

# Arista FlexRoute™ Engine

One of the foundations for Arista Networks Cloud-Grade Routing is the R-series portfolio. Introduced in 2016, the R-series platforms provide industry leading performance, scale and port-density. The feature richness and flexibility of the platforms allows them to be deployed in a wide range of open networking solutions including large scale layer-2 and layer-3 cloud designs, provider edge where scaleable L2 and L3 VPN services are required, in a traffic engineered MPLS core where high density 400G/100G is required, metro-aggregation for the backhaul of Ethernet services, in the Telco data center for EVPN overlay solutions & lately AI spine for modern workloads in data centers with a lot of successful deployments that are running live today. The portfolio serves as a universal platform for routing & switching use cases inside and outside of the data center. Multiple generations of silicon, in R-series, have enabled delivery of 2x-3x incremental increase in throughput and deliver maximum performance for modern networks in space and power efficient form factors.

One of the key innovations that have been applied on the R-series is Arista's FlexRoute™ technology, which enables optimized IP forwarding capacity in hardware in an algorithmic way, to address use cases around growth beyond Internet route scale.

This white paper details the FlexRoute technology.



*Arista 7280R3/R3A fixed platforms*



*Arista 7800R3/R3A modular platforms*

## Arista FlexRoute Engine

The Arista FlexRoute Engine provides support for IP forwarding capacities in hardware with sufficient headroom for future growth in both IPv4 and IPv6 route scale to more than 5 million routes. The innovative FlexRoute Engine with its patented algorithmic approach to building layer-3 forwarding tables on Arista R-series platforms is unique to Arista.

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.

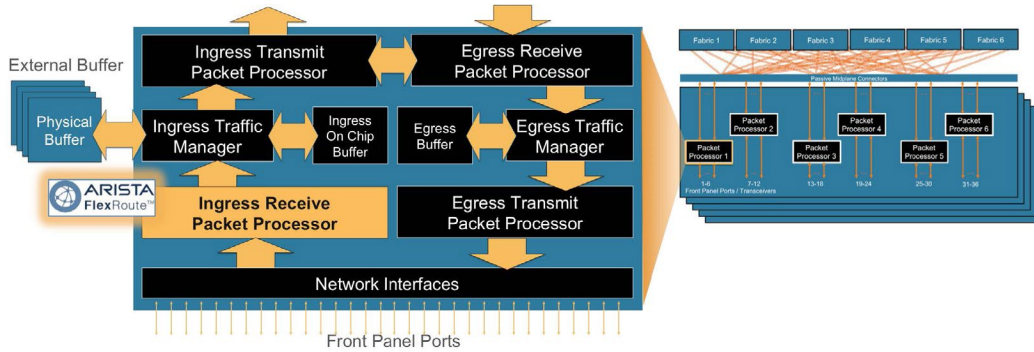


Figure 1: Arista FlexRoute Engine within the packet processor on linecards

Internally FlexRoute uses an algorithmic approach for storing prefixes & performing lookups, it splits the overall lookup process depending on the target prefix among multiple HW tables that reside within the packet processor pipeline to get a better overall utilization of system resources. When compared to typical longest prefix match (LPM) approaches, FlexRoute uses less active silicon (lower activity factor) combined with a more efficient use of the transistors (denser storage) to hold the 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.

With FlexRoute, all IP forwarding lookups & optimizations are done inside the packet processor without going out to external memories, this allows for fully utilizing the available interfaces on the packer processor as customer useable ports rather than being used as interfaces to connect to external co-processors or memories that offload IP forwarding lookups outside the packet processor.

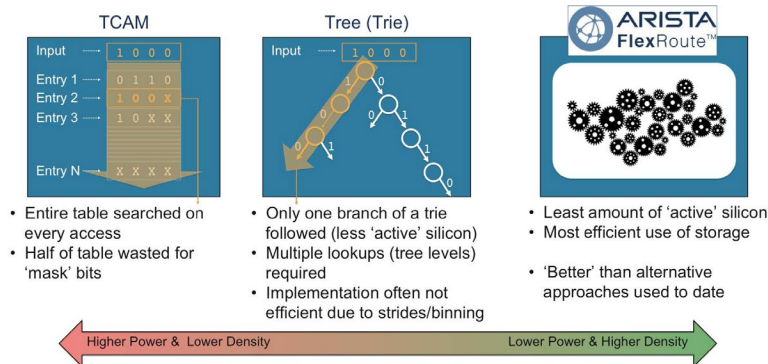


Figure 2: Arista FlexRoute Engine for longest-prefix-match lookups compared to alternatives

The algorithms used to perform the IP forwarding 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. It is also optimized around an aggressive expansion of IPv6 announcements (most prefix announcements are /32 and /48). In comparison to the typical ways of increasing LPM tables which

revolve around 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

Let's look at the history & expectations of growth of the internet routing table and how FlexRoute provides a solution with headroom for years of growth.

### 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 more than a decade. In January 2023 Geoff, as part of APNIC Labs, published an analysis of the Internet routing table in 2022 [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-2019	Jan-2020	Jan-2021	Jan-2022	Jan-2023
IPv4 prefixes	760,000	814,000 (+7%)	860,000 (+6%)	906,000 (+5%)	940,000 (+4%)
IPv6 prefixes	62,400	79,400 (+27%)	105,500 (+33%)	146,500 (+39%)	172,400 (+18%)
<b>Total (IPv4+IPv6)</b>	<b>822,400</b>	<b>893,400 (+9%)</b>	<b>965,500 (+8%)</b>	<b>1,052,500 (+9%)</b>	<b>1,112,400 (+6%)</b>

The analysis provides very useful insights on how the number of prefixes in the internet have been growing & the difference in growth rates for IPv4 versus IPv6 prefixes; Taking into account the Regional Internet Registry prefix allocations and actual prefix route announcements (e.g. more specific prefixes advertised) and how that trend of growth has changed over the past few years, the same report provides predictions for the future expected growth. It is not very easy to get accurate predictions for the future specially with how growth rates has changed in the last couple of years versus how the growth rates have been in previous years, that rate of growth for IPv6 is a little harder to predict as well, 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 & 4 from [1])**

Metric	Jan-2023 (actual)	Jan-2024 (prediction)	Jan-2025 (prediction)	Jan-2026 (prediction)	Jan-2027 (prediction)	Jan-2028 (prediction)	Jan-2029 (prediction)
IPv4 prefixes	944,000	977,000 (+4%)	1,005,000 (+3%)	1,028,000 (+2%)	1,045,000 (+1.6%)	1,057,000 (+1.1%)	1,062,000 (+0.5%)
IPv6 prefixes (L)	172,000	199,000 (+16%)	226,000 (+14%)	252,000 (+11.5%)	279,000 (+10.7%)	305,000 (+9.3%)	332,000 (+8.8%)
IPv6 prefixes (E)	172,000	243,000 (+41%)	320,000 (+31%)	412,000 (+28.75%)	554,000 (+34.4%)	723,000 (+30.5%)	959,000 (+32.6%)
<b>Total (linear IPv6)</b>	<b>1,116,000</b>	<b>1,176,000 (+5%)</b>	<b>1,231,000 (+5%)</b>	<b>1,280,000 (+4%)</b>	<b>1,324,000 (+3.5%)</b>	<b>1,362,000 (+2.8%)</b>	<b>1,394,000 (+2.5%)</b>
<b>Total (exponential IPv6)</b>	<b>1,116,000</b>	<b>1,220,000 (+9%)</b>	<b>1,325,000 (+8.6%)</b>	<b>1,440,000 (+8.6%)</b>	<b>1,599,000 (+11%)</b>	<b>1,780,000 (+11%)</b>	<b>2,021,000 (+13.5%)</b>

While the predictions in [1] summarized in Table 2 are derived & based on recent growth rates, the underlying data analysis can be summarized that IPv4 growth rate is expected to slow down while IPv6 is expected to pick up in a bit higher rate, but overall that shows there is more than 5 years' of headroom before the total of IPv4 and IPv6 prefix announcements cumulatively exceeds 2 million prefixes, 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.

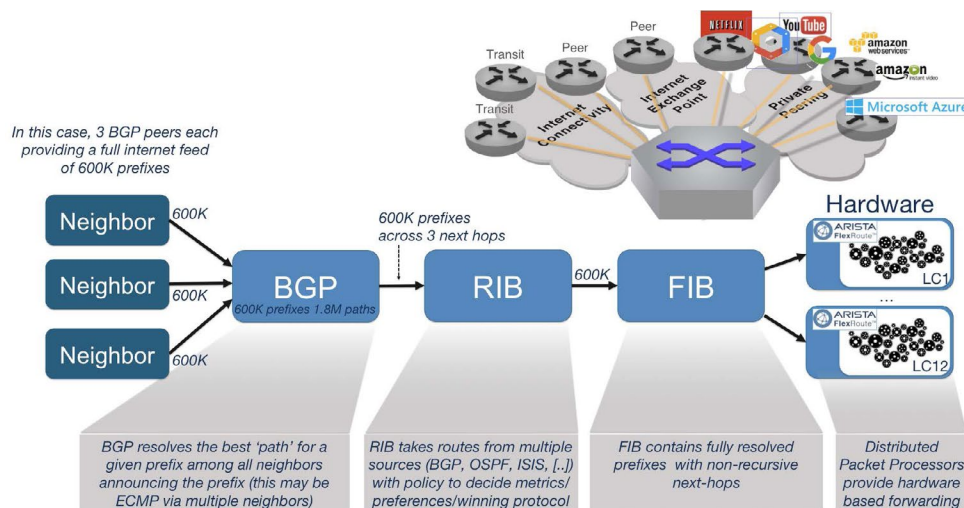


Figure 3: Prefixes received in BGP and their resolution from BGP to RIB to FIB

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.

The increased RIB scale does not necessarily mean an increase in the consumption of data plane HW resources, but it does impact overall convergence. FlexRoute provides a unique way for managing data plane HW resources and Arista's robust BGP stack and underlying EOS operating system architecture provides the necessary foundation for scale, convergence and availability.

## Flexible HW resources Allocations

The most common approach in networking silicon when it comes to managing HW resources is that a forwarding pipeline typically has predefined rigid boundaries for various tables that serve different lookups for packet forwarding.

For example, there is a HW table of a fixed size dedicated for LPM to do IP lookups, another table for MPLS lookups, another table for MAC address lookup & various tables for other lookup functions each with a fixed size. However, real world deployments are never identical to each other when it comes to what is expected from a device in a given role in a network. Some deployments focus on layer-2 services, so it would benefit from higher MAC address scale while other deployments could benefit from larger IP forwarding scale like internet peering applications.

Arista's R3/R3A series platforms optimize the HW resource allocations, by providing in EOS, a set of built-in profiles known as MDB profiles, MDB refers to Modular DataBase. It is a collection of HW tables that serve the different lookup functions expected from a router, however the MDB profiles adjust the HW table boundaries to allocate more resources to a given functionality that suits better the expected use case.

This capability of tuning the available hardware resources in the R3/R3A series has unlocked even more potential for FlexRoute. A deployment that requires the highest possible FIB capacity can benefit from this flexibility by allocating more hardware resources for layer-3 lookups and FlexRoute would utilize the algorithmic approach to maximize the utilization of the allocated resources.

### 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:

#### Real World Example #1: Internet Edge Router

In this deployment an R3 device is acting as an internet facing edge router, peering with few upstream neighbors that are advertising a full internet routing table - 10 copies in this example. The consumed resources included below are using an internet routing table snapshot from February 2023, which sums up to ~969K IPv4 prefixes plus ~179K IPv6 prefixes.

In addition to the internet routing table, there is a private peering VRF that is receiving additional private prefixes from upstream private peers. Those prefixes are also sent to a pair of route reflectors as VPNv4 & VPNv6 prefixes. There are a total of ~296K IPv4 prefixes plus ~80K IPv6 prefixes in this VRF. The prefixes are also received via 10 upstream BGP peers (i.e 10 copies).

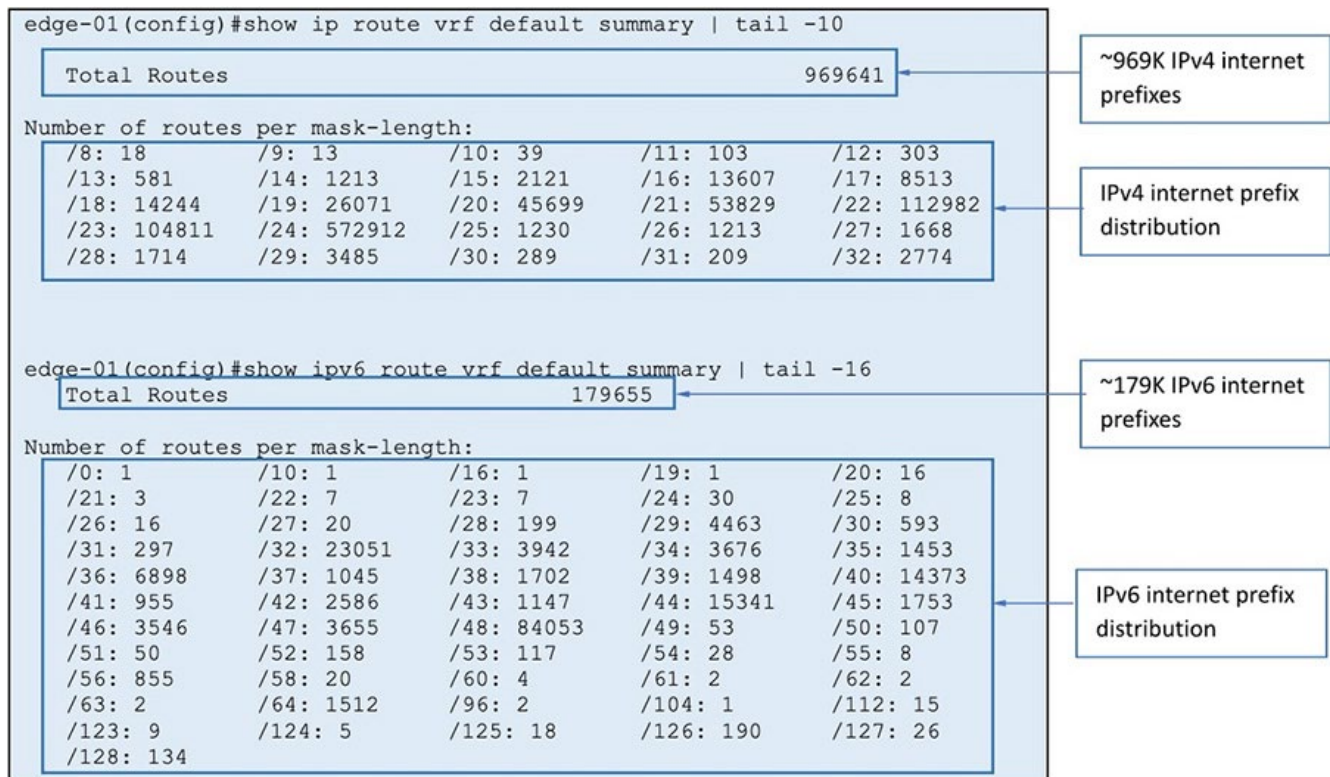


Figure 4: IPv4 routes in the FIB

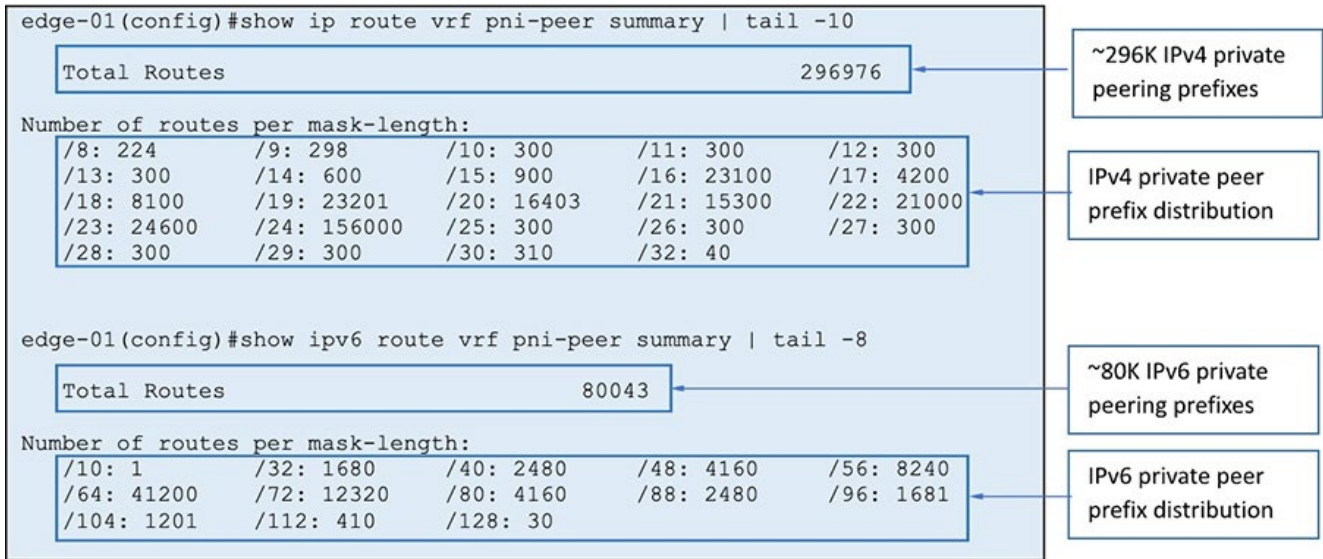


Figure 5: IPv6 routes in the FIB

Below snapshots show the number of BGP prefixes received from different BGP peers for both internet routing & private peering, for IPv4 & IPv6 respectively.

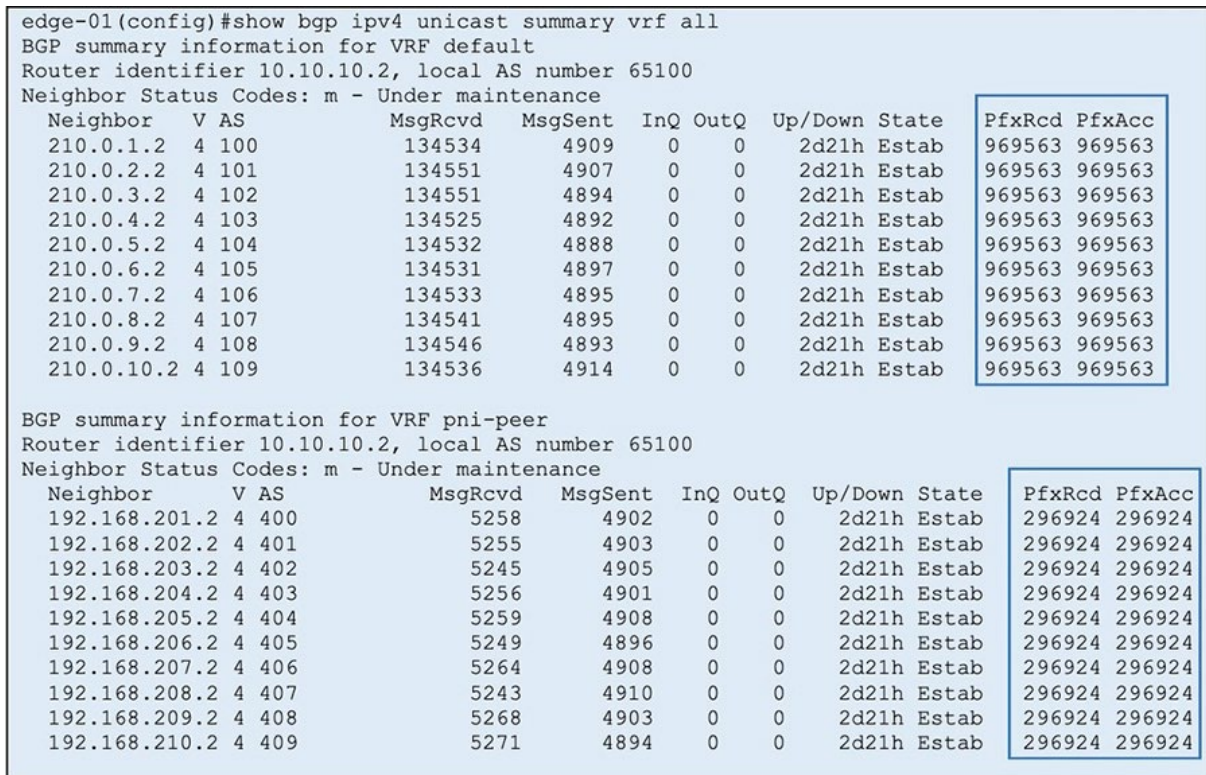


Figure 6: IPv4 BGP paths stored in the RIB

```

edge-01(config)#show bgp ipv6 unicast summary vrf all
BGP summary information for VRF default
Router identifier 10.10.10.2, local AS number 65100
Neighbor Status Codes: m - Under maintenance
Neighbor      V AS      MsgRcvd  MsgSent  InQ  OutQ  Up/Down  State  PfxRcd  PfxAcc
2001:fe0:cafe::2 4 100      46714    4923    0    0    2d21h  Estab  179610 179610
2001:fe1:cafe::2 4 101      46718    4919    0    0    2d21h  Estab  179610 179610
2001:fe2:cafe::2 4 102      46719    4910    0    0    2d21h  Estab  179610 179610
2001:fe3:cafe::2 4 103      46706    4915    0    0    2d21h  Estab  179610 179610
2001:fe4:cafe::2 4 104      46707    4912    0    0    2d21h  Estab  179610 179610
2001:fe5:cafe::2 4 105      46720    4921    0    0    2d21h  Estab  179610 179610
2001:fe6:cafe::2 4 106      46719    4909    0    0    2d21h  Estab  179610 179610
2001:fe7:cafe::2 4 107      46708    4919    0    0    2d21h  Estab  179610 179610
2001:fe8:cafe::2 4 108      46708    4915    0    0    2d21h  Estab  179610 179610
2001:fe9:cafe::2 4 109      46710    4902    0    0    2d21h  Estab  179610 179610

BGP summary information for VRF pni-peer
Router identifier 10.10.10.2, local AS number 65100
Neighbor Status Codes: m - Under maintenance
Neighbor      V AS      MsgRcvd  MsgSent  InQ  OutQ  Up/Down  State  PfxRcd  PfxAcc
2001:fe0:cafe::2 4 400      5106     4901    0    0    2d21h  Estab  80000 80000
2001:fe1:cafe::2 4 401      5104     4907    0    0    2d21h  Estab  80000 80000
2001:fe2:cafe::2 4 402      5101     4914    0    0    2d21h  Estab  80000 80000
2001:fe3:cafe::2 4 403      5112     4906    0    0    2d21h  Estab  80000 80000
2001:fe4:cafe::2 4 404      5108     4901    0    0    2d21h  Estab  80000 80000
2001:fe5:cafe::2 4 405      5104     4915    0    0    2d21h  Estab  80000 80000
2001:fe6:cafe::2 4 406      5106     4918    0    0    2d21h  Estab  80000 80000
2001:fe7:cafe::2 4 407      5092     4907    0    0    2d21h  Estab  80000 80000
2001:fe8:cafe::2 4 408      5104     4895    0    0    2d21h  Estab  80000 80000
2001:fe9:cafe::2 4 409      5113     4903    0    0    2d21h  Estab  80000 80000
    
```

Figure 7: IPv6 BGP paths stored in the RIB

In summary, this deployment has a total of ~1.26M IPv4 prefixes and ~260K IPv6 prefixes that combined add up to ~1.52M routes installed in the FIB, while the RIB has a total of ~12.6M IPv4 paths & ~2.26M IPv6 paths for a total of ~15.26M BGP paths.

The below snapshot shows the data plane HW resources consumed using the “show hardware capacity” command in Arista EOS operating system.

Table	Feature	Chip	Used	Used	Free
EcmpLevel1	Routing	Jericho2	0	0%	16384
EcmpLevel2	Routing	Jericho2	5	0%	8186
EcmpLevel3	Routing	Jericho2	0	0%	8192
FecLevel1	Routing	Jericho2	0	0%	104856
FecLevel2	Routing	Jericho2	85	0%	104767
FecLevel3	Routing	Jericho2	4098	3%	100758
Routing	Resource1	Jericho2	211	20%	813
Routing	Resource2	Jericho2	142	27%	370
Routing	Resource3	Jericho2	1387	22%	4757
Routing	Resource4	Jericho2	16038	48%	16730
Routing	Resource5	Jericho2	9033	55%	7351
Routing	Resource6	Jericho2	119088	51%	110288
Routing	V4Hosts		0	0%	104448
Routing	V4Routes		0	0%	393158
Routing	V6Hosts		0	0%	98289
Routing	V6Routes		0	0%	98289

Figure 8: HW resource utilization with default MDB profile

The highest resource consumed in this case is at 55% which leaves more than enough space to hold future increase in internet prefixes. The highlighted capacities in this example are before enabling FlexRoute.

### Real World Example #2: Internet Edge Router - flexible resource allocation

The internet edge router example deployment shown here is mostly layer-3 deployment with minimal MPLS & layer-2 requirements. A typical edge device will have its overall HW resources pre-allocated to accommodate variable possible use cases but in this case allocating a large portion of the HW resources for MPLS might not be needed. This is where flexible resource allocations in EOS & R3 platforms via MDB profiles becomes a benefit.

Since the deployment is mostly layer-3, we can choose to configure the L3-XXXL MDB profile to re-allocate more resources to be used as Routing resources. The new HW resources consumed for the same deployment with the new profile is shown below.

```
edge-01(config)#show platform sand mdb profile
System profile: l3-xxxl
```

```
edge-01(config)#show hardware capacity | egrep "Routing|Table|--"
```

Table	Feature	Chip	Used	Used	Free
EcmpLevel1	Routing	Jericho2	0	0%	16384
EcmpLevel2	Routing	Jericho2	5	0%	8186
EcmpLevel3	Routing	Jericho2	0	0%	8192
FecLevel1	Routing	Jericho2	0	0%	52426
FecLevel2	Routing	Jericho2	85	0%	52339
FecLevel3	Routing	Jericho2	4098	7%	48330
Routing	Resource1	Jericho2	190	18%	834
Routing	Resource2	Jericho2	129	25%	383
Routing	Resource3	Jericho2	1188	19%	4956
Routing	Resource4	Jericho2	13593	41%	19175
Routing	Resource5	Jericho2	7959	48%	8425
Routing	Resource6	Jericho2	120780	46%	141364
Routing	V4Hosts		0	0%	104448
Routing	V4Routes		0	0%	786374
Routing	V6Hosts		0	0%	104448
Routing	V6Routes		0	0%	196593

Figure 9: HW resource utilization with L3-XXXL MDB profile

Enabling the new profile, the highest resource consumed for Routing decreased from 55% to 48%. This reduction in utilization is due to increased HW resources allocated for routing. One more resource to highlight is the "v4Routes" resource, which has its allocated capacity increased, compared to the default MDB profile shown earlier. This resource hasn't been utilized in this example yet but the next example shown demonstrates the utilization.

### Real World Example #3: Internet Edge Router - FlexRoute

For the same deployment example used here, FlexRoute can be enabled to optimize the HW resources consumption used for Routing.

FlexRoute can be turned on to target optimizations for specific prefix-lengths. Additionally as explained earlier, FlexRoute can be turned on with internet prefix distribution. EOS has a built-in "internet profile" that is tuned towards real internet prefix distribution. It targets a specific set of prefix lengths that are the most dominant in the current internet routing table & apply the optimizations on those specific prefix lengths.



```
edge-01(config)#show running-config | grep optimize
ip hardware fib optimize prefixes profile internet
```

```
edge-01(config)#show hardware capacity | egrep "Routing|Table|--"
```

Table	Feature	Chip	Used	Used	Free
EcmpLevel1	Routing	Jericho2	0	0%	16379
EcmpLevel2	Routing	Jericho2	5	0%	8186
EcmpLevel3	Routing	Jericho2	0	0%	8192
FecLevel1	Routing	Jericho2	0	0%	52421
FecLevel2	Routing	Jericho2	85	0%	52339
FecLevel3	Routing	Jericho2	4	0%	52424
Routing	Resource1	Jericho2	90	8%	934
Routing	Resource2	Jericho2	54	10%	458
Routing	Resource3	Jericho2	514	8%	5630
Routing	Resource4	Jericho2	5152	15%	27616
Routing	Resource5	Jericho2	3261	19%	13123
Routing	Resource6	Jericho2	45516	17%	216628
Routing	V4Hosts		0	0%	104448
Routing	V4Routes		487922	62%	298452
Routing	V6Hosts		0	0%	74613
Routing	V6Routes		0	0%	74613

Figure 10: HW resource utilization with FlexRoute enabled

After enabling FlexRoute with current routing scale, EOS has triggered the algorithmic optimizations to pick the relevant prefixes, pack them efficiently based on the learned prefixes, then program them to the relevant routing resources in the hardware. As shown in the table above, the highest resource utilization decreased from 48% to 19%! The V4Routes resource is being utilized by FlexRoute. This allows for a substantial headroom for future growth and preserves customer investment in key deployments.

### Arista EOS, State and NetDB

At the core of the Arista R3/R3A 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 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 fast convergence in mind.

As networks evolve, state streaming has become a necessity in today's networks. External SDN controllers are utilized by large network operators to gather network & routing table state in real-time and execute network wide decisions that get programmed in the routers. EOS has been built around programmability and state sharing allows for unlocking such use cases at scale.

## Conclusion

Arista's FlexRoute Engine provides support for scale beyond 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 5 million routes. The innovative FlexRoute Engine with its patented algorithmic approach to building layer-3 forwarding tables on R3/R3A platforms is unique to Arista and a key enabler for customers building their next generation networks.

## References

[1] BGP in 2022 – The Routing Table, Geoff Huston (APNIC): <https://labs.apnic.net/index.php/2023/01/09/bgp-in-2022-the-routing-table/>

### 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](mailto: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, 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



Copyright © 2023 Arista Networks, Inc. All rights reserved. CloudVision, and EOS are registered trademarks and Arista Networks is a trademark of Arista Networks, Inc. All other company names are trademarks of their respective holders. Information in this document is subject to change without notice. Certain features may not yet be available. Arista Networks, Inc. assumes no responsibility for any errors that may appear in this document. August 2023 02-0064-03