Best Practices in IPv4 Anycast Routing

Gaurab Raj Upadhaya

Packet Clearing House

What *isn't* Anycast?

- Not a protocol, not a different version of IP, nobody's proprietary technology.
- Doesn't require any special capabilities in the servers, clients, or network.
- Doesn't break or confuse existing infrastructure.

Packet Clearing House

What *is* Anycast?

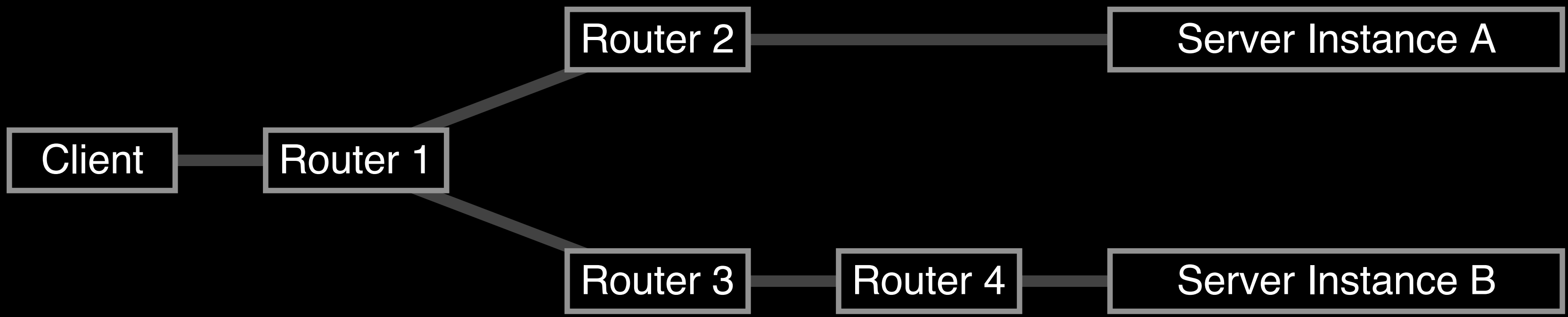
- Just a configuration methodology.
- Mentioned, although not described in detail, in numerous RFCs since time immemorial.
- It's been the basis for large-scale content-distribution networks since at least 1995.
- It's gradually taking over the core of the DNS infrastructure, as well as much of the periphery of the world wide web.

Packet Clearing House

How Does Anycast Work?

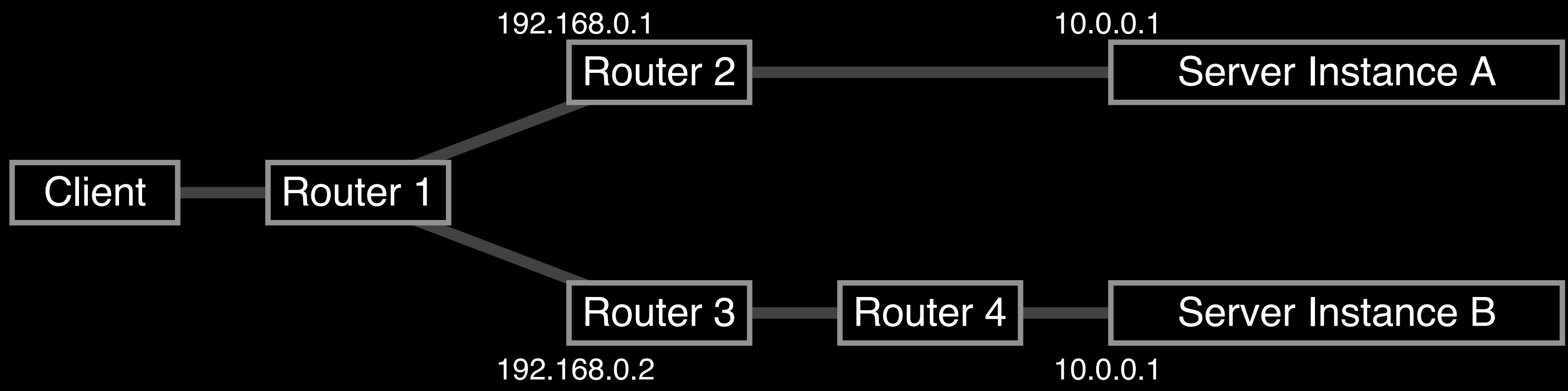
- The basic idea is extremely simple:
- Multiple instances of a service share the same IP address.
- The routing infrastructure directs any packet to the topologically nearest instance of the service.
- What little complexity exists is in the optional details.

Example



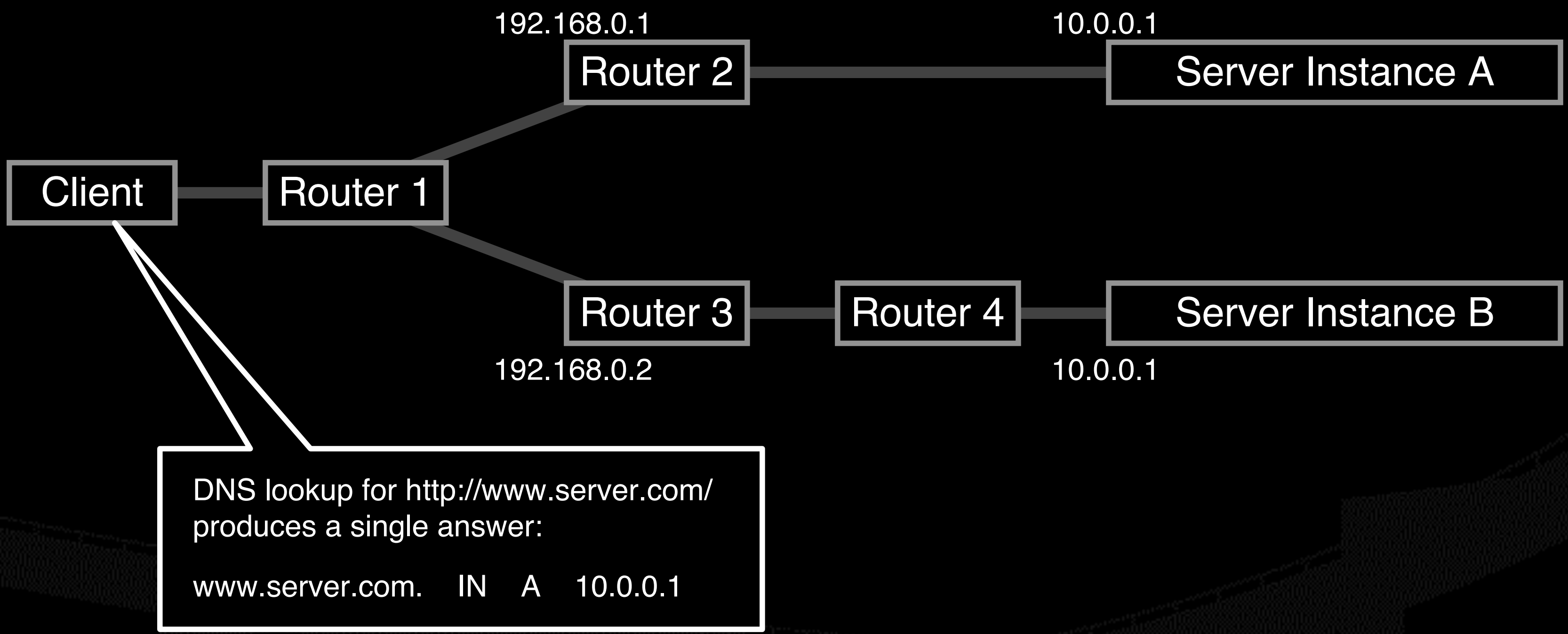
Router 2
Server Instance A
Client
Router 1
Router 3
Router 4
Server Instance B

Example

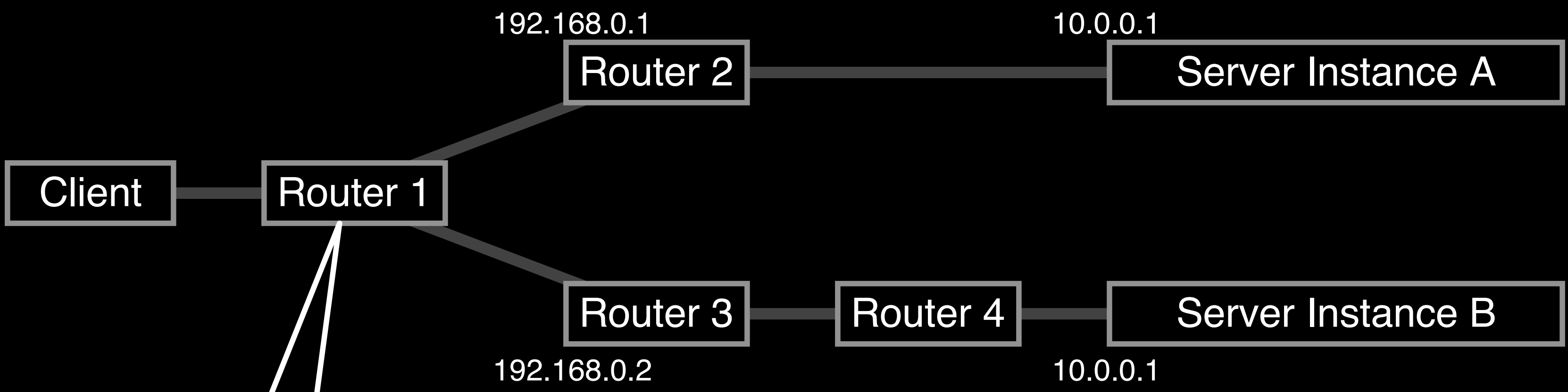


192.168.0.1
10.0.0.1
Router 2
Server Instance A
Client
Router 1
Router 3
Router 4
Server Instance B
192.168.0.2
10.0.0.1

Example


192.168.0.1
Router 2
10.0.0.1
Server Instance A
Client
Router 1
Router 3
Router 4
Server Instance B
192.168.0.2
10.0.0.1
DNS lookup for http://www.server.com/ produces a single answer:
www.server.com. IN A 10.0.0.1

Example

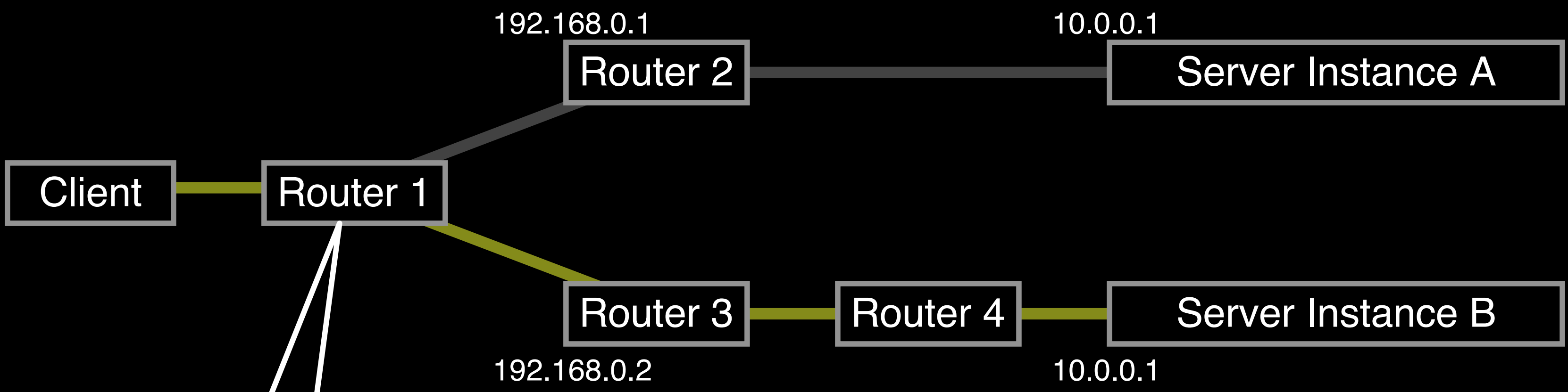


Routing Table from Router 1:

Destination	Mask	Next-Hop	Distance
192.168.0.0	/29	127.0.0.1	0
10.0.0.1	/32	192.168.0.1	1
10.0.0.1	/32	192.168.0.2	2

Packet Clearing House

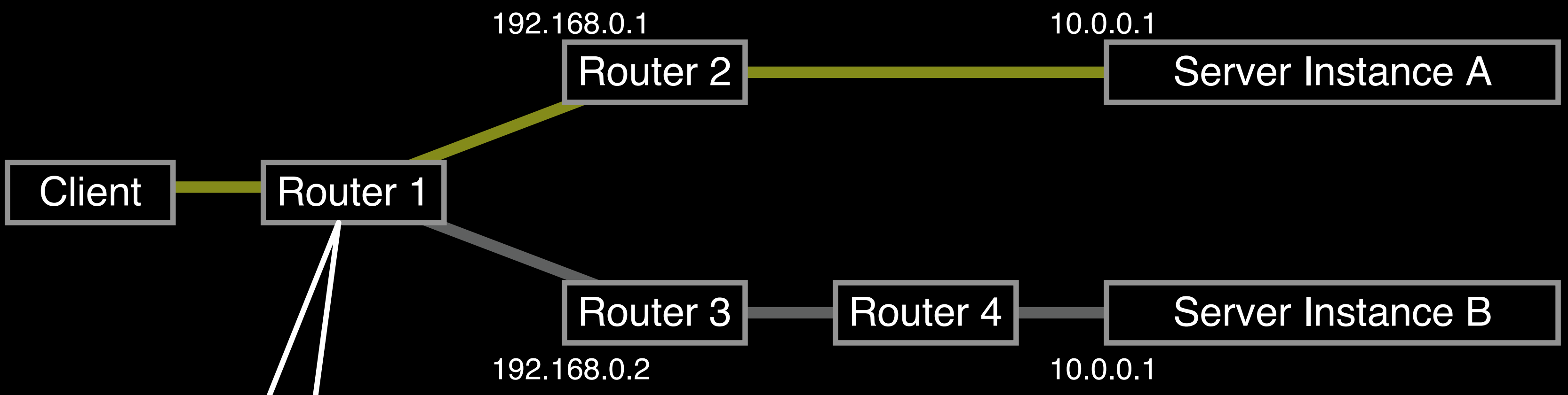
Example



Routing Table from Router 1:

Destination	Mask	Next-Hop	Distance
192.168.0.0	/29	127.0.0.1	0
10.0.0.1	/32	192.168.0.1	1
10.0.0.1	/32	192.168.0.2	2

Example

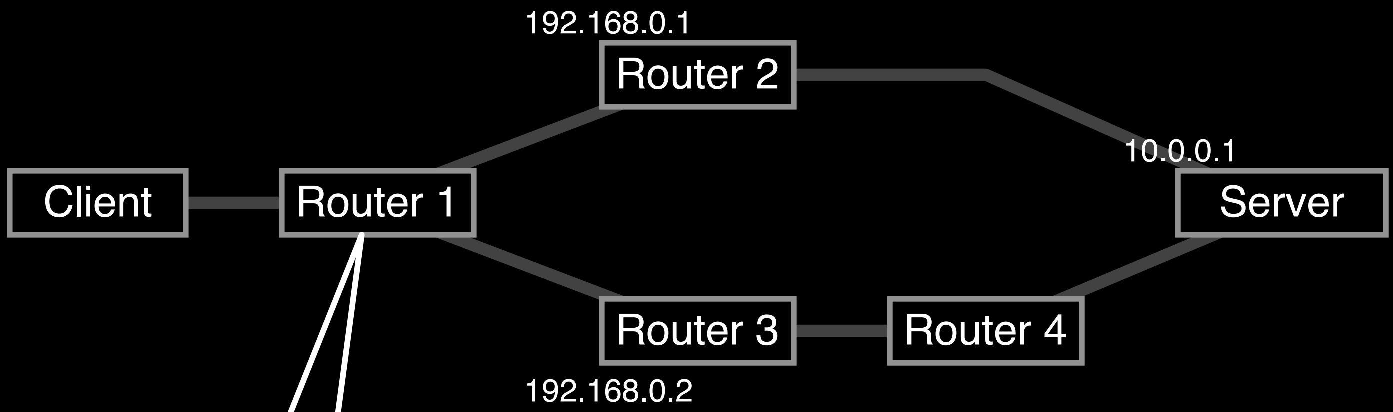


Routing Table from Router 1:

Destination	Mask	Next-Hop	Distance
192.168.0.0	/29	127.0.0.1	0
10.0.0.1	/32	192.168.0.1	1
10.0.0.1	/32	192.168.0.2	2

Example

What the routers think the topology looks like:



Routing Table from Router 1:

Destination	Mask	Next-Hop	Distance
192.168.0.0	/29	127.0.0.1	0
10.0.0.1	/32	192.168.0.1	1
10.0.0.1	/32	192.168.0.2	2

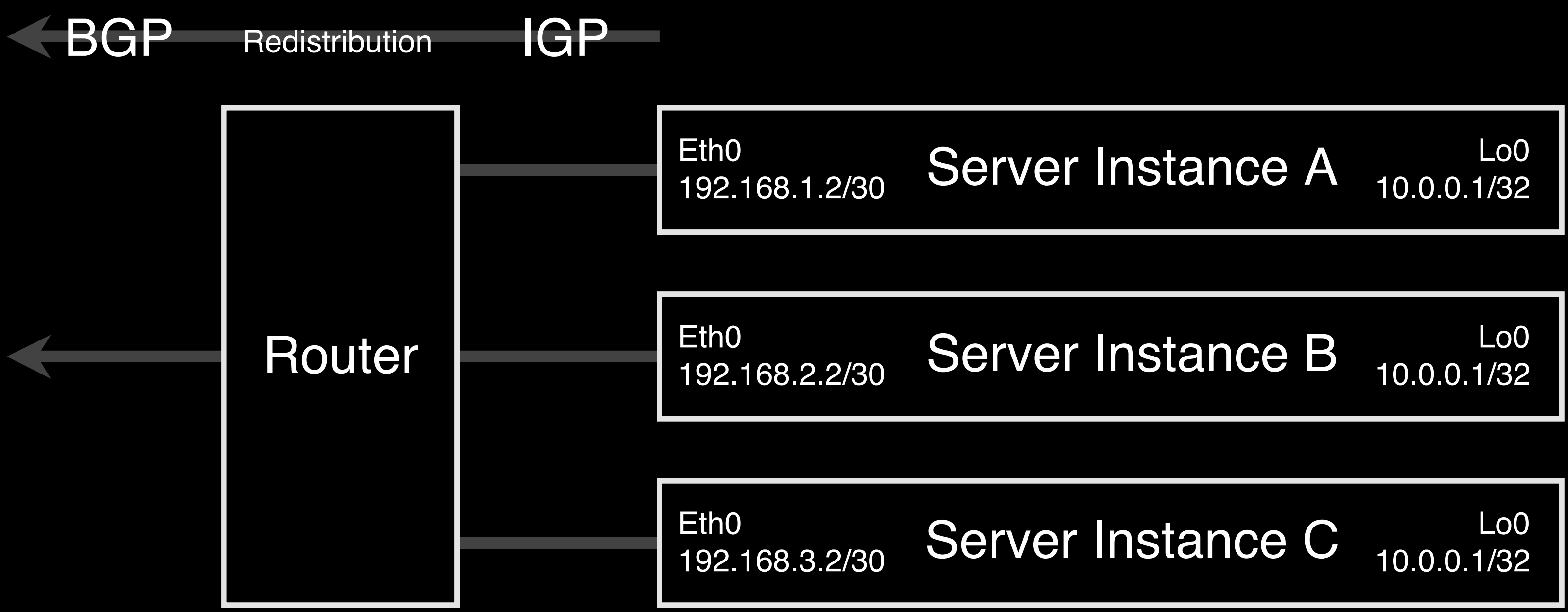
Building an Anycast Server Cluster

- Anycast can be used in building either local server clusters, or global networks, or global networks of clusters, combining both scales.
- F-root is a local anycast server cluster, for instance.

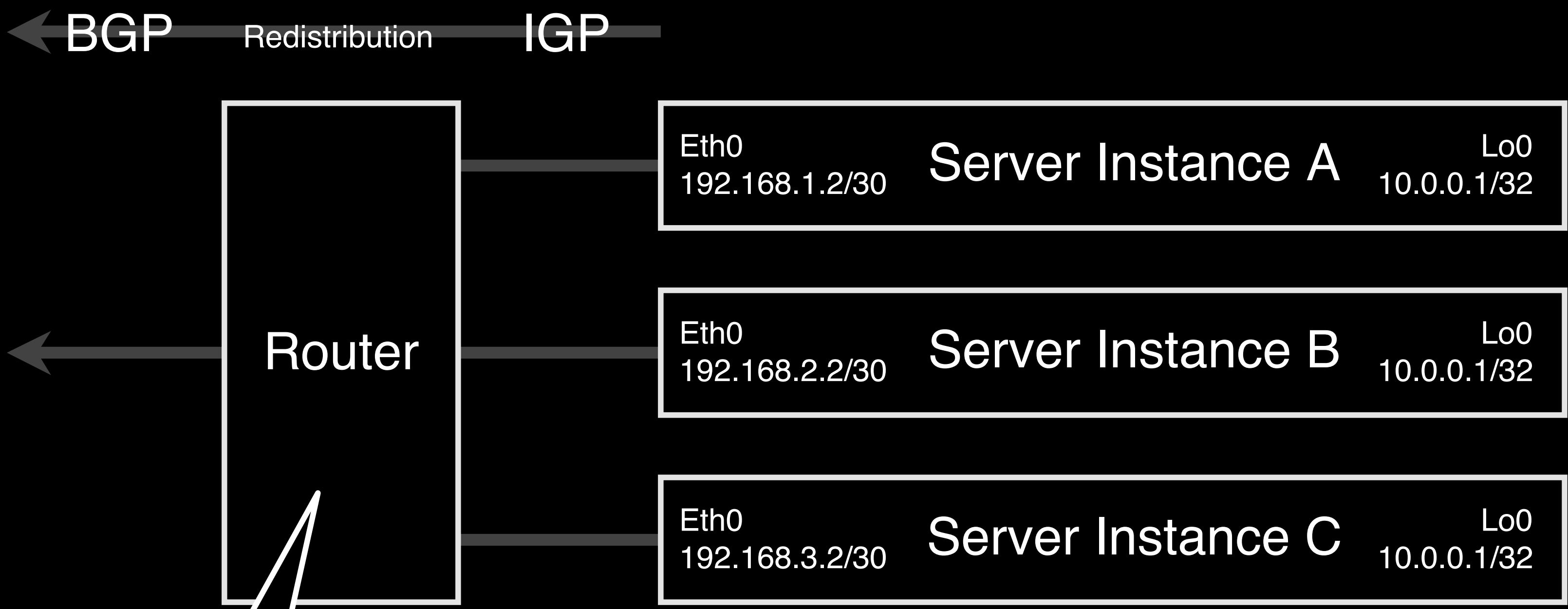
Building an Anycast Server Cluster

- Typically, a cluster of servers share a common virtual interface attached to their loopback devices, and speak an IGP routing protocol to an adjacent BGP-speaking border router.
- The servers may or may not share identical content.

Example

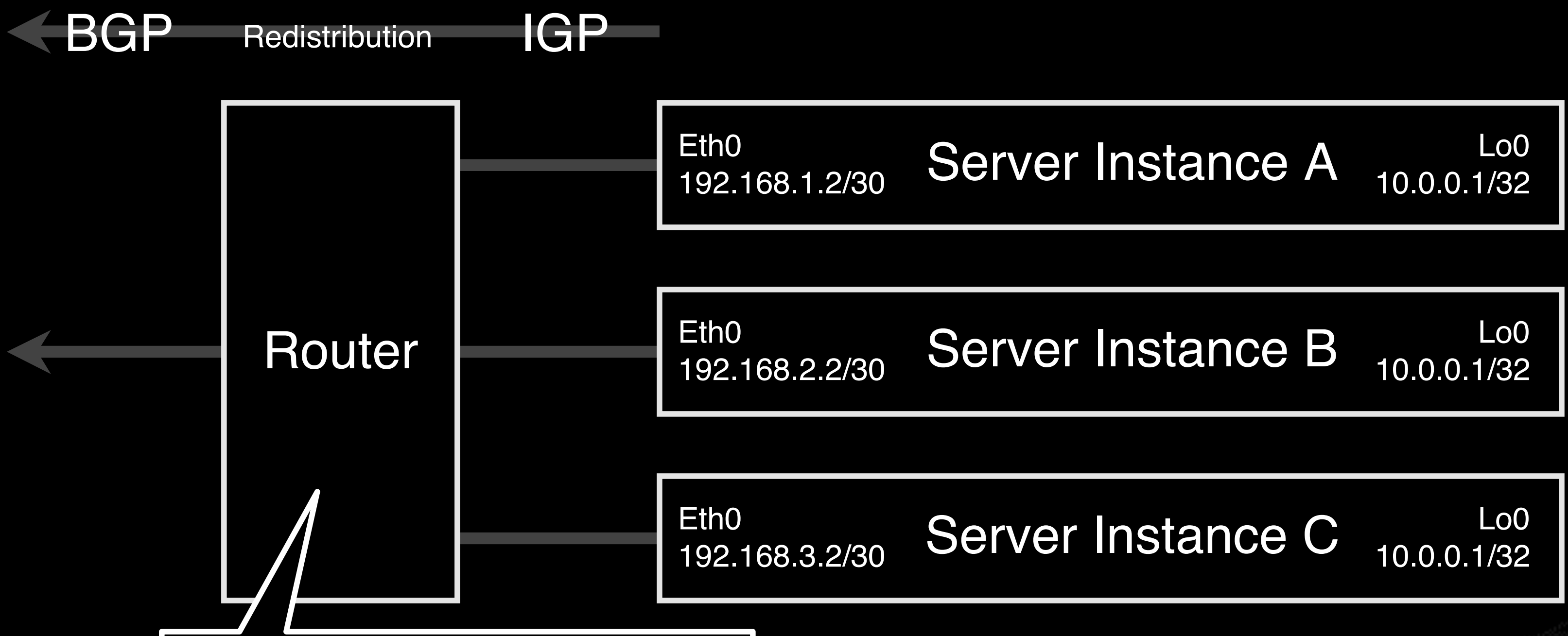

BGP
Redistribution
IGP
Router
Eth0
192.168.1.2/30
Server Instance A
Lo0
10.0.0.1/32
Eth0
192.168.2.2/30
Server Instance B
Lo0
10.0.0.1/32
Eth0
192.168.3.2/30
Server Instance C
Lo0
10.0.0.1/32

Example



Destination	Mask	Next-Hop	Dist
0.0.0.0	/0	127.0.0.1	0
192.168.1.0	/30	192.168.1.1	0
192.168.2.0	/30	192.168.2.1	0
192.168.3.0	/30	192.168.3.1	0
10.0.0.1	/32	192.168.1.2	1
10.0.0.1	/32	192.168.2.2	1
10.0.0.1	/32	192.168.3.2	1

Example



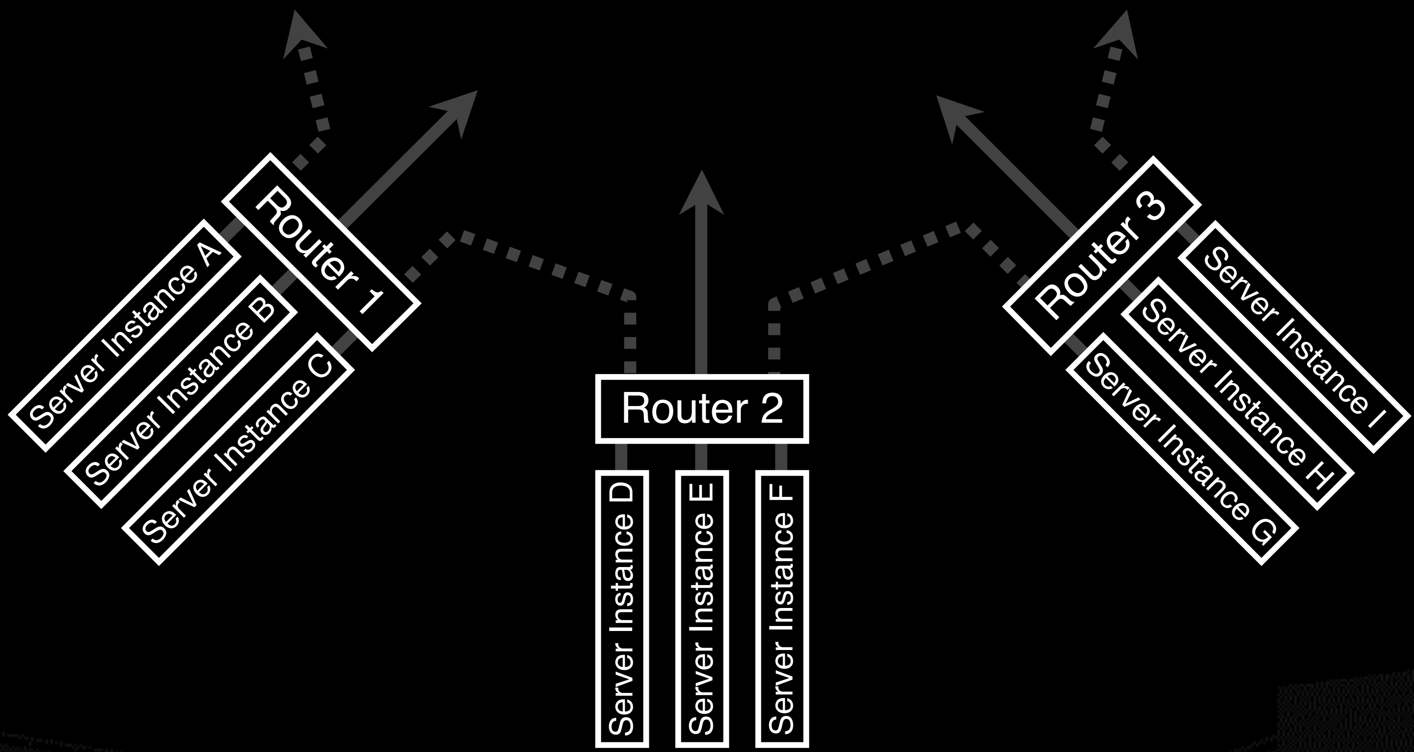
Destination	Mask	Next-Hop	Dist
0.0.0.0	/0	127.0.0.1	0
192.168.1.0	/30	192.168.1.1	0
192.168.2.0	/30	192.168.2.1	0
192.168.3.0	/30	192.168.3.1	0
10.0.0.1	/32	192.168.1.2	1
10.0.0.1	/32	192.168.2.2	1
10.0.0.1	/32	192.168.3.2	1

} Round-robin load balancing (last three rows)

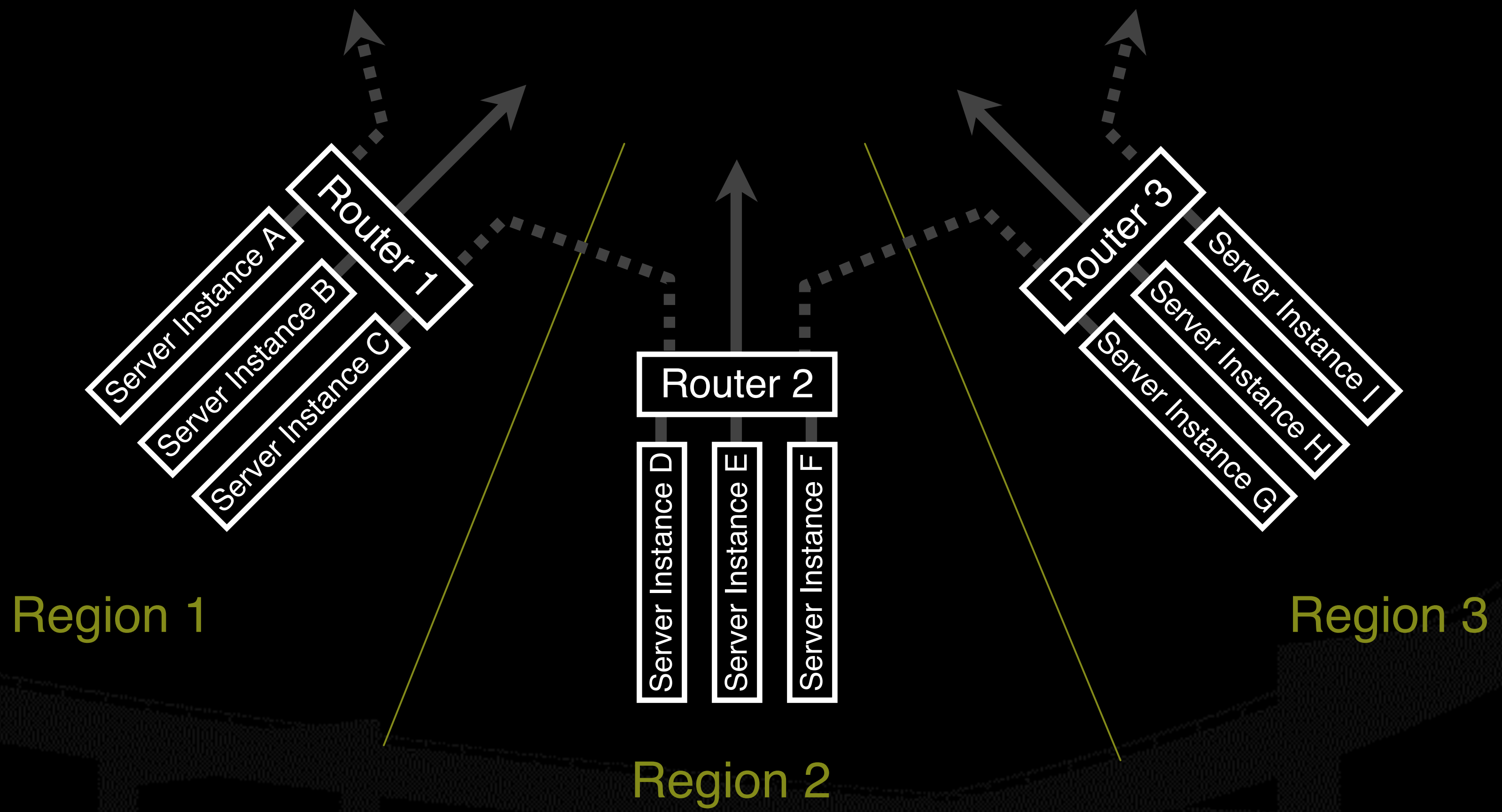
Building a Global Network of Clusters

- Once a cluster architecture has been established, additional clusters can be added to gain performance.
- Load distribution, fail-over between clusters, and content synchronization become the principal engineering concerns.

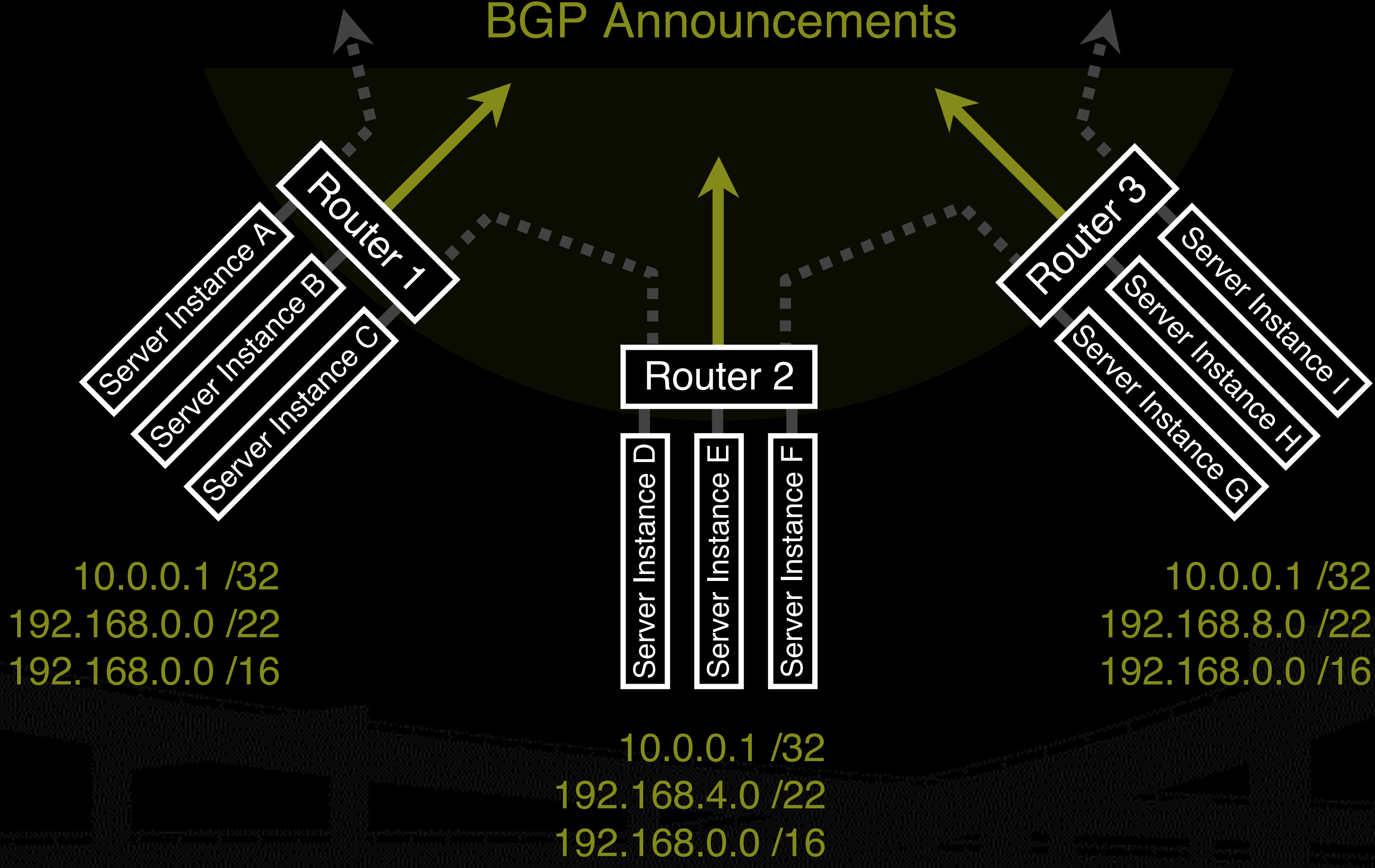
Example


Router 1
Server Instance A
Server Instance B
Server Instance C
Router 2
Server Instance D
Server Instance E
Server Instance F
Router 3
Server Instance G
Server Instance H
Server Instance I

Example


Router 1
Server Instance A
Server Instance B
Server Instance C
Router 2
Server Instance D
Server Instance E
Server Instance F
Router 3
Server Instance G
Server Instance H
Server Instance I
Region 1
Region 2
Region 3

Example

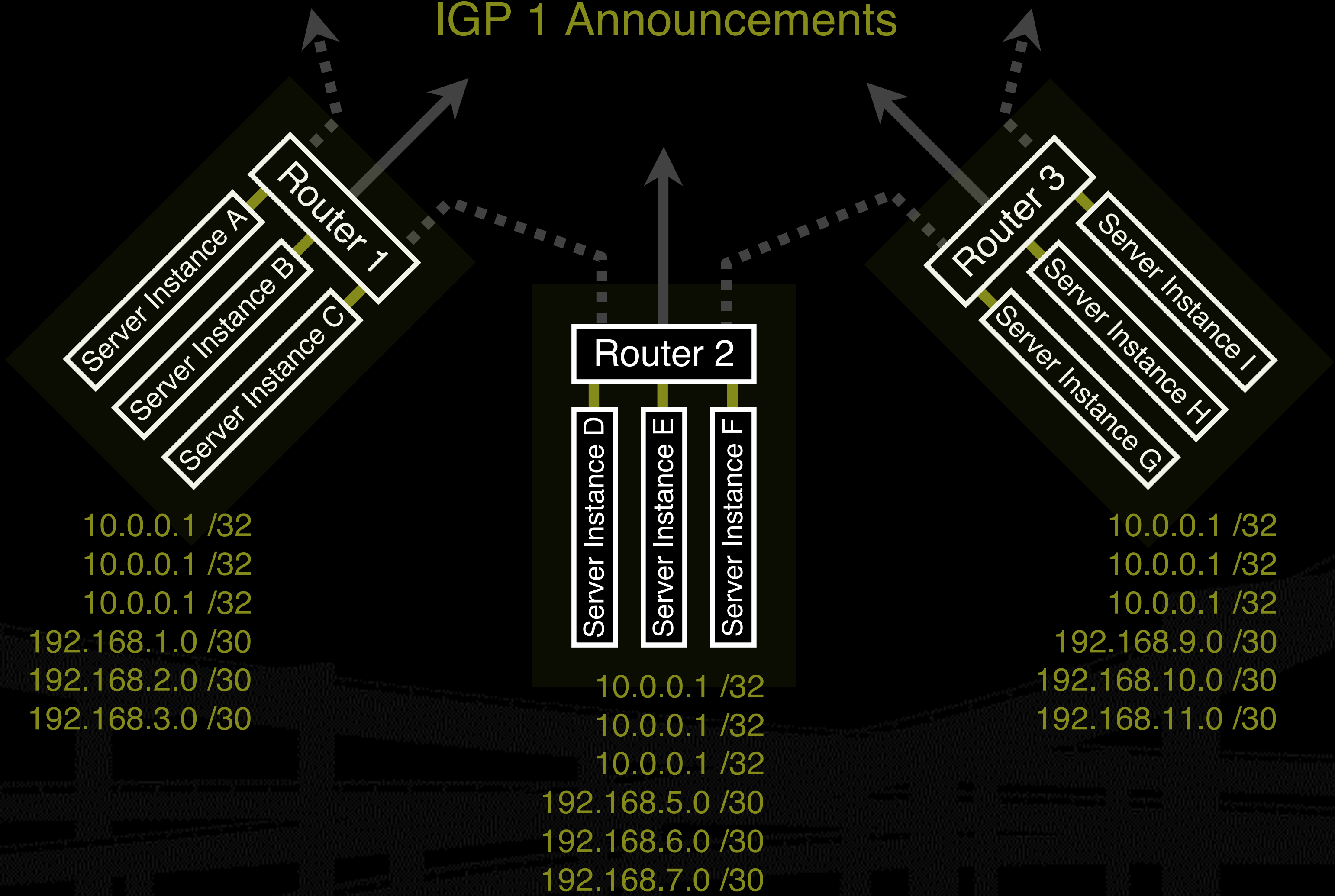


BGP Announcements
Router 1
Router 2
Router 3
Server Instance A
Server Instance B
Server Instance C
Server Instance D
Server Instance E
Server Instance F
Server Instance G
Server Instance H
Server Instance I
10.0.0.1 /32
192.168.0.0 /22
192.168.0.0 /16
10.0.0.1 /32
192.168.4.0 /22
192.168.0.0 /16
10.0.0.1 /32
192.168.8.0 /22
192.168.0.0 /16

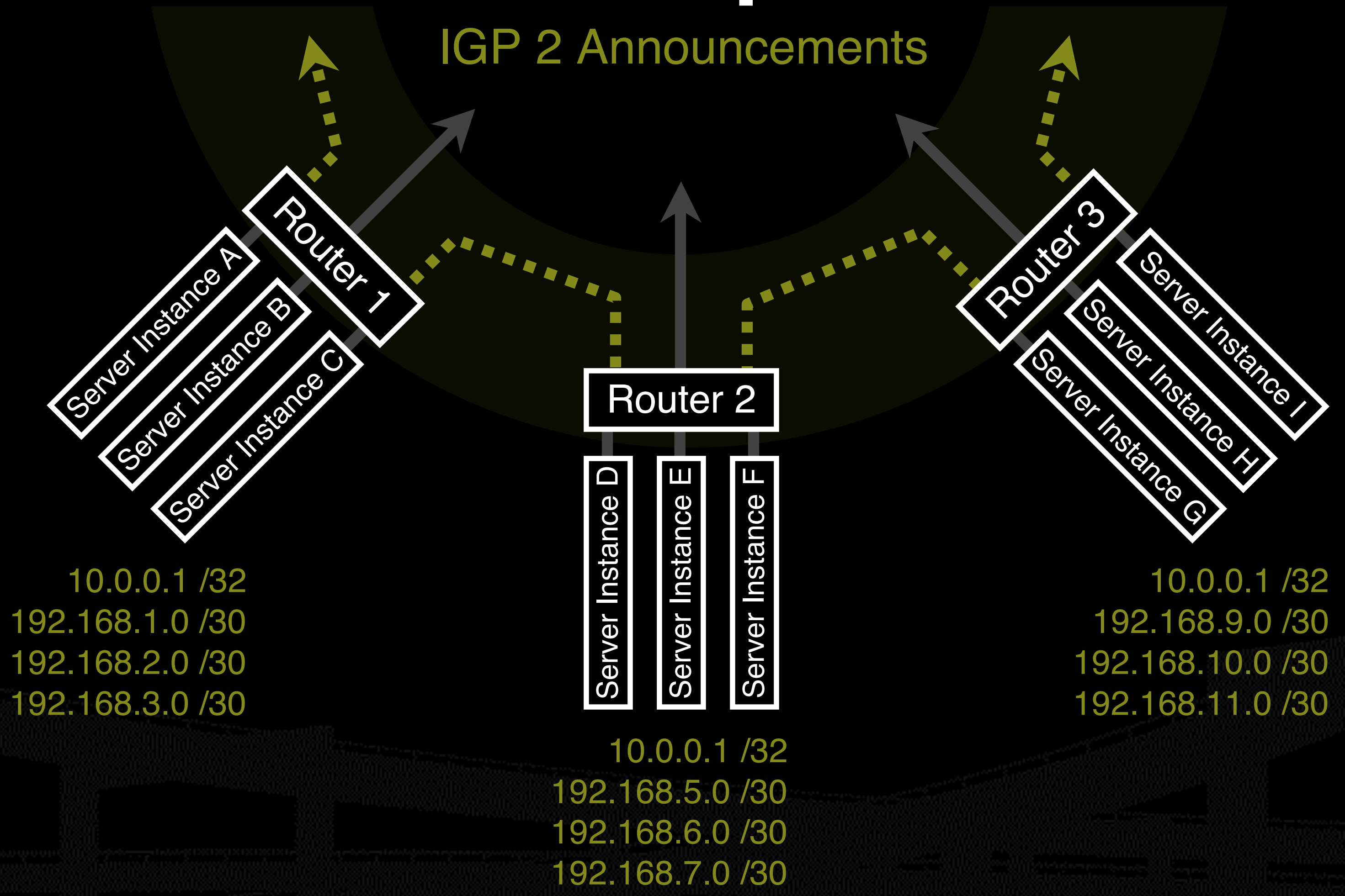
PCH Packet Clearing House

Example

IGP 1 Announcements



Example



IGP 2 Announcements
Router 1
Server Instance A
Server Instance B
Server Instance C
Router 2
Server Instance D
Server Instance E
Server Instance F
Router 3
Server Instance G
Server Instance H
Server Instance I
10.0.0.1 /32
192.168.1.0 /30
192.168.2.0 /30
192.168.3.0 /30
10.0.0.1 /32
192.168.5.0 /30
192.168.6.0 /30
192.168.7.0 /30
10.0.0.1 /32
192.168.9.0 /30
192.168.10.0 /30
192.168.11.0 /30

Performance-Tuning Anycast Networks

- Server deployment in anycast networks is always a tradeoff between absolute cost and efficiency.
- The network will perform best if servers are widely distributed, with higher density in and surrounding high demand areas.
- Lower initial cost sometimes leads implementers to compromise by deploying more servers in existing locations, which is less efficient.

Caveats and Failure Modes

- DNS resolution fail-over
- Long-lived connection-oriented flows
- Identifying which server is giving an end-user trouble

DNS Resolution Fail-Over

- In the event of poor performance from a server, DNS servers will fail over to the next server in a list.
- If both servers are in fact hosted in the same anycast cloud, the resolver will wind up talking to the same instance again.
- Best practices for anycast DNS server operations indicate a need for two separate overlapping clouds of anycast servers.

Packet Clearing House

Long-Lived Connection-Oriented Flows

- Long-lived flows, typically TCP file-transfers or interactive logins, may occasionally be more stable than the underlying Internet topology.
- If the underlying topology changes sufficiently during the life of an individual flow, packets could be redirected to a different server instance, which would not have proper TCP state, and would reset the connection.
- This is not a problem with web servers unless they’re maintaining stateful per-session information about end-users, rather than embedding it in URLs or cookies.
- Web servers HTTP redirect to their unique address whenever they need to enter a stateful mode.
- Limited operational data shows underlying instability to be on the order of one flow per ten thousand per hour of duration.

Identifying Problematic Server Instances

- Some protocols may not include an easy in-band method of identifying the server which persists beyond the duration of the connection.
- Traceroute always identifies the ***current*** server instance, but end-users may not even have traceroute.

Packet Clearing House

A Security Ramification

- Anycast server clouds have the useful property of sinking DOS attacks at the instance nearest to the source of the attack, leaving all other instances unaffected.
- This is still of some utility even when DOS sources are widely distributed.

PCH Anycast Service

- We provide anycast service for 14 ccTLDs and two gTLDs.
- At few selected locations, we provice connectivity to the anycast instance of the i.root-servers.net
- We have plans to anycast the SIP registry of the INOC DBA (www.pch.net/inoc-dba).

Packet Clearing House

PCH Anycast Network

- We look at a few things
  - Uniformity
  - Maximum Reachibility
  - No recurring cost
  - Easy way to manage with minimal staff and attention
  - Parallel operation of our route collection system

PCH
Packet Clearing House

Uniformity

Topology

- Redundant transit at every location
  - Four global Transit nodes
    - San Francisco and London are equivalent
    - Ashburn and Hongkong are equivalent
- Tunnel mesh
  - Dual Tunnel Hub in different continent for management
- Redundant private hubs

PCH
Packet Clearing House

Current (as of earlier this year) anycast Footprint

Packet Clearing House

New Sites in the Pipeline

How we do it.

- We have two routers with different ASN connected to the IX. We run multiple peering sessions with each peer.
- Transit is generally provided through a separate link.
- We have a /23 assigned for our own anycasting
- The Routers announce the /23 as well as management address at each location. Global Nodes announce all networks.

What on the host ?

- Use rsync to sync the anycast nodes
  - Using AXFR/IXFR is fine with DNS, but we also need to sync other stuff, so we use rsync every hour.
- Run quagga on our servers
  - Runs iBGP with the routers. If the host goes down, the iBGP sessions goes down, thus the router withdraws the network from the peers.
- For the i.root-servers.net and the .biz servers, we run BGP with their blades.

What does all of these do for networks?

- Distributing DNS servers or other static system across the network
  - Inject a /32 for your DNS servers into the IGP and then put multiple servers everywhere. The customers don't need to change DNS server IP each time they change locations
  - Sink DoS traffic to the closest node
  - Netflow collection to the closest node
  - Standardization of router and system configs.

Questions ?

Thank You

Gaurab Raj Upadhaya
Peering and Network Group
Packet Clearing House
gaurab@pch.net

With acknowledgements to Bill Woodcock.
The anycast tutorial can be found at
http:// www.pch.net / resources / tutorials / anycast