Arista FlexRoute™ Engine

Arista Networks' award-winning Arista 7500 Series was introduced in April 2010 as a revolutionary switching platform, which maximized datacenter 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 datacenter 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.

In 2016, the Arista 7500R Universal Spine platform delivered 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.

Just three years later, the Arista 7500R3/7800R3 Universal Spine platform delivers a five-fold increase in bandwidth and performance, with noticeable improvements in scale and system throughput to meet large scale deployments.

FlexRoute™ is the technology that enables IP forwarding capacity in excess of 2.5M+ prefixes in hardware on the Arista 7500R3/7800R3 Universal Spine and Arista 7280R3 Universal Leaf platforms. This whitepaper details the FlexRoute Engine.

Arista 7800R Series Modular Data Center Switches

Arista 7500R3 Series Chassis (left to right) - 7512R, 7508R and 7504R

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 2.5 million routes. The innovative FlexRoute Engine with its patented algorithmic approach to building layer 3 forwarding tables on Arista 7500R3/7800R3 and 7280R3 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.

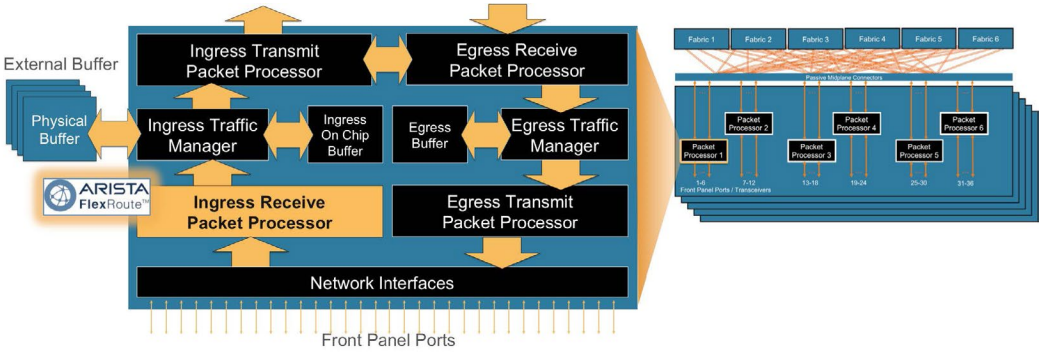


Figure 1: Arista FlexRoute Engine within the packet processor on linecards

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.

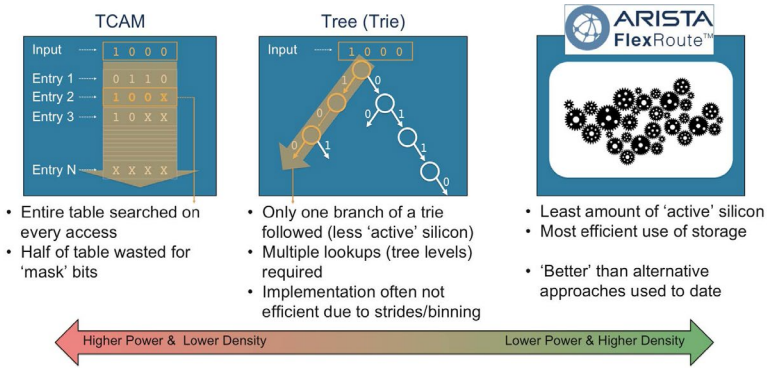


Figure 2: Arista FlexRoute Engine for longest-prefix-match lookups compared to alternatives

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 deaggregating 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 the continuous growth of the internet routing table. Let's look back at the how the internet routing table has evolved to its current size (June 2019: ~856K prefixes [785K IPv4, ~71K IPv6]) and how it is expected to evolve in future. [2]

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 more than a decade. In January 2019 Geoff, as part of APNIC Labs, published an analysis of the Internet routing table in 2018[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-2016	Jan-2017	Jan-2018	Jan-2019
IPv4 prefixes	587,000	646,000 (+10%)	699,000 (+8%)	760,000 (+9%)
IPv6 prefixes	27,200	34,800 (+28%)	45,700 (+31%)	62,400 (+37%)
Total (IPv4+IPv6)	**614,200**	**504,700 (+11%)**	**744,700 (+9%)**	**822,400 (+10%)**

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: Forecasting the IPv4 and IPv6 BGP Table (source: Geoff Huston / APNIC Labs Table 3 & 5 from [1])

Metric	Jan-2019 (actual)	Jan-2020 (prediction)	Jan-2021 (prediction)	Jan-2022 (prediction)	Jan-2023 (prediction)	Jan-2024 (prediction)
IPv4 prefixes	755000	810,000 (+7%)	864,000 (+7%)	919,000 (+6%)	974,000 (+6%)	1028,000 (+6%)
IPv6 prefixes (L)	62,000	75,000 (+21%)	89,000 (+19%)	102,000 (+15%)	116,000 (+14%)	130,000 (+12%)
IPv6 prefixes (E)	62,000	83,000 (+34%)	109,000 (+31%)	145,000 (+33%)	192,000 (+32%)	255,000 (+33%)
Total (linear IPv6)	**817,000**	**885,000 (+8%)**	**953,000(+8%)**	**1,021,000 (+7%)**	**1,090,000 (+7%)**	**1,158,000(+6%)**
Total (exponential IPv6)	**817,000**	**893,000 (+9%)**	**973,000 (+9%)**	**1064,000 (+9%)**	**1,166,000 (+10%)**	**1,283,000 (+10%)**

While the predictions in [1] summarized in Table 2 are predictions, the underlying data clearly shows there is more than 2 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.

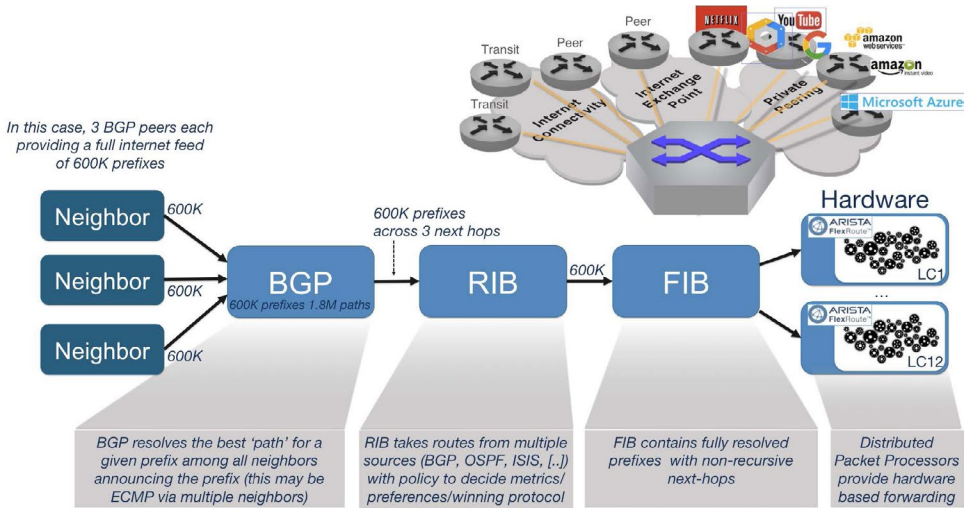


Figure 3: Prefixes received in BGP and their resolution from BGP to RIB to FIB

Regardless of 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 Resources Utilization

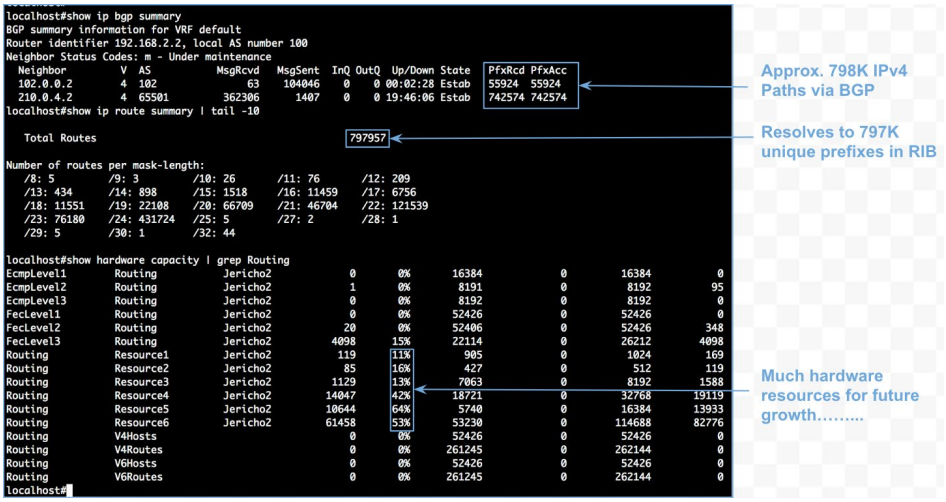
Arista's innovative FlexRoute Engine is designed and built around the internet routing table and prefix distribution with capacity of over 2.5 million prefixes for IPv4 and IPv6 combined. FlexRoute is enabled via a FlexRoute license and the following CLI commands:

```
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 ~798K prefixes received from two BGP neighbors that resulted in ~797K unique prefixes in the routing table (RIB). The highest-capacity hardware resource in this case is at 64% usage. The "show hardware capacity" EOS command shows the resource utilization:



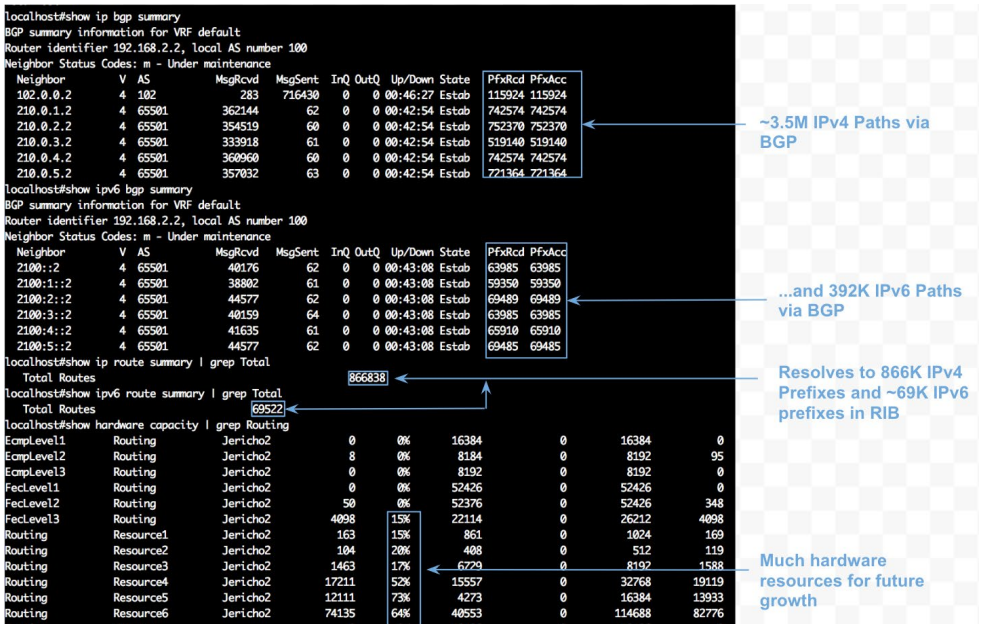

```
localhost#show ip bgp summary
BGP summary information for VRF default
Router identifier 192.168.2.2, local AS number 100
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  102.0.0.2        4  102              63    104046    0    0 00:02:28 Estab   55924  55924
  210.0.4.2        4  65501        362306      1407    0    0 19:46:06 Estab   742574 742574
localhost#show ip route summary | tail -10

   Total Routes                                797957

Number of routes per mask-length:
   /8: 5          /9: 3          /10: 26        /11: 76        /12: 209
   /13: 434       /14: 898       /15: 1518      /16: 11459     /17: 6756
   /18: 11551     /19: 22108     /20: 66709     /21: 46704     /22: 121539
   /23: 76180     /24: 431724    /25: 5         /27: 2         /28: 1
   /29: 5         /30: 1         /32: 44

localhost#show hardware capacity | grep Routing
EcmpLevel1        Routing          Jericho2            0      0%      16384          0      16384      0
EcmpLevel2        Routing          Jericho2            1      0%       8191          0       8192     95
EcmpLevel3        Routing          Jericho2            0      0%       8192          0       8192      0
FecLevel1         Routing          Jericho2            0      0%      52426          0      52426      0
FecLevel2         Routing          Jericho2           20      0%      52406          0      52426    348
FecLevel3         Routing          Jericho2         4098     15%      22114          0      26212   4098
Routing           Resource1        Jericho2          119     11%        905          0       1024    169
Routing           Resource2        Jericho2           85     16%        427          0        512    119
Routing           Resource3        Jericho2         1129     13%       7063          0       8192   1588
Routing           Resource4        Jericho2        14047     42%      18721          0      32768  19119
Routing           Resource5        Jericho2        10644     64%       5740          0      16384  13933
Routing           Resource6        Jericho2        61458     53%      53230          0     114688  82776
Routing           V4Hosts                              0      0%      52426          0      52426      0
Routing           V4Routes                             0      0%     261245          0     262144      0
Routing           V6Hosts                              0      0%      52426          0      52426      0
Routing           V6Routes                             0      0%     261245          0     262144      0
localhost#
```

Figure 4: A router connected to Internet2




```
localhost#show ip bgp summary
BGP summary information for VRF default
Router identifier 192.168.2.2, local AS number 100
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  102.0.0.2        4  102              283    716430    0    0 00:46:27 Estab   115924 115924
  210.0.1.2        4  65501         362144        62    0    0 00:42:54 Estab   742574 742574
  210.0.2.2        4  65501         354519        60    0    0 00:42:54 Estab   752370 752370
  210.0.3.2        4  65501         333918        61    0    0 00:42:54 Estab   519140 519140
  210.0.4.2        4  65501         360960        60    0    0 00:42:54 Estab   742574 742574
  210.0.5.2        4  65501         357032        63    0    0 00:42:54 Estab   721364 721364
localhost#show ipv6 bgp summary
BGP summary information for VRF default
Router identifier 192.168.2.2, local AS number 100
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  2100::2          4  65501           40176        62    0    0 00:43:08 Estab   63985  63985
  2100:1::2        4  65501           38802        61    0    0 00:43:08 Estab   59350  59350
  2100:2::2        4  65501           44577        62    0    0 00:43:08 Estab   69489  69489
  2100:3::2        4  65501           40159        64    0    0 00:43:08 Estab   63985  63985
  2100:4::2        4  65501           41635        61    0    0 00:43:08 Estab   65910  65910
  2100:5::2        4  65501           44577        62    0    0 00:43:08 Estab   69485  69485
localhost#show ip route summary | grep Total
   Total Routes                                866838
localhost#show ipv6 route summary | grep Total
   Total Routes                                69522
localhost#show hardware capacity | grep Routing
EcmpLevel1        Routing          Jericho2            0      0%      16384          0      16384      0
EcmpLevel2        Routing          Jericho2            8      0%       8184          0       8192     95
EcmpLevel3        Routing          Jericho2            0      0%       8192          0       8192      0
FecLevel1         Routing          Jericho2            0      0%      52426          0      52426      0
FecLevel2         Routing          Jericho2           50      0%      52376          0      52426    348
FecLevel3         Routing          Jericho2         4098     15%      22114          0      26212   4098
Routing           Resource1        Jericho2          163     15%        861          0       1024    169
Routing           Resource2        Jericho2          104     20%        408          0        512    119
Routing           Resource3        Jericho2         1463     17%       6729          0       8192   1588
Routing           Resource4        Jericho2        17211     52%      15557          0      32768  19119
Routing           Resource5        Jericho2        12111     73%       4273          0      16384  13933
Routing           Resource6        Jericho2        74135     64%      40553          0     114688  82776
```

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 ~3.7M IPv4 and ~277K IPv6 paths that results in ~752K IPv4 and ~69K IPv6 prefixes in the routing table (RIB). The highest-utilized hardware resource in this case is 65%:

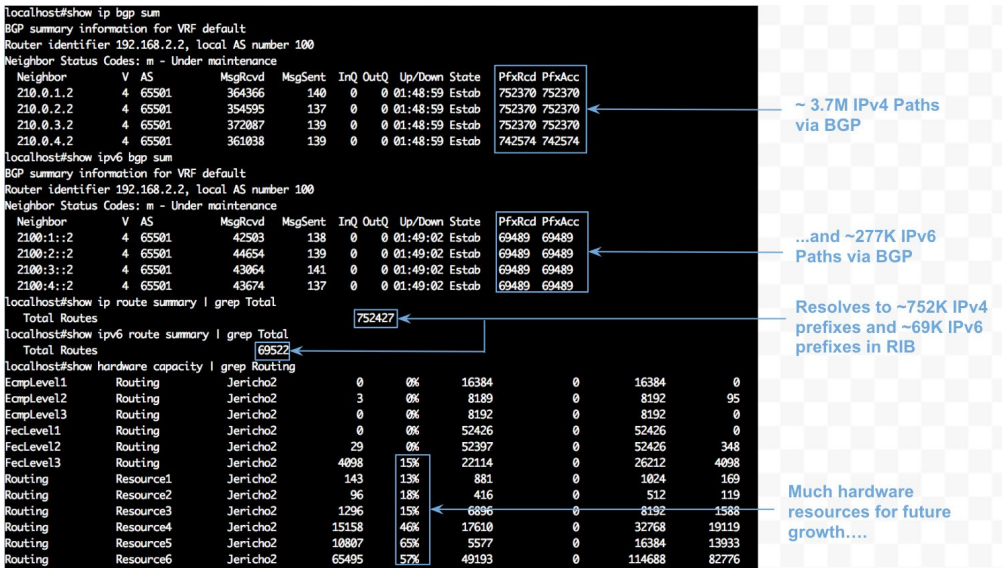


Figure 6: A cloud titan provider with four full feeds (4 transit providers) for both IPv4 & IPv6

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:

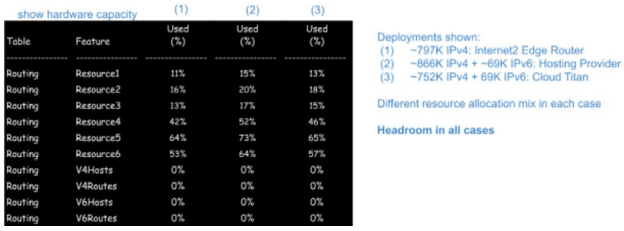


Figure 7: Summary of hardware resource utilization across the examples

Arista's work on the algorithms and techniques around FlexRoute will continue, with additional capacity enhancements planned.
Body Subhead style Level 2

Arista EOS, SysDB and NetDB

At the core of the Arista 7500R3/7800R3 and 7280R3 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.

Conclusion

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 2.5 million routes. The innovative FlexRoute Engine with its patented algorithmic approach to building layer 3 forwarding tables on Arista 7500R3/7800R3 and 7280R3 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 2018, Geoff Huston (APNIC): https://labs.apnic.net/?p=1195

[2] IPv6 CIDR REPORT: http://www.cidr-report.org/v6/as2.0/ and CIDR REPORT: https://www.cidr-report.org/as2.0/

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@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