Best Practices in DNS Service-Provision Architecture

Version 1.0

February 2006

Bill Woodcock

Packet Clearing House

It’s all Anycast

Large ISPs have been running production anycast DNS for more than a decade.

Which is a very long time, in Internet years.

95% of the root nameservers are anycast.

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.

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.

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.

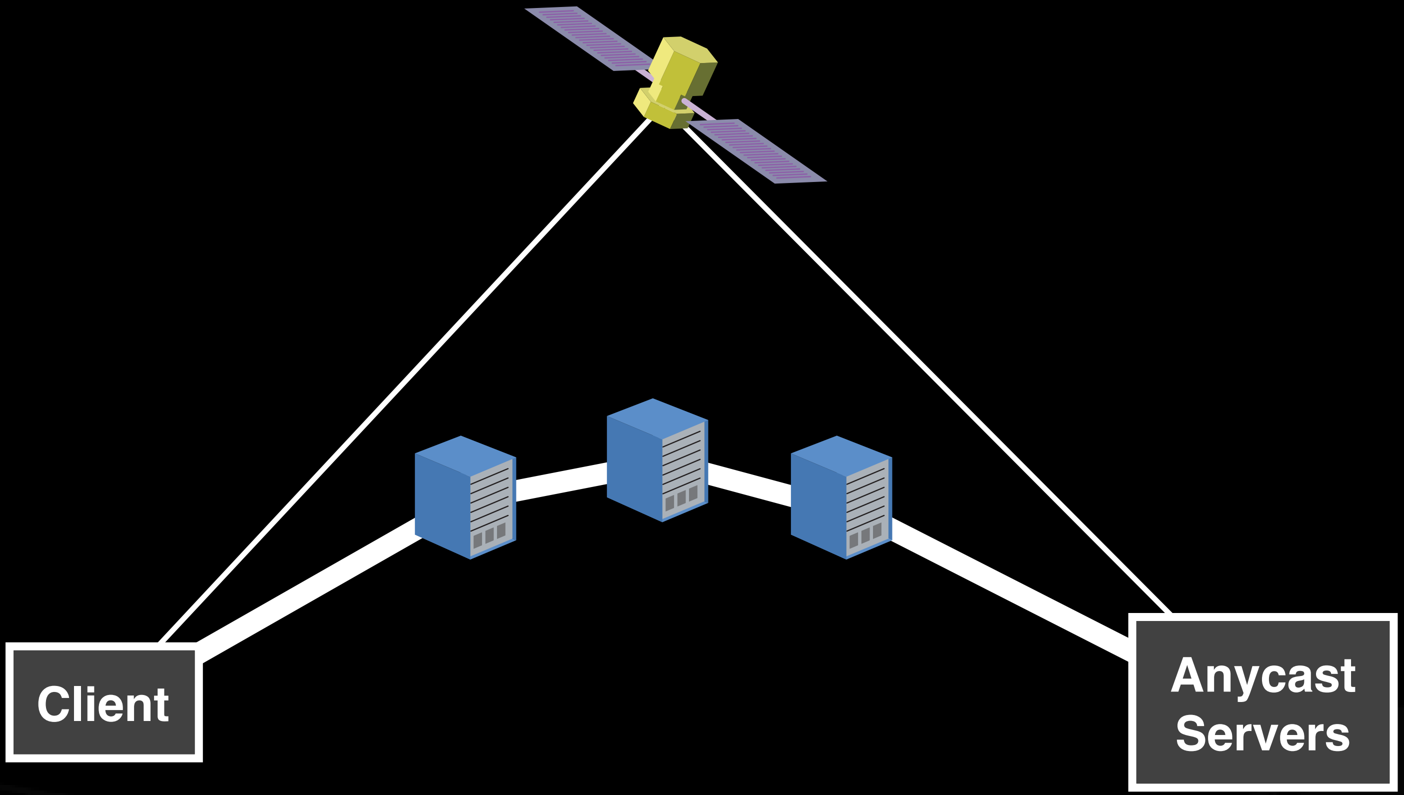
They don’t necessarily yield the same answer.

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

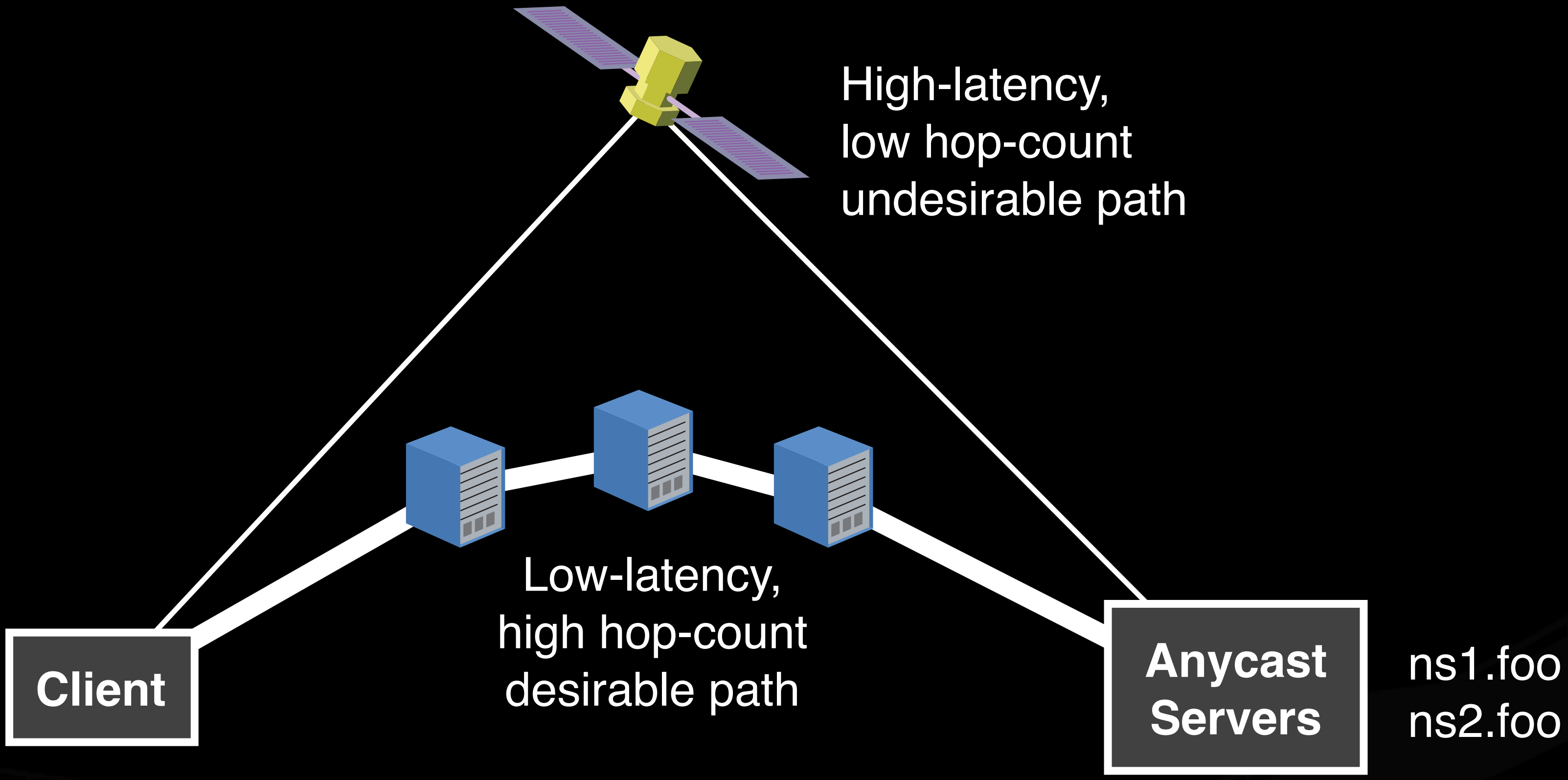
Neither the DNS service provider nor the user are likely to care about hop-count.

Both care a great deal about latency.

How The Conflict Plays Out

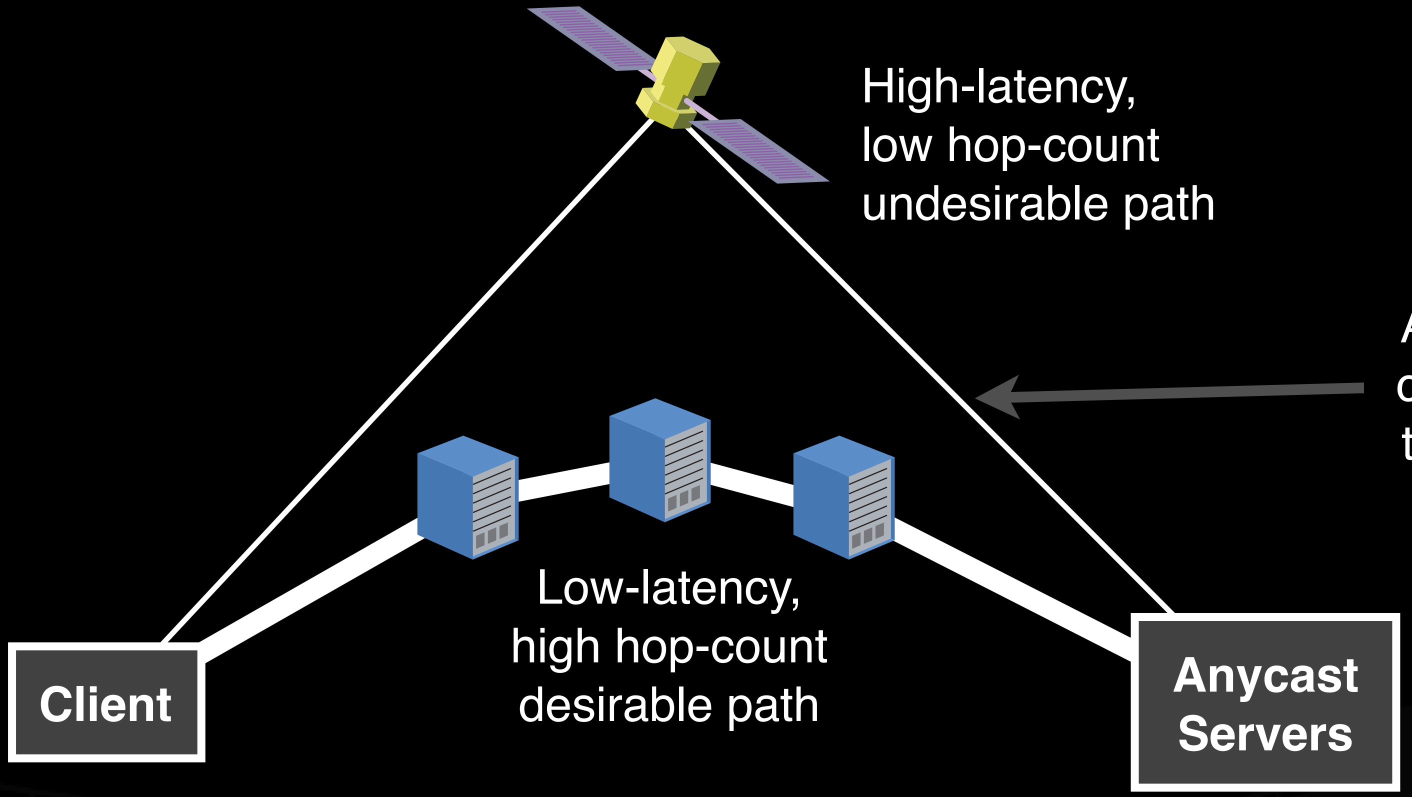


How The Conflict Plays Out



Two servers with the same routing policy

How The Conflict Plays Out



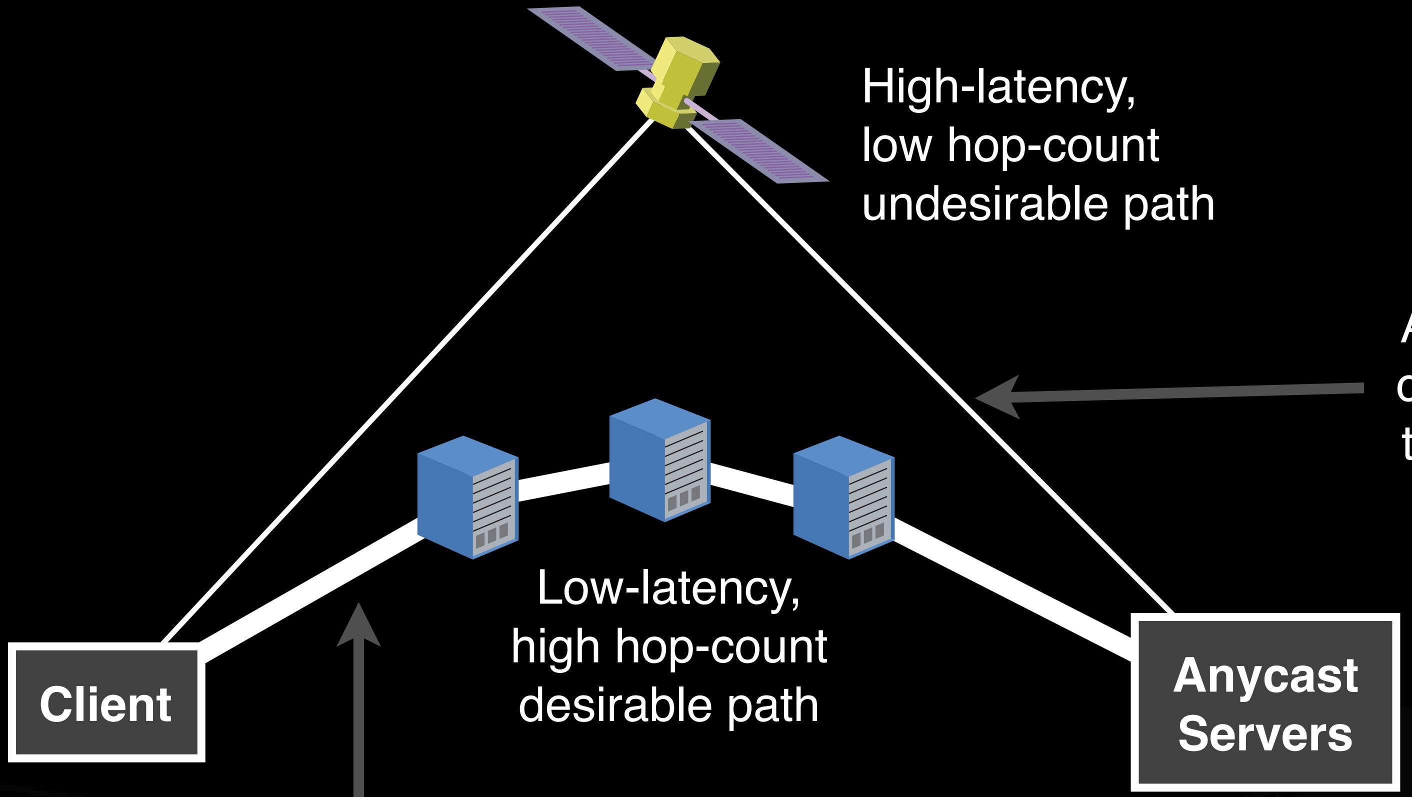
High-latency,
low hop-count
undesirable path

Anycast
chooses
this one

ns1.foo
ns2.foo

Two servers with the
same routing policy

How The Conflict Plays Out

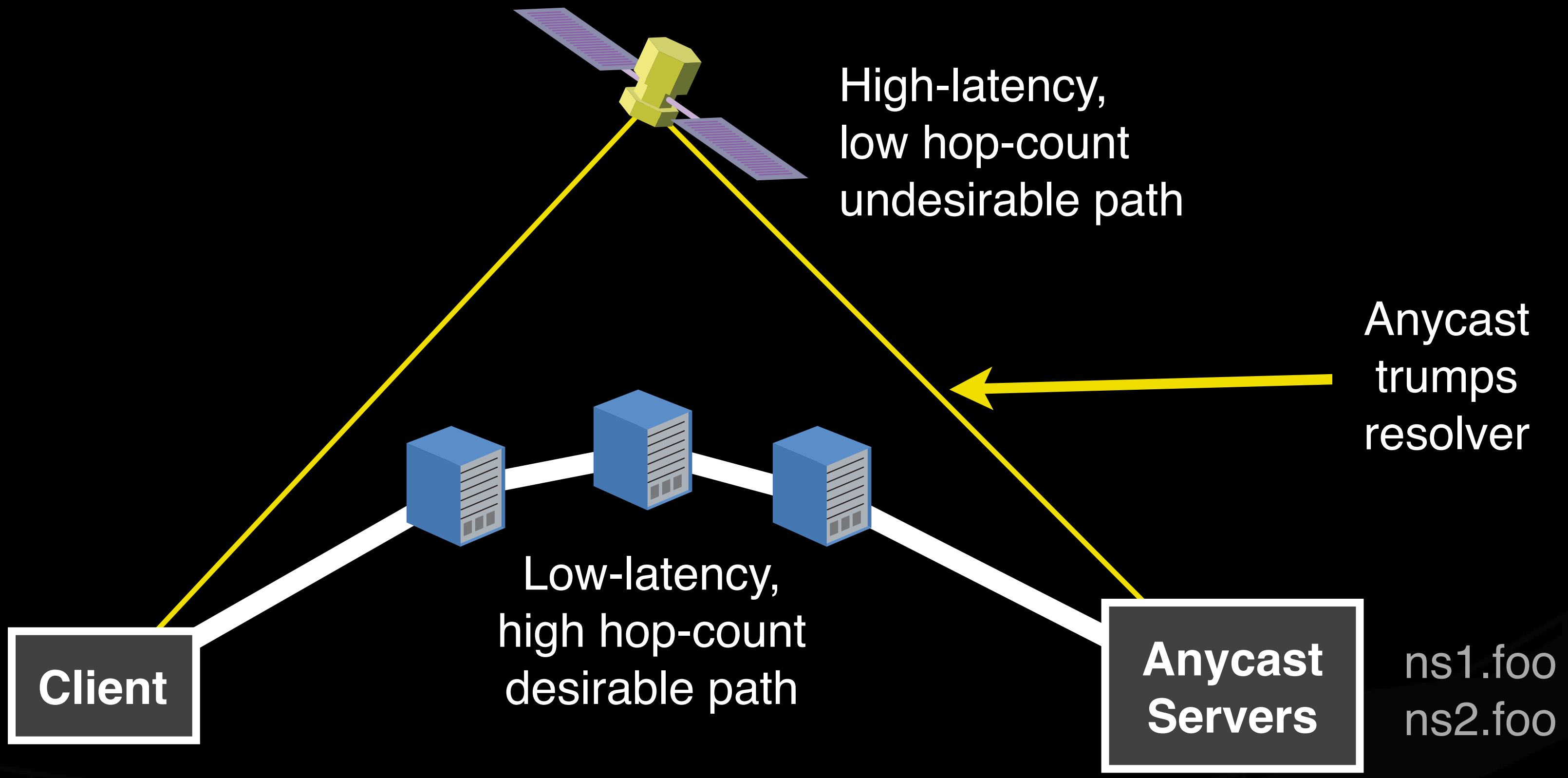


ns1.foo
ns2.foo

Resolver
chooses
this one

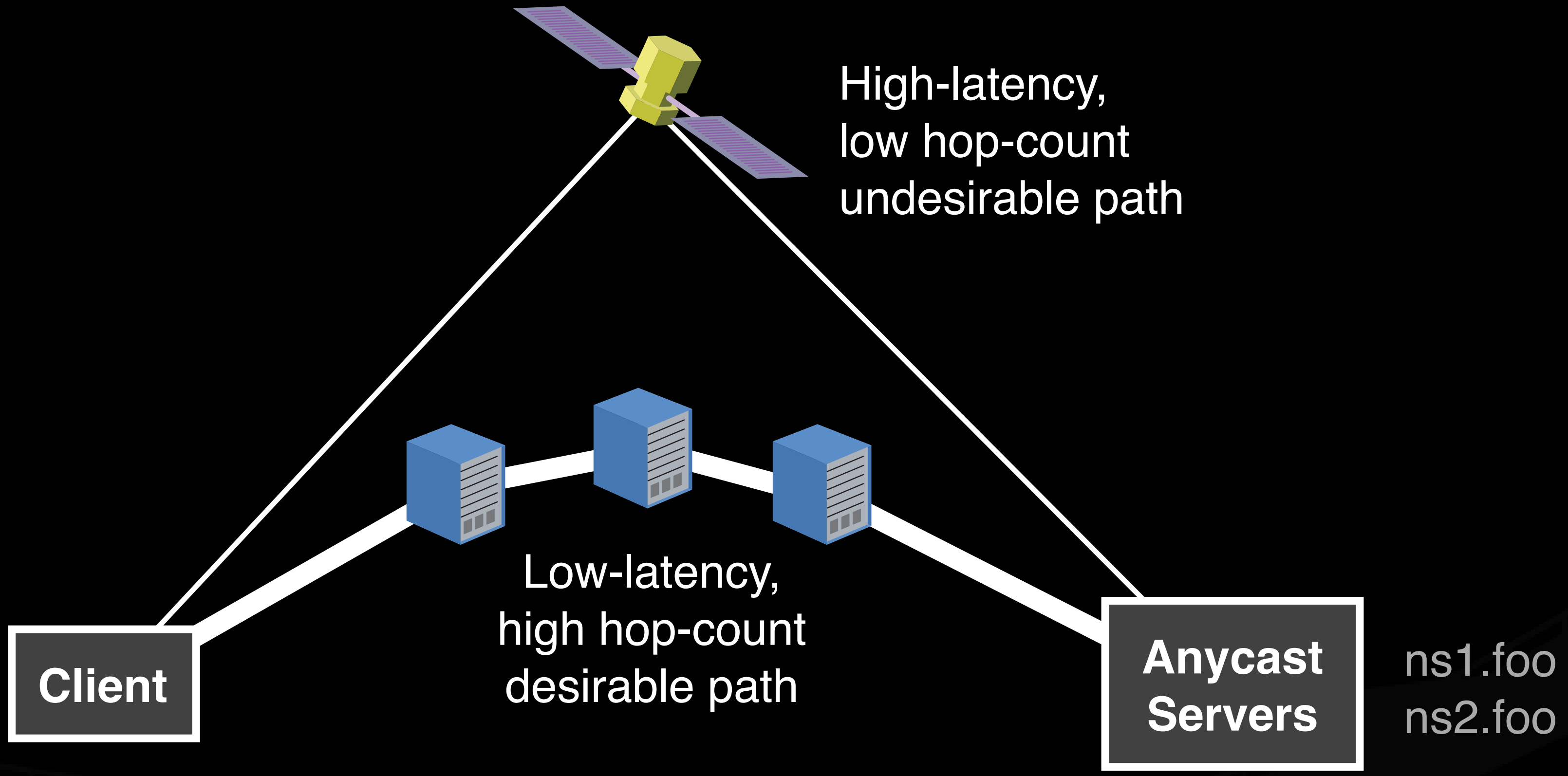
Two servers with the
same routing policy

How The Conflict Plays Out



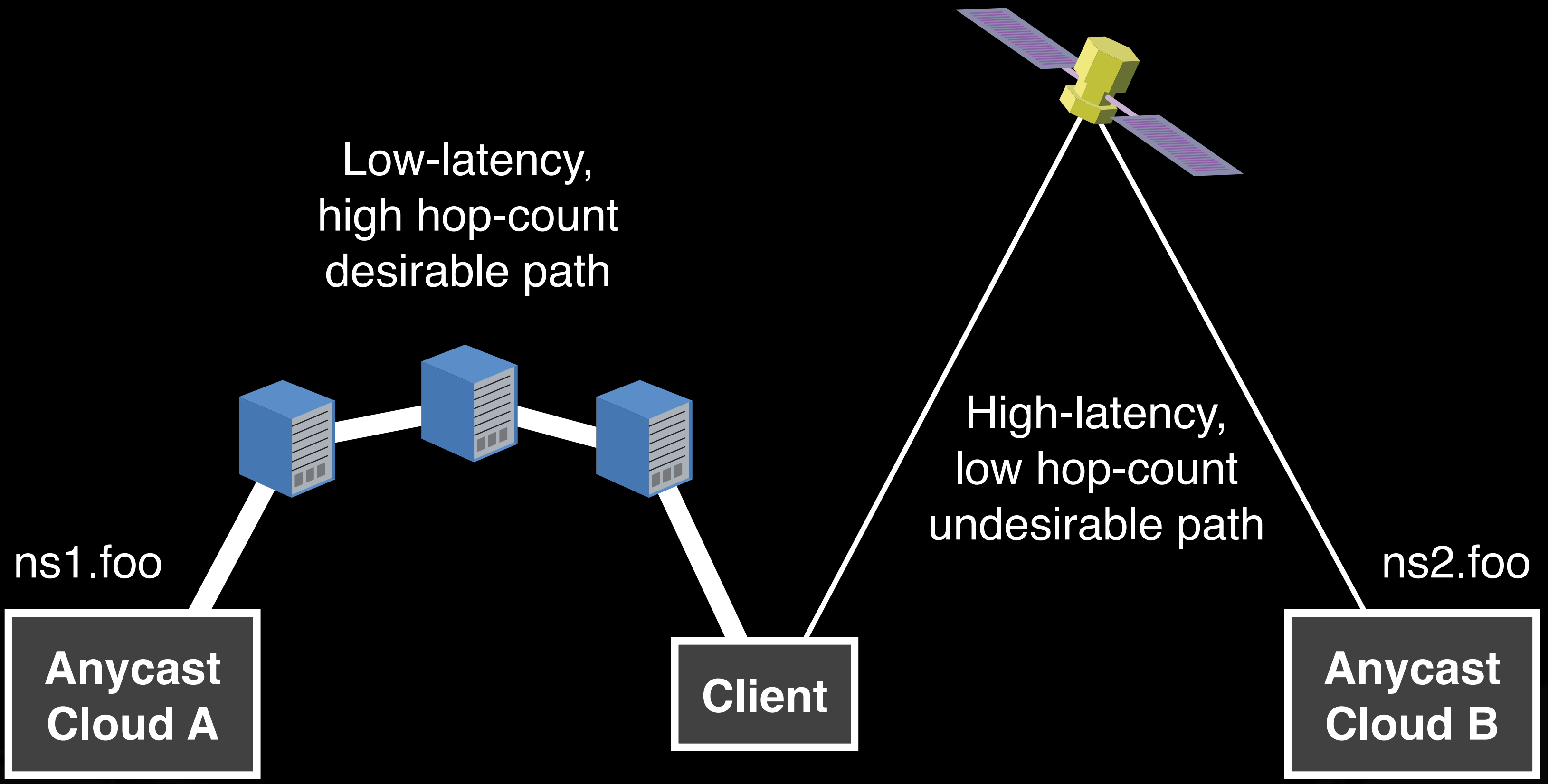
Two servers with the same routing policy

Resolve the Conflict



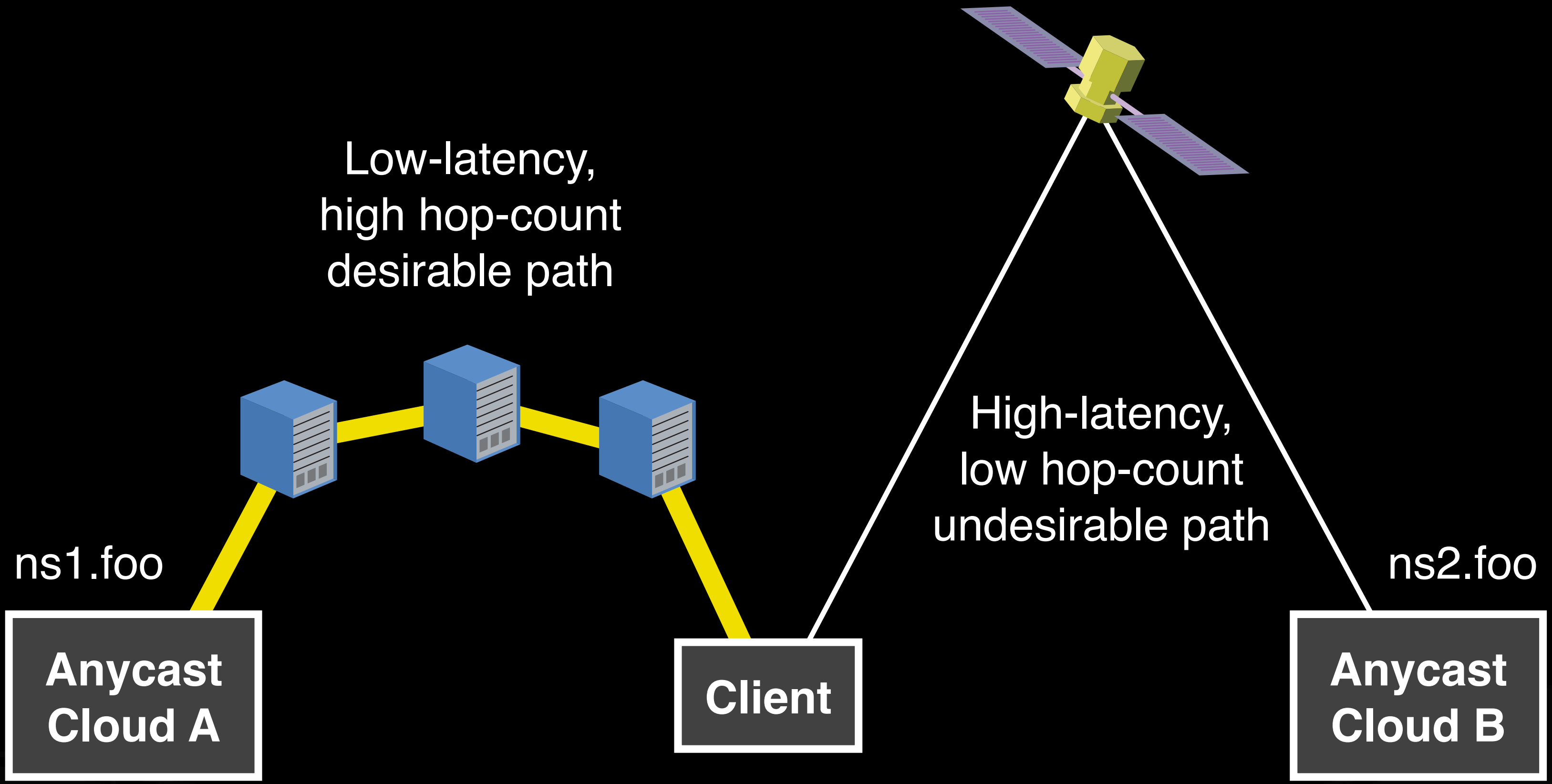
The resolver uses different IP addresses for its fail-over mechanism, while anycast uses the same IP addresses.

Resolve the Conflict



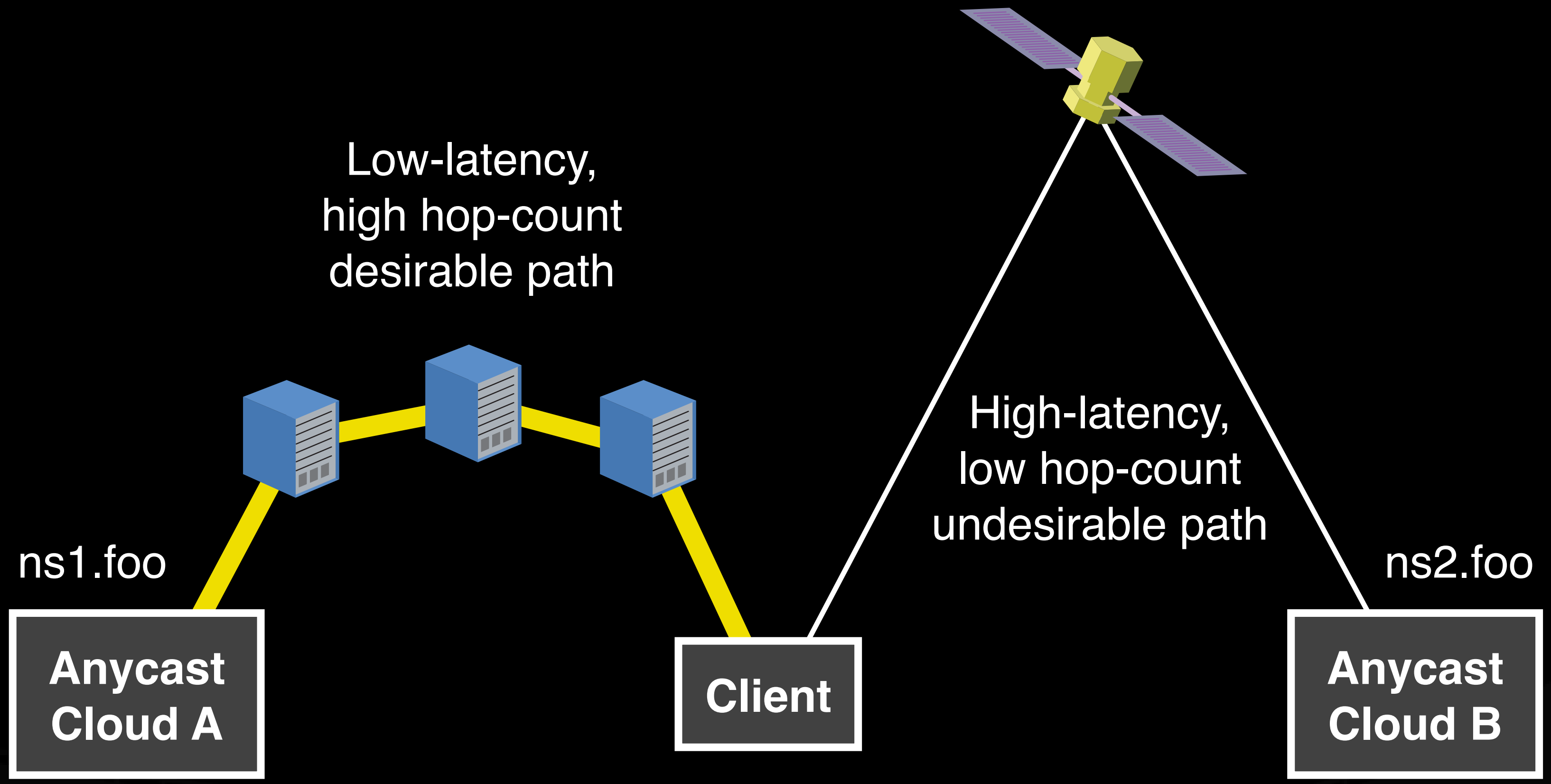
Split the anycast deployment into “clouds” of locations, each cloud using a different IP address and different routing policies.

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.

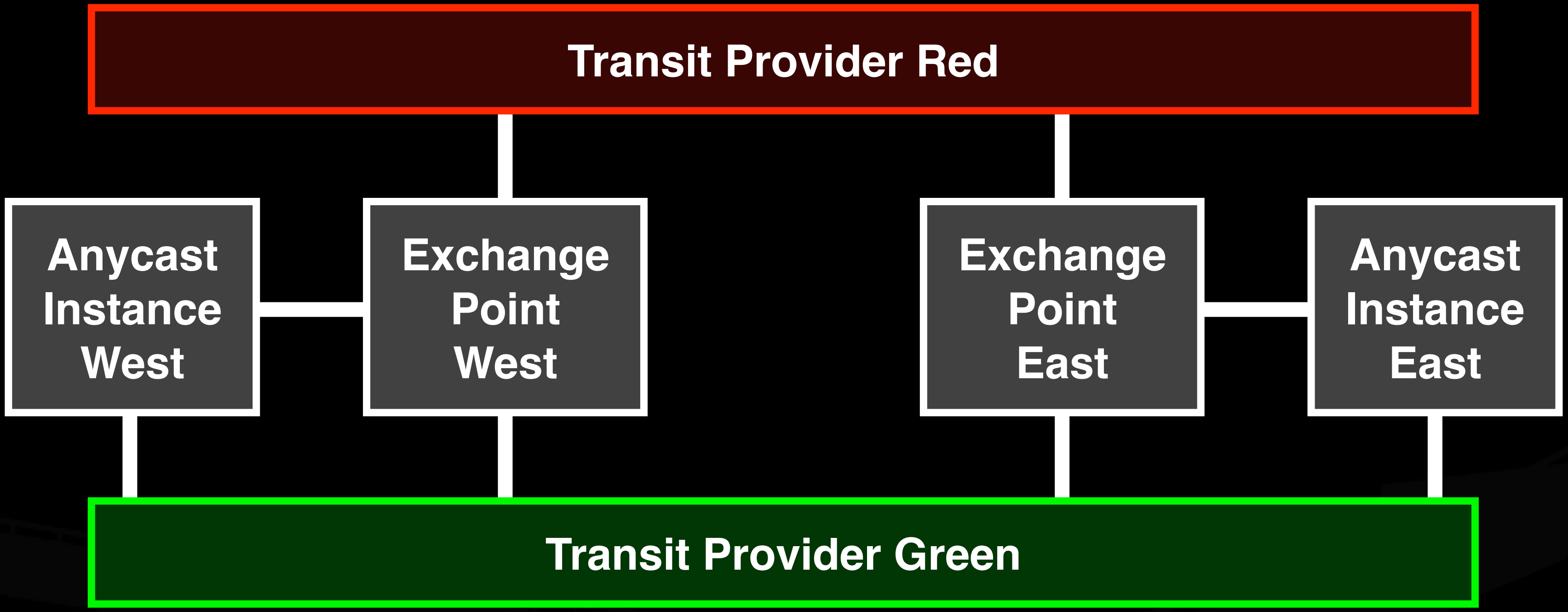
ISPs’ Economic Incentives (Grossly Simplified)

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.

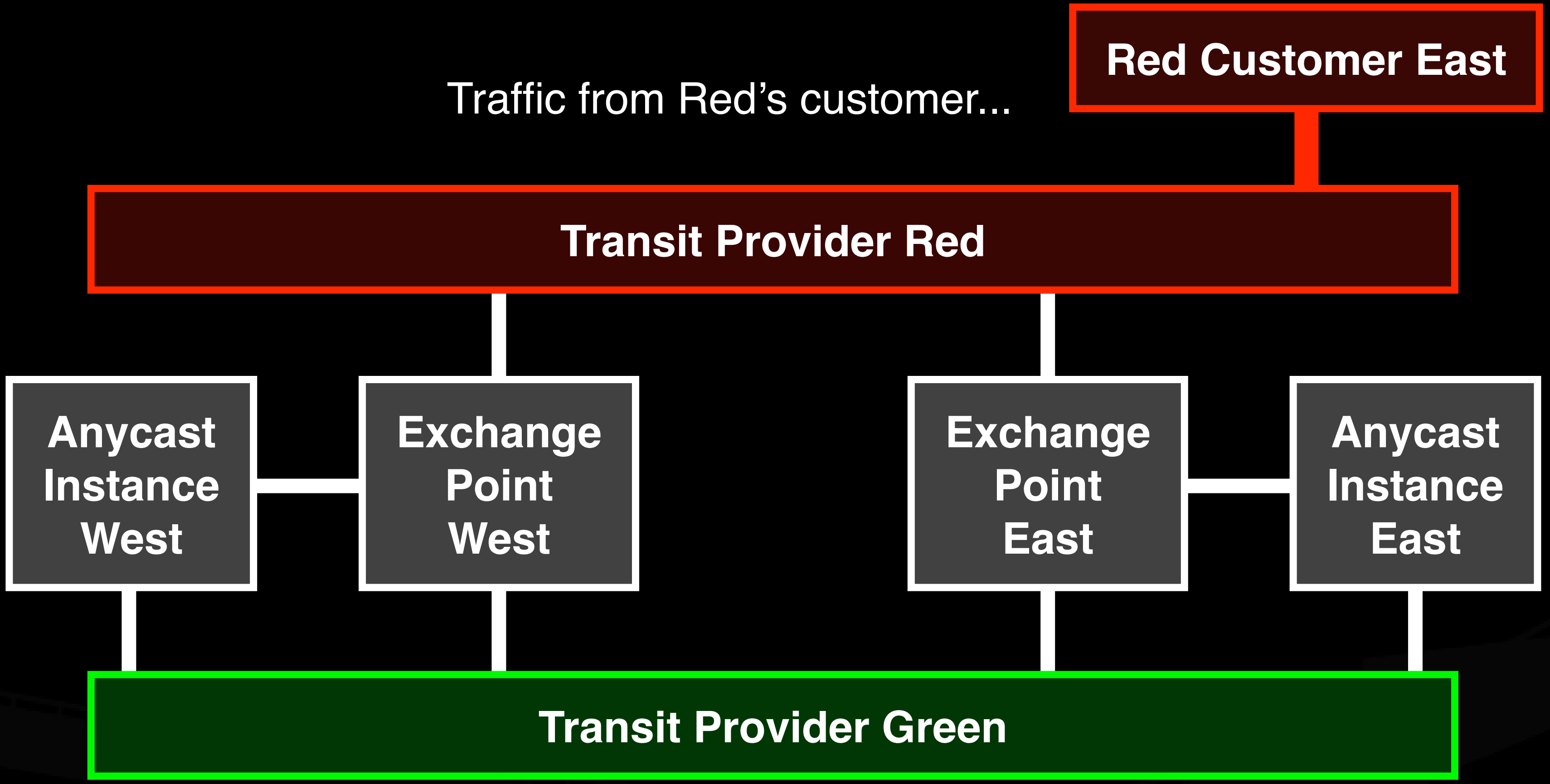
Normal Hot-Potato Routing

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

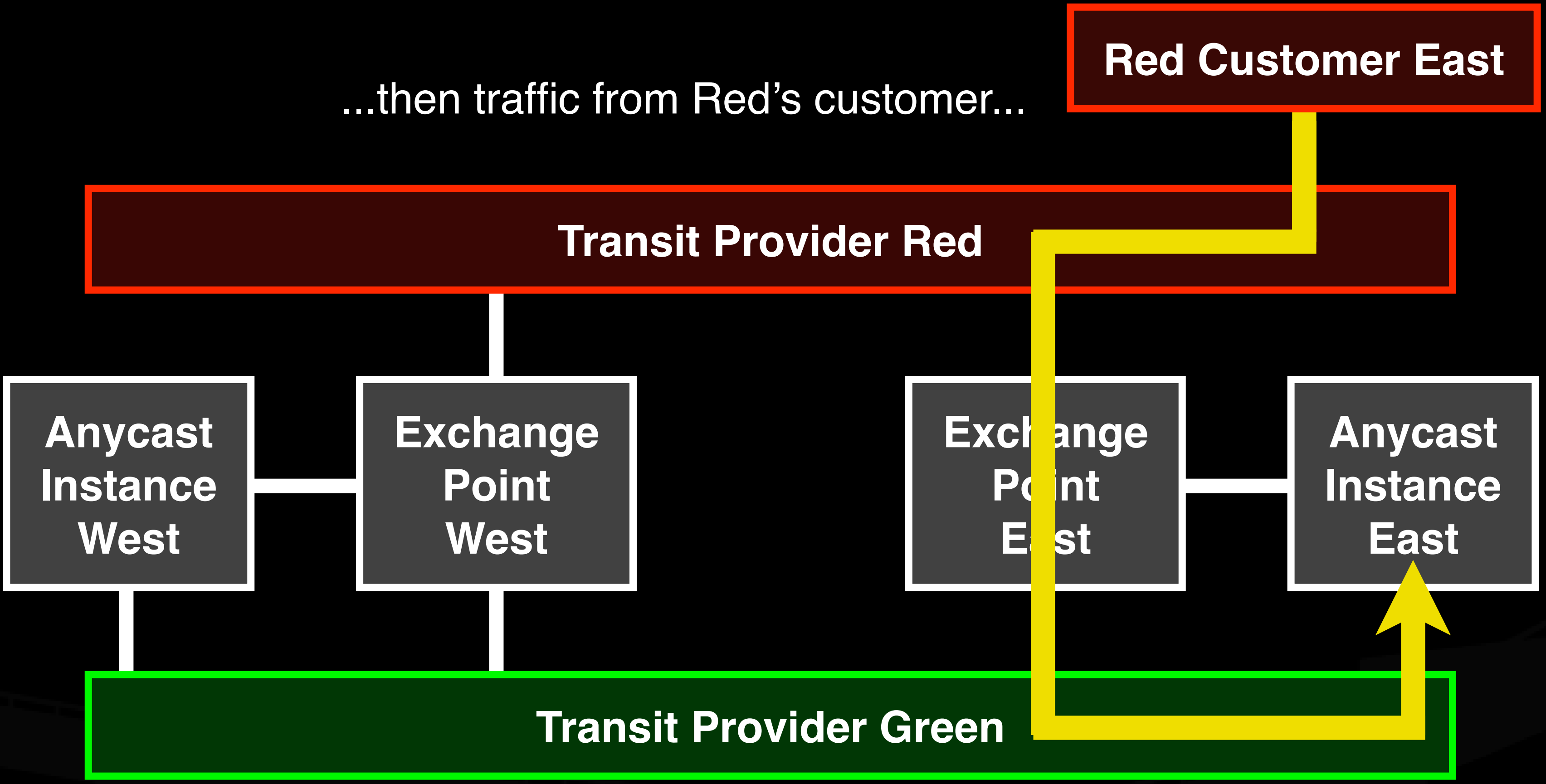


...but is a customer of large Transit Provider Green...

Normal Hot-Potato Routing



Normal Hot-Potato Routing



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

Red Customer East

Transit Provider Red

Anycast Instance West

Exchange Point West

Exchange Point East

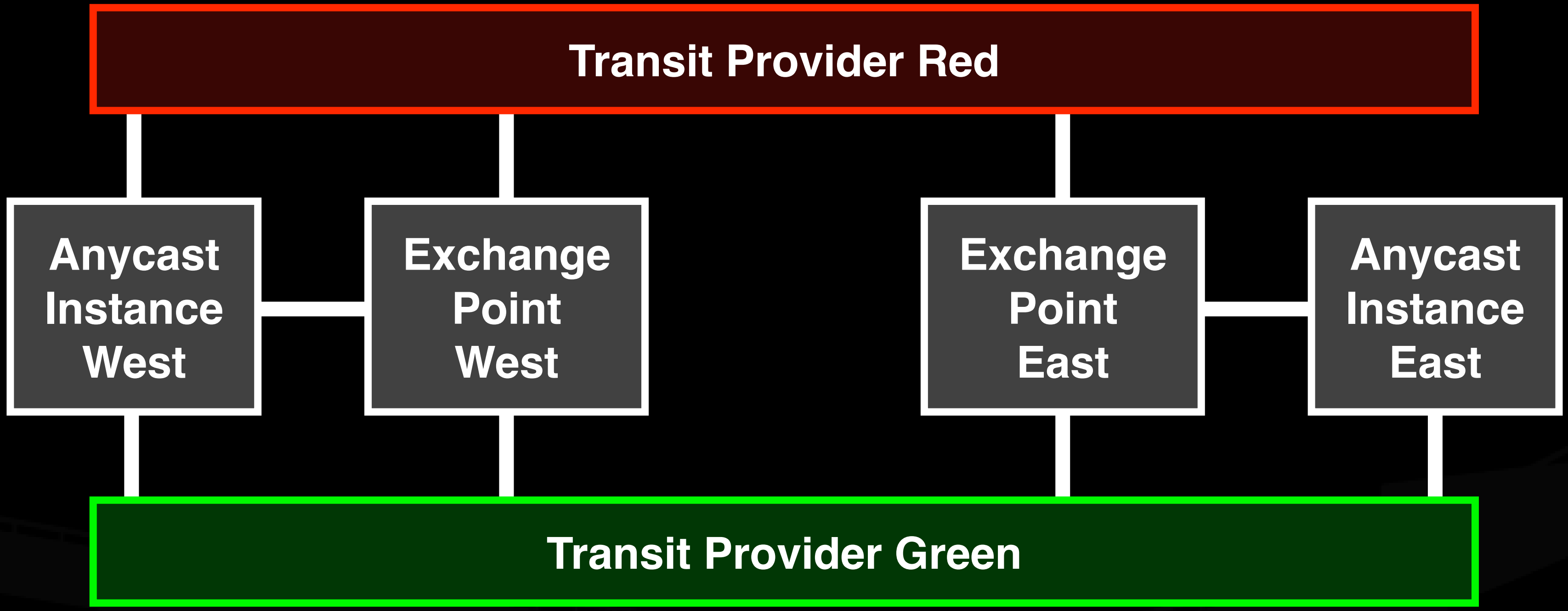
Anycast Instance East

Transit Provider Green

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

How the Conflict Plays Out

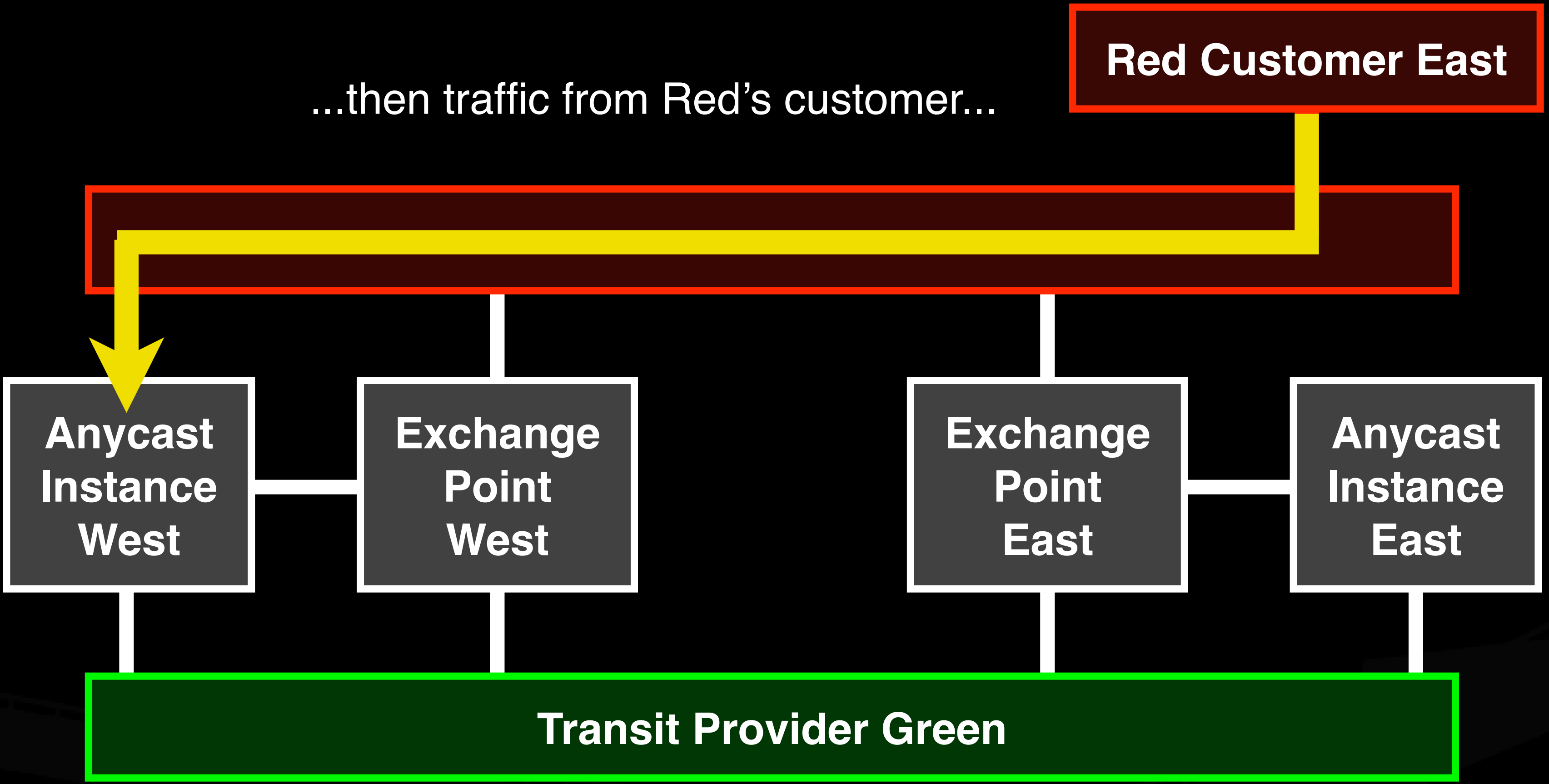
But if the anycast network is a customer of both large Transit Provider Red...



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

How the Conflict Plays Out

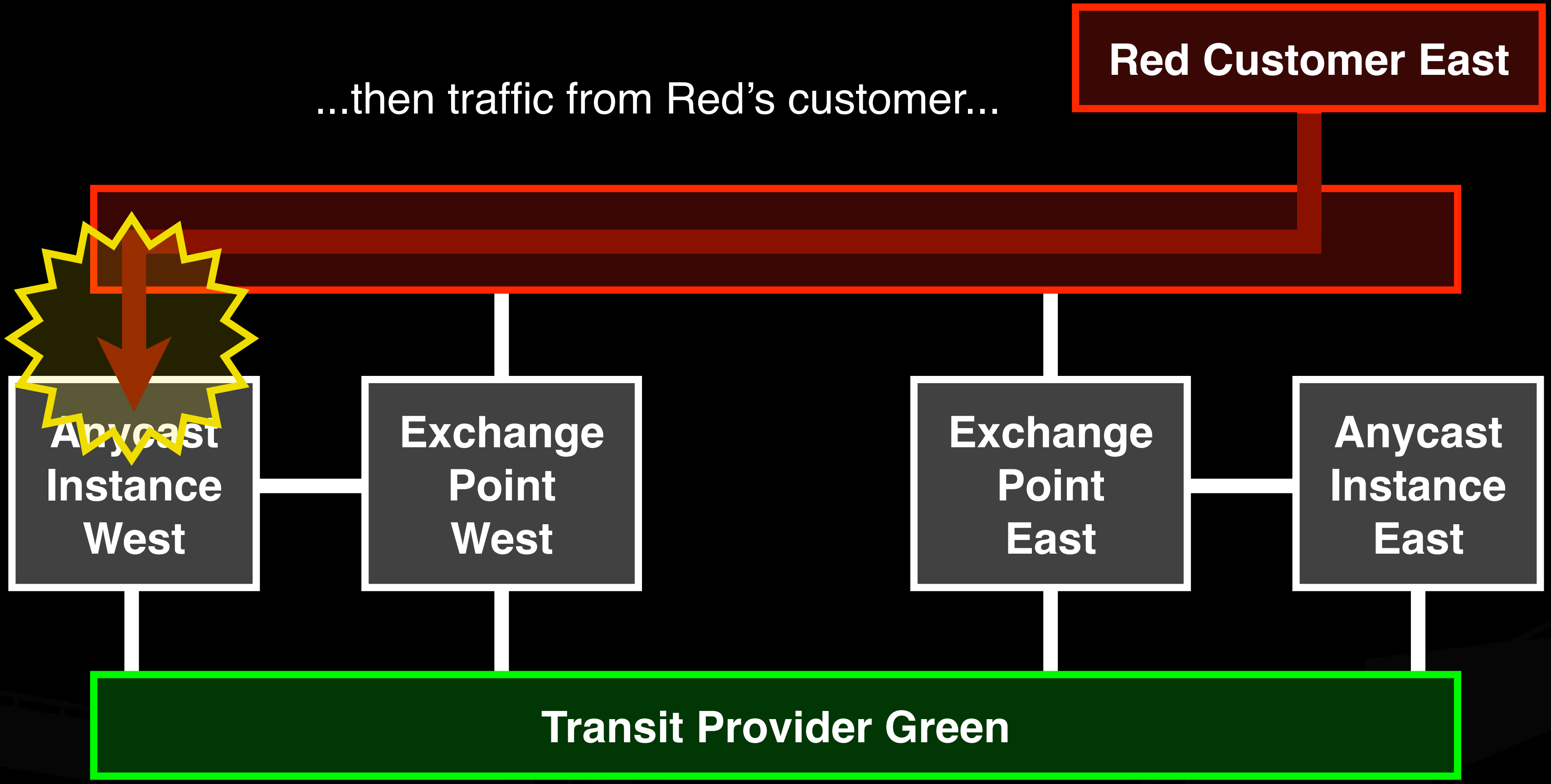
...then traffic from Red's customer...



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

Packet Clearing House

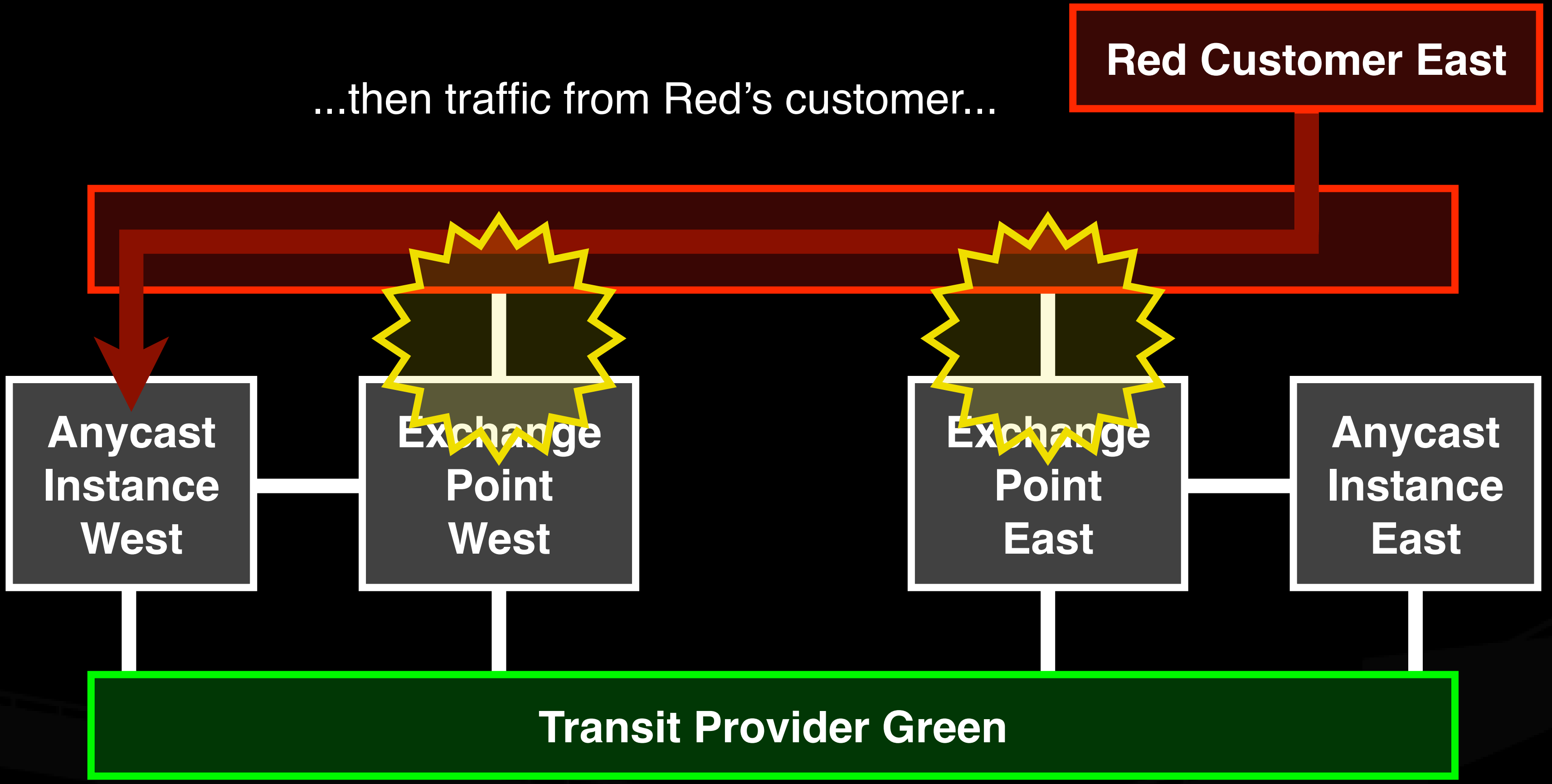
How the Conflict Plays Out



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

How the Conflict Plays Out

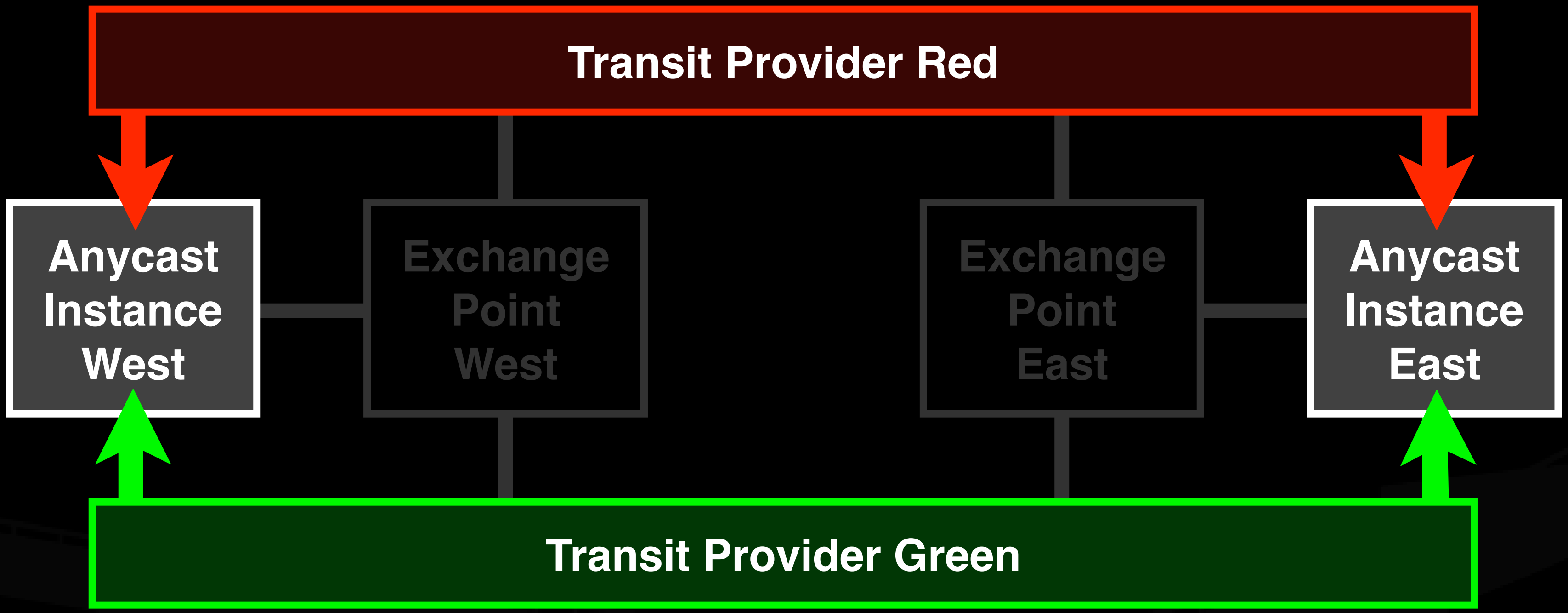
...then traffic from Red’s customer...



...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.

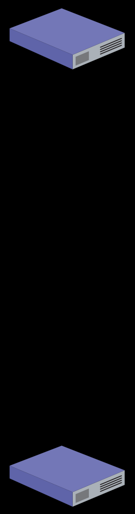


This caution is not necessary with small transit providers who don’t have the capability of backhauling traffic to the wrong region on the basis of policy.

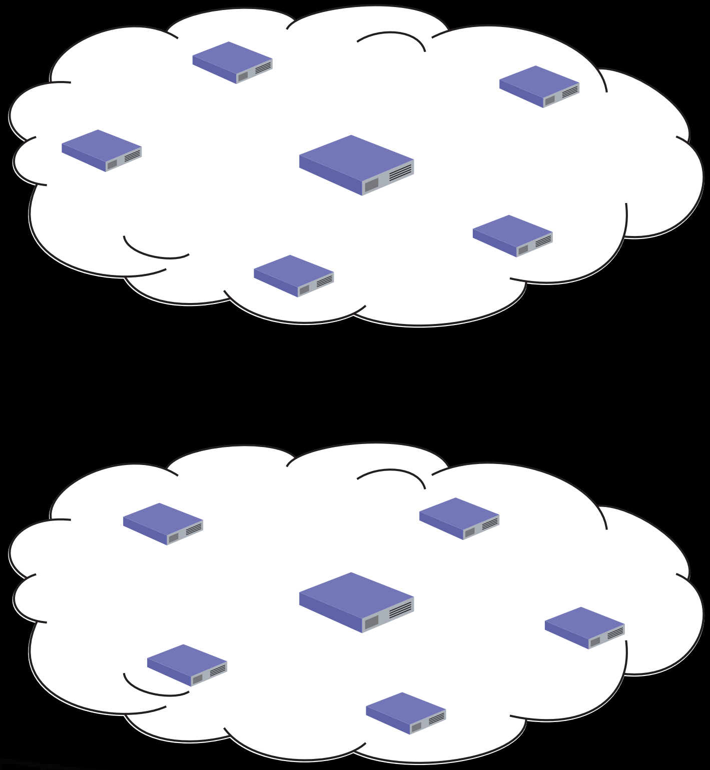
Putting the Pieces Together

- 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

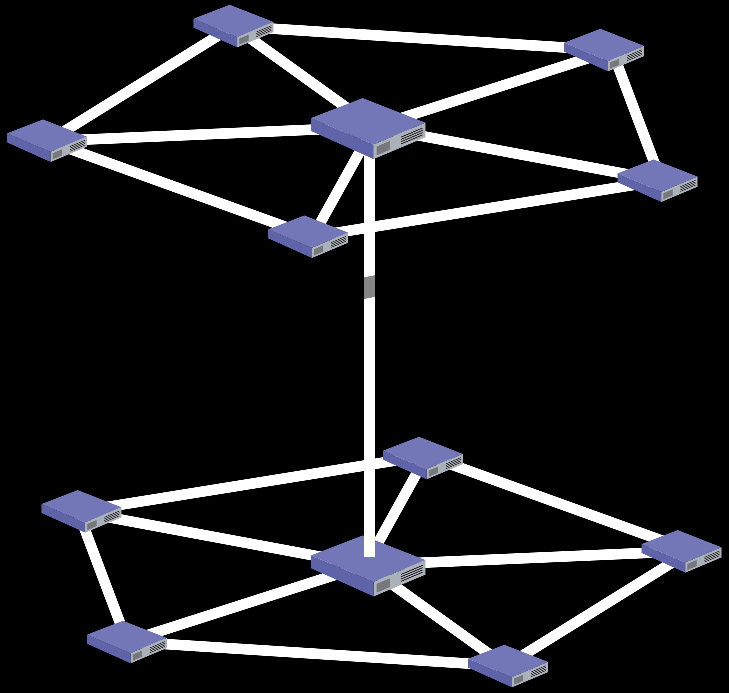


An A Cloud and a B Cloud



A Network Topology

“Dual Wagon-Wheel”

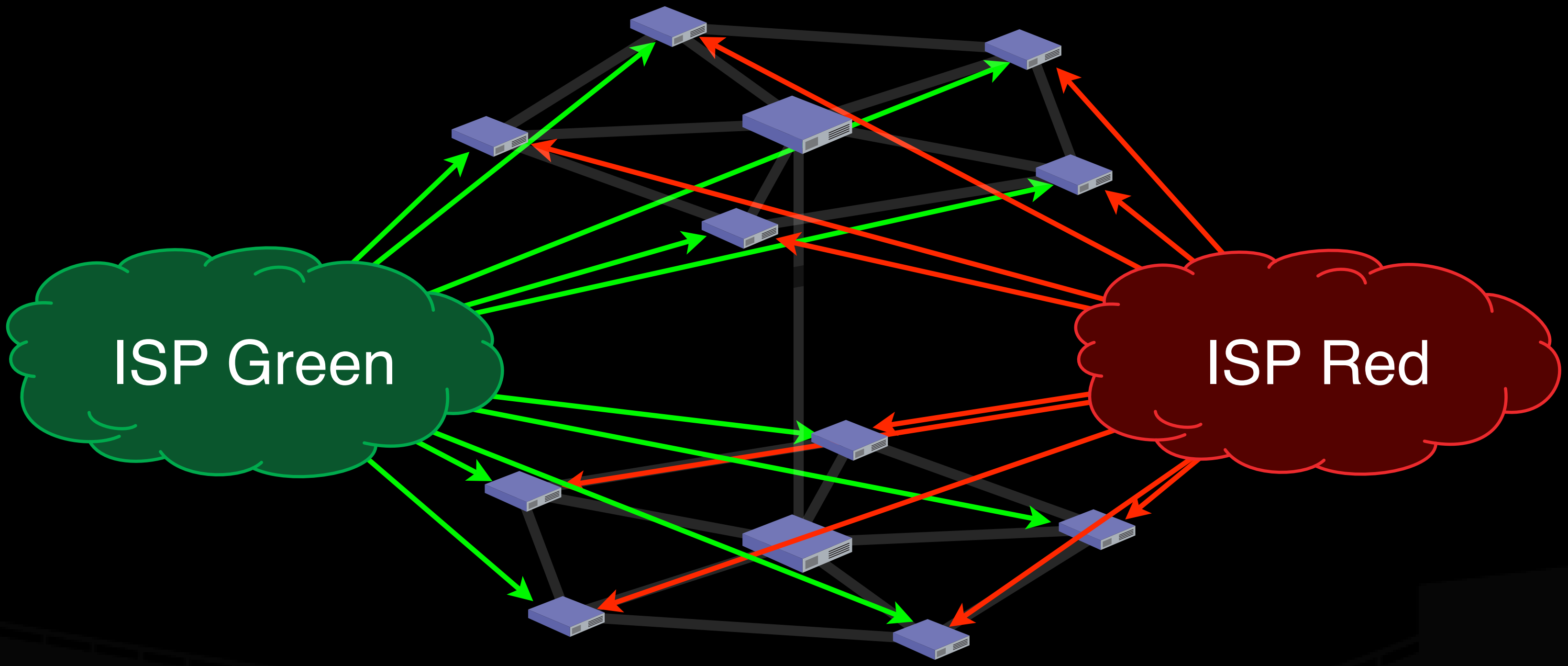


A Ring

B Ring

Redundant Transit

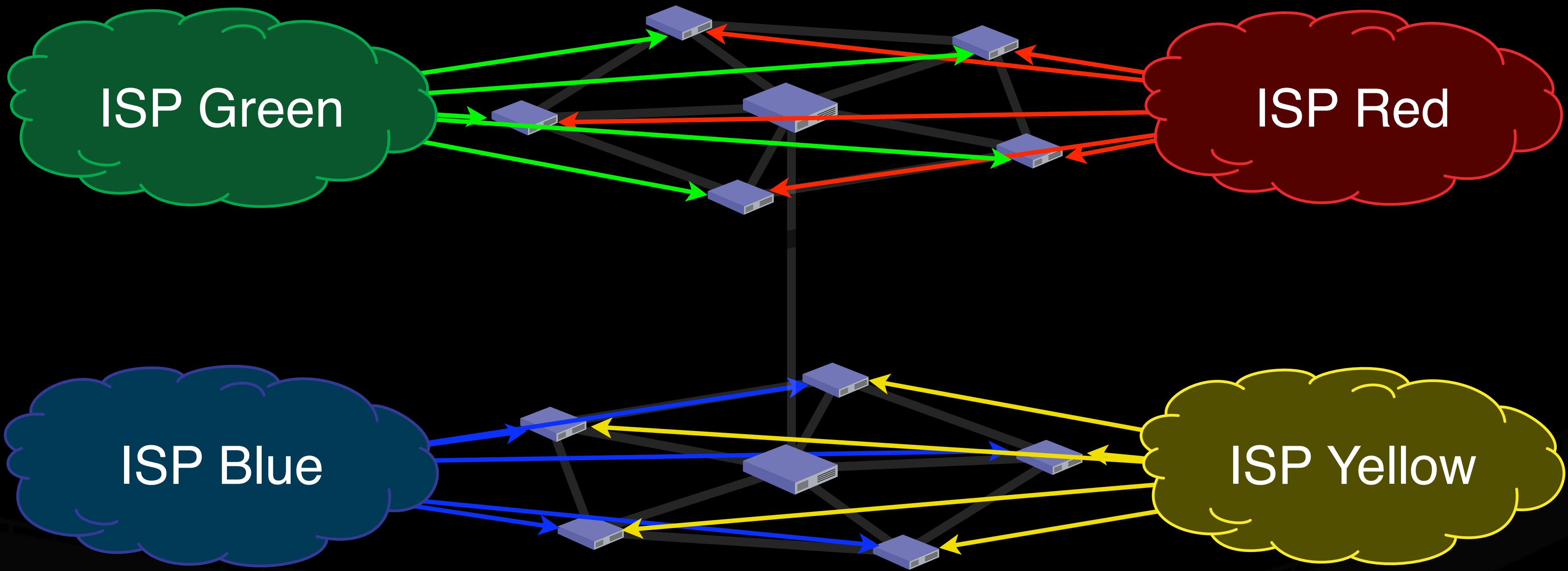
Two ISPs



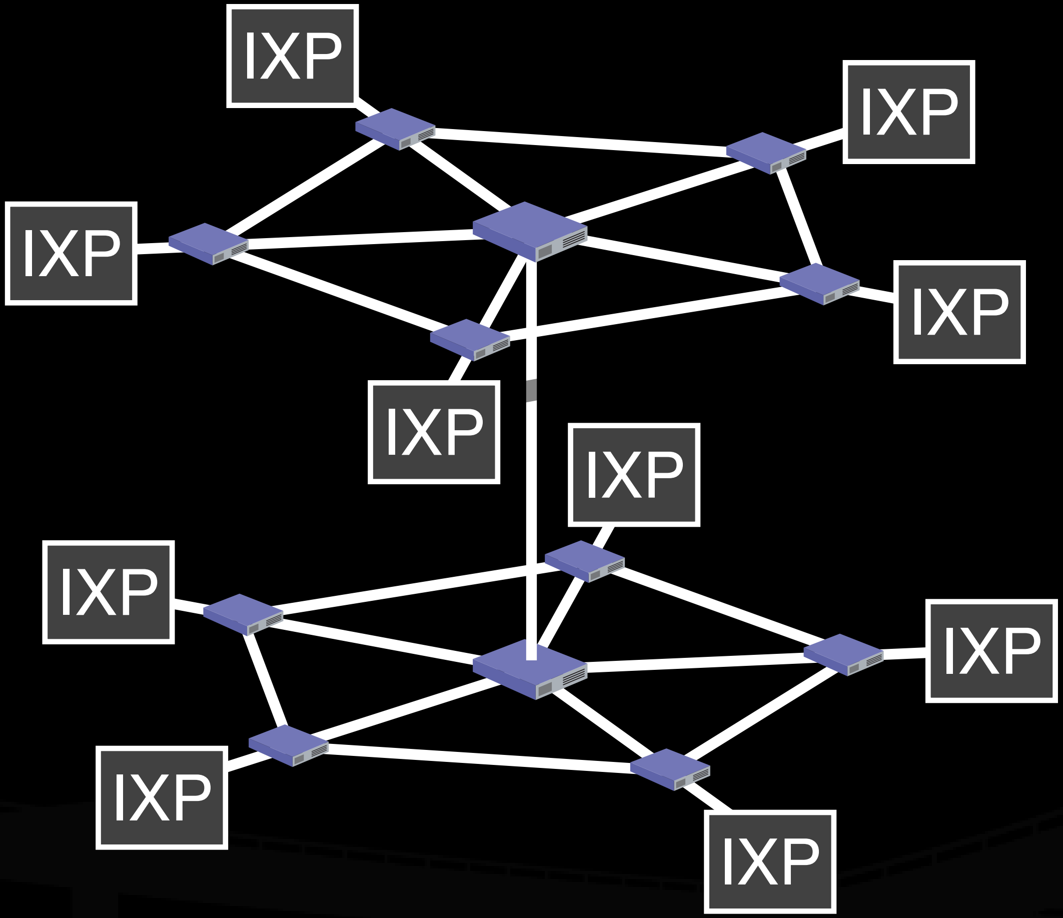
Packet Clearing House

Redundant Transit

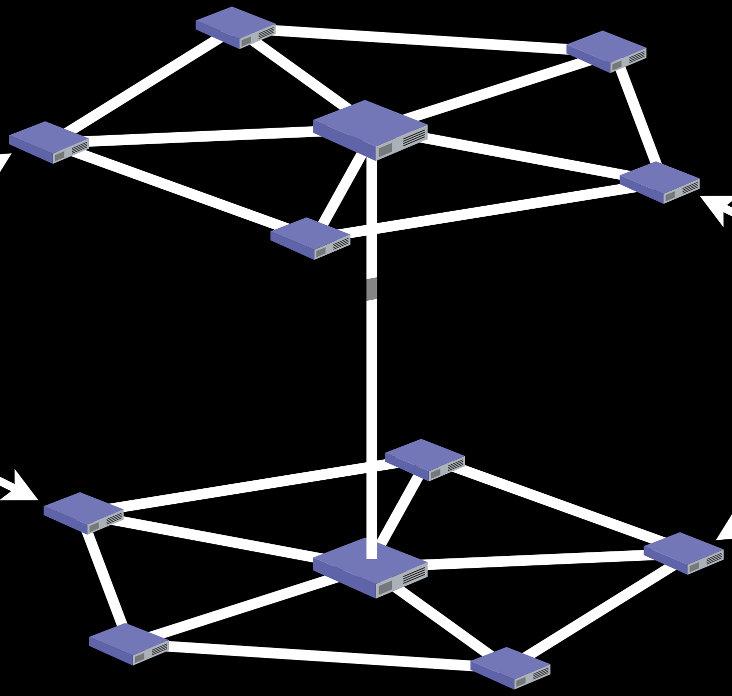
Or four ISPs



Local Peering



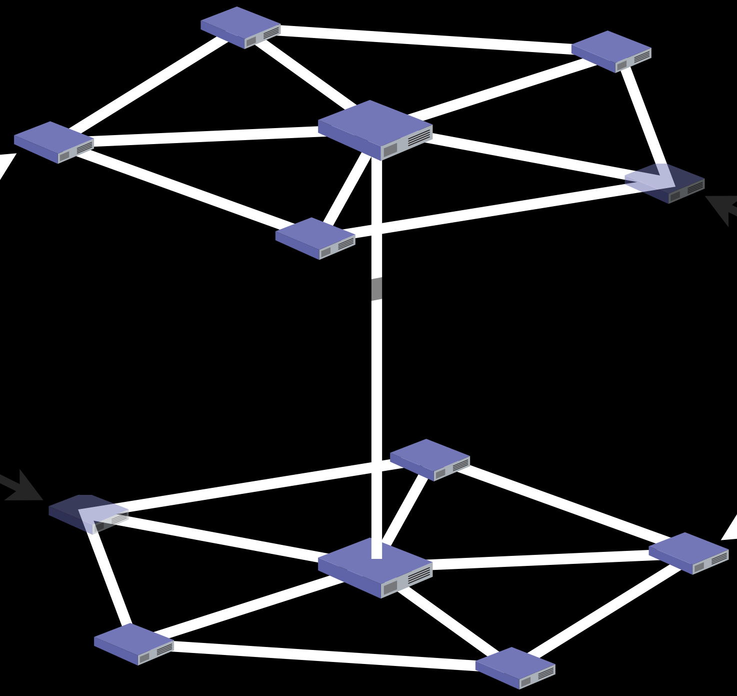
Resolver-Based Fail-Over



Customer Resolver Server Selection

Customer Resolver Server Selection

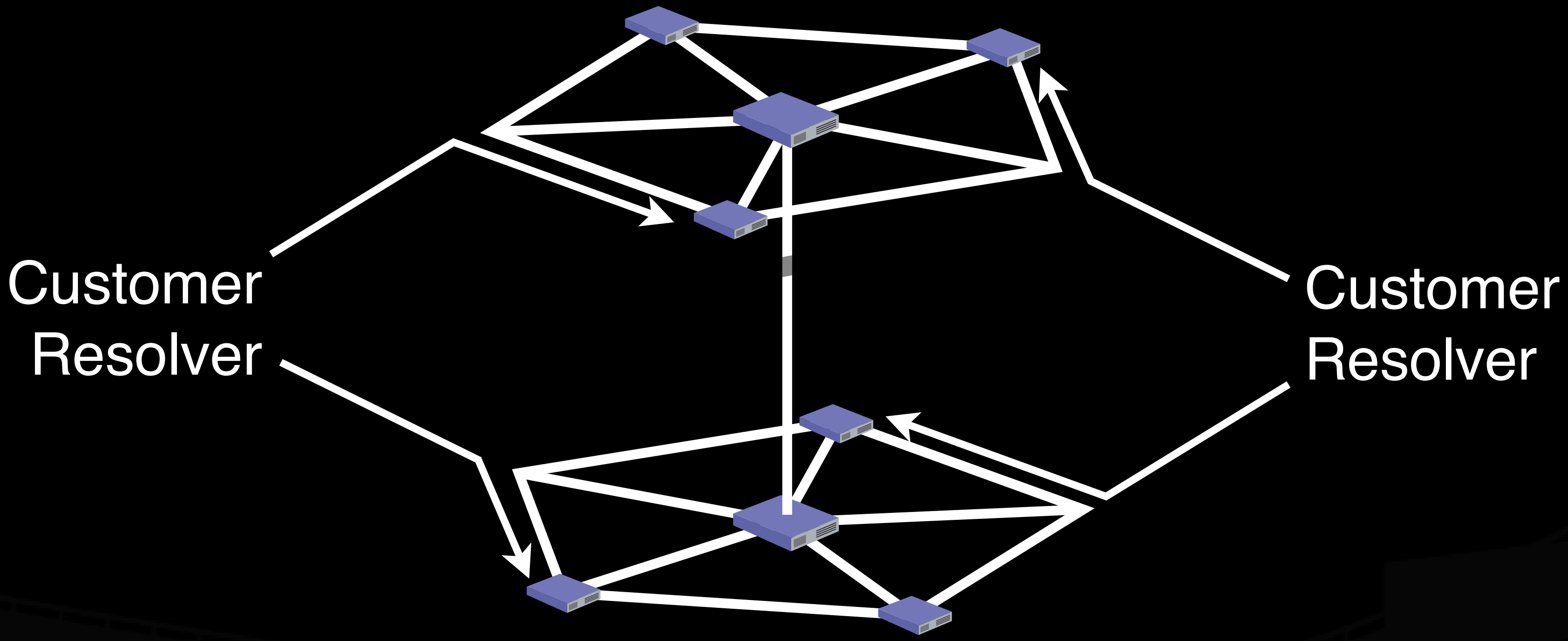
Resolver-Based Fail-Over



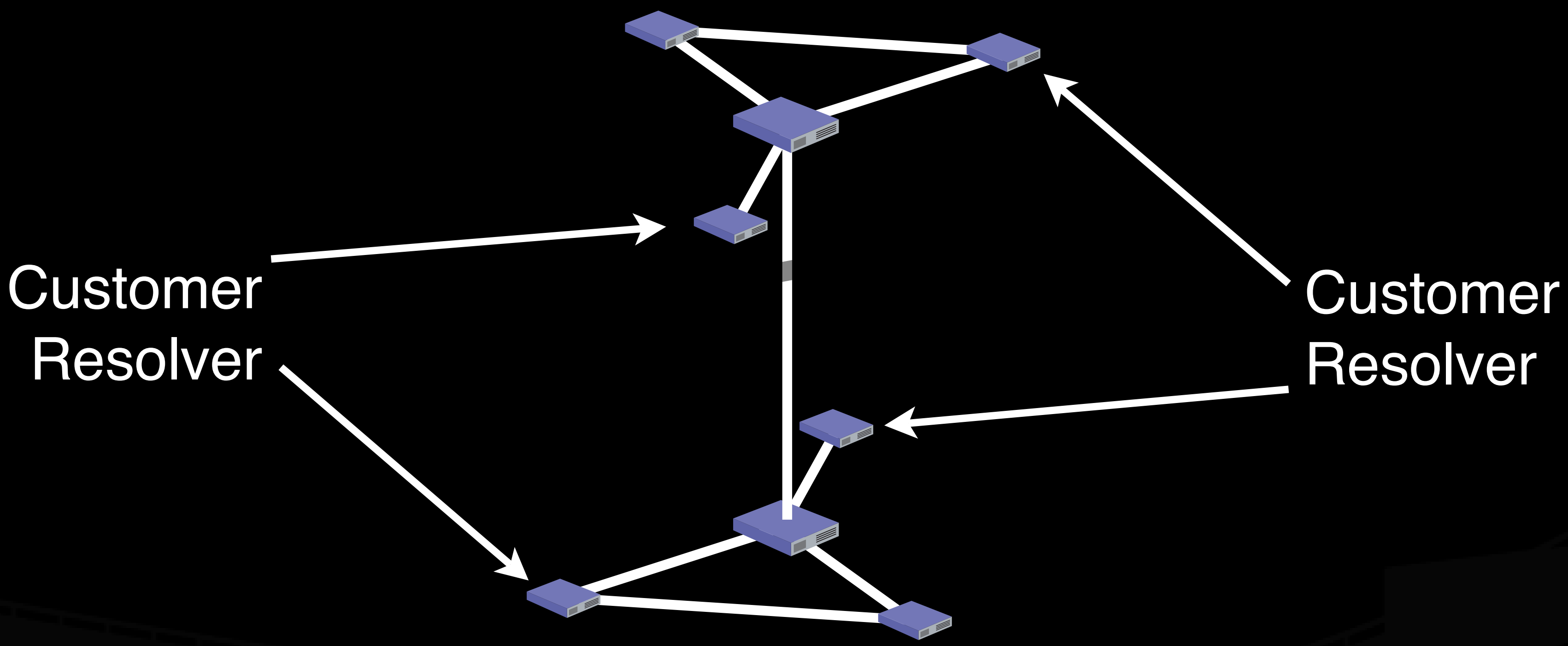
Customer
Resolver
Server
Selection

Customer
Resolver
Server
Selection

Internal Anycast Fail-Over



Global Anycast Fail-Over



Thanks, and Questions?

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

http:// www.pch.net / resources / papers / dns-service-architecture

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