# BGP Scaling Techniques Revisited

John G. Scudder
Internet Engineering Group, LLC
122 S. Main, Suite 280
Ann Arbor, MI 48104
*Email: jgs@ieng.com*

Rohit Dube
High Speed Networks Research, Bell Labs
4C-508, 101 Crawfords Corner Rd
Holmdel, NJ 07733
*Email: rohit_dube@acm.org*

## 1 Introduction

This note adds definitions and clarification to "A Comparison of Scaling Techniques for BGP" [1], corrects some minor errors and clarifies points which may not have been clear in the original paper. It also adds a new analysis of the scaling properties of route-reflectors and confederations, leading to a new conclusion that route-reflection and confederations scale equally well in the general case and that the choice of the scaling technique employed by a network needs to be made on a basis other than the scalability metric we have analyzed.

## 2 Terminology

[1] measures "scaling" in terms of the maximum number of BGP sessions any given router in the network must support. This is probably the best gross metric of how a BGP network scales. For convenience, we term this scaling metric the "BGP session degree" of a network.

In addition, we introduce the term "CBGP" to mean "confederation-EBGP".

## 3 Corrections

### 3.1 Tree Topologies

[1] states in Section 2.2 (paragraph 2) that sub-ASes within a confederation can be sub-divided. The terminology in this section is not as clear as it should be — a confederation can arrange its sub-ASes into any general graph. A specific realization of this graph could be a tree, the root of which is the hub and its children would be the spokes. Each of these spokes could serve as the hubs for a second level of spokes and so on.

Route-reflector clusters are typically deployed in a tree-topology. This is not strictly necessary as valid non-tree configurations are quite possible with route-reflectors.

### 3.2 Unique Problems

The "persistent loop" problem presented for route reflectors in Section 3.3 can also be contrived with confederations, and thus isn't actually unique. Consider the case when RR1 and R4 comprise one sub-AS, and RR2 and R3 comprise a second sub-AS. Instead of a peering session between RR1 and RR2, we have a CBGP peering session between R3 and R4. It is possible for a loop to form in a manner similar to that presented for route reflectors. (However, we note that owing to the typical deployment model for confederations, there may be less risk of such a misconfiguration in the field.)

The online version of Figure 4 of [1] has dashed lines showing the BGP sessions (these are the non-straight lines in the figure). The printing process seems to have inadvertently blurred the space between the dashes making dashed lines look like solid lines.

In Section 3.3 (Figure 5) of [1], attention should be paid to "all other things being equal". In general, nearest exit routing based on IGP cost to the BGP nexthop will lead to optimal routes. Only in the case where there is a tie on the IGP cost will the sub-optimal routing case present itself.

### 3.3 Off By One

[1] has an off-by-one error in the description of the confederation network in Section 3.4. Since we have 19 spokes, with two border routers per spoke and two CBGP sessions per spoke border router, there are $19 \times 2 \times 2 = 76$ CBGP sessions in the network. Thus, if each router in the hub network is to support four CBGP sessions, as described, there must be 19 routers in the hub network, not 18 as previously presented, and thus there are 399 and not 398 routers in the entire network.

## 4 BGP Session Degree Comparison

The topologies compared in [1] reflect current practice for the deployment of route-reflectors and confederations, and further, each topology reflects the simplest deployable topology with currently-available implementations. However, we note that it is possible to deploy route-reflectors in a three-level hierarchy with exactly the same number of BGP sessions as the corresponding network with confederations —

Assume, as in the confederation case in [1], that a route-reflector based network of 399 routers has 19 POPs comprising 20 routers each and a single core backbone comprising 19 routers. Each POP contains two route-reflection servers and 18 route-reflection clients. Each route-reflection server is in turn a route-reflection client of two core routers. We thus have a three-level route-reflection hierarchy, instead of the two-level hierarchy previously presented. Each of the route-reflection servers in the POPs thus has 18 IBGP sessions to its clients, plus one to its peer route-reflection server, plus two to the core routers of which it is a route-reflection client, for a total of 21 IBGP sessions. Each router in the core has 18 IBGP sessions to its peer core routers, plus four IBGP sessions to its route-reflection clients, for a total of 22 IBGP sessions. The "leaf" route-reflection clients require only 19 IBGP sessions, 17 to the other clients and 2 to the route-reflection servers.

(Note that it would also be possible to omit the internal full mesh, requiring only two IBGP sessions of each client, at the cost of some additional latency.)

If the three-level deployment does a better job of minimizing the BGP session degree of the network, why are route-reflectors not deployed in this way? We believe that it is because it is not necessary to do so. Current high-end backbone routers are capable of supporting the tens of BGP sessions needed to deploy route-reflection with just two levels of hierarchy. The shallower hierarchy (a) is simpler, which is a desirable trait, and (b) induces less latency in the propagation of routing updates.

It is not possible to simplify a confederation deployment to the level of the previously-presented route-reflector deployment. However, we note that this is an implementation consideration. With a BGP implementation capable of participating in two different sub-ASes simultaneously, it would be possible to construct a confederation deployment analogous to the route-reflector deployment.

## 5 Revised Conclusion

[1] presented the two commonly-used BGP scaling techniques, and concluded that confederations were more scalable in terms of the BGP session degree. We now conclude that route-reflectors and confederations are identical in terms of their scaling properties. The only difference is that the simplest route-reflector topology with two tiers has a higher BGP session degree compared to the simplest confederations topology.

Both confederations and route-reflectors can support multiple levels of hierarchy and in general, each level of hierarchy further reduces the BGP session degree of the network. However, in the real world, two- and three-tier hierarchies suffice to construct very large networks. The choice between route-reflectors and confederations is therefore driven by other considerations such as incremental deployability and flexibility of configuration. The former plays in favor of route-reflectors as mentioned in [1]. We briefly consider the latter in terms of (a) running multiple IGPs in the same AS and (b) specifying BGP policies at the cluster or sub-AS boundaries with-in an AS.

With respect to running multiple IGPs, confederations have more flexibility since confederation members are not forbidden from rewriting the BGP next hop when sending a route into CBGP. The effect of this is that different sub-ASes need not have detailed knowledge of one another's topologies, and may run independent IGPs. (Doing so may result in sub-optimal routing, however, as discussed in [1].)

We note that it is possible in principle to deploy route reflectors with multiple IGPs, one per cluster. This requires that all routers within the AS be made aware of all potential BGP nexthops in order for the prefixes being advertised to be considered reachable. Currently known deployments of route-reflectors do not use multiple IGPs in an AS so the complete mechanism has not been developed and we refrain from discussing it here. Finally, we note that it might be reasonable to allow route reflectors to rewrite BGP next hops just as confederation members are permitted to do, in which case the two would again be perfectly comparable.

With respect to policies, both confederations and route-reflectors can support policies at cluster or sub-AS boundaries. These policies may serve to limit inter-cluster or inter-sub-AS control traffic, and as such may enhance scalability. There are minor differences in the exact policies which may be applied in each case, but the major difference may be the mental model – since confederations model the confederation as a group of smaller ASes, with CBGP behaving much like EBGP, it may be more natural for a network administrator to apply policy controls at sub-AS boundaries.

## References

[1] R. Dube. A Comparison of Scaling Techniques for BGP. *ACM Computer Communications Review*, 29(3), 1999.