Packet Clearing House

Best Practices in DNS Service-Provision Architecture

Version 1.2

Bill Woodcock

Packet Clearing House

Packet Clearing House

Nearly all DNS is Anycast

Large ISPs have been anycasting recursive DNS servers for more than twenty years.

Which is a very long time, in Internet years.

All but one of the root nameservers are anycast.

All the large gTLDs are anycast.

Reasons for Anycast

Transparent fail-over redundancy

Latency reduction

Load balancing

Attack mitigation

Configuration simplicity (for end users) or lack of IP addresses (for the root)

No Free Lunch

The two largest benefits, fail-over redundancy and latency reduction, both require a bit of work to operate as you'd wish.

Packet Clearing House

Fail-Over Redundancy

DNS resolvers have their own fail-over mechanism, which works... um... okay.

Anycast is a very large hammer.

Good deployments allow these two mechanisms to reinforce each other, rather than allowing anycast to foil the resolvers' fail-over mechanism.

Packet Clearing House

Resolvers' Fail-Over Mechanism

DNS resolvers like those in your computers, and in referring authoritative servers, can and often do maintain a *list* of nameservers to which they'll send queries.

Resolver implementations differ in how they use that list, but basically, when a server doesn't reply in a timely fashion, resolvers will try another server from the list.

Anycast Fail-Over Mechanism

Anycast is simply layer-3 routing.

A resolver's query will be routed to the topologically nearest instance of the anycast server visible in the routing table.

Anycast servers govern their own visibility.

Latency depends upon the delays imposed by that topologically short path.

Conflict Between These Mechanisms

Resolvers measure by latency.

Anycast measures by hop-count.

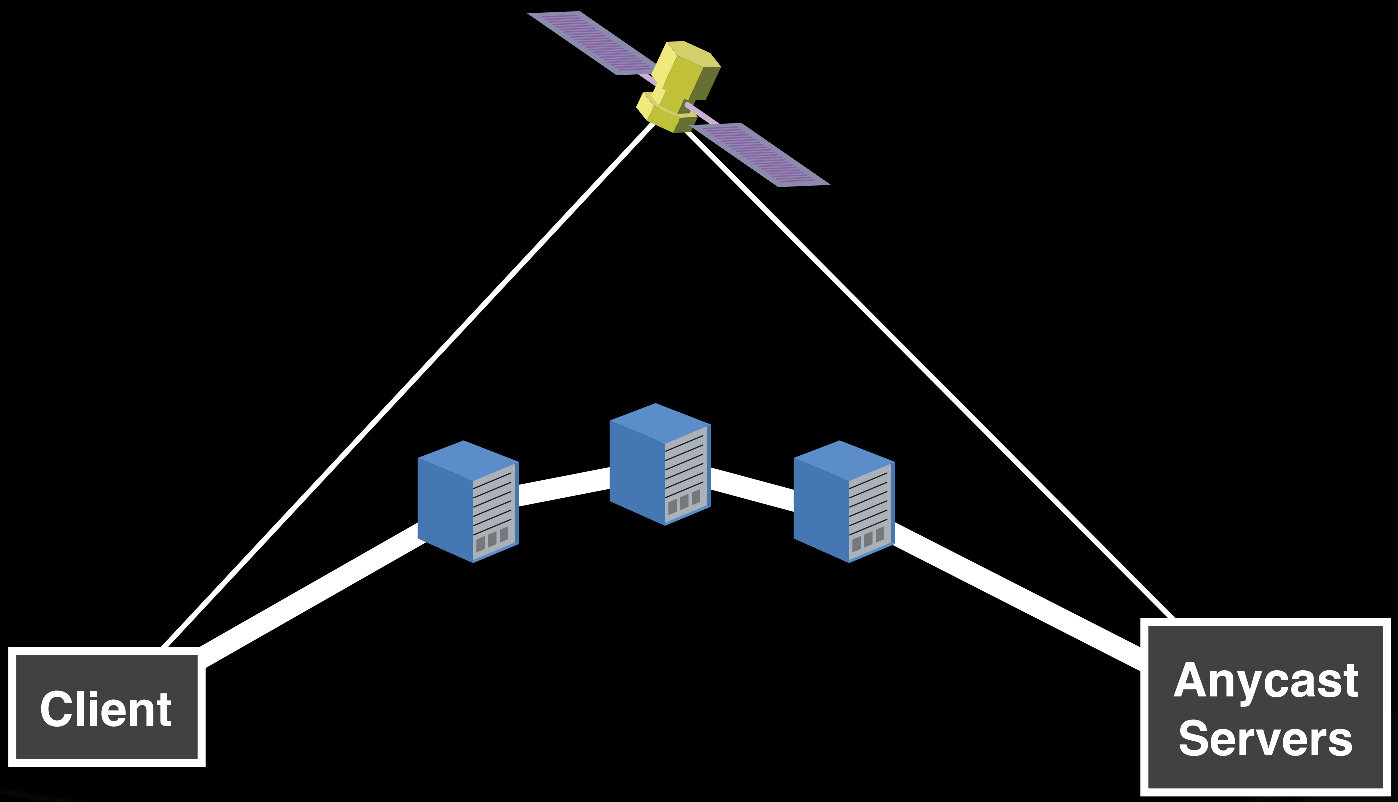
They don't necessarily yield the same answer.

Anycast always trumps resolvers, if it's allowed to.

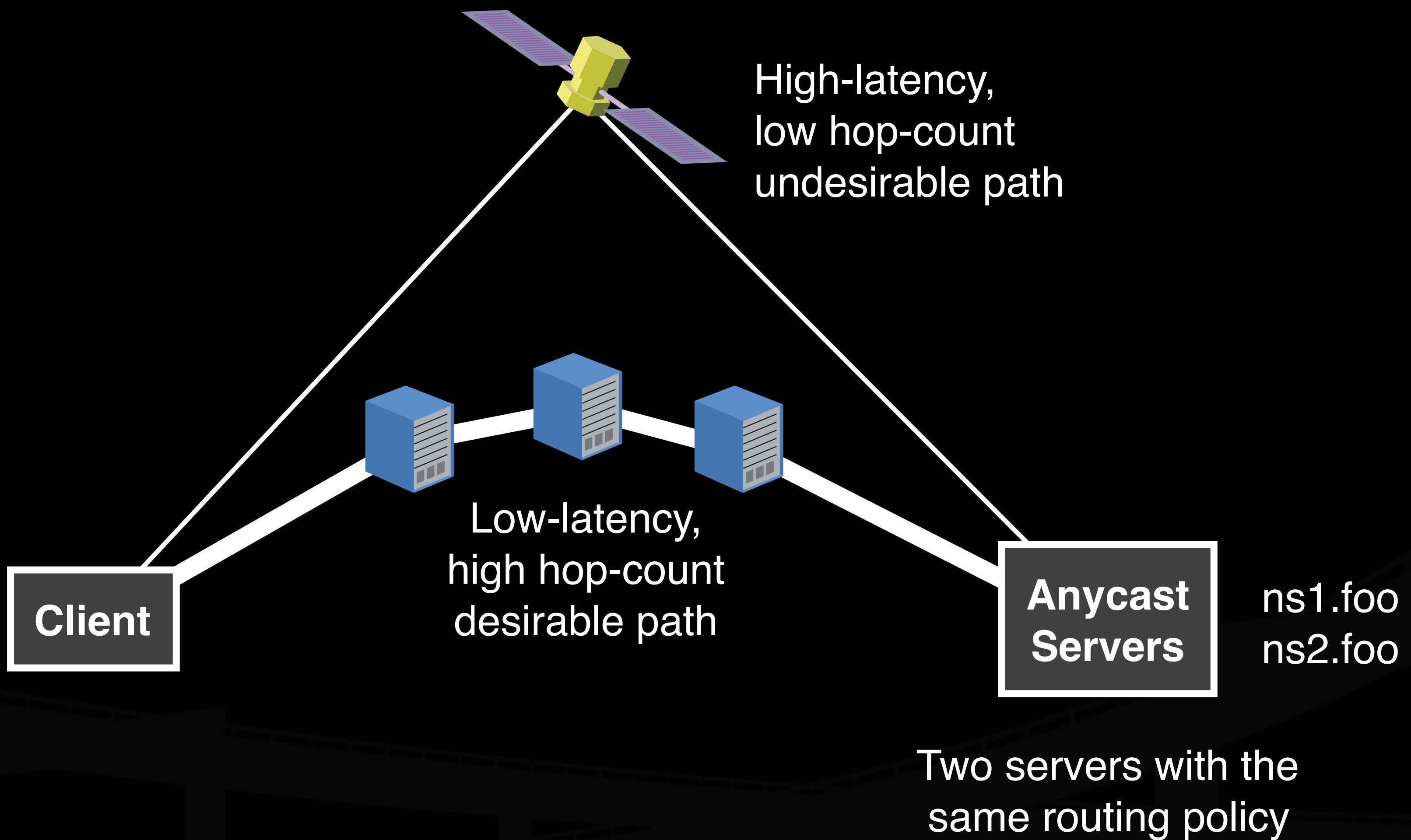
Neither the DNS service provider nor the user are likely to care about hop-count.

Both care a great deal about latency.

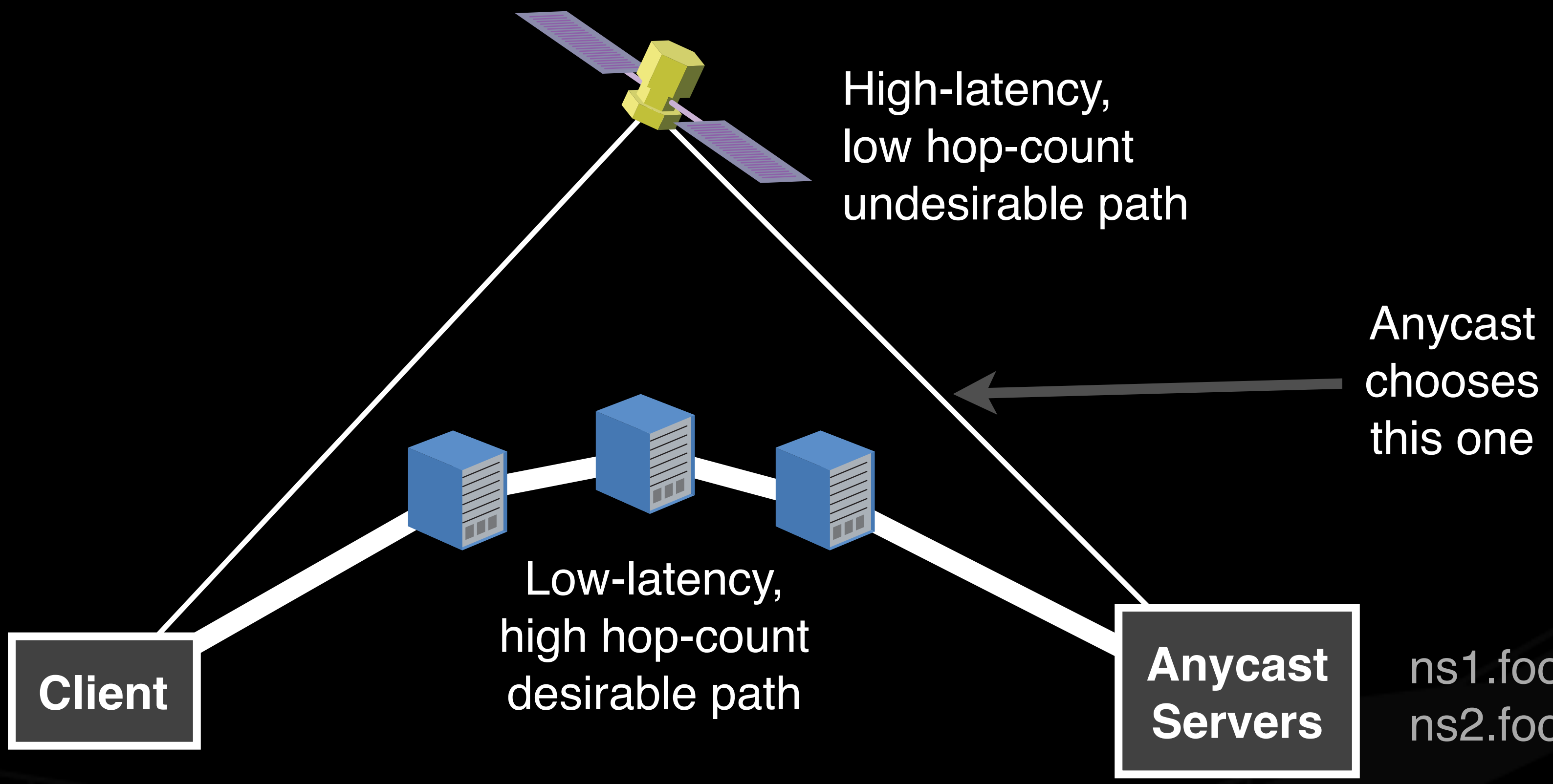
How The Conflict Plays Out



How The Conflict Plays Out

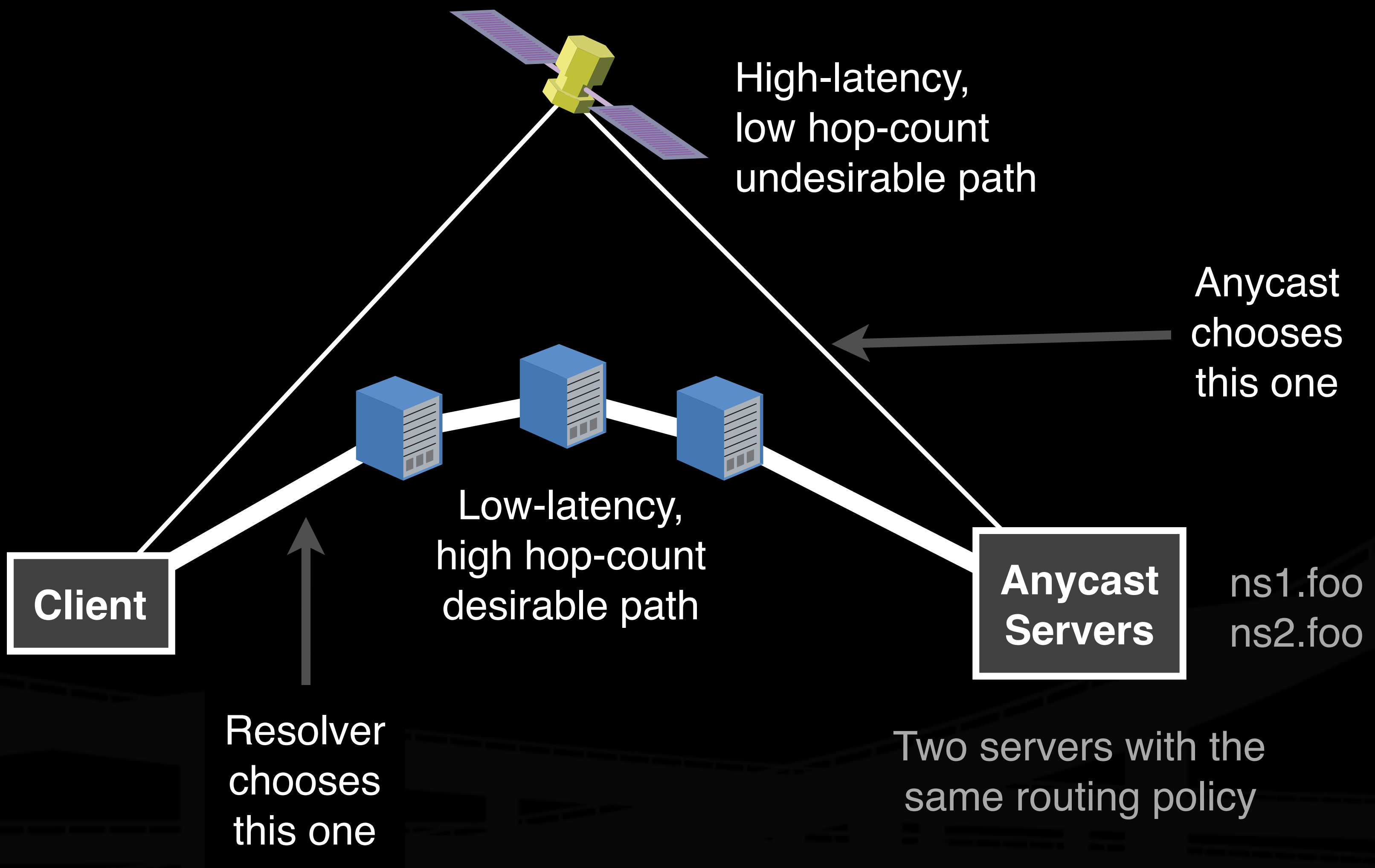


How The Conflict Plays Out



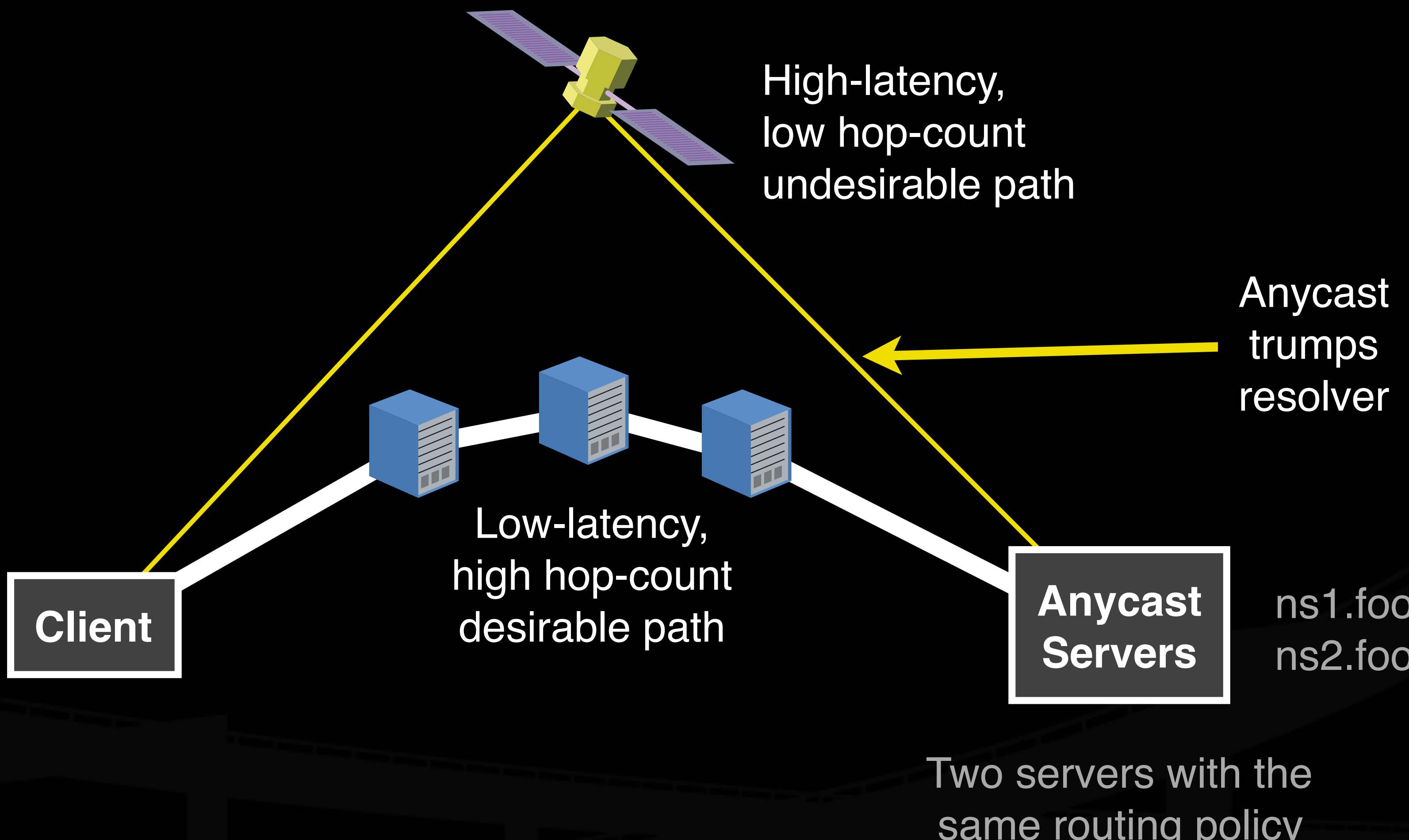
Two servers with the same routing policy

How The Conflict Plays Out

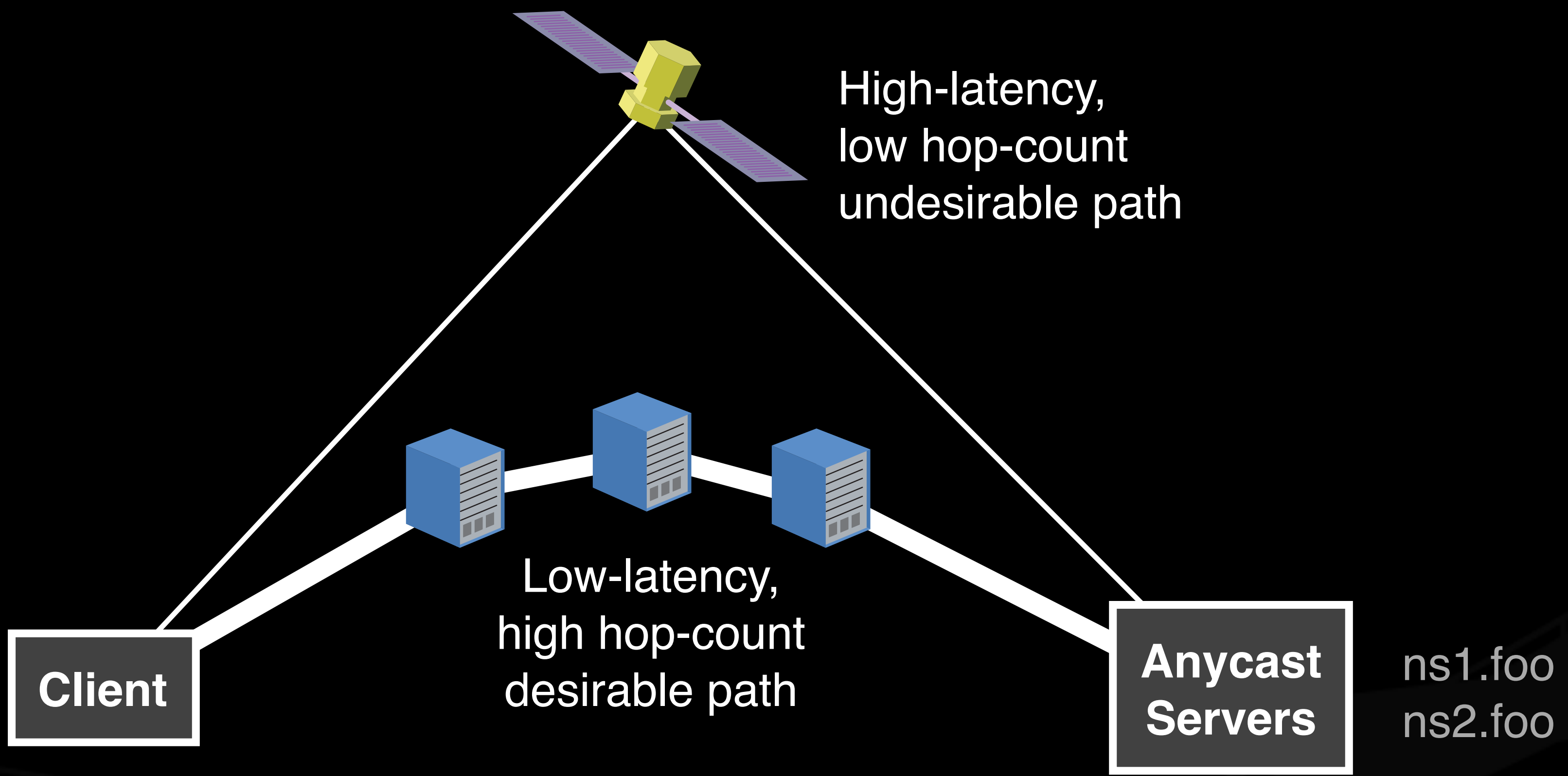


Packet Clearing House

How The Conflict Plays Out

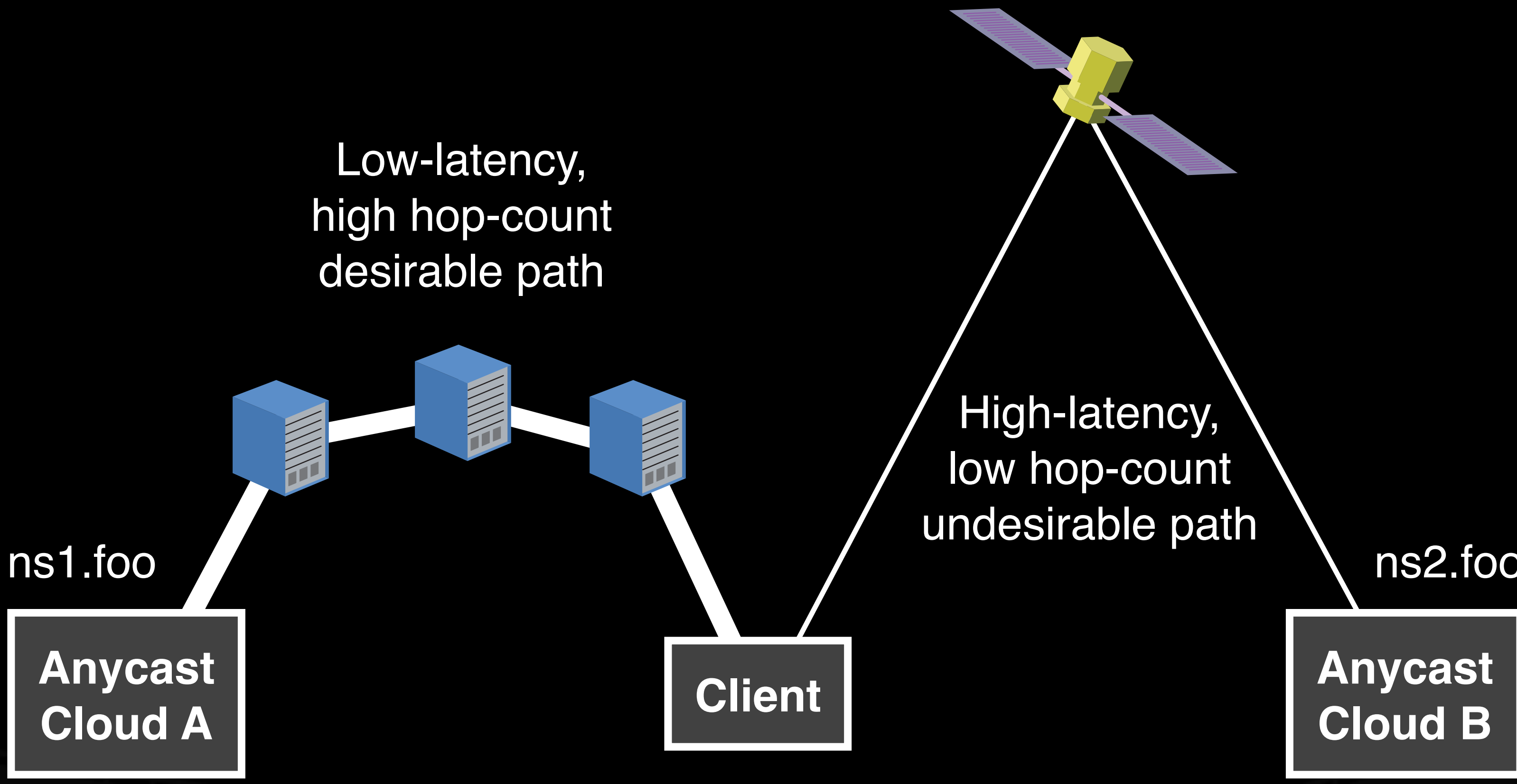


Resolve the Conflict



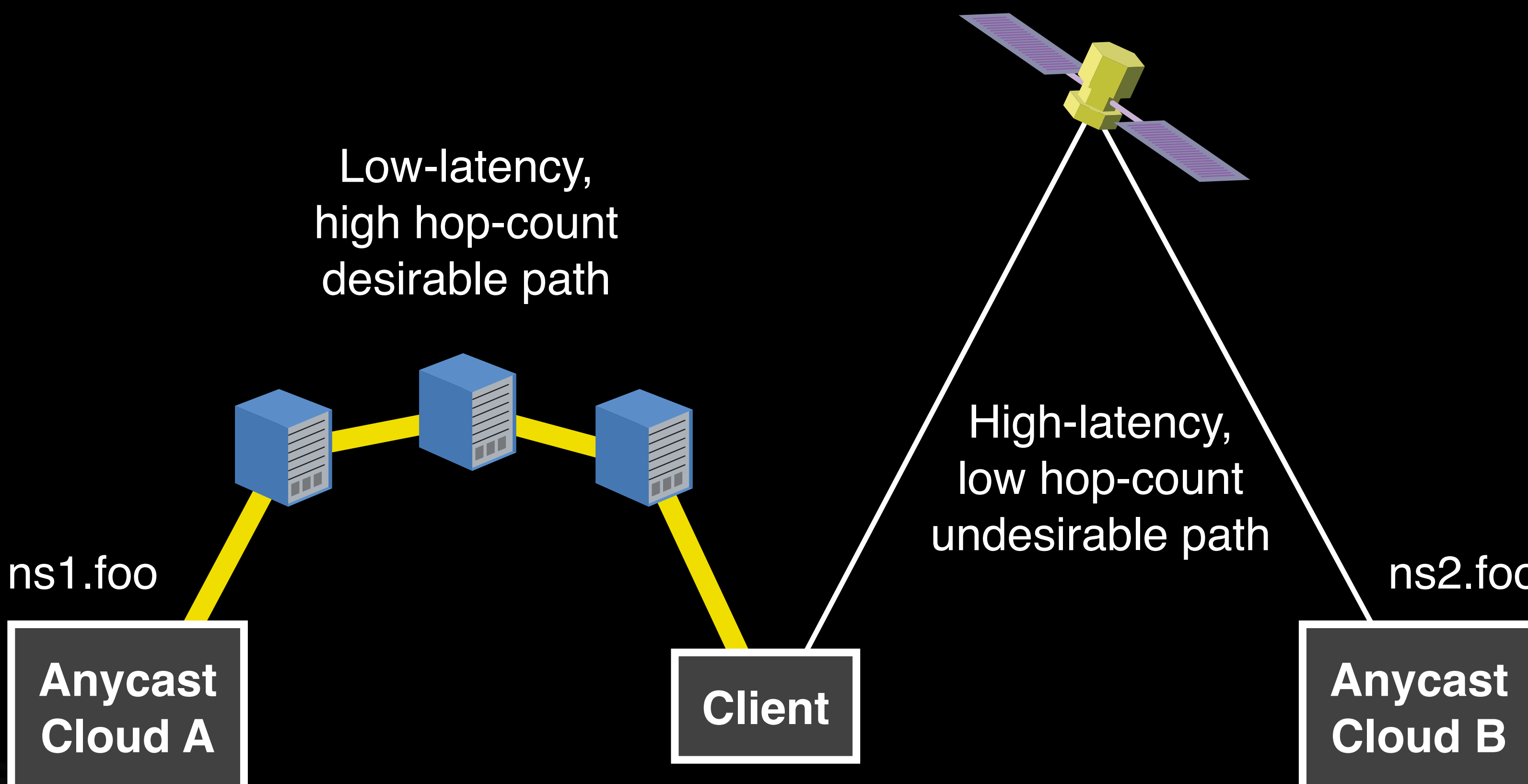
The resolver uses different IP addresses for its fail-over mechanism, while anycast uses the same IP addresses.

Resolve the Conflict



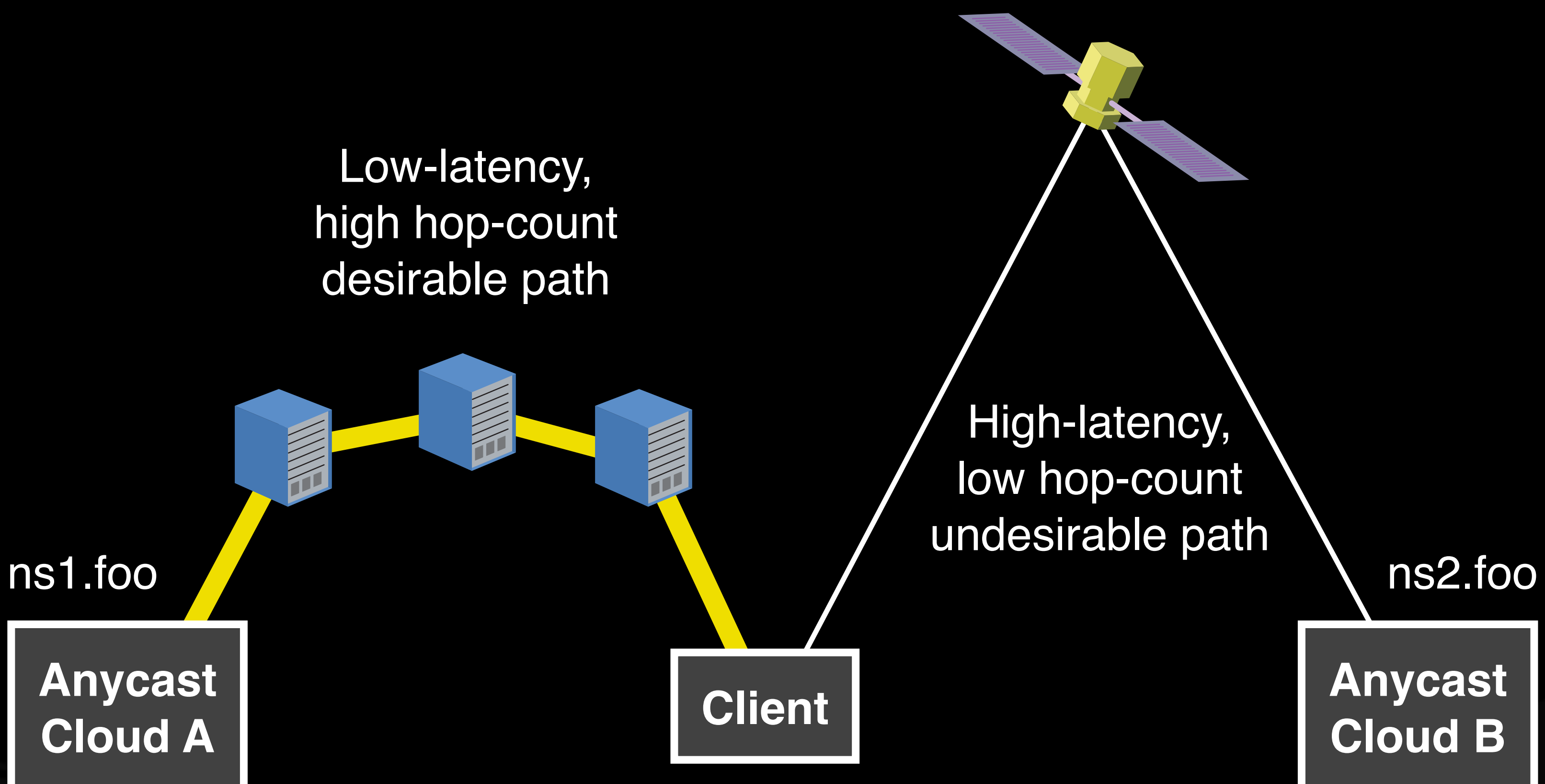
Split the anycast deployment into “clouds” of locations, each cloud using a different IP address and different routing policies.

Resolve the Conflict



This allows anycast to present the nearest servers,
and allows the resolver to choose the one which performs best.

Resolve the Conflict



These clouds are usually referred to as “A Cloud” and “B Cloud.” The number of clouds depends on stability and scale trade-offs.

Latency Reduction

Latency reduction depends upon the native layer-3 routing of the Internet.

The theory is that the Internet will deliver packets using the shortest path.

The reality is that the Internet will deliver packets according to ISPs' policies.

Latency Reduction

ISPs’ routing policies differ from shortest-path where there’s an economic incentive to deliver by a longer path.

ISPs’ Economic Incentives (Grossly Simplified)

ISPs have high cost to deliver traffic through transit.

ISPs have a low cost to deliver traffic through their peering.

ISPs receive money when they deliver traffic to their customers.

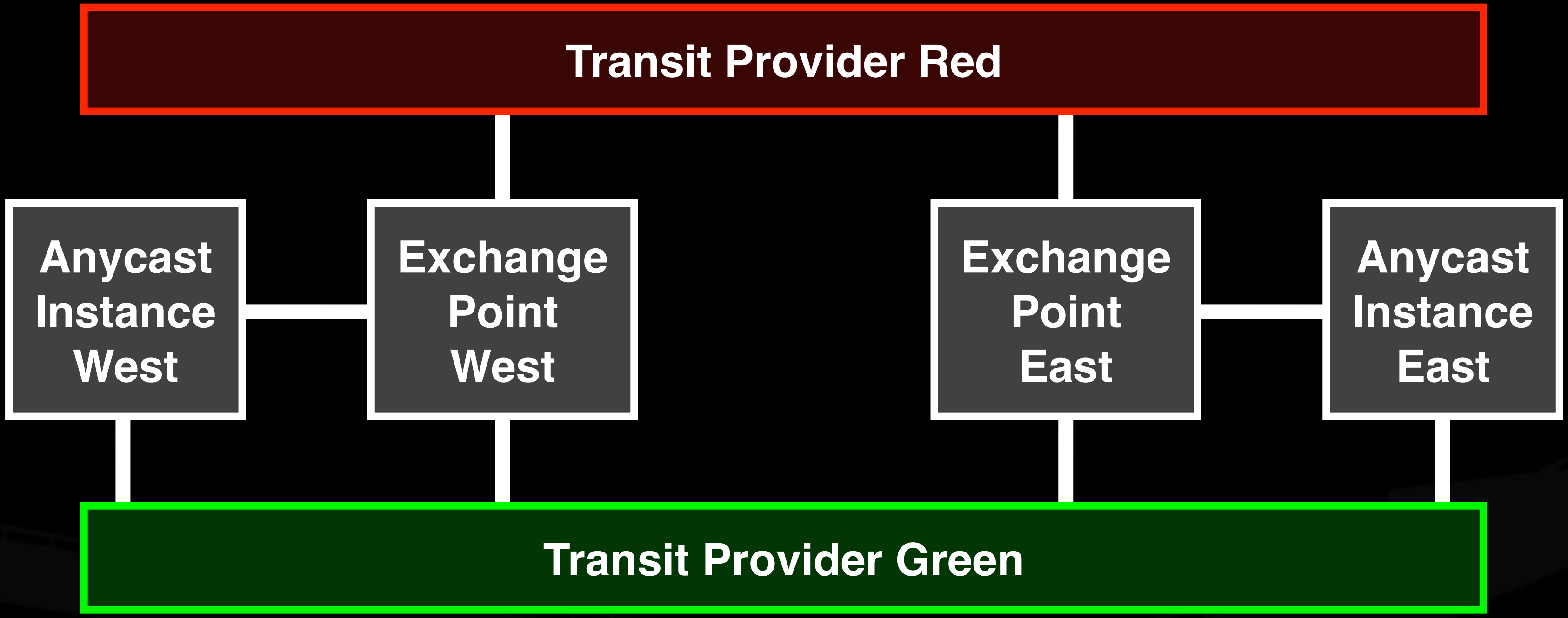
ISPs’ Economic Incentives (Grossly Simplified)

Therefore, ISPs will deliver traffic to a customer across a longer path, before by peering or transit across a shorter path.

If you are both a customer, and a customer of a peer or transit provider, this has important implications.

Normal Hot-Potato Routing

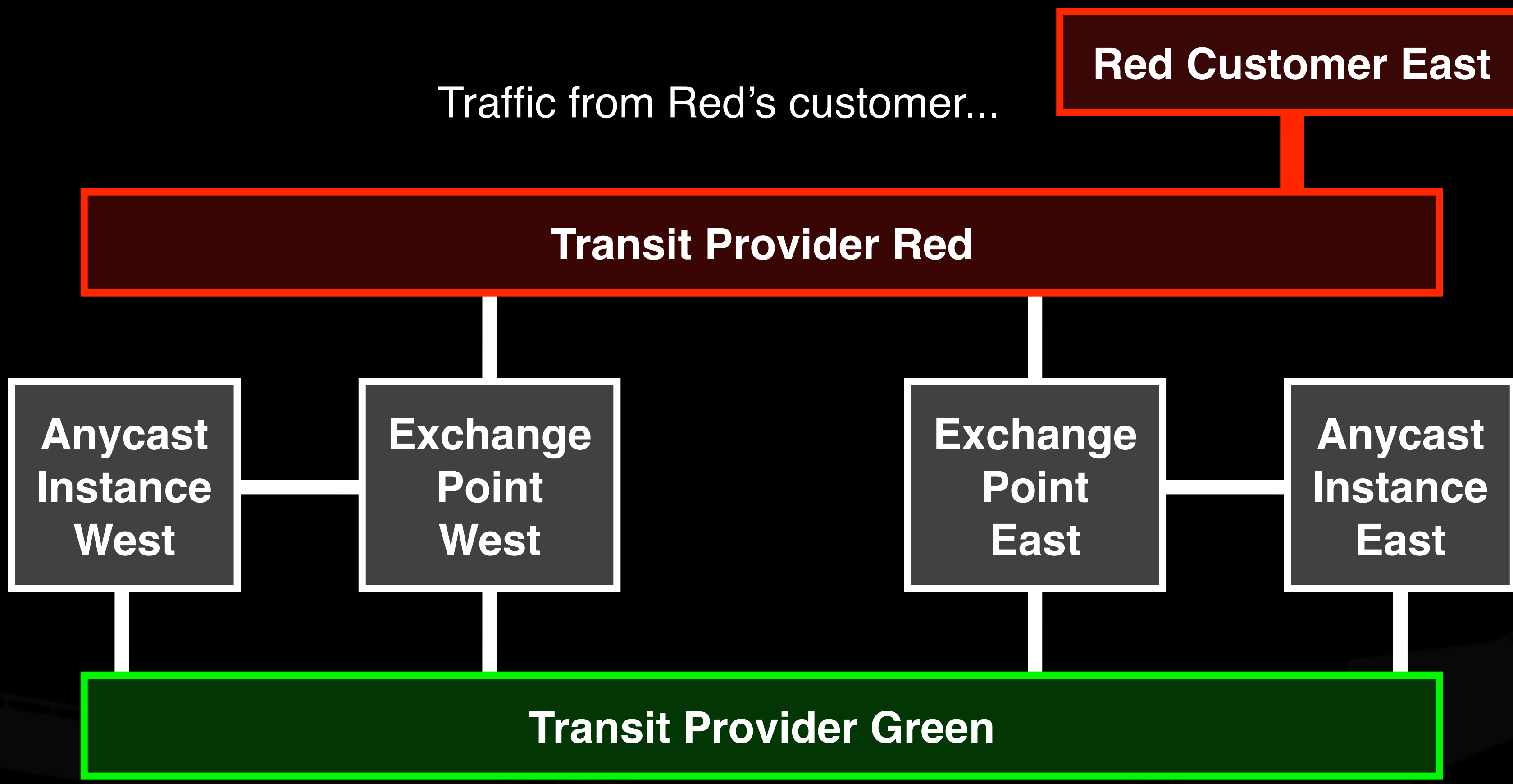
If the anycast network is not a customer of large Transit Provider Red...



...but is a customer of large Transit Provider Green...

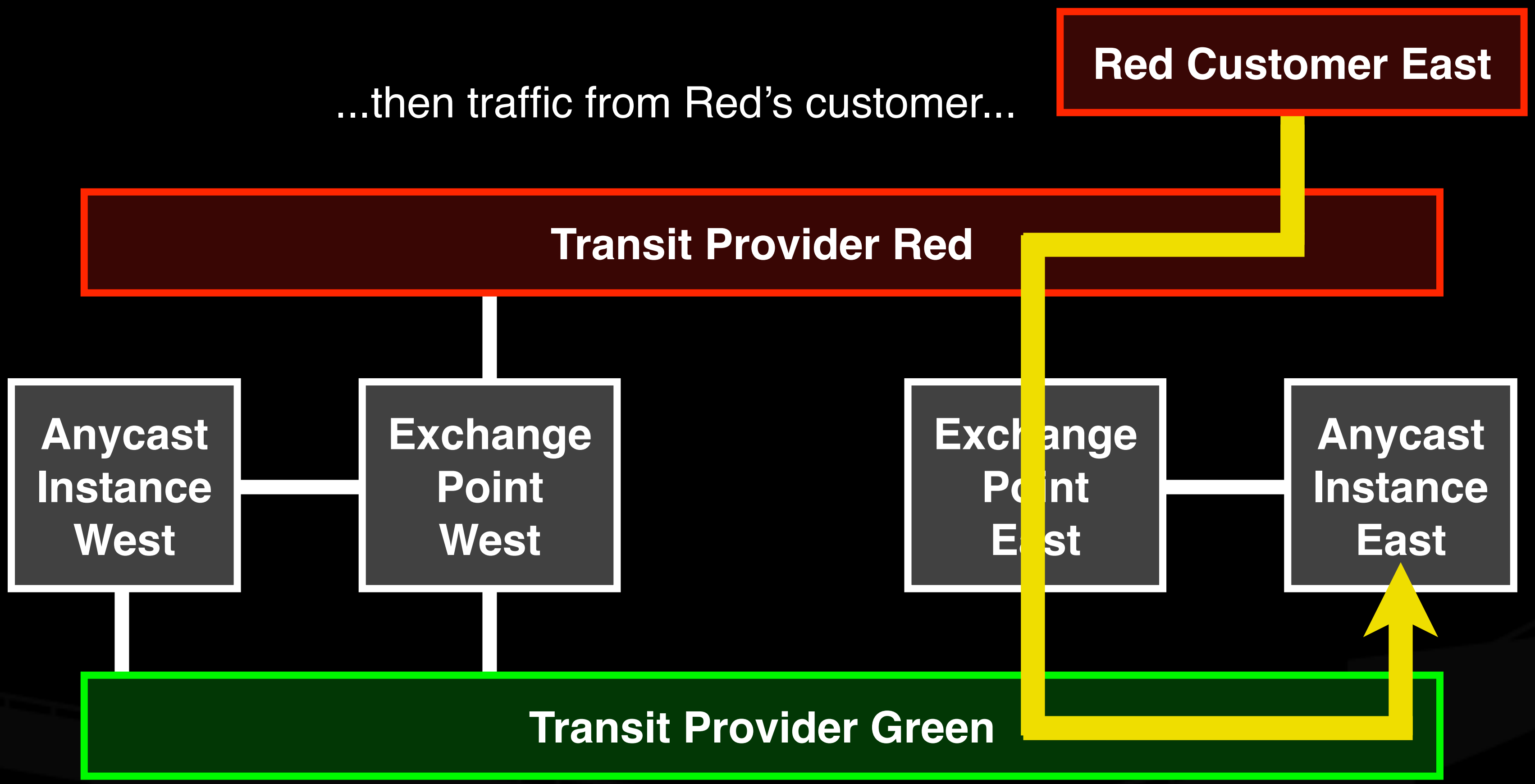
Packet Clearing House

Normal Hot-Potato Routing



Normal Hot-Potato Routing

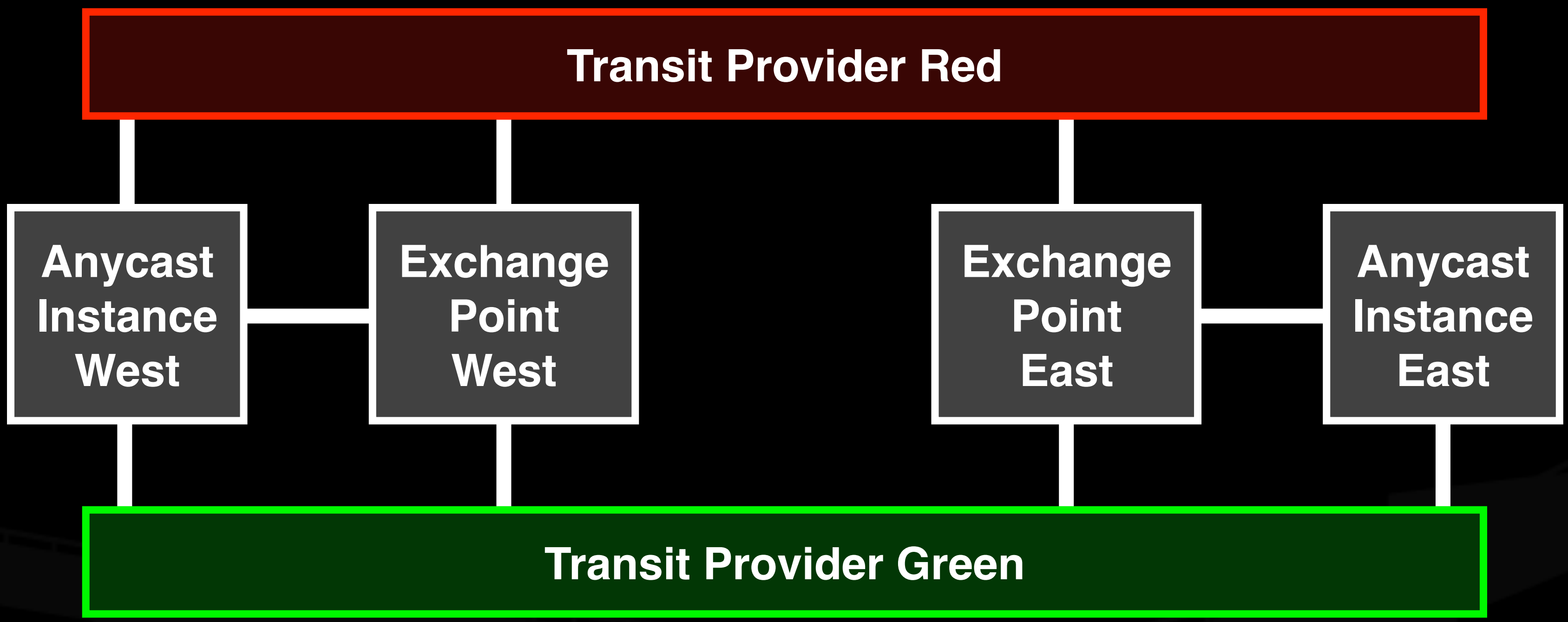
...then traffic from Red's customer...



...is delivered from Red to Green via local peering, and reaches the local anycast instance.

How the Conflict Plays Out

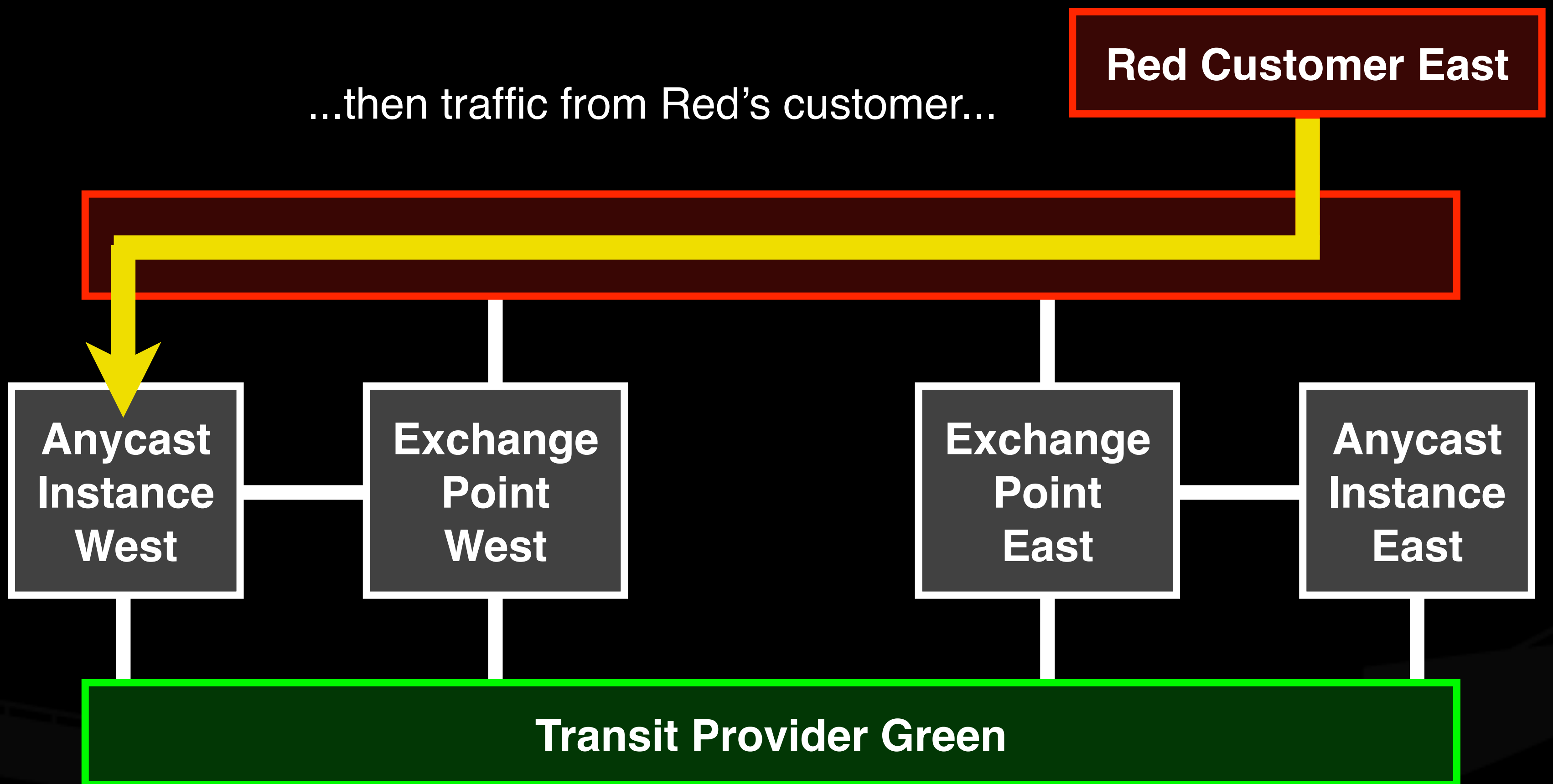
But if the anycast network is a customer of both large Transit Provider Red...



...and of large Transit Provider Green, but not at all locations...

How the Conflict Plays Out

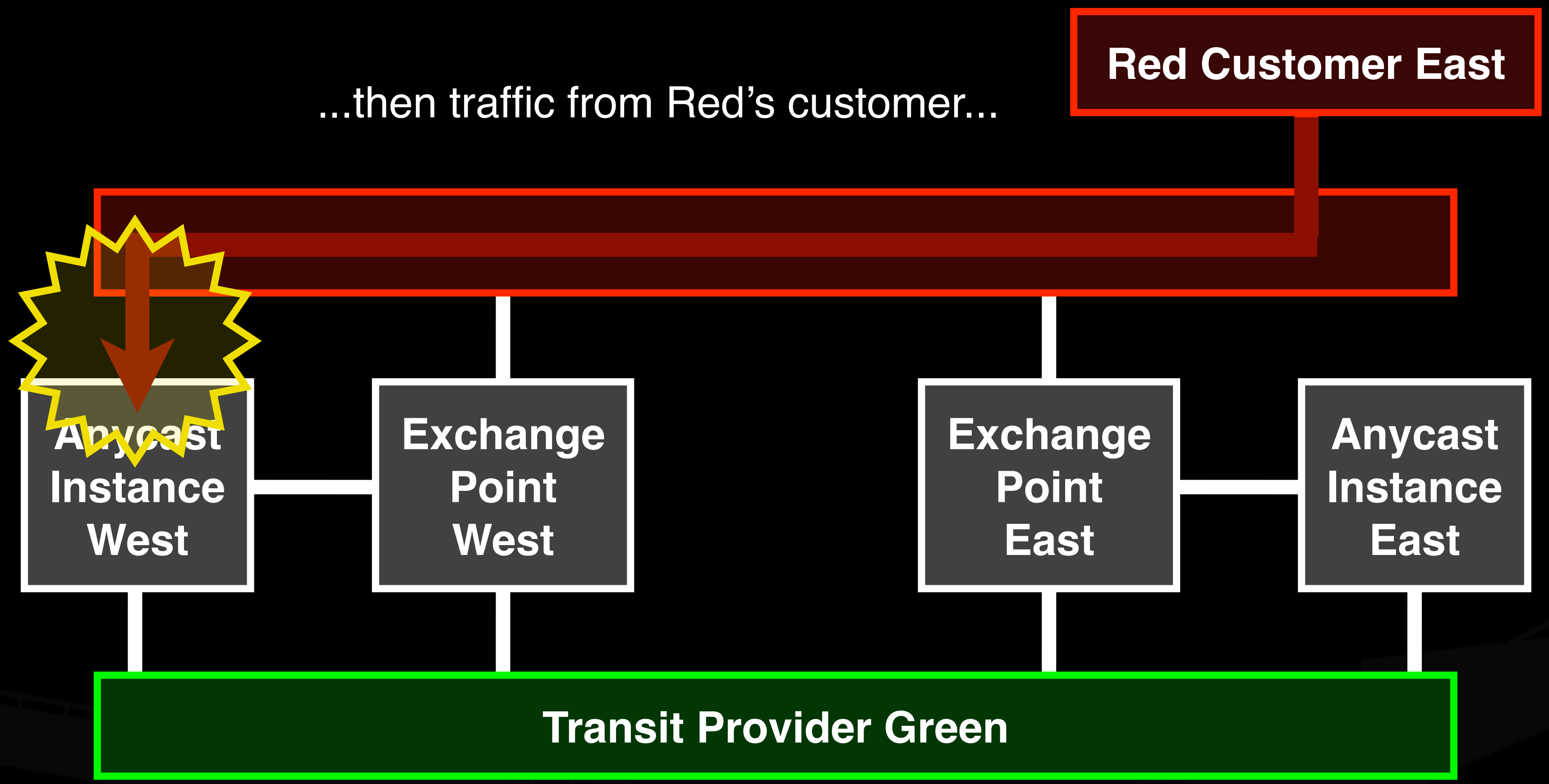
...then traffic from Red's customer...



...will be misdelivered to the remote anycast instance...

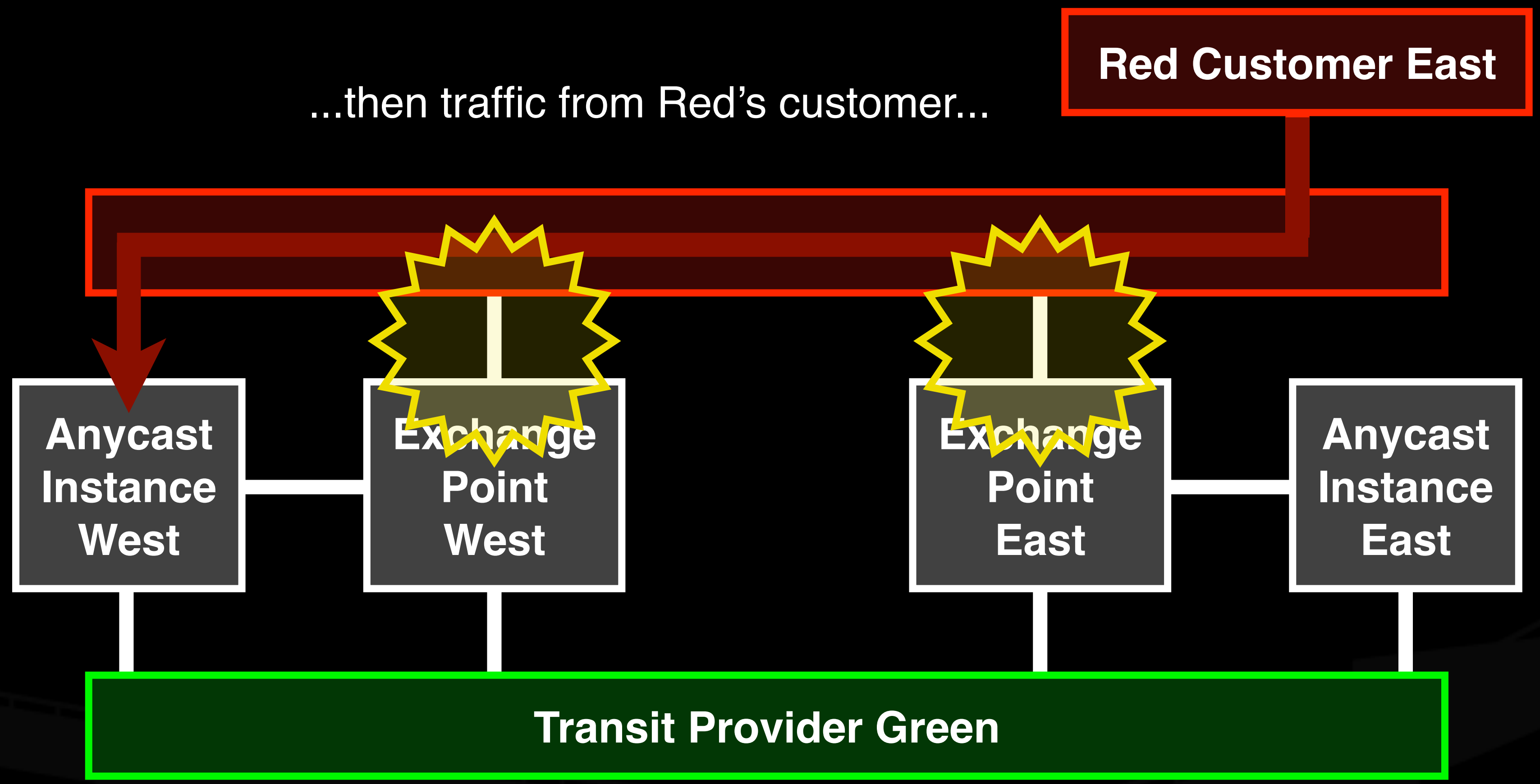
PCH
Packet Clearing House

How the Conflict Plays Out



...will be misdelivered to the remote anycast instance, because a customer connection...

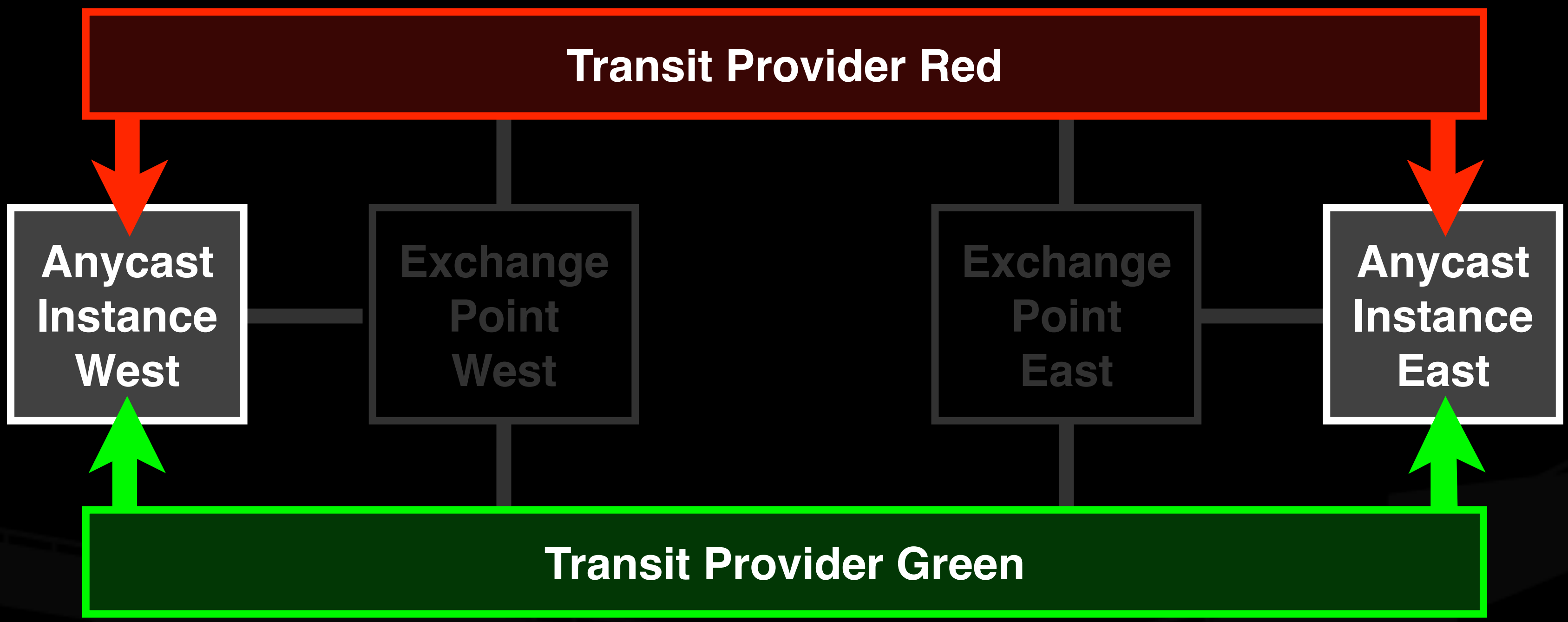
How the Conflict Plays Out



...will be misdelivered to the remote anycast instance, because a customer connection is preferred for economic reasons over a peering connection.

Resolve the Conflict

Any two instances of an anycast service IP address must have the same set of large transit providers at all locations.

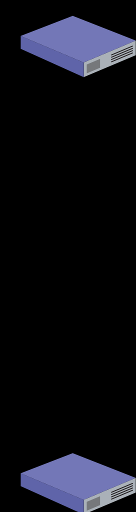


This caution is not necessary with small transit providers who don't have the capability of backhauling traffic to the wrong region on the basis of policy.

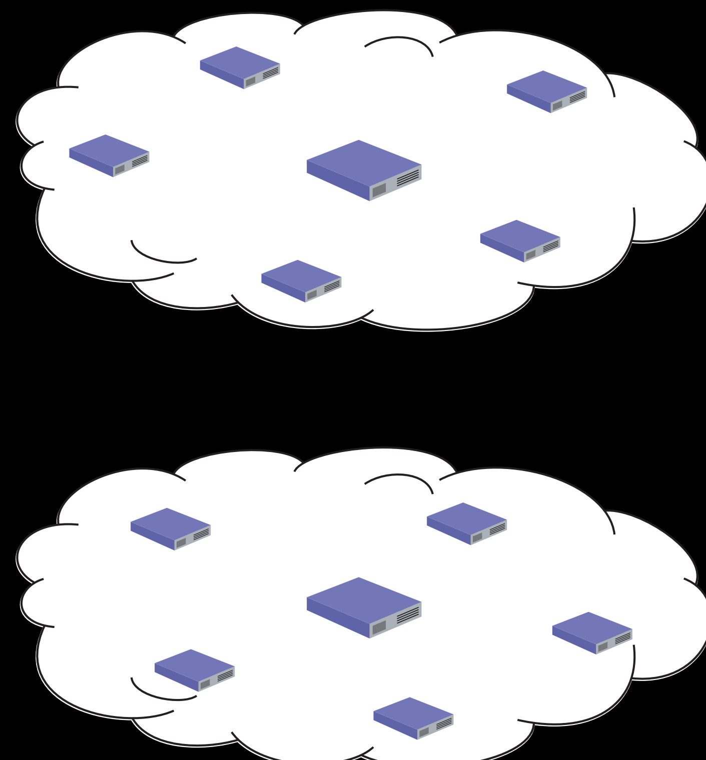
Putting the Pieces Together

- We need an A Cloud and a B Cloud.
- We need a redundant pair of the same transit providers at most or all instances of each cloud.
- We need a redundant pair of hidden masters for the DNS servers.
- We need a network topology to carry control and synchronization traffic between the nodes.

Redundant Hidden Masters

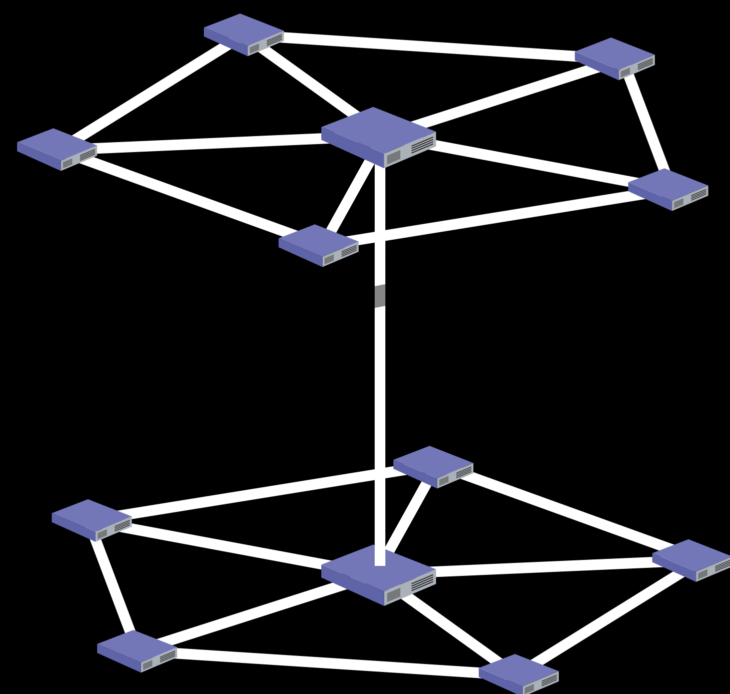


An A Cloud and a B Cloud



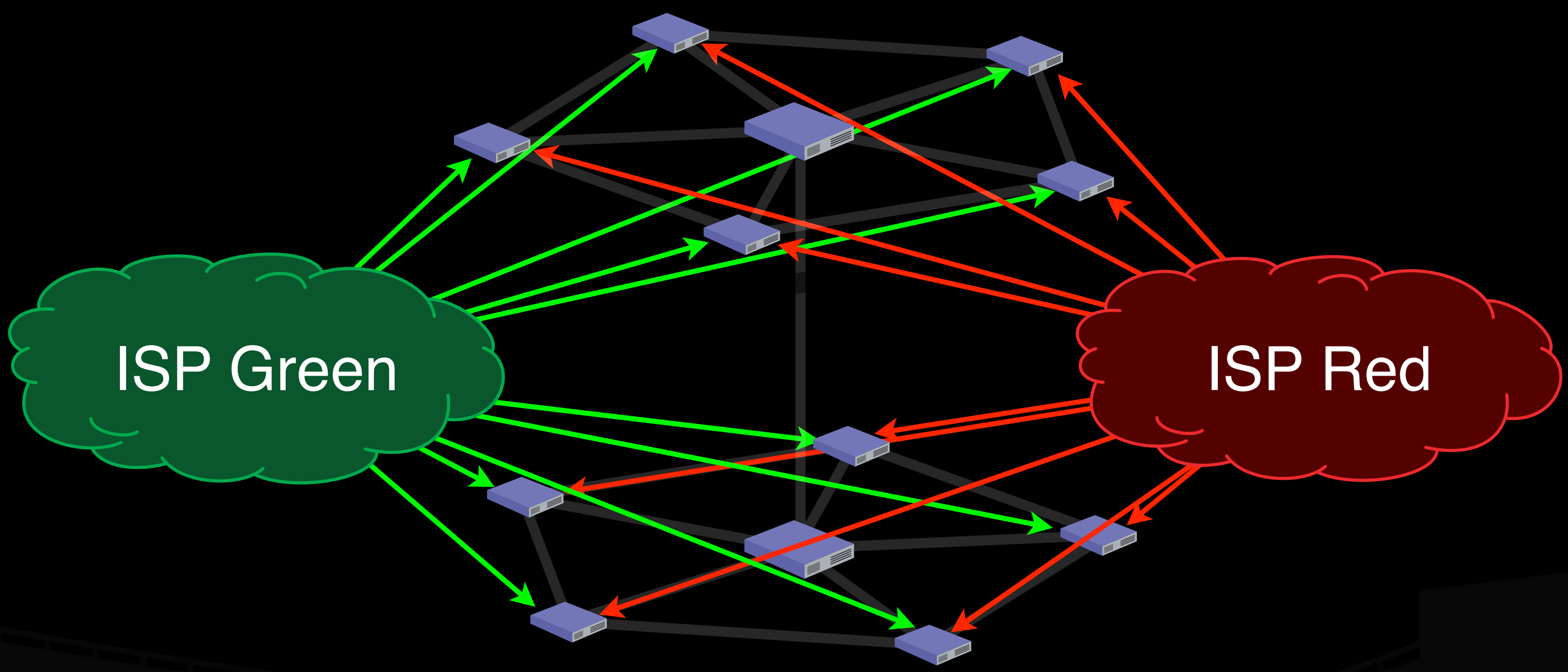
A Network Topology

"Dual Wagon-Wheel"



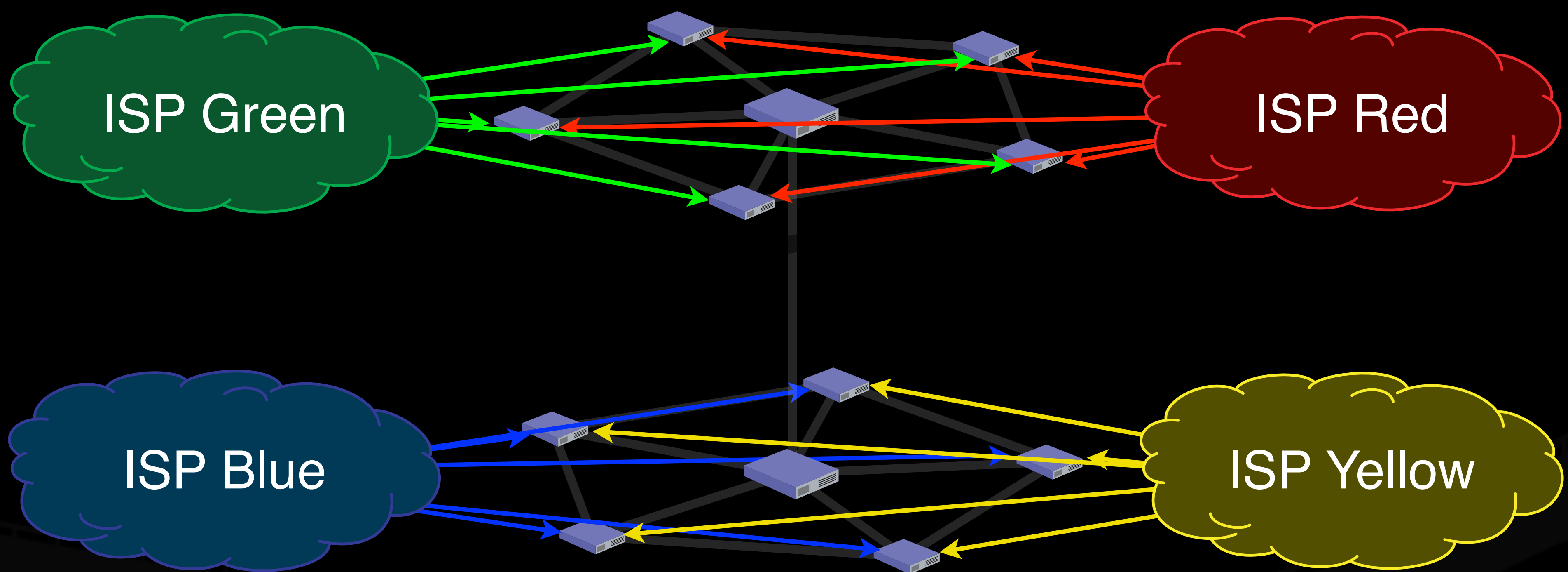
Redundant Transit

Two ISPs



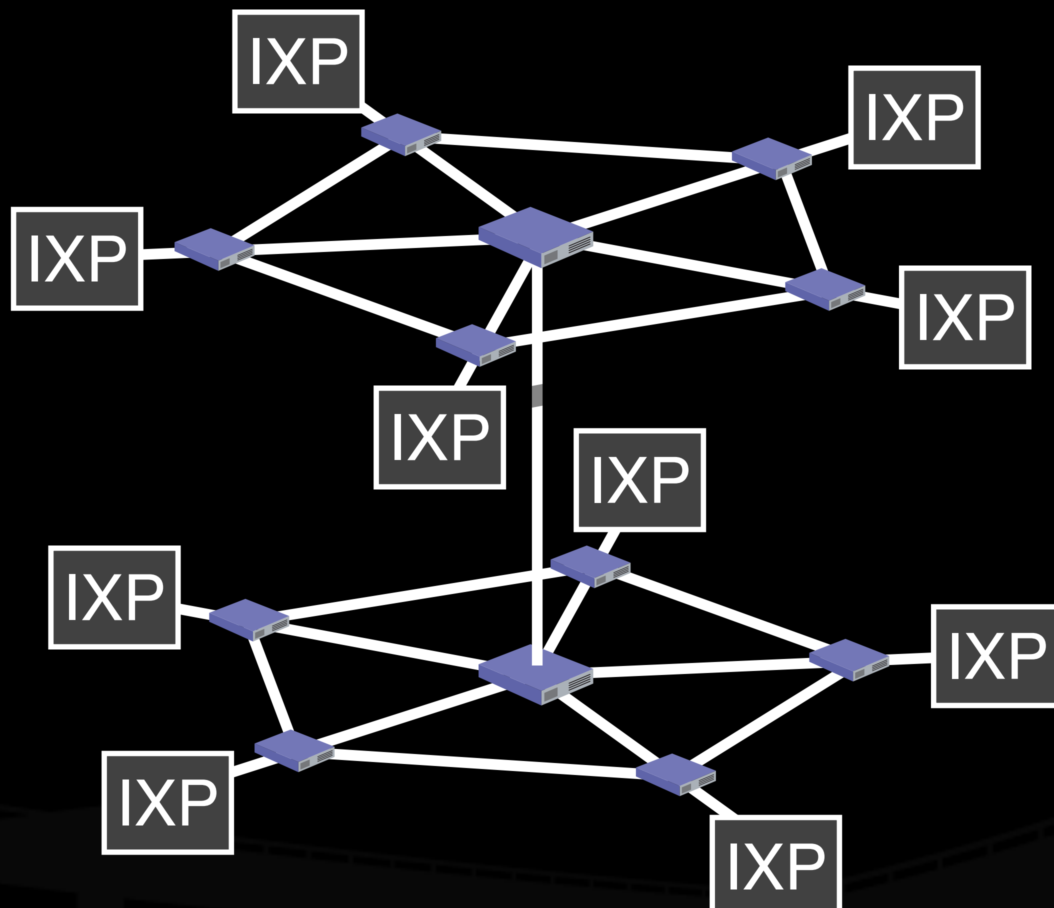
Redundant Transit

Or four ISPs

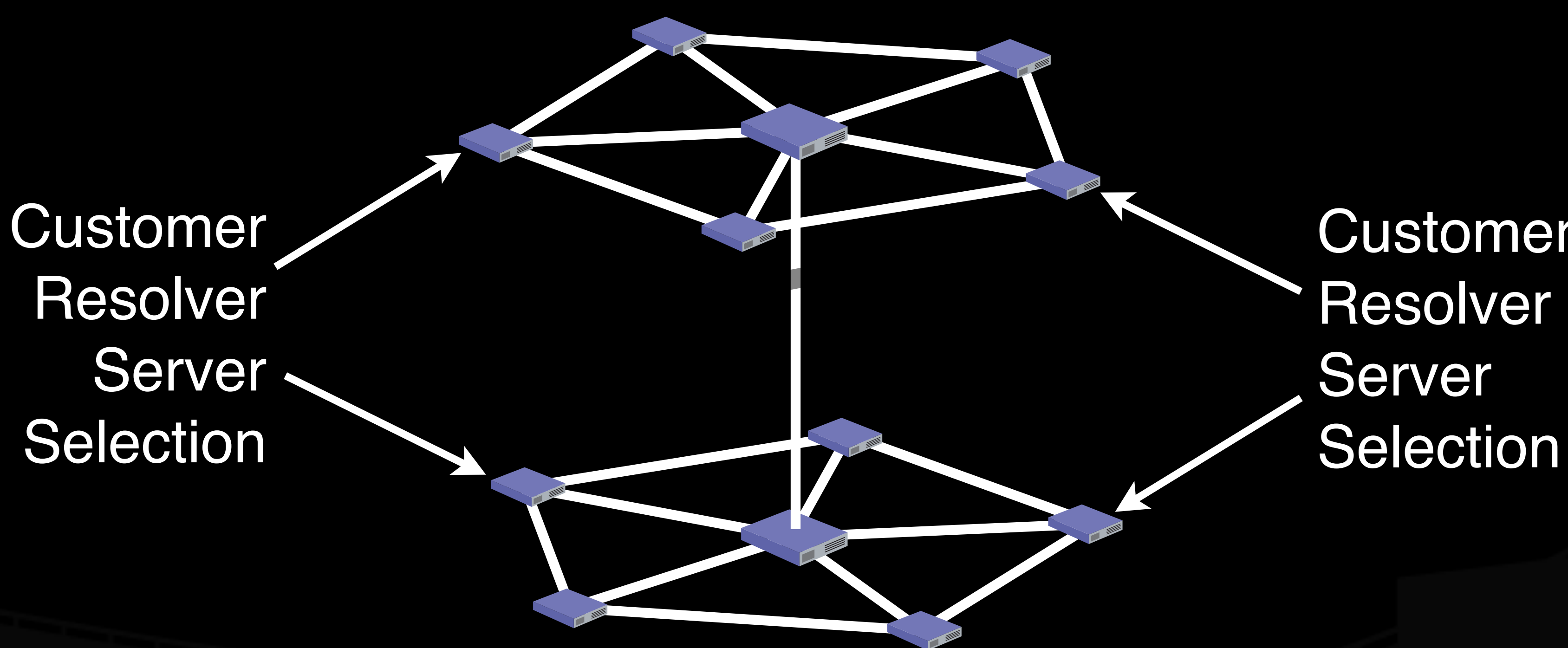


PCH
Packet Clearing House

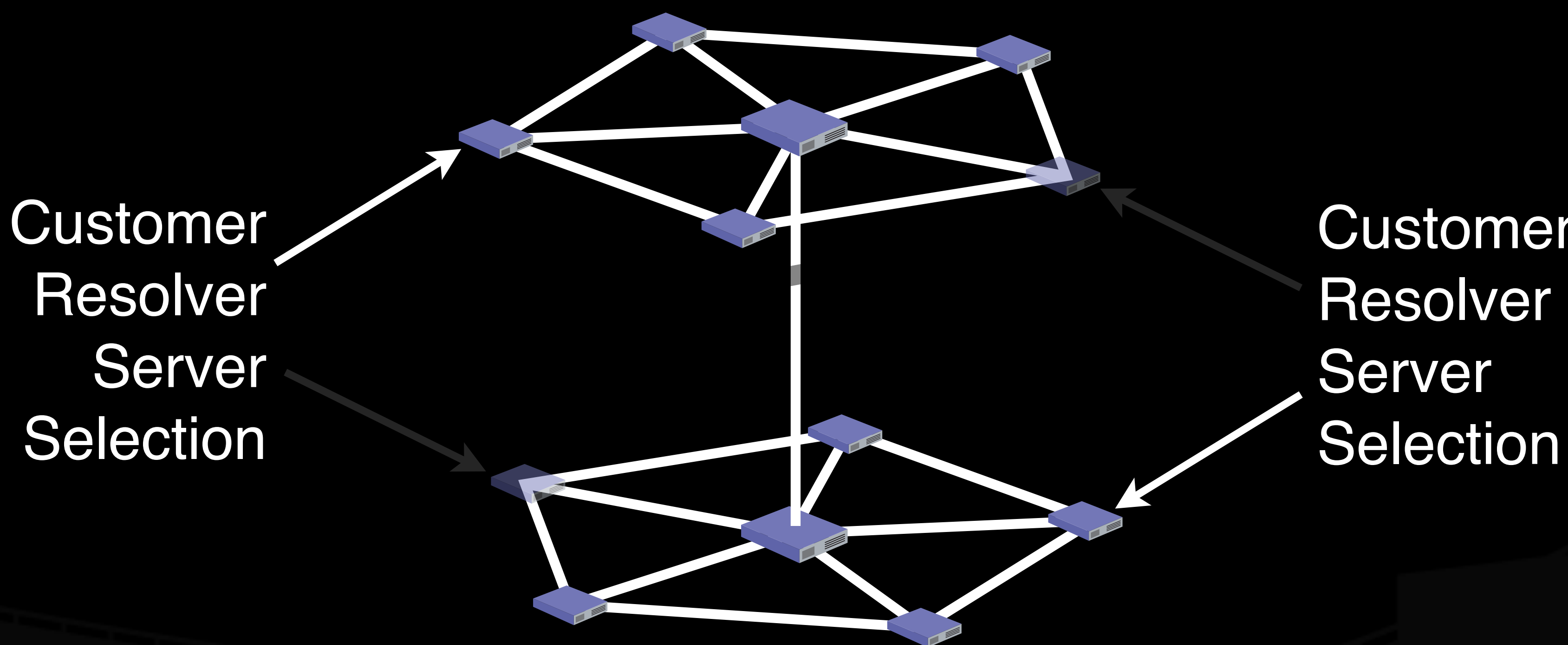
Local Peering



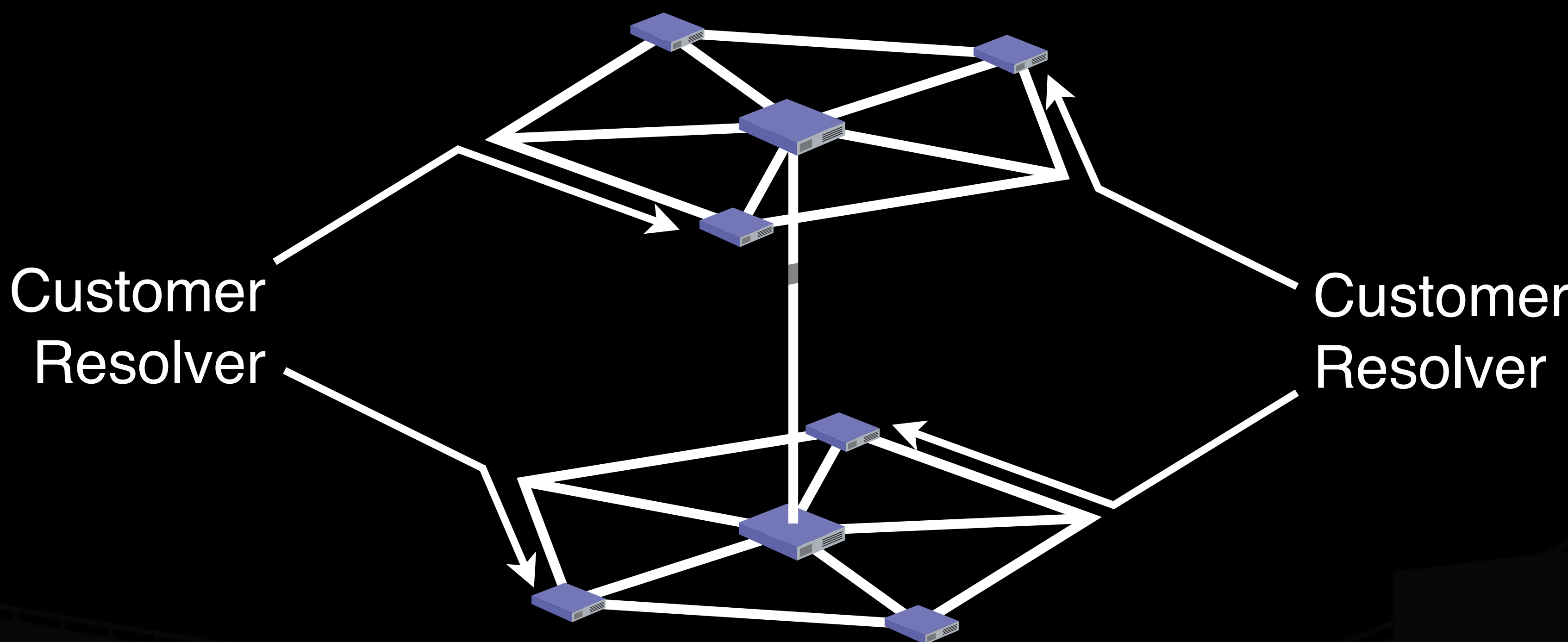
Resolver-Based Fail-Over



Resolver-Based Fail-Over

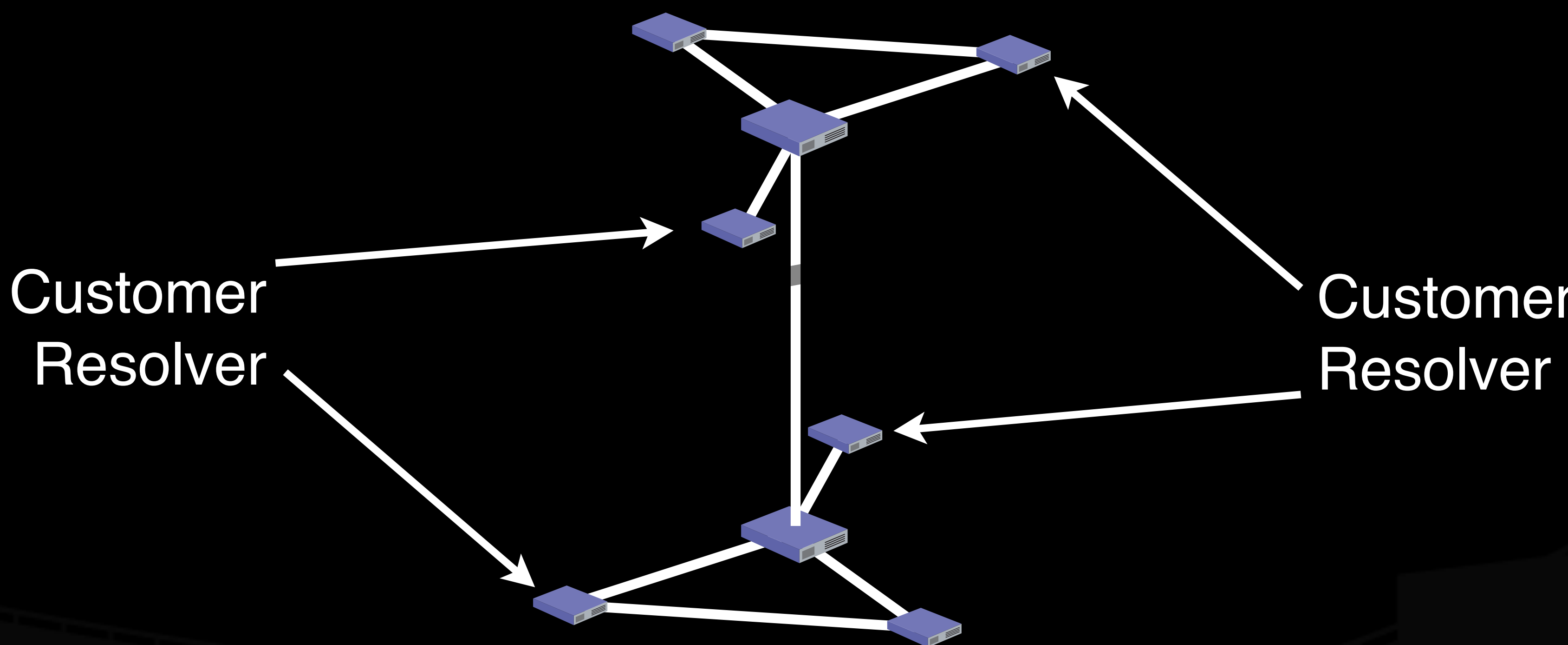


Internal Anycast Fail-Over



PCH
Packet Clearing House

Global Anycast Fail-Over



Unicast Attack Effects

Traditional unicast server deployment...

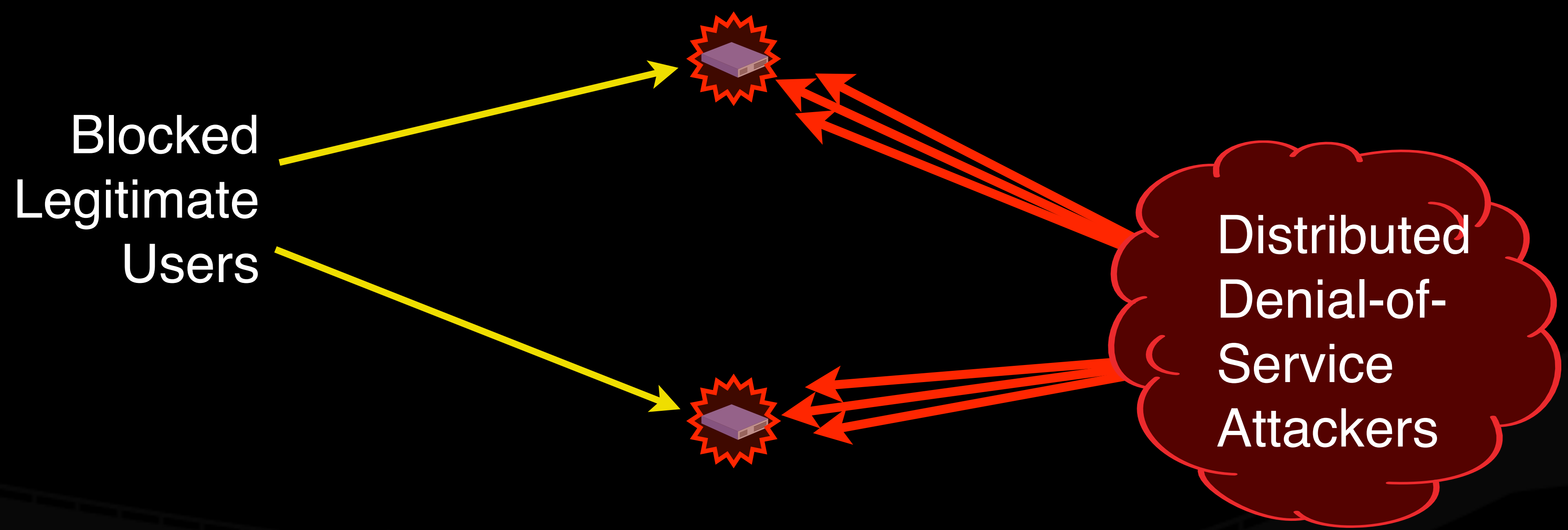
Unicast Attack Effects

Traditional unicast server deployment...

...exposes all servers to all attackers.

Unicast Attack Effects

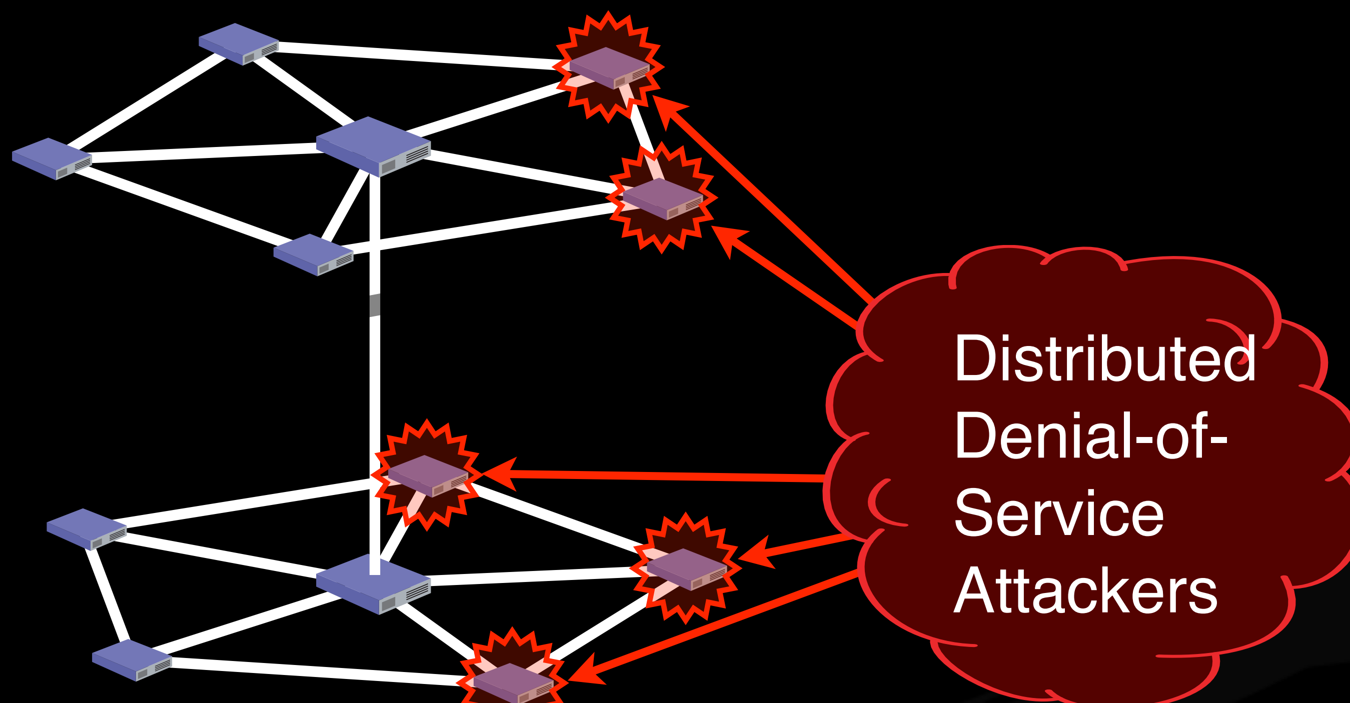
Traditional unicast server deployment...



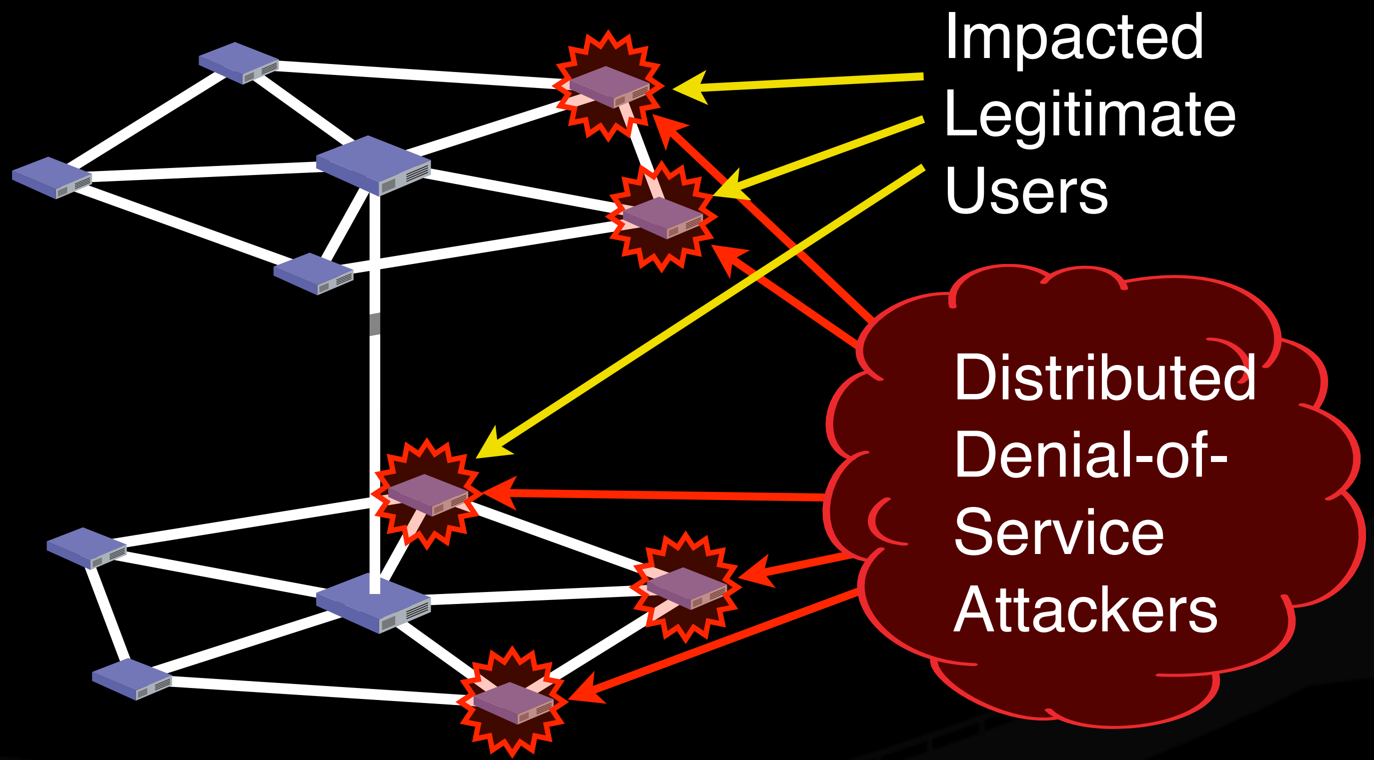
...exposes all servers to all attackers,
leaving no resources for legitimate users.

Packet Clearing House

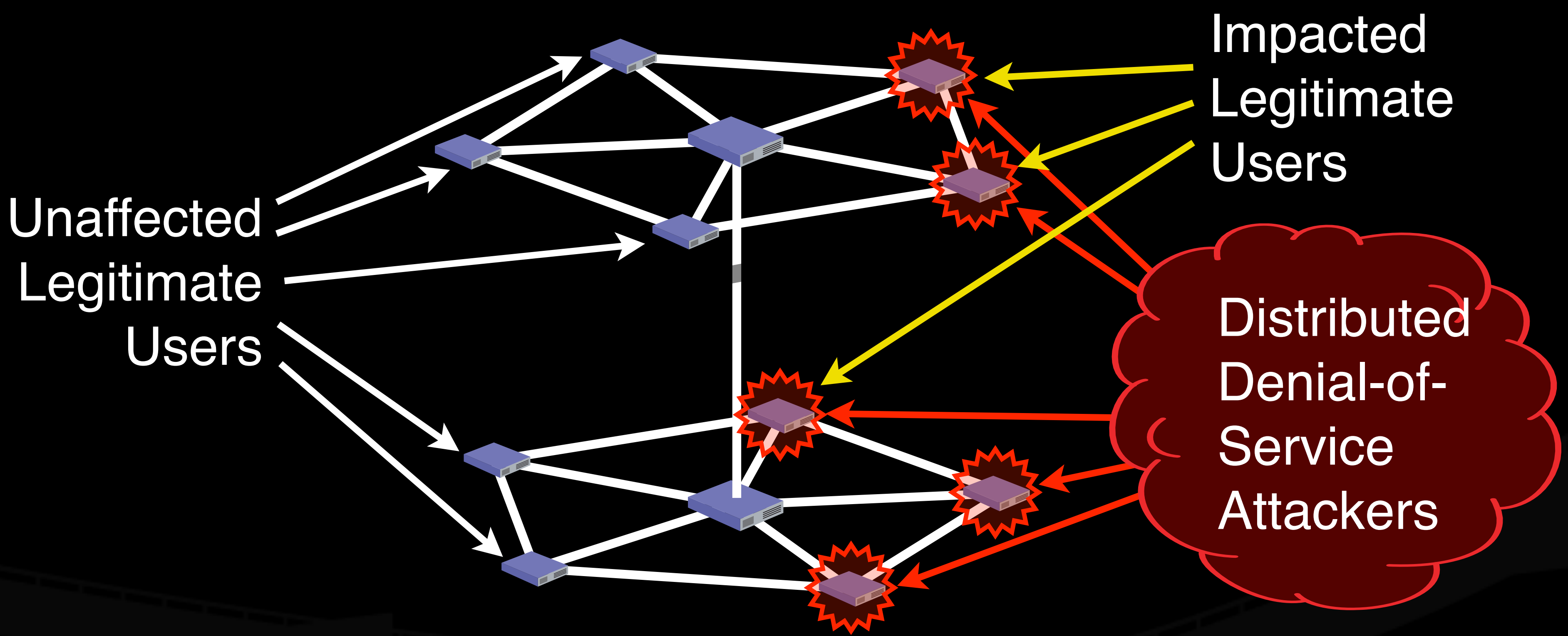
Anycast Attack Mitigation



Anycast Attack Mitigation



Anycast Attack Mitigation



Packet Clearing House

Thanks, and Questions?

Copies of this presentation can be
found in PDF and QuickTime formats at:

https:// pch.net / resources / papers / dns-service-architecture

Bill Woodcock
Research Director
Packet Clearing House
woody@pch.net

Overview of PCH DNS Anycast Service and Infrastructure

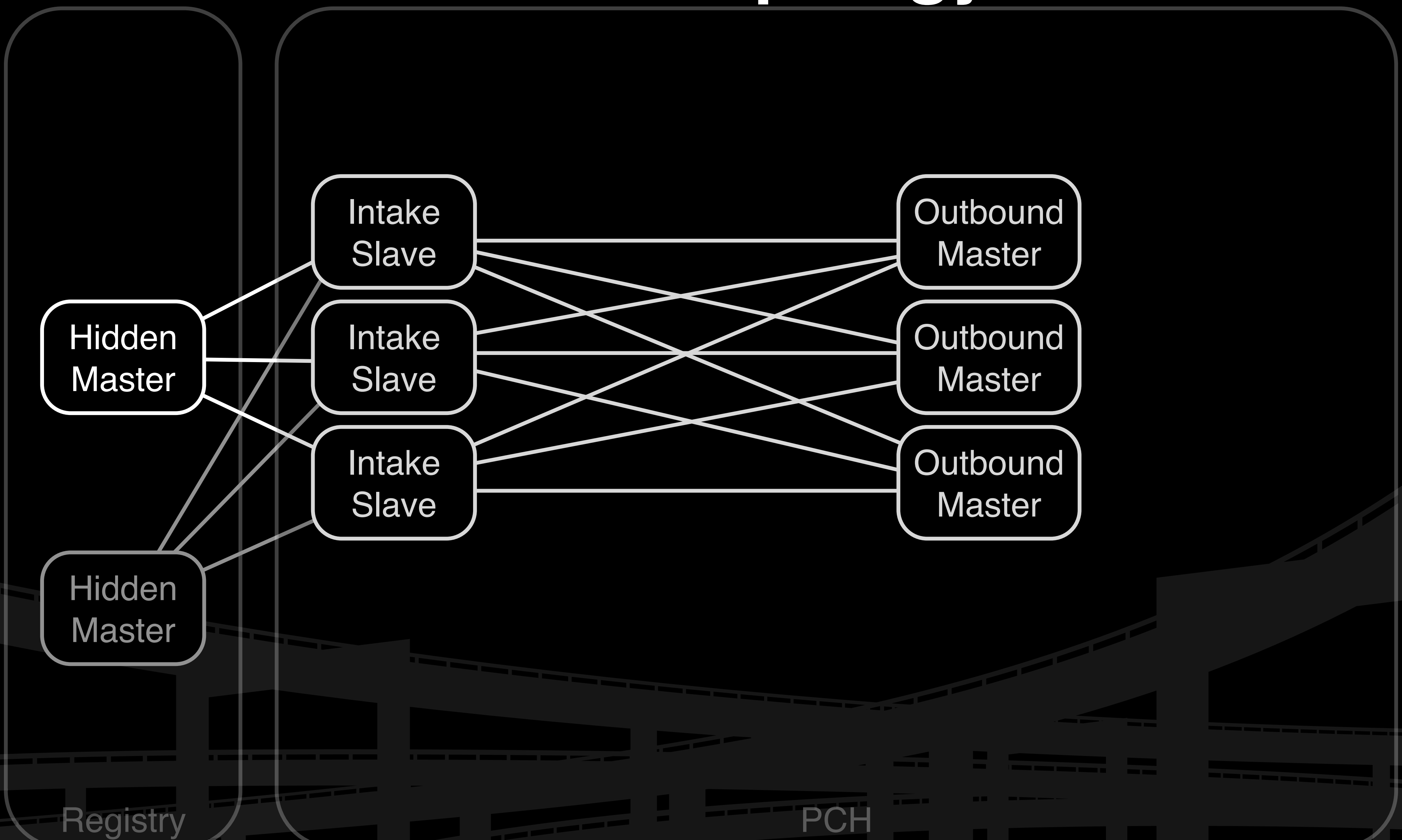
Bill Woodcock

March, 2016

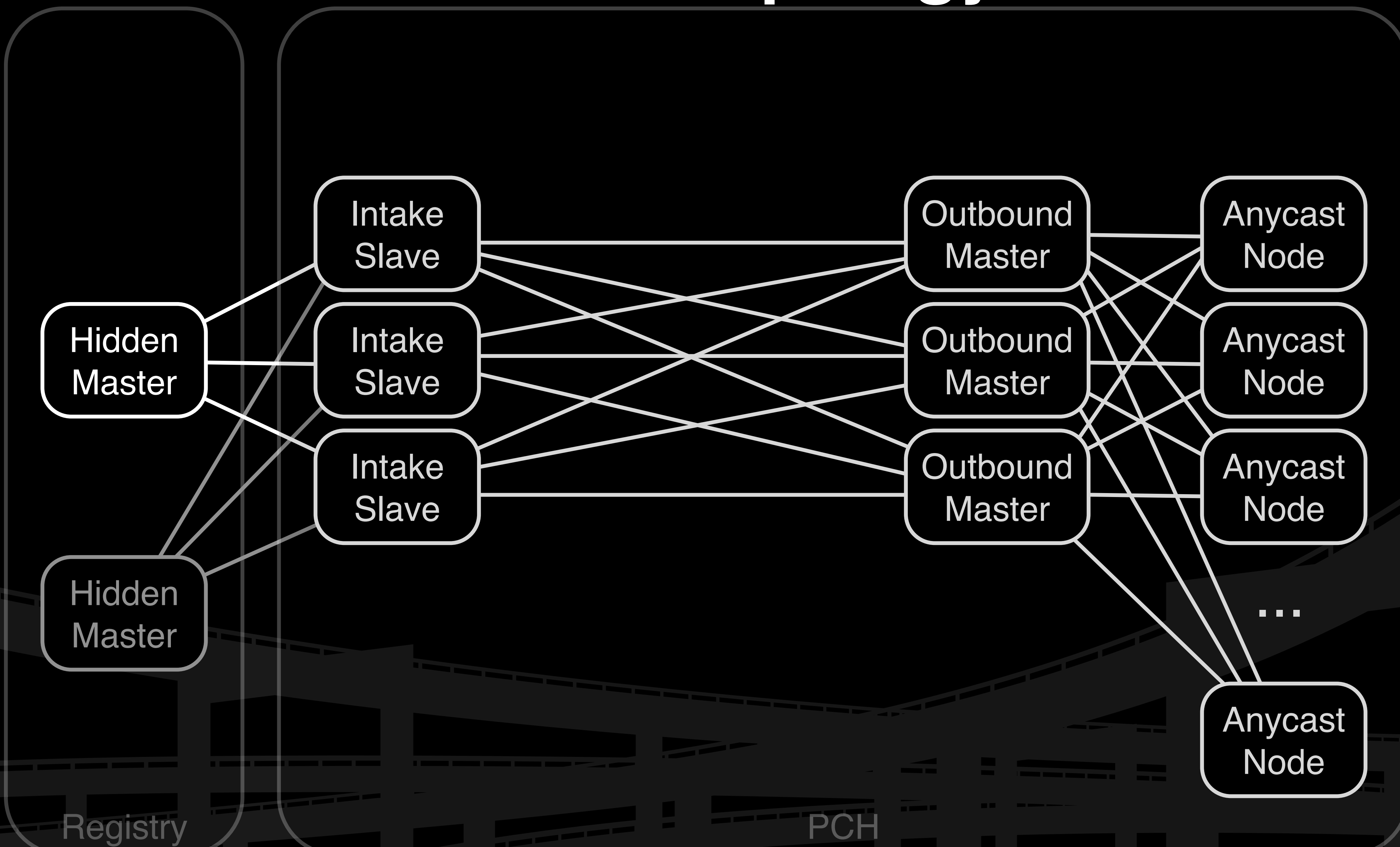
Overall Topology

Overall Topology

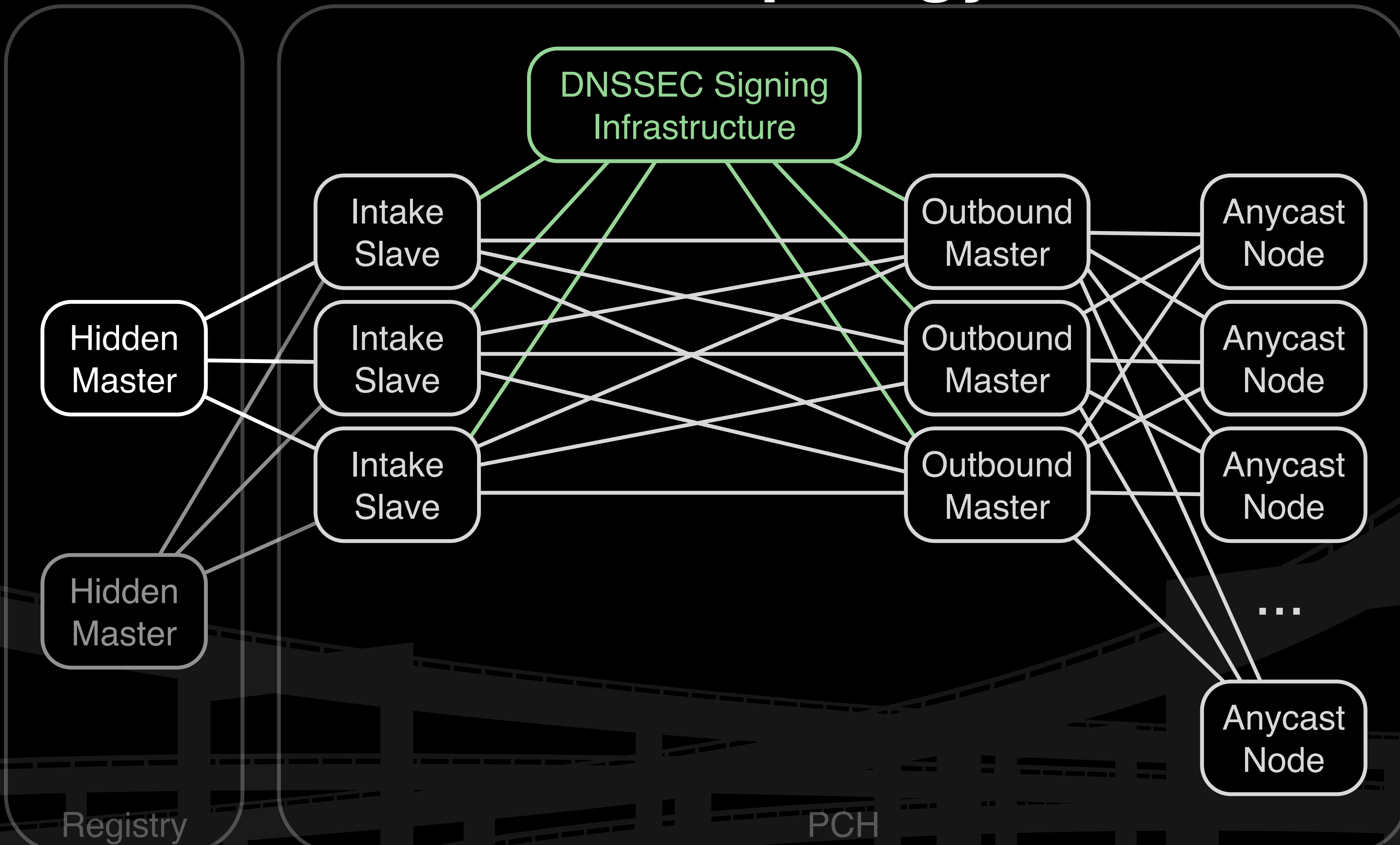
Overall Topology



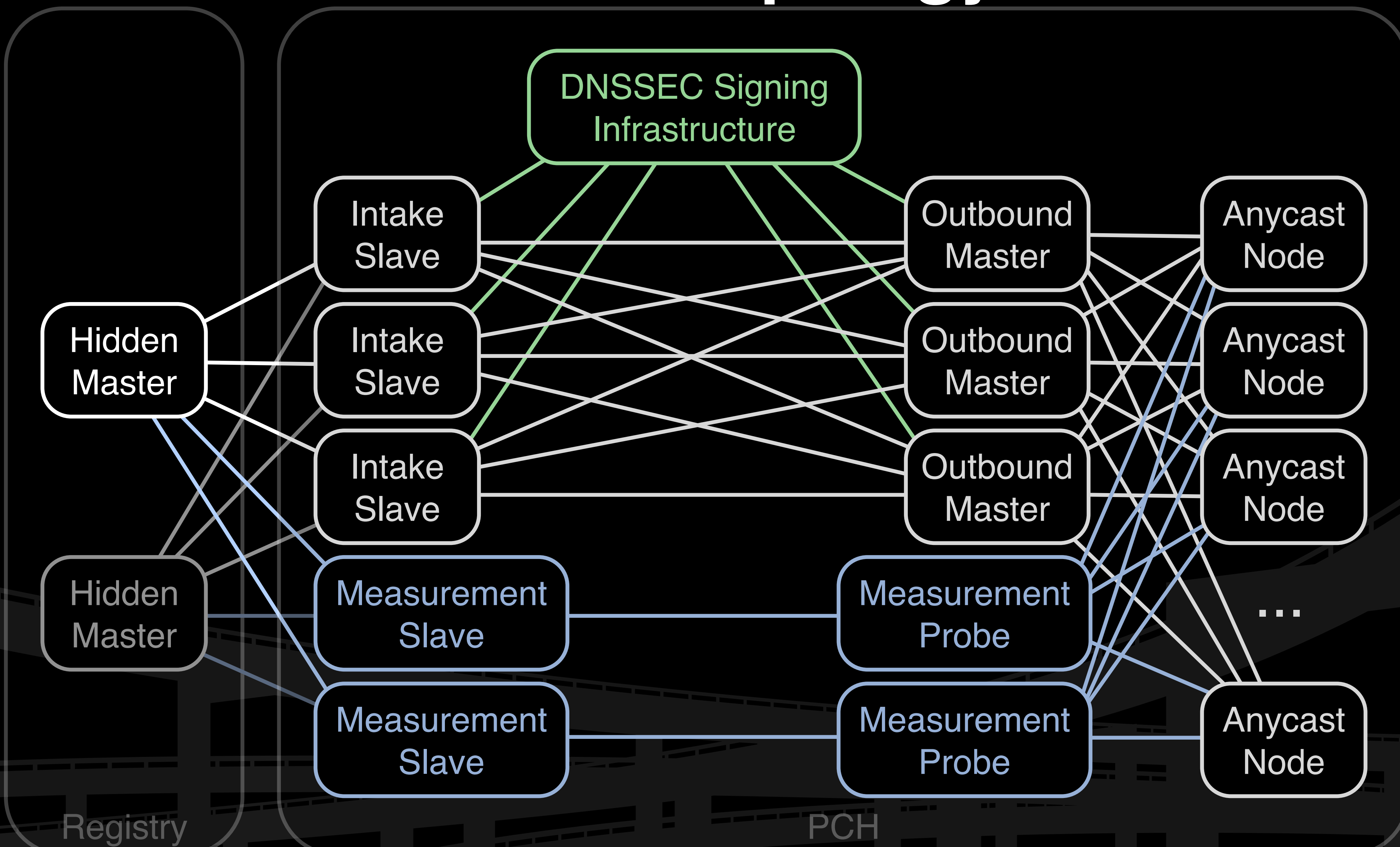
Overall Topology



Overall Topology

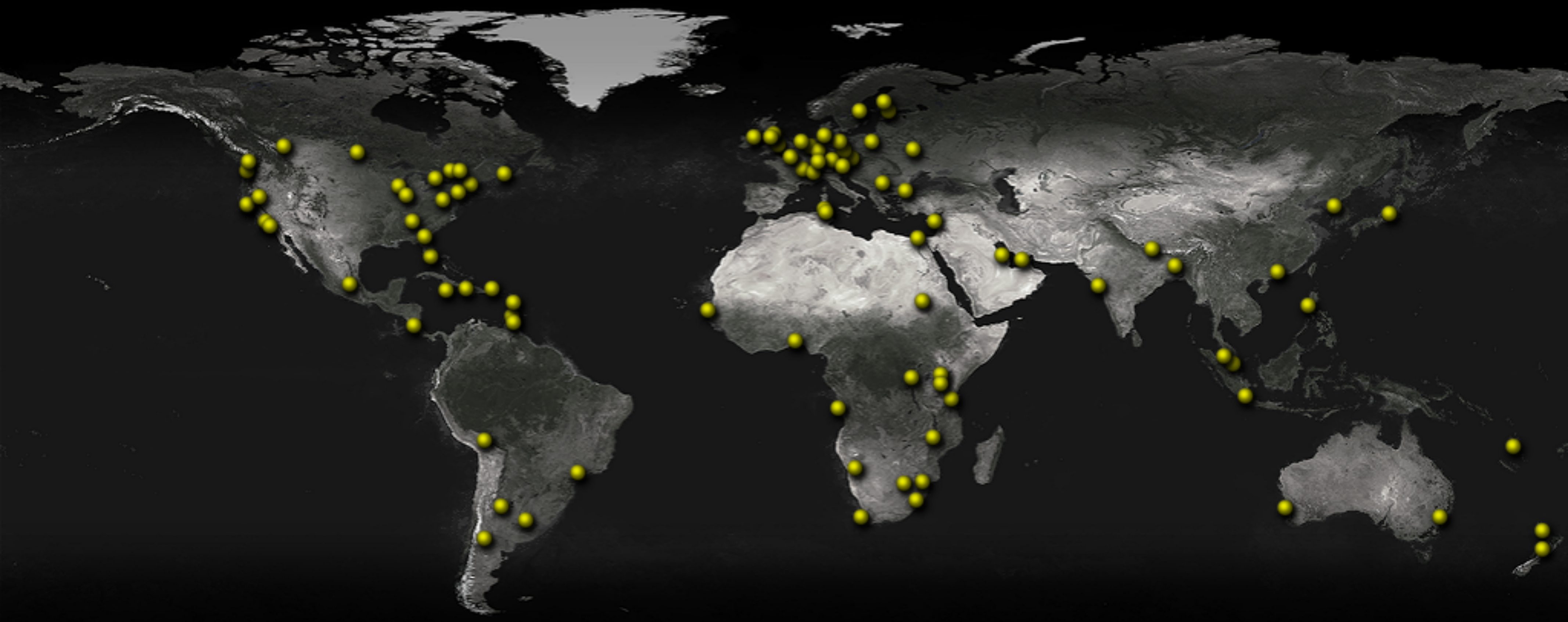


Overall Topology



Packet Clearing House

130 Locations



Anycast Node Construction

135 locations at the moment, adding a new one about every ten days.

70% are “small” 250Mbps
20% are “medium” 20-60Gbps
10% are “large” 60-120Gbps

All installations are preconfigured. Small are self-installed by the local host, while medium and large are installed by PCH staff.

PCH
Packet Clearing House

Small (70%)

2Gbps peering
1Gbps transit

Cisco 2921 Router
250Mbps throughput

Internally-integrated
Cisco UCS-E160D-M2 x86 server
64GB RAM, 2x 1TB SATA drives

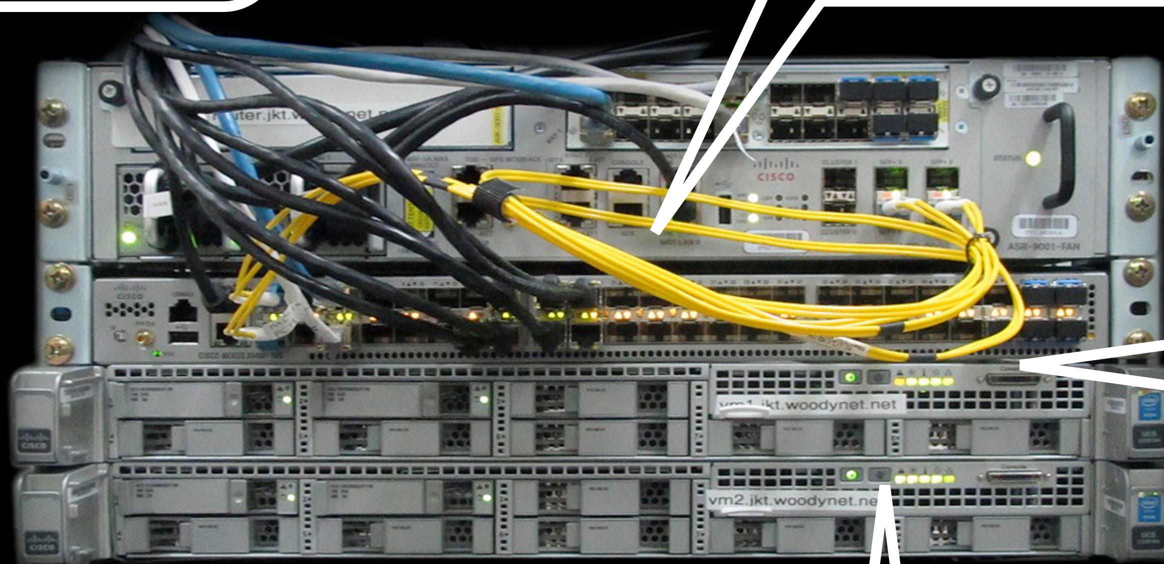
All-in-one enclosure, ships preconfigured in a single shipping crate, requires only three patch cords and one power cord to bring up.

Packet Clearing House

Medium (20%)

10-40Gbps peering
10-20Gbps transit

Cisco ASR9001 Router

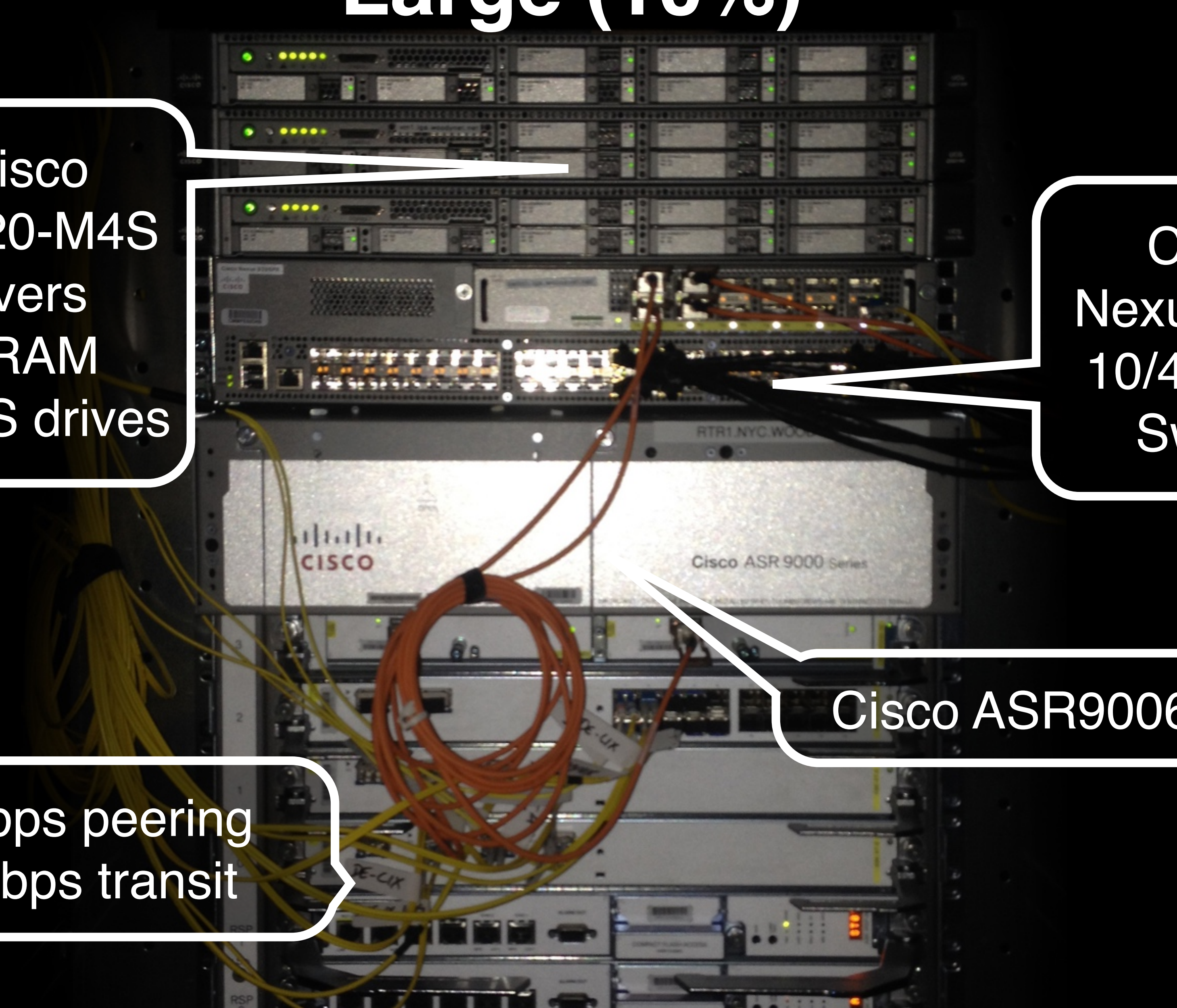


Cisco
Nexus 3548
10Gbps
Switch

Two Cisco UCSC-C220-M4S x86 servers
768GB RAM, 8x 1TB SAS drives

PCH
Packet Clearing House

Large (10%)



3x-8x Cisco
UCSC-C220-M4S
x86 servers
768GB RAM
8x 1TB SAS drives

Cisco
Nexus 9396
10/40Gbps
Switch

Cisco ASR9006 Router

40-80Gbps peering
20-40Gbps transit

Packet Clearing House

Making Our Own Bandwidth

Essentially all Internet bandwidth (more than 98%) is produced by “peering” in Internet Exchange Points.

Bandwidth is transported from IXPs to the point of consumption, increasing in cost, and suffering loss and latency along the way. This is called “transit.”

Unlike other DNS service providers, we are not dependent upon transit. We serve data exclusively from within IXPs, producing essentially all of our own bandwidth at higher quality and lower cost.

Nondiscriminatory Access

Other DNS service providers dependency upon transit makes registries' zone data a pawn in local transit politics.

By contrast, through our open peering, PCH makes the zones of the registries we serve equally available to all networks and users at no cost.

We already have nearly 8,000 direct connections with other networks in 130 locations on six continents, and add more every day.

New Zone Autoconfiguration

From trusted registries, over authenticated transport, we autoconfigure new zones.

If we see an update pertaining to a zone that we're not configured for, we automatically configure that zone across our infrastructure.

If the zone goes stale, we check whether it's delegated to our servers from the root. If not, we deconfigure it and stop serving it.

AXFR to IXFR

When registries serve us zone data via AXFR or we perform a DNSSEC full-zone signing, we convert to IXFR within our infrastructure, optimizing performance, particularly to our remotest anycast nodes.

Thanks, and Questions?

Bill Woodcock
Executive Director
Packet Clearing House
woody@pch.net