سینداد
فناوران شبکه

IPv4 Anycast Routing

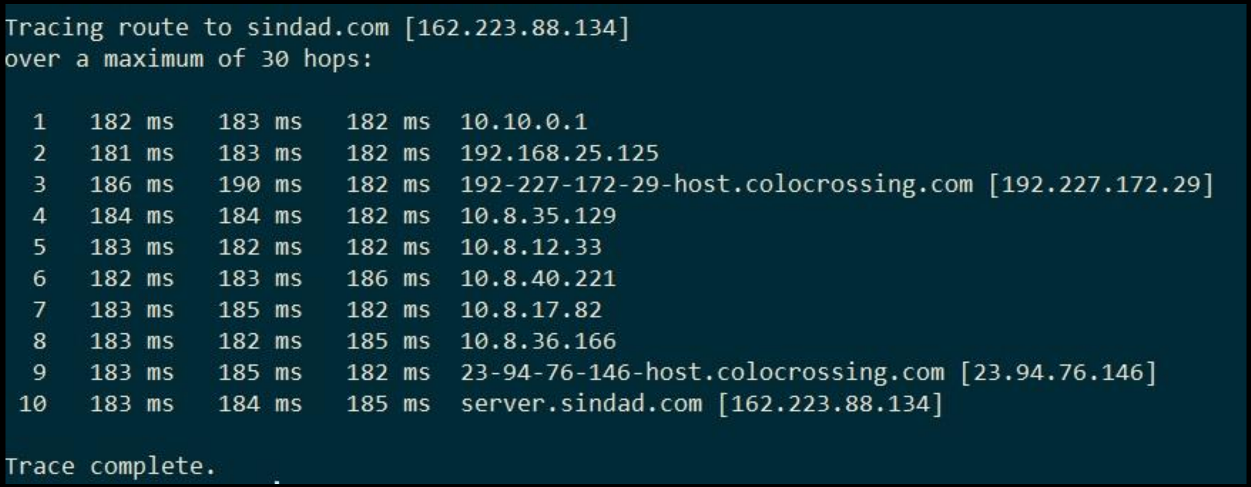
Reza Jamali
Sindad
2018 July

In Real World How WebSites act

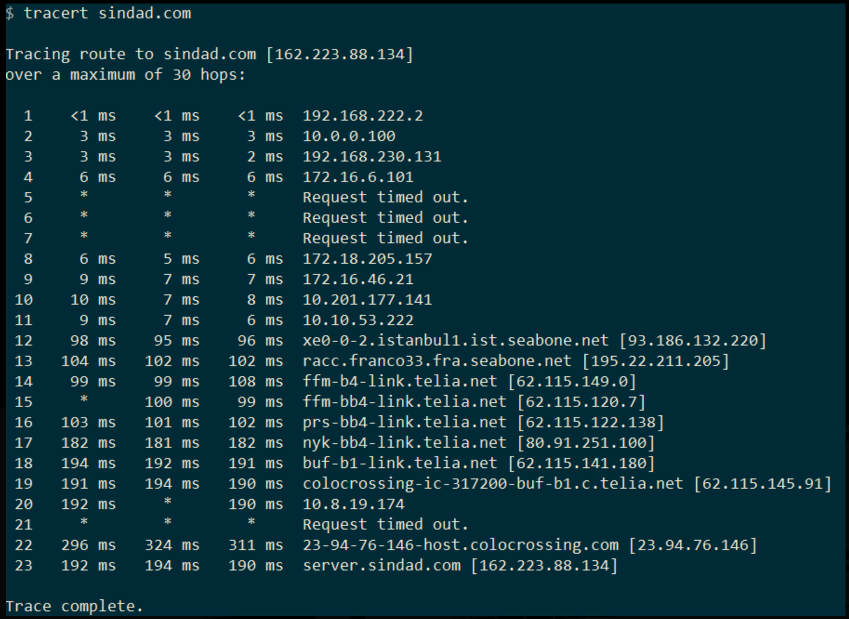
- If you try to access sindad.com from US the request will be routed to the Colocrossing data center. If you try to access sindad.com from Asia, again the request will be routed to the same Colocrossing Data Center.

Two major problems associated with this architecture.

- If Colocrossing Data Center goes down, then my site won’t be accessible.
- Second problem is if a user from Asia, access my site, that user has to unnecessarily suffer a latency of few hundred milliseconds. The problem is with everyone. Say a person accessing my site from US, he will still suffer a little latency as his packets needs to travel all the way to Colocrossing.



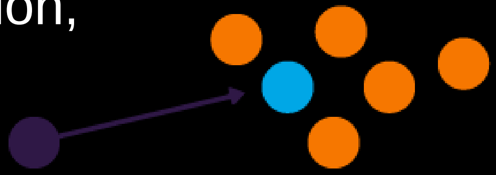
```
Tracing route to sindad.com [162.223.88.134]
over a maximum of 30 hops:

  1   182 ms   183 ms   182 ms  10.10.0.1
  2   181 ms   183 ms   182 ms  192.168.25.125
  3   186 ms   190 ms   182 ms  192-227-172-29-host.colocrossing.com [192.227.172.29]
  4   184 ms   184 ms   182 ms  10.8.35.129
  5   183 ms   182 ms   182 ms  10.8.12.33
  6   182 ms   183 ms   186 ms  10.8.40.221
  7   183 ms   185 ms   182 ms  10.8.17.82
  8   183 ms   182 ms   185 ms  10.8.36.166
  9   183 ms   185 ms   182 ms  23-94-76-146-host.colocrossing.com [23.94.76.146]
 10   183 ms   184 ms   185 ms  server.sindad.com [162.223.88.134]

Trace complete.
```



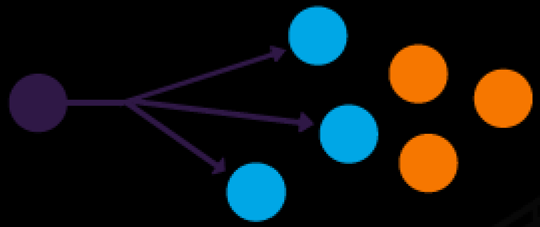
```
$ tracert sindad.com

Tracing route to sindad.com [162.223.88.134]
over a maximum of 30 hops:

  1    <1 ms    <1 ms    <1 ms  192.168.222.2
  2     3 ms     3 ms     3 ms  10.0.0.100
  3     3 ms     3 ms     2 ms  192.168.230.131
  4     6 ms     6 ms     6 ms  172.16.6.101
  5     *        *        *     Request timed out.
  6     *        *        *     Request timed out.
  7     *        *        *     Request timed out.
  8     6 ms     5 ms     6 ms  172.18.205.157
  9     9 ms     7 ms     7 ms  172.16.46.21
 10    10 ms     7 ms     8 ms  10.201.177.141
 11     9 ms     7 ms     6 ms  10.10.53.222
 12    98 ms    95 ms    96 ms  xe0-0-2.istanbul1.ist.seabone.net [93.186.132.220]
 13   104 ms   102 ms   102 ms  racc.franco33.fra.seabone.net [195.22.211.205]
 14    99 ms    99 ms   108 ms  ffm-b4-link.telia.net [62.115.149.0]
 15     *      100 ms    99 ms  ffm-bb4-link.telia.net [62.115.120.7]
 16   103 ms   101 ms   102 ms  prs-bb4-link.telia.net [62.115.122.138]
 17   182 ms   181 ms   182 ms  nyk-bb4-link.telia.net [80.91.251.100]
 18   194 ms   192 ms   191 ms  buf-b1-link.telia.net [62.115.141.180]
 19   191 ms   194 ms   190 ms  colocrossing-ic-317200-buf-b1.c.telia.net [62.115.145.91]
 20   192 ms     *      190 ms  10.8.19.174
 21     *        *        *     Request timed out.
 22   296 ms   324 ms   311 ms  23-94-76-146-host.colocrossing.com [23.94.76.146]
 23   192 ms   194 ms   190 ms  server.sindad.com [162.223.88.134]

Trace complete.
```

Addressing Method

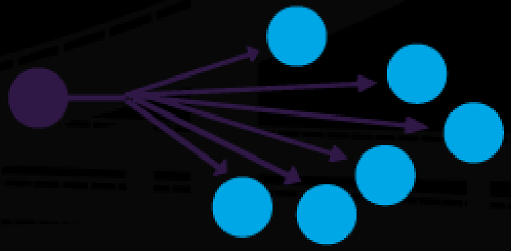
- Unicast addressing: uses a *one-to-one* association,



- Multicast addressing: uses a one-to-unique many association



- Broadcast addressing: uses a one-to-all association



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.

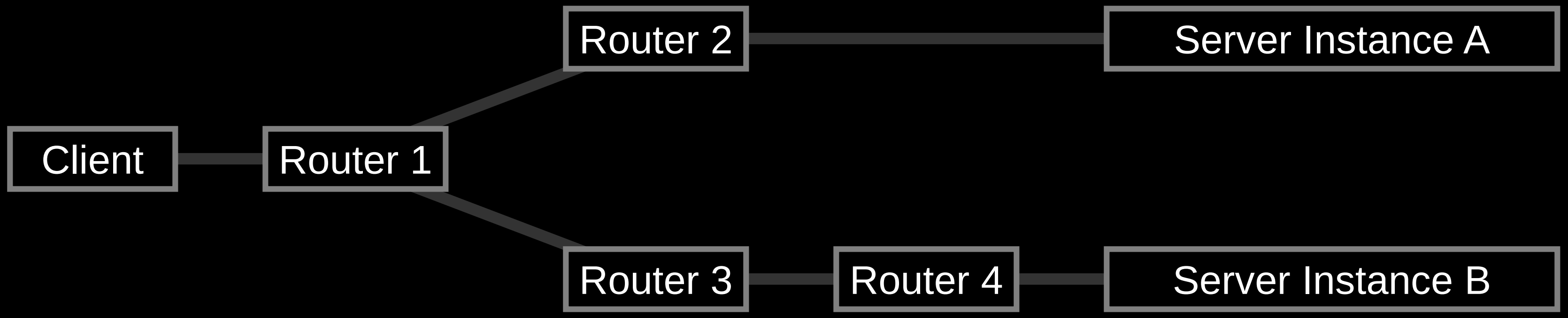
What *is* Anycast?

- Just a configuration methodology.
- Anycast described in fallowing RFCs 4786 -7049 - 1546.
- 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.

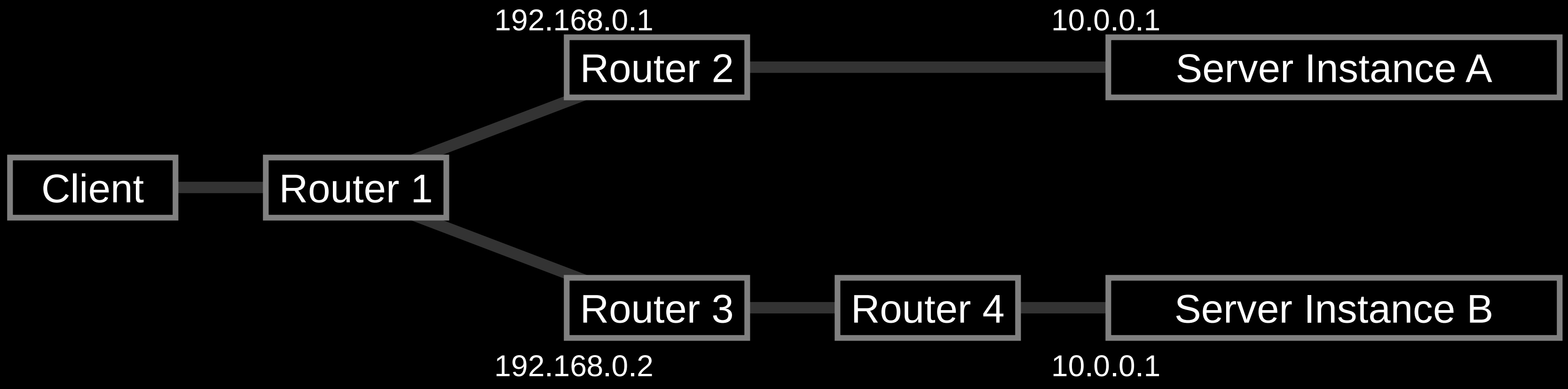
How Does Anycast Work?

- 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.

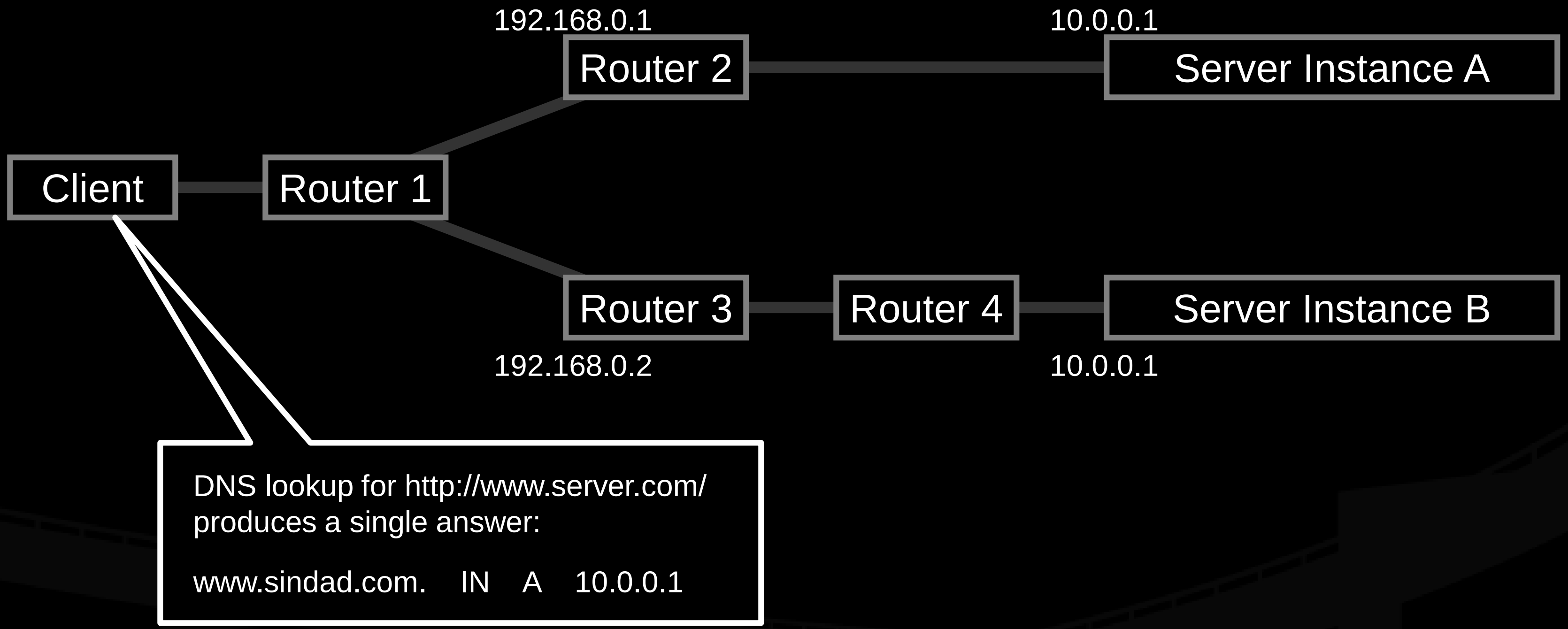
Example


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

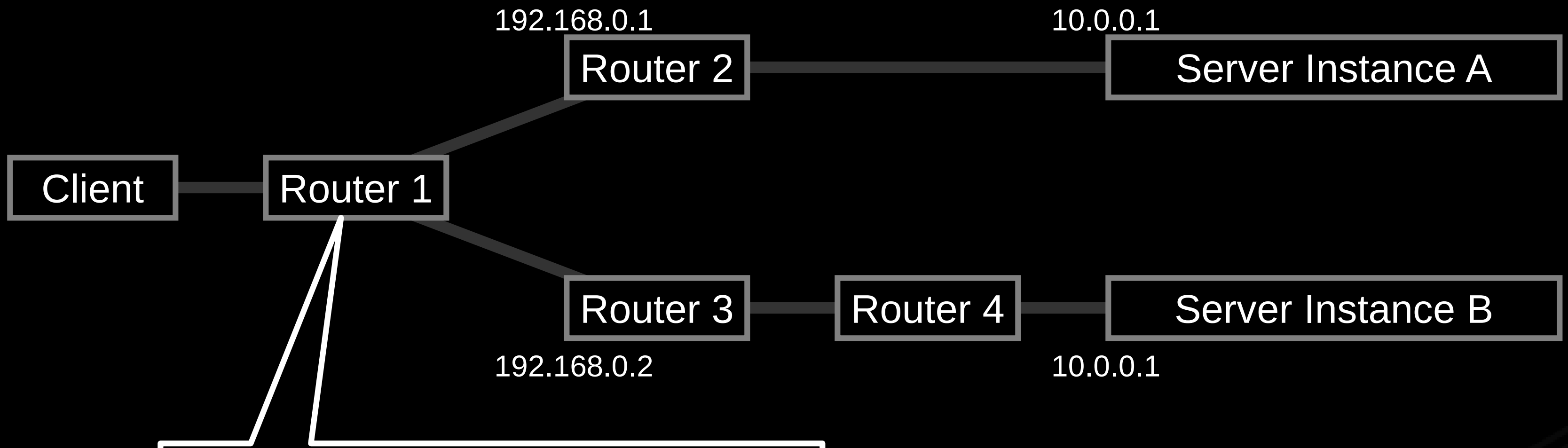
Example



DNS lookup for http://www.server.com/
produces a single answer:

www.sindad.com.   IN   A   10.0.0.1

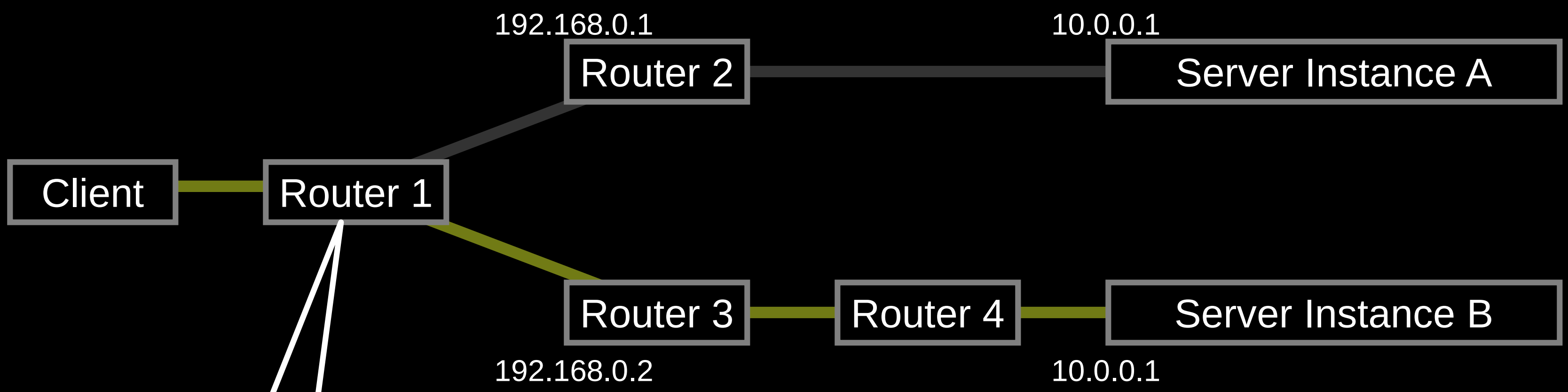
Example



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

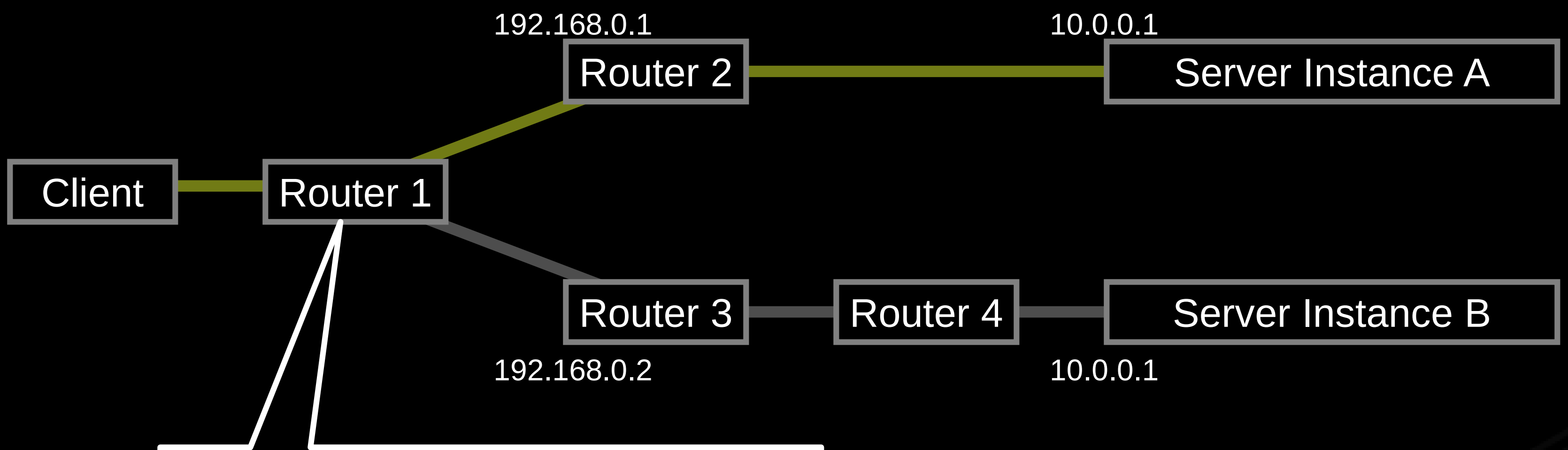
Example



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

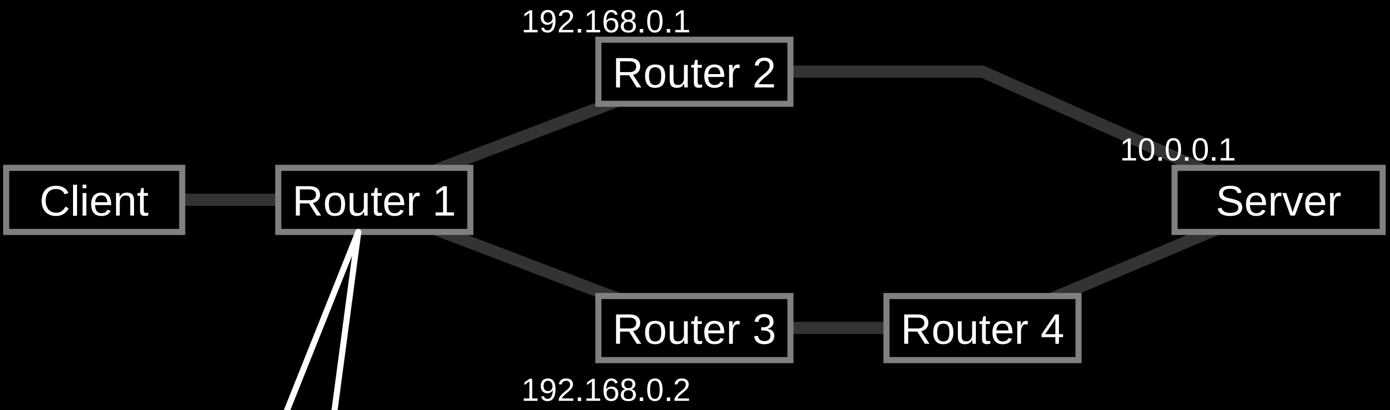


Routing Table from Router 1:

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:



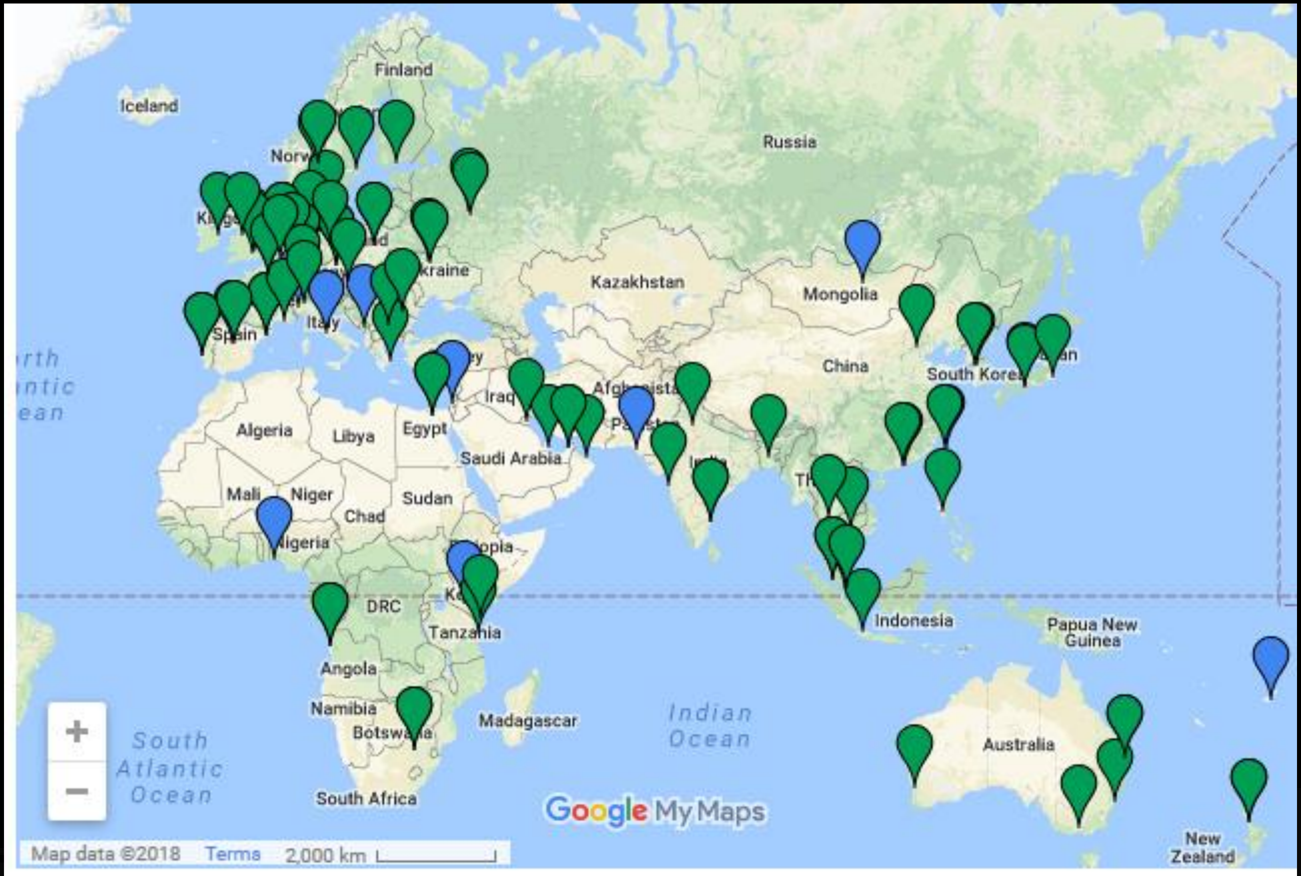
Routing Table from Router 1:

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.

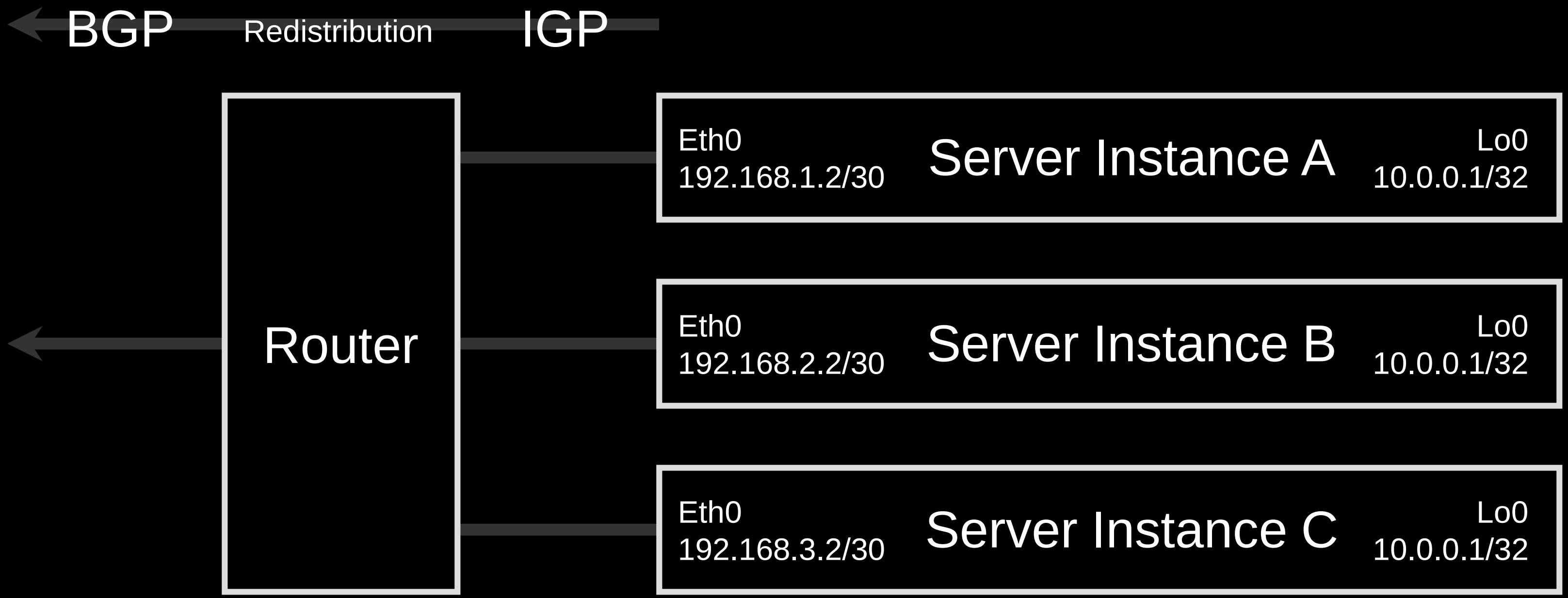


f.root-servers.net [192.5.5.241]

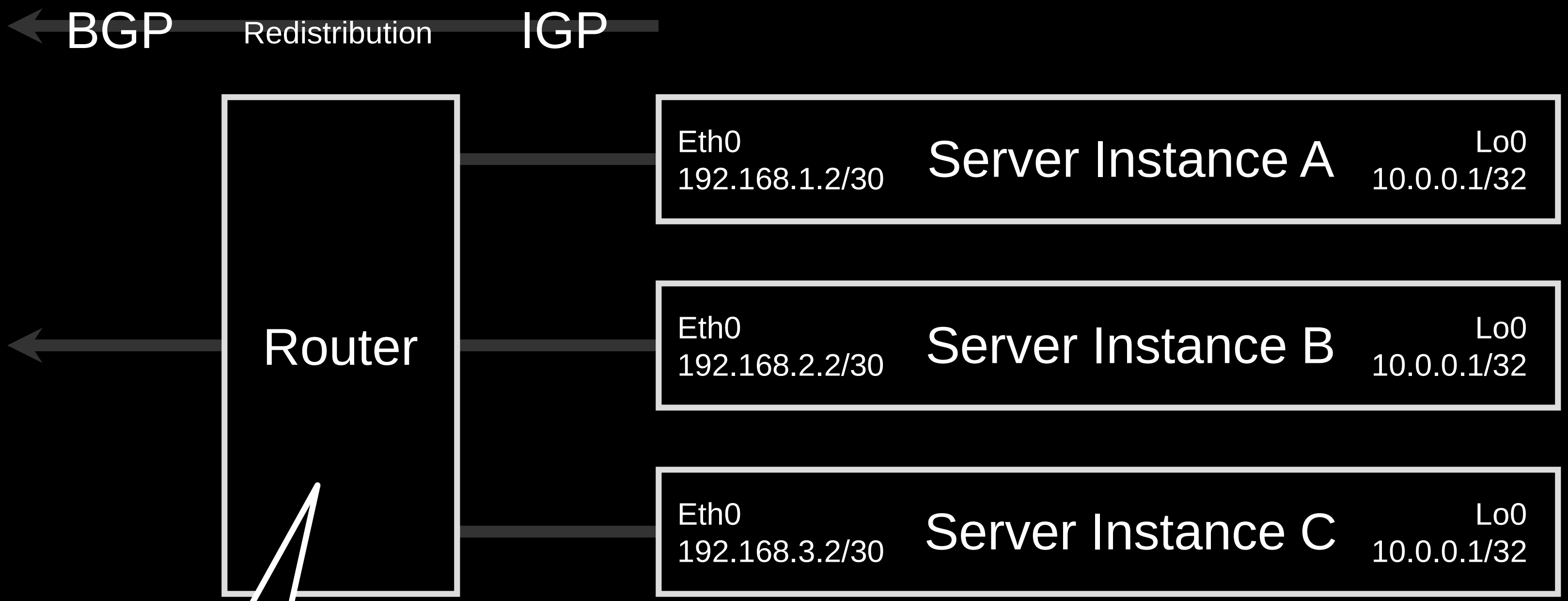
Building an Anycast Server Cluster

- 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



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

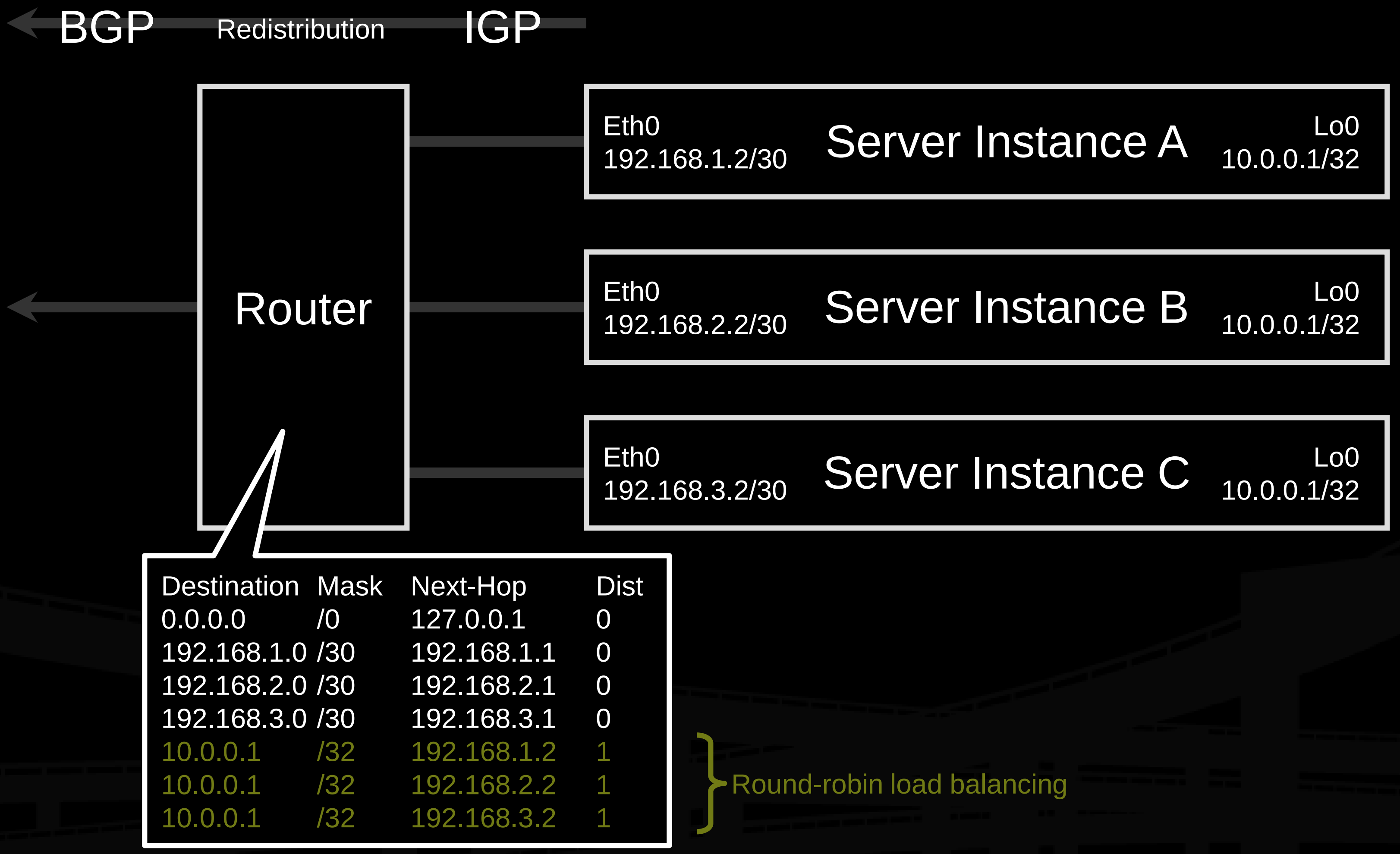
Eth0
192.168.3.2/30

Server Instance C

Lo0
10.0.0.1/32

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

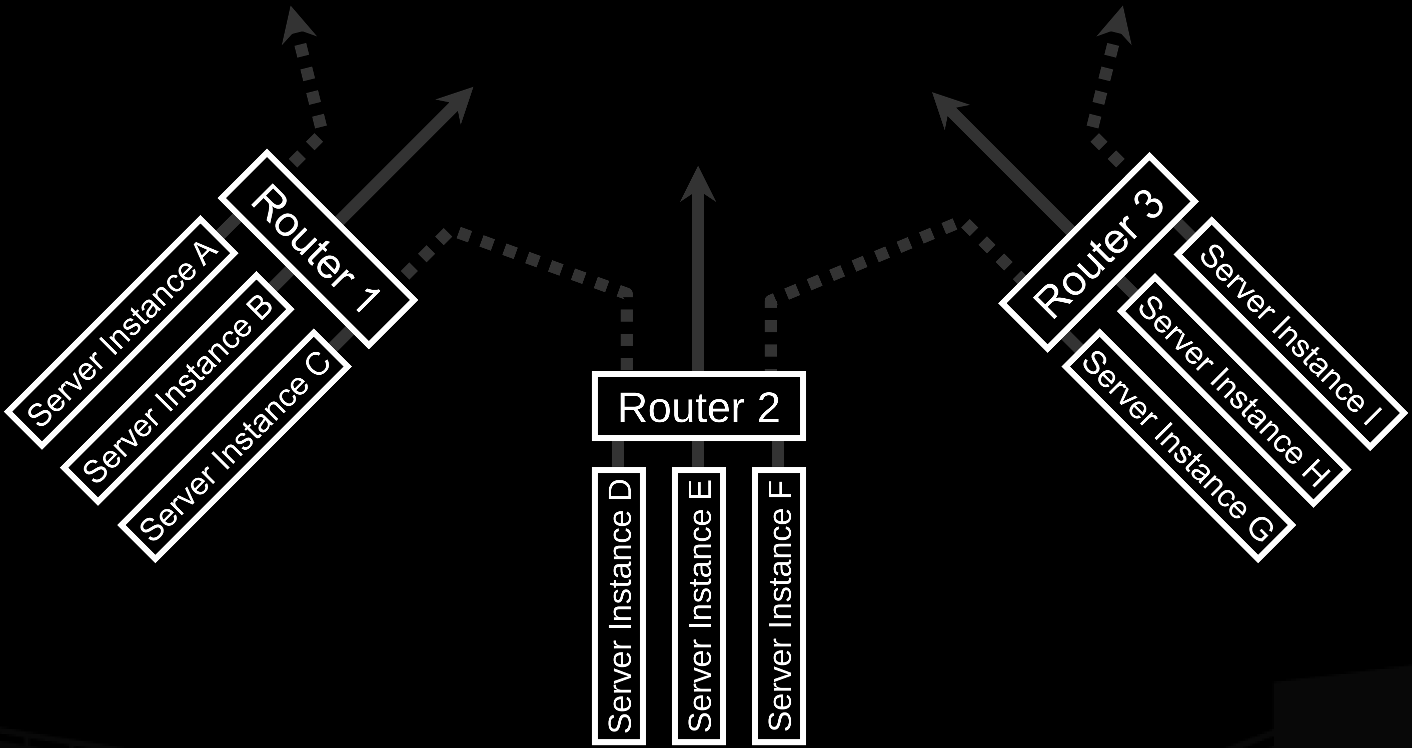


Destination	Mask	Next-Hop	Dist
0.0.0.0	/0	127.0.0.1	0
192.168.1.0	/30	192.168.1.1	0
192.168.2.0	/30	192.168.2.1	0
192.168.3.0	/30	192.168.3.1	0
10.0.0.1	/32	192.168.1.2	1
10.0.0.1	/32	192.168.2.2	1
10.0.0.1	/32	192.168.3.2	1

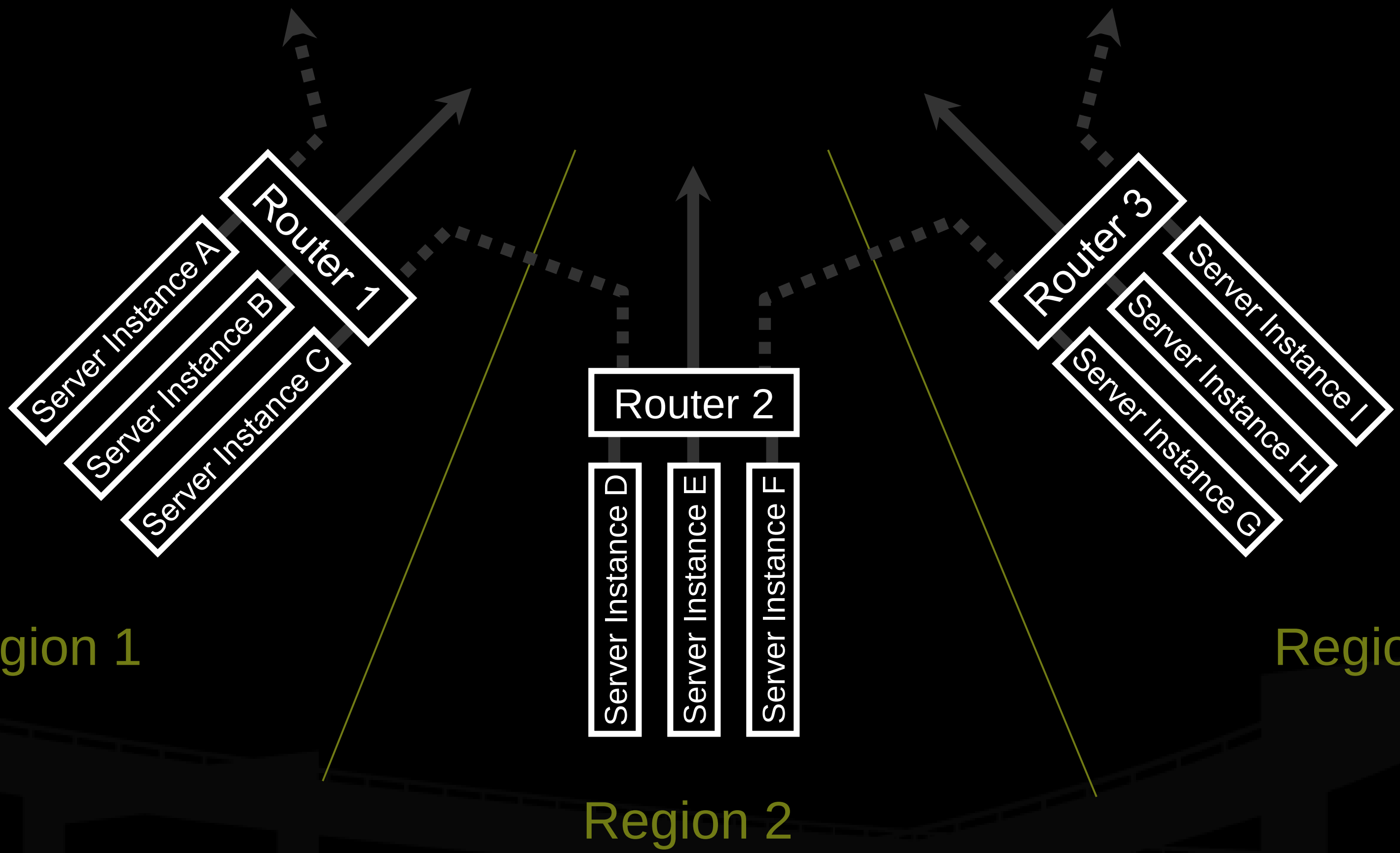
Building a Global Network of Clusters

- 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

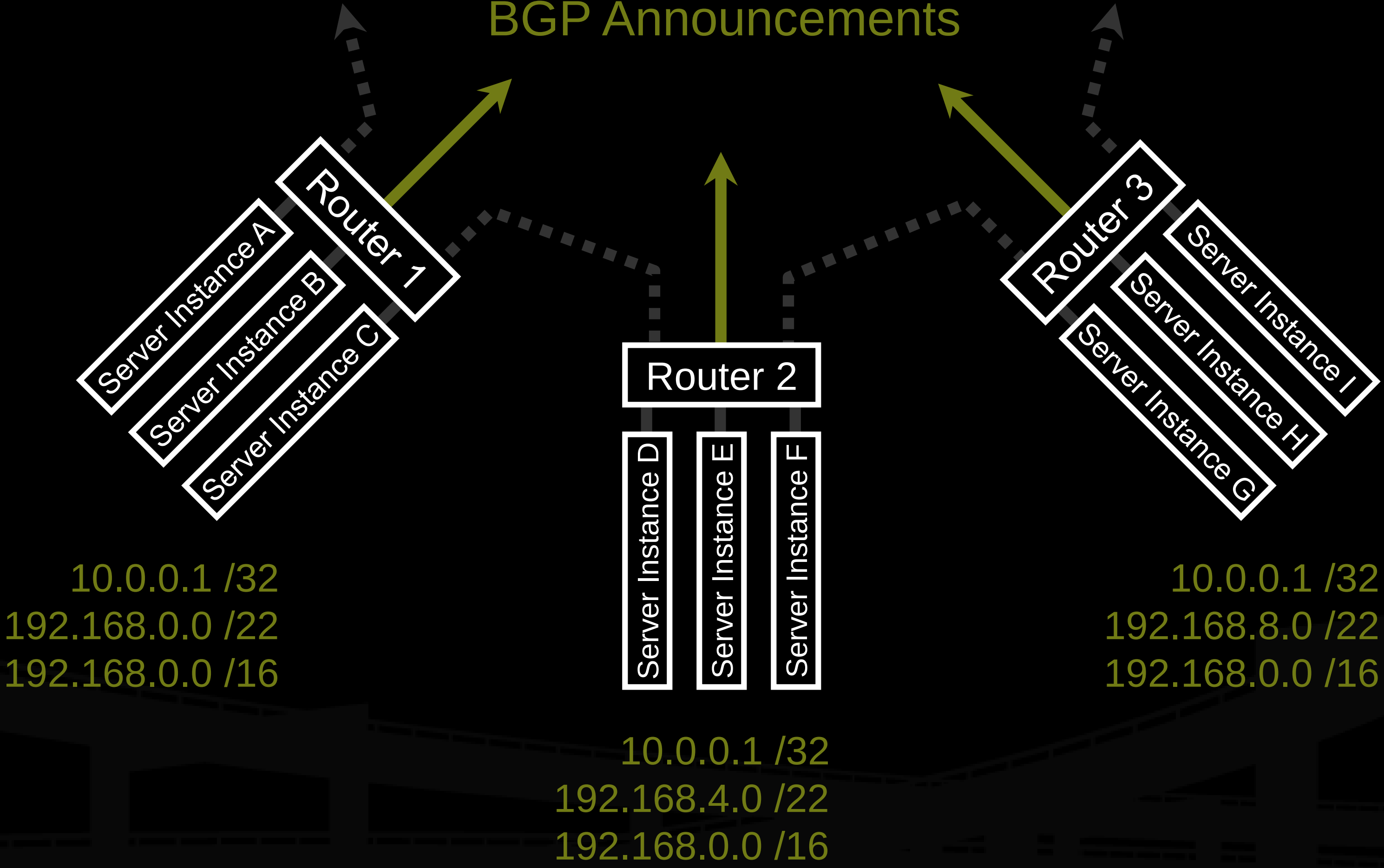

Router 1
Server Instance A
Server Instance B
Server Instance C
Router 2
Server Instance D
Server Instance E
Server Instance F
Router 3
Server Instance G
Server Instance H
Server Instance I

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 I
Server Instance H
Server Instance G
Region 1
Region 2
Region 3

Example

BGP Announcements



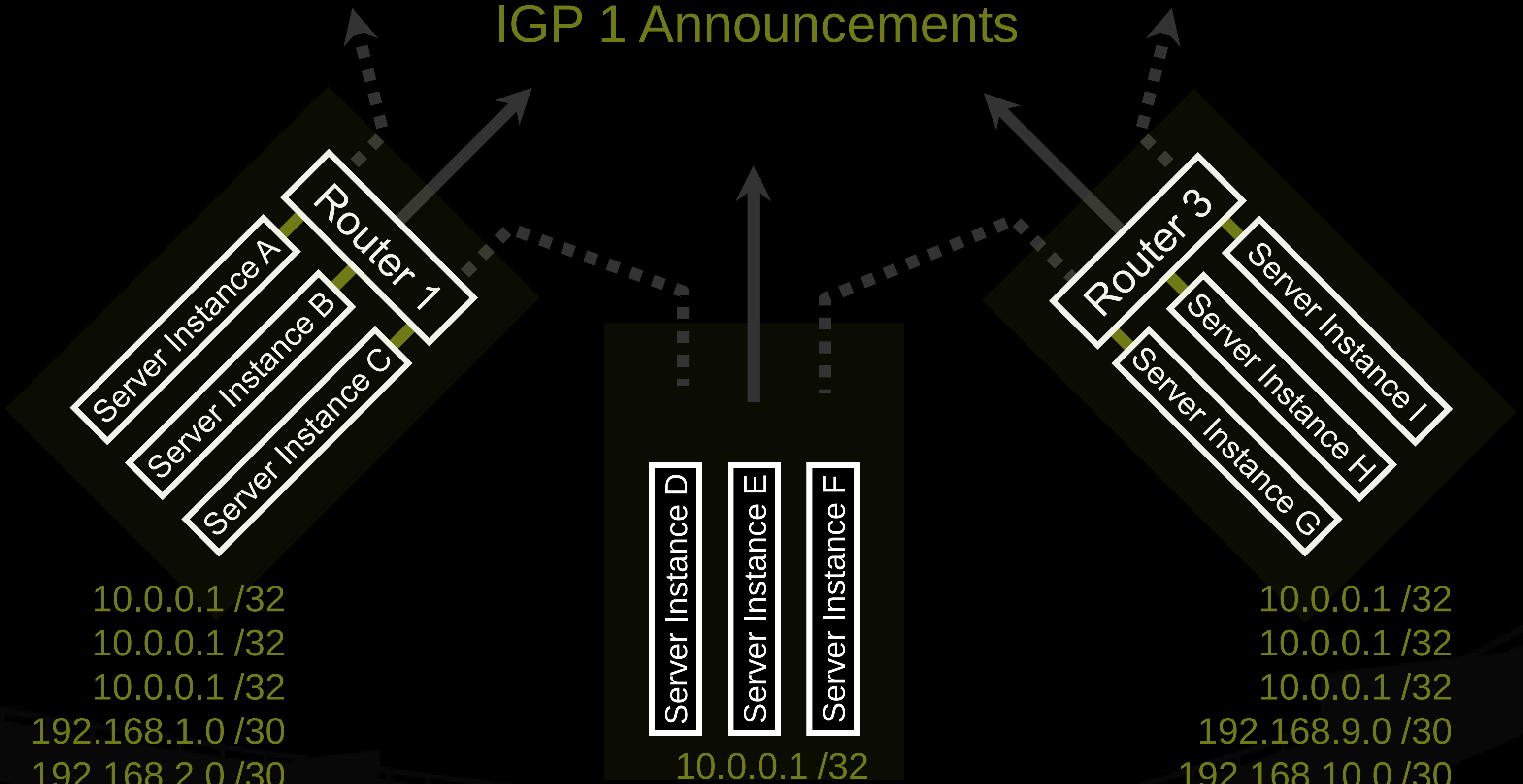
10.0.0.1 /32
192.168.0.0 /22
192.168.0.0 /16

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

Example

IGP 1 Announcements



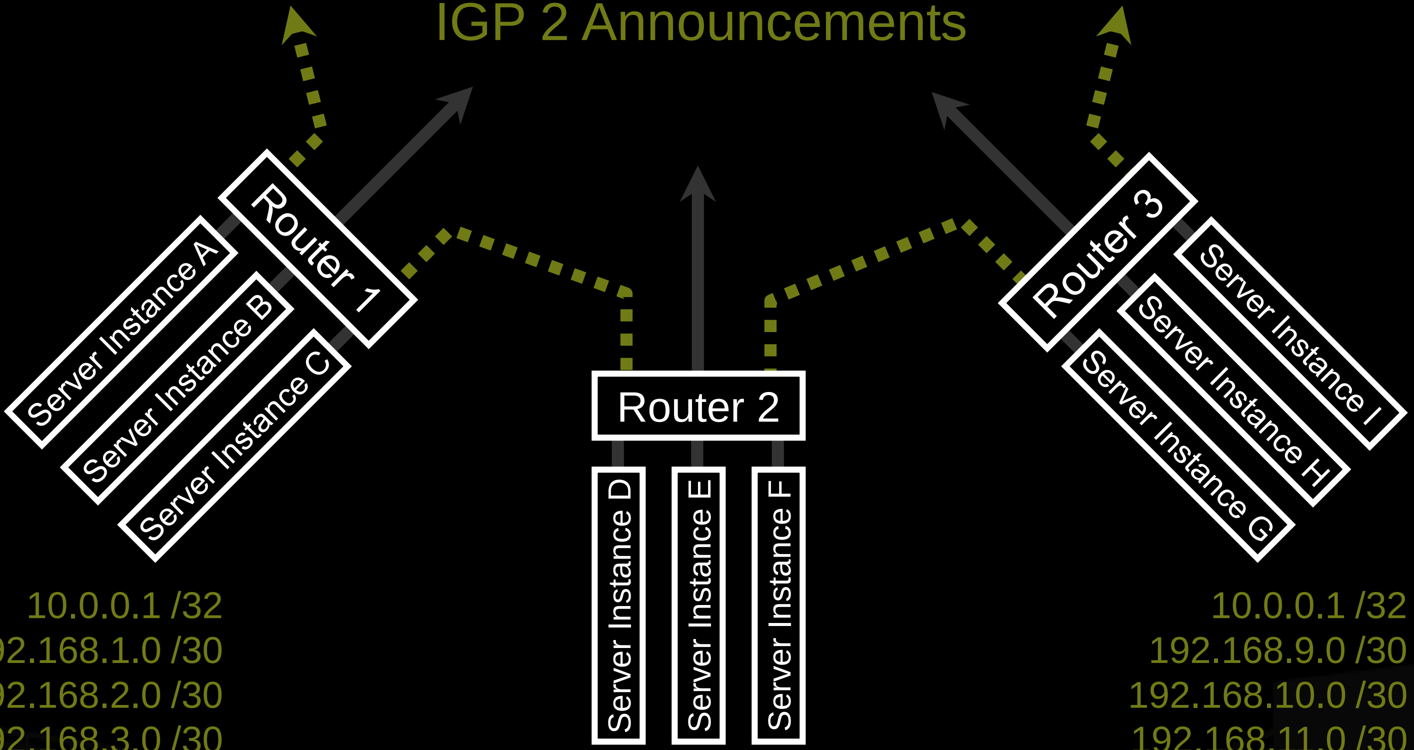
10.0.0.1 /32
10.0.0.1 /32
10.0.0.1 /32
192.168.1.0 /30
192.168.2.0 /30
192.168.3.0 /30

10.0.0.1 /32
10.0.0.1 /32
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
10.0.0.1 /32
10.0.0.1 /32
192.168.9.0 /30
192.168.10.0 /30
192.168.11.0 /30

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

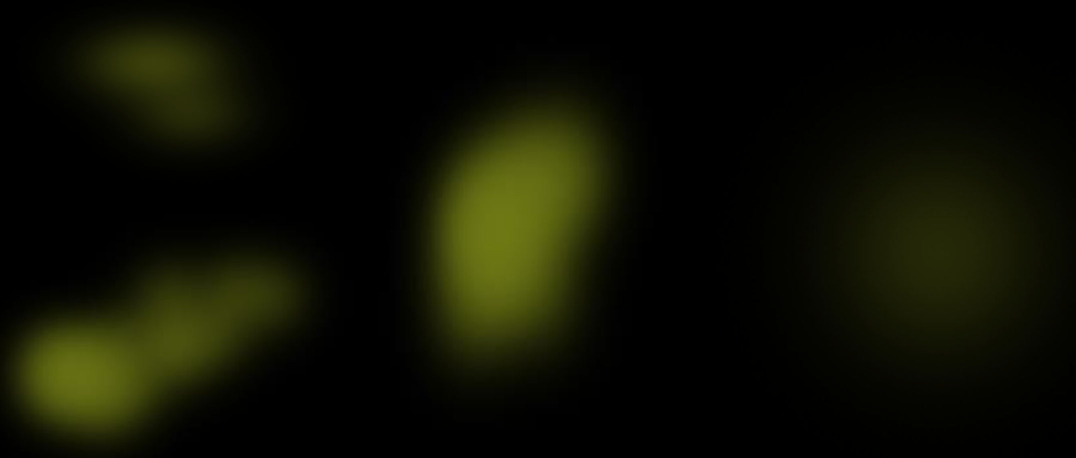
10.0.0.1 /32
192.168.9.0 /30
192.168.10.0 /30
192.168.11.0 /30

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.

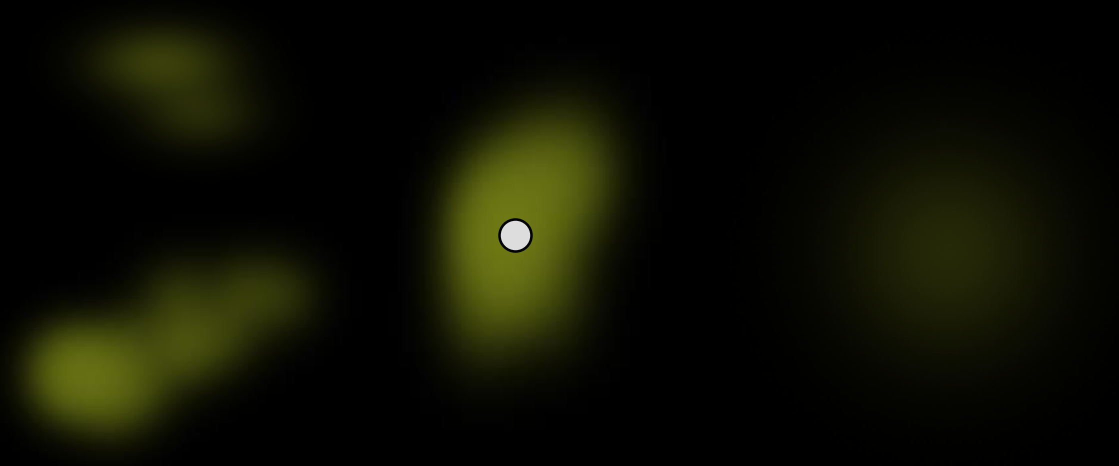
Example

Geographic plot of user population density



Example

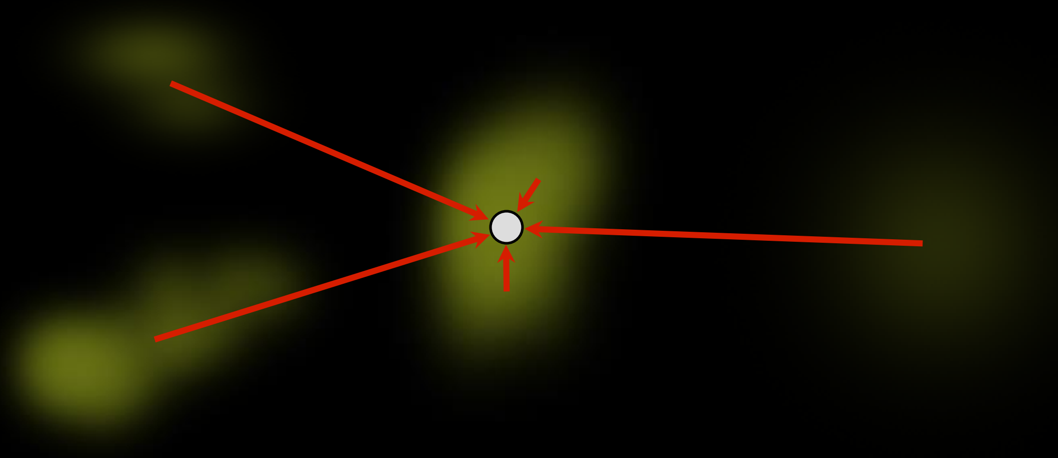
Geographic plot of user population density



Server deployment

Example

Geographic plot of user population density

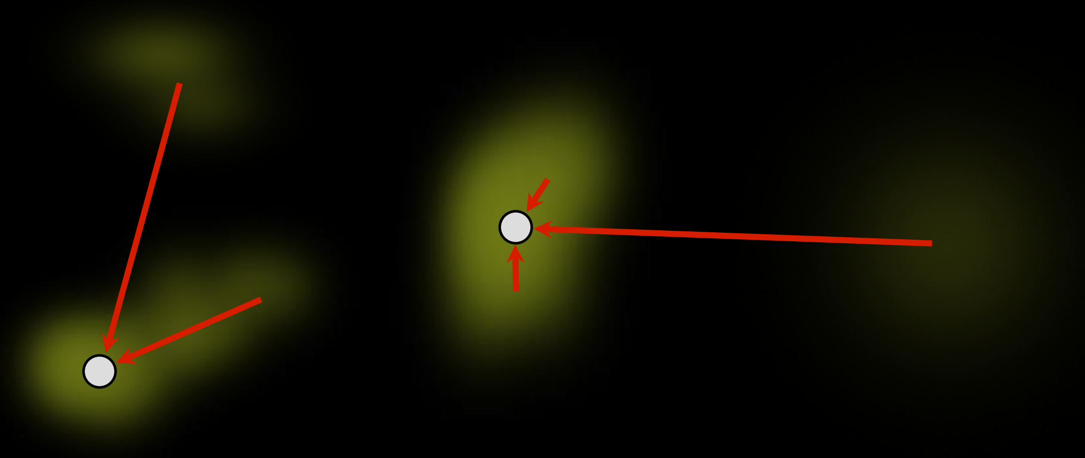


Server deployment

Traffic Flow

Example

Geographic plot of user population density

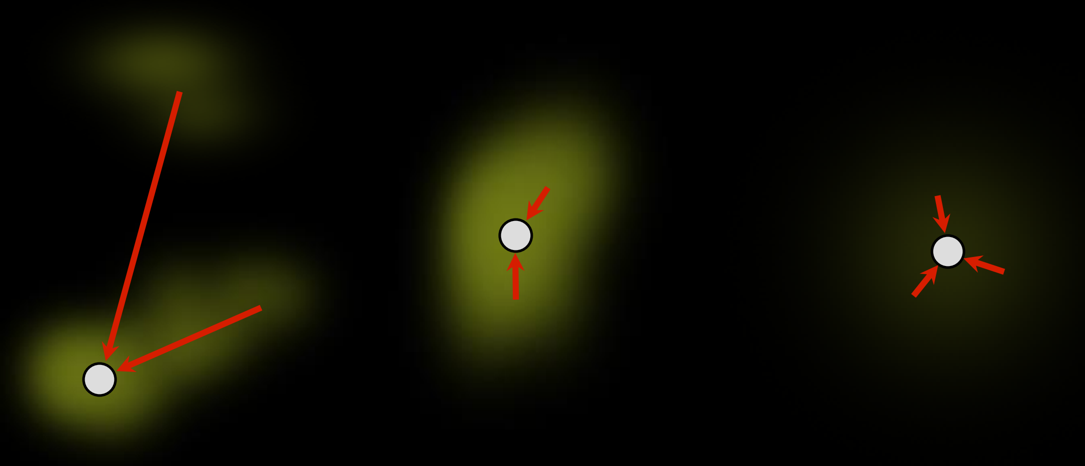


Server deployment

Traffic Flow

Example

Geographic plot of user population density

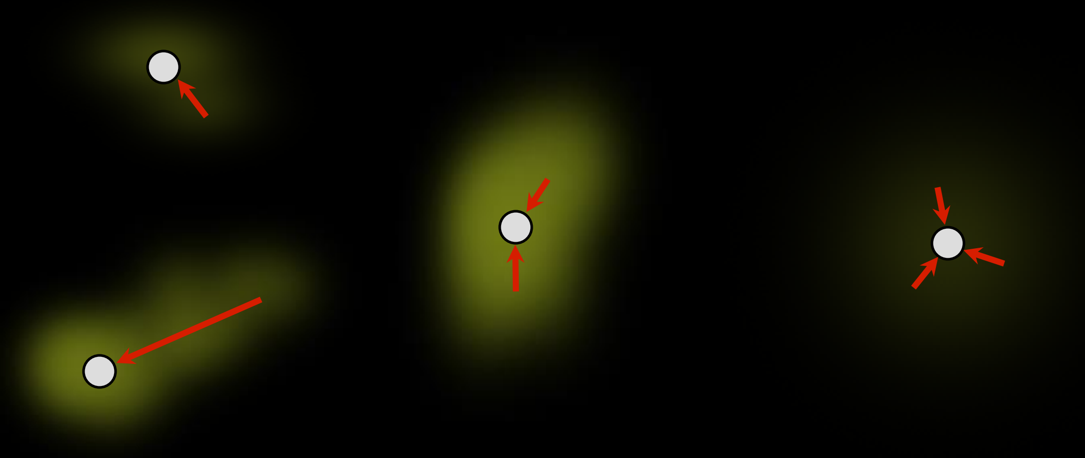


Server deployment

Traffic Flow

Example

Geographic plot of user population density

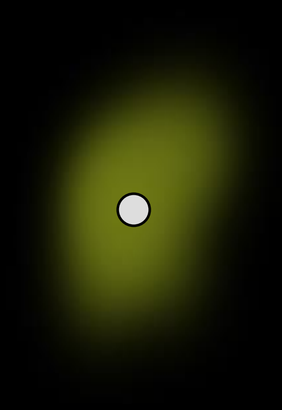


Server deployment

Traffic Flow

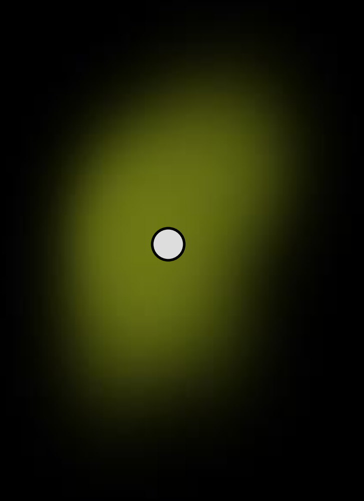
Example

Drawing traffic growth away from a hot-spot



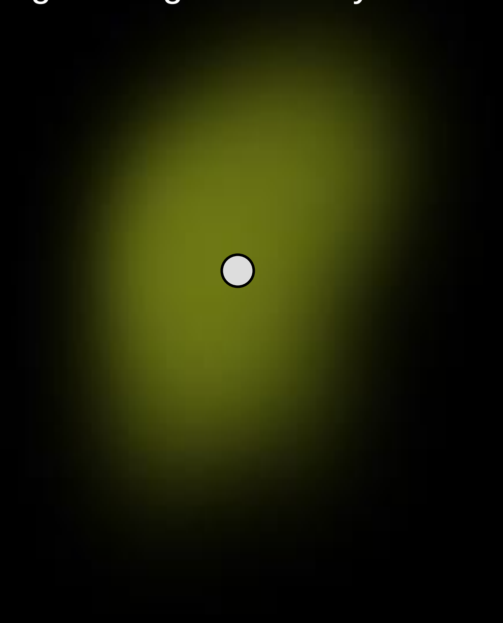
Example

Drawing traffic growth away from a hot-spot



Example

Drawing traffic growth away from a hot-spot

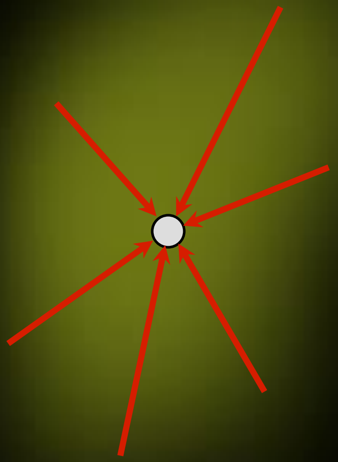


Example

Drawing traffic growth away from a hot-spot

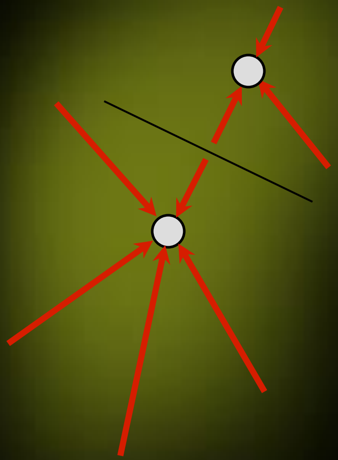
Example

Drawing traffic growth away from a hot-spot



Example

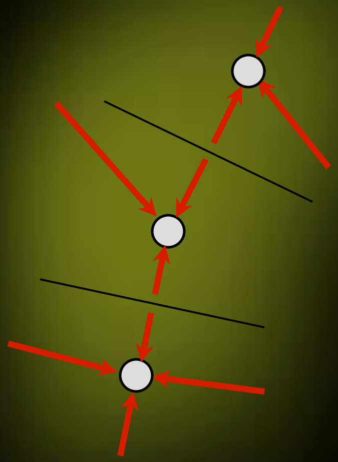
Drawing traffic growth away from a hot-spot



Topological watershed

Example

Drawing traffic growth away from a hot-spot



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.

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.

Identifying Problematic Server Instances

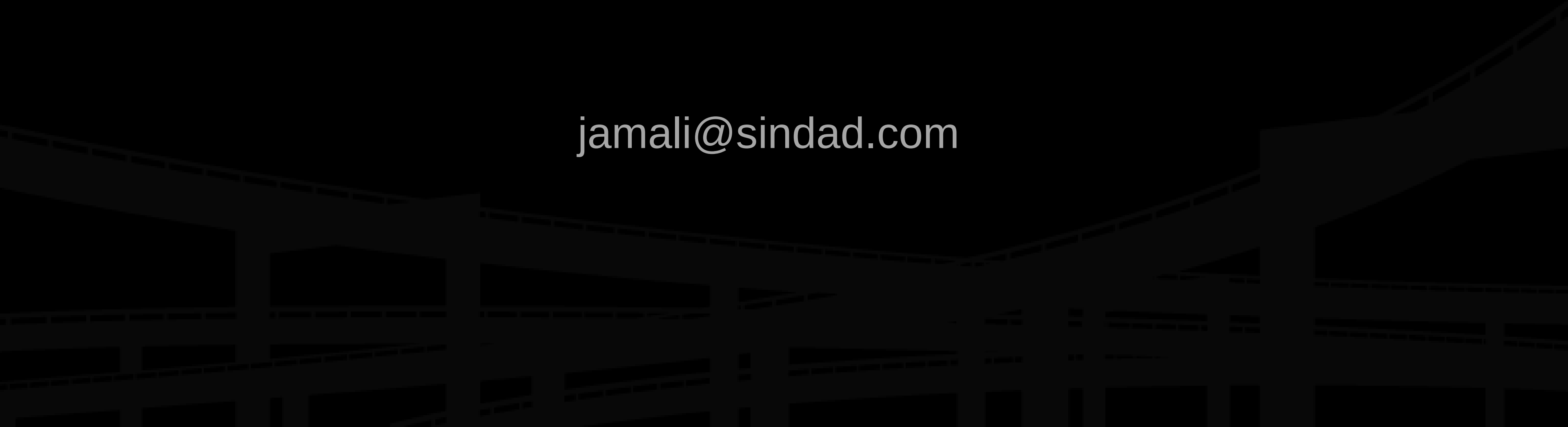
- Some protocols may not include an easy in-band method of identifying the server which persists beyond the duration of the connection.
- Traceroute always identifies the ***current*** server instance, but end-users may not even have traceroute.

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.

سینداد
فناوران شبکه

Thank You.



jamali@sindad.com