Modernizing the BGP Route Reflection Architecture

Achieving Convergence and Service Optimization Through Virtual Route Reflectors

A technical paper prepared for presentation at SCTE TechExpo24

Mark Goodwin
Lead IP Design Engineer
Cox Communications, Inc.
Mark.Goodwin@cox.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

1. Introduction

The Border Gateway Protocol (BGP) is a critical routing protocol used in large-scale networks, including the Internet. From the viewpoint of an organization's networking, BGP serves both internal and external purposes. External BGP (EBGP) is a prevalent method used for establishing connections to networks outside of your organization, serving the external use case. Conversely, for internal use, the Internal BGP (IBGP) is commonly adopted. This document will concentrate on the modern approach for scaling IBGP networks from the perspective of the architecture, the platform, and the routing.

The traditional method for building IBGP networks incorporates the IBGP full mesh model. This model is an excellent strategy for networks ranging from small to medium size. In this framework, all IP routing data is disseminated among all devices within the IBGP network. Essentially, each device maintains a logical connection or IBGP session with every other device. The total connectivity for the IBGP model is calculated using the formula $\frac{N(N-1)}{2}$ which is referred to as the n squared formula. Deployments with the IBGP full mesh model offer numerous benefits.

As the IBGP network expands, however, it faces several challenges. A fully distributed or regionalized route reflection model for IBGP is proposed to address these challenges, offering enhanced efficiencies and scalability for the IBGP network. This paper will outline the challenges associated with large-scale IBGP and illustrate how the route reflection model effectively overcomes these challenges.

The traditional platforms for IBGP network consist of routing devices that serve as the network nodes. These nodes boast considerable computational power, but this capacity must be distributed among various processes within the routing device. As the number of network devices escalates and the network's complexity intensifies, it imposes a significant burden on the computational power, especially the processes designated for BGP processing. This paper proposes the use of x86 servers to handle the heavy workload of BGP processing. X86 based servers have significantly more computational power than the router platforms and offer substantial computational resources for BGP scalability. This paper will depict the x86 based server as a streamlined approach for achieving high BGP scalability.

In the IBGP full mesh network, routing is optimized as each node possesses a comprehensive map of the IBGP network and employs the shortest path computation from the IGP for efficient routing through the IBGP network. However, updates or modifications to the IBGP full mesh architecture can often lead to suboptimal routing. This paper will delve into some of the corner cases where this may occur and propose relevant solutions to preserve optimal routing.

This paper explores a modern approach to scaling IBGP networks, focusing on design principles and optimization for platform, architecture, and routing. We delve into the architectural considerations and propose enhancements to improve scalability and create efficiencies.

2. The I-BGP Full Mesh Design

The I-BGP full mesh is a widely adopted practice in traditional service provider networks. These networks typically consist of a core backbone network interconnected with edge metro networks that serve various subscriber markets. In this design, each node (router) within the network establishes a BGP session with every other node. This full mesh requirement complements the default behavior of BGP routing, which restricts I-BGP routers from advertising learned routes to other I-BGP peers. By maintaining direct sessions, the network prevents routing loops and ensures operational efficiencies, including load balancing, optimized routing, and rapid convergence.

The BGP multipath feature is essential within the iBGP full mesh design. Once the Interior Gateway Protocol (IGP) identifies equal paths to a pair of loopback addresses (typically associated with the participating routers), BGP enables equal-cost load balancing for traffic between the border devices (located at the network boundary) and the edge devices (serving subscriber markets). With each router having a complete view of BGP routing information, all available paths to external destinations are known. Load balancing ensures that routers distribute traffic equally across multiple paths, providing redundancy and fault tolerance. This critical feature improves resource utilization and guarantees high reliability—when one path fails, traffic automatically reroutes to an alternative path.

Given that all devices participate in a full mesh of iBGP sessions, routing is fairly optimized within the iBGP design. To illustrate, consider a simple example with three peering routers at the network border (Los Angeles, Dallas, and Miami) and an edge router serving a subscriber market in Atlanta. Common content providers exist in each peering center. The desired behavior for inbound traffic to the Atlanta subscriber market is to receive the traffic from the Miami peering center—this route offers the most efficiency from a latency and routing perspective. This behavior is inherent within the iBGP design, as the routers in the Atlanta subscriber market receive the same content provider routes from each peering center and resolve the next hop to the closest geographically located peering center based on IGP metrics, which in this case would be the Miami location. From an operational perspective, this approach is highly efficient.

One final benefit of the iBGP full mesh design is its rapid convergence during failures. In this design, routers can swiftly adjust their routing tables due to direct exchanges of updates with all peers. This stands in contrast to alternative iBGP designs that introduce intermediary hops. Additionally, with multiple active paths in the Forwarding Information Base (FIB), immediate traffic switchover occurs upon detecting a node or link failure. The event-driven mechanisms within the iBGP full mesh—such as direct peer-to-peer connections and seamless interaction with the underlying IGP—minimize disruptions during network failures. The IBGP full mesh is illustrated in Figure 1 below.

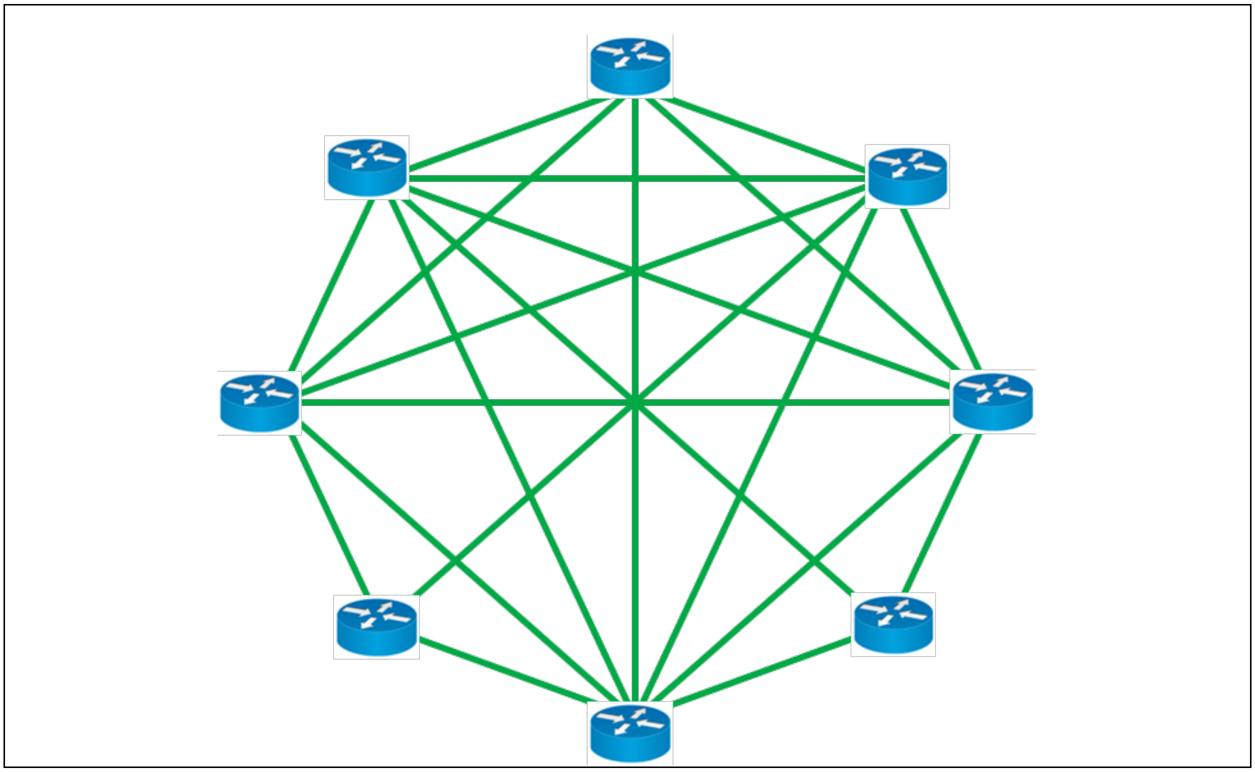

Figure 1 - IBGP Full Mesh Design

3. I-BGP Full Mesh Scalabliity Concerns

As Service Provider networks expand, certain challenges can diminish the benefits of an I-BGP full mesh. The scalability of I-BGP networks is quantified by the formula N(N-1)/2, commonly referred to as the n-squared problem in large networks. This issue arises due to the exponential growth in the number of I-BGP connections as the network size increases. For example, consider a network with 8 routers. In an I-BGP full mesh, the total number of connections would be 24. However, if the network were to expand to 12 routers, the total connections would increase to 66—a substantial 175% increase! Such rapid growth becomes cost-prohibitive in terms of router resources.

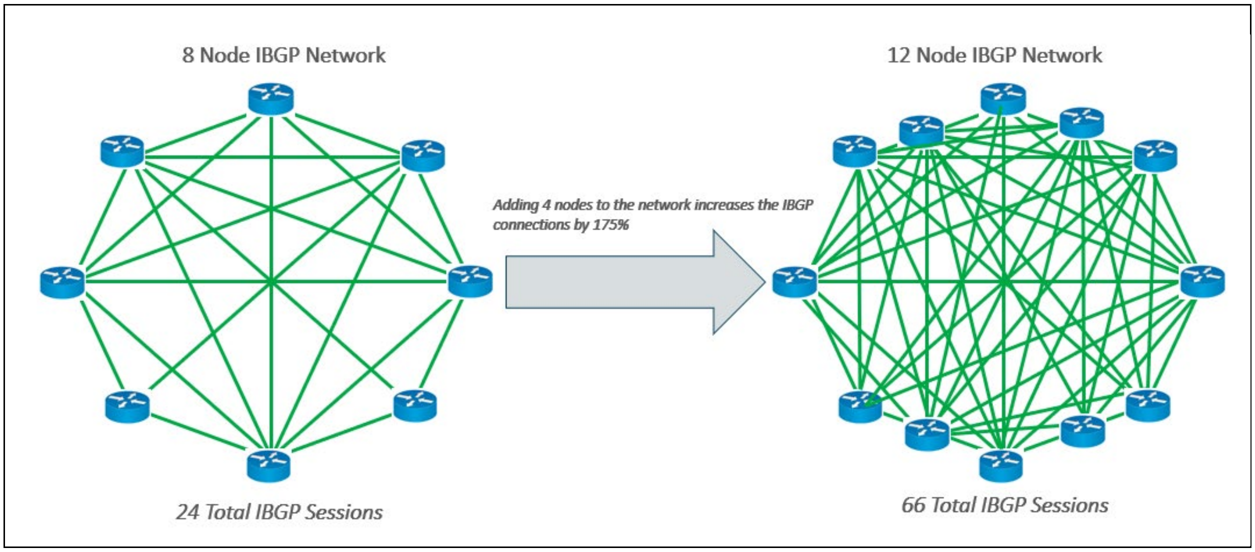

Figure 2 - IBGP Scalability Concerns

As Service Provider networks expand, the increasing number of I-BGP connections places additional demands on router resources, particularly the CPU and memory. These resources play a critical role in handling essential BGP tasks such as control plane processing, managing BGP update messages, and maintaining peer status with other routers. However, it's important to note that these resources are not exclusively allocated to BGP processing; they must also handle other network functions. Consequently, in a large network with a high volume of BGP connections, this resource sharing can lead to suboptimal performance for BGP processes. As an example, when CPU cycles and memory are divided among various networking protocols and tasks, contention arises which results in BGP processes experiencing delays or inefficiencies due to resource competition. Also, slower convergence times during route changes can result from resource limitations, affecting overall network stability and responsiveness.

Despite significant progress in router platform technology, these platforms are still not fully optimized for efficient processing of large-scale I-BGP networks. As networks expand, the demand for robust BGP solutions grows. To address this, we propose a hierarchical route reflection model that leverages x86-based servers as BGP router reflectors. By distributing the load and enhancing scalability, this approach aims to improve BGP performance in large networks. The hierarchical model will organize BGP route reflectors into tiers which reduce the number of direct peer connections and enhance manageability. The use of x-86 based servers will provide flexibility, cost-effectiveness, and the ability to handle BGP processing efficiently.

4. BGP Architecture Optimization with Hierarchical Route Reflection

The scalability challenges posed by the iBGP full mesh can be effectively tackled with a hierarchical route reflection architecture. This architecture introduces new BGP attributes, namely the originator ID and cluster list, which are instrumental in preventing routing loops. As a result, the stringent requirements of BGP routing are relaxed, eliminating the need for full mesh connectivity between all iBGP nodes. This leads to a significant reduction in the total iBGP peering sessions across the network.

To illustrate, consider an iBGP network with 12 nodes. In this scenario, the total peering sessions would amount to 66. However, if this network were to be transformed into a route reflection architecture with 2 centralized route reflectors, the total number of peering sessions would be reduced to 25. This equates to an approximate reduction of over 62% in the total number of peering sessions to manage.

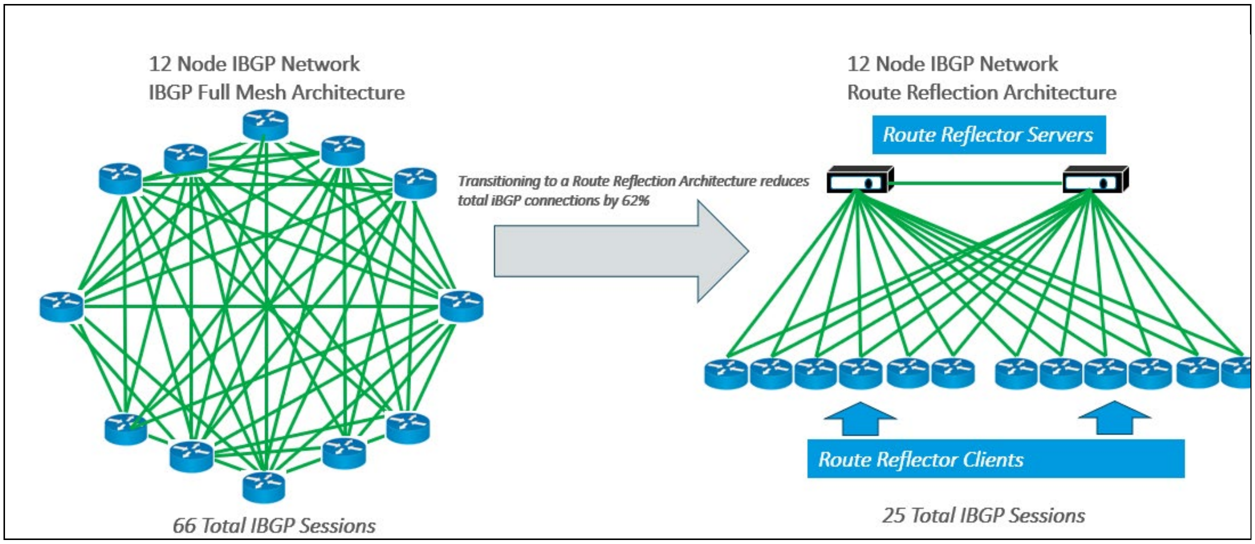

Figure 3 - Route Reflection Architecture Reduces IBGP connections.

Moreover, the route reflection architecture requires fewer overall routes to be maintained in each router's Routing Information Base (RIB). For instance, in a network with 8 nodes where each node is advertising 10K identical routes, an iBGP full mesh would necessitate each node to maintain 7 copies of the route, or 70,000 routes in its local RIB. By transitioning to the route reflection architecture, this number is reduced to a single copy of the route, or 10,000 routes in its local RIB.

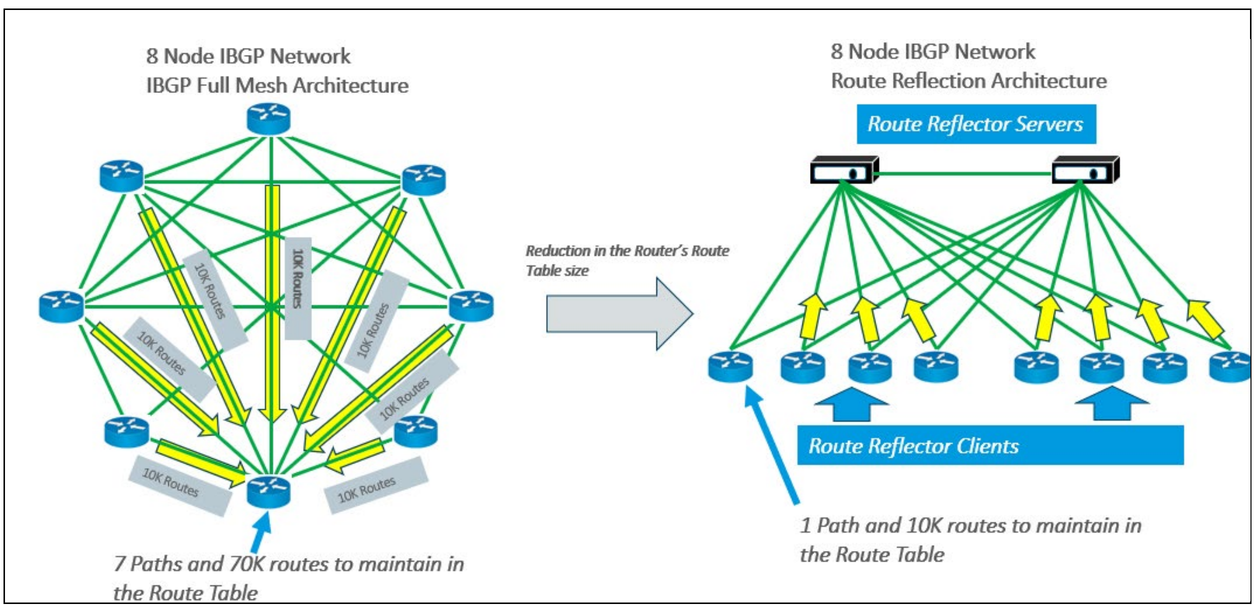

Figure 4 - Route Reflection Architecture Reduces Route Table size.

These examples underscore the efficiency brought about by the route reflection architecture. There are three potential options for deploying this architecture, each with its own set of advantages and considerations. It's important to evaluate these options in the context of the specific network requirements and constraints.

4.1. Centralized Route Reflection Architecture

The centralized route reflection architecture is a design strategy that advocates for a single pair of route reflectors housed in a centralized data center for the iBGP network. This model offers the advantage of cost-effectiveness, owing to the reduced hardware and operational costs associated with managing only a single pair of route reflectors. Additionally, it simplifies operations and retains all the benefits of a hierarchical route reflection architecture.

However, this model does present a potential risk in the form of a single point of failure. Therefore, it's crucial to have a robust disaster recovery plan in place to address any disruptions that might occur due to a failure in the primary data center. One effective strategy to mitigate this risk is to establish a pair of backup route reflectors in a separate data center. This ensures network resilience and continuity, even in the event of unforeseen circumstances affecting the primary data center.

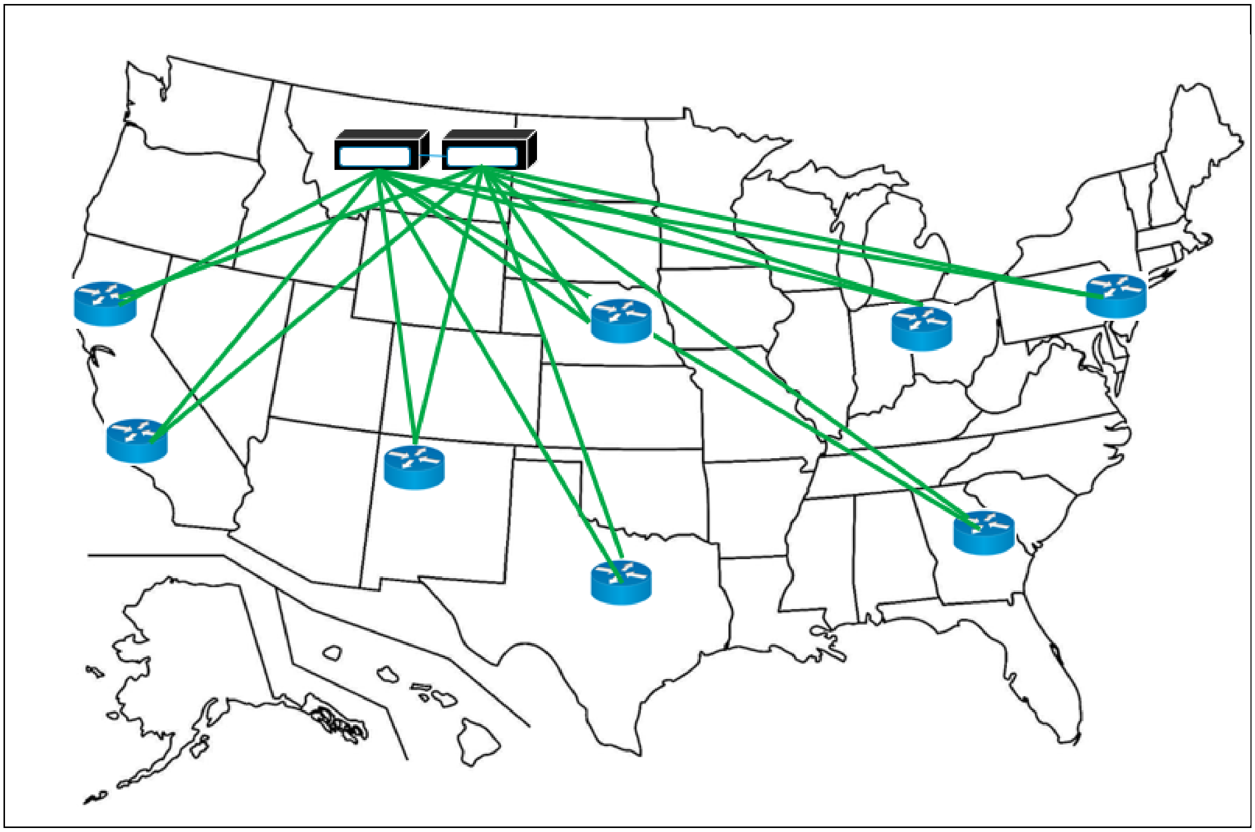

Figure 5 - Centralized Route Reflection Architecture

4.2. Regionalized Route Reflection Architecture

The Regionalized Route Reflection Architecture is a strategic design approach that suggests the deployment of a pair of route reflectors in two or more geographically redundant data centers. This design strategy is a cost-effective solution that provides high resilience in the event of a failure, ensuring uninterrupted network service.

In this architecture, the iBGP full mesh established between the geographically redundant route reflectors which ensures optimal path selection and prevents routing loops.

The edge devices in the subscriber markets are configured as route reflector clients. These edge devices, which could be routers or switches, connect the subscriber markets to the network. They receive and propagate routing information from their associated route reflectors, thus playing a crucial role in maintaining the network's routing efficiency and stability.

This architecture allows for efficient routing information exchange and provides a robust failover mechanism, making it an attractive option for network design in large-scale, distributed environments. It balances the need for network resilience and cost-effectiveness, making it a viable choice for many organizations.

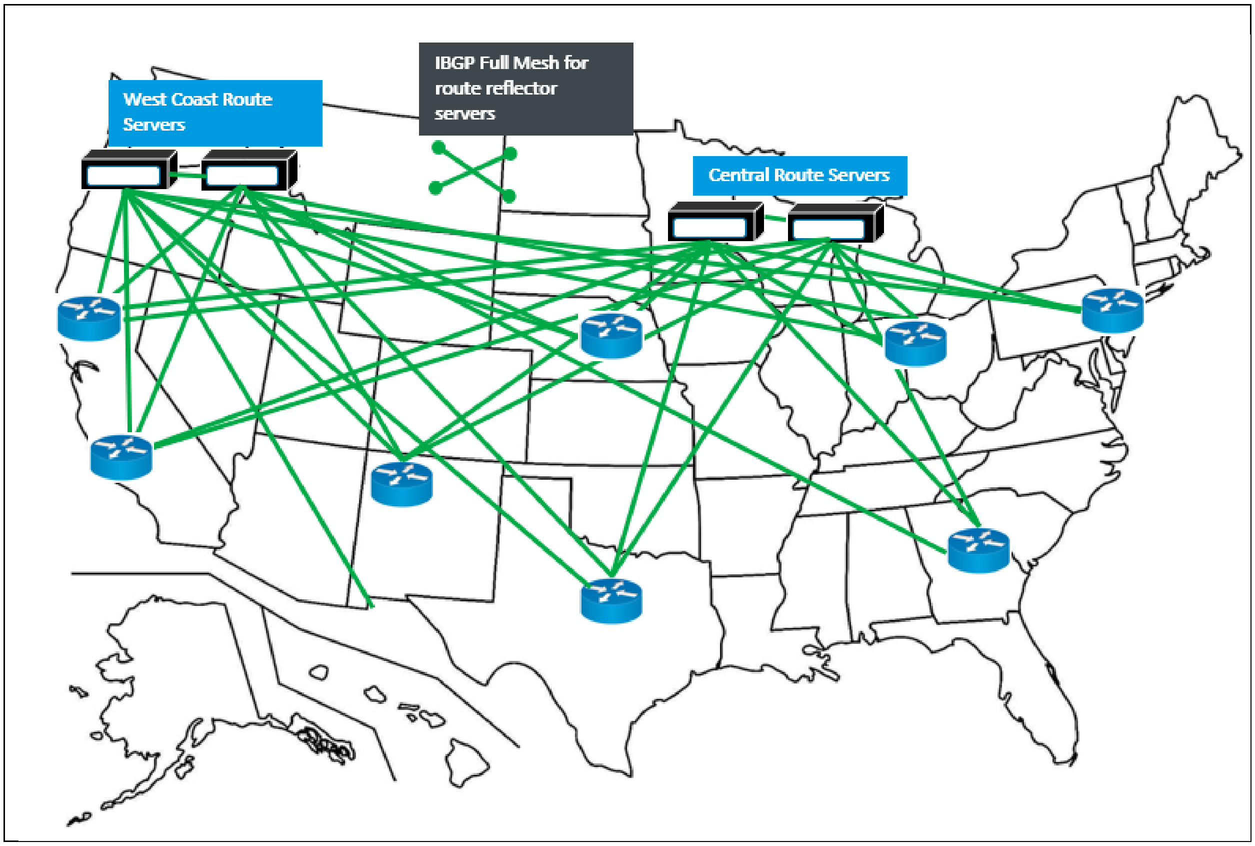

Figure 6 - Geographically Redundant Route Reflection Architecture

4.3. Fully Distributed Route Reflection Architecture

This strategy suggests the addition of a pair of route reflectors to the edge devices at each of the subscriber markets. It necessitates the establishment of full mesh iBGP sessions between the route reflectors, with the edge devices functioning as route reflector clients. This approach offers enhanced service flexibility and closely emulates the iBGP full mesh architecture, which is currently prevalent in many network designs.

Moreover, this design ensures efficient convergence times due to the full mesh configuration of the route reflectors. This results in rapid propagation of routing information across the network, thereby minimizing downtime and enhancing network performance.

However, this option does come with its own set of challenges. It presents a high operational and financial cost. The implementation of this strategy will effectively double the number of devices in the network, leading to an increase in the operational workload. This could potentially strain resources and require additional investment in network management and maintenance.

Furthermore, the increased complexity of the network could also lead to a higher likelihood of configuration errors and potential network vulnerabilities. Therefore, while this option offers numerous benefits, these factors must be carefully considered during the network design and planning phase.

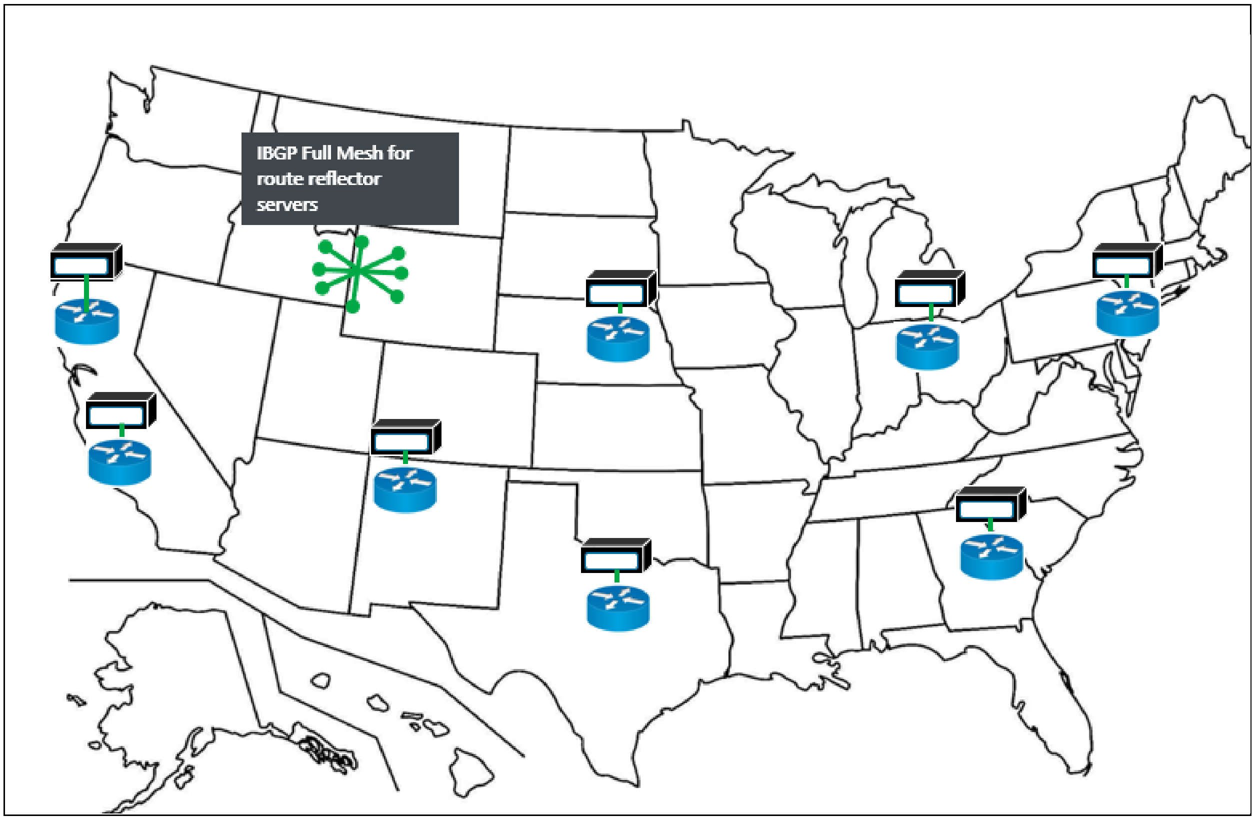

Figure 7 - Fully Distributed Route Reflection Architecture

5. Platfrom Optimization with Virtualized Route Reflectors

5.1. The Virtualized Route Reflectors offer increased compute power

The x86 servers employed for virtual route reflectors offer a significant increase in compute power compared to the control plane cards found in even the most advanced router control plane modules. For instance, common edge routing devices from leading industry vendors typically feature a CPU with up to an 8-core 2.7GHz configuration.

In contrast, a standard build of a Dell 750 server utilizes two Intel Xeon CPUs, each boasting 40 cores. This not only allows for parallel processing but also represents a substantial increase in CPU performance over traditional routing platforms. It’s worth noting that there are CPUs available in the market that offer even more power than the ones used in the Dell 750 server build.

Moreover, industry routing vendors offer virtualized implementations of their platforms. These implementations leverage the power of multiple CPUs, with each CPU core dedicated to the parallel processing of Border Gateway Protocol (BGP) updates. This approach optimizes the handling of BGP updates, enhancing the overall efficiency and performance of the network.

5.2. The Virtualized Route Reflectors offer increased scalability

Modern router platforms can hold a substantial number of routes in their route tables, typically exceeding 20 million. However, this is a finite resource. Due to the optimized data structures that routers employ to achieve faster convergence times; they cannot utilize all available memory. This limitation is a trade-off for the speed and efficiency of network operations.

In contrast, virtualized route reflectors, which run on server platforms, are optimized differently. They do not have the same constraints as those seen on router control plane cards. These virtualized systems leverage the extensive computational resources and memory management capabilities of modern servers. As a result, they can scale to accommodate a significantly larger number of routes, reaching into the tens of millions.

This scalability of virtualized route reflectors provides a distinct advantage in large-scale network environments. It allows for more extensive and complex routing operations without compromising on performance or efficiency. Furthermore, the use of server-based systems for route reflection introduces the benefits of server technology, such as higher processing power and more efficient memory usage, into the network routing domain. This fusion of technologies is a testament to the ongoing evolution of network infrastructure, offering increased scalability and improved network performance.

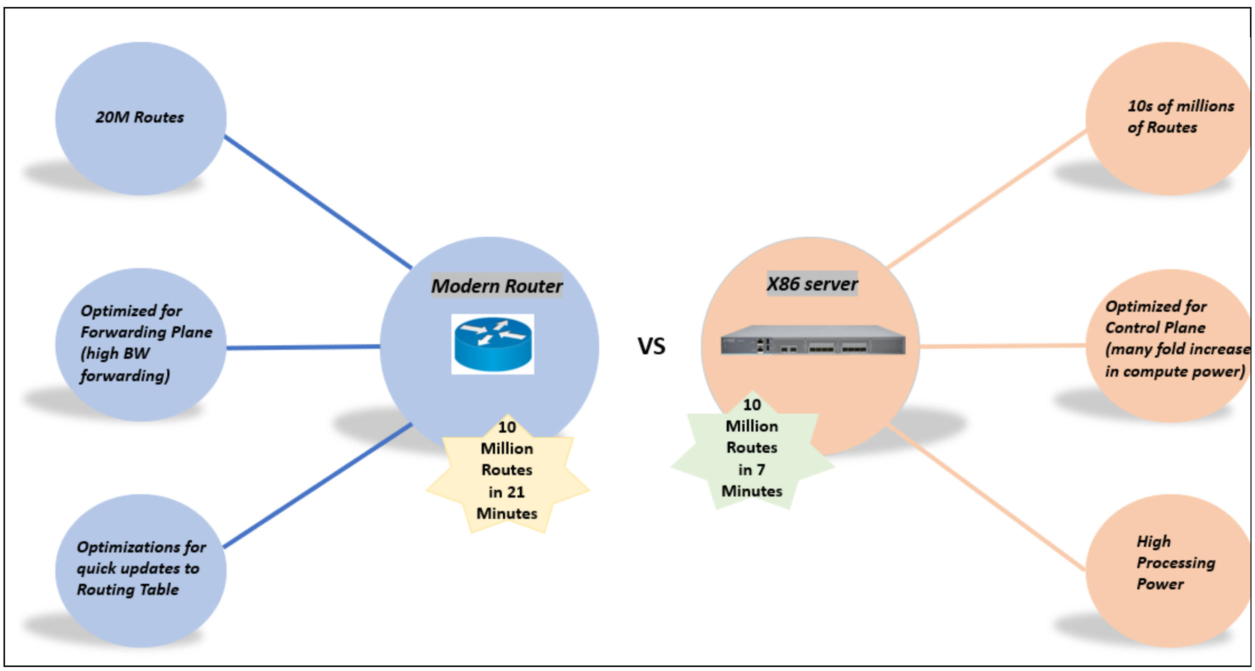

Figure 8 - Route Reflector Platform Optimization with Virtual Route Reflectors

6. Routing Optimization with virtualized Route Reflection

6.1. Use BGP Add-Path for Load Balancing

One crucial consideration when transitioning from a full iBGP mesh to a virtualized route reflection architecture is the potential loss of load balancing for equal-cost network links. BGP route reflectors typically select the single best path from the available paths and only reflect this best path to the route reflector clients. This can lead to suboptimal routing and slower convergence times.

For instance, consider a scenario where redundant edge routers advertise the same route to a virtual route reflector. The route reflector, using its best path selection process, will choose only one of these routes and then re-advertise it to the border routers at the network's edge. Ideally, the border router should load balance incoming traffic to the pair of redundant edge routers. However, if it receives a single BGP path for the route, it will only forward traffic to one router, leading to an imbalance.

This issue is addressed with the BGP Add-Path capability. This feature allows BGP to advertise not just the best path downstream, but also a user-configured number (N) of the best paths. In the scenario described above, the virtual route reflector would advertise both the best path and the second-best path to the border routers. As a result, the border routers would receive both paths and continue to load balance traffic based on the multipath configuration to the egress edge routers.

This approach effectively mitigates the routing limitations of the route reflection, enabling more efficient use of network resources and improving overall network performance. However, it's important to note that the implementation of BGP Add-Path requires careful planning and configuration to ensure optimal results.

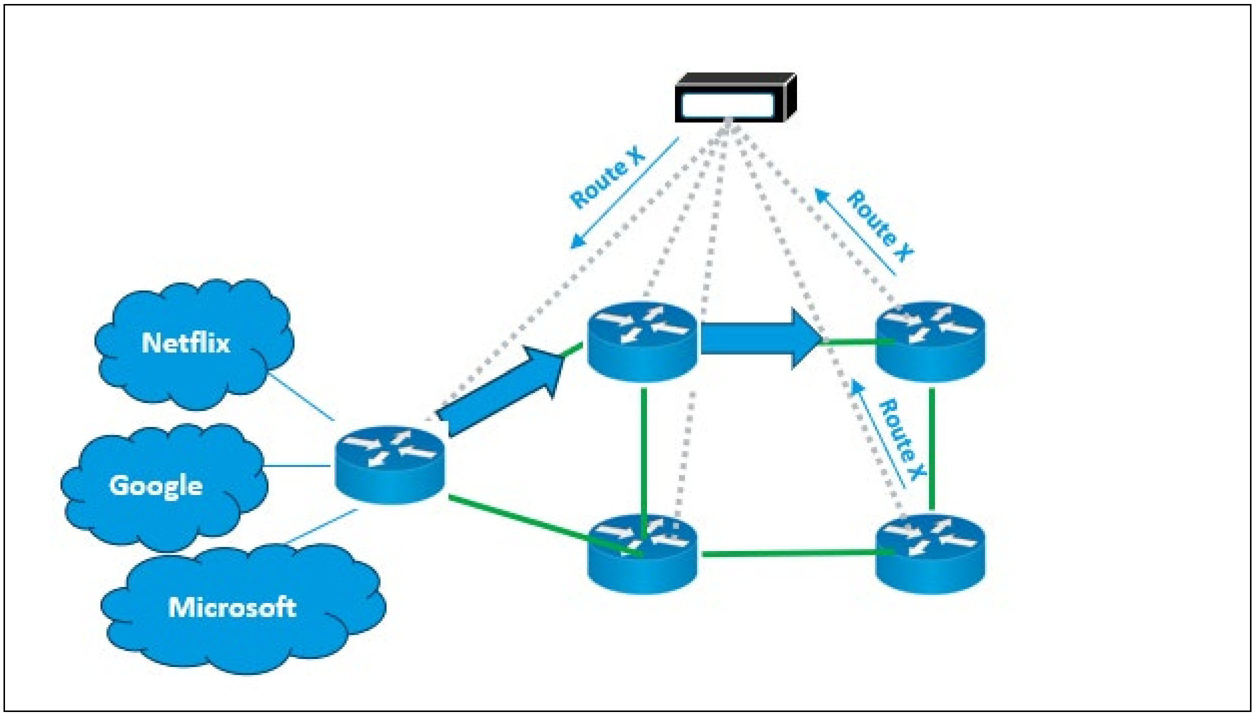

Figure 9 - Default Behavior of Route Reflector will impact Load Balancing

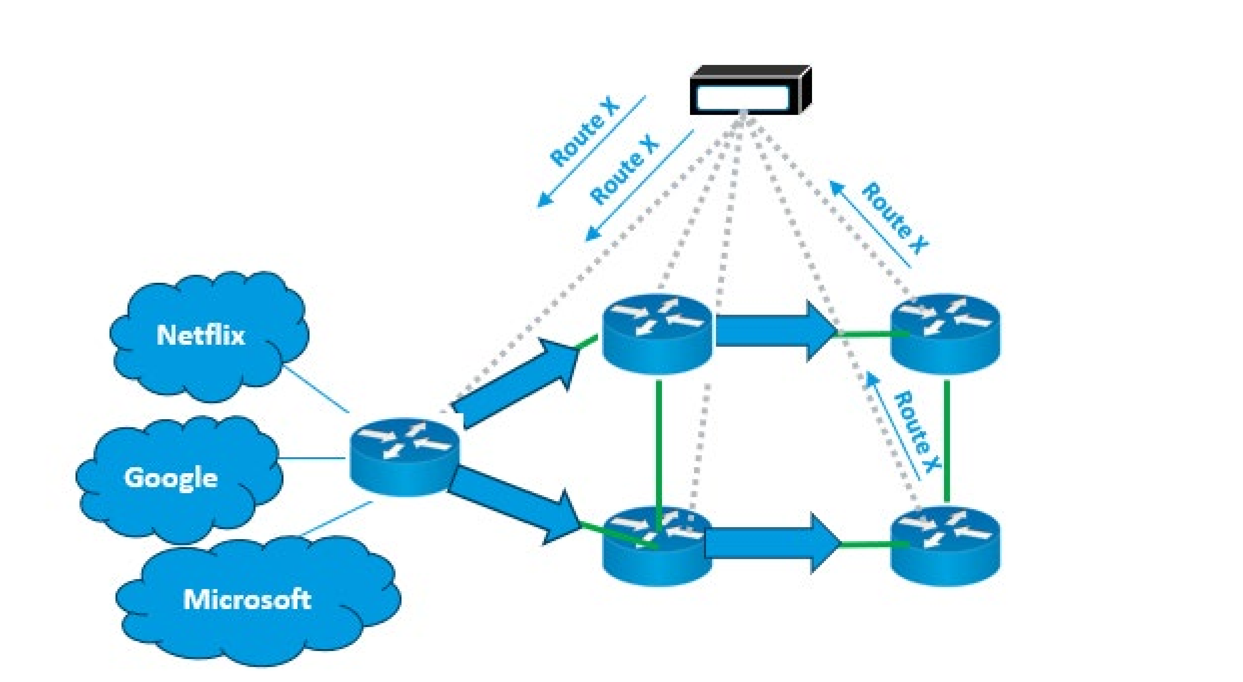

Figure 10 - Use BGP Add-Path to maintain Load Balancing

7. Conclusion

In conclusion, the transition towards a virtualized route reflection architecture signifies a pivotal advancement in the realm of network technology. This shift offers a multitude of benefits, including increased route scalability and compute power, which are essential for handling the growing demands of modern networks.

Moreover, this architecture enhances service convergence, enabling a more efficient and streamlined network operation. By allowing for the simultaneous processing of multiple routes and leveraging the power of modern servers, it significantly improves network performance.

Furthermore, the use of virtualized route reflectors introduces a new level of flexibility and adaptability into network design. It allows for more efficient use of network resources, better management of network traffic, and improved resilience against network failures.

This evolution not only optimizes current network operations but also paves the way for future innovations in network technology, setting a new standard for network performance and reliability. As we continue to push the boundaries of what's possible, the virtualized route reflection architecture stands as a beacon of progress in our ongoing journey to redefine network technology.

Abbreviations

BGP	Border Gateway Protocol
EBGP	Exterior Border Gateway Protocol
IBGP	Interior Border Gateway Protocol
RR	Route Reflector
IGP	Interior Gateway Protocol
FIB	Forwarding Information Base
RIB	Route Information Base
CPU	Central Processing Unit

Bibliography & References

https://www.cisco.com/c/en/us/products/collateral/routers/asr-9000-series-aggregation-services-routers/asr-9000-series-rsp5x-ds.html

https://www.juniper.net/us/en/products/routers/mx-series/mx2000-universal-routing-platforms-datasheet.html

https://i.dell.com/sites/csdocuments/Product_Docs/en/poweredge-R750-spec-sheet.pdf

https://www.ece.nus.edu.sg/stfpage/bsikdar/papers/gbcom04s.pdf

Understanding BGP Route Reflection: Simplifying Scalable Network Designs - j2sw Blog