SOFTLAYER®

IPv4 Address Conservation Method for Hosting Providers

Wen Temitim and Christopher Papandreou

NANOG 58, June 2013

The Issue

- When allocating subnets for servers in a hosting environment, an IP is bound to a router and the process consumes 3 IP addresses out of the subnet for assignment on the router (Network, Broadcast and Gateway).

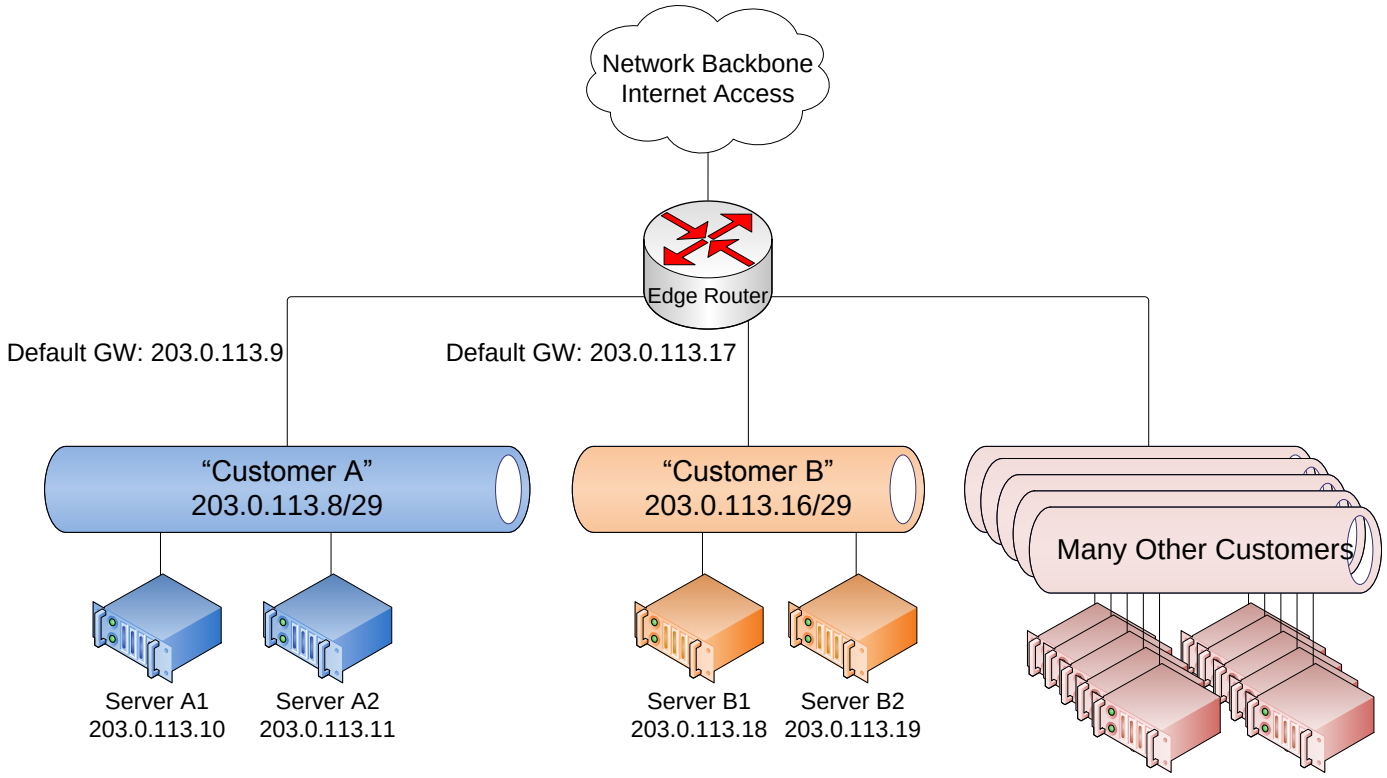
- If one or two servers are assigned to this subnet the most common address allocation for this subnet is a /29.

The Issue cont'd

- **This is a small subnet size but for a subnet that has just two servers 6 IP addresses are wasted to allow connectivity.**

- **This may not seem like a lot but if this same scenario is multiplied across thousands of subnets across different routers it is a very inefficient waste of IP space. In markets like APAC and Europe where APNIC's and RIPE's maximum allocation size is a /22 this eats into IP space very quickly.**

The Issue, Illustrated



6 Public IPs "wasted"
unavailable for customer use:
203.0.113.8, 203.0.113.9, 203.0.113.15
IPs not in use:
203.0.113.12, 203.0.113.13, 203.0.113.14

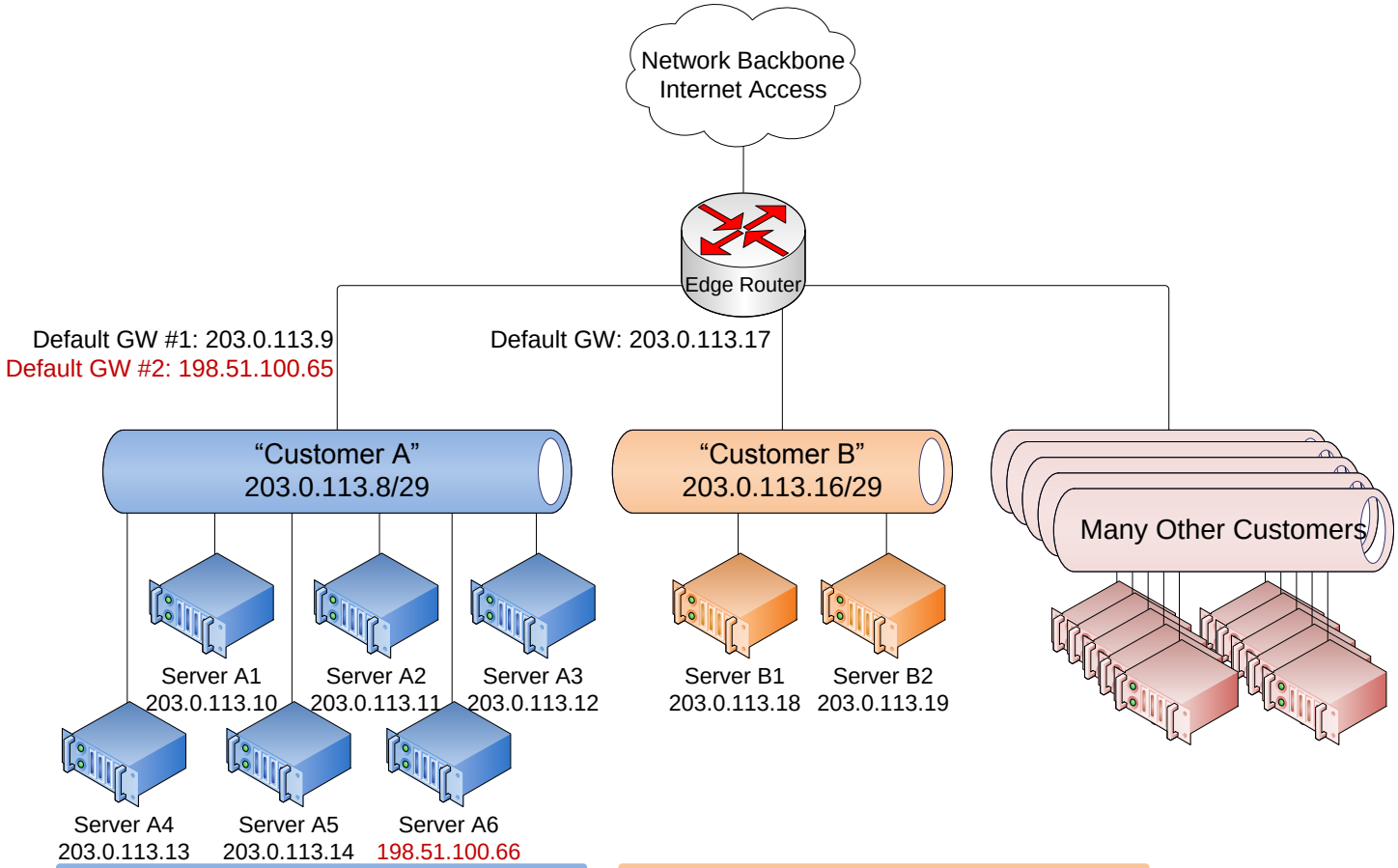
6 Public IPs "wasted"
unavailable for customer use:
203.0.113.16, 203.0.113.17, 203.0.113.23
IPs not in use:
203.0.113.20, 203.0.113.21, 203.0.113.22

The Issue cont'd

- **What happens when the customer that has two servers grows to have 6 servers?**

- **An additional subnet has to be allocated for these servers and if this growth was months after the initial allocation the next contiguous block may not be available to convert the /29 into a /28.**

The Issue cont'd, Illustrated



The Issue cont'd

- What about if you have a redundant environment that relies on a FHRP such as HSRP or VRRP? Even more addresses are used to setup connectivity.

- This may not seem like a big deal now but what happens in X years when you have to go to an IP broker to receive more IPs?

The Solution?

- **Move to IPv6?!?!**

- **We've been dual stack for years but you can't force everyone else to move which means our customers will still want IPv4 connectivity so they can reach their customers.**

Is there a more efficient way?

- **Why does the gateway have to be a publicly routable resource?**
- **The only place the gateway IP address is really used is as a link local address for servers to get a MAC address and as the source IP for router to directly connected host communication.**
- **The actual forwarding of packets destined for remote networks don’t actually reference the IPv4 address.**

IPv4 link local gateway

- **Why not use a “reserved pool” of addresses to act as a link local gateway?**

RFC1918 –
10.0.0.0/8
172.16.0.0/12
192.168.0.0/16

RFC3927 –
169.254.0.0/16

RFC6598 –
100.64.0.0/10 ← winner!

Why not RFC 1918?

- Published in 1996.
- You're likely already using some or all of the space for other uses.
- Potentially confuse Ops staff and customers.

Why not RFC 3927?

- Aren't there IPv4 link local addresses already?
- Yes there are (RFC3927) and you could use them but if you have systems that are multi-homed to different networks you can cause issues.
- If the second interface is configured as a DHCP client and doesn't receive an IP address the OS will configure an IPv4 link local address in the 169.254.0.0/16 range. Your more specific IPv4 link local address should be okay but there is a possibility for issues.

Is it ok to use RFC6598 for this?

- Aren't these blocks reserved for other purposes?
- Yes they are. However in this scenario they are used for link local connectivity so if configured properly they should not interfere with their other uses on your network. In fact, if your provisioning system will allow it, you can designate a small part of this block for just this purpose, and re-use it on each router.

Is it really worth changing?

/best movie trailer guy voice – *Imagine a world where you can no longer receive IP addresses from RIRs…*

- **If there is still demand for IPv4 addresses from your customers, you will have a few options**
  - **- Turn them away**
  - **- Use some sort of NAT gateway (costly and in a hosting environment it will have to be 1:1 NAT since a majority will want 80 and 443)**
  - **- Buy IPs from a broker**

Cost of IP addresses

- There has been some recent press about Microsoft purchasing ARIN space from Nortel for $7.5 million which worked out to $11.25/IP (http://www.networkworld.com/community/blog/microsoft-pays-nortel-75-million-ipv4-address)

- If all of the RIRs are out of IPs, the price per IP will most likely be much higher, but let's use $11.25/IP for our example.

Cost of IP addresses (cont'd)

- Let's say you have a router that has a 1000 SVIs, and each SVI has a primary and secondary address block on it.
- Each block equates to 3 wasted IPs (Network/Gateway/Broadcast).
- If you were paying $11.25 per IP, the cost of this overhead would be:

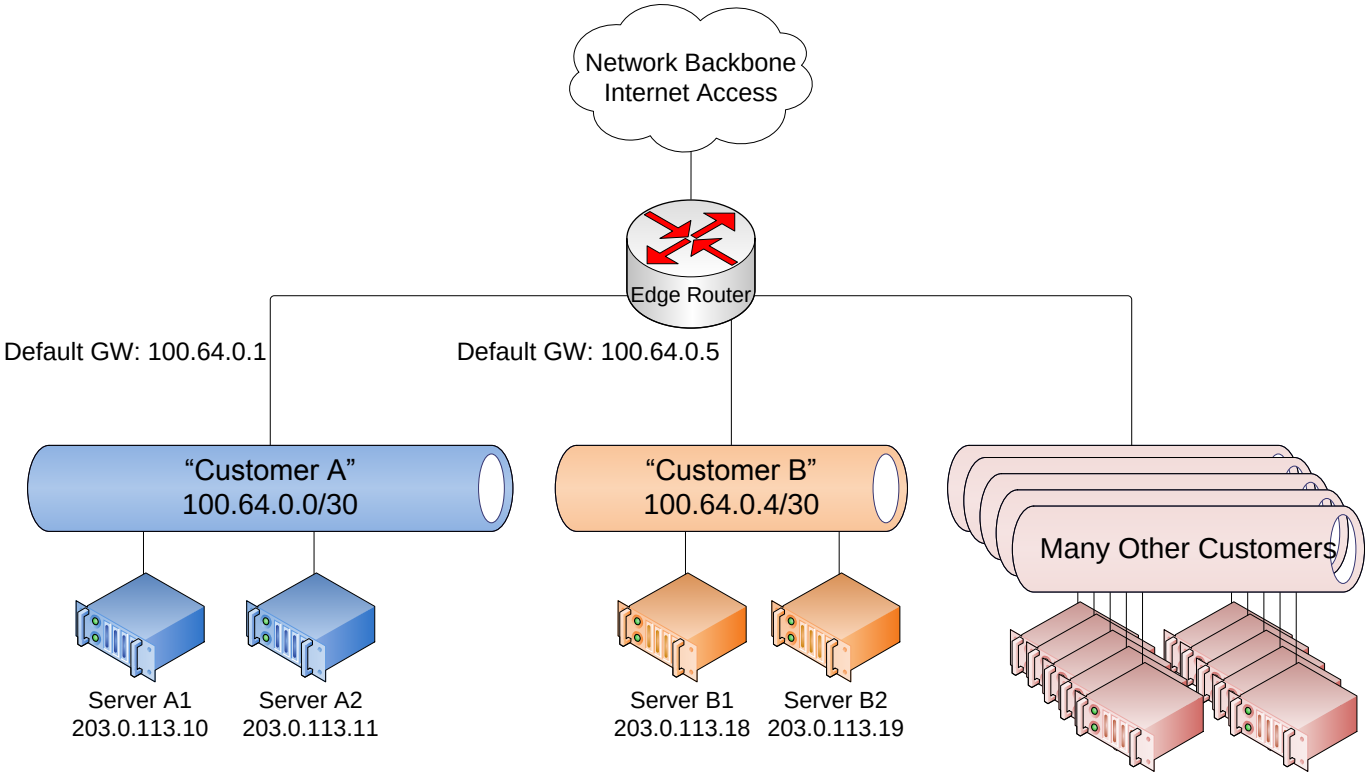
  - *1000 (SVIs) * 2 (blocks per SVI) * 3 (wasted IPs per SVI) * $11.25 = $67,500.00*

- If you are a big network and have a large number of routers, this could get very expensive very quickly.

So how does it work?

- **There are some rules:**
  - Adding one of these subnets will cause unicast RPF to allow traffic sourced from this subnet (you're using uRPF right?) so you need to have an ACL that prevents this .
  - Do not advertise the space into your IGP (or at least tag it so it doesn't get exported into the global routing table).
  - The OS needs to be capable of a binding a /32 subnet mask. Most modern OS's support this: Win7 and on, *nix, FreeBSD, etc. (Win2k3 won't allow it through the GUI, but it works through 'netsh' commands).
  - You're router must be able to support static routes to an interface.

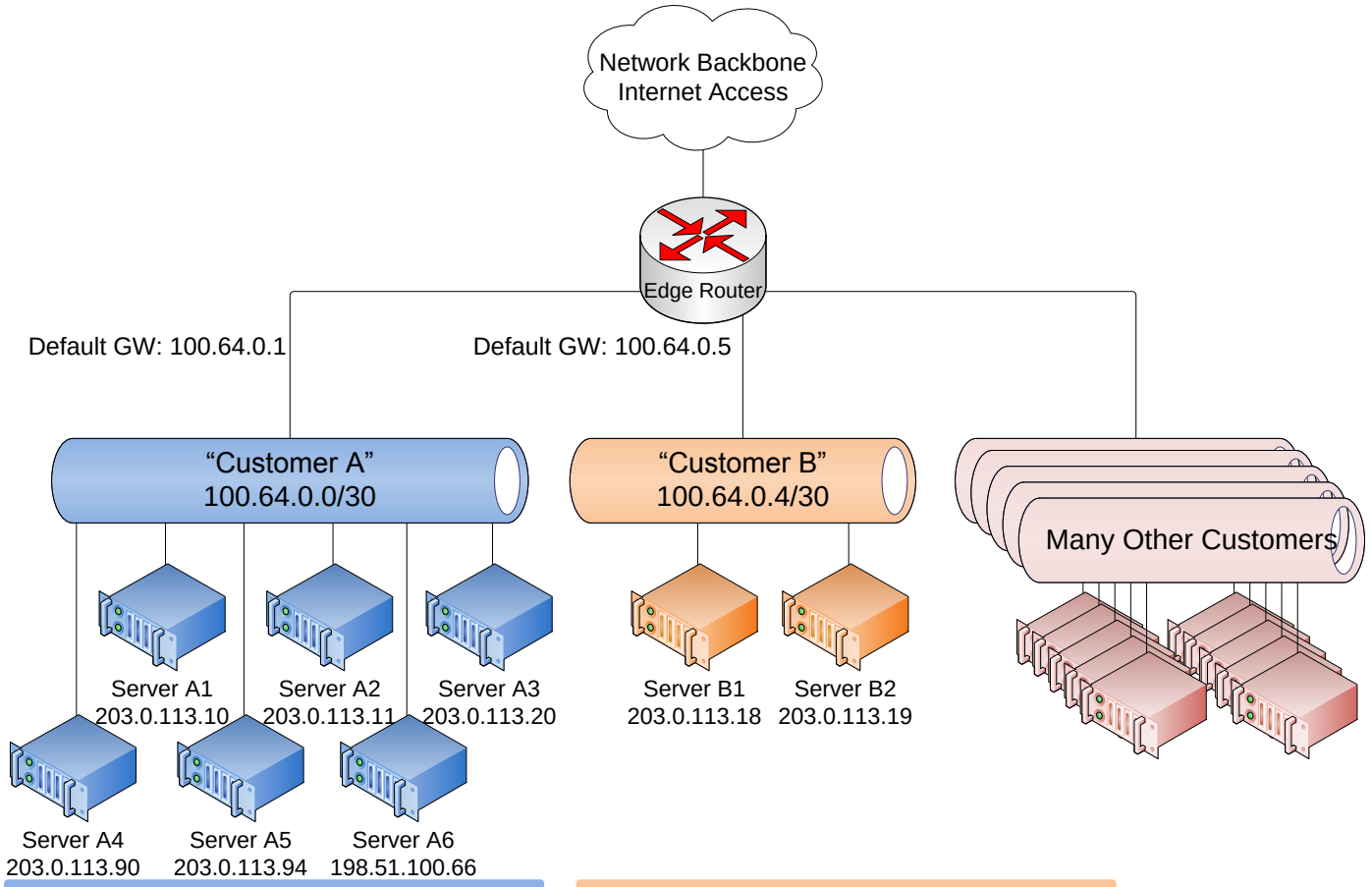
Initial Configuration



No Public IPs "wasted"
ip route 203.0.113.10 255.255.255.255 <CustAInt>
ip route 203.0.113.11 255.255.255.255 <CustAInt>
Servers default GW = 100.64.0.1 static via NIC

No Public IPs "wasted"
ip route 203.0.113.18 255.255.255.255 <CustBInt>
ip route 203.0.113.19 255.255.255.255 <CustBInt>
Servers default GW = 100.64.0.5 static via NIC

Scaling Illustration



No Public IPs "wasted"

ip route 203.0.113.10 255.255.255.255 <CustAInt>
ip route 203.0.113.11 255.255.255.255 <CustAInt>
ip route 203.0.113.20 255.255.255.255 <CustAInt>
ip route 203.0.113.90 255.255.255.255 <CustAInt>
ip route 203.0.113.94 255.255.255.255 <CustAInt>
ip route 198.51.100.66 255.255.255.255 <CustAInt>
Servers default GW = 100.64.0.1 static via NIC

No Public IPs "wasted"

ip route 203.0.113.18 255.255.255.255 <CustBInt>
ip route 203.0.113.19 255.255.255.255 <CustBInt>
Servers default GW = 100.64.0.5 static via NIC

Configuration Steps

1. Configure your routing policy to prevent the link local gateway from being advertised.

2. Update your outbound ACL.

3. Configure your router interface.

4. Configure your server.

Routing policy update

```
!
ip prefix-list BGP-DONT-ANNOUNCE seq 10 permit 100.64.0.0/10 le 32
!
route-map BGP-DONT-ANNOUNCE deny 5
 match ip address prefix-list BGP-DONT-ANNOUNCE
!
```

Update ACL

- Your ACL may need to be more specific if you don't have unicast RPF configured.

- This ACL is very strict, and will not allow you to ping your gateway but ARP will still work. This is so that customers on different interfaces on the same router can't ping each others gateway but it can be modified if it is too strict.

```
ip access-list extended BLOCK_BAD_SUBNETS
 deny   ip any 10.0.0.0 0.255.255.255
 deny   ip any 172.16.0.0 0.15.255.255
 deny   ip any 192.168.0.0 0.0.255.255
 deny   ip any 100.64.0.0 0.63.255.255
 deny ip 10.0.0.0 0.255.255.255 any
 deny ip 172.16.0.0 0.15.255.255 any
 deny ip 192.168.0.0 0.0.255.255 any
 deny ip 100.64.0.0 0.63.255.255 any
 permit ip any any
 !
```

Router Interface (SVI)

```
interface Vlan111
 ip address 100.64.0.1 255.255.255.252
 ip access-group BLOCK_BAD_SUBNETS in
 no ip redirects
 no ip unreachables
 ip verify unicast source reachable-via rx
!

ip route 203.0.113.5 255.255.255.255 Vlan111
```

Router Interface (SVI & HSRP)

```
interface Vlan111
  ip address 100.64.0.2 255.255.255.248
  ip access-group BLOCK_BAD_SUBNETS in
  no ip redirects
  no ip unreachables
  ip verify unicast source reachable-via rx
  standby 10 ip 100.64.0.1
  standby 10 preempt
!
ip route 203.0.113.5 255.255.255.255 Vlan111
```

Server Config

- **Centos Example – Please note that updates need to be made to /etc/sysconfig/network files for changes to be persistent after a reboot.**

ifconfig eth0 203.0.113.5 netmask 255.255.255.255

route add -host 100.64.0.1/32 dev eth0
route add default gw 100.64.0.1

But Wait . . .

Benefits

- **This will allow you to allocate a single /32 to your customer servers – no IP waste**

- **You can still allocate a /29 (or larger) and statically route it to the interface and all of the IP addresses will still be usable.**

Additional Benefits

- Best practices state that when you add an IP address to a router interface you should block this IP from being reachable at your edge so that DoS attacks can't be targeted at your infrastructure.

- This can mean maintaining a list of tens of thousands of edge ACL lines or blackhole routes (1 */31 + 1 */32 per subnet)

- Since you are now using gateway addresses that shouldn't be in the public routing table this should help prevent those types of external attacks. You're on your own for internal attacks.

Caveats

- Even though your router may have a huge FIB you need to validate that it can handle a large number of static routes in its config.
- If there is a lot of server to server communication this can cause sub-optimal forwarding in your environment since traffic may have to hairpin off the router interface. Softlayer has a separate layer3 backend network so it's not a problem for us.
- There are also some techniques to allow servers in different subnets but in the same layer2 domain to talk to each other directly but it requires a lot of upkeep by the server admins.

Your example was IOS based, JunOS is king in my shop.

- So currently there isn't a static route to interface option in JunOS.
- We've spoken with our Juniper reps and there are open Engineering requests
- If you know of a way to do this we'd love to hear it! If not, bug your Juniper rep for ER# 29720!

Other Vendor Support

- As part of our next generation hardware testing, we've also tested some gear from Arista and found an interesting feature.
- Their multi-chassis link aggregation support allows for a unified forwarding plane so there is no need for the active/backup concept in HSRP/VRRP.
- Their layer 3 redundancy uses a feature called "virtual arp".
- You bind a shared gateway ip between both chassis, but they are both active and respond to arp as the gateway (https://eos.aristanetworks.com/2011/05/active-active-router-redundancy-l3-anycast-gateway/)
- This allows you to implement the methods described in this presentation, but also allows for an interesting technology refresh option.

Technology refresh option

- **This functionality allows for an upgrade path from a single chassis or unified control plane hardware configuration, to redundant hardware using separate control planes without affecting customer IP allocation.**

Configuration Example

Single Chassis SVI Configuration:

```
Interface Vlan3
  ip address 203.0.113.1 255.255.255.248
!
```

- **The problem with converting this to HSRP/VRRP is most vendor implementations require that the virtual address be in the same subnet. This means you would have to ask you customer for some IP addresses back so you could bind the subnets on the interface.**

Configuration Example (cont'd)

HSRP with Cisco:

```
(config)#int Vlan3
(config-if)#ip add 100.64.0.2 255.255.255.248
(config-if)#standby 1 ip 203.0.113.1
% Warning: address is not within a subnet on this interface
(config-if)#end

#show standby Vlan3
Vlan3 - Group 1
  State is Init (interface down)
  Virtual IP address is 203.0.113.1 (wrong subnet for this interface)
  Active virtual MAC address is unknown
    Local virtual MAC address is 0000.0c07.ac01 (v1 default)
```

- The state will never change to active with this misconfiguration. This apparently used to work in previous IOS code, but was "fixed" in newer code.

Configuration Example (cont'd)

VRRP with JUNOS:

```
# set interfaces xe-0/0/0 unit 0 family inet address 100.64.0.1/24 vrrp-group 1 virtual-address 203.0.113.1

{master}[edit]
# show | compare
[edit interfaces]
+   xe-0/0/0 {
+       unit 0 {
+           family inet {
+               address 100.64.0.2/29 {
+                   vrrp-group 1 {
+                       virtual-address 203.0.113.1;
+                   }
+               }
+           }
+       }
+   }

{master}[edit]
# commit check
[edit interfaces xe-0/0/0 unit 0 family inet address 100.64.0.2/29]
  'vrrp-group 1'
    virtual address must share same mask with interface ip
error: configuration check-out failed
```

Configuration Example (cont'd)

Virtual ARP with Arista EOS:

Device 1:

```
ip virtual-router mac-address aa:bb:cc:dd:ee:ff
interface Vlan3
   ip address 100.64.0.2/29
   ip virtual-router address 203.0.113.1
!
ip route 203.0.113.0/29 Vlan3
```

Device 2:

```
ip virtual-router mac-address aa:bb:cc:dd:ee:ff
interface Vlan3
   ip address 100.64.0.3/29
   ip virtual-router address 203.0.113.1
!
ip route 203.0.113.0/29 Vlan3
```

Technology Refresh Option

- **We've discussed this with a couple of different vendors and it couldn't hurt to get additional support from the community, so if it this seems like an interesting option please bring it up to your account team!**

Questions? Thoughts?

- **Please contact us:**
  wtemitim@softlayer.com
  chrisp@softlayer.com