

**Jeff Tantsura**

*Sr. Principal Network Architect Azure Networking*

Microsoft


**The only constant in life is change**





The background features a dark blue field with a faint, light blue grid of geometric shapes (squares, circles, lines) resembling a circuit board or network diagram. Overlaid on this are several thick, vibrant blue lines. Two lines are horizontal on the left side, one above the other. From the right end of the upper horizontal line, a thick blue line curves upwards and to the right, crossing over another thick blue line that is also curving upwards. From the right end of the lower horizontal line, a thick blue line curves downwards and to the right. Small, solid blue dots are placed at various points along these lines: one on the upper horizontal line, one on the lower horizontal line, one on the upper curve, one on the lower curve, and one on the line that crosses over the other curve.

# Reliability is a constant struggle – for everyone!

### Global DNS outage hits Microsoft Azure customers

Update: An Azure DNS outage, which affected Microsoft customers and services across the globe for several hours, now seems to be mostly mitigated.

By  Mary Jo Foley for All About Microsoft | September 15, 2016 -- 13:50 GMT (06:50 PDT) | Topic: Cloud

4  in 348   

It's been a rough morning for Microsoft Azure customers worldwide.

**RELATED STORIES**

**VERITAS** Data Management  
Veritas must reassure APAC customers about private equity ownership

BUSINESS INSIDER TECH INSIDER

### Amazon's Cloud Crash Disaster Permanently Destroyed Many Customers' Data

the two-way BREAKING NEWS FROM NPR

AMERICA

### Amazon And The \$150 Million Typo

March 3, 2017 · 10:54 AM ET

### Microsoft confirms Azure outage was human error

19 December 2014 | By Peter Judge

Data Center ▶ Cloud

### Google Cloud rolls back changes after 18-hour load balancer brownout

VMs across US, Europe and Asia all unable to "connect to backends"

By [Simon Sharwood, APAC Editor](#) 31 Aug 2017 at 03:03 11  SHARE ▼

# Agenda

Azure is one of the biggest networks in the world

Changes at scale and without any negative impact?

- Abstract/compartmentalize/contain
- Simulate
- Emulate
- Validate

# Agenda

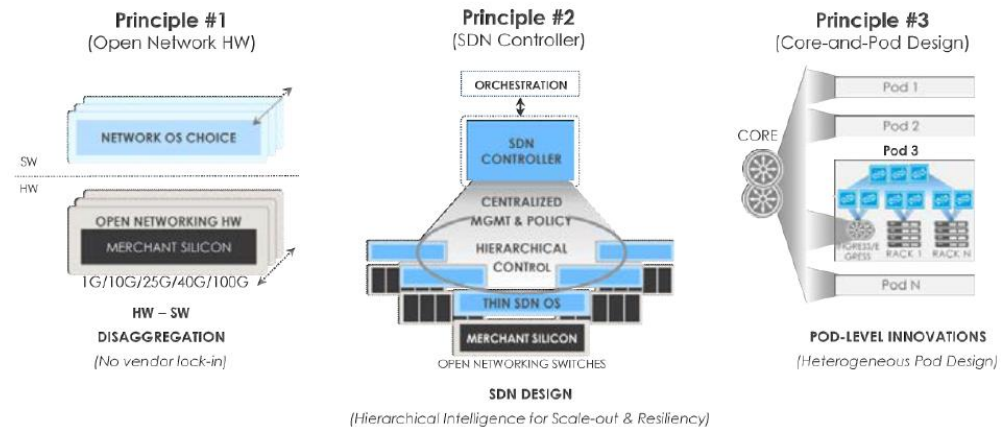
Azure is one of the biggest networks  
in the world

# What does it mean to be “hyperscale”?

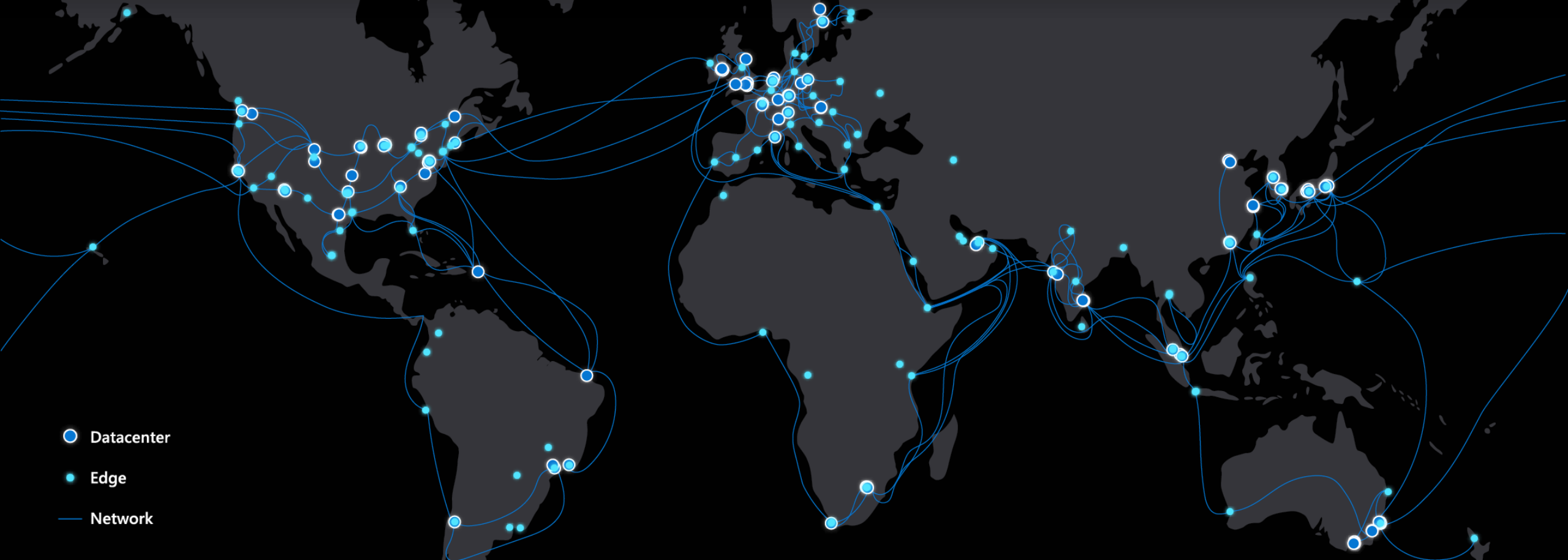
Architecture’s ability to scale with increasing demand.

- Common scale infrastructure
- Dynamic and automated provisioning
- Diverse workload mix
- Service Level Agreements
  - Consistency, Low-latency, high-throughput

## HYPERSCALE DESIGN PRINCIPLES



# Microsoft global network



**61** Azure regions

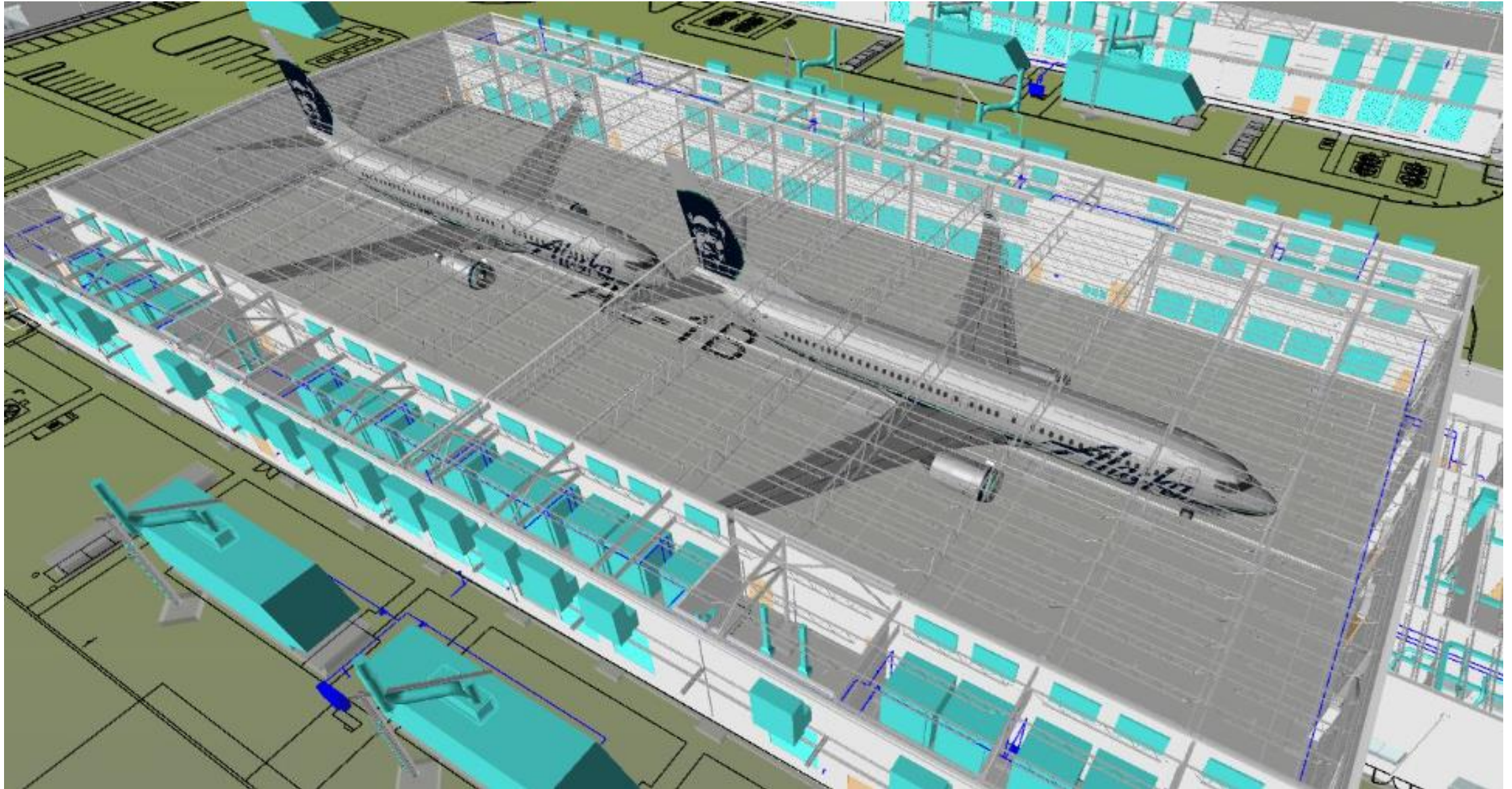
**175k+** miles of fiber + subsea cables

**185+** Network edge sites

**200+** Express route partners

**20k+** peering connections

# Microsoft builds (very) large Data Centers





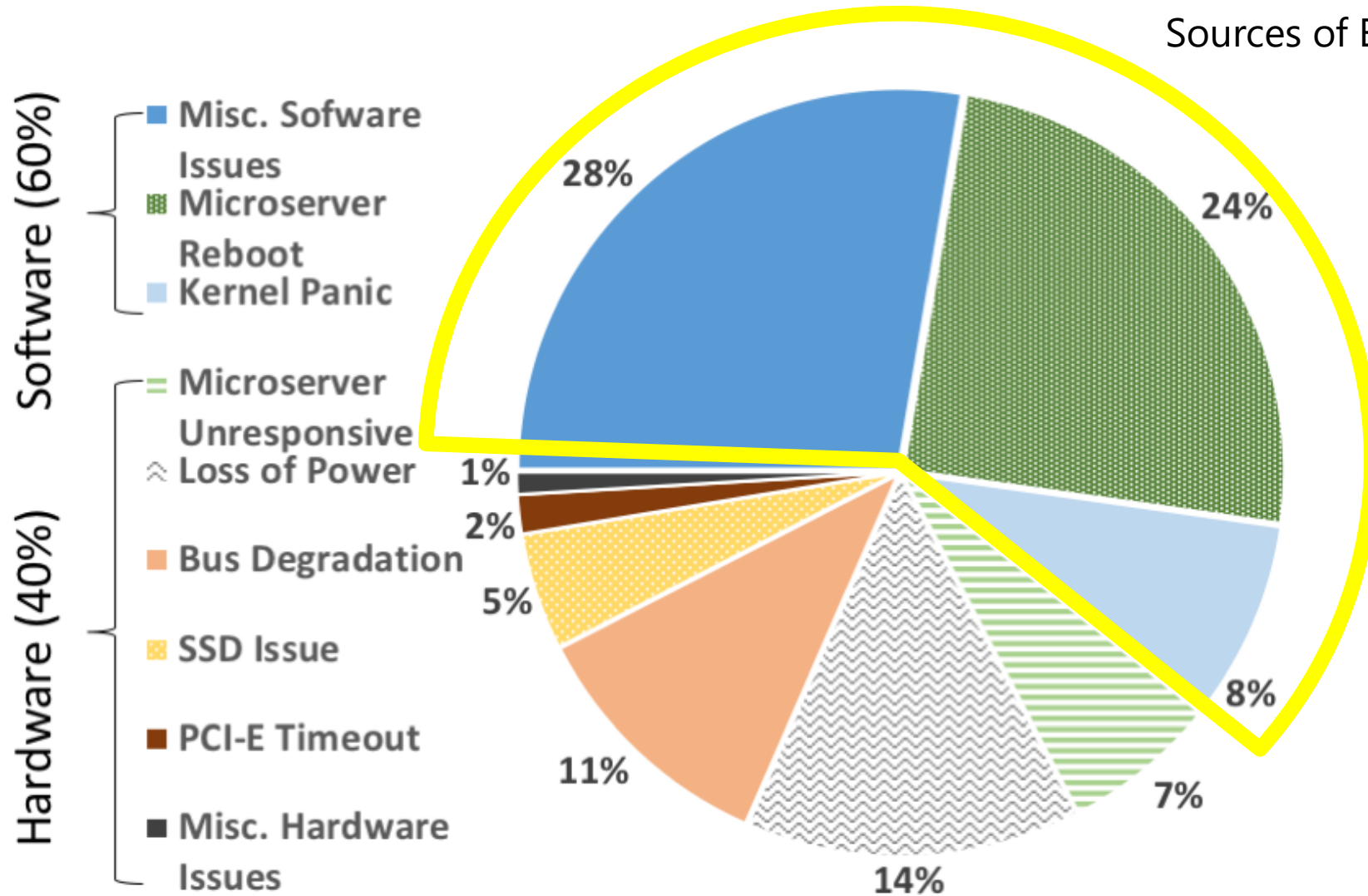






# Changes cause outages

Sources of Error in Facebook (2018)



• Data from <https://research.fb.com/building-switch-software-at-facebook-scale/>

**How do we build and  
operate the most  
reliable hyperscale  
network?**

# Agenda

Changes at scale and without any negative impact?

- Abstract/compartmentalize/contain

# Azure DC networking – building blocks

Amount of  
details



rack



Degree of  
abstraction

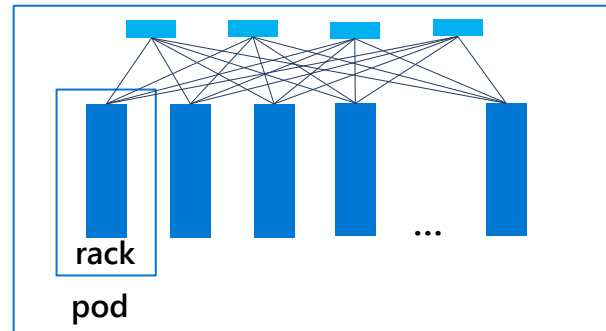


# Azure DC networking – building blocks

Amount of details



POD



Degree of abstraction



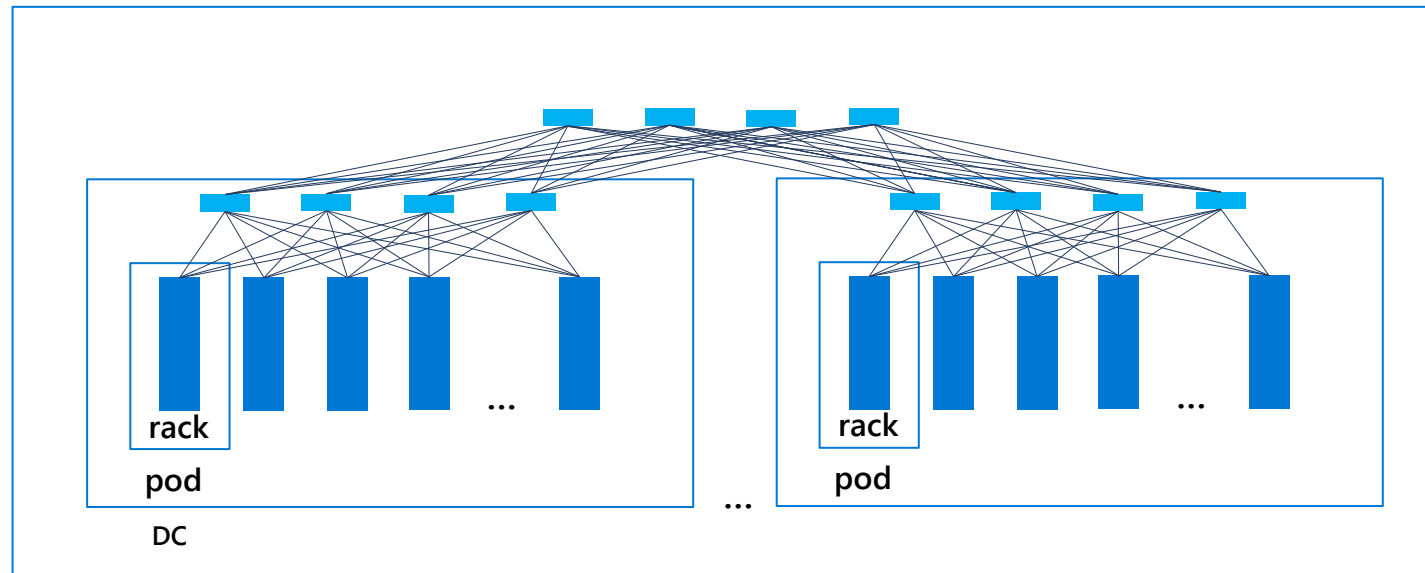
# Azure DC networking – building blocks

Amount of details



DC

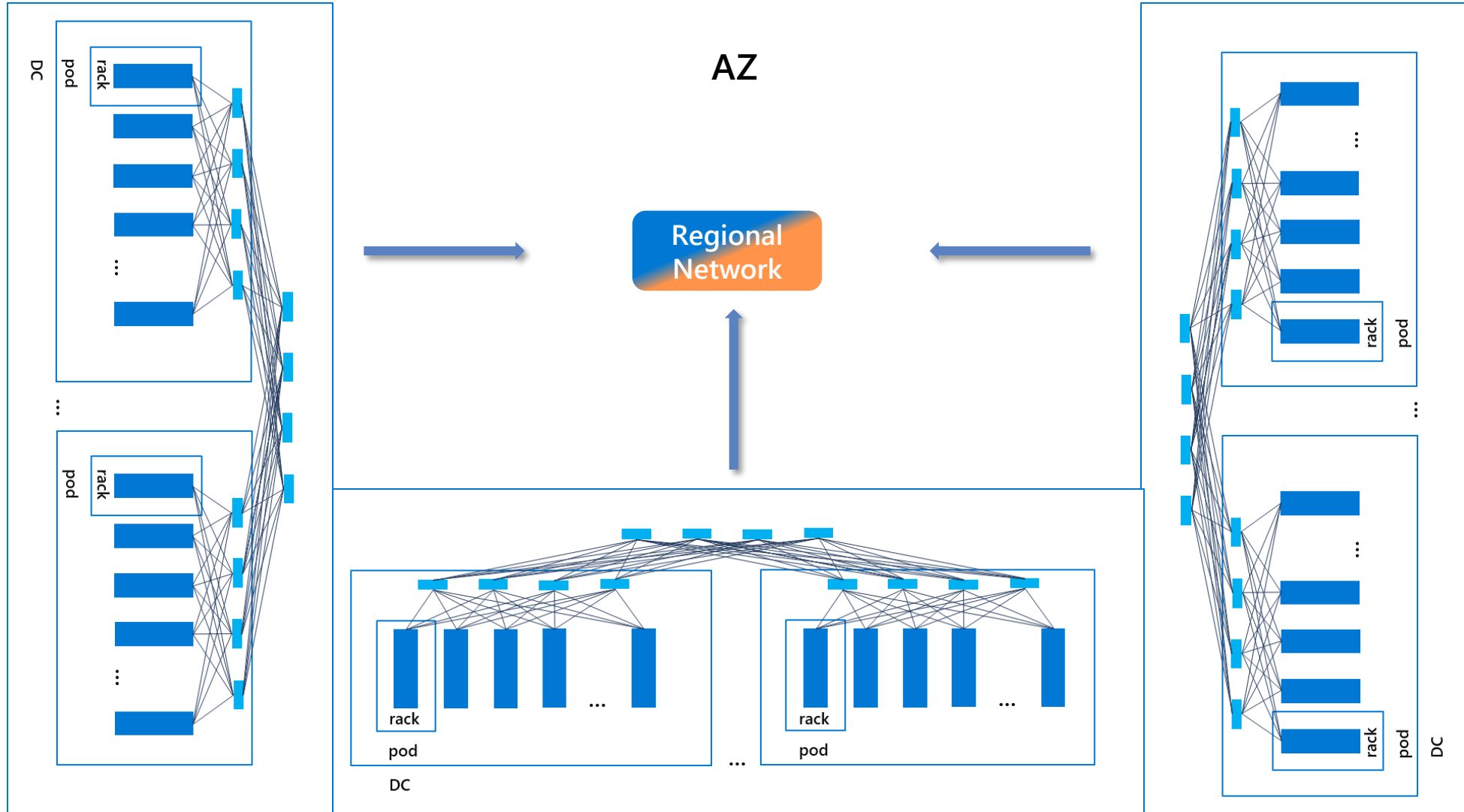
Degree of abstraction





# Azure DC networking – building blocks

Amount of details



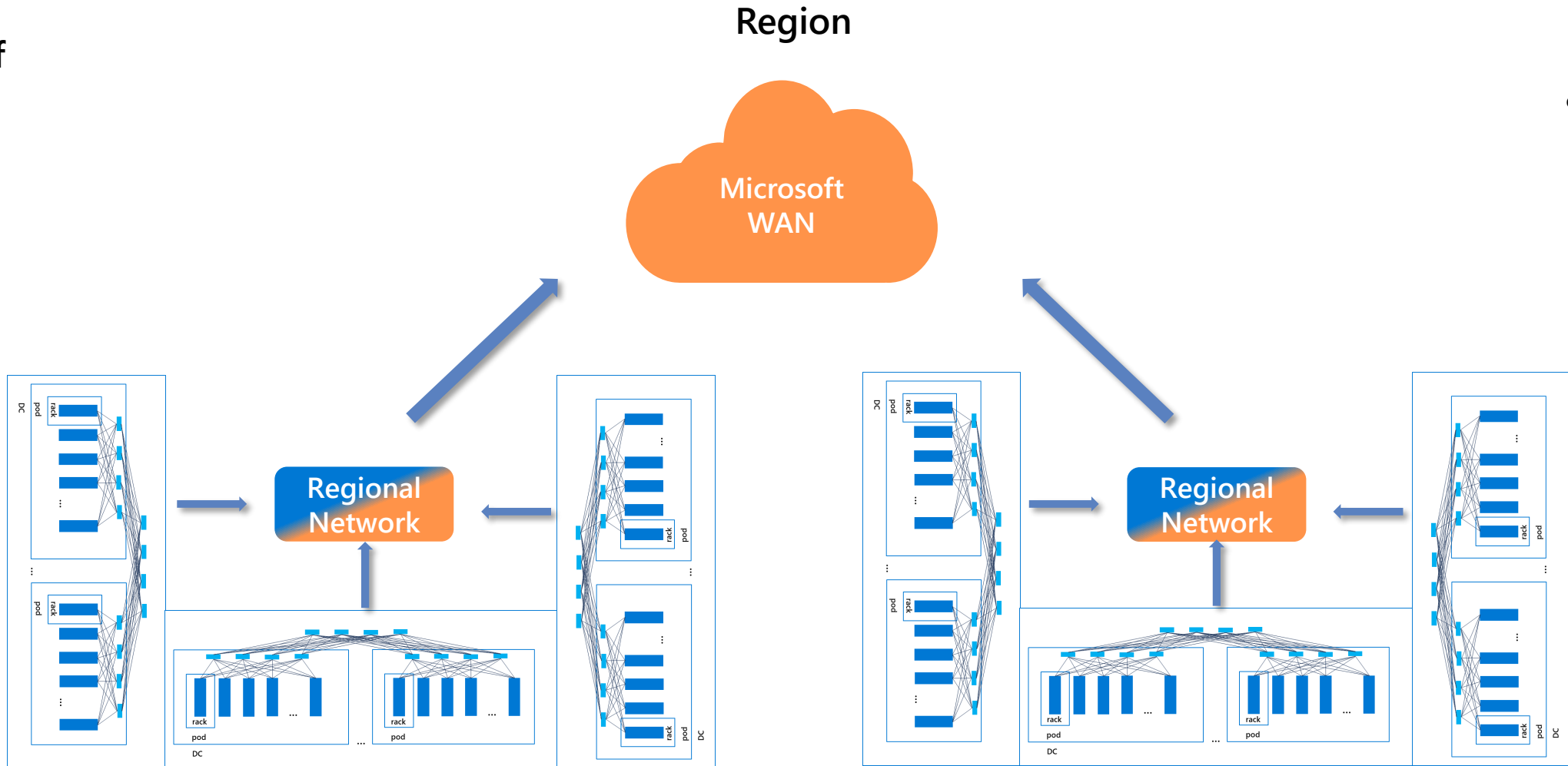
Degree of abstraction



# Azure DC networking – building blocks

Amount of details

Degree of abstraction



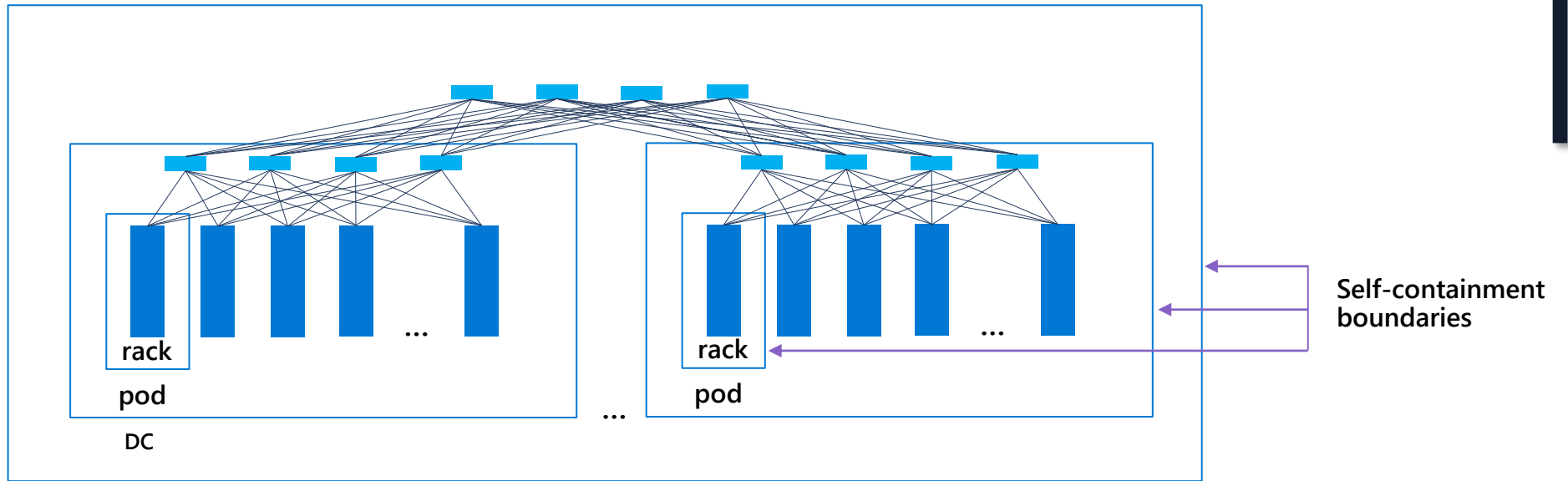
# Azure DC networking – building blocks

Amount of details

**self-containment**  
[ˌselfkənˈtɑːnmənt]  
NOUN

- 1.the condition of being complete, or having all that is needed, in itself.
- 2.the quality of not depending on or being influenced by others; independence.

Degree of abstraction



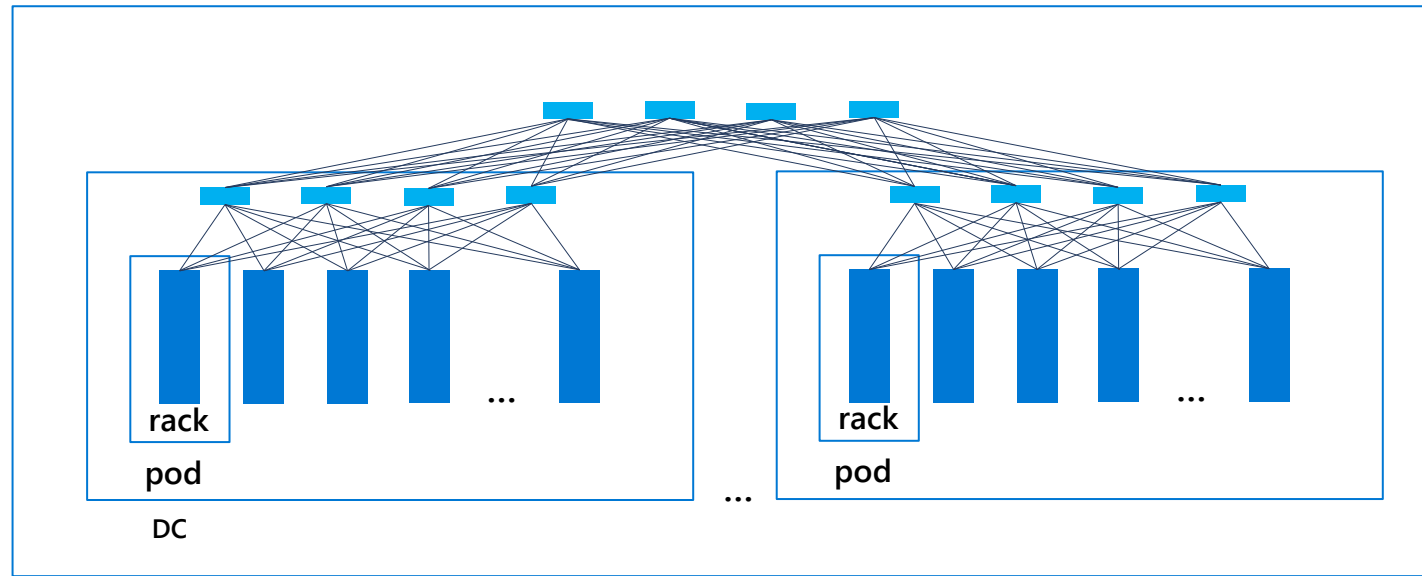
# Azure DC networking – routing

Amount of details

Requirements:

- reduce number of routes to the strictly necessary minimum
- enough details to prevent blackholing
- in multi-planar topology POD local backup is always preferred
- parallel plane backup path is preferred over “plane merge” point (beyond DC)

Degree of abstraction



# Azure DC networking – routing

Amount of  
details

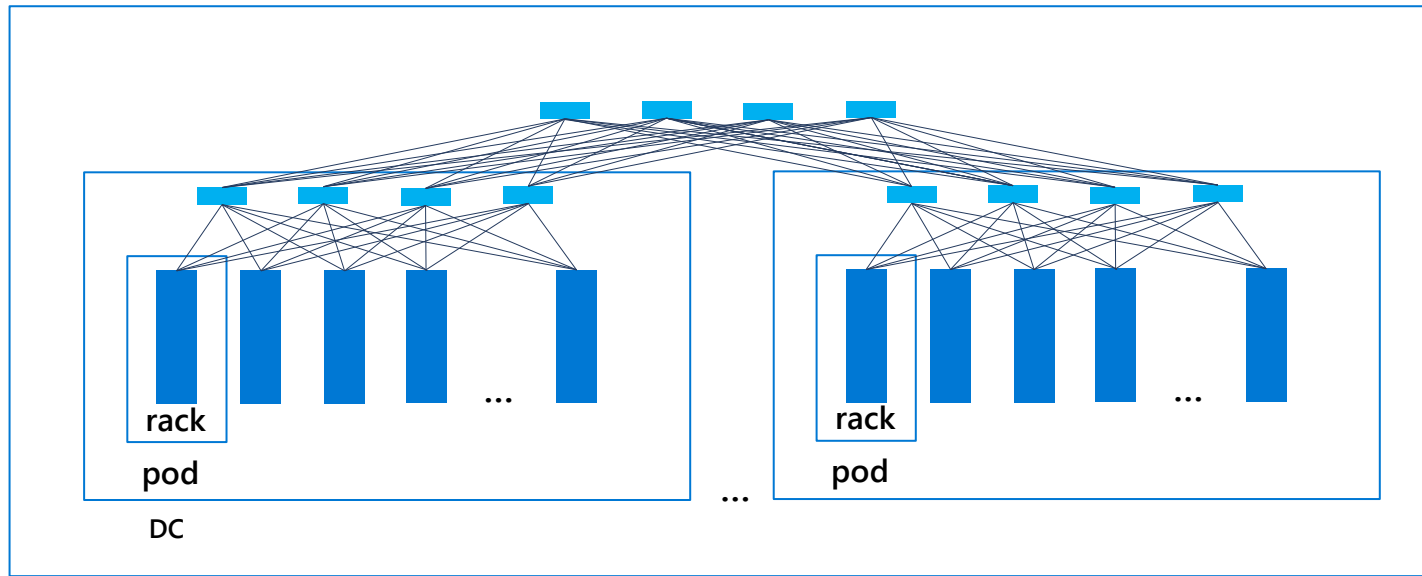


Naïve approach (too little information)

Path Vector + aggregation:

- blackholing on next-next-hop failure

Degree of  
abstraction



# Azure DC networking – routing

Amount of details

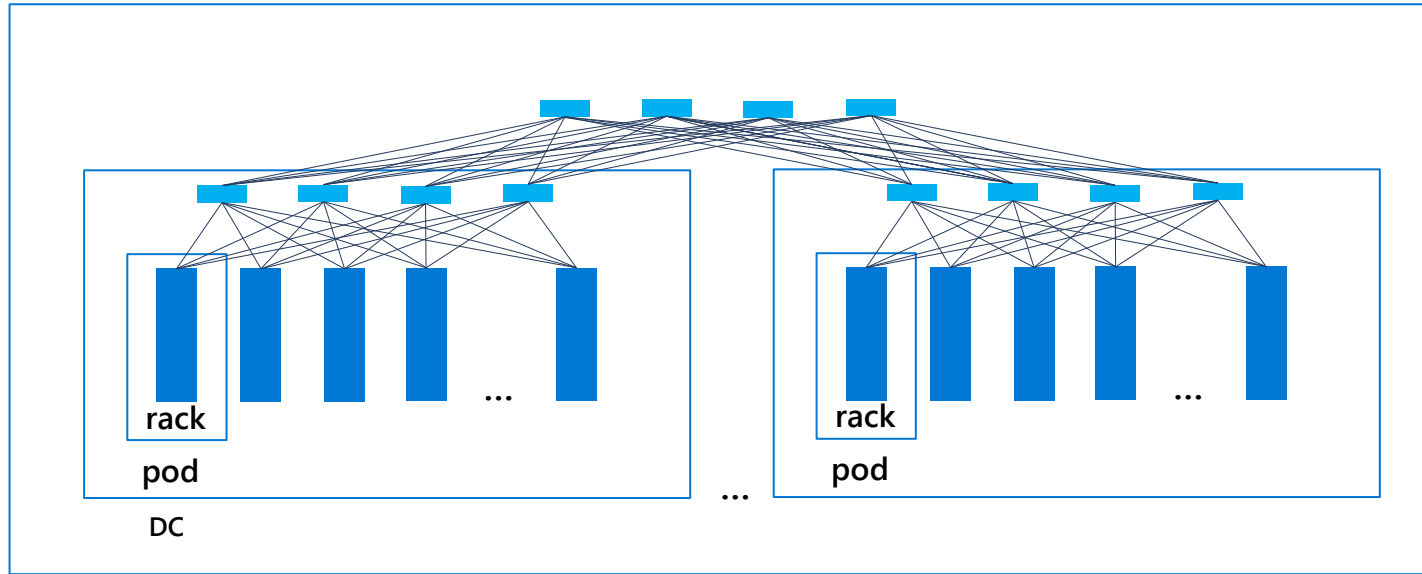


Naïve approach (too much information)

Link State:

- excessive flooding
- blast radius

Degree of abstraction



# Azure DC networking – routing

Amount of details

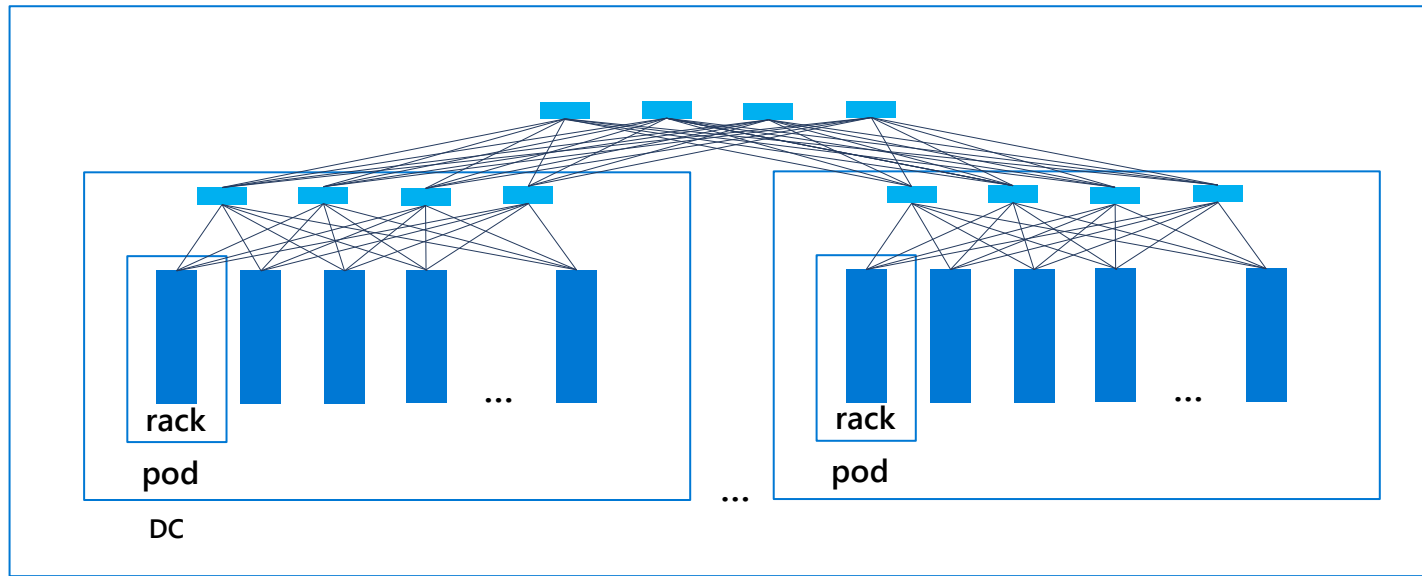


Pragmatic approach

Path Vector + once bounce:

- local traffic always stays local, longer backup path through bounce
- backup path is always available (with minimum path hunting)

Degree of abstraction



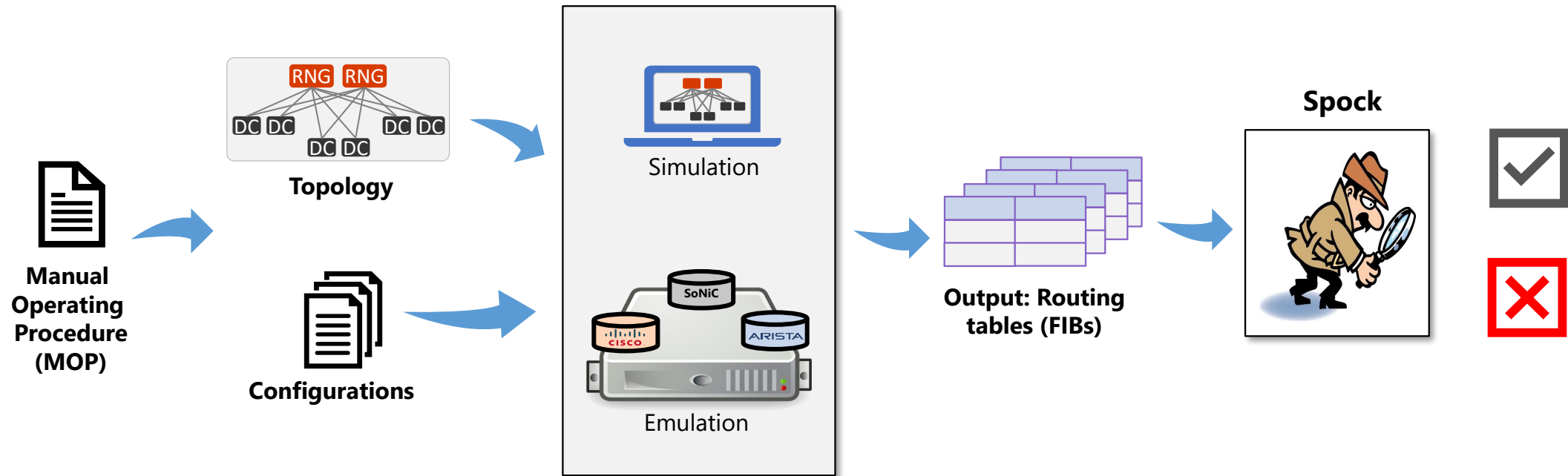
# Agenda

Changes at scale and without any negative impact?

- Simulate
- Emulate
- Validate



# Network Change Verification System (NCVS)



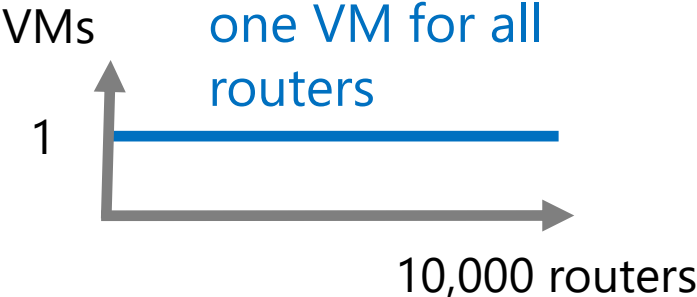
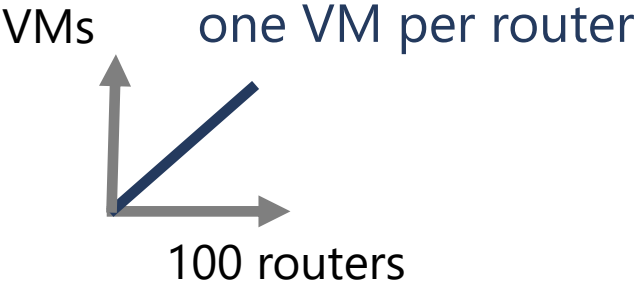
# Emulation vs Simulation

## Distributed Emulation

## Efficient Simulation



**Scalability:**



**Coverage:**

subset of a datacenter

entire Azure region



**Speed:**

non-interactive (~ 1 hour)

interactive (~ minutes)



**Fidelity:**

Bug compatible with actual devices

Validated daily: NLS vs production ribs across all Azure switches

# Agenda

Changes at scale and without any negative impact?

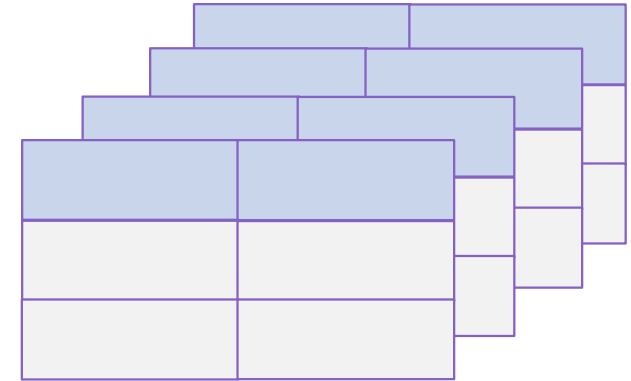
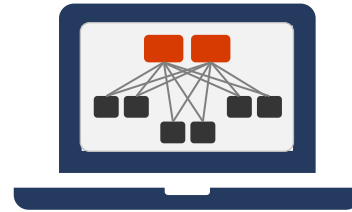
- Simulate
- Emulate
- Validate

# Test in a high-fidelity simulator before you fly!

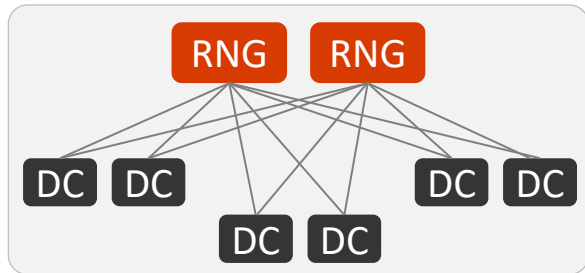


# Network Logic Solver

Configuration files



Output: Routing tables  
(RIB/FIB)



Azure region topology

Scales to Azure  
Regions

Models routing protocols behavior

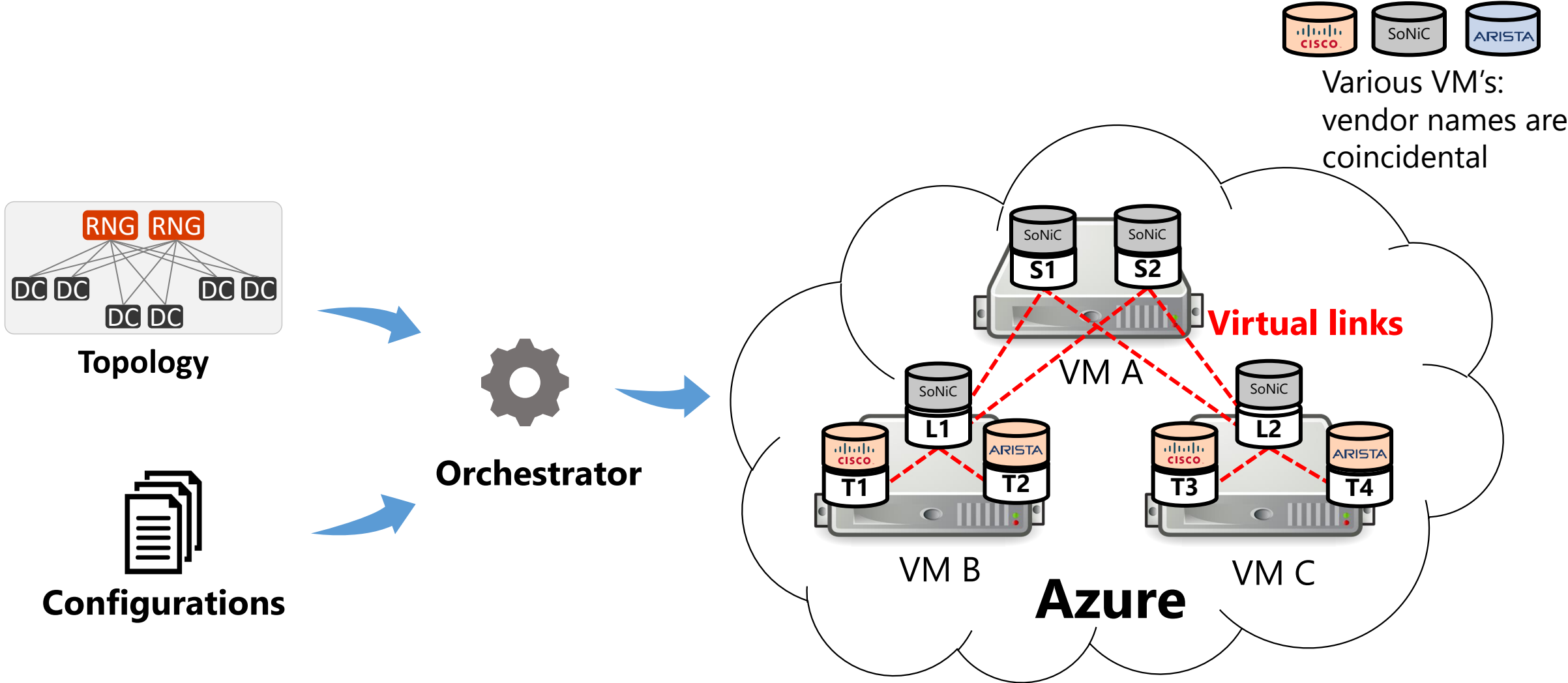
Efficiently simulates control  
plane

# Agenda

Changes at scale and without any negative impact?

- Simulate
- Emulate
- Validate

# Network Emulation



# Open Network Emulator (ONE) – digital tween

**fast**

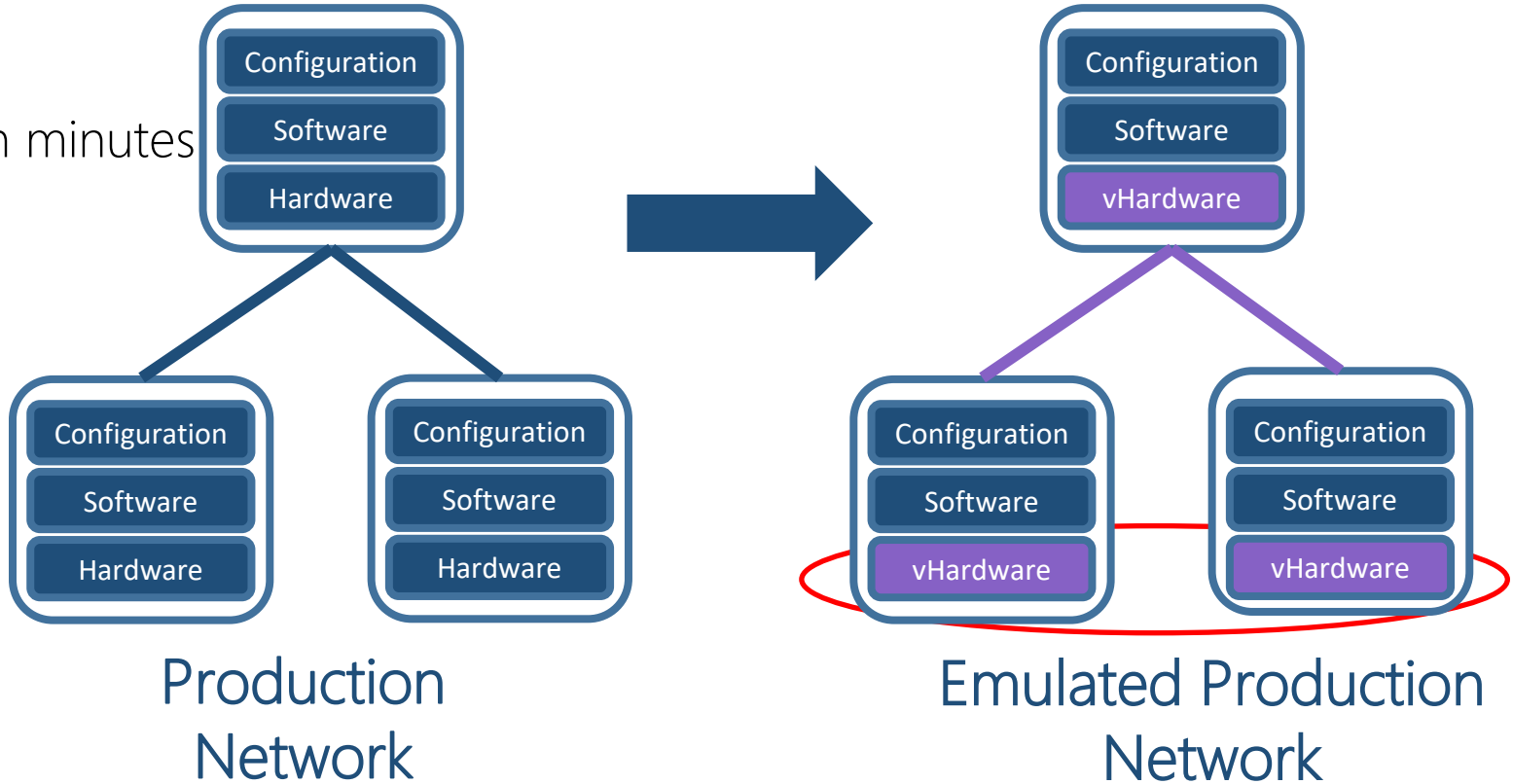
network with 1000s of devices created in minutes

**seamless**

push-button deployment

**high fidelity**

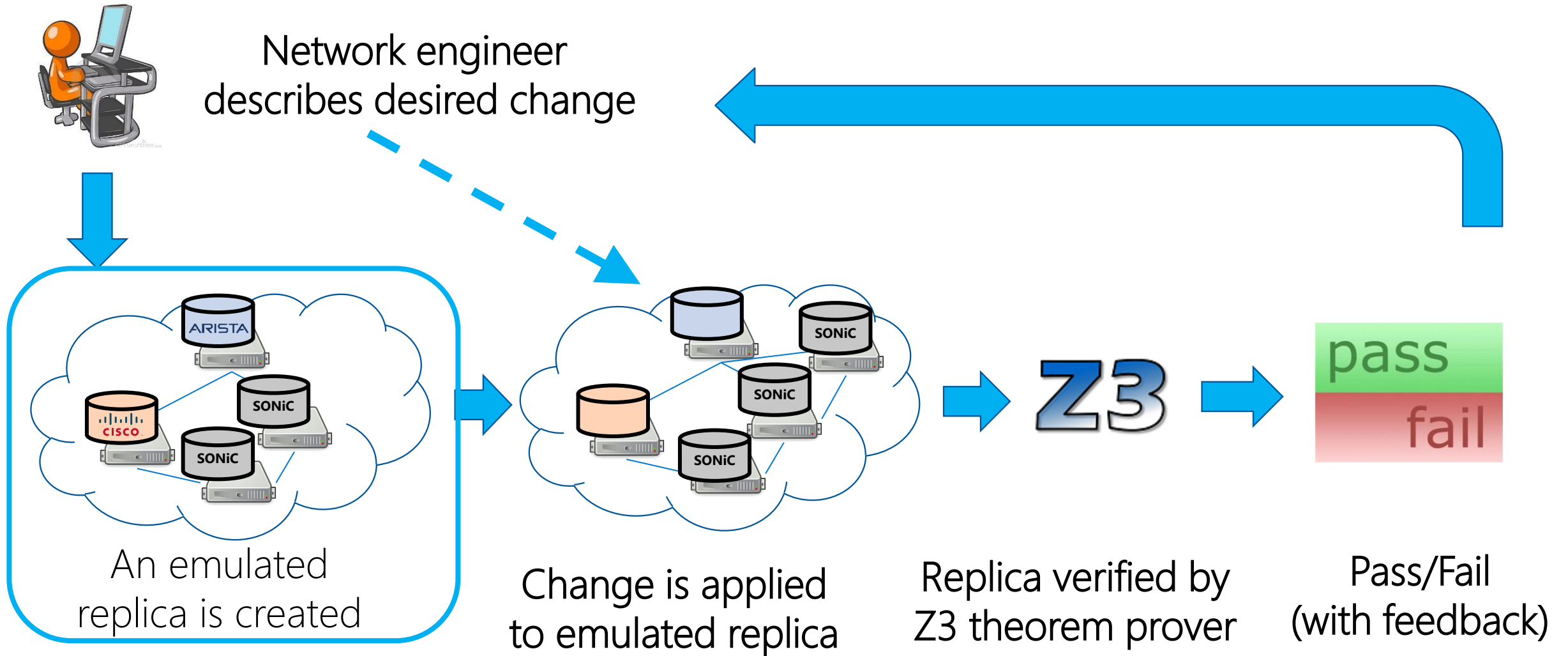
devices work exactly as production  
support from multiple vendors



**30+ million core-hours on Azure Network Emulator yearly runs**



# ONE typical usage scenario



# Agenda

Changes at scale and without any negative impact?

- Simulate
- Emulate
- Validate

# Formal methods: Forwarding Information Base (FIB) Verifier

Continuously check for "intent gap"

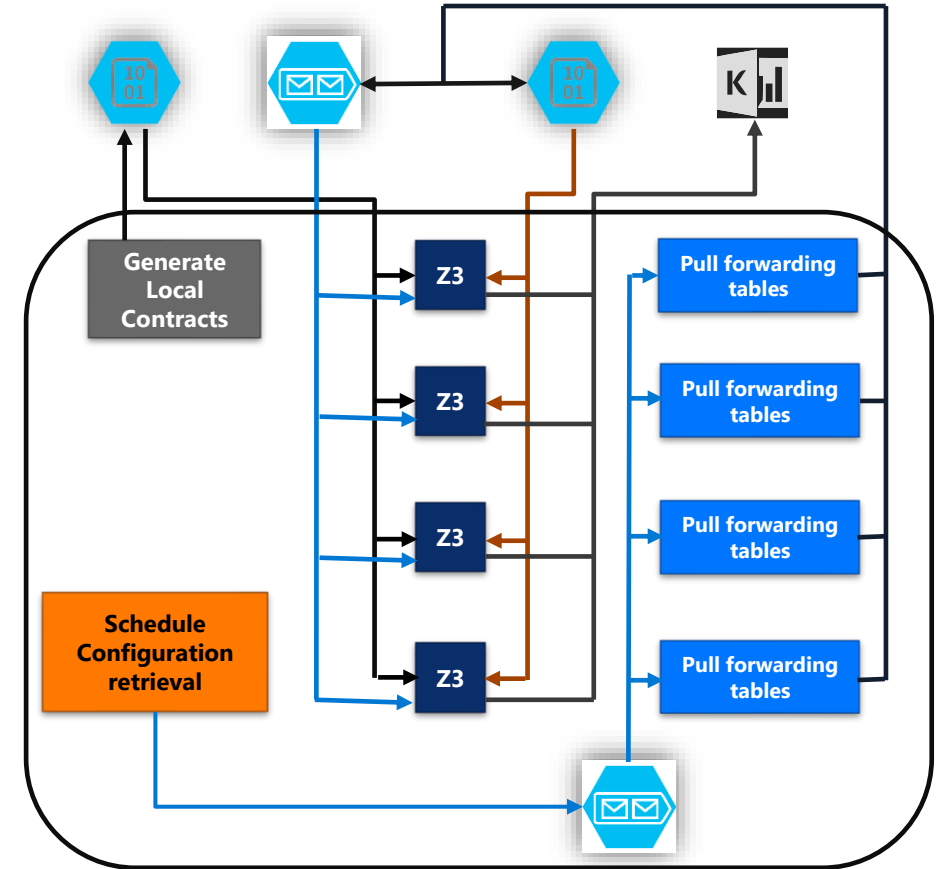
Derive routing invariants from topology and architectural metadata

- local correctness of each device
- global reachability properties

Validate invariants against actual FIBs

Errors trigger workflow for automated or manual remediation

Continuous check on over 100K+ devices in fleet



# Spock Overview

## What is Spock

- Network testing framework in C#
- Invariants as test methods with assertions (e.g., `Assert.IsTrue`)
- Special API + DSL for symbolic reasoning: `Assert.AllPackets(...)`
- Mixes concrete and symbolic testing

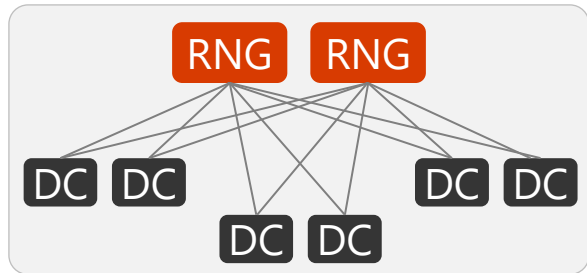
## How does Spock verify user assertions

- Classical model checking algorithms from the 90s
- Careful implementation and domain optimizations

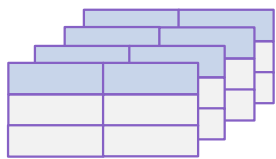
# Spock Architecture

```
[SpockTestMethod(scope: Datacenter)]  
public void TorReachabilityTest(Datacenter dc) {  
    foreach (var t0 in dc.T0s) {  
        foreach (var prefix in t0.VlanAddresses) {  
            var relevant = ContainedBy(prefix);  
            var reachable = Reachable(dc.T0s, t0);  
            var invariant = If(relevant, reachable);  
            Assert.AllPackets(invariant);  
            ...  
        }  
    }  
}
```

**Input:** Specification as a test



**Input:** Azure region topology



**Input:** Routing tables (FIBs)

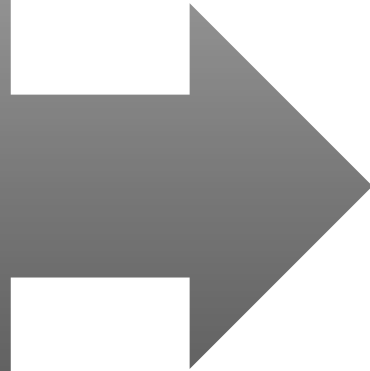


```
Passed: BgpSessionTest  
Failed: TorReachabilityTest  
Datacenter: dc01  
Packet: 10.12.1.3  
Path: dc01-0108-01t0,  
       dc01-0108-12t1  
...
```

**Output:** Test result

# The only constant in life is change

- Abstract
- Simulate
- Emulate
- Validate



- Grow at hyper speed
- 1000's of changes a day
- Reliability
- Agility

**We are hiring!**

Thank you!

MPLSSD&AI<sup>★</sup>NETWORLD22  
5/6/7APRIL