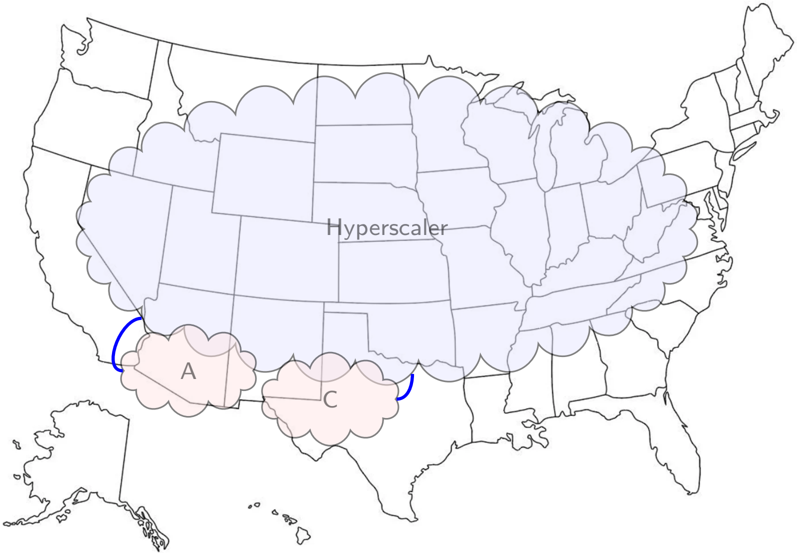
Thrive, Not Just Survive

Engineering Resilience in Content Provider Networks

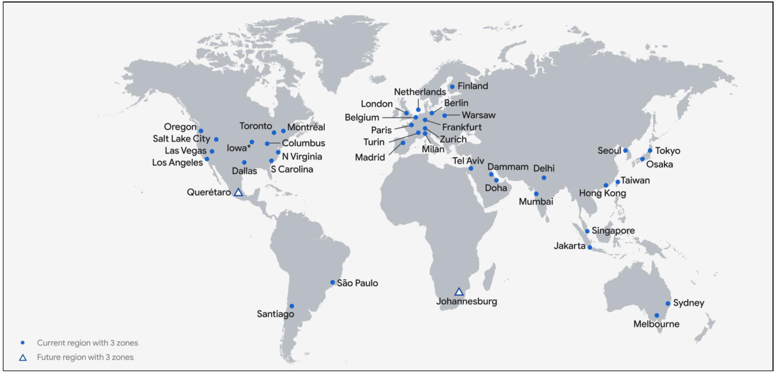
Ramesh Govindan

December 20, 2023

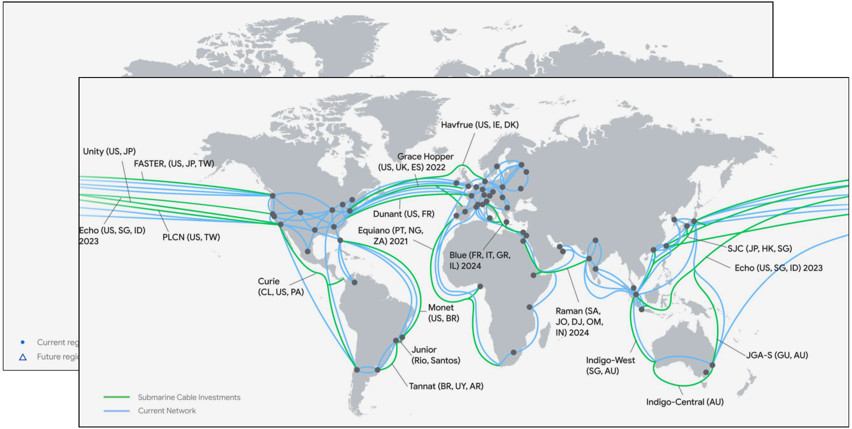
The Internet Today



Hyperscalers



Hyperscalers



Survivability: The Availability Perspective

Survivability

- Focus on resilience to network failures

Availability

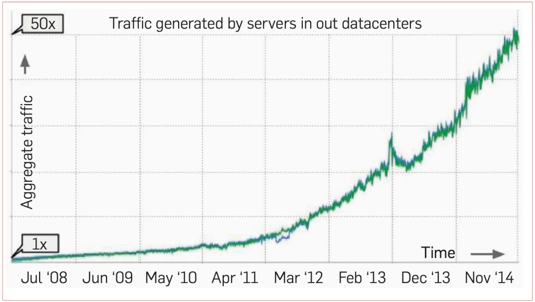
- Quantifiable outcome of resilience efforts

Availability is Quantifiable

Downtime per month

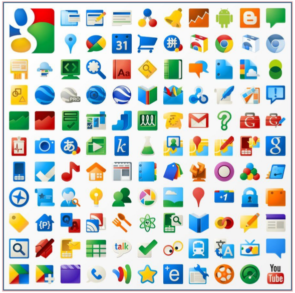
Three-nines	25 mins
Four-nines	4 mins
Five-nines	25 seconds

Achieving High Availability is Hard



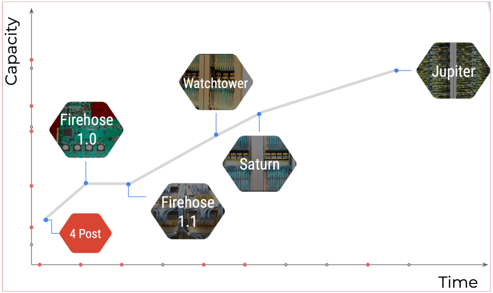
Rapid traffic growth

Achieving High Availability is Hard

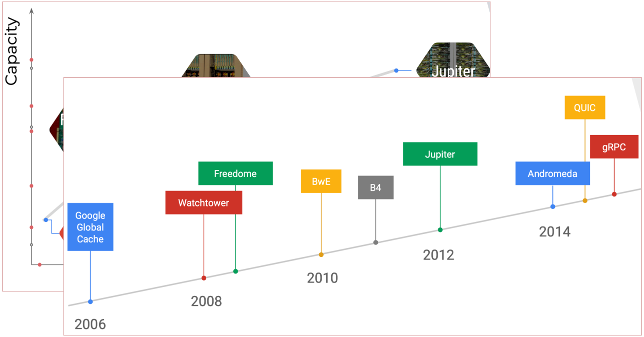


Services drive this growth

Achieving High Availability is Hard

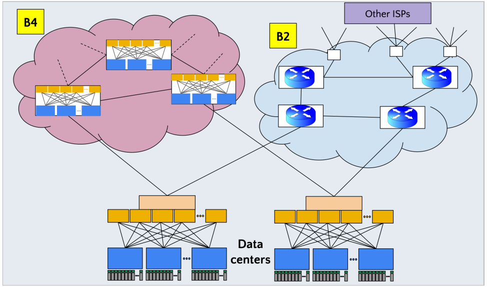


Achieving High Availability is Hard



Need high feature velocity to meet growth

Impediments to High Network Availability

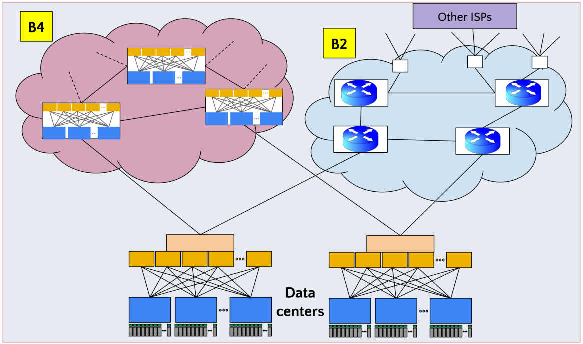


Analysis of large network failures

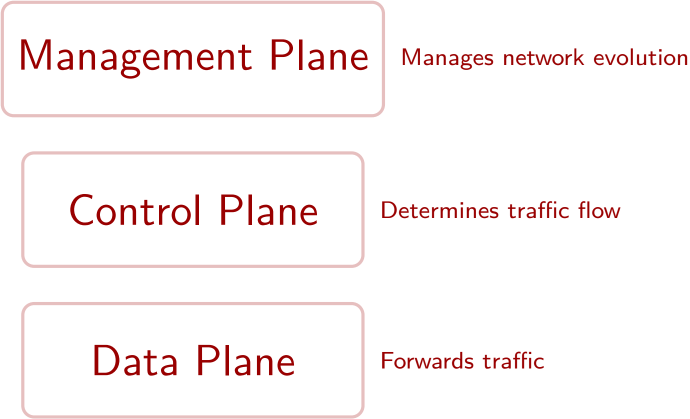
- Failure taxonomy
- Principles for high availability

R. Govindan *et al.*, "Evolve or Die: High-Availability Design Principles Drawn from Google's Network Infrastructure," in *Proceedings of the ACM Conference of the Special Interest Group on Data Communication (SIGCOMM '16)*, Florianópolis, Brazil, Aug. 2016

Background: Google’s Network



Background: Network Decomposition

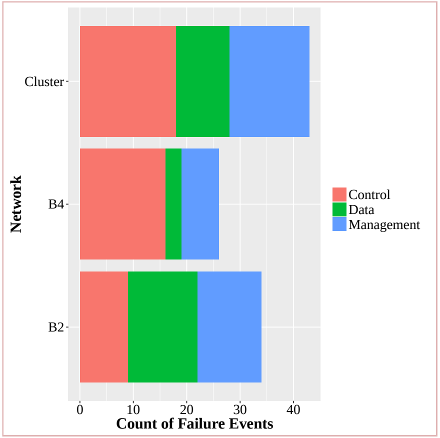


Management Plane — Manages network evolution

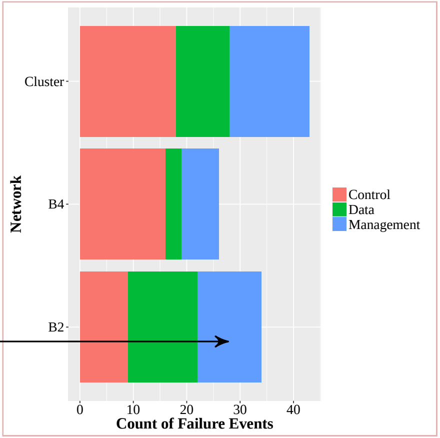
Control Plane — Determines traffic flow

Data Plane — Forwards traffic

Where do failures occur?

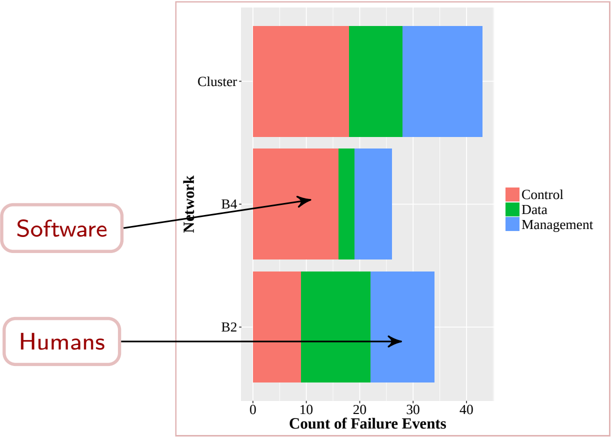


Where do failures occur?



Humans

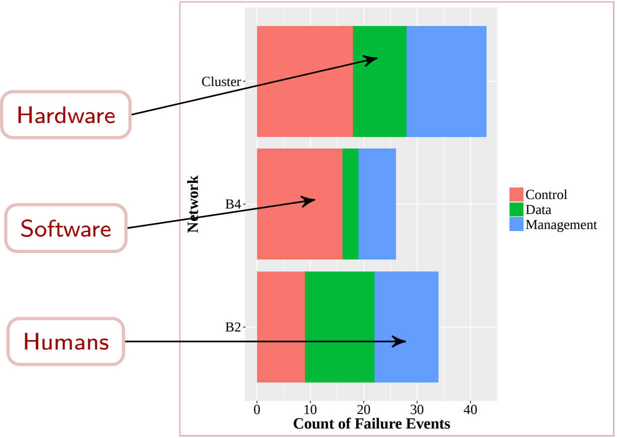
Where do failures occur?



Software

Humans

Where do failures occur?

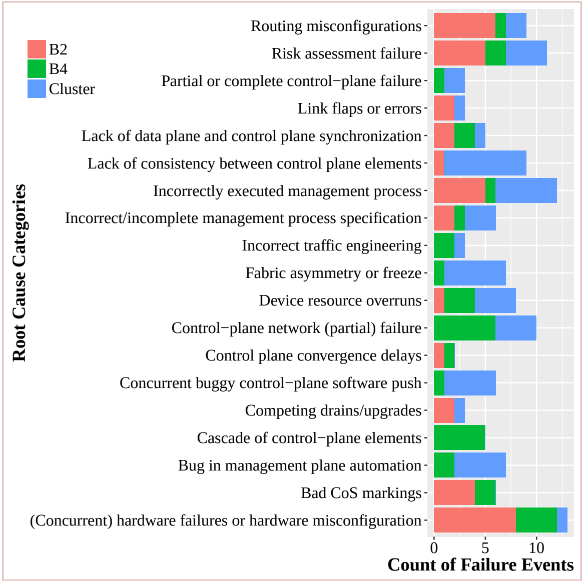


Hardware

Software

Humans

What are the root causes?



Other Findings

Lessons

- 70% of failures during management operations
- most due to concurrent failures
- large, cascading failures frequent

No silver bullet

- Defense in depth

Defense in Depth

Design

Performance evaluation
Code reviews
Regression testing

Defense in Depth

Design

Performance evaluation
Code reviews
Regression testing

Deployment

Property checking
Testing, canarying
Progressive rollout

Defense in Depth

Design

Performance evaluation
Code reviews
Regression testing

Runtime

Fast recovery
Fallbacks
Graceful degradation

Deployment

Property checking
Testing, canarying
Progressive rollout

Defense in Depth

Design

Performance evaluation
Code reviews
Regression testing

Runtime

Fast recovery
Fallbacks
Graceful degradation

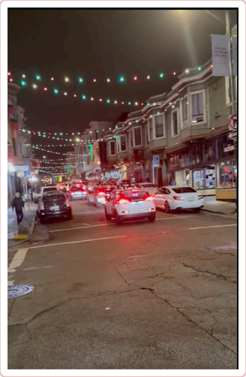
Deployment

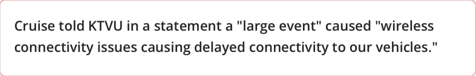
Property checking
Testing, canarying
Progressive rollout

Operation

Incremental updates
Programmed management
Postmortems

Towards a Brave New World





Cruise told KTVU in a statement a "large event" caused "wireless connectivity issues causing delayed connectivity to our vehicles."

Connected autonomous vehicles

- a significantly harder problem
- humans in the control loop
- stakeholders with competing incentives