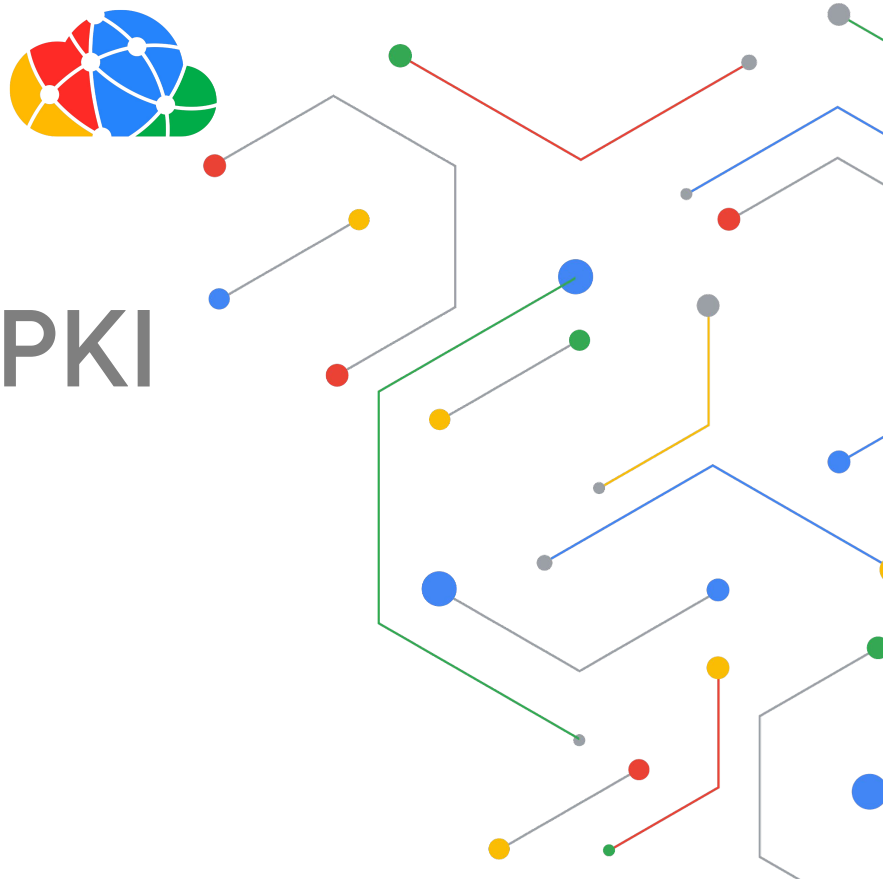
Google Cloud

“Without networking, there is no cloud.”

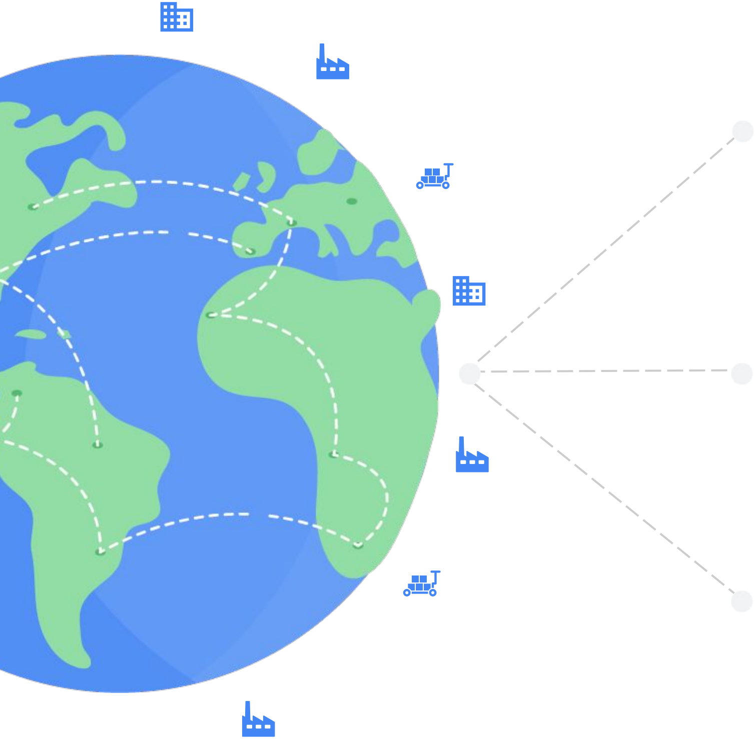


Cloud SDN

BGP Peering and RPKI

Shaowen Ma
Group Product Manager
shaowen@google.com

Agenda



Cloud SDN

Underlay B2/B4/Jupiter/Andromeda

Cloud BGP Peer and RPKI

Cloud Global Network and Interconnect

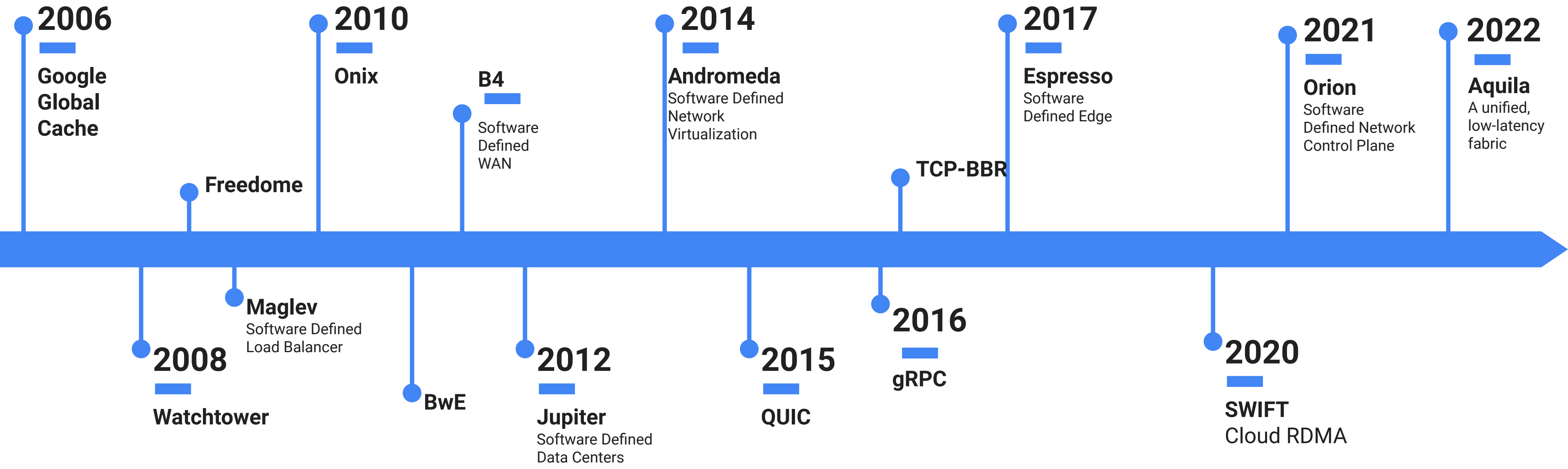
BGP Peering and RPKI

Summary

Many Cloud Innovations

A snapshot

Google innovations in networking



2006
Google Global Cache

Freedome

2008
Watchtower

Maglev
Software Defined Load Balancer

2010
Onix

BwE

B4
Software Defined WAN

2012
Jupiter
Software Defined Data Centers

2014
Andromeda
Software Defined Network Virtualization

2015
QUIC

TCP-BBR

2016
gRPC

2017
Espresso
Software Defined Edge

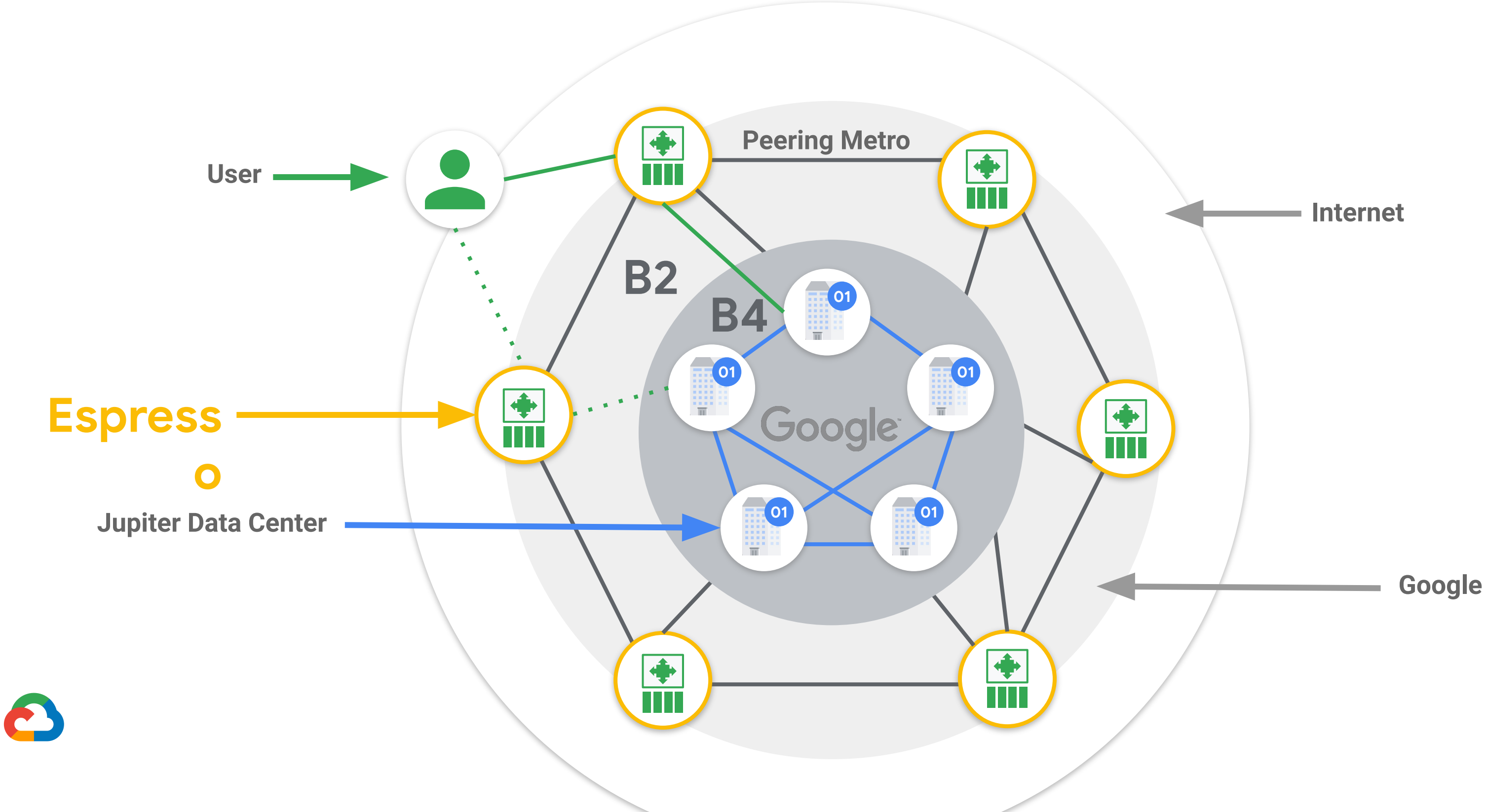
2020
SWIFT
Cloud RDMA

2021
Orion
Software Defined Network Control Plane

2022
Aquila
A unified, low-latency fabric

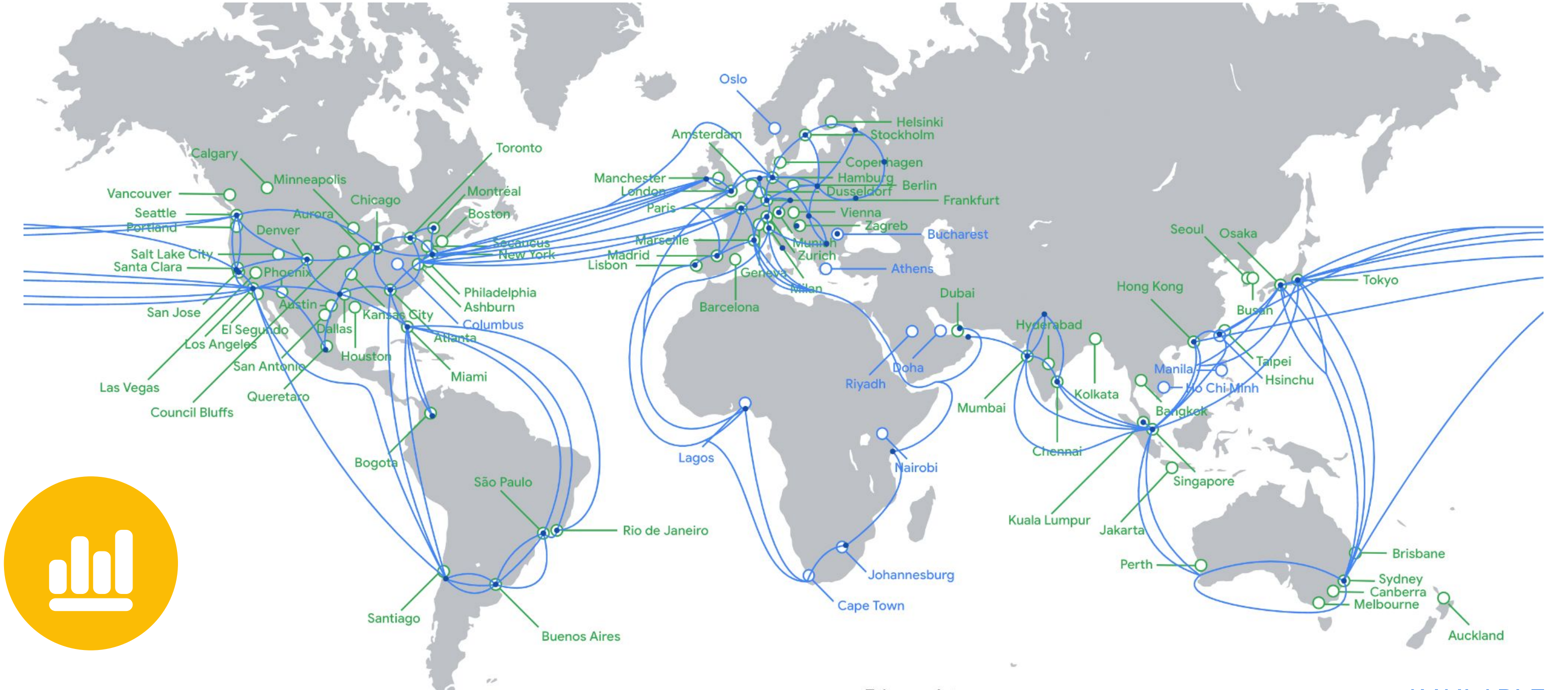
Google Network: Architect Overview

Software Defined from the Inside Out



Cloud Connect/BGP Peering

Google Global Network



○ Dedicated Interconnect

• Edge point of presence

○ Dedicated Interconnect (In Progress)

— Network

Google Cloud

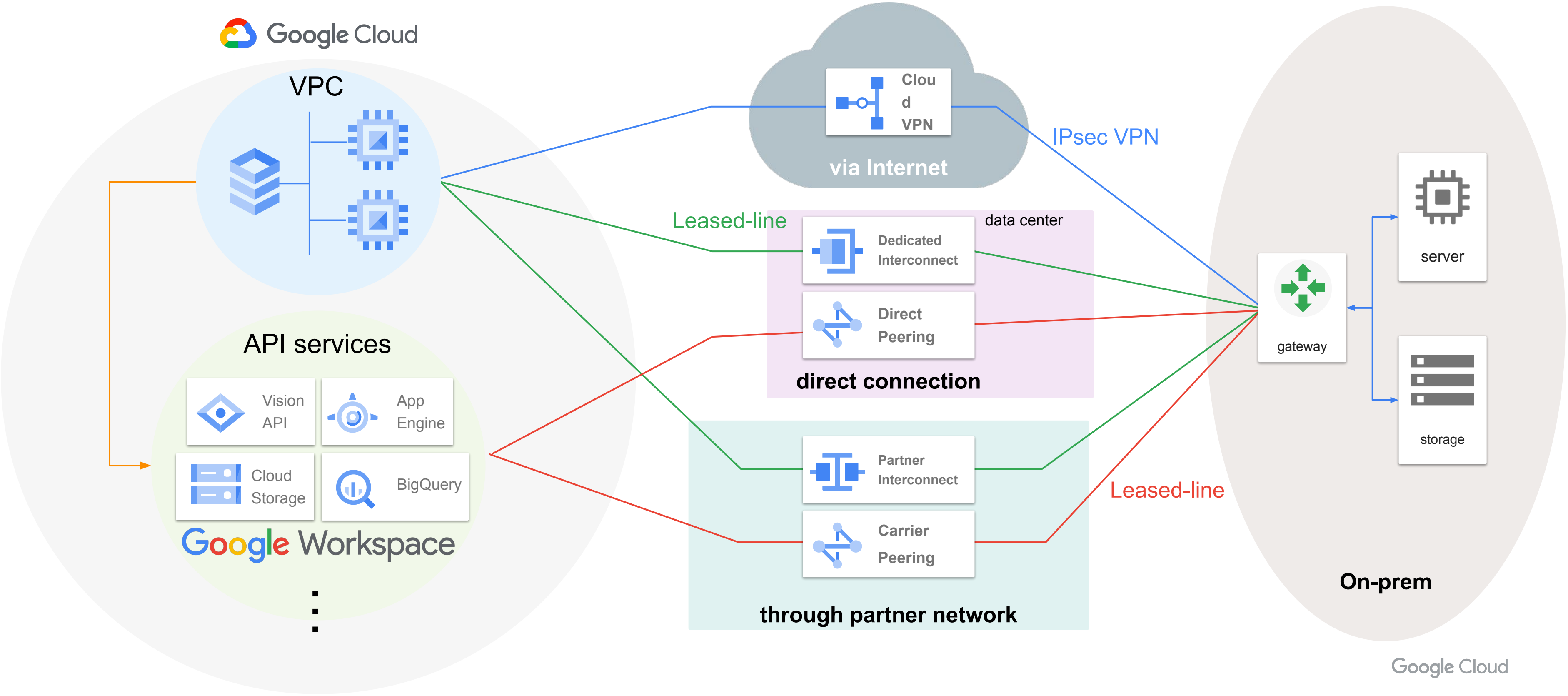
34
REGIONS

103
ZONES

173
EDGE LOCATIONS

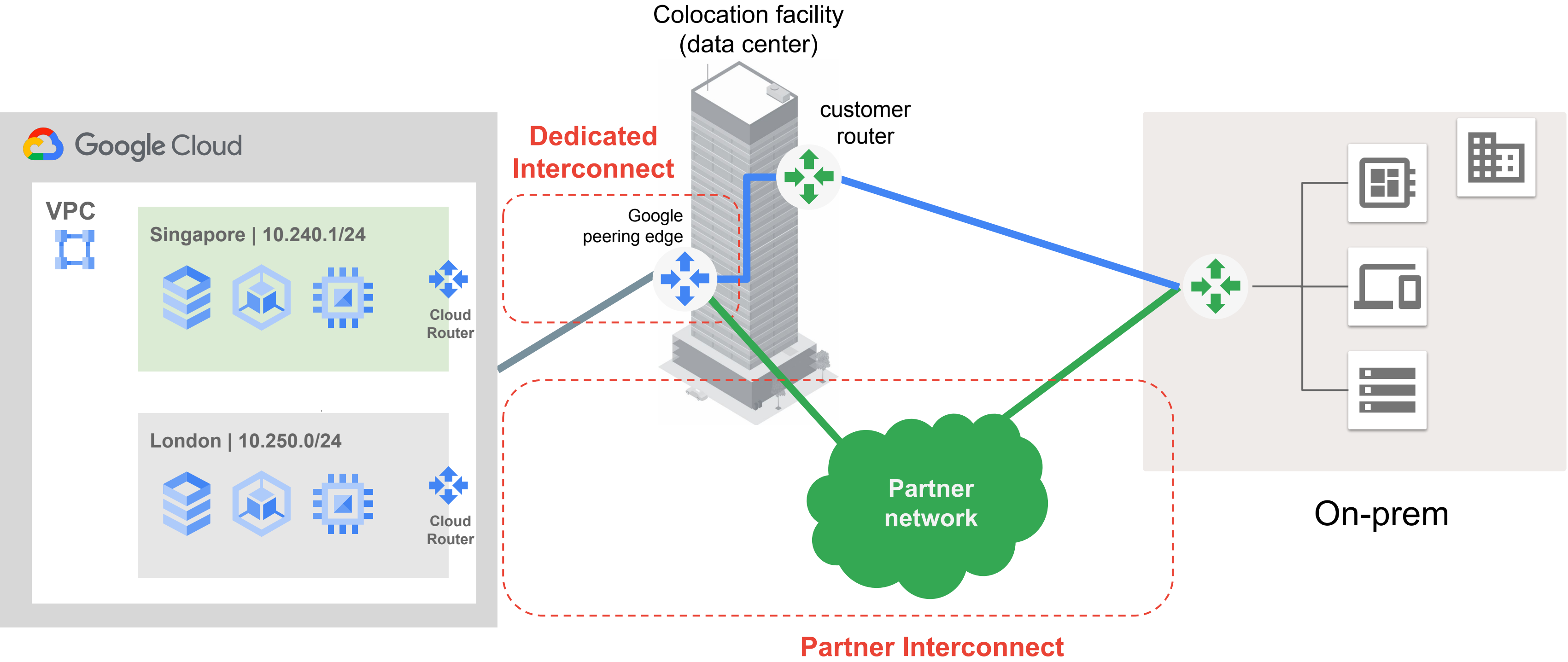
AVAILABLE IN
200+
COUNTRIES AND TERRITORIES

Connecting to Google Cloud



Google Cloud

Google Cloud Interconnect



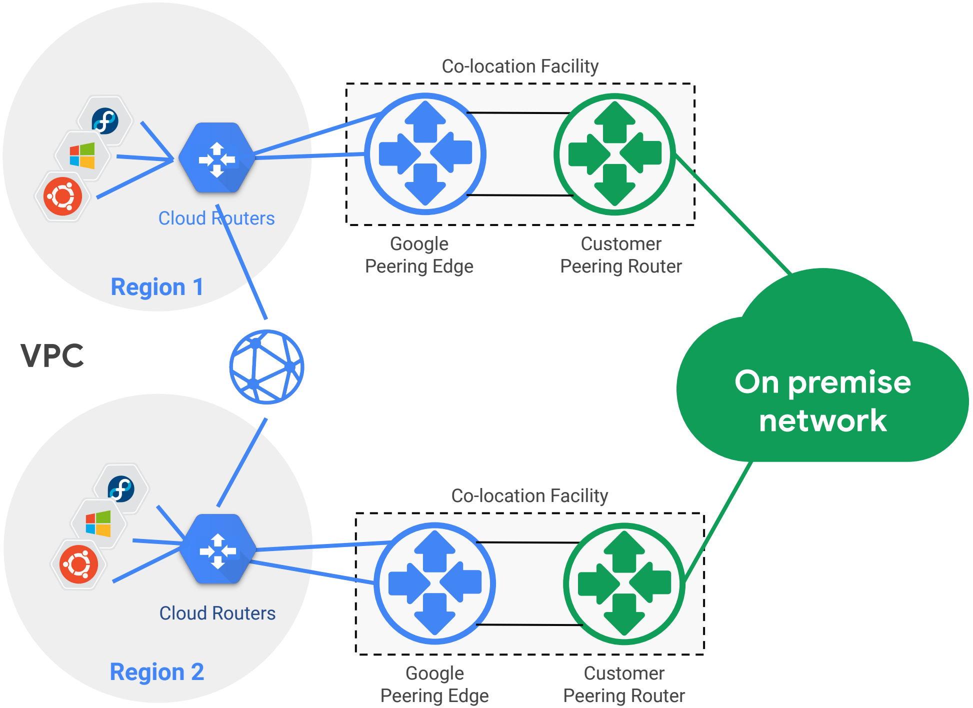
Network setup for high availability

99.99% Availability from **four interconnects** in **two metros**, to **two cloud regions**

99.9% Availability for **single region, single metro.**

BGP Graceful Restart between Cloud Router & Customer Router

Cloud Router's **Global Routing feature** will advertise routes to all projects



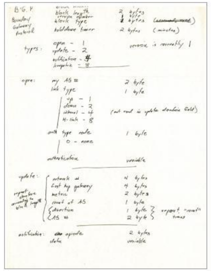
Google Cloud

BGP Peering and RPKI

BGP Hijack/Leak Outage

1989

BGP RFC 1105
Two Napkin Protocol,
Start Internet from 80K Host



B.G.P.
Boundary Gateway Protocol

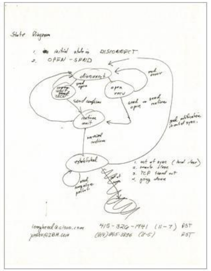
block length — 2 bytes
version number — 1 byte
block type — 1 bytes (~~[illegible]~~)
holddown timer — 2 bytes (minutes)

types: open - 1, update - 2, notification - 4, keepalive - 8 — version is currently 1

open: my AS # — 2 byte
link type — 1 byte
up - 1, down - 2, internal - 4, H-link - 8 (not used in update direction field)
auth type code — 1 byte
0 - none
authentication — variable

update: network # — 4 bytes
first hop gateway — 4 bytes
metric — 2 bytes
count of AS — 1 byte
direction — 1 byte
AS # — 2 byte
repeat "count" times
repeat structure according to block length

notification: opcode — 2 bytes
data — variable



lougheed@cisco.com 415-326-1941 (11-7) PST
yakov@IBM.com (914)945-3896 (8-5) EST

1997

AS7007 Incident
mistakenly (re)announces 72,000+ routes
Black Hole
Murphy's Law Strikes

2008

ISP
Accidentally announced Youtube Prefix for Block/ Black Hole Internal
BGP Hijack

2017

BGP Leak to Verizon
Accidentally 160K Prefixes, Major Impact Japan(KDDI/NTT Com)
BGP Leak

Visa/MasterCard and Major Bank
ISP leaks 36 Prefixes for Payment Service
FinTech Attacked

2018

Cryptocurrencies
ISP Hijack Amazon DNS Route 53, Targeting myetherwallet.com
FinTech BGP Hijack

2019

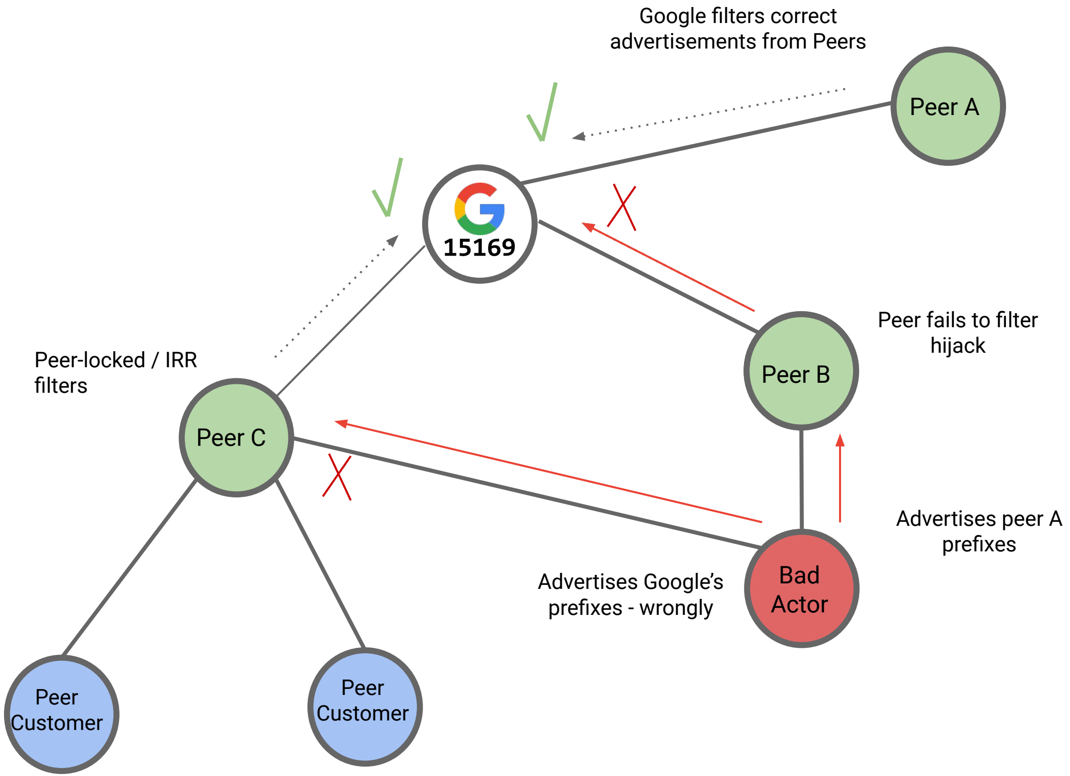
Verizon BGP leak
BGP optimizers make it worse
Cloudflare customer outages, AWS issues, Facebook outage.
BGP Leak

2020

ISP,
ISP hijacks internet traffic for Google, AWS, Cloudflare, and others
BGP Hijack

Google Peering, A better Internet for cloud

- The Internet remains key to Cloud customers
- Resilience of Internet routing needs to get better
- Enhanced monitoring and filtering
- Active community role (MANRS, Peer Lock)
- RPKI


Google filters correct advertisements from Peers
Peer A
15169
Peer B
Peer fails to filter hijack
Peer-locked / IRR filters
Peer C
Advertises peer A prefixes
Advertises Google's prefixes - wrongly
Bad Actor
Peer Customer
Peer Customer

Route filtering – validate all incoming routes

Multiple filtering mechanisms to address different kinds of BGP hijacks

Filtering based on routing data in public Internet Routing Registries (IRRs)

- filtering on **allowed peer announcements** (based on AS-SET expansion)
- challenge: IRR has high coverage, but data can be stale, invalid, contradictory
- rolled out widely across most ISP peering sessions in *reject* mode

Route origin validation (ROV) based on RPKI

- filtering on **allowed origin** (prevents many misconfiguration hijacks)
- challenge: lower coverage, e.g., < 40% IPv4 space is registered
- active filtering in pilot – targeted for rollout to most peering sessions

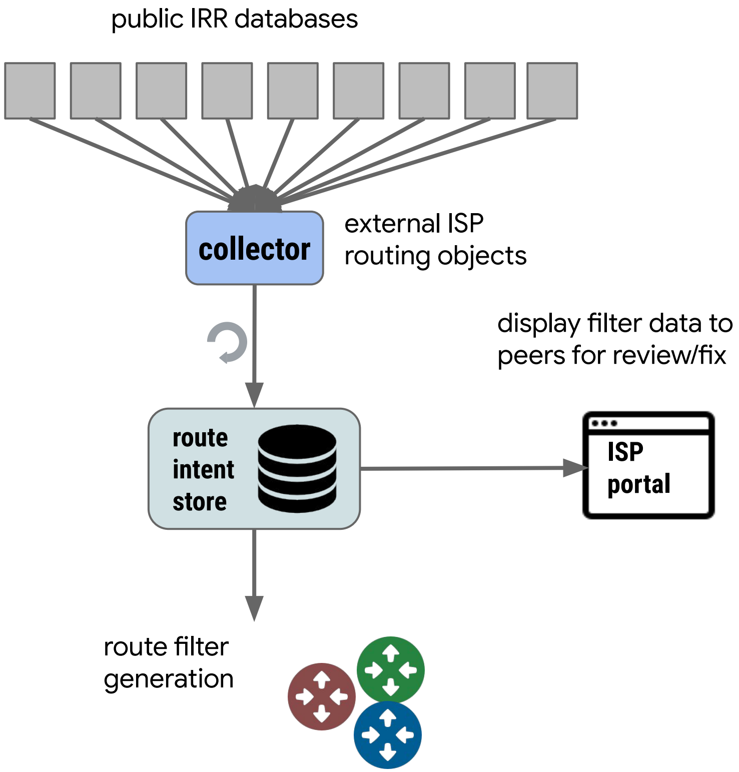
IRR-based route filtering

How IRR filtering works in Google:

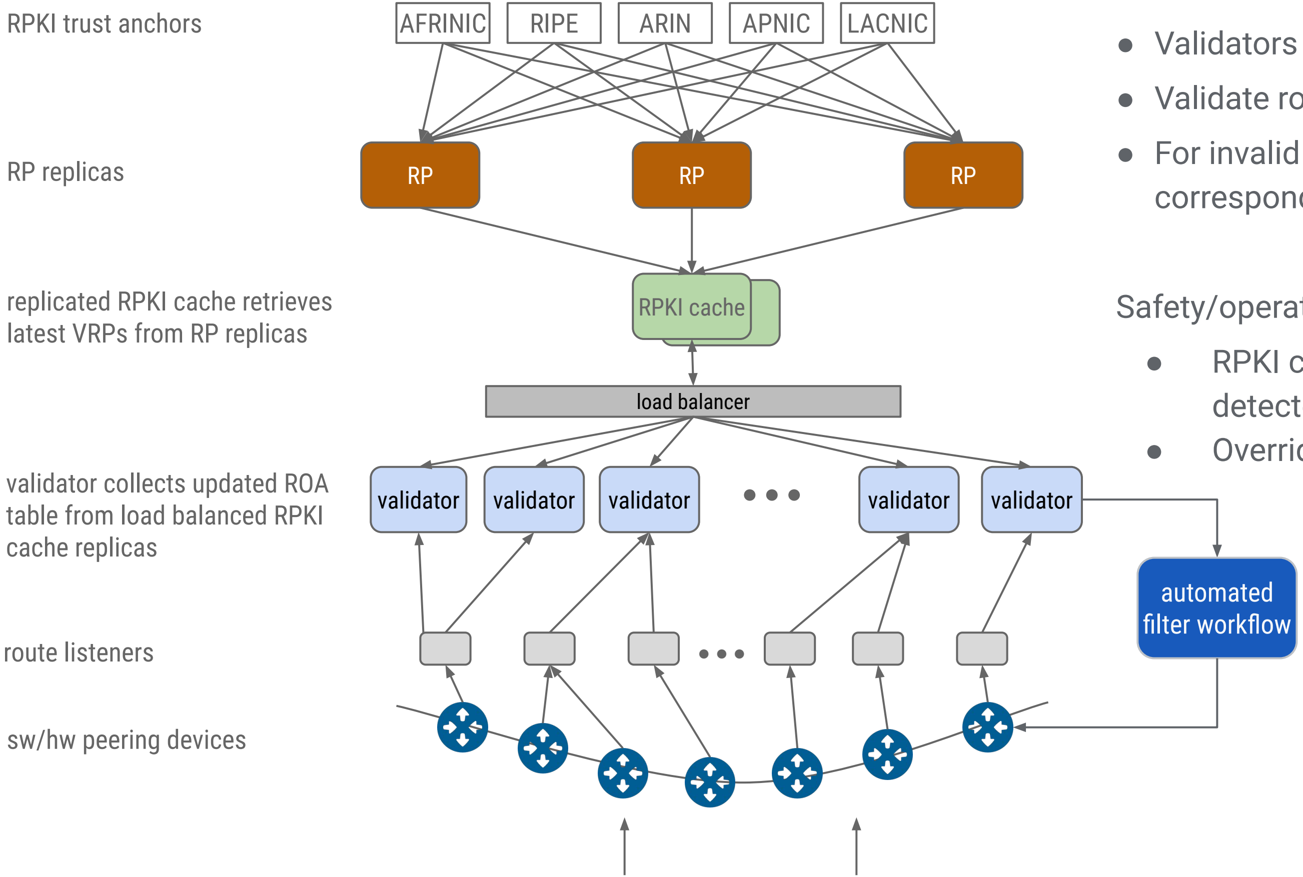
- collect and process public routing data (IRRs, peeringDB)
  - currently pull from ~25 IRRs
- build per-ASN *allow-list* of routes peers are expected to advertise
  - check for high traffic impact for any single ASN
  - check for connectivity via alternative routes
- rollout allow-lists on peering edge devices
- treat any received route that does not match the allow-list as invalid
  - *depref*: send traffic over alternate/transit routes (serves as grace period to update routing data)
  - *reject*: drop route (traffic must take alternate path)

Considerations

- subject to IRR data availability / accuracy
- need to consider filter scale on routers
- weekly rollout schedule for updated filters



RPKI origin validation



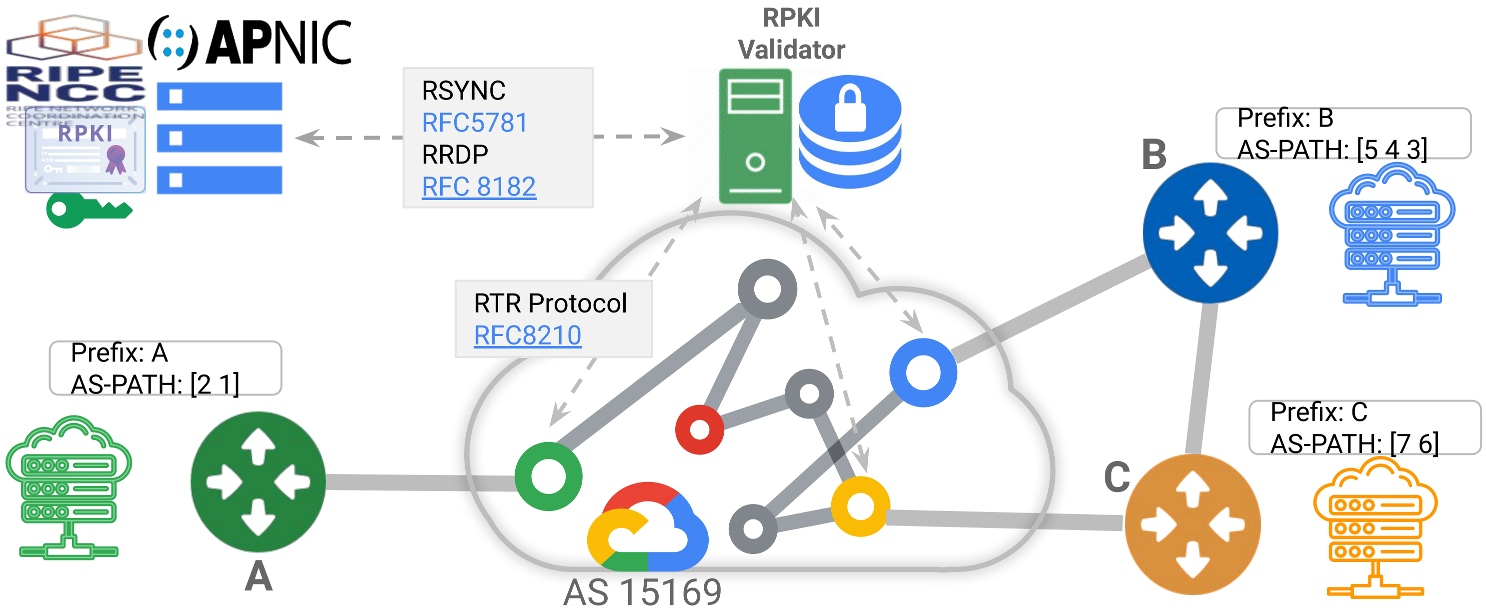
- Validators monitor routes at peering edge
- Validate routes against current RPKI cache
- For invalid routes, initiate filter installation on corresponding sessions

Safety/operational mechanisms:

- RPKI cache: fail-static if large changes detected from RPs
- Overrides via local RPKI additions

Google Cloud

Google Peering, RPKI, Very easy to fake



Prefix origin extended communities

"ext:4300:000000000000" = Origin valid
"ext:4300:000000000001" = Origin not found
"ext:4300:000000000002" = Origin invalid

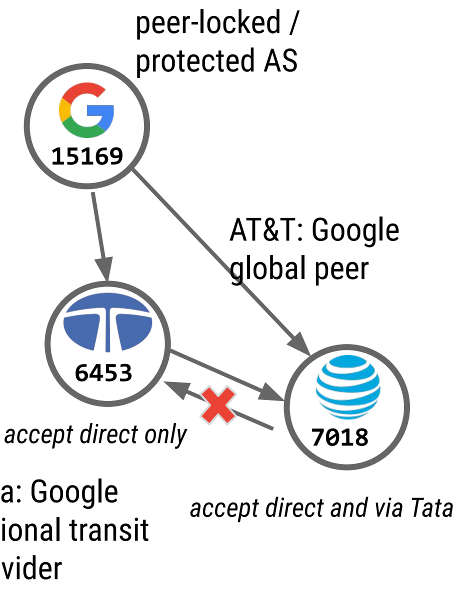
✓ Prefix: B
AS-PATH: [7 6 3]

✓ Prefix: B
AS-PATH: [7 5 4 3]

✗ Prefix: B
AS-PATH: [7 6]

1
2

Google Peering, RPKI with Peer Lock



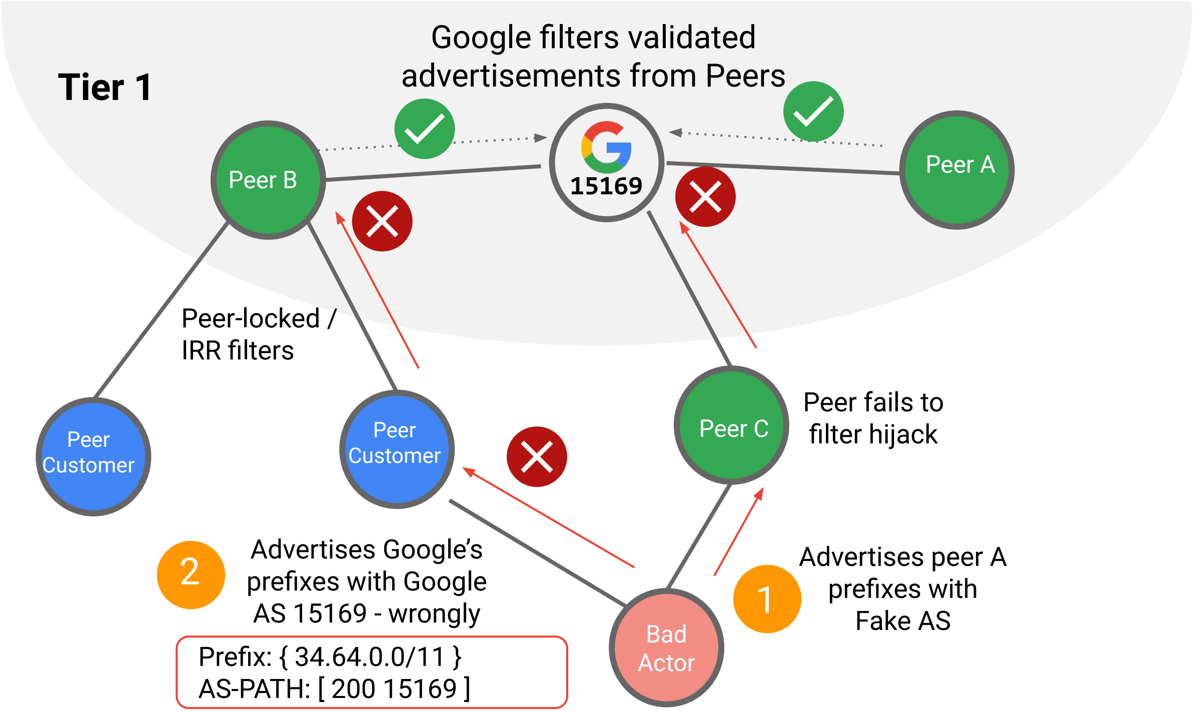
peer-locked / protected AS

AT&T: Google global peer

accept direct only

Tata: Google regional transit provider

accept direct and via Tata



Google filters validated advertisements from Peers

Peer-locked / IRR filters

Peer fails to filter hijack

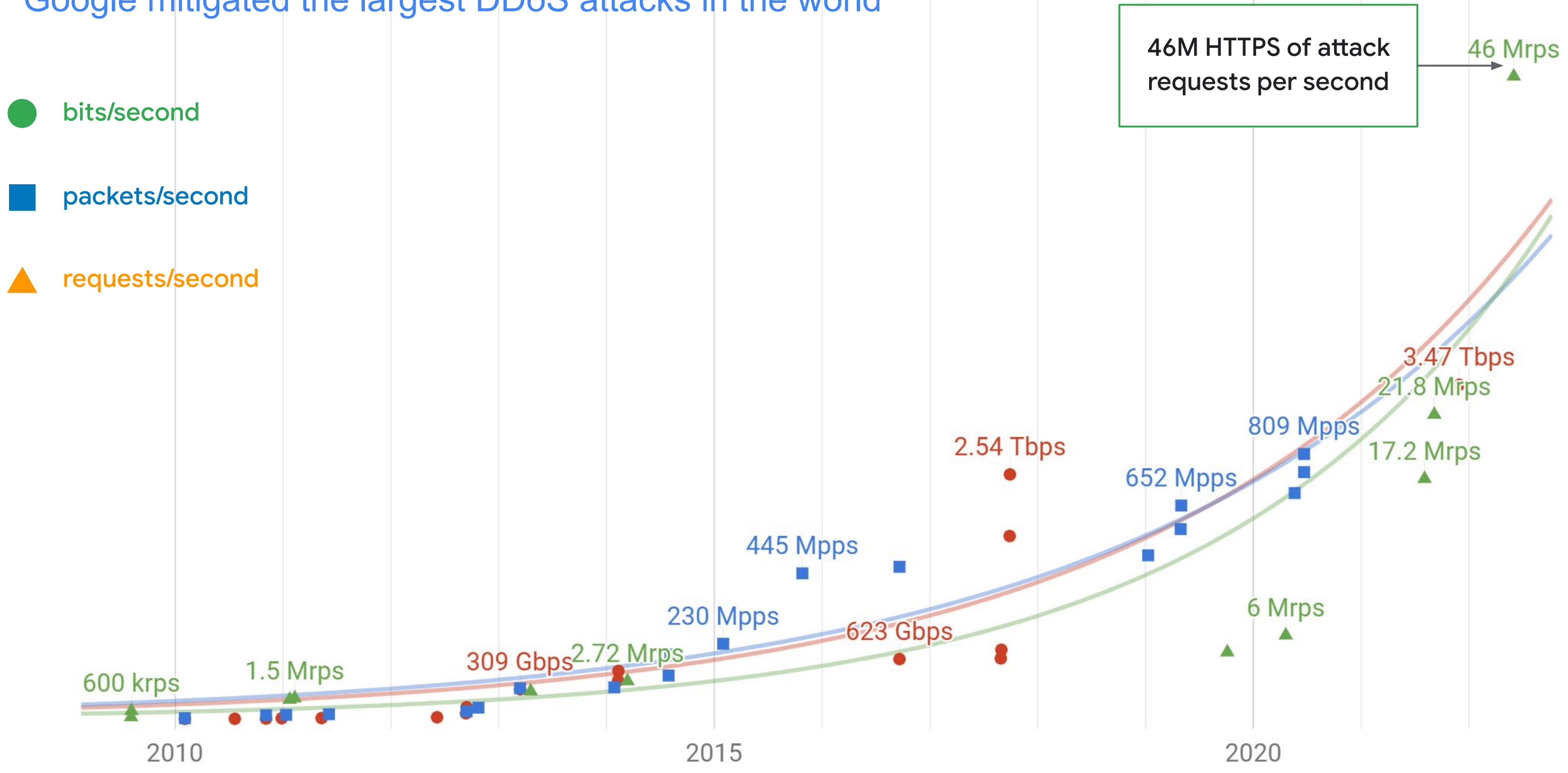
2 Advertises Google's prefixes with Google AS 15169 - wrongly

Prefix: { 34.64.0.0/11 }
AS-PATH: [200 15169]

1 Advertises peer A prefixes with Fake AS

Protection at Planetary Scale

Google mitigated the largest DDoS attacks in the world



46M HTTPS of attack requests per second

Multiple solutions required

Publish route intent

- Register Google/GCP routes in public registries
- Enables other networks to validate Google routes
- Prevents propagation of hijacks; **protects connectivity *to* Google**

Validate received route announcements

- Work with peers and customers to properly register routes
- Deploy filtering systems to accept only valid routes
- Prevents accepting bogus routes; **protects connectivity *from* Google/GCP**

Detect disruptions in the Internet

- Deploy first- and third-party monitoring systems to alert on hijacks in external networks
- Proactively mitigate when significant problems are detected
- Reduces repair time, but often depends on actions by external networks

Accelerate progress via collaborations

- Leverage MANRS as a collaboration vehicle
- Work with other providers to align on common solutions and policies
- Share experience and information via MANRS forums

Thank you!

Google Cloud