ARISTA WHITE PAPER

Arista FlexRoute™ Engine

Arista Networks' award-winning Arista 7500 Series was introduced in April 2010 as a revolutionary switching platform, which maximized data center performance, efficiency and overall network reliability. It raised the bar for switching performance, being five times faster, one-tenth the power draw and one-half the footprint compared to other modular data center switches.

In 2013, the Arista 7500E Series delivered a three-fold increase in density and performance, with no sacrifices on features and functionality and with complete investment protection.

Just three years later the Arista 7500R Universal Spine platform delivers more than a 3.8X increase in performance and density with significant increases in features and functionality including support for full internet table routing capacity.

FlexRoute™ is the technology that enables IP forwarding capacity in excess of 1M+ prefixes in hardware on the Arista 7500R Universal Spine and Arista 7280R Universal Leaf platforms. This whitepaper details the FlexRoute Engine.

ARISTA

ARISTA FLEXROUTE ENGINE

The Arista FlexRoute Engine provides support for the full internet routing table, in hardware, with IP forwarding at Layer 3 and with sufficient headroom for future growth in both IPv4 and IPv6 route scale to more than 1 million routes. The innovative FlexRoute Engine with its patented algorithmic approach to building layer 3 forwarding tables on Arista 7500R and 7280R Universal Spine and Leaf platforms is unique to Arista and a key enabler in calling these platforms routers.

On the hardware side, FlexRoute performs a longest-prefix-match (LPM) layer 3 lookup for IPv4 and IPv6 as part of the ingress packet processing on the distributed packet processor(s) on every linecard (Figure 1.) or system.

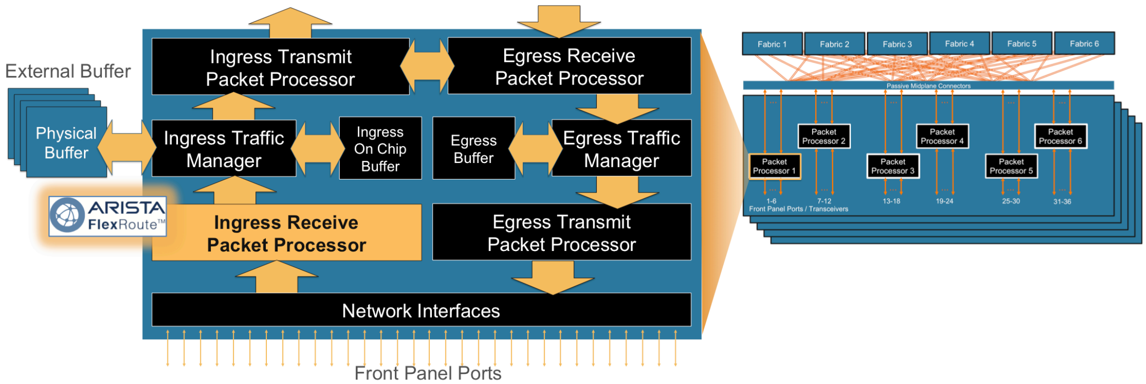


Figure 1: Arista FlexRoute Engine within the packet processor on linecards

Internally FlexRoute uses an algorithmic approach to performing lookups. When compared to legacy LPM approaches, FlexRoute uses less active silicon (lower *activity factor*) combined with a more efficient use of the transistors (denser storage) to hold the LPM forwarding tables. The result is dramatically lower power, a higher number of ports and greater throughput when compared to alternate approaches on the same process node.

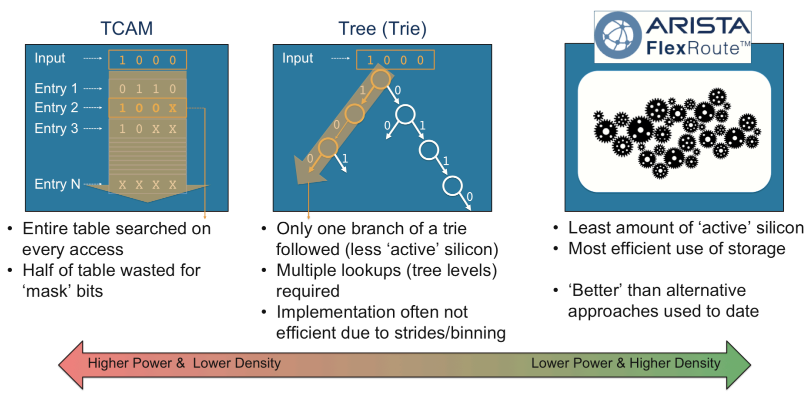


Figure 2: Arista FlexRoute Engine for longest-prefix-match lookups compared to alternatives

The algorithms used to perform the LPM lookup are optimized based on the historic growth of the internet routing table and known trends of how the routing table is expected to evolve. For example, FlexRoute is optimized on the continued and expected acceleration of de-aggregation of the IPv4 prefix space (e.g. /23 prefixes de-aggregating to 2 x /24s). It is also optimized around an aggressive expansion of IPv6 announcements (most prefix announcements are /32 and /48). In comparison to the legacy ways of increasing LPM tables which either involve increasing the size of tables and memories (more transistors, more power/heat, lower port density) or increasing the depth of lookups in a tree structure (lower performance), the algorithmic approach used in FlexRoute becomes *more efficient* with these trends and the evolution of the internet routing table.

PATHS, PREFIXES AND INTERNET GROWTH

At Arista, we're confident the algorithmic techniques used to build the LPM in FlexRoute will provide many years of headroom for continued growth of the internet routing table. Without stating how far beyond 1M+ prefixes it can scale or how the efficiency evolves over time, lets look back at the how the internet routing table has evolved to its current size (May 2016: ~649K prefixes [610K IPv4, ~29K IPv6]) and how it is expected to evolve in future.

PAST, PRESENT AND FUTURE INTERNET GROWTH

Geoff Huston, the Chief Scientist at APNIC, the Asia Pacific Regional Internet Registry has been providing research, analysis and commentary on the global internet routing table for close to a decade. In January 2016 Geoff, as part of APNIC Labs, published an analysis of the Internet routing table in 2015[1] building upon previous years' analysis and commentary on the topic.

The exact number of IPv4 and IPv6 prefixes that make up the internet varies depending on location and localized summarization, however the broad number of prefixes is quite clear, so too are the trends. Using the passive measurement point of the global routing table from AS131072 and its data from the perspective of Australia and Japan in the APNIC region, the data collected shows IPv4 and IPv6 prefix space expansion as follows:

Table 1: Historic growth of IPv4 and IPv6 announcements (source: Geoff Huston / APNIC Labs Table 1 & 2 from [1])

Metric	Jan-2013	Jan-2014	Jan-2015	Jan-2016
IPv4 prefixes	441,000	488,000 (+10%)	530,000 (+9%)	587,000 (+11%)
IPv6 prefixes	11,900	16,700 (+40%)	21,000 (+26%)	27,200 (+30%)
Total (IPv4+IPv6)	**452,000**	**504,700 (+11%)**	**551,000 (+9%)**	**614,200 (+11%)**

Taking into account the Regional Internet Registry prefix allocations and actual prefix route announcements (e.g. more specific prefixes advertised) and how that trend has increased over time, with a view to what future prefix announcements, updates and de-aggregation will likely happen based on historic trends, the same report provides predictions for the future expected growth. IPv6 is a little harder to predict, so the report provides predictions based both on linear growth *(L)* and exponential growth *(E)*, with the reality most likely somewhere between the two:

Table 2: Future predicted growth of IPv4 and IPv6 announcements (source: Geoff Huston / APNIC Labs Tables 3 & 4 from [1])

<table>
<tr><th>Metric</th><th>Jan-2016
(actual)</th><th>Jan-2017
(prediction)</th><th>Jan-2018
(prediction)</th><th>Jan-2019
(prediction)</th><th>Jan-2020
(prediction)</th><th>Jan-2021
(prediction)</th></tr>
<tr><td>IPv4 prefixes</td><td>586,879</td><td>629,000 (+7%)</td><td>675,000 (+7%)</td><td>722,000 (+7%)</td><td>769,000 (+7%)</td><td>816,000 (+6%)</td></tr>
<tr><td>IPv6 prefixes (L)</td><td rowspan="2">27,241</td><td>30,421 (+12%)</td><td>35,113 (+15%)</td><td>39,806 (+13%)</td><td>44,498 (+12%)</td><td>49,203 (+11%)</td></tr>
<tr><td>IPv6 prefixes (E)</td><td>37,968 (+39%)</td><td>51,303 (+35%)</td><td>69,322 (+35%)</td><td>93,669 (+35%)</td><td>126,671 (+35%)</td></tr>
<tr><td>Total (linear IPv6)</td><td rowspan="2">614,120</td><td>659,421 (+7%)</td><td>710,113 (+8%)</td><td>761,806 (+7%)</td><td>813,498 (+7%)</td><td>865,203 (+6%)</td></tr>
<tr><td>Total
(exponential IPv6)</td><td>666,968 (+9%)</td><td>726,303 (+9%)</td><td>791,322 (+9%)</td><td>862,669 (+9%)</td><td>942,671 (+9%)</td></tr>
</table>

While the predictions in [1] summarized in Table 2 are predictions, the underlying data clearly shows there is more than 5 years' of headroom before the total of IPv4 and IPv6 prefix announcements cumulatively exceeds 1 million entries, even with an aggressive expansion rate.

BGP PATHS, ROUTES AND FORWARDING ENTRIES

There are often misconceptions on how prefixes and paths in BGP relate to entries stored in forwarding tables.

For example, if you receive transit capacity from three upstream providers (BGP neighbors), each sending 600K prefixes in BGP, there are 1.8 million paths (600K x 3 neighbors) but this still 600K unique prefixes, not 1.8 million prefixes. That some prefixes are preferred via one neighbor or another would be resolved at the BGP level, or if there are multiple equal-cost paths for a prefix, the route prefix would be via equal-cost-multi-pathing (ECMP), however the result is still that there are still only 600K prefixes just that some prefixes point at one next-hop or another, or a group of next-hop entries in the ECMP case.

The relationship between prefixes received in BGP and how they are stored in the routing table (RIB) and forwarding table (FIB) is shown in figure 3.

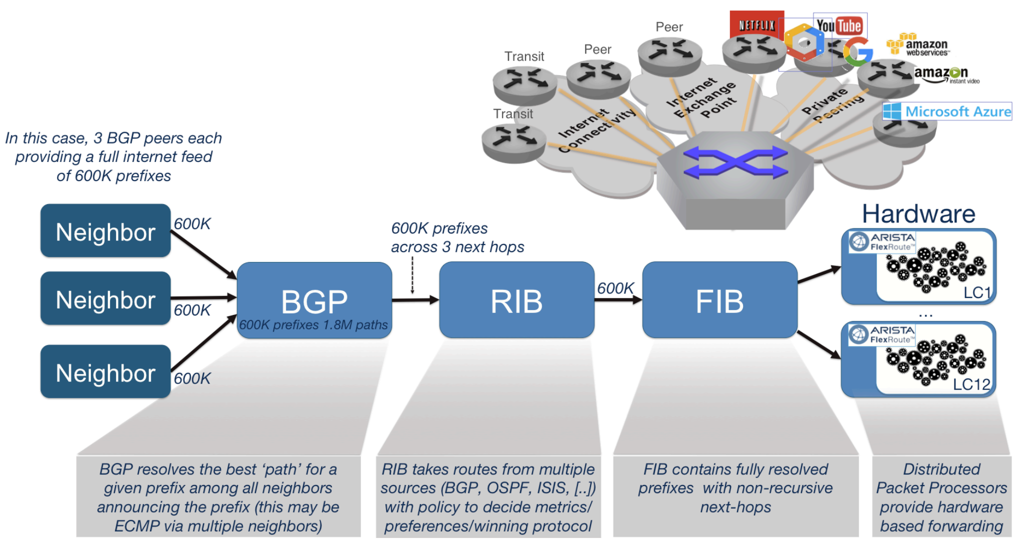


Figure 3: Prefixes received in BGP and their resolution from BGP to RIB to FIB

Regardless of number of the number of full tables received from transit providers, numbers of peers, or even someone inadvertently announcing prefixes they aren't meant to, there is no increase in the number of prefixes as a result of multiple transit or peering providers.

REAL WORLD FLEXROUTE RESOURCE UTILIZATION

Arista's innovative FlexRoute Engine is designed and built around the internet routing table and prefix distribution with capacity of over 1 million prefixes for IPv4 and IPv6 combined. FlexRoute is enabled via a FlexRoute license and the following CLI commands:

```
arista(config)# ip hardware fib optimize prefixes profile internet
arista(config)# ipv6 hardware fib optimize prefixes profile internet
```

Real world examples of the hardware capacity and resources utilized in multiple deployments are shown below.

REAL WORLD EXAMPLE 1: INTERNET2 EDGE ROUTER (IPV4 ONLY)

In this deployment (an Internet2 edge router) of IPv4, there are ~595K prefixes received from two BGP neighbors that resulted in ~579K unique prefixes in the routing table (RIB). The highest-capacity hardware resource in this case is at 62% usage. The *"show hardware capacity" EOS* command shows the resource utilization:

```
localhost# show ip bgp summary
BGP summary information for VRF default
Router identifier xxx.yyy.yy.8, local AS number zzzzz
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State  PfxRcd PfxAcc
  xxx.yy.yy.2      4  zzzzz           5306       593    0    0 04:07:41 Estab  16863  16863
  xxx.yy.yy.5      4  zzzzz         179287       588    0    0 04:52:20 Estab  577115 577115


localhost# show ip route summary | tail -10
Total Routes                                        579300

Number of routes per mask-length:
   /0: 1          /8: 18         /9: 13         /10: 35        /11: 101
   /12: 267       /13: 504       /14: 1020      /15: 1716      /16: 12838
   /17: 7334      /18: 12446     /19: 25153     /20: 36598     /21: 38647
   /22: 64463     /23: 55814     /24: 321823    /25: 75        /26: 115
   /27: 200       /28: 21        /29: 30        /30: 44        /31: 6
   /32: 18


localhost# show hardware capacity | grep Routing
Forwarding Resources Usage

Table    Feature    Chip          Used  Used      Free   Committed   Best Case      High
                               Entries   (%)   Entries     Entries         Max Watermark
                                                                       Entries
-------- ---------- --------- -------- ----- --------- ----------- ----------- ---------
Routing  Resource1                 662   32%      1386           0        2048       664
Routing  Resource2                 411   40%       613           0        1024       419
Routing  Resource3               12258   37%     20510           0       32768     12304
Routing  V4Hosts                     0    0%     57344           0       57344         0
Routing  V4Routes               494832   62%    291596           0      786432    495235
Routing  V6Hosts                     0    0%     57344           0       57344         0
Routing  V6Routes                    0    0%    291596           0      786432         0
```

Approx. 595K IPv4 paths via BGP

Resolves to ~579K unique prefixes in RIB

Much hardware resources for future growth …

Figure 4: A router connected to Internet2

REAL WORLD EXAMPLE 2: LARGE HOSTING PROVIDER OFFERING FULL IPV4/IPV6

In this deployment, a large hosting provider is using the device as an edge router, with full internet IPv4 and IPv6 as well as many internal prefixes. In this case ~2.9M IPv4 and ~204K IPv6 paths are received via BGP which results in ~854K IPv4 and ~45K IPv6 prefixes in the routing table (RIB). The ~900K aggregate prefixes result in 83% usage of the highest-utilized hardware routing resource:

```
rbx-testjericho-a72# show ip bgp summary
BGP summary information for VRF default
Router identifier xxx.yyy.y.5, local AS number zzzzz
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State  PfxRcd PfxAcc
  xxx.yyy.y.11     4  zzzzz          14113       376    0    0 06:12:50 Estab  216183 216174
  xxx.yyy.y.12     4  zzzzz         222636       376    0    0 06:13:05 Estab  597610 597610
  xxx.yyy.y.13     4  zzzzz         164351       219    0    0 03:36:13 Estab  813392 813383
  xxx.yyy.y.100    4  zzzzz           1073       221    0    0 03:37:31 Estab  40387  40387
  xxx.yyy.y.101    4  zzzzz         185725       221    0    0 03:37:29 Estab  580821 580821
  xxx.yyy.y.102    4  zzzzz         185499       221    0    0 03:37:26 Estab  621090 621090
rbx-testjericho-a72# show ipv6 bgp summary
BGP summary information for VRF default
Router identifier xxx.yyy.y.5, local AS number zzzzz
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State  PfxRcd PfxAcc
  xxxx:xxxx::1024  4  zzzzz          60125       379    0    0 06:14:11 Estab  36509  15738
  xxxx:xxxx::1025  4  zzzzz          63947       377    0    0 06:14:22 Estab  39870  39870
  xxxx:xxxx::1026  4  zzzzz          47078       219    0    0 03:36:09 Estab  42678  42678
  xxxx:xxxx::1000  4  zzzzz           1978       222    0    0 03:37:15 Estab  15738  15738
  xxxx:xxxx::1001  4  zzzzz          46925       222    0    0 03:37:11 Estab  26840  26840
  xxxx:xxxx::1002  4  zzzzz          48853       222    0    0 03:37:08 Estab  42580  42580
rbx-testjericho-a72# show ip route summary | grep Total
Total Routes                                     854393
rbx-testjericho-a72# show ipv6 route summary | grep Total
 Total Routes                           45678
rbx-testjericho-a72# show hardware capacity | grep Routing
Forwarding Resources Usage

Table    Feature    Chip          Used  Used      Free   Committed   Best Case      High
                               Entries   (%)   Entries     Entries         Max Watermark
                                                                       Entries
-------- ---------- --------- -------- ----- --------- ----------- ----------- ---------
Routing  Resource1                1319   64%       729           0        2048      1320
Routing  Resource2                 809   79%       215           0        1024       814
Routing  Resource3               24102   73%      8666           0       32768     24104
Routing  V4Hosts                     1    0%     57343           0       57344         1
Routing  V4Routes               644336   83%    124302           0      786432    644364
Routing  V6Hosts                     0    0%     57344           0       57344         0
Routing  V6Routes                17792   12%    124302           0      786432     17795
```

~2.9M IPv4 paths via BGP

… and ~204K IPv6 paths via BGP

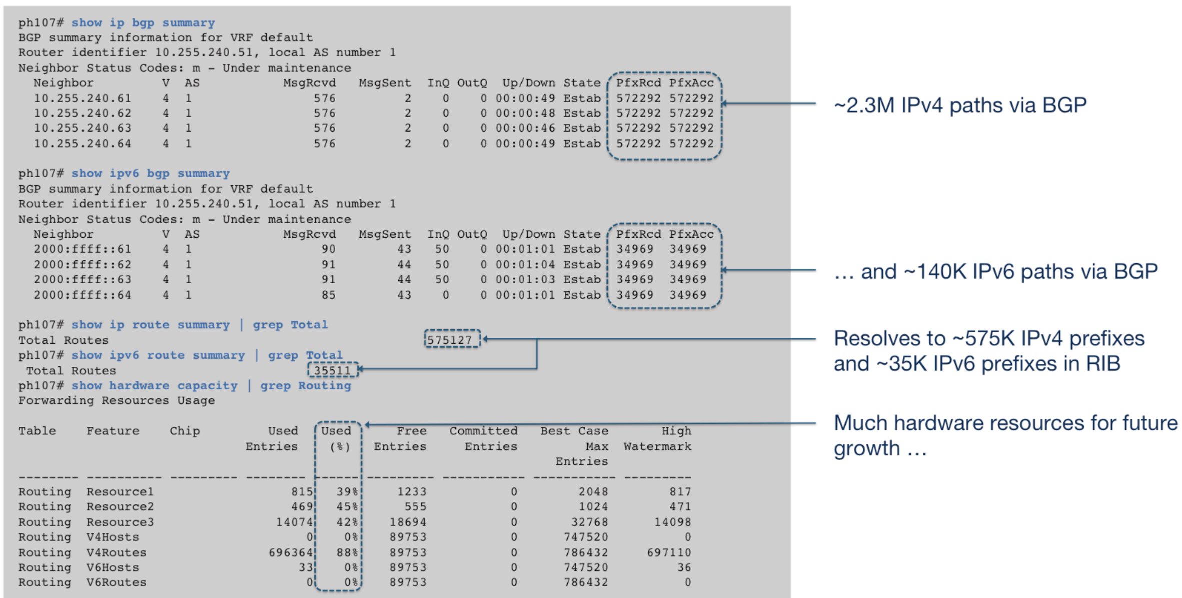
Resolves to ~854K IPv4 prefixes and ~45K IPv6 prefixes in RIB

Much hardware resources for future growth …

Figure 5: A large hosting provider with both IPv4 & IPv6

REAL WORLD EXAMPLE 3: CLOUD TITAN FULL IPV4/IPV6 INTERNET EDGE ROUTER

In this deployment, a cloud titan is using the device as an edge router, with both IPv4 and IPv6 via multiple transit providers. In this case there are four full feeds for both IPv4 and IPv6 with ~2.3M IPv4 and ~140K IPv6 paths that results in ~575K IPv4 and ~35K IPv6 prefixes in the routing table (RIB). The highest-utilized hardware resource in this case is 88%:

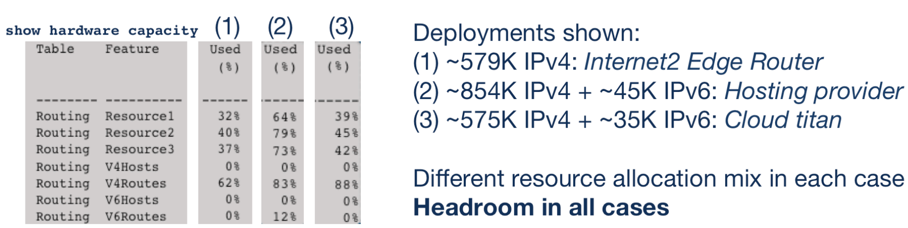


```
ph107# show ip bgp summary
BGP summary information for VRF default
Router identifier 10.255.240.51, local AS number 1
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State  PfxRcd PfxAcc
  10.255.240.61    4  1                576         2    0    0 00:00:49 Estab  572292 572292
  10.255.240.62    4  1                576         2    0    0 00:00:48 Estab  572292 572292
  10.255.240.63    4  1                576         2    0    0 00:00:46 Estab  572292 572292
  10.255.240.64    4  1                576         2    0    0 00:00:49 Estab  572292 572292

ph107# show ipv6 bgp summary
BGP summary information for VRF default
Router identifier 10.255.240.51, local AS number 1
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State  PfxRcd PfxAcc
  2000:ffff::61    4  1                 90        43   50    0 00:01:01 Estab  34969  34969
  2000:ffff::62    4  1                 91        44   50    0 00:01:04 Estab  34969  34969
  2000:ffff::63    4  1                 91        44   50    0 00:01:03 Estab  34969  34969
  2000:ffff::64    4  1                 85        43    0    0 00:01:01 Estab  34969  34969

ph107# show ip route summary | grep Total
Total Routes                                          575127
ph107# show ipv6 route summary | grep Total
 Total Routes                              35511
ph107# show hardware capacity | grep Routing
Forwarding Resources Usage

Table    Feature    Chip          Used   Used     Free   Committed   Best Case      High
                               Entries    (%)  Entries     Entries         Max Watermark
                                                                       Entries
-------- ---------- --------- -------- ------ -------- ----------- ----------- ---------
Routing  Resource1                 815    39%     1233           0        2048       817
Routing  Resource2                 469    45%      555           0        1024       471
Routing  Resource3               14074    42%    18694           0       32768     14098
Routing  V4Hosts                     0     0%    89753           0      747520         0
Routing  V4Routes               696364    88%    89753           0      786432    697110
Routing  V6Hosts                    33     0%    89753           0      747520        36
Routing  V6Routes                    0     0%    89753           0      786432         0
```


Figure 6: A cloud titan provider with four full feeds (4 transit providers) for both IPv4 & IPv6

HARDWARE RESOURCE SUMMARY

Due to the algorithmic approach, exactly which resources are used varies across deployments. In the examples provided there is more than sufficient capacity to forward using the full internet routing table, with forwarding resource headroom for many years of future growth:



```
show hardware capacity   (1)    (2)    (3)
 Table    Feature        Used   Used   Used
                          (%)    (%)    (%)

 -------- ----------    ------ ------ ------
 Routing  Resource1       32%    64%    39%
 Routing  Resource2       40%    79%    45%
 Routing  Resource3       37%    73%    42%
 Routing  V4Hosts          0%     0%     0%
 Routing  V4Routes        62%    83%    88%
 Routing  V6Hosts          0%     0%     0%
 Routing  V6Routes         0%    12%     0%
```


Figure 7: Summary of hardware resource utilization across the examples

Arista's work on the algorithms and techniques around FlexRoute will continue, with additional capacity enhancements planned.

ARISTA EOS, SYSDB AND NETDB

At the core of the Arista 7500R and 7280R Universal Spine and Leaf platforms is Arista EOS® (Extensible Operating System). EOS is built on the strong foundations of a multi-process state-sharing architecture with modularity, programmability, fault containment and resiliency as the core software building blocks.

System state is stored in a highly efficient, centralized System Database (SysDB) and accessed using an automated publish/subscribe/notify model and internally NetDB is used to enable scaling of the routing stack to support millions of routes and hundreds of neighbors with faster convergence than traditional routers and legacy approaches to control-plane state on routers would otherwise.

While many network vendors claim they have a fast, scalable and robust control-plane, the fine print is that it can take seconds to react to failures and minutes for routes to be programmed in hardware. Arista EOS scales with industry-leading convergence and route programming, sub-second (typically milliseconds) reaction times to disruptions. In contrast to legacy approaches a key consideration of FlexRoute has been the ability to support fast prefix programming in the dataplane and make-before-break programming of the forwarding tables in hardware that doesn't disrupt adjacent entries.

SUMMARY

Arista's FlexRoute Engine provides support for the full internet routing table in hardware, with IP forwarding at Layer 3 and with sufficient headroom for future growth in both IPv4 and IPv6 route scale to more than 1 million routes. The innovative FlexRoute Engine with its patented algorithmic approach to building layer 3 forwarding tables on Arista 7500R and 7280R Universal Spine and Leaf platforms is unique to Arista and a key enabler in calling these platforms routers.

REFERENCES

[1] Analysis of the Internet Routing table in 2015, Geoff Huston (APNIC): https://labs.apnic.net/?p=767

Santa Clara—Corporate Headquarters
5453 Great America Parkway
Santa Clara, CA 95054
Tel: 408-547-5500
www.arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office
1390 Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A & B, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062