Maintain large configurations in a new way
(The BGP walk for the new kid on block)

Peter Lundqvist
peter@arista.com

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

BGP Peering been little the Giants game, earlier…

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

RIB Concept with Peering

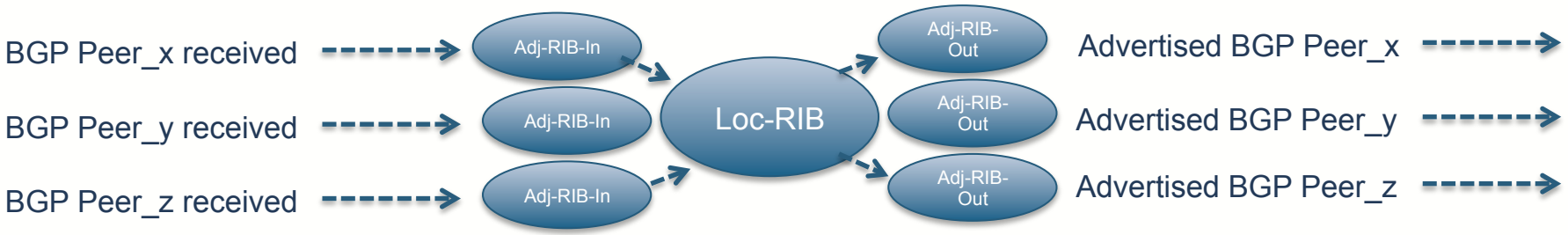
(The basics, I know…)

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

What about this RIB thing

- Routing Information Base (RIB) divided into 3 “tables”
  - Adj-RIB-In
    - The copy of the received state table from peer, Route-map inbound policies affect this
  - Loc-RIB
    - The RIB after policies and path-selection for the node itself
  - Adj-RIB-Out
    - The advertised state table to peer, Route-map outbound policies affect this. This is not copy, just a pointer to the Loc-RIB



Confidential. Copyright © Arista 2016. All rights reserved.

Where do we come from ?
(What have we learn ?)

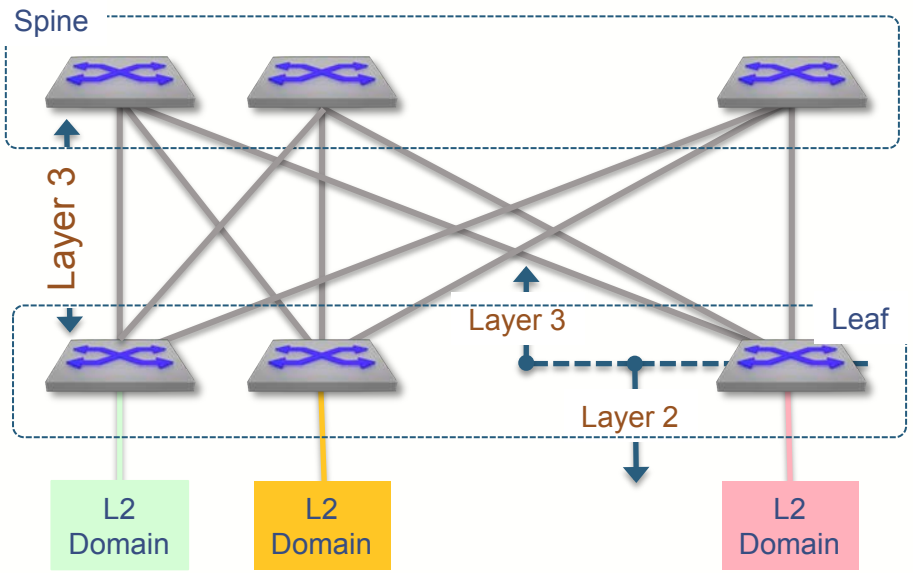
Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Data Center IP Fabric Design Principles #1

- Provide Stability at Scale
  - If Spine more than 2 nodes -> Layer 3 Leaf-Spine architecture
  - **Routing protocol between leaf and spine nodes**
  - Leaf nodes running Default Gateway for the hosts/subnets within the rack
- Benefits
  - **Standard mature control protocols between the leaf and spine**
  - Stability by reducing scope of the Layer 2 broadcast domains
  - Limits the MAC table sizes of the spine for improved scaling

We do Datacenter… trust the fact if L2 fabric would work/scale, we would build it.

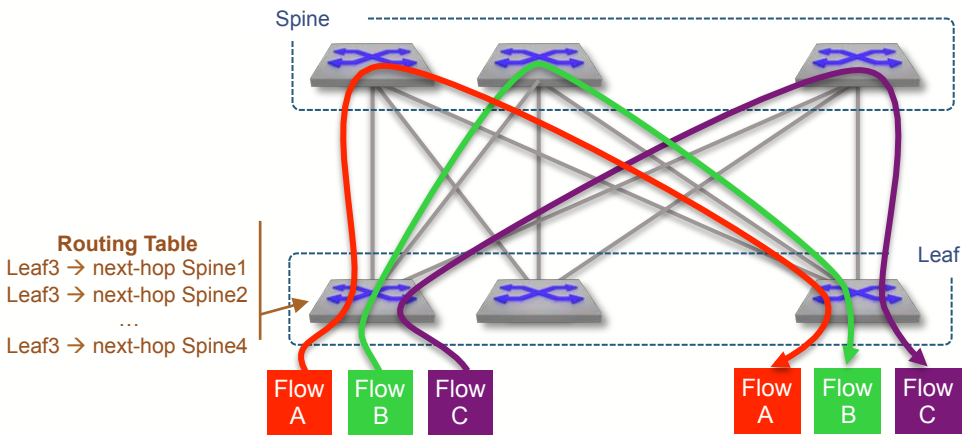


Confidential. Copyright © Arista 2016. All rights reserved.

L3 Equal Cost “Multipathing” Principles #2

- Its not that difficult actually
  - Each leaf node has multiple paths of equal “cost” to each individual spine
  - ECMP load balance flows across the multiple spine node, loose a spine means loose bandwith but NOT connectivity
  - Switch load-balancing algorithm is configurable regards 5-tuple plus possible seed bits to avoid polarization…
  - Advanced features to go beyond with BGP BW Communities and UCMP weights to dictate the flows

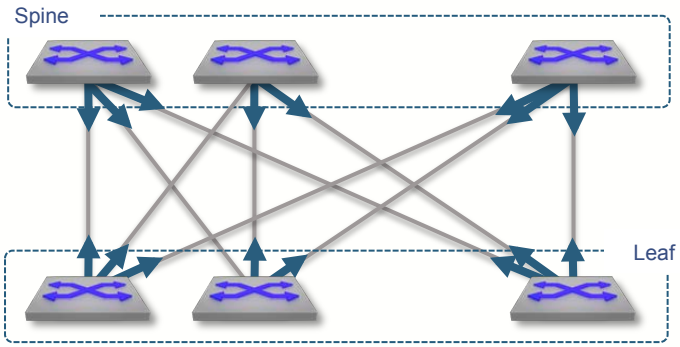
Active paths the best redundancy, since proof of the pudding that it works is daily business



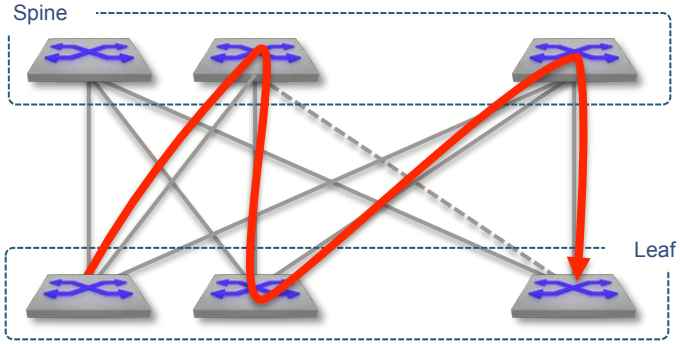
Confidential. Copyright © Arista 2016. All rights reserved.

Routing Protocol Selection Principles #3

- Link State Protocols (OSPF or IS-IS)
  - Absolute no problem if Spine is small
  - Fabric wide topology knowledge on each node (LSDB)
  - Link-state flooding, periodic updates
  - NOTE: Non-deterministic path during transient events (micro loop), leafs can be become a transit node



Link state flooding



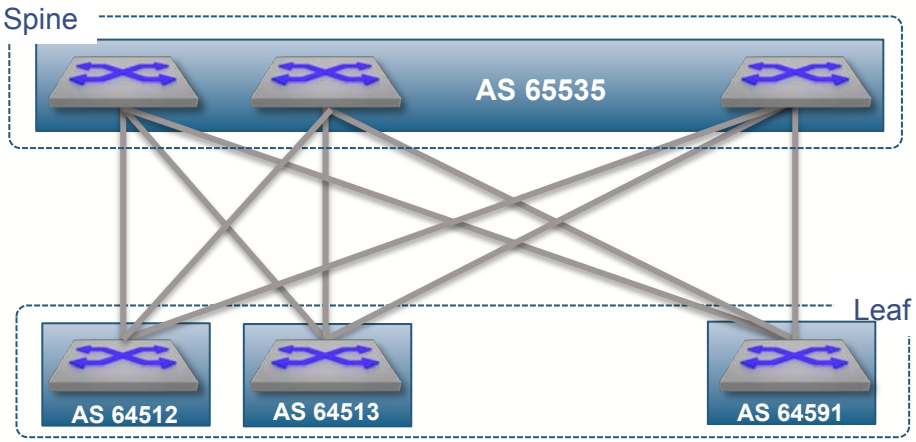
Transient events can result in leaf as a transit node

Confidential. Copyright © Arista 2016. All rights reserved.

Routing Protocol Selection…

- eBGP as the Routing protocol for the IP fabric
  - Control of routing advertisements to the leaf, via policies. Adj-RIB-Out vs In, simple as that
  - BGP Updates yes, but no flooding
  - No periodic Database update/sync
  - Usage of private ASN range
    - 2 Byte ASN 64512-65535
    - 4 Byte ASN 4200000000-4294967294
  - Leaf and Spine peer over IP interfaces on each point-to-point link with direct peering on each link

Come on, we all know it… BGP works !



Confidential. Copyright © Arista 2016. All rights reserved.

Lesson from L3 ECMP DC
(BGP done little different)

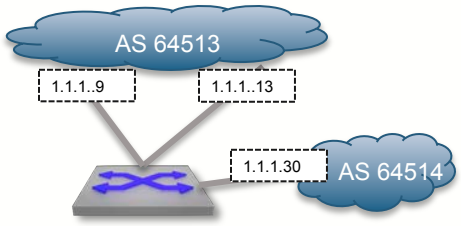
Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Multipath

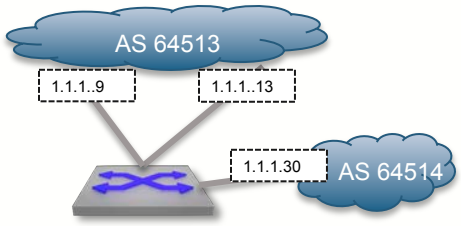
- EOS default multipath relaxed, which means that multiple entries installed in the RIB, if Multipath “maximum-paths x” been configured
- This is useful in DC and L3 ECMP design, however not common with Peering and BGP edge:
  - Peering normal design redundant on multiple BGP edge borders
  - Default behavior other OS only install multiple entries for routes learn from the same peer AS

EOS Multipath Default



RIB
Route 1.1.2/24
AS 64514 Nexthop 1.1.1.30
AS 64513 Nexthop 1.1.1.9
AS 64513 Nexthop 1.1.1.13

Other OS Multipath Default



RIB
Route 1.1.2/24
AS 64513 Nexthop 1.1.1.9
AS 64513 Nexthop 1.1.1.13

Confidential. Copyright © Arista 2016. All rights reserved.

Multipath…

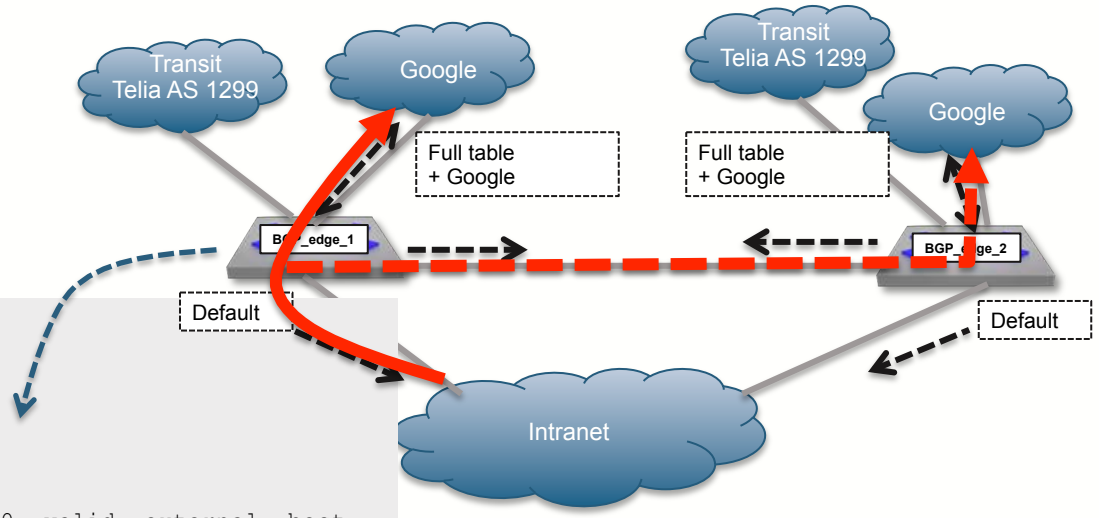
- EOS default behavior with Multipath, ASN (64513&64514) ignored and all 3 routes installed

```
veos4(config)#sh ip bgp 1.1.2.0/24 det
BGP routing table information for VRF default
Router identifier 1.1.1.4, local AS number 64512
Route status: [a.b.c.d] - Route is  queued for advertisement to peer.
BGP routing table entry for 1.1.2.0/24
 Paths: 3 available
  64513
    1.1.1.9 from 1.1.1.9 (1.1.1.1)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external, ECMP head, ECMP, best, ECMP contributor
      Community: 64513:11
      Rx SAFI: Unicast
  64513
    1.1.1.13 from 1.1.1.13 (1.1.1.2)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external, ECMP, ECMP contributor
      Community: 64513:10
      Rx SAFI: Unicast
      Not best: Router ID tie-break configured
  64514
    1.1.1.30 from 1.1.1.30 (1.1.1.6)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external, ECMP, ECMP contributor
      Rx SAFI: Unicast
      Not best: Router ID tie-break configured
 Advertised to 2 peers:
  peer-group 64513:
    1.1.1.13
  peer-group edge:
    1.1.1.30
```

Confidential. Copyright © Arista 2016. All rights reserved.

Special convergence scenarios

- Best Path selection picks eBGP but installs iBGP in the RIB (and FIB)




```
veos1(config)#sh ip bgp x.x.x.x/xx
BGP routing table information for VRF default
Router identifier 1.1.1.1, local AS number 64513
BGP routing table entry for x.x.x.x/xx
 Paths: 2 available
  64512
    1.1.1.10 from 1.1.1.10 (1.1.1.4)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external, best
 64512
    1.1.1.14 from 1.1.1.18 (1.1.1.2)
      Origin IGP, metric 0, localpref 100, weight 0, valid, internal, backup

veos1(config)#sh ip ro 1.1.2.0/24
(…)
Gateway of last resort:
 B E     x.x.x.x/xx [109/0] via 1.1.1.10, Ethernet1
                            via 1.1.1.22, Ethernet4, backup
```

```
router bgp 64513
   router-id 1.1.1.1
   bgp additional-paths install
(…)
```

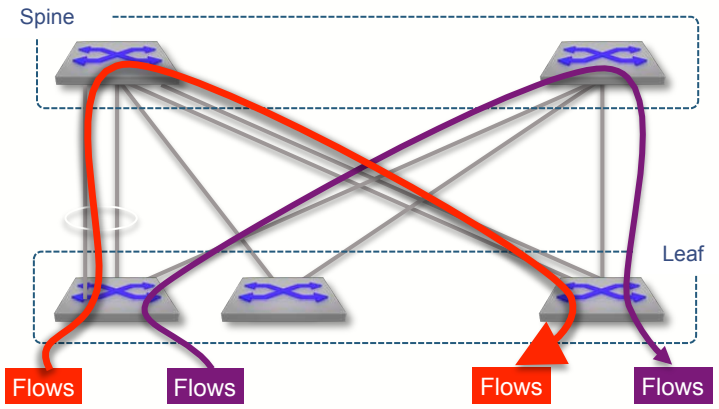
Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Handle Speed difference

- Fix suddenly BW variations example link issues
- Offload or move traffic to/from certian NEXT_HOP (BGP Edge or specific Leaf&Spine)

```
veos3b(config)#sh ip bgp 1.1.1.6/32
BGP routing table information for VRF default
Router identifier 1.1.1.7, local AS number 64515
BGP routing table entry for 1.1.1.6/32
 Paths: 2 available
  64512 64514
    1.1.1.42 from 1.1.1.42 (1.1.1.5)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external,
      ECMP head, ECMP, UCMP, best, ECMP contributor
      Extended Community: Link-Bandwidth-AS:64512:125000000.000000(Bps)
      Rx SAFI: Unicast
  64512 64514
    1.1.1.40 from 1.1.1.40 (1.1.1.4)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external,
      ECMP, UCMP, ECMP contributor
      Extended Community: Link-Bandwidth-AS:64512:1250000000.000000(Bps)
      Rx SAFI: Unicast
(…)

veos3b(config)#sh ip route 1.1.1.6/32
(…)
 B E    1.1.1.6/32 [200/0] via 1.1.1.40, Ethernet1, weight 10/11
                           via 1.1.1.42, Ethernet2, weight 1/11
```



Routing Table
Leaf3 → next-hop Spine1 10/11 flows
Leaf3 → next-hop Spine2 1/11 flows
…

Confidential. Copyright © Arista 2016. All rights reserved.

Merchant Silicon #1

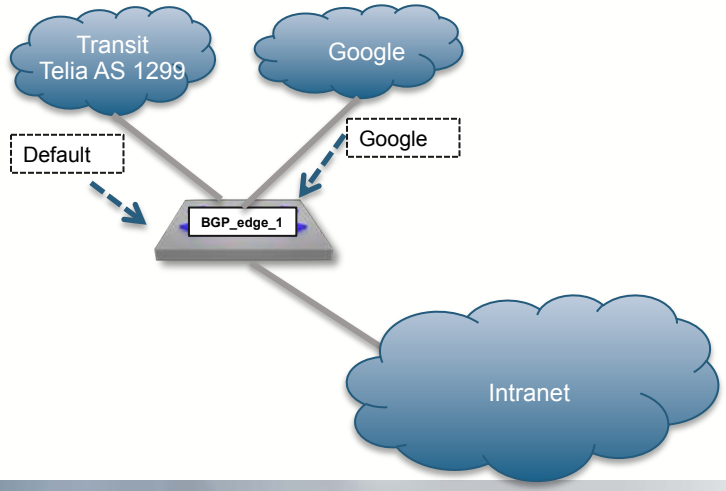
(Our first step with Arad)

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Selective Route Download (SRD)

- If you are the last stop station, no need to have full stations list, this is the reality for most Datacenters
- Of course there can be a gain have full feed in the RIB for visability, statistics and troubleshooting, but why have full table in the FIB?
- For the most part a default to transit, and more specific to private peering

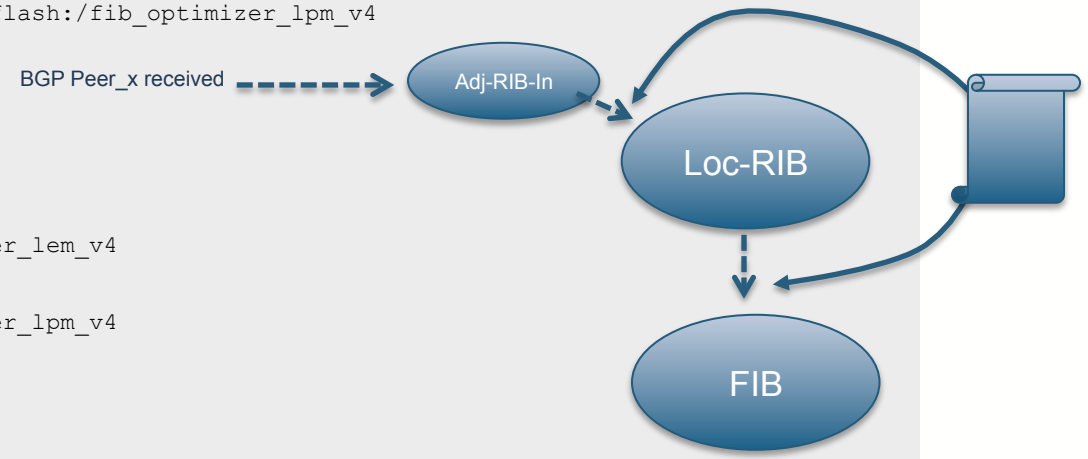


Confidential. Copyright © Arista 2016. All rights reserved.

SRD…

- The prefix-list can been stored on the flash as a simple file, or write into the CLI…

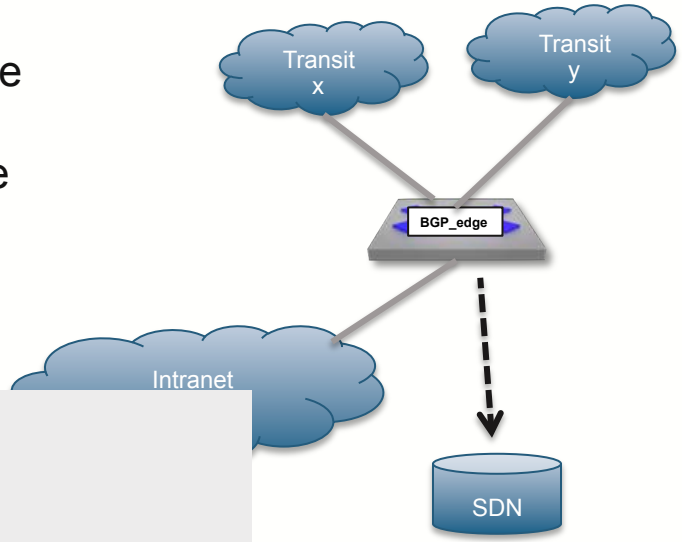
```
(…)
ip as-path access-list ASN_DC permit ^43650$ any
ip as-path access-list ASN_TRANSIT permit ^1299$ any
ip prefix-list FILTER_LEQ_24 seq 10 permit 0.0.0.0/0 le 25
ip prefix-list fib_optimizer_lem_v4 source flash:/fib_optimizer_lem_v4
ip prefix-list fib_optimizer_lpm_v4 source flash:/fib_optimizer_lpm_v4
!
route-map SRD permit 10
   match as-path ASN_DC
!
route-map SRD permit 20
   match as-path ASN_TRANSIT
!
route-map SRD permit 30
   match ip address prefix-list fib_optimizer_lem_v4
!
route-map SRD permit 40
   match ip address prefix-list fib_optimizer_lpm_v4
!
router bgp 8403
   no bgp enforce-first-as
   bgp route install-map SRD
(…)
```



Confidential. Copyright © Arista 2016. All rights reserved.

BGP ADD-PATH

- With ADD-PATH all routes before path-selection can be relay to example a route-server.
- Route-server would not be able to calculate alternative paths unless ADD-PATH been used
- Then policy could be changed to move certain traffic towards alternative path




```
veos2.02:54:52(config)#sh ip bgp 1.1.1.7/32 detail
BGP routing table information for VRF default
Router identifier 1.1.1.2, local AS number 64513
Route status: [a.b.c.d] - Route is  queued for advertisement to peer.
BGP routing table entry for 1.1.1.7/32
 Paths: 2 available
  64512 64515
    1.1.1.10 from 1.1.1.17 (1.1.1.1)
      Origin IGP, metric 0, localpref 100, weight 0, received 03:39:30 ago, valid, internal, best
      Rx path id: 0x1
      Rx SAFI: Unicast
  64512 64512 64515
    1.1.1.22 from 1.1.1.17 (1.1.1.1)
      Origin INCOMPLETE, metric 101, localpref 100, weight 0, received 00:03:41 ago, valid, internal
      Rx path id: 0x2
      Rx SAFI: Unicast
      Not best: Another route from the same AS is a better BGP route
(…)
```

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Merchant Silicon #2

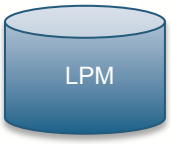
(Begin to Scale using Jericho)

Confidential. Copyright © Arista 2016. All rights reserved.

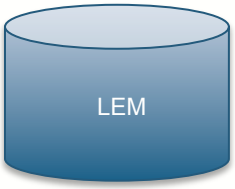
ARISTA

FIB Optimization

- Arad/Jericho based hardware (7500/7280) FIB are divided into two tables (LPM&LEM).
- LPM design to carry variable prefix-length, LEM fixed size example MAC
- To scale the FIB numbers for both IPv4 and IPv6, the solution is to move prefixes from LPM to LEM, example the most commonly prefixes…




```
petrus:Downloads plundqvi$ more route.txt | grep -v /22
B E    2.16.40.0/24 [200/0] via 80.249.209.103, Port-Channel1 AMSIX
 B E    2.22.230.0/24 [200/0] via 80.249.208.168, Port-Channel1 AMSIX
 B E    2.84.64.0/19 [200/2000] via 80.249.208.179, Port-Channel1 AMSIX
 B E    2.88.192.0/19 [200/0] via 80.249.212.145, Port-Channel1 AMSIX
 B E    2.88.240.0/20 [200/0] via 80.249.212.145, Port-Channel1 AMSIX
 B E    2.100.0.0/14 [200/0] via 80.249.209.34, Port-Channel1 AMSIX
 B E    2.156.0.0/16 [200/0] via 80.249.211.79, Port-Channel1 AMSIX
 B E    2.157.0.0/16 [200/0] via 80.249.211.79, Port-Channel1 AMSIX
 B E    2.176.0.0/16 [200/0] via 80.249.208.30, Port-Channel1 AMSIX
 B E    2.177.0.0/16 [200/0] via 80.249.208.30, Port-Channel1 AMSIX
 B E    2.178.0.0/16 [200/0] via 80.249.208.30, Port-Channel1 AMSIX
 B E    2.190.0.0/16 [200/0] via 80.249.208.30, Port-Channel1 AMSIX
 B E    2.191.0.0/16 [200/0] via 80.249.208.30, Port-Channel1 AMSIX
 B E    2.176.0.0/12 [200/0] via 80.249.208.30, Port-Channel1 AMSIX
 B E    5.1.32.0/21 [200/0] via 80.249.211.91, Port-Channel1 AMSIX
(...)
```



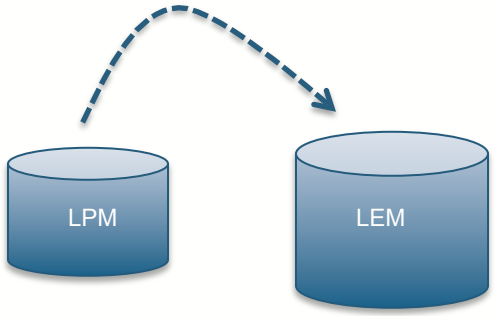

```
petrus:Downloads plundqvi$ more route.txt | grep /22
 B E    2.20.24.0/22 [200/0] via 80.249.208.168, Port-Channel1 AMSIX
 B E    5.32.132.0/22 [200/0] via 80.249.212.43, Port-Channel1 AMSIX
 B E    5.61.212.0/22 [200/0] via 80.249.212.81, Port-Channel1 AMSIX
 B E    5.101.104.0/22 [200/10] via 80.249.211.98, Port-Channel1 AMSIX
 B E    5.144.188.0/22 [200/0] via 80.249.211.74, Port-Channel1 AMSIX
 B E    5.154.56.0/22 [200/0] via 80.249.211.91, Port-Channel1 AMSIX
 B E    5.154.88.0/22 [200/0] via 80.249.212.81, Port-Channel1 AMSIX
 B E    23.215.108.0/22 [200/0] via 80.249.208.94, Port-Channel1 AMSIX
 B E    23.239.28.0/22 [200/0] via 80.249.210.82, Port-Channel1 AMSIX
 B E    31.22.64.0/22 [200/101] via 80.249.208.42, Port-Channel1 AMSIX
 B E    31.24.48.0/22 [200/0] via 80.249.208.197, Port-Channel1 AMSIX
 B E    31.222.96.0/22 [200/0] via 80.249.212.81, Port-Channel1 AMSIX
 B E    31.222.104.0/22 [200/0] via 80.249.212.81, Port-Channel1 AMSIX
 B E    31.222.112.0/22 [200/0] via 80.249.212.81, Port-Channel1 AMSIX
 B E    31.222.116.0/22 [200/0] via 80.249.212.81, Port-Channel1 AMSIX
 B E    37.25.40.0/22 [200/0] via 80.249.211.91, Port-Channel1 AMSIX
 B E    37.131.172.0/22 [200/0] via 80.249.211.20, Port-Channel1 AMSIX
```

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

FIB Optimization IPv4

- What prefixes to move to LEM ?
- In this IPv4 RIB, most learn routes /24 => LEM



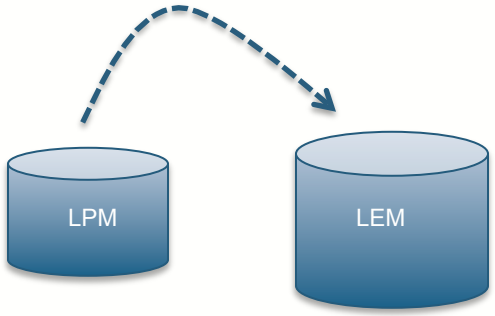

```
bgp-edge2>sh ip route sum
   Route Source                                    Number Of Routes
------------------------------------ ------------------------
   connected                                                   14
   static (persistent)                                          0
   static (non-persistent)                                      0
   VXLAN Control Service                                        0
   static nexthop-group                                         0
   ospf                                                         0
     Intra-area: 0 Inter-area: 0 External-1: 0 External-2: 0
     NSSA External-1: 0 NSSA External-2: 0
   ospfv3                                                       0
   bgp                                                     349512
     External: 1781 Internal: 347731
   isis                                                         0
     Level-1: 0 Level-2: 0
   rip                                                          0
   internal                                                    27
   attached                                                    61
   aggregate                                                    0

   Total Routes                                            349614

Number of routes per mask-length:
   /8: 9          /9: 5          /10: 20        /11: 62        /12: 153
   /13: 271       /14: 564       /15: 1026      /16: 6962      /17: 4524
   /18: 7277      /19: 15269     /20: 22422     /21: 23962     /22: 40007
   /23: 32956     /24: 193252    /25: 195       /26: 100       /27: 57
   /28: 76        /29: 75        /30: 65        /31: 19        /32: 225
```

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

FIB Optimization IPv6

- What prefixes to move to LEM ?
- In this IPv6 RIB, most learn routes /48 => LEM




```
bgp-edge2>sh ipv6 route sum
   Route Source                Number Of Routes
------------------------------ ----------------
   connected                                13
   static (persistent)                       0
   static (non-persistent)                   0
   static nexthop-group                      0
   ospf                                      0
   bgp                                   24452
   isis                                      0
     Level-1: 0 Level-2: 0
   internal                                 15
   attached                                103
   aggregate                                 0

   Total Routes                          24583

Number of routes per mask-length:
   /10: 1        /16: 1        /20: 7        /22: 4        /23: 4
   /24: 14       /25: 5        /26: 14       /27: 13       /28: 61
   /29: 1090     /30: 82       /31: 77       /32: 6571     /33: 304
   /34: 208      /35: 211      /36: 984      /37: 89       /38: 496
   /39: 124      /40: 1127     /41: 60       /42: 165      /43: 15
   /44: 660      /45: 99       /46: 351      /47: 196      /48: 11026
   /49: 3        /50: 2        /52: 6        /55: 1        /56: 30
   /60: 1        /62: 1        /63: 1        /64: 283      /65: 1
   /96: 1        /118: 1       /120: 1       /124: 4       /125: 2
   /126: 44      /127: 16      /128: 23
```

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

FIB Optimization…

- IPv4 /24 about to be moved from LPM => LEM

```
bgp-edge(config)#ip hardware fib optimize prefix-length 24 ?
  0     Prefix length 0
  1     Prefix length 1
  10    Prefix length 10
  11    Prefix length 11
  12    Prefix length 12
  13    Prefix length 13
  14    Prefix length 14
  15    Prefix length 15
  16    Prefix length 16
  17    Prefix length 17
  18    Prefix length 18
  19    Prefix length 19
  2     Prefix length 2
  20    Prefix length 20
  21    Prefix length 21
  22    Prefix length 22
  23    Prefix length 23
  25    Prefix length 25
  26    Prefix length 26
  27    Prefix length 27
  28    Prefix length 28
  29    Prefix length 29
  3     Prefix length 3
  30    Prefix length 30
  31    Prefix length 31
  32    Prefix length 32
  4     Prefix length 4
  5     Prefix length 5
  6     Prefix length 6
  7     Prefix length 7
  8     Prefix length 8
  9     Prefix length 9
  <cr>
```

Confidential. Copyright © Arista 2016. All rights reserved.

Flexroute FIB Optimization

- Going from manual tuning of tables to dynamic way
- Flexroute enable automatically tuning and distribution the installation of prefixes to tables
- This allows roughly support for 1M+ prefixes in the FIB
- Result in this case
  - IPv4 /20 and /24 automatically moved to LEM from LPM.
  - IPv6 /48 also automatically moved to LEM

```
bgp-edge2(config)#ip hardware fib optimize prefixes profile internet
bgp-edge2(config)#ipv6 hardware fib optimize prefixes profile internet


bgp-edge2(config)#show ip bgp sum
BGP summary information for VRF default
Router identifier 109.105.96.55, local AS number 2603
Neighbor Status Codes: m - Under maintenance
 Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State  PfxRcd PfxAcc
 109.105.97.24    4  2603          286314        96    0    0 00:19:38 Estab  348325 348325
(…)


bgp-edge2#show platform sand l3 summary
Ipv4:
  Routes     : 349919 backlog: 0 unprogrammed: 0
  Adjacencies: 380    backlog: 0 unprogrammed: 0
Ipv6:
  Routes     : 24533 backlog: 0 unprogrammed: 0
  Adjacencies: 380    backlog: 0 unprogrammed: 0
Mpls:
  Routes     : 0     backlog: 0 unprogrammed: 0
  Adjacencies: 0     backlog: 0 unprogrammed: 0
Lem:
  IPv4 Host in Lem          : disabled
  IPv4 Prefix-lengths in Lem: 20 24
  IPv4 Routes per prefix-len: [/20]=46475 [/24]=250987
  IPv6 Host in Lem          : disabled
  IPv6 Prefix-lengths in Lem: 48
  IPv6 Routes per prefix-len: [/48]=10988
(…)
```

Confidential. Copyright © Arista 2016. All rights reserved.

Going from GateD…

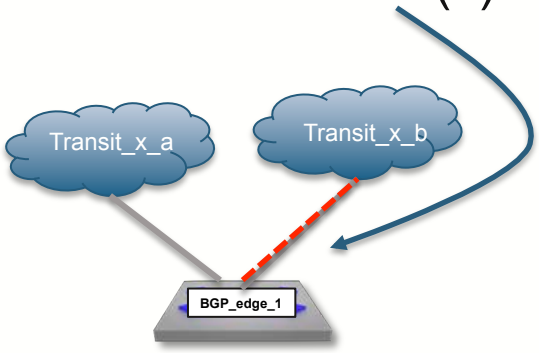
(Say what you do, do what you say)

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

TCP statistics related to BGP session

- BGP session performance all about TCP mechanism
- TCP statistics related to the peer as important to monitor as RIB table(s)




```
bgp-edge(config)#sh ip bgp neighbors 3.3.3.3
(...)
TCP Socket Information:
TCP state is ESTABLISHED
  Recv-Q: 0/32768
  Send-Q: 0/32768
  Outgoing Maximum Segment Size (MSS): 8948
  Total Number of TCP retransmissions: 0
  Options:
    Timestamps enabled: yes
    Selective Acknowledgments enabled: yes
    Window Scale enabled: yes
    Explicit Congestion Notification (ECN) enabled: no
  Socket Statistics:
    Window Scale (wscale): 7,7
    Retransmission Timeout (rto): 212.0ms
    Round-trip Time (rtt/rtvar): 15.0ms/13.0ms
    Delayed Ack Timeout (ato): 40.0ms
    Congestion Window (cwnd): 10
    TCP Throughput: 47.72 Mbps
    Advertised Recv Window (rcv_space): 17896
```

Confidential. Copyright © Arista 2016. All rights reserved.

BGP best path selection

- Generic attributes vs vendor specific attributes

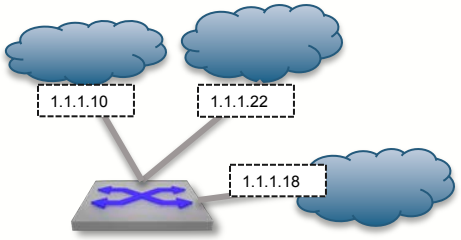
Path-Selection	BGPv4 RFC 4271	Arista	Cisco	Juniper
Non-standard		Highest Weight	Highest Weight	Route Preference
1	Highest LOCAL_PREF	Highest LOCAL_PREF	Highest LOCAL_PREF	Highest LOCAL_PREF
Non-standard			Interior > Exterior > Exterior via Interior	
2	Shortest AS_PATH	Shortest AS_PATH	Shortest AS_PATH	Shortest AS_PATH
3	Lowest ORIGIN: IGP >EGP > Incomplete	Lowest ORIGIN: IGP >EGP > Incomplete	Lowest ORIGIN: IGP >EGP > Incomplete	Lowest ORIGIN: IGP >EGP > Incomplete
4	Lowest MED	Lowest MED	Lowest MED	Lowest MED
Non-standard				Interior > Exterior > Exterior via Interior
5	eBGP>iBGP	eBGP>iBGP	eBGP>iBGP	eBGP>iBGP
6	Lowest IGP metric to NEXT_HOP	Lowest IGP metric to NEXT_HOP	Lowest IGP metric to NEXT_HOP	Lowest IGP metric to NEXT_HOP
Non-standard	Multipath	Multipath	Multipath	Multipath
Non-standard	Oldest active route	Oldest active route	Oldest active route	Oldest active route
7	Lowest ORIGINATOR_ID, or Router-ID	Lowest ORIGINATOR_ID, or Router-ID	Lowest ORIGINATOR_ID, or Router-ID	Lowest ORIGINATOR_ID, or Router-ID
8 (RFC4456 standard)	Shortest CLUSTER_LIST	Shortest CLUSTER_LIST	Shortest CLUSTER_LIST	Shortest CLUSTER_LIST
9	Lowest peering address	Lowest BGP peering address	Lowest BGP peering address	Lowest BGP peering address
Etc…				

Confidential. Copyright © Arista 2016. All rights reserved.

Trace best path selection

- BGP path selection visible

```
veos1(config)#  sh ip bgp 0.0.0.0/0 det
BGP routing table information for VRF default
Router identifier 1.1.1.1, local AS number 64513
Route status: [a.b.c.d] - Route is  queued for advertisement to peer.
BGP routing table entry for 0.0.0.0/0
 Paths: 3 available
  64512 64514
    1.1.1.10 from 1.1.1.10 (1.1.1.4)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external, best
      Rx SAFI: Unicast
  64512 64514
    1.1.1.22 from 1.1.1.22 (1.1.1.5)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external
      Rx SAFI: Unicast
      Not best: Router ID
  64512 64514
    1.1.1.14 from 1.1.1.18 (1.1.1.2)
      Origin IGP, metric 0, localpref 100, weight 0, valid, internal
      Rx SAFI: Unicast
      Not best: eBGP path preferred
 Advertised to 1 peers:
  peer-group mlag:
    1.1.1.18
```



Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Mode of regular-expressions

(That thing with religions…)

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

AS_PATH

- Default EOS mode run with Gated DFA POSIX regular-expression (asn)
- Cisco by default using String mode
- EOS can be run with either asn | string mode

```
bgp-edge(config)#ip as-path regex-mode ?
  asn     Match AS path as AS numbers
  string  Match AS path as a string

bgp-edge(config)#sh ip as-path access-list
ip as-path regex-mode asn
ip as-path access-list 2222 permit _2222$ any
ip as-path access-list two permit ^2222_ any
```

Confidential. Copyright © Arista 2016. All rights reserved.

AS_PATH, asn or string

- With asn mode each entry return complete AS number
- Below returns explicit AS 645 => no hits

```
bgp-edge(config)#ip as-path regex-mode asn
! Some as-path ACLs may get disabled due to invalid regexes under new regex mode

bgp-edge(config)#sh ip bgp regexp 645
BGP routing table information for VRF default
Router identifier 1.1.1.6, local AS number 1111
Route status codes: s - suppressed, * -  valid, > - active, # - not installed, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP,  b - backup, L = labeled-unicast
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST -  Cluster List, LL Nexthop - Link Local Nexthop

        Network               Next Hop              Metric  LocPref Weight  Path
```

Confidential. Copyright © Arista 2016. All rights reserved.

AS_PATH, asn or string…

- string mode each entry just characters
- below means any AS with characters “645” = hits

```
bgp-edge(config)#ip as-path regex-mode string
! Some as-path ACLs may get disabled due to invalid regexes under new regex mode

bgp-edge(config)#sh ip bgp regexp 645
BGP routing table information for VRF default
Router identifier 1.1.1.6, local AS number 1111
Route status codes: s - suppressed, * -  valid, > - active, # - not installed, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP,  b - backup, L = labeled-unicast
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST -  Cluster List, LL Nexthop - Link Local Nexthop

        Network               Next Hop              Metric  LocPref Weight  Path
 * >     1.0.0.0/8           -                   0       0       -       64512 i
 * >     1.1.1.1/32          1.1.1.29            0       100     0       64512 64513 64513 i
 * >     1.1.1.4/32          1.1.1.29            0       100     0       64512 i
 * >     1.1.1.7/32          1.1.1.29            0       100     0       64512 64515 I
(…)
```

Confidential. Copyright © Arista 2016. All rights reserved.

AS_PATH, asn or string…

- A classic POSIX expression is to pick out odd or even numbers, so if you want AS_PATH that start with even number its ^[02468] or you want all AS that ends with an odd number [13579]$, then string mode is the way

```
bgp-edge(config)#sh ip bgp regexp ^[02468]
BGP routing table information for VRF default
Router identifier 1.1.1.6, local AS number 1111
Route status codes: s - suppressed, * -  valid, > - active, # - not installed, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP,  b - backup, L = labeled-unicast
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST -  Cluster List, LL Nexthop - Link Local Nexthop

         Network                Next Hop              Metric  LocPref Weight  Path
 * >      1.0.0.0/8           -                  0       0       -       64512 i
 * >      1.1.1.1/32          1.1.1.29           0       100     0       64512 64513 64513 i
 * >      1.1.1.4/32          1.1.1.29           0       100     0       64512 i
 * >      1.1.1.7/32          1.1.1.29           0       100     0       64512 64515 i
 * >      1.1.1.8/30          1.1.1.29           0       100     0       64512 i
 * >      1.1.1.12/30         1.1.1.29           0       100     0       64512 I
(…)
 * >      2.0.0.0/8           3.3.3.3            0       100     0       2222 ?
 * >      2.2.0.0/24          3.3.3.3            0       100     0       2222 i
```

Confidential. Copyright © Arista 2016. All rights reserved.

AS_PATH, asn or string…

- But it doesn't have to be that complicated of course
- In most cases characters will be used, then modes are relative careless since it will generate same result

```
bgp-edge#sh ip bgp regexp _64513_   <- all routes with AS 64513 anywhere in the AS_PATH
bgp-edge#sh ip bgp regexp ^64513_   <- all routes with AS 64513 as last entry in the AS_PATH
bgp-edge#sh ip bgp regexp _64513$   <- all routes with AS 64513 as the origin in the AS_PATH

bgp-edge(config)#sh ip bgp regexp (64513)+$ <- single or multiple entry of AS 64513
BGP routing table information for VRF default
Router identifier 1.1.1.6, local AS number 1111
Route status codes: s - suppressed, * -  valid, > - active, # - not installed, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP,  b - backup, L = labeled-unicast
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST -  Cluster List, LL Nexthop - Link Local Nexthop

         Network                Next Hop              Metric  LocPref Weight  Path
 * >      1.1.1.1/32          1.1.1.29         0       100     0       64512 64513 64513 i
 * >      1.1.1.100/32        1.1.1.29         0       100     0       64512 64513 i
 * >      1.1.1.111/32        1.1.1.29         0       100     0       64512 64513 i
 * >      1.1.2.0/24          1.1.1.29         0       100     0       64512 64513 i
 * >      1.1.3.0/24          1.1.1.29         0       100     0       64512 64513 i
```

Confidential. Copyright © Arista 2016. All rights reserved.

Protect your RIB

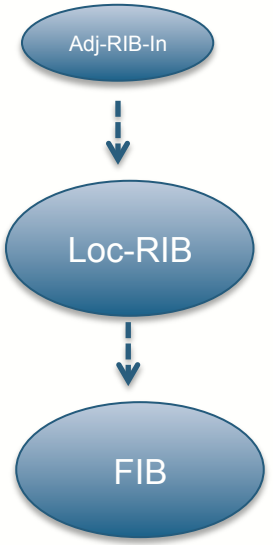
(Ohhh… there can be mistakes ?)

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Maximum-routes

- So, timers are nice but why not separate pre/post Adj-RIB-In
- Maximum-routes
  - all routes included rejected routes due to Policy, include possible temporary mistakes
- Maximum-accepted-routes
  - no count on routes rejected due to policy, this reflect true increase of routes




```
router bgp 1111
(…)
   neighbor ebgp idle-restart-timer 60
   neighbor ebgp maximum-routes 1000
   neighbor ebgp maximum-accepted-routes 900
(…)
```

Confidential. Copyright © Arista 2016. All rights reserved.

Maximum-routes…

- Session is brought down to idle state for timer defined (or idle until manual reset the BGP session state with “clear ip bgp nei x.x.x.x”)

```
bgp-edge(config-router-bgp)#sh ip bgp nei 3.3.3.3
BGP neighbor is 3.3.3.3, remote AS 2222, external link
  BGP version 4, remote router ID 2.2.2.2, VRF default
  Inherits configuration from and member of peer-group ebgp
  Negotiated BGP version 4
  Last read never, last write never
  Hold time is 180, keepalive interval is 60 seconds
  Configured hold time is 180, keepalive interval is 60 seconds
  Connect timer is inactive
  Configured idle-restart time is 61 seconds
  Idle-restart timer is active, time left: 00:00:34
  BGP state is Idle, Peer exceeded max-accepted-routes
  Peering failure hint: Peer exceeded max-accepted-routes
  Number of transitions to established: 10
  Last state was Established
  Last event was RecvUpdate
  Last sent notification:Cease/maximum number of prefixes reached,
  Last time 00:00:03, First time 00:09:12, Repeats 64
(…)
```

Confidential. Copyright © Arista 2016. All rights reserved.

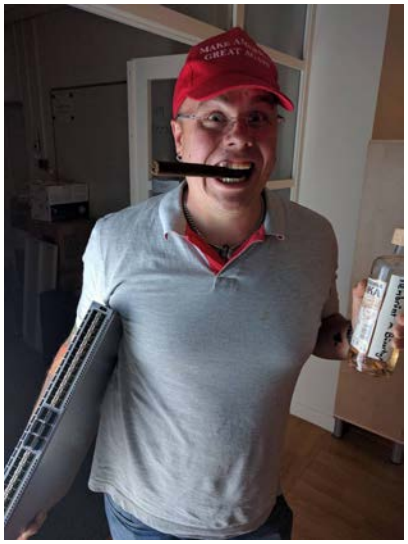
Scale policies
(So it begins…)

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Route-Maps simple don't scale…

- The game is repeat, repeat, and repeat…
- You have to be able to reuse
- Answer is sub-route-map
- The walking "pest" below ☺



```
(….)
route-map akamai-in permit 2
!
route-map akamai-in deny 5
   sub-route-map sanity
!
route-map akamai-in permit 10
   sub-route-map peering-in
   continue
!
route-map akamai-in permit 20
   match as-path radb-as-akamai
!
route-map akamai-in deny 50
ut
!
route-map amazon-in deny 5
   sub-route-map sanity
!
route-map amazon-in permit 10
   sub-route-map peering-in
   continue
!
route-map amazon-in permit 20
   match as-path radb-as-amazon
!
route-map amazon-in deny 50
(...)
```

Confidential. Copyright © Arista 2016. All rights reserved.

Sub-Route-Maps

+ The gain is that it avoids repeating

- Must understand the result order from called policy, which can be missleading (permit can actually be a deny)

```
(…)
route-map as2222 deny 5   <-- Here is actually the deny of the routes
   sub-route-map global
(…)
route-map global permit 5 <--- Note the permit
   match ip address prefix-list 1918
(…)
ip prefix-list 1918 seq 5 permit 10.0.0.0/8 le 32
ip prefix-list 1918 seq 10 permit 172.16.0.0/12 le 32
ip prefix-list 1918 seq 15 permit 192.168.0.0/16 le 32
ip prefix-list 1918 seq 20 permit 100.64.0.0/10 le 32
(...)
```

Confidential. Copyright © Arista 2016. All rights reserved.

Sub-Route-Maps example

Router receives good routes but also crap routes like martians, RFC1918, /32 etc…

(…)

	Network	Next Hop	Metric	LocPref	Weight	Path
* >	1.0.0.0/8	-	0	0	-	64512 i
* >	1.1.1.4/32	1.1.1.29	0	100	0	64512 i
* >	1.1.1.6/32	-	0	0	-	i
* >	1.1.1.8/30	1.1.1.29	0	100	0	64512 i
* >	1.1.1.12/30	1.1.1.29	0	100	0	64512 i
* >	1.1.1.28/30	-	1	0	-	i
*	1.1.1.28/30	1.1.1.29	0	100	0	64512 i
* >	1.1.1.32/30	-	1	0	-	i
* >	1.1.1.40/31	1.1.1.29	0	100	0	64512 i
* >	2.2.0.0/24	3.3.3.3	0	100	0	2222 i
*** >**	**2.2.2.2/32**	**3.3.3.3**	**0**	**100**	**0**	**2222 i**
*** >**	**2.2.2.100/32**	**3.3.3.3**	**0**	**100**	**0**	**2222 i**
*** >**	**3.3.3.2/31**	**-**	**1**	**0**	**-**	**i**
*** >**	**10.0.0.0/8**	**3.3.3.3**	**0**	**100**	**0**	**2222 ?**
*** >**	**172.16.0.0/12**	**3.3.3.3**	**0**	**100**	**0**	**2222 ?**

(…)

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Sub-Route-Maps…

Let's build RFC1918, Martian and prefix-lenght control with global policies plus control AS_PATH which is peering specific

```
(…)
ip as-path access-list 2222 permit _2222$ any

ip prefix-list 1918 seq 5 permit 10.0.0.0/8 le 32
ip prefix-list 1918 seq 10 permit 172.16.0.0/12 le 32
ip prefix-list 1918 seq 15 permit 192.168.0.0/16 le 32
ip prefix-list 1918 seq 20 permit 100.64.0.0/10 le 32

ip prefix-list lenght seq 5 permit 0.0.0.0/0 ge 25

ip prefix-list martian seq 5 permit 0.0.0.0/8 le 32
ip prefix-list martian seq 10 permit 127.0.0.0/8 le 32
ip prefix-list martian seq 15 permit 169.254.0.0/16 le 32
ip prefix-list martian seq 20 permit 192.0.2.0/24 le 32
ip prefix-list martian seq 25 permit 224.0.0.0/4 le 32
ip prefix-list martian seq 30 permit 240.0.0.0/4 le 32
(…)
```

```
(...)
route-map as2222 deny 5
   sub-route-map global
!
route-map as2222 permit 50
   match as-path 2222
   set community 2222
   set local-preference 101
!
route-map global permit 5
   match ip address prefix-list 1918
!
route-map global permit 10
   match ip address prefix-list martian
!
route-map global permit 15
   match ip address prefix-list lenght
!
```

Apply as2222 on ASN 2222 peer

```
!
router bgp 1111
   router-id 1.1.1.6
   maximum-paths 4 ecmp 4
   bgp additional-paths install
   neighbor ebgp peer-group
   neighbor ebgp remote-as 2222
   neighbor ebgp remove-private-as
   neighbor ebgp soft-reconfiguration inbound all
   neighbor ebgp ebgp-multihop
   neighbor ebgp route-map as2222 in
   neighbor ebgp send-community
(…)
```

Confidential. Copyright © Arista 2016. All rights reserved.

Sub-Route-Maps…

Much better (no RFC1918, Martians and /32 from External peer etc…)

```
(…)
         Network                Next Hop            Metric  LocPref Weight  Path
 * >     1.0.0.0/8              -                   0       0       -       64512 i
 * >     1.1.1.4/32             1.1.1.29            0       100     0       64512 i
 * >     1.1.1.6/32             -                   0       0       -       i
 * >     1.1.1.8/30             1.1.1.29            0       100     0       64512 i
 * >     1.1.1.12/30            1.1.1.29            0       100     0       64512 i
 * >     1.1.1.28/30            -                   1       0       -       i
 *       1.1.1.28/30            1.1.1.29            0       100     0       64512 i
 * >     1.1.1.32/30            -                   1       0       -       i
 * >     1.1.1.40/31            1.1.1.29            0       100     0       64512 i
 * >     2.2.0.0/24             3.3.3.3             0       100     0       2222 i
 * >     3.3.3.2/31             -                   1       0       -       i
(…)
```

Confidential. Copyright © Arista 2016. All rights reserved.

Scale Prefix-list&AS-PATH lists

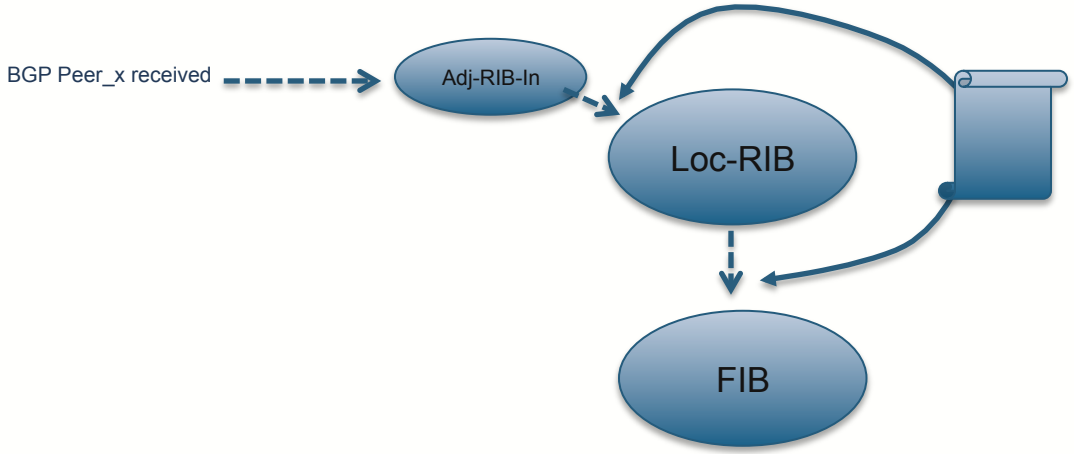
(Apply lesson learn from large DCs)

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

Prefix-lists

- These lists can be looooong with BGP peering devices
- EOS have ability to call prefix-list outside the CLI configuration
- Concept was originally to speed up the ability control what Routes from RIB to be installed in the FIB,
- The gain is both speed and flexibility when apply these with scripts
- The File can be local or remote



Confidential. Copyright © Arista 2016. All rights reserved.

URL based prefix-lists…

- The list been stored on the flash as a file… but can be read from CLI (or from bash) using “detail”.
- “refresh” knob reads&update changes to file

```
!
ip prefix-list martian2 source flash:/martian.txt
!
route-map xas2222 deny 10
   match ip address prefix-list martian2
!
```

```
bgp-edge(config)#sh ip prefix-list martian2 det
ip prefix-list martian2 source flash:/martian.txt
! Num entries: 6. Update time: 2017-02-14 20:39:36
    seq 5 permit 0.0.0.0/8 le 32
    seq 10 permit 127.0.0.0/8 le 32
    seq 15 permit 169.254.0.0/16 le 32
    seq 20 permit 192.0.2.0/24 le 32
    seq 25 permit 224.0.0.0/4 le 32
    seq 30 permit 240.0.0.0/4 le 32
(…)

bgp-edge(config)#
bgp-edge(config)#refresh ip prefix-list martian2
refreshed ip prefix-list martian2 source flash:/martian.txt
Num entries: 6
bgp-edge(config)#
```

Confidential. Copyright © Arista 2016. All rights reserved.

URL based AS_PATH reg-exp lists

- AS_PATH Reg-exp can be VERY long in the CLI and resource heavy to install with CPU spikes
- If do the same trick as with prefix-list, much can be offload with easier provisioning.

```
(...)
ip as-path access-list radb-as-akamai permit ^20940(_20940)*$ any
ip as-path access-list radb-as-akamai permit ^20940(_[0-9]+)*_(12222|16625|16702|18680)$ any
ip as-path access-list radb-as-akamai permit ^20940(_[0-9]+)*_(18717|20189|21342|21357)$ any
ip as-path access-list radb-as-akamai permit ^20940(_[0-9]+)*_(21399|22207|23454|23455)$ any
ip as-path access-list radb-as-akamai permit ^20940(_[0-9]+)*_(23903|24319|31107|31108)$ any
ip as-path access-list radb-as-akamai permit ^20940(_[0-9]+)*_(31109|31110|31377|33905)$ any
ip as-path access-list radb-as-akamai permit ^20940(_[0-9]+)*_(34164|34850|35204|35993)$ any
ip as-path access-list radb-as-akamai permit ^20940(_[0-9]+)*_(35994|39836|43639)$ any
ip as-path access-list radb-as-alltele permit ^44581(_44581)*$ any
ip as-path access-list radb-as-alltele permit ^44581(_[0-9]+)*_(13104|23456|31568|41012)$ any
ip as-path access-list radb-as-alltele permit ^44581(_[0-9]+)*_(42347|42823|50171|50904)$ any
ip as-path access-list radb-as-alltele permit ^44581(_[0-9]+)*_(50924|51821|62463)$ any
ip as-path access-list radb-as-amazon permit ^16509(_16509)*$ any
ip as-path access-list radb-as-amazon permit ^16509(_[0-9]+)*_(7224|8987|9059|10124)$ any
ip as-path access-list radb-as-amazon permit ^16509(_[0-9]+)*_(14618|17493|38895|39111)$ any
ip as-path access-list radb-as-amazon permit ^16509(_[0-9]+)*_(62785)$ any
ip as-path access-list radb-as-apple permit ^714(_714)*$ any
ip as-path access-list radb-as-apple permit ^714(_[0-9]+)*_(6185)$ any
ip as-path access-list radb-as-blix permit ^50304(_50304)*$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(8896|12654|15149|15888)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(16185|16186|18541|21030)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(22717|22989|23421|23456)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(26167|26414|27176|28824)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(29017|29300|29458|29479)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(29997|31005|31018|32748)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(34822|34989|35642|35687)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(35742|39392|39783|41869)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(42400|42473|43646|44511)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(44654|46805|48376|48630)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(48708|49262|49788|50531)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(50562|50582|51850|51872)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(51886|53861|55803|56334)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(56809|56867|57202|57381)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(57423|57963|57997|58298)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(59608|59767|59802|59863)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(60068|60526|60584|60717)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(60848|61122|61126|61227)$ any
ip as-path access-list radb-as-blix permit ^50304(_[0-9]+)*_(61238|61292|62248|62271)$ any
ip as-path access-list radb-as-cloudflare permit ^13335(_13335)*$ any
ip as-path access-list radb-as-cloudflare permit ^13335(_[0-9]+)*_(23456|33612)$ any
ip as-path access-list radb-as-cn permit ^42695(_42695)*$ any
ip as-path access-list radb-as-comhem permit ^39651(_39651)*$ any
ip as-path access-list radb-as-comhem permit ^39651(_[0-9]+)*_(25037|28942|56563)$ any
(...)
```

Confidential. Copyright © Arista 2016. All rights reserved.

ARISTA

URL based AS_PATH reg-exp lists…

- Speed to change/apply impressive !!!
- 7600 lines divided into 47 files
- In brief proof of the pudding
  - C code beats Python regards speed

```
(...)
ip as-path access-list new-radb-as-197541 source file:/tmp/new-radb-as-197541
ip as-path access-list new-radb-as-akamai source file:/tmp/new-radb-as-akamai
(...)
```

URL based AS-PATH ACL install file

```
[admin@nv442 tmp]$ time cat tacc_parse_conf | FastCli -A
real      0m0.711s
user      0m0.044s
sys       0m0.012s
```

CLI way (copy file running-config)

```
[admin@nv442 tmp]$ time cat orig_parse_conf | FastCli -A
real      11m32.254s
user      0m0.052s
sys       0m0.016s
```

Confidential. Copyright © Arista 2016. All rights reserved.