A Comparison of Scaling Techniques for BGP

Rohit Dube
High Speed Networks Research
Room 4C-508, Bell Labs, Lucent Technologies
101 Crawfords Corner Road, Holmdel, NJ 07733
Email: rohitd@dnrc.bell-labs.com

Abstract

BGP is **the** inter-domain routing protocol used in the Internet today. During the course of its evolution, the Internet has gone from being a simple and small network to one that is run at its core by large service providers constantly battling with bigger and bigger topologies forcing the routing community to invent ways of scaling both interior and exterior routing protocols. *Route-reflectors* and *confederations* have turned out to be the weapons of choice in scaling BGP to these large topologies. This paper takes a close look at these two mechanisms and seeks to compare them.

1 Introduction

The Border Gateway Protocol (BGP) [1], [2], [3] is the pervasive inter-domain routing protocol in the Internet today. Before the recent explosive growth of the service providers topologies, BGP was typically used in a configuration where all the border routers imported routes from external Autonomous Systems (ASes) and then distributed them to all the routers within their own AS. This distribution was accomplished using a full-mesh of Internal-BGP (IBGP) peerings amongst all the routers in the AS. Once this flat topology hit the scaling limit (both administrative and the cpu/memory ceiling), mechanisms were devised to reduce the number of peering sessions per router. There are three such mechanisms deployed in the Internet today - route-reflectors [4], confederations [5] and route-servers [6]. Of these three, route-reflectors and confederations are the dominant mechanisms having been implemented by multiple vendors and deployed by the biggest Internet and Network Service Providers (ISPs and NSPs). In this paper we analyze and compare route-reflectors and confederations. We start by describing these mechanisms followed by a detailed comparison. We conclude with a summary of our observations and pointers for future work.

2 IBGP, Route-reflectors and Confederations

Consider the scaled down BGP topology depicted in figure 1. Routers *r1* through *r6* form an ISP backbone. In order to provide consistent loop free routing, each of these routers maintains IBGP peering sessions with all the others. When one of these routers learns a prefix, say from an External-BGP (EBGP) peer, it runs the BGP decision algorithm and installs the best route to the prefix into its routing table. If this best route is in turn *not* learnt from an IBGP peer, the prefix is propagated to all the IBGP peers. In the stable state this provides all the BGP routers in the network with all possible routes to a prefix. (Further details on this can be found elsewhere in the literature [1], [2], [3]).

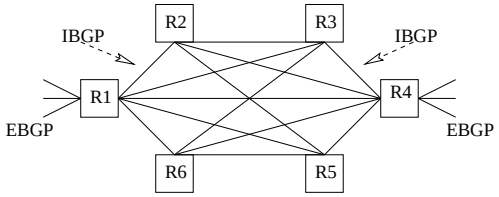

Figure 1: Full-mesh IBGP

Given that each BGP router has to peer with all the BGP routers within the AS, it is easy to see that as the number of BGP routers in the AS grows, the number of peering sessions each router needs to maintain increases to $(n-1)$ with a total of $n \times (n-1)/2$ IBGP peerings in the network, where n is the number of BGP routers. Maintaining these peering sessions gets quickly out of hand with increasing n, both for the network administrators and the the router hardware.

2.1 Route-reflection

Route-reflectors tackle this scaling problem by dividing the IBGP topology into *clusters*. A cluster consists of one or more BGP routers acting as server(s) and the remaining as client(s). The servers are fully meshed with each other and also have peering sessions with all the clients. The clients may or may not peer with each other. Further, clients in a cluster can act as servers for *sub-clusters* provided a strict ancestor-descendant relationship is maintained between the cluster and its sub-clusters and that the servers of all the sub-clusters of a cluster are fully-meshed in a peer-peer relationship. The sub-clusters can have their own sub-sub-clusters and so on. Note that the servers of the top-level clusters of the hierarchy form a full-mesh amongst themselves (i.e they are in a peer-peer relationship with each other).

The client-server relationship described above is used to break the *"don't propagate IBGP routes"* rule on the route-reflector servers. The server is allowed to *reflect* routes from a non-client (i.e. an IBGP router in a peer-peer relationship) to all its clients and from a client to all the other clients as well as non-clients. It is helpful to think of the server as a

proxy agent which disseminates routes between its servers and peers on one side and clients on the other (in both directions).

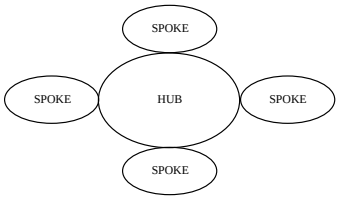

Figure 2: Hub and Spoke topology

ISPs typically deploy route-reflectors in a two-level hierarchy similar to the *hub-and-spokes* network in figure 2. The hub consists of all the route-reflector servers arranged in a full-mesh. These servers are physically located in a point-of-presence (POP) facility, typically in pairs for redundancy. Each of these facilities also contain the client routers as shown in figure 3 which represents a scaled down version of a large ISPs POP.

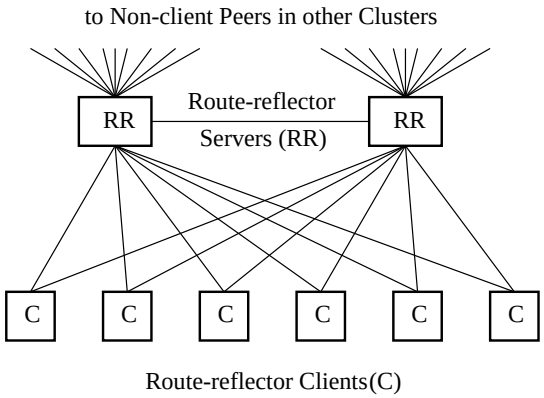

Figure 3: Route-reflector based POP

2.2 Confederations

Confederations tackle the same scaling problem by dividing an AS into *sub-ASes*. Each of these sub-ASes is fully meshed inside with regular IBGP sessions and is a flat BGP network. At the boundary between two sub-ASes, a modified form of EBGP is used (called confederation-EBGP) which adds the local sub-AS to the AS Path for loop detection within the confederation (the AS Path is a record of the ASes a prefix has traversed and is ordinarily used by EBGP to detect looping updates). To the outside world, this confederation of sub-ASes, looks like a single regular BGP network. At the boundary between a confederation of sub-ASes and a regular AS, the peering is a standard EBGP session, except that the router in the confederation does some extra work in order to hide its internal structure from the outside world.

Each sub-AS in the confederation can be further subdivided to form a confederation of sub-sub-ASes (and so on). The sub-ASes therefore form either an ancestor-descendant relationship (when an AS contains a confederation of sub-ASes) or a peer-peer relationship (when two sub-ASes belong to the same parent confederation).

With respect to deployment, confederation are typically made to fit the hub-and-spoke topology of figure 2. A central sub-AS forms the hub and spans the geography of the ISPs network. Metropolitan or larger areas typically form the spoke sub-ASes. Hierarchy within the hub or the spoke is typically not used.

3 Similarities and Differences

As may be evident by now, route-reflection and confederations solve the IBGP scaling problem in ways which are very similar on some counts but dissimilar on others. In this section we analyze the two approaches with respect to their underlying philosophies, deployment scenarios, problems unique to these approaches and scalability.

3.1 Underlying Philosophy

Route-reflection primarily works by changing the behavior of *IBGP* sessions. The main idea is that of selectively propagating updates over IBGP sessions from the routers designated to be route-reflector servers. On the other hand, confederations work by breaking up an AS into smaller, more manageable sub-ASes, in the process changing the behavior of *EBGP* sessions.

3.2 Deployment

In the field, route-reflectors have proven to be more popular than confederations. This is probably because deploying route-reflectors requires a software upgrade only on the routers which are to be designated as servers. The clients can be oblivious of the fact that some of the updates they receive are reflected. Confederations, on the other hand, require all routers to be able to process the segment type extensions to the AS Path attribute. This forces a topology moving from a full-mesh IBGP network to confederations to perform a fork-lift software upgrade of all the routers. For most existing networks, this is likely too high a barrier to entry. (For details on attribute extensions related to route-reflectors and confederations see [4] and [5]. In the interest of brevity and clarity, we have deliberately culled the details from this manuscript).

Interestingly, both route-reflectors and confederations are typically deployed in the hub-and-spoke topology discussed earlier in figure 2. In both cases, hierarchy is not used within the hub or the spokes. The only difference is that while for route-reflectors the boundary of the hub and the spokes is made up of routers (i.e route-reflector servers), with confederations the boundary is actually a confederation-EBGP session between routers in different sub-ASes.

3.3 Unique Problems

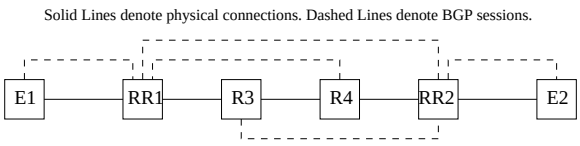

Figure 4: Persistent Loop

Because IBGP was originally designed around the *"never propagate routes learnt from another IBGP peer"* idea and because route-reflection breaks this rule, route-reflection based topologies can encounter *persistent loops* and other problems as described in [7]. We briefly repeat a simple example demonstrating this. Consider the network in figure 4. RR1 and RR2 are route-reflector servers with R3 and R4 as clients respectively (i.e. RR1 and R4 form a cluster and RR2 and R3 form another). E1 and E2 are in a separate AS and peer with RR1 and RR2 respectively. If E1 and E2 advertize the same prefix, RR1 readvertizes the prefix to R4 which now has a path to the prefix going through R3 to RR1 to E1. Similarly R3 gets a path to the prefix through R4 to RR2 to E2. R3 and R4 end up pointing to each other for the prefix in question, creating a persistent loop.

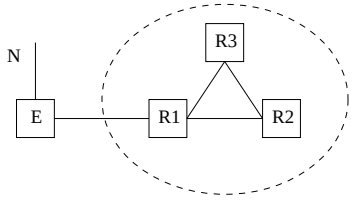

Figure 5: Sub-optimal Routing

Use of confederations on the other hand can lead to sub-optimal routing within an AS. Consider the topology in figure 5. R1, R2 and R3 are routers in the same confederation and each of them belong to a separate sub-AS. Router E is in an AS of its own and has a regular EBGP session with R1. E advertizes the network N to R1 which readvertizes it to R2 and R3. R2 and R3 also readvertize the route to N to each other. So R3 has a route to N from both R2 and R1. All other things being equal, R3 may choose the longer route through R2 to reach N. This is because while tie-breaking between routes to the same prefix, most BGP implementations do not take into account the length of the sub-AS path (some vendors solve this problem by providing a knob to take the sub-AS path length into account).

3.4 Scalability

Currently, large ISP networks run between 300 and 500 routers using one of these two approaches to reduce peering requirements. In the following paragraphs the maximum number of peering sessions is calculated for a hub-and-spoke network of approximately 400 routers –

Assume that a route-reflector based network has 20 POPs each with 2 servers and 18 clients, for a total of 400 routers. Each server therefore sees $18 + 1 + 19 \times 2 = 57$ IBGP peering sessions. The clients in each POP (assuming that they are fully meshed) see 19 IBGP sessions, one to each router in the POP.

Similarly assume that a confederation based network of 398 router has 20 sub-ASes, one of which is the hub containing 18 routers and the remaining 19 are spokes each containing 20 routers. Further assume that 2 routers from each spoke sub-AS peer with 2 router of the central sub-AS. Each router on the spoke sub-AS boundary therefore sees 19 IBGP sessions and 2 confederation-EBGP sessions for a total of 21 BGP sessions. Each router in the hub has 17 IBGP sessions and 4 confederation-EBGP sessions for a total of 21 sessions. The routers not on the boundary of the spoke sub-ASes see only the 19 IBGP sessions.

Clearly, the number of BGP sessions that need to be maintained by the confederations based network is much lesser than the route-reflector based network. The comparison is a little bit unfair as the confederation based network has one spoke less than the route-reflector based network. Yet, the general result holds. In both cases, the routers on the boundary of the spoke have the effect of condensing routes for the rest of the spoke. However with confederations, the hub itself does a lot of condensing before passing on the routes to the spokes (accounting for the reduction in the number of peering sessions) and the total number of updates seen by both the spoke-boundary routers as well as the spoke-internal routers is much smaller with confederations than route-reflectors. It should be noted that by imposing additional hierarchy, a topology with route-reflectors can be tailored to reduce the maximum number of BGP peerings on any router in the network.

4 Conclusion and Future Work

Both route-reflectors and confederations have proven themselves in the field and look very similar when deployed, but they have distinct advantages over each other. Route-reflectors are *backward compatible* and can therefore be deployed in a network incrementally without requiring a forklift upgrade. Confederations on the other hand reduce the number of BGP peering sessions much better (at least for the canonical hub-and-spoke topology).

Several questions remain unanswered in this article and present an opportunity for extensive simulation. For instance, how far in terms of the number of BGP sessions and the total number of BGP updates can the two-level hierarchy for route-reflectors and the similar hub-and-spoke for confederations scale? Or, how do the two techniques compare in terms of convergence time in the face of failures? In addition, the effect of these mechanisms on the stability of the network as a whole is not clear and should be looked at more closely. [8], [9] analyze the general problem of instability in the Internet, but they don't specifically identify the role of network architecture with respect to this instability. As the size of the ISP networks increase, the importance of this particular problem will grow.

Acknowledgements

We would like to thank Vab Goel for describing the Sprint Network, Joe Malcolm for describing the UUNET network and Jeff Young for describing the Cable and Wireless (formerly MCI) network and Tony Przgyienda and the CCR reviewers for reviewing this paper.

References

[1] Y. Rekhter and T. Li. A Border Gateway Protocol (BGP-4), March 1995. IETF RFC 1771.

[2] B. Halabi. *Internet Routing Architectures.* Cisco-Press, 1997.

[3] J.W. Stewart III. *BGP4: Inter-Domain Routing in the Internet.* Addison-Wesley, 1998.

[4] T. Bates and R. Chandra. BGP Route Reflection: An alternative to full mesh IBGP, June 1996. IETF RFC 1966.

[5] P. Traina. Autonomous System Confederations for BGP, June 1996. IETF RFC 1965.

[6] D. Haskin. A BGP/IDRP Route Server Alternative to a full mesh routing, October 1995. IETF RFC 1863.

[7] R. Dube and J.G. Scudder. Route Reflection Considered Harmful, November 1998. IETF Draft draft-dube-route-reflection-harmful-00.txt.

[8] C. Labovitz, G.R. Malan, and F. Jahanian. Internet Routing Instability. In *SIGCOMM Conference*. ACM, 1997.

[9] C. Labovitz, G.R. Malan, and F. Jahanian. Origins of Internet Routing Instability. In *INFOCOM Conference*. IEEE, 1999.