Deploying IP Anycast

Kevin Miller

Carnegie Mellon Network Group

kcm@cmu.edu

NANOG 29 – October 2003

Carnegie Mellon

Overview

- Why anycast?
  - Server load balancing
  - Service reliability
  - Client transparency
  - Locality / latency improvements
  - Distributed response to DoS

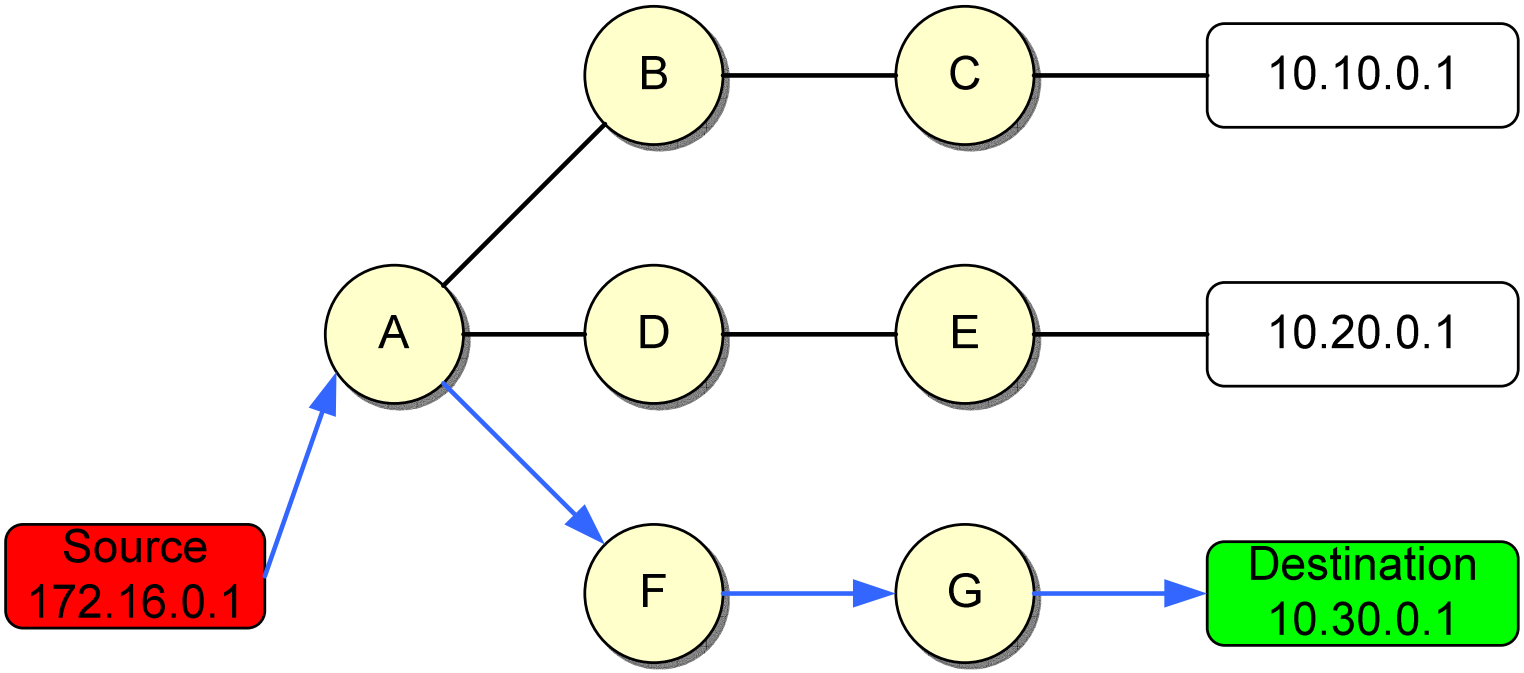
- Assume experience with unicast routing
- http://www.net.cmu.edu/pres/anycast

Agenda

- **What is Anycast?**
- Deploying IPv4 anycast services
- Anycast usage case studies
- Advanced Topics

Not Unicast

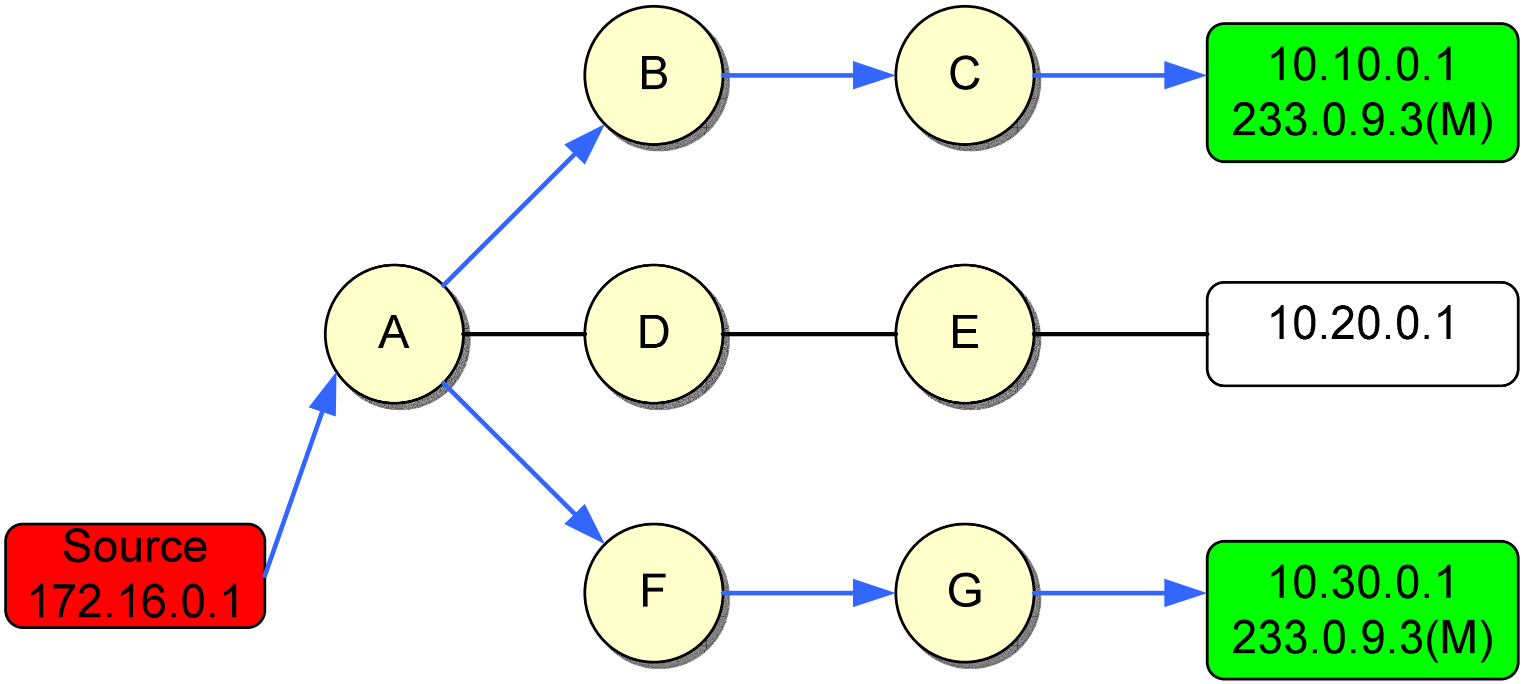
- Unicast: Single host receives all traffic



Carnegie Mellon

Not Multicast

- Multicast: Many hosts receive (all) traffic to multicast group



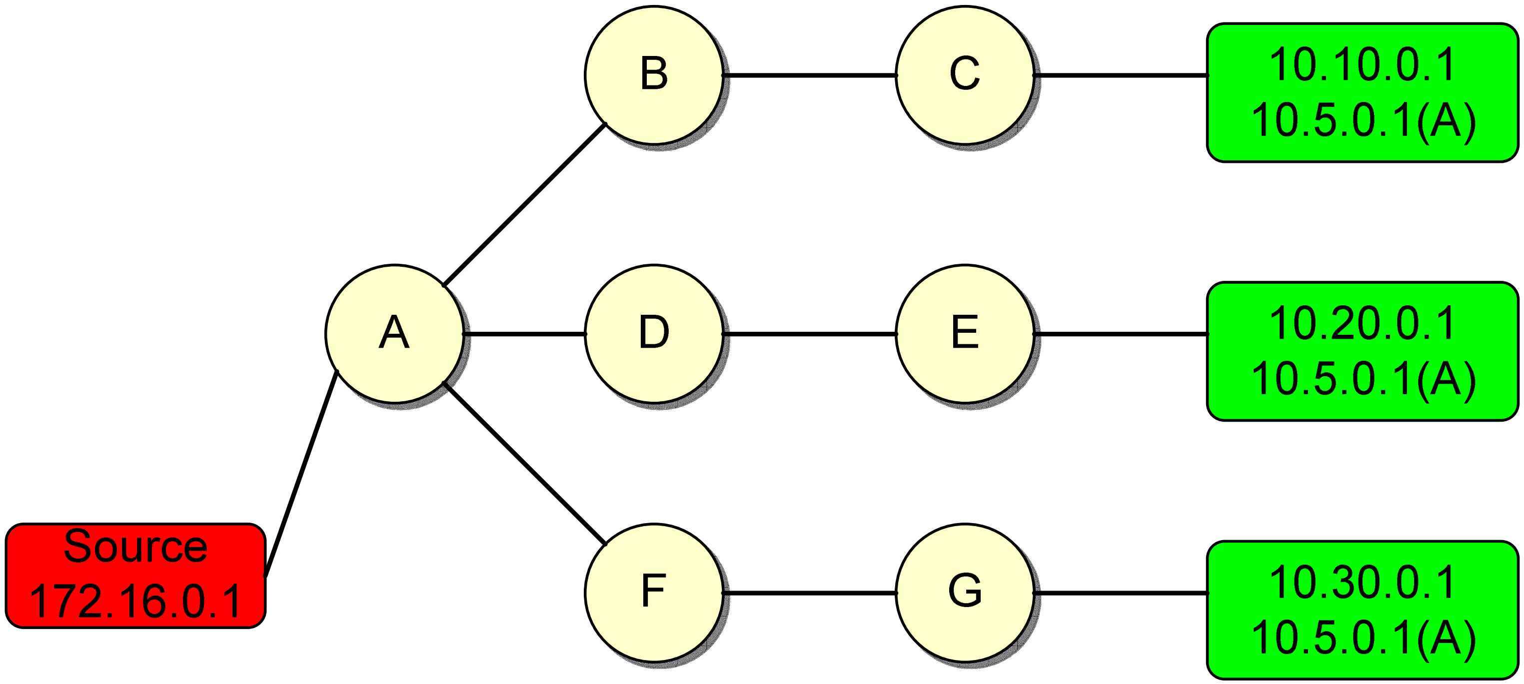
Carnegie Mellon

Anycast

- Multiple nodes configured to accept traffic on single IP address
- Usually, **one node** receives each packet
  - Packet could be dropped like any other
  - Preferably only one node receives packet, but no absolute guarantee
- The node that receives a specific packet is determined by routing.

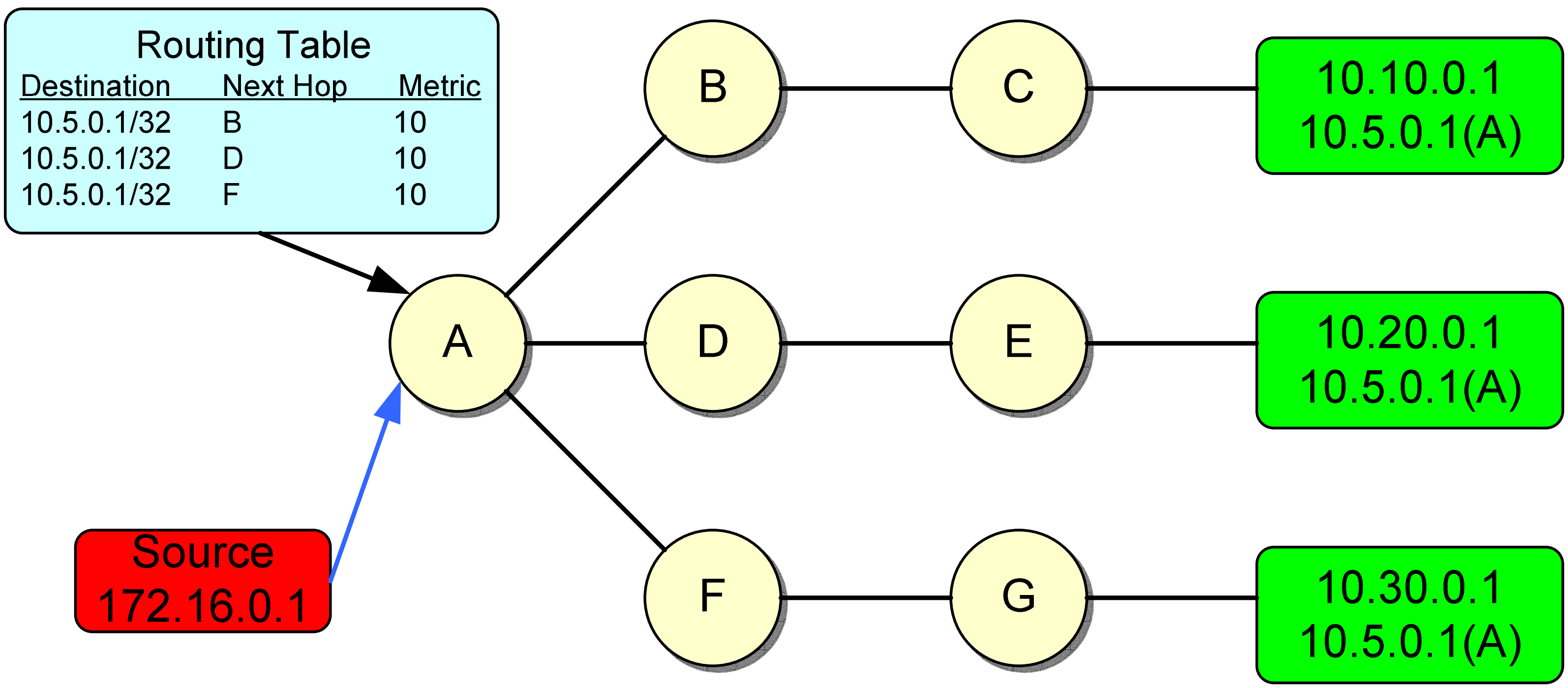
Anycast

- Three nodes configured with anycast address (10.5.0.1)



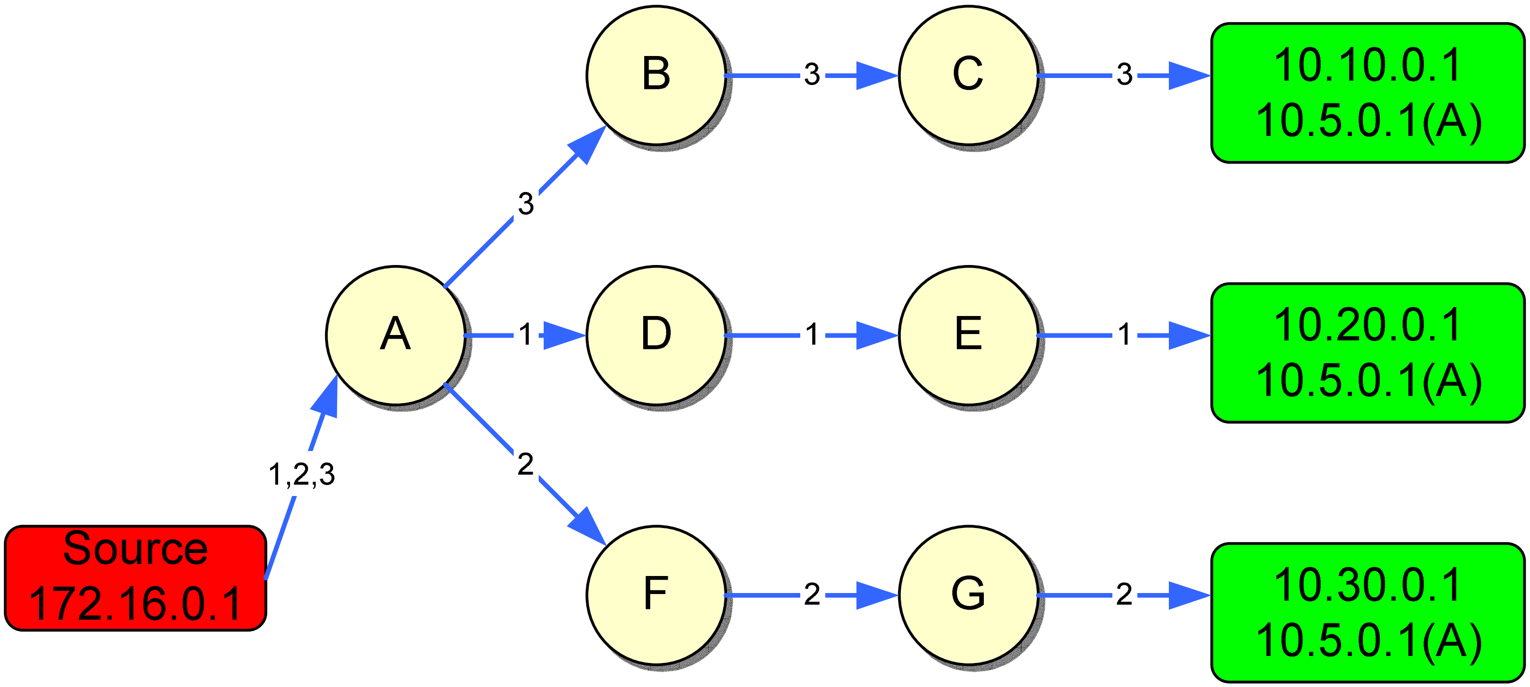
Anycast

- Potentially equal-cost multi-path



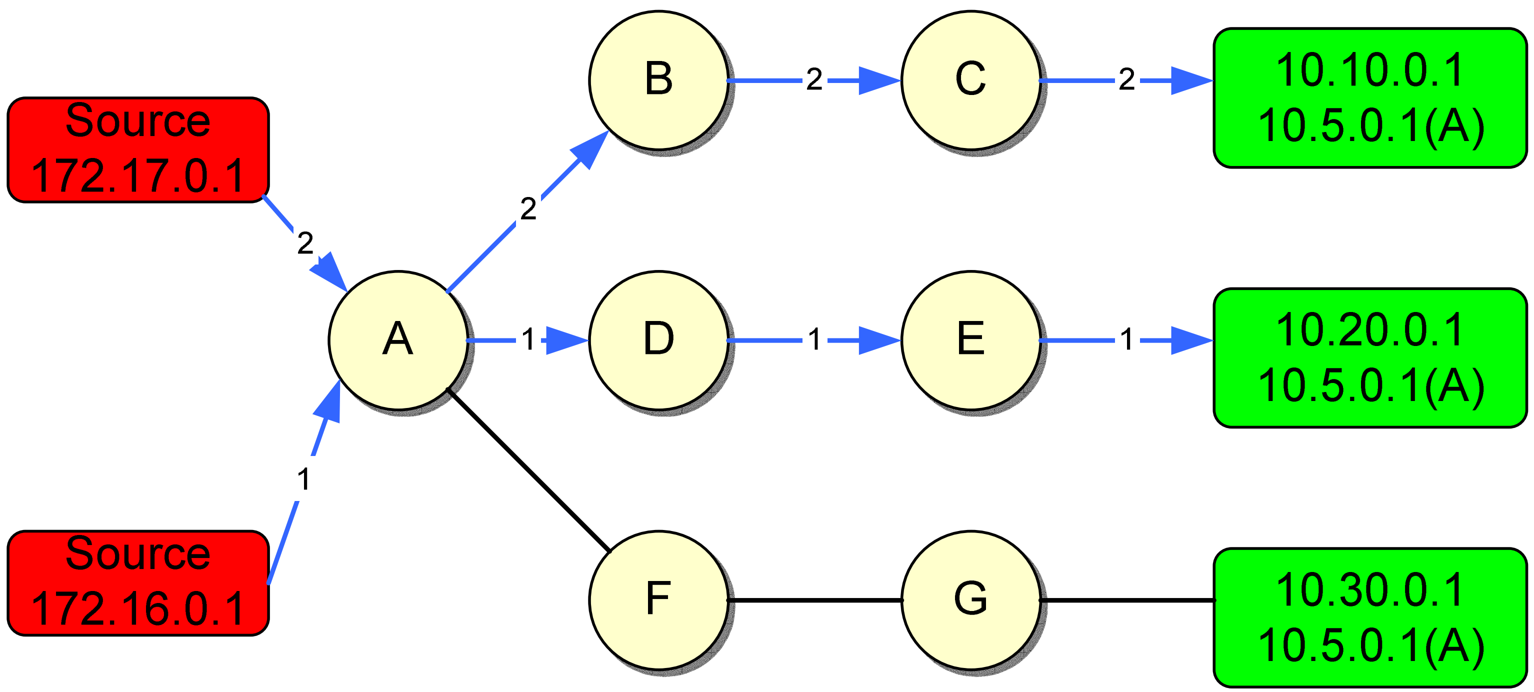
Anycast

- Sequential packets may be delivered to different anycast nodes



Anycast

- Traffic from different nodes may follow separate paths



Carnegie Mellon

Anycast

- **Server receiving a packet is determined by unicast routing**
- Sequential packets from a client to an anycast address may be delivered to different servers
- Best used for single request/response type protocols

Anycast

- Clients, servers, and routers require no special software/firmware
- Does not negatively interfere with existing networks
- Just leveraging existing infrastructure

Anycast Documented

- Concept discussed in RFC1546 (11/93)
  - Current practices have evolved from operational experience
  - CIDR eliminated a hurdle from 1546
- Evolution is briefly documented in RFC2101 (2/97)
- Anycast DNS noted in RFC2181 (7/97)
  - Reply source address must match request dest address

Anycast Documented

- IPv6 anycast – different, will discuss later
  - Architecture (RFC1884, now RFC3513)
  - Reserved anycast addresses (RFC2526)
  - Anycast v4 prefix for 6to4 routers (RFC3068)
  - Source address selection (RFC3484)
  - DHCP (RFC3315)
- Anycast authoritative name service (RFC3258)
- Anycast for multicast RP (RFC3446)
- ISC Technote (ISC-TN-2003-1)
- Term 'anycast' used in 51 RFCs total

Agenda

- **Deploying IPv4 anycast services**
  - – Address selection
  - – Host configuration
  - – Service configuration
  - – Network configuration
  - – Monitoring and using anycasted service

Address Selection

- Current practice is to assign anycast addresses from unicast IP space
- Designate small subnet(s) for anycast use
  - Consider best practices of inter-domain routing announcements
  - /24 is a popular selection
  - Subnet may not be attached to any interface

Host Configuration

- Hosts need to be configured to accept traffic to anycast address
- Want to maintain a unique management address on each host
- Typically, anycast addresses are configured as additional loopbacks
- Make sure ingress filters are updated!

Configuring Addresses

Linux

```
# ifconfig lo:1 10.5.0.1 netmask 255.255.255.255 up
# ifconfig lo:1
lo:1      Link encap:Local Loopback
          inet addr:10.5.0.1  Mask:255.255.255.255
          UP LOOPBACK RUNNING  MTU:16436  Metric:1
```

OpenBSD

```
# ifconfig lo0 alias 10.5.0.1 netmask 255.255.255.255
# ifconfig lo0
lo0: flags=8049<UP,LOOPBACK,RUNNING,MULTICAST> mtu 33224
        inet 127.0.0.1 netmask 0xff000000
        inet 10.5.0.1 netmask 0xffffffff
```


Carnegie Mellon

Configuring Addresses

Solaris

```
# ifconfig lo0 addif 10.5.0.1/32 up
Created new logical interface lo0:1
# ifconfig lo0:1
lo0:1: flags=1000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4>
        inet 10.5.0.1 netmask ffffffff
```

Network Configuration

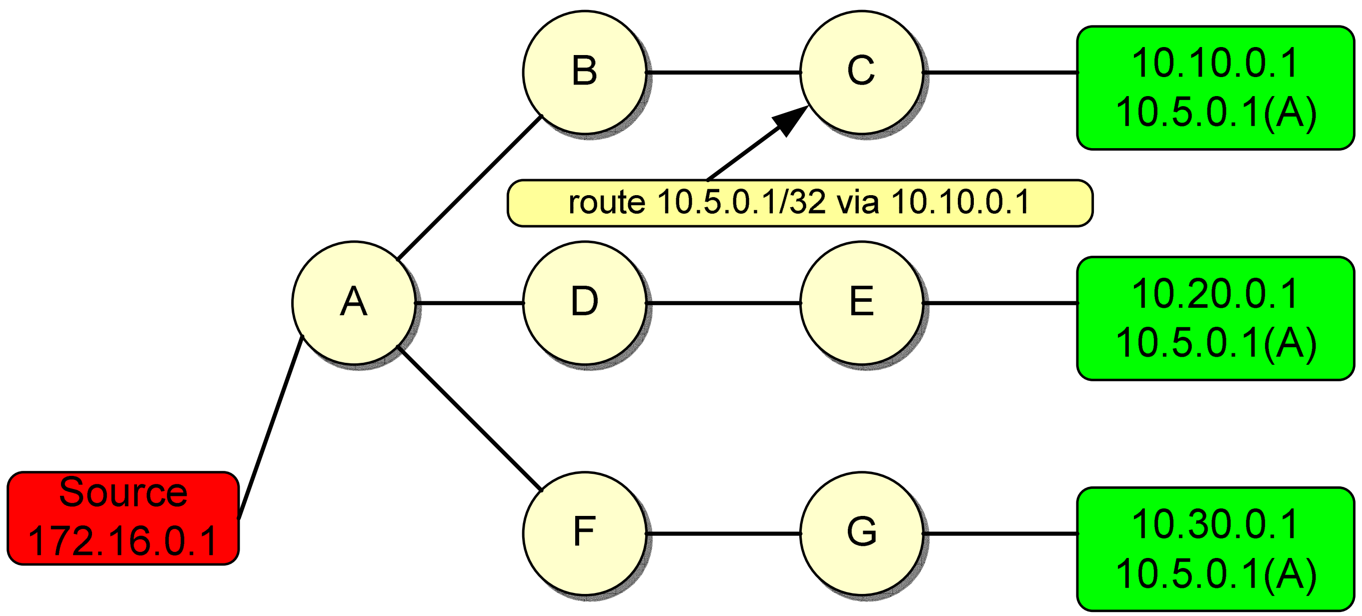
- Correctly configuring the network may be the trickiest aspect of anycast
- Intra-domain vs. inter-domain configuration

Intra-Domain Configuration

- If the anycasted service is entirely within your routing domain, only intra-domain consideration is needed
  - All anycast nodes are within domain
  - Or multiple "intra-domain" locations
- Need to configure routing to deliver anycast traffic to servers

Static IGP Routes

- Simple: configure static routes on first-hop routers (host routes)
- Ensure routes are propagated through domain



Carnegie Mellon

Static IGP Routes

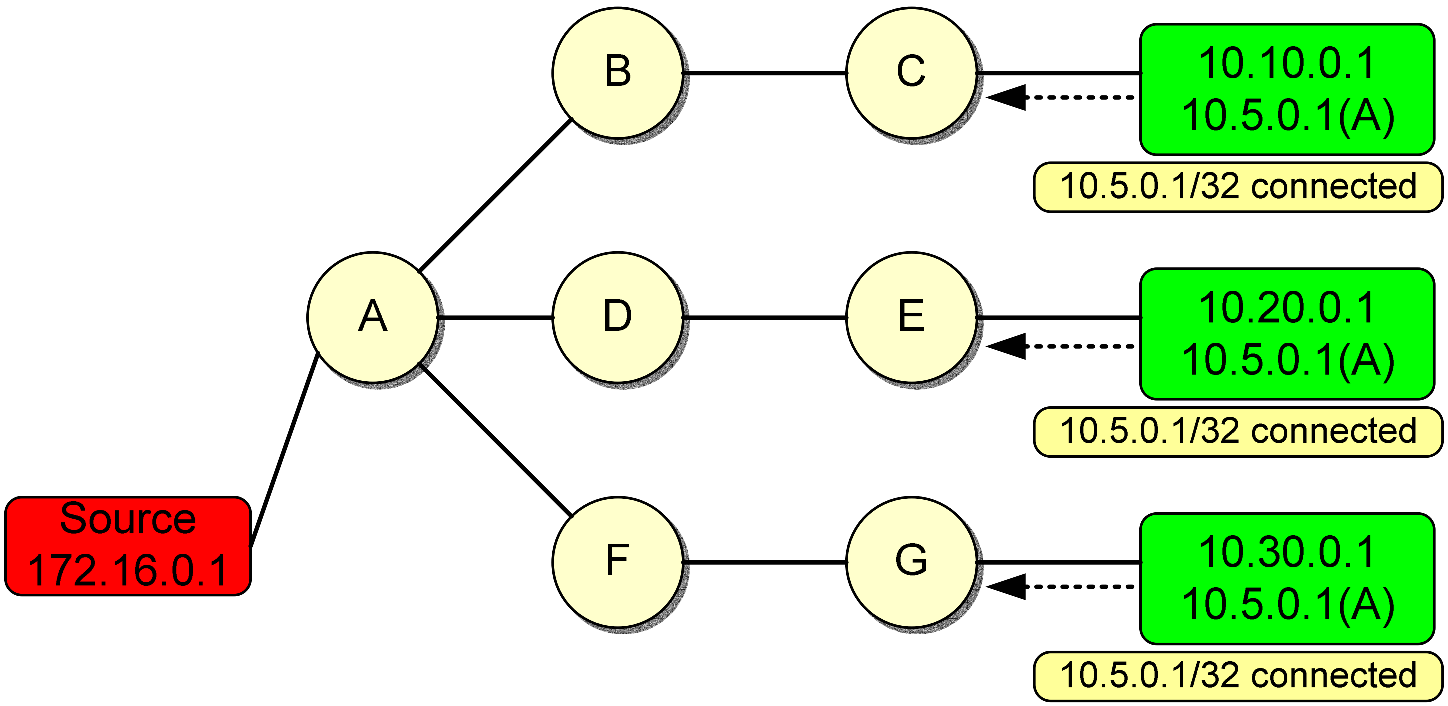
- Simple to configure
- Doesn’t respond quickly to server failure
- Provides the ability to relocate servers without service outage, though

Dynamic IGP Routes

- Run a host-based routing daemon on anycast servers
  - GateD
  - Zebra/Quagga
- Host itself is route originator
- When host is down, route is withdrawn
- Leverages routing infrastructure

Dynamic IGP Routes

- Each host announces route to IGP cloud



Dynamic IGP Routes

- Configuration obviously specific to IGP
  - Connected route redistribution, if anycast addresses are host loopbacks
- Host up doesn't imply service is up
  - Want a mechanism for withdrawing routes automatically when service is unusable

Inter-Domain Configuration

- Follow traditional BGP operating rules
  - Announce from a consistent origin AS
  - Advertise the service/anycast supernet
  - Limit route flapping
  - Provider-independent IP space

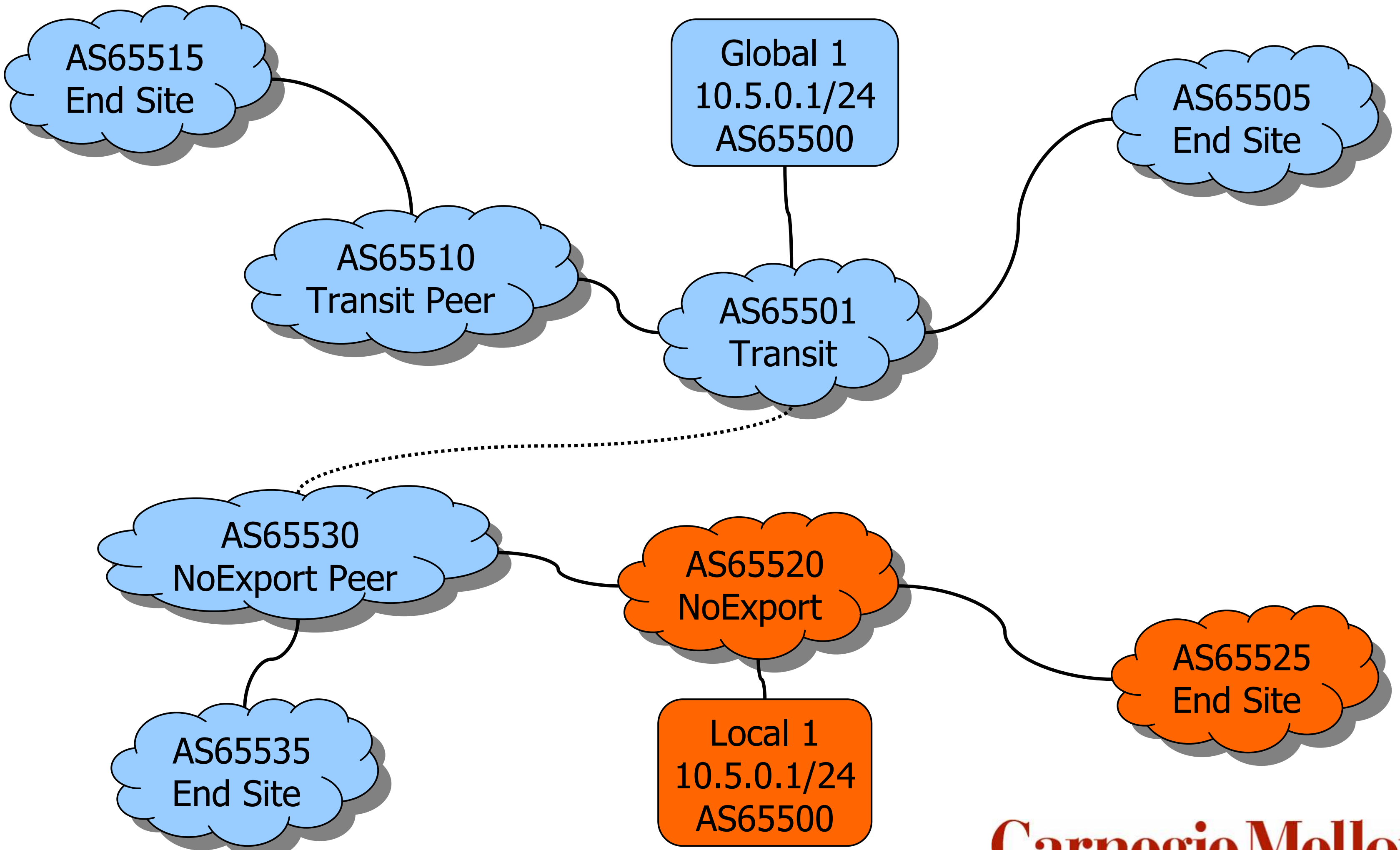
Inter-Domain Configuration

- Intra-domain routing must be correct
  - Servers can be iBGP peered; ‘network’ style announcement on the host
  - Can use IGP with redistribution
- Withdraw routes when service is unavailable at a particular location

Inter-Domain Configuration

- Some deployments distinguish “global” nodes from “local” nodes
  - Global nodes are announced without limitation; upstream provides transit
  - Local nodes add “no-export” community to limit the clients that will use the node
- Why?
  - Money (global/local imply different relationships)
  - Node stability, capabilities (due to service area)

Inter-Domain Configuration



Carnegie Mellon

Service Configuration

- Obviously depends on implementation
- Configure service to listen on anycast IP
  - Most require no special configuration
  - Verify that service responds **from** anycast address when queried
  - May want to limit service to listen **only** on anycast IP address
- Assume: identical service by each server

Anycast Address Use

- Clients told to use anycast addresses
- Anycast service configured by name
  - Authoritative DNS, Syslog, Kerberos
  - Can use round-robin DNS for additional redundancy

```
;; ANSWER SECTION:
ns1.example.com.    86400  IN     A      10.5.0.1
```

Anycast Address Use

- Caching DNS Service
  - Assign server addresses by DHCP, PPP, word of mouth
  - Note the poor behavior of OS stub resolvers
    - The first configured DNS Server is tried on **every query**
    - Results in multi-second delays for many queries
    - Perfect opportunity for anycast

Monitoring

- Monitoring is more complicated
- Could monitor the unique (non-anycast) IPs, but doesn't verify the actual service
- Monitoring the anycast (service) IP can't be done centrally
- Distributed monitoring needed for distributed service
- Also want to monitor routes

Agenda

- What is Anycast?
- Deploying IPv4 anycast services
- **Anycast usage case studies**
- Advanced Topics

Anycast in Action

- Authoritative DNS
  - AS 112
  - Root Servers: F, I, K, others
  - .ORG Top Level Domain
- Caching DNS
- Anycast for Multicast RP
- Anycast Sink Holes
- 6to4 routers (RFC3068)

AS112 Project

- Problem: Many clients try queries and updates for/to RFC1918/link local reverse zones
- Goal: Reduce unnecessary root server load from these queries/updates
- Solution: Delegate reverse zones to anycasted black-hole servers.
- www.as112.net

AS112 Project

- Black-hole servers use IPs in
  `192.175.48.0/24`
- Announced from 16 locations worldwide
- Common origin AS112

Configuring BGP

One Vendor..

```
router bgp 112
  bgp router-id 192.175.48.254
  network 192.175.48.0
  neighbor PEER_IP remote-as PEER_AS
  neighbor PEER_IP ebgp-multihop
  neighbor PEER_IP next-hop-self
```

[http://www.chagreslabs.net/jmbrown/research/as112/]

Another Vendor..

```
policy-statement advertise-aggregate {
  term first-term {
    from protocol aggregate;
    then accept;
  }
  term second-term {
    from route-filter 192.175.48.0/24 longer;
    then reject;
  }
}
[continued]
```


Carnegie Mellon

Configuring BGP

Another Vendor..

```
# set routing-options aggregate route 192.175.48.0/24

[edit protocols bgp group 112]
# set export advertise-aggregate
# set type external
# set peer-as PEER_AS
# set neighbor PEER_IP
```

AS112 Project

BIND Configuration

```
zone “10.in-addr.arpa” { type master; file “db.RFC-1918”; };
zone “254.169.in-addr.arpa” { type master; file “db.RFC-1918”; };
...
```

Zone File: db.RFC-1918

```
$TTL 300
@ IN SOA prisoner.iana.org. hostmaster.root-servers.org. (
                         2002040800 30m 15m 1w 1w )
     NS blackhole-1.iana.org.
     NS blackhole-2.iana.org.
```

[http://www.chagreslabs.net/jmbrown/research/as112/]

Root Servers

- Problems:
  - Low concentration of root servers outside the US (high latency, higher cost links)
  - DoS attacks hurt the root servers and infrastructure in between
  - Can't just add more NS records to root zone
- Goal: Add root servers to underrepresented areas of the world
- Solution: Use inter-domain anycast to serve existing root server IP addresses

Root Servers

- "F" root server (ISC) anycasted
  - First cloned node announced Nov. 2002
  - Now have 12 locations
  - Common origin AS3557
  - Second hop AS for local nodes also assigned to the ISC

Global Node AS Path
`... 3557 3557 3557`

Local Node AS Path (Ex)
`... 23709 3557`

.ORG Top Level Domain

- Recent NANOG discussion analyzing anycast use for .ORG
- Suggestion of an outage at one location
  - "the monitors that test each of the anycast nodes reported no outages"
- DNS provided by 2 anycast servers (`204.74.112.1, 204.74.113.1`)
- Two /24s; different transit providers
- Eleven total second-hop ASs

.ORG Top Level Domain

- Highlights anycast lessons
  - – Operators will encounter anycast
    - Different locations, different experiences
  - – Service availability and routing announcements are coupled
  - – Consider reliability mechanisms built into service – they can help or hurt

Caching DNS

- Problems:
  - Hosts respond poorly when caching nameserver is unreachable
  - Caching NS is hard to re-IP (static configs)
- Goal: Always have caching DNS service on first client-configured IP
- Solution: Use anycasted servers; configure anycast IPs on clients

Caching DNS

- We designated `128.2.1.0/26` for intra-domain anycast use (from our IP space)
- Two caching server IPs: `128.2.1.10`, `128.2.1.11`
- Using BIND9
- Configured on 4 servers; 6 interfaces
- Addresses assigned by DHCP, PPP

Caching DNS

- Each server runs host-based routing daemon (Quagga) to join OSPF cloud
- Using OSPF NSSA areas to hosts
  - Minimizes the number of routes on the servers
  - Enables multiple interfaces on servers in separate NSSA areas but no forwarding through server

Caching DNS Config

BIND 9 Changes

```
options {
        listen-on { 128.2.1.10; 128.2.1.11; };
        query-source address 128.2.4.21;
};
```

Upstream Router Changes

```
router ospf 1
 area 0.0.0.0 authentication message-digest
 area 128.2.4.0 authentication message-digest
 area 128.2.4.0 nssa default-information-originate no-summary
 network 128.2.4.0 0.0.0.63 area 128.2.4.0
 network 128.2.0.0 0.0.255.255 area 0.0.0.0
```

Caching DNS Config

BIND 9 Changes

```
options {
        listen-on { 128.2.1.10; 128.2.1.11; };
        query-source address 128.2.4.21;
};
```

Note generic lesson:
Make sure servers aren't sourcing non-response traffic from anycast addresses.

Upstrea[illegible]

```
router os[illegible]
 area 0.0[illegible]est
 area 128.2.[illegible]-digest
 area 128.2.4.0 nssa default-information-originate no-summary
 network 128.2.4.0 0.0.0.63 area 128.2.4.0
 network 128.2.0.0 0.0.255.255 area 0.0.0.0
```

Caching DNS Config

Upstream Router, Said Another Way

```
[edit protocols ospf]
area 4 {
  nssa {
    area-range 128.2.4.0/26;
    default-lsa {
      default-metric metric;
      type-7;
    }
    no-summaries;
  }
  authentication-type md5;
  interface interface;
}
```

Host-Based Router Config

quagga.conf

```
interface eth0
 ip address 128.2.4.21/26
!
interface lo:1
 ip address 128.2.1.10/32
!
interface lo:2
 ip address 128.2.1.11/32
```

ospfd.conf

```
interface eth0
 ip ospf authentication message-digest
 ip ospf message-digest-key 1 md5 [key]
!
router ospf
 ospf router-id 128.2.4.21
 ospf abr-type cisco
 compatible rfc1583
 area 128.2.4.0 authentication message
 area 128.2.4.0 nssa
 network 128.2.4.21/26 area 128.2.4.0
 redistribute connected
 distribute-list 50 out connected
!
access-list 50 permit host 128.2.1.10
access-list 50 permit host 128.2.1.11
```

Multicast RP

- Problem:
  - PIM-SM specifies one active RP per multicast group at a time
  - A routing domain may be too large for this to be feasible (RP on the other coast)
  - Slow failover if RP fails
  - Not directly possible for shared-tree load balancing
- Goal: Multiple RPs for same group within a routing domain
- Solution: Use anycast addresses for RP (RFC3446)

Carnegie Mellon

Multicast RP

- Designate more than one RP
- Assign anycast address as loopback on each RP
- Configure all other routers to use anycast address as RP for all groups
- Setup MSDP mesh among all RPs (using **unique** addresses)
  - RP address cannot be used in SA messages

Multicast RP

RP Routers

```
interface Loopback0
 description Router Management
 ip address 10.2.4.249 255.255.255.255
 ip pim sparse-mode
interface Loopback1
 description Anycast RP Interface
 ip address 10.2.1.130 255.255.255.255
 ip pim sparse-mode
!
ip msdp peer 10.2.4.248 connect-source Loopback0
ip msdp peer 10.2.4.250 connect-source Loopback0
ip msdp mesh-group CMU-MSDP 10.2.4.248
ip msdp mesh-group CMU-MSDP 10.2.4.250
ip msdp cache-sa-state
ip msdp originator-id Loopback0
```

Non-RP Routers

```
ip pim rp-address 10.2.1.130 override
ip pim accept-rp 10.2.1.130
```

Multicast RP

RP Routers

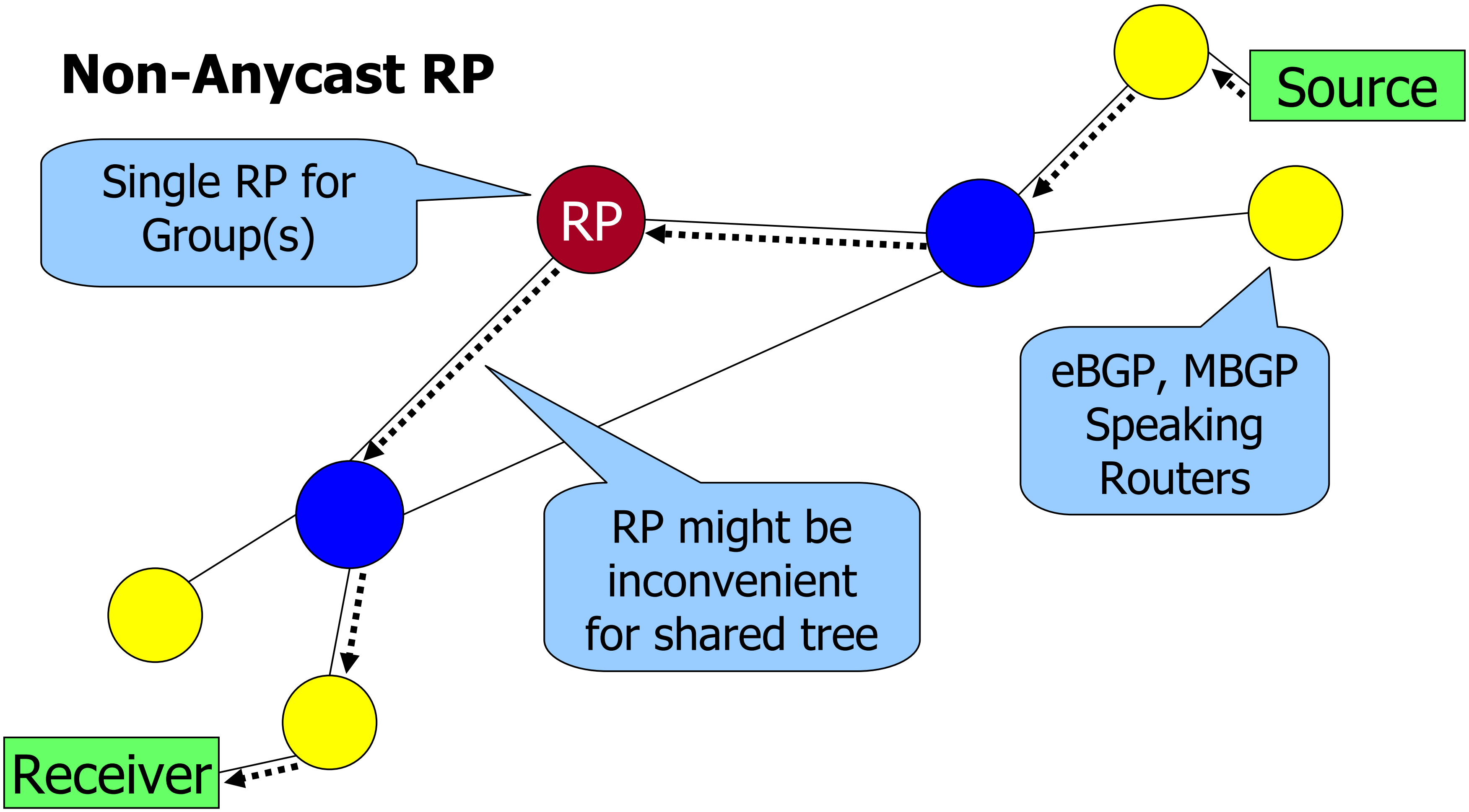
```
[edit interfaces lo0 unit 0 family inet]
# set address 10.2.4.249/32
# set address 10.2.1.130/32
[edit protocols pim]
rp {
  local {
    address 10.2.1.130;
  }
}
interface all {
  mode sparse;
  version 2;
}
[edit protocols msdp]
group CMU-MSDP {
  local-address 10.2.4.249;
  mode mesh-group;
  peer 10.2.4.248;
  peer 10.2.4.250;
}
```

Non-RP Routers

```
[edit protocols pim]
rp {
  static {
    address 10.2.1.130 {
      version 2;
    }
  }
}
```

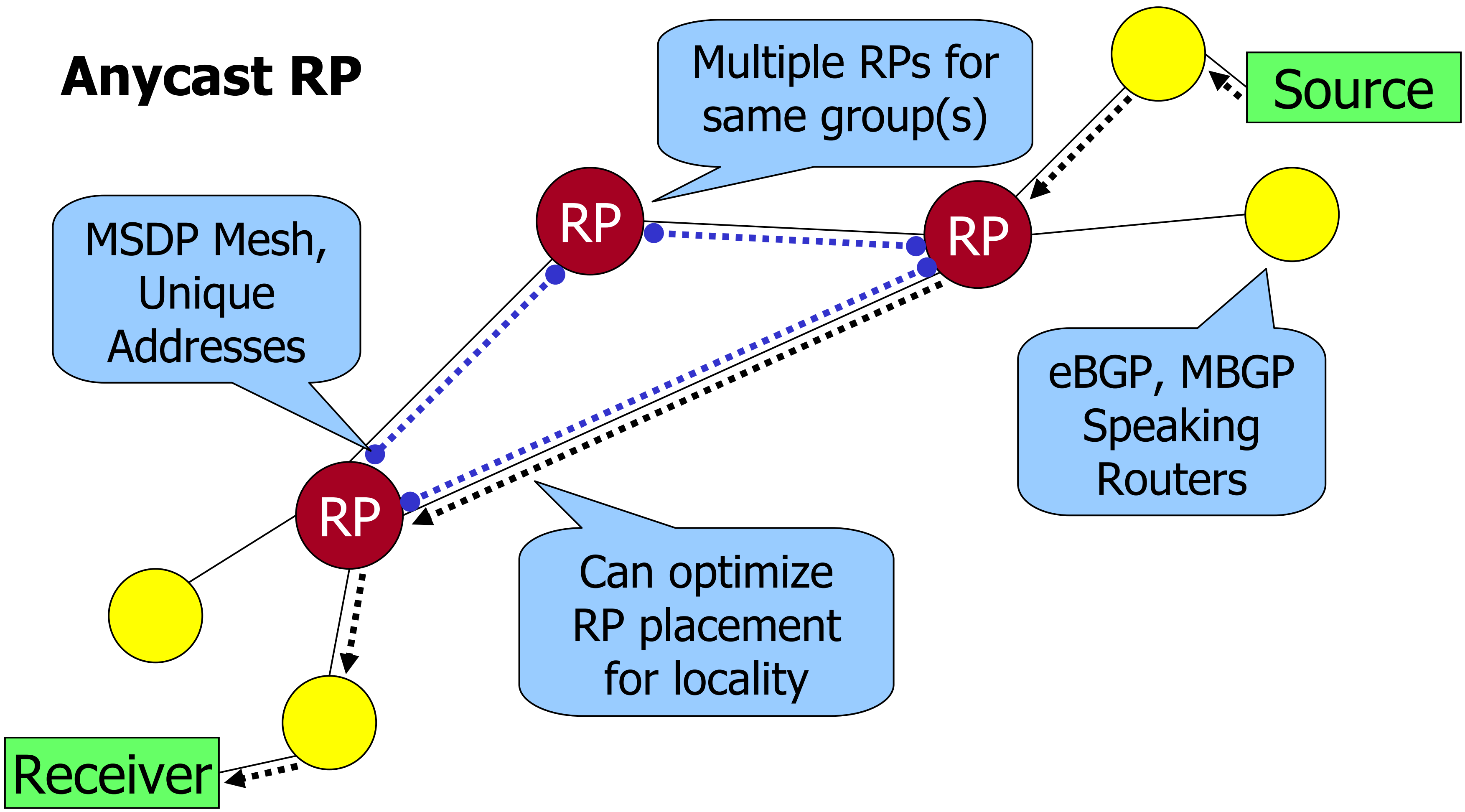
Multicast RP

Non-Anycast RP



Carnegie Mellon

Multicast RP



Carnegie Mellon

Anycast Sinkholes

- Problem: Homeless network traffic (e.g. turbo worms, backscatter, etc) can cause problems for core routers to sink; sinkholes help but also don't want to send traffic across large network to sink
- Goal: Want to be able to forward traffic to multiple sinkhole points for analysis
- Solution: Use anycast to enable distributed sinkholes throughout a large network

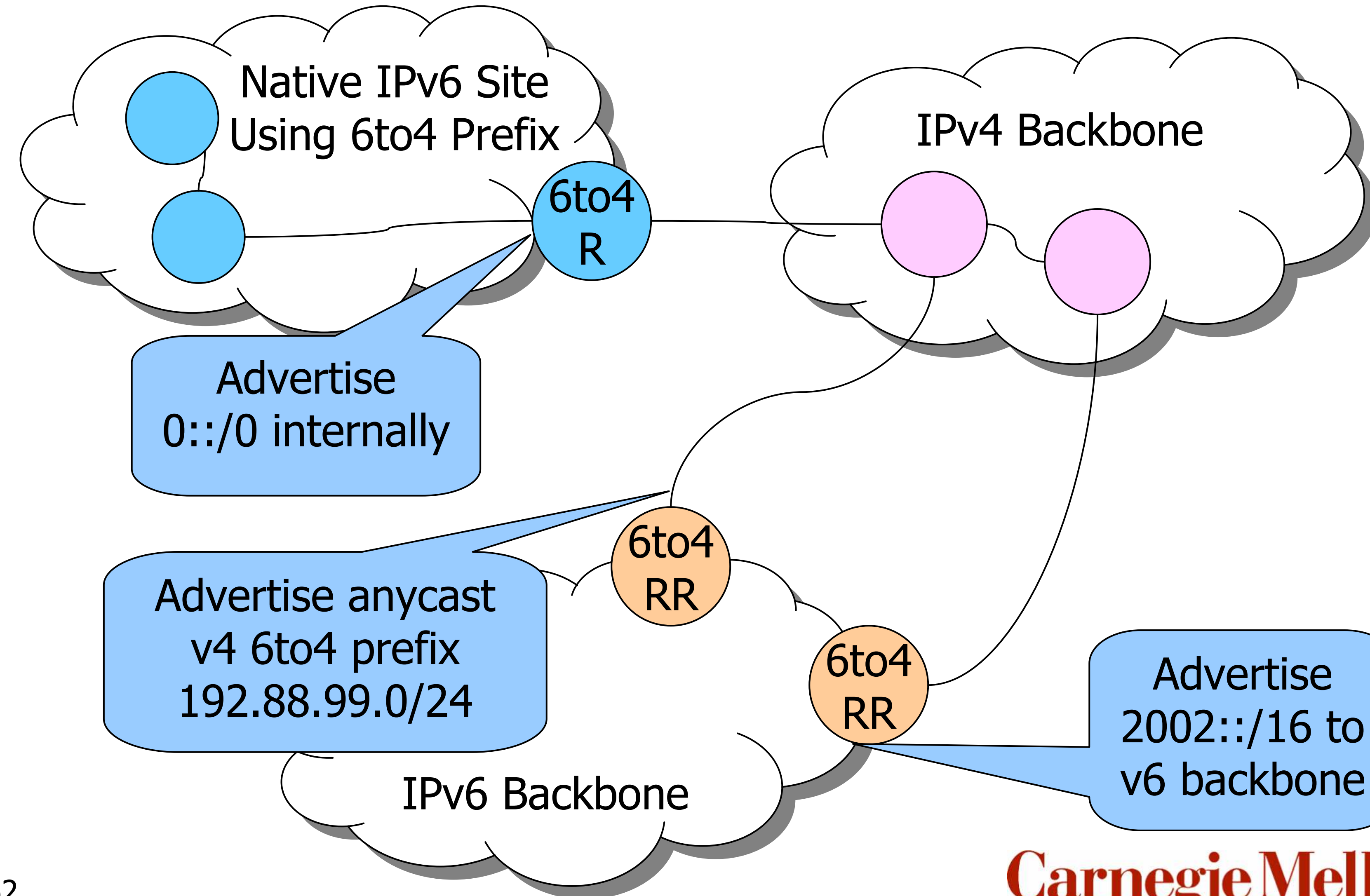
Anycast Sinkholes

- Traffic can be directed to sinkhole via:
  - default route
  - pieces of unused IP space
  - BGP next-hop triggering, ex: for DoS victims
- Multiple sinkholes can be deployed using anycast as sinkhole destination address
- Very good slides by Greene, McPherson (see resources page for links)

6to4 Routers

- Problem: Connecting islands of v6 across existing v4 infrastructure involves 6to4 relay routers
- Goal: Provide an easy way for end sites to locate relays into the native v6 world
- Solution: Use a well-known IPv4 anycast prefix for 6to4 relay routers

6to4 Routers



TCP-Based Services

- Unwise to use anycast for long-term TCP services, due to route changes
- Experience shows that routes are generally stable, though
  - Especially inter-domain, due to routing protocols
  - Equal cost load balancing would cause problems
  - But, routers often do flow path caching

TCP-Based Services

- Very few knobs to direct traffic in response to server load, as well

> "as long as you don't make silly assumptions about client locality based on "which anycasted server heard it", such that you give back incoherent answers in hopes that they will be somehow client-optimal, bgp-anycast isn't even controversial at this point in time."
>
> - Paul Vixie, 4/03

Other (Potential) Uses

- NTP/Time
- Syslog
- RADIUS
- Kerberos
- Single packet request-response UDP protocols are “easy”

Agenda

- What is Anycast?
- Deploying IPv4 anycast services
- Anycast usage case studies
- **Advanced Topics**
  - Multi-homed hosts
  - IPv6

Multi-Homed Hosts

- Multi-homing at the host physical interface
- Can be used with anycast addressing
- Special case: single multi-homed host configured with anycast address
  - More appropriately a ‘service’ address
  - Server redundancy, no service separation
  - Much of the same configuration
  - Additional complications with default route

IPv6 Anycast

- IPv6 Anycast, per RFC3513, is different from “shared-unicast” addressing (what we’re calling anycast)
  - 3513: Eliminate constraints on routing infrastructure, upper-layer protocols
  - Decouple Anycast from any thought about TCP/UDP (and still make it work)
  - “Shared Unicast” IPv6 would generally map from v4 experiences
- Hagino, Ettikan draft addresses the differences and limitations

3513 vs Shared Unicast

Issue	RFC3513 Anycast	Shared Unicast
Identifying anycast dest.	Same address format as unicast	Same address format as unicast
Deterministic packet delivery	None – seq. packets may reach diff. hosts	None – seq. packets may reach diff. hosts
Anycast host addresses	Disallowed; routers only	No restriction
Anycast as source addr.	Disallowed	Required for current operational use
IPsec	Difficult – instability of addressing and routing	Difficult – instability of addressing and routing

3513 vs Shared Unicast

Issue	RFC3513 Anycast	Shared Unicast
Identifying anycast dest.	Same ad unicast	as
Deterministic packet delivery	None – s reach dif	ay
Anycast host addresses	Disallowed; routers only	No restriction
Anycast as source addr.	Disallowed	Required for current operational use
IPsec	Difficult – instability of addressing and routing	Difficult – instability of addressing and routing

Why?
Questions about how hosts announce routes into domain.
Shared-unicast solutions apply.

3513 vs Shared Unicast

Issue	RFC3513 Anycast	Shared Unicast
Identifying anycast dest.	Same address format as unicast	Same address format as unicast
Deterministic packet delivery	None – se reach diff.	
Anycast host addresses	Disallowe	
Anycast as source addr.	Disallowed	Required for current operational use
IPsec	Difficult – instability of addressing and routing	Difficult – instability of addressing and routing

Why?
Anycast addresses don't uniquely identify source node. Trying to define most generic solution at IP layer.

Carnegie Mellon

IPv6 Anycast Improvements

- Allow hosts to have 3513 Anycast addresses
  - Just need to define mechanism(s) for announcing routes into domain
- Provide deterministic endpoint with anycast
  - Could use routing header specifying non-anycast address as intermediate hop
- Anycast address as source address?
  - Source address is unique machine address
  - Home address option of Anycast address
  - Would break semantic equality of source address/home address

IPv6 Anycast Protocol Issues

- 3513 and UDP
  - DNS, etc. require matching source address as queried destination
  - "Use better security and drop the checks"
- 3513 and TCP
  - Connections identified by address/port pair of source/destination
  - Provide a means for changing address of connection/initiating new connection?

IPv6 Anycast General Issues

- With emphasis on route aggregation, questions arise about global inter-domain shared-unicast
- 3513 Anycast specifically: host routes must be carried within aggregation domain encompassing all interfaces of specific Anycast IP

IPv6 Anycast Addresses

- Reserved addresses/ranges (3513)

Subnet-Router Anycast Address

n bits	128-n bits
subnet prefix	**0000000000000**

Reserved Anycast Addresses

	n bits	121-n bits	7 bits
Non-EUI-64	**subnet prefix**	**11111111111**	**anycast ID**

	64 bits	57 bits	7 bits
EUI-64 “Local”	**subnet prefix**	**111111011..11**	**anycast ID**

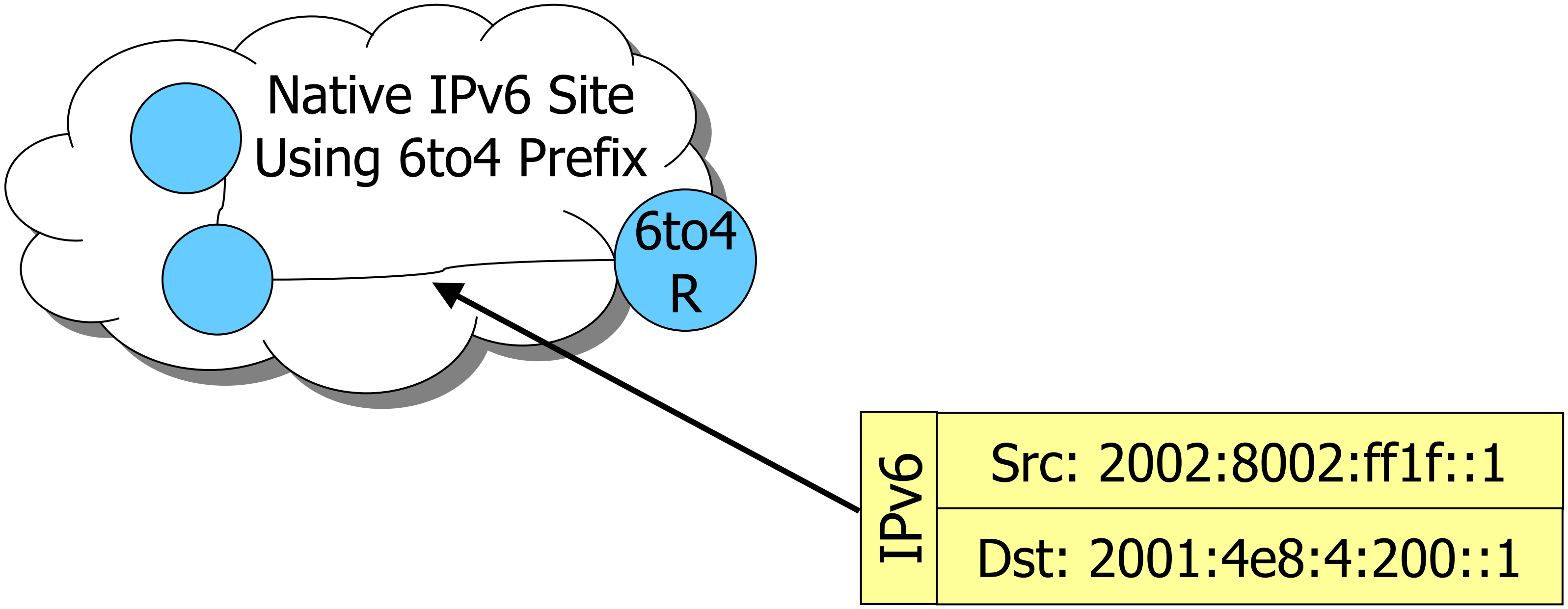
Summary

- Anycast is relatively simple to deploy in existing networks
- Operators are finding new uses for it in different areas
- Look for some changes as v6 comes around

Questions?

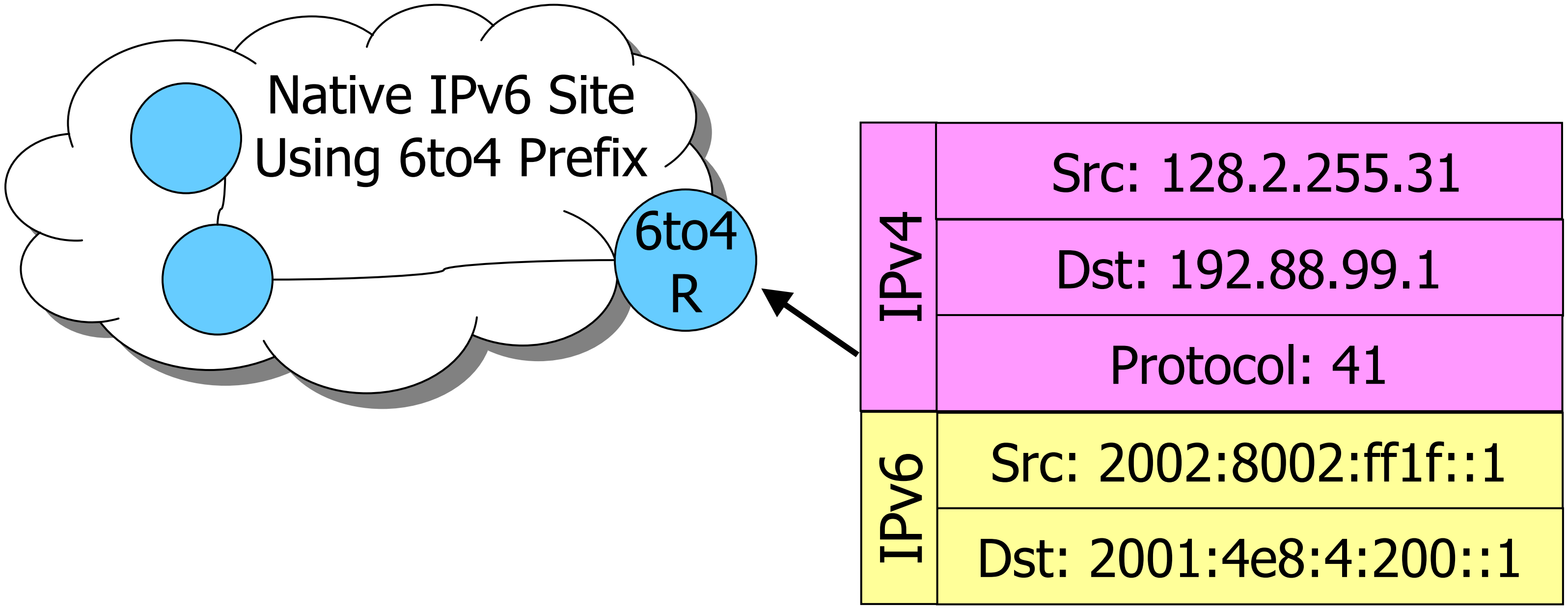
- Presentation resources: http://www.net.cmu.edu/pres/anycast
- Kevin Miller: kcm@cmu.edu

6to4 Routers



- Hosts generate native IPv6 packets

6to4 Routers

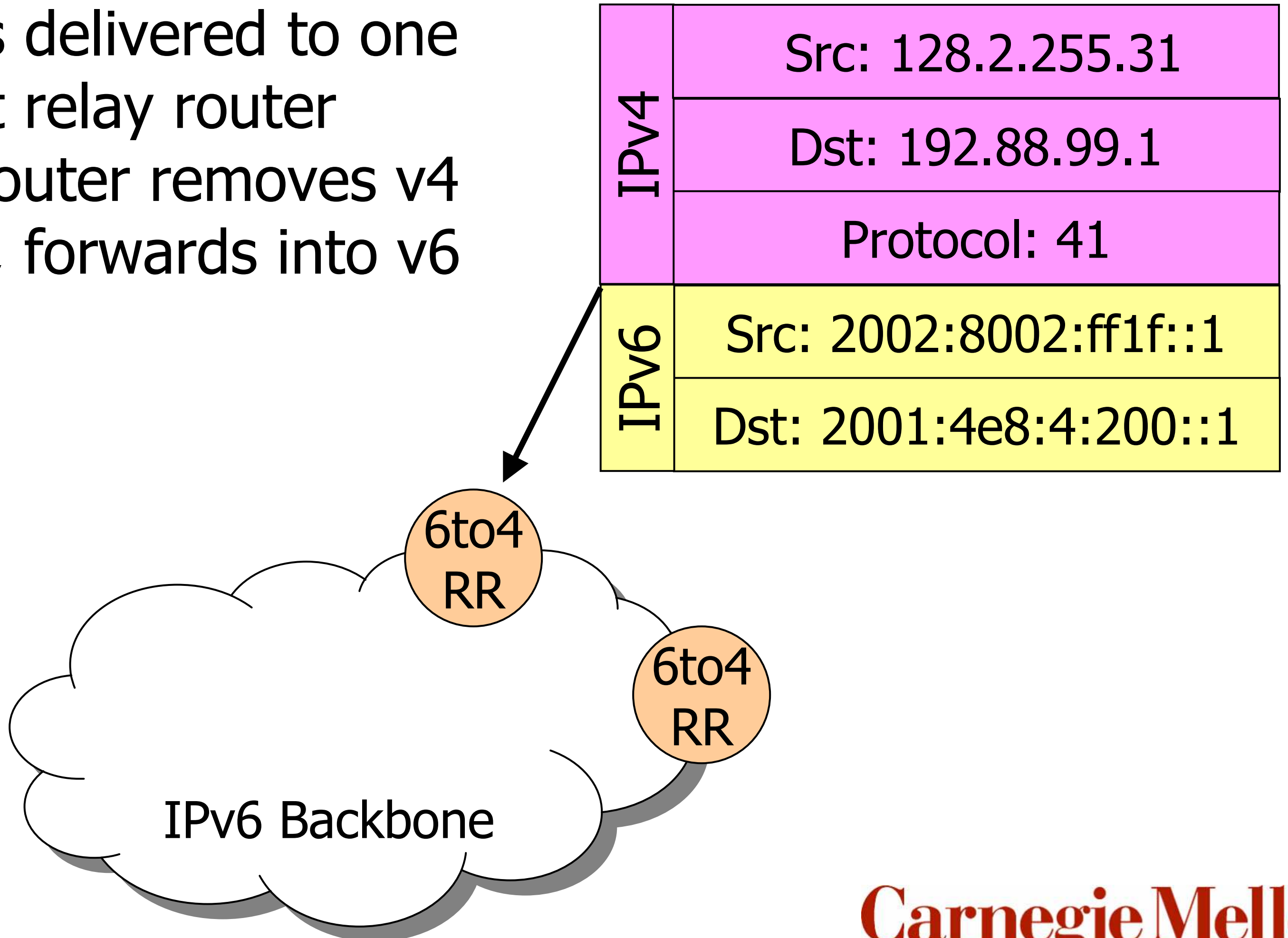


- 6to4 Router encapsulates packet to send over v4 backbone
- Note: External v4 address dictates v6 prefix

Carnegie Mellon

6to4 Routers

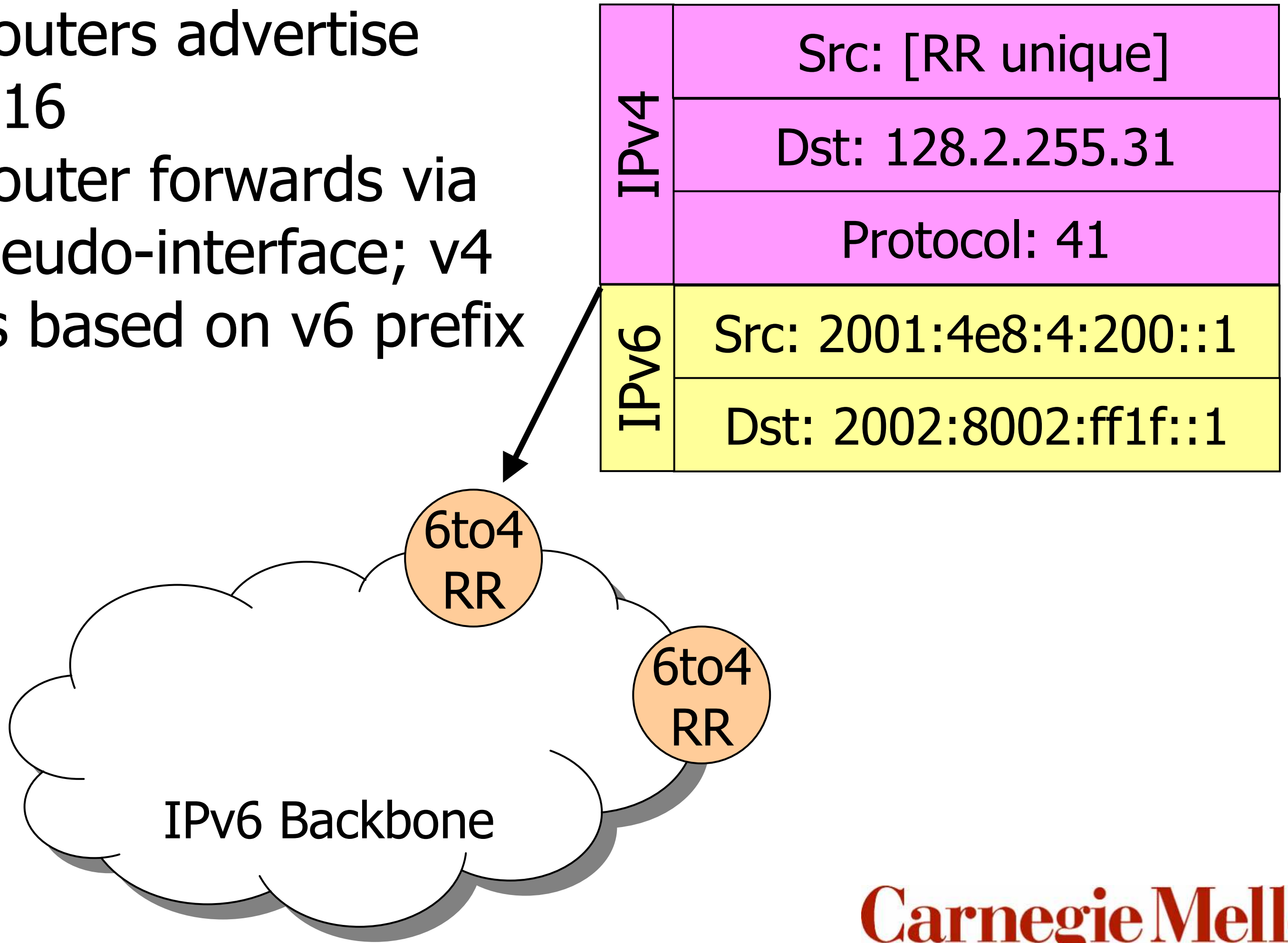
- Packets delivered to one anycast relay router
- Relay router removes v4 header, forwards into v6



Carnegie Mellon

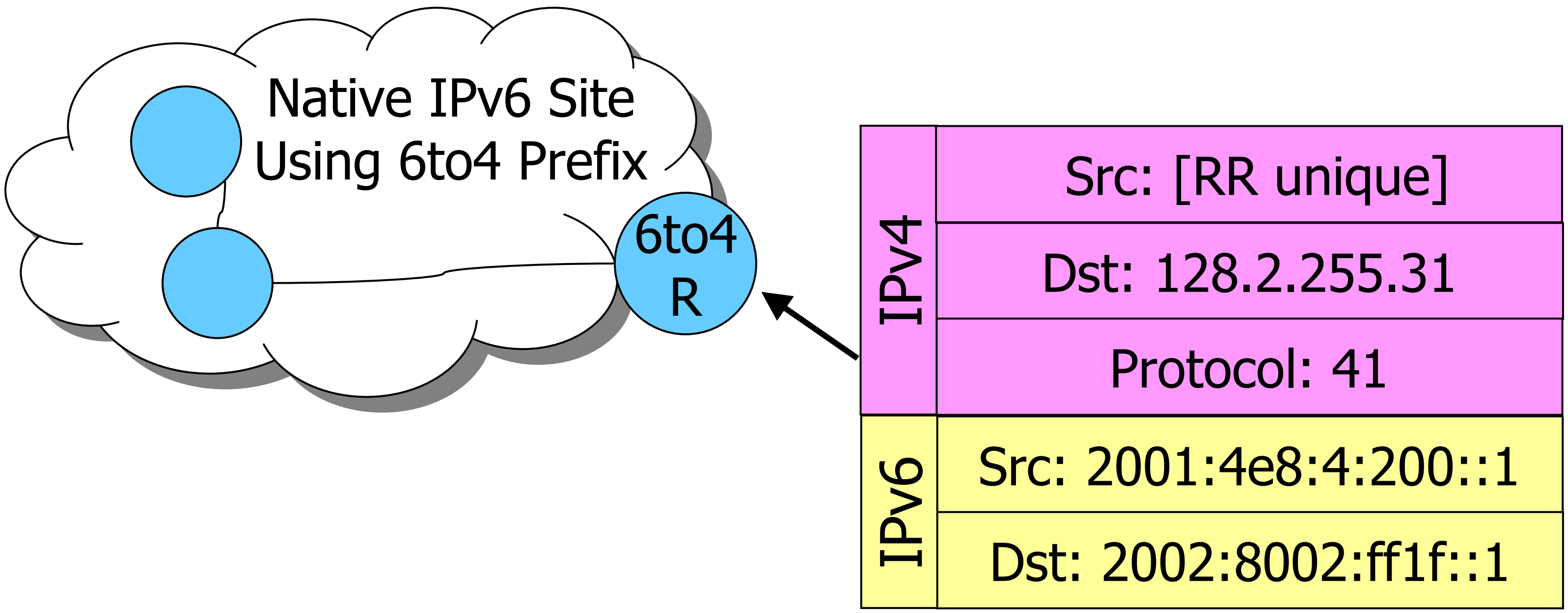
6to4 Routers

- Relay routers advertise 2002::/16
- Relay router forwards via 6to4 pseudo-interface; v4 address based on v6 prefix



Carnegie Mellon

6to4 Routers



- 6to4 Router removes v4 header; forwards v6 packet locally

Carnegie Mellon

Identifying Specific Node

- F Root Server
  - `dig hostname.bind @f.root-servers.net chaos txt`
- K Root Server
  - `dig id.server @k.root-servers.net chaos txt`
- .ORG TLD Servers
  - `dig whoareyou.ultradns.net @tld1.ultradns.net`
  - `dig whoareyou.ultradns.net @tld2.ultradns.net`