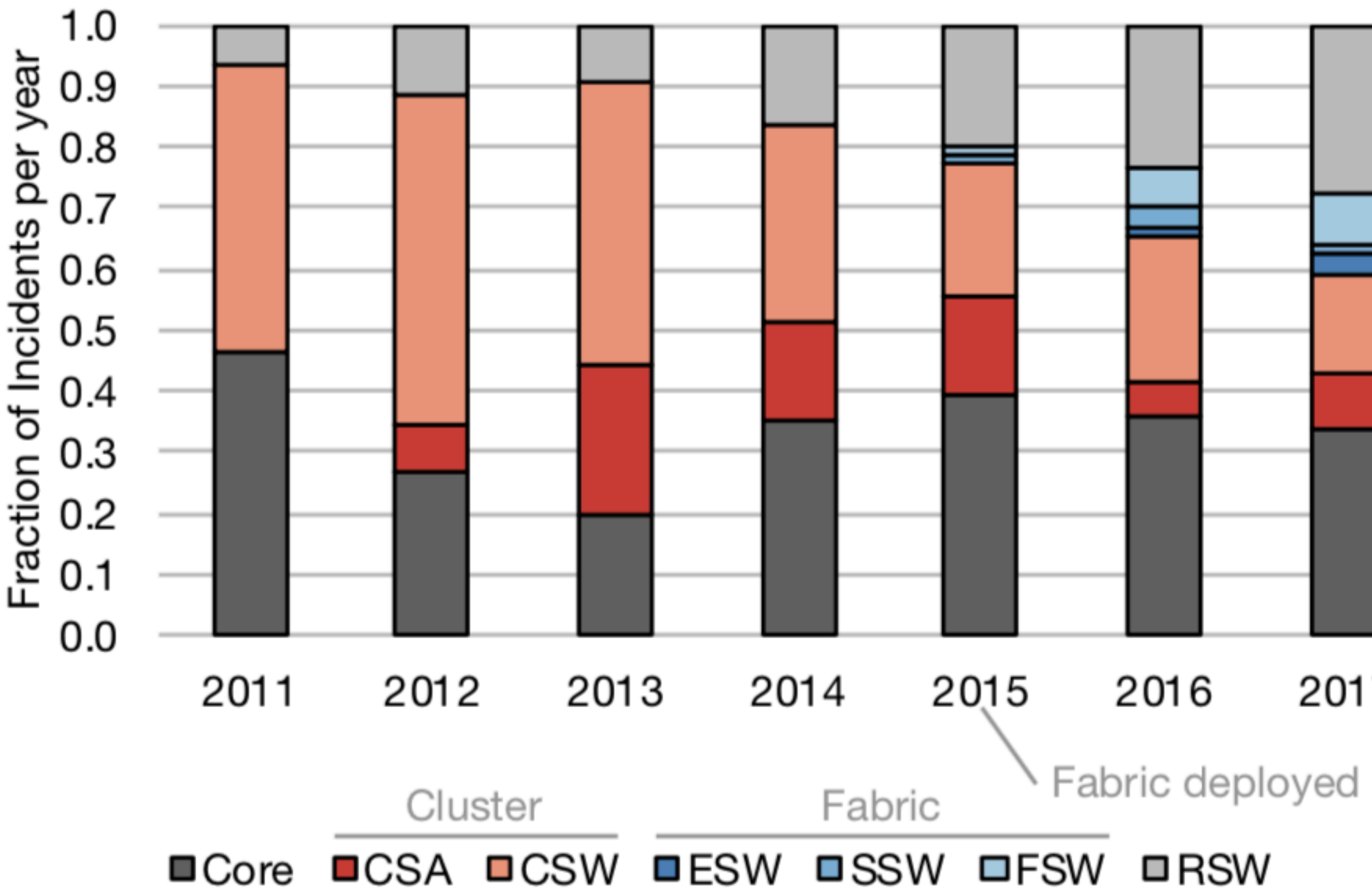
First and last hop reliability



First and last hop reliability

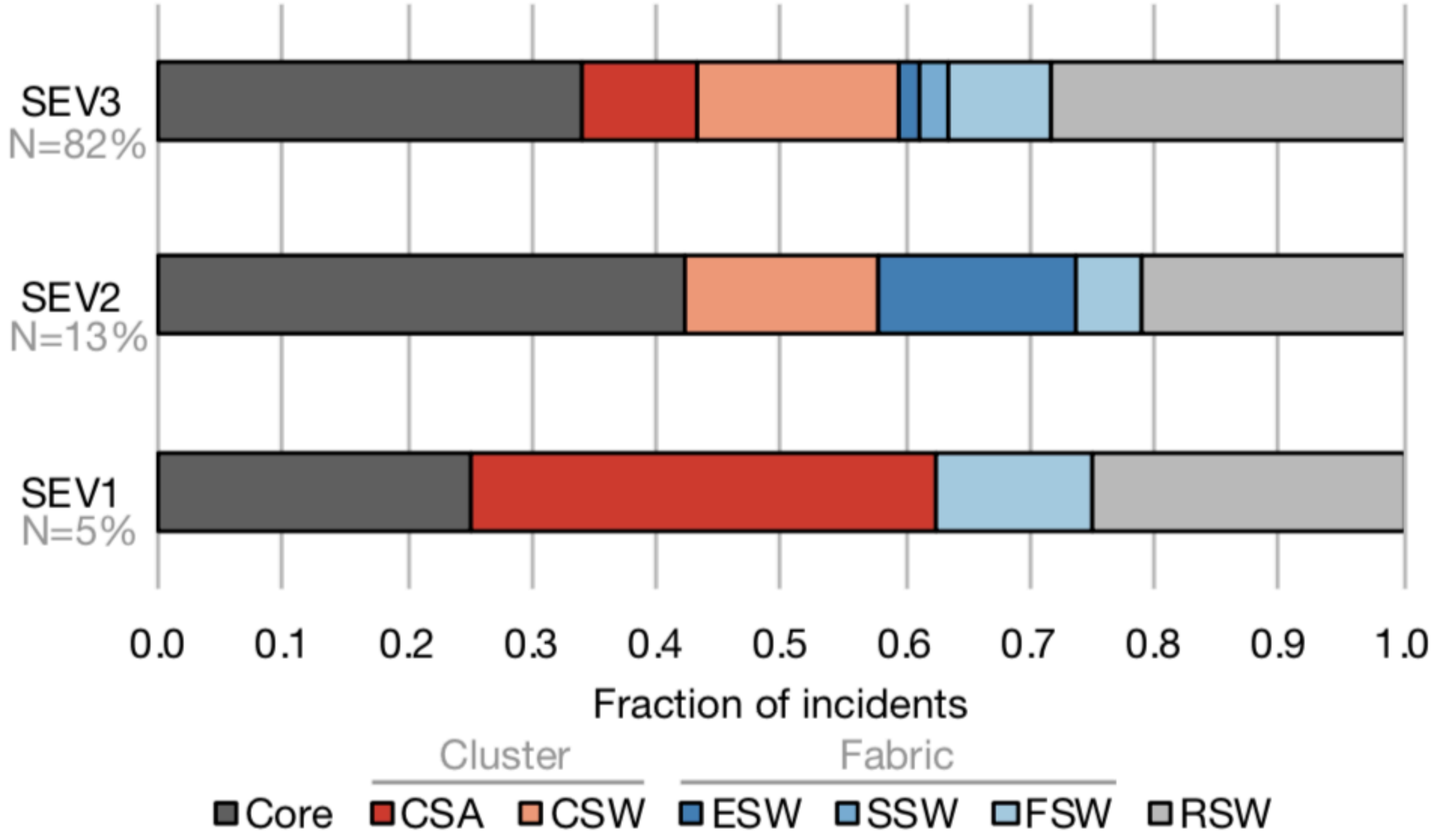


First and last hop reliability



Rack switches make up 82% of network devices

Main cause across all severities



Implications for data center networks

- More redundant switches one approach

Implications for data center networks

- More redundant switches one approach
- Make software more resilient

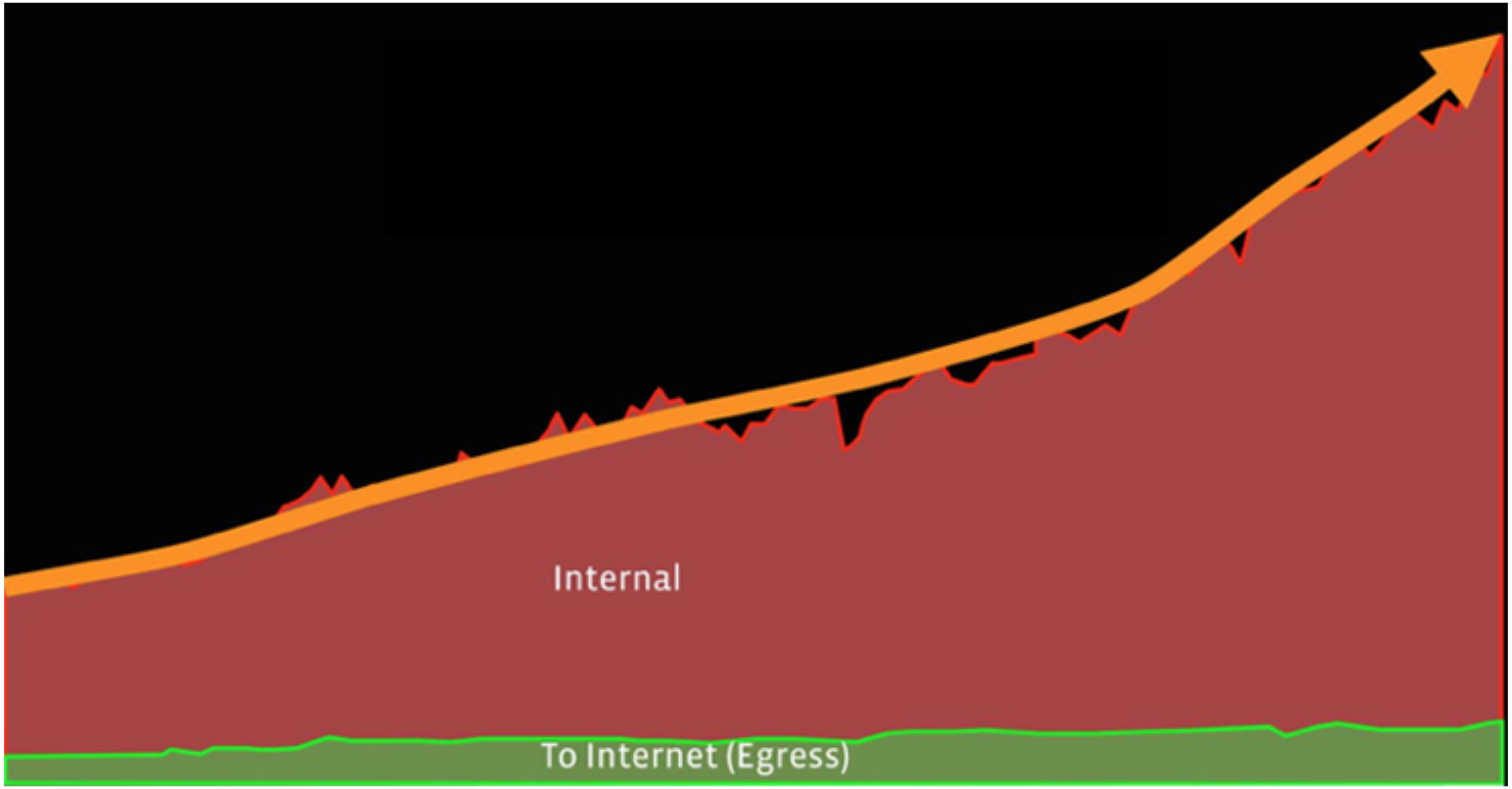
Implications for data center networks

- More redundant switches one approach
- Make software more resilient
- More aggressive automated repairs

Outline

- Introduction to data center networks
- Intra data center networks
- **Inter data center networks**
- Concluding thoughts

Backbone traffic growth



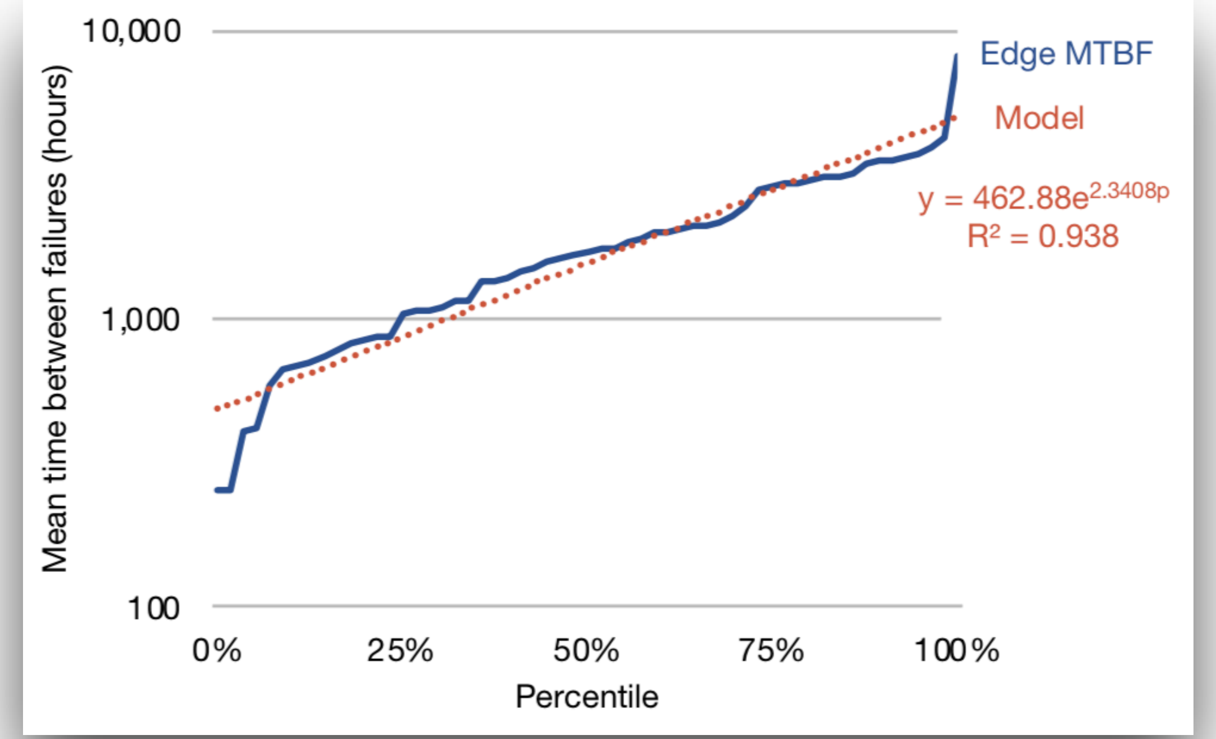
Data center backbones

- Shared resource
- Frequent link failure
- Capacity planning dictates reliability

Methodology: Measuring backbone reliability

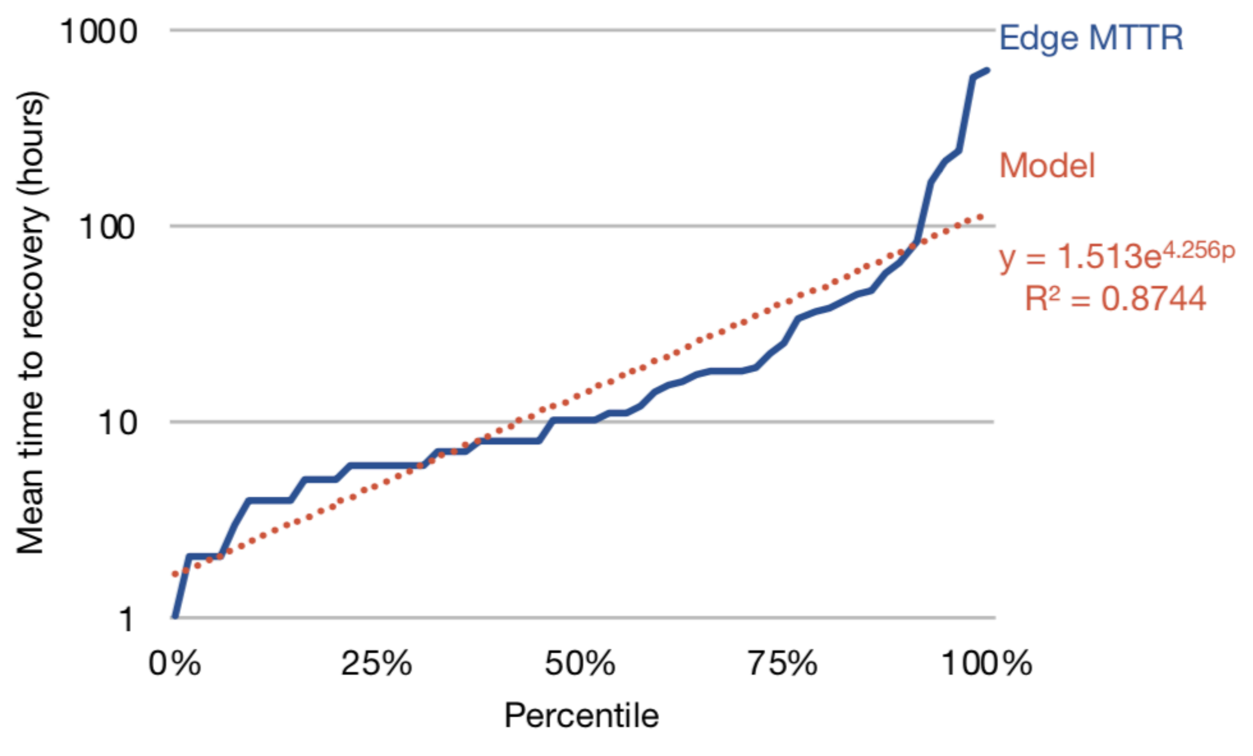
- Email sent for maintenance and outages
- Parsed and logged into a database
- Used to compute reliability statistics:
  - Mean time between failure (MTBF)
  - Mean time to repair (MTTR)
  - Over 18 months of data

Edge node MTBF distribution



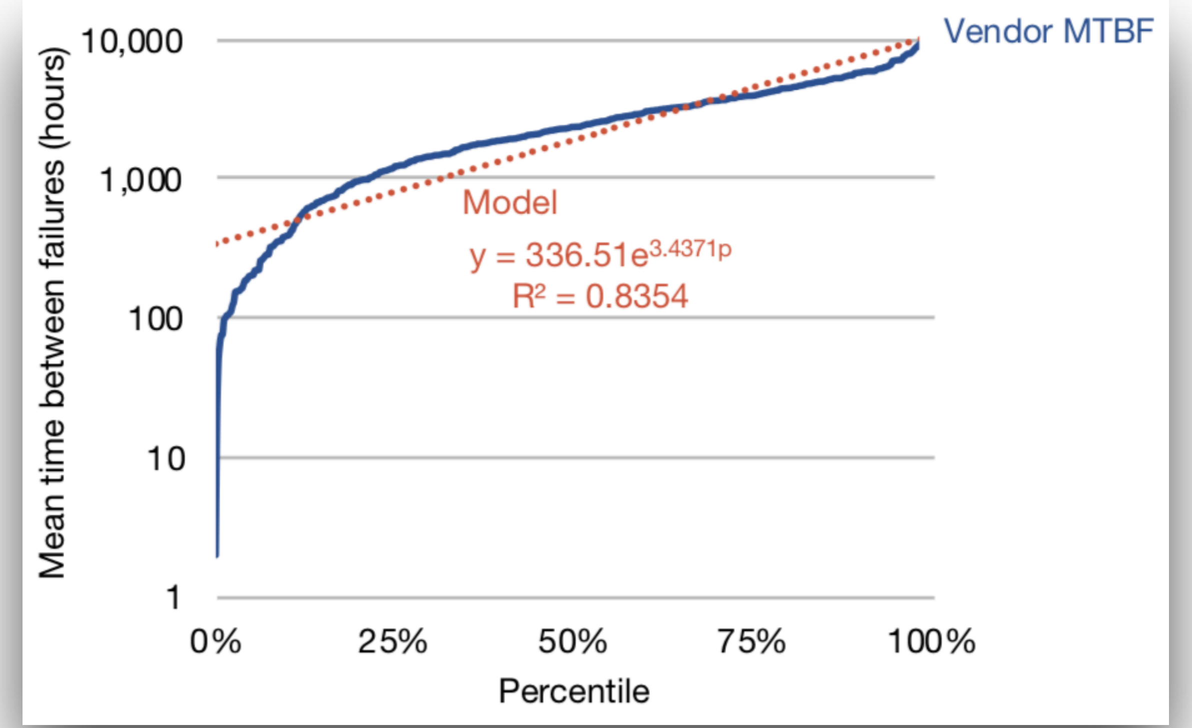
Typical edge node failure is in the order of months

Edge node MTTR distribution



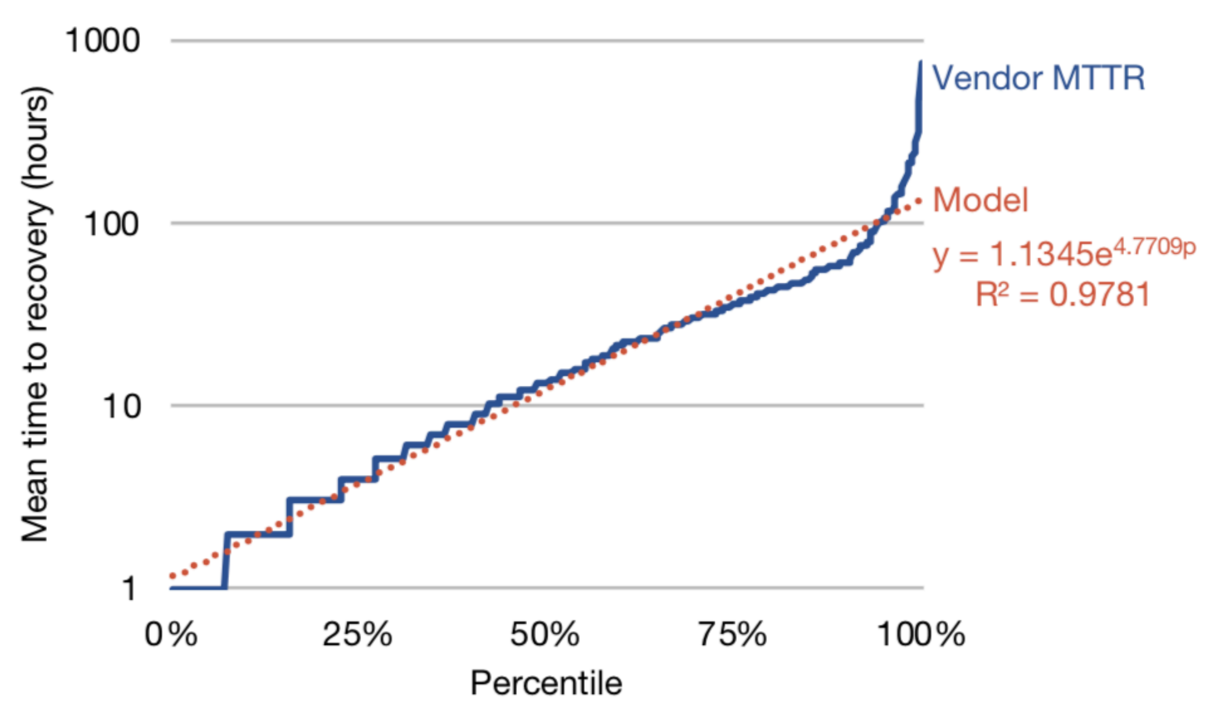
Edge node mean time to repair is in the order of hours

Fiber vendor MTBF distribution



Typical vendor link failure is in the order of months

Fiber vendor MTTR distribution



Vendor MTBF and MTTR span multiple orders of magnitude

Outline

- Introduction to data center networks
- Intra data center networks
- Inter data center networks
- **Concluding thoughts**

Conclusion

- First and last hop reliability forces to rethink how network and software share the task of reliability

Conclusion

- First and last hop reliability forces to rethink how network and software share the task of reliability
- Reliable backbone planning is a key enabler for geo replication and software management flexibility

Strengths

- Based on Facebook's data
  - 7 years of intra data center data
  - 18 months of inter data center data
- Large scale data center reliability
  - Common challenge across the industry.

Weaknesses

- Why do rack switch incidents increase?
- Logged versus unlogged failures
- Technology changes over time
- More engineers making changes
- Switch maturity

Brainstorming and Discussion

Brainstorming and Discussion

- What will happen to the backbone networks and core switches?
- Will they see a similar shift from proprietary to more customizable software?

Brainstorming and Discussion

- How to make better reports?
- Can we automate how they are written?

Brainstorming and Discussion

- Why are the rack switch incidents increasing over time?

The End

Thank You