Improving Routing Scalability with Aggressive Route Aggregation

[1,2]Wei Zhang, [1,3,4] Jun Bi, [1,3,4] Jianping Wu
[1] *Department of Computer Science, Tsinghua University, Beijing, China*
[2] *Wuhan Commanding Communications Institution, Wuhan, China*
[3] *Network Research Center, Tsinghua University, Beijing, China*
[4] *Tsinghua National Laboratory for Information Science and Technology (TNList), China*
zw@netarchlab.tsinghua.edu.cn, junbi@tsinghua.edu.cn, jianping@cernet.edu.cn

doi:10.4156/jcit.vol6. issue2.16

Abstract

The size of the Boarder Gateway Protocol (BGP) routing table is continuously increasing, which incurs more memory demands in routers and increasing BGP update messages between peering routers as well. The RIB/FIB does not scale well with the Internet size, which is known as the routing scalability problem of the Internet. There are two major facts relevant to this problem: firstly more and more newly assigned IP blocks come into the routing system and in most cases are not aggregatable with the previously assigned addresses in their origin ASes; secondly, many BGP prefixes are de-aggregated into fragments or overlapped with numerous trivial prefixes and these fragmental prefixes are widespread throughout the Internet without effective aggregation. In this paper we address the second point: improving the routing scalability via aggressive aggregation strategies.

Our observation of the BGP routing information suggests that there exists a routing "core" of the Internet which is composed of a few "tier-1" Internet Service Providers (ISPs). Numerous "edge" autonomous systems (ASes) get transit routes to the other region of the Internet through the "core". We propose the "core" to conduct aggressive aggregation before advertising routes to their customers. With this strategy, a considerable portion of the fragmental prefixes can be eliminated from the perspective of the "edge" ASes. Our simulation confirms that this strategy will gain admirable improvements on reducing the size of routing tables for most part of the Internet.

Keywords*: Internet, Routing, Aggregation, Scalability*

1. Introduction

The Internet routing scalability problem has been highlighted by IETR/IRTF in recent years. As the Internet evolves along a path of seeming inexorable growth, the BGP routing table size and the BGP update churn become two prominent Internet scaling issues. This dramatic growth of the routing table can decrease the packet forwarding speed and demand more router memory space.

To the Internet operators, the most concerned issue is the impact of the rapid growing size of the BGP routing table. In some sense, the routing scalability problem showed up as routing table inflation, which implies the growth of routing convergence time, and the cost, power (and hence, heat dissipation) and ASIC real estate requirements of core router hardware. Intensive studies have been conducted by various research groups and an influx of outputs was formulated in RFC 4984[1] as a basic understanding of this issue. The root causes of the routing scalability problem are interweaving of architectural deficiencies and practical manipulation of route services provisioning: e.g. the IP address allocation practice and Provider Independent (PI) addresses assignment, multihoming, longest prefix match routing mechanism, etc. All above mentioned factors are against IP prefixes aggregation within the routing system.

Based on our observation of the routing information, we believe an aggressive aggregation strategy can help to reduce the BGP RIB/FIB. We will illustrate where and how to apply this strategy. Our evaluation shows that only a few "tier-1" ASes need to be the subjects of our new routing scheme before the majority of the ASes may enjoy considerable benefits.

According to our understanding, the few "tier-1" ASes constructed a so-called "transit core" of the Internet. This core is densely connected if not "full-meshed". A large portion of the BGP routes must be proliferated to the whole world through the "core". At present, ruled by CIDR and the longest prefix match routing mechanism, the BGP prefixes tend to be de-aggregated by ISPs in order to implement traffic engineering (TE) or multihoming. Unfortunately, the BGP control plane has a “flat” structure and these de-aggregated prefixes can be proliferated to every BGP routers although in most cases which is not necessary. We argue that for a given ISP, in the case of multihoming or TE, locally de-aggregated prefixes may be only meaningful to its neighboring ASes and not necessarily the same significant to every remote ASes. Therefore we hope that fragmental prefixes can be aggregated into less specific ones before they have been advertised to numerous “remote” ASes.

Motivated by this concern, we propose the following aggressive route aggregation strategy: the tier-1 ISPs in the “transit core” of Internet, aggressively aggregated prefixes before advertise “remote” routes to its local customers. In other words, the “core” ASes refrain to advertise “trivial” prefixes to their customers. Therefore, the “core” ASes will block the more specific prefixes from proliferating to every BGP routers located at the “edge” of the Internet. This strategy can reduce the size of routing tables from the perspective of the numerous “edge” ASes.

One may instinctively feel that our scheme constitutes a routing hierarchy (differentiating ASes into “core” and “edge”), and may incur path stretch consequently. But before we jump to a pessimistic conclusion we need to analyze the inter-domain paths in real life. We argue that any inter-domain path through the “transit core” is the least wanted for each edge ASes, and normally much longer than the average ASes distance in terms of AS hops. Based on this assumption, the stretch is limited. What is more, if we take the perspective of an edge AS, the most concerned issue is the price instead of other things (e.g. the stretch or the provider’s traffic engineering demand). Anyway, we estimate an upper bound of the stretch and conservatively the average stretch which is by large affordable.

We also discuss the impact of this strategy to tier-1 ISPs in the “core”. If a “core” ISP wants to induct its customer’s traffic from different AS border routers (ASBRs), it normally may advertise de-aggregated prefixes to corresponding BGP peers. In our scheme, this advantage can be preserved only if the de-aggregated prefixes are put into corresponding BGP community advertisements.

The rest of the paper is organized as follow: section 2 illustrates our observation and understanding on the BGP route aggregation; section 3 gives a brief description of our proposal; section 4 evaluates the gains via simulation; section 5 analyzes the possible stretch in our scheme; section 6 discusses the impacts of this strategy; section 7 compares our method with other related works. Finally, we conclude our contributions in section 8.

2. Observations

According to BGP routing information collected by [2,3,4], we examined the composition of the Route Information Base (RIB) and Forwarding Information Base (FIB) respectively as well as the BGP updates collected from monitoring vantage points, and observed the following:

- Poor aggregation: too many long prefixes in the RIB/FIB.

In figure 1, we plot the FIB observed from AS 6447 with the distribution of different mask lengths. The average prefix length is about 22.3, with /24 prefixes account for about half of the FIB size. As well known, the /24 prefixes is the longest prefixes that conventionally allowed in BGP which means most of the BGP routes are not aggregated from their originate ASes to the observations points. The BGP routing scheme cannot be scalable as long as there is no effective aggregation mechanism.

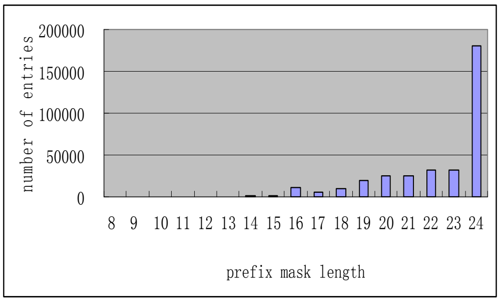

Figure 1. Prefix mask length distribution in FIB

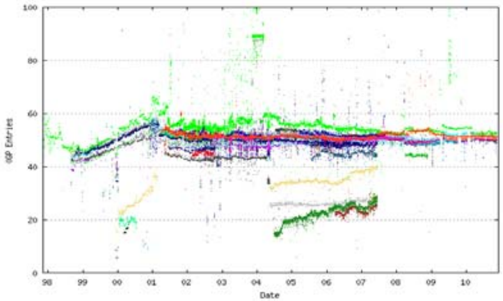

Figure 2. percentage of covered prefixes

- Many “covered” prefixes in FIB.

We define a more specific prefix to be the “covered” prefix of a less specific prefix (namely “covering” prefix). E.g. “10.4.0.0/16” covers “10.4.1.0/24” and “10.4.2.0/24”. In a typical BGP router, about 50% of the RIB/FIB entries are routes with overlapped prefixes, since an advertisement might be associated with a sub-span of address space of another advertisement. [2] Plotted the percentage of “covered” prefixes in the FIB from 1998 to 2010, as shown in figure 2. In figure 2, different observation points were plotted in different colors. Although there are some discrepancies between observation points, it is obvious that in most ASes approximately 50% of FIB entries are “covered” prefixes. Especially after 2008, the discrepancy converges to a more obvious consistency.

- Route redundancy: overlapped prefixes with the same AS path.

Among the routes of the more specific prefixes (covered prefixes), too many (about 40%) have the match path or origin AS with their respective aggregation prefixes (covering prefixes), as shown in Figure 3 plotted by [2]. There is an increasing trend in the track, and the noises were caused by the routing dynamics (instable routes). Obviously the redundancy of the routing information is remarkable. Although the origin AS may have justifiable reasons to de-aggregate its prefixes, the redundant information should not be proliferated to every BGP routers. B. Zhang et al. proposes FIB aggregation in [5] to get rid of the redundancy, but their techniques cannot reduce the RIB size and BGP updates. What is more, the FIB aggregation is costly when take the computing/updating overhead into consideration.

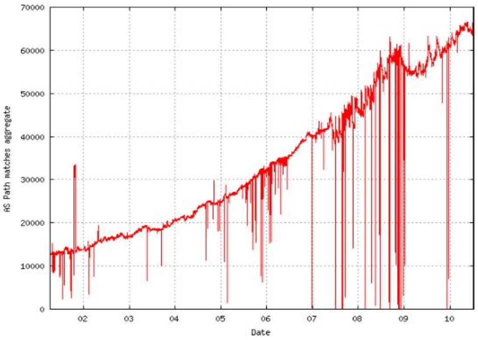

Figure 3. AS path match the covering prefixes

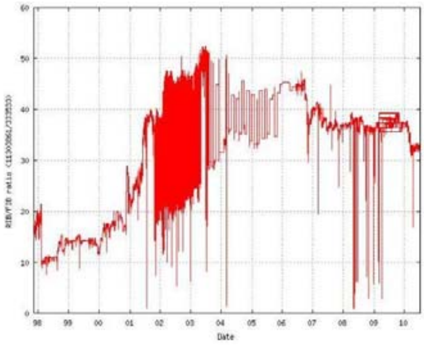

Figure 4. RIB/FIB ratio

- A high RIB/FIB ratio

For some ASes, we observed the RIB/FIB ratio can be as high as more than 30 as shown in Figure 4 plotted by [2]. The RIB/FIB ratio is related to the number of BGP peering sessions of a given AS. Consequently, the BGP update overhead might be a heavy burden in this case (have to handle more than 14M RIB entries).

- The “core” ASes shown up in many AS paths

As we further analyze the AS path of each RIB/FIB entries, we find that a few “core” AS numbers have shown up in many AS paths. From the perspective of the observation point, there are a considerable portion of routes being received from this “transit core”, and the “core” ASes fail to aggregate them when advertise these routes to their customers.

The above observations gave us the following fundamental understanding: numerous “edge” ASes have strong incentives and justifiable reasons (e.g. traffic engineering or anti prefix hijack etc.) to de-aggregate their assigned address blocks when they advertise BGP prefixes as origin ASes. This assumption has been confirmed by previous works [5, 15, 18, 20]. The de-aggregated prefixes contribute a significant portion of the BGP RIB/FIB. The “core” ASes fail to aggregate these aggregatable prefixes before advertising them to the other “edge” ASes. Therefore the routing structure is fairly “flat” and consequently not scales well.

3. Our proposal

Based on our observation and understanding, we propose an aggressive route aggregation strategy that all the “core” ASes refrain from advertising more specific prefixes to their customers. Since many routes traverse the “transit core”, the ISPs in the “core” hold advantageous positions to aggregate those redundant routes.

In our scheme, the tier-1 ISP may aggregate all BGP prefixes received from peering tier-1 ASes when advertises them to its customers, as long as the aggregated prefix will not cover the span of addresses that are unreachable. This strategy can be easily implemented if the tier-1 ISP simply set more aggressive aggregation rules into the RIB-out policy to each customer.

In this case, the aggressive route aggregation only needs to be mandatory to a few “tier 1” ISPs (since they have no overt providers). Definitely, their ASes are supposed to be located in the “transit core” and hopefully with a densely connected mesh topology (if not full meshed). Then this “transit core” can prevent much specific routing information from propagating to the whole BGP world. The “edge” ASes will not change, except some of them might need to configure BGP communities with their providers if the provider needs to de-aggregate some prefixes in order to implement its inbound traffic engineering. We will discuss the situation in section 6.

All in all, we propose to establish a straightforward “core/edge separating” routing structure in the Internet. In the “core”, the peering tier-1 ASes exchange routes with each other with or without aggregation. Their customers (non tier-1 ASes) will advertise prefixes to their providers (tier-1 ASes) as normal. On the other hand, the only change that we are going to introduce into the routing scheme is that all tier-1 ASes advertise aggressively aggregated routes to their customers. The consequence of this strategy is that the tier-1 ASes will have all BGP routes as usual; the “edge” ASes will reduce their RIB size to some extend that depends on the aggregation efficiency of their providers.

We evaluated our scheme via simulation, and the methodology and outcome of this strategy have been demonstrated in the following section.

4. Evaluation

4.1. Methodology

To simulate our strategy within the “transit core”, we used the 14 tier-1 ASes defined in [14] and further extend our “core” to include other tier-1 ASes defined by [4]. Finally altogether 30 ASes had been chosen as our baseline set. In future, new tier-1 ASes may join the “core”. However, the evolution of the “core” is relatively stable. One thing we can count on is the number of “tier-1” ASes and their topology significance will not change too much.

We then used BGP RIBs collected by Routeviews [3]. Our data set came from 92 of monitors. In order to evaluate the outcome in the “edge” networks we removed the monitors located in the “core” (the 30 tier-1 ASes chosen previously), therefore we had 62 observation points dwelled in 50 distinct “edge” ASes.

From the BGP dump data collected from these observation points we checked the AS Path attribute of each prefix. Then we conducted the following algorithm: step 1, go through the BGP RIB of the “edge” AS, if the AS path attribute of a given prefix consists of any “core” AS which we defined beforehand, we put the prefix into the candidate aggregate set of the “core” AS. After step 1, we got candidate aggregation sets for all providers of the “edge” AS. In step 2, we try to aggressively aggregate prefixes in each candidate aggregation set. Multiple prefixes can be aggregated as one less specific prefix provided they meet the binary boundary requirement and there is no “hole” in the aggregated prefix. For overlapped prefixes, the covered one will be simply aggregated into the covering prefix. The AS-path attribute of the new aggregated prefix needs to be overwritten. We set the AS-path as a concatenation of two parts: the AS-set part and AS-sequence part. The AS-set part contains the group of ASes from the origin AS to the “core” AS that aggregate the routes. The AS-sequence part represents an ordered sequence of ASes from the aggregation point to the “edge” AS (where our observation point located). The new AS path may help us evaluate the path stretch of this routing scheme.

4.2. Estimated gains

In our simulation, we calculated the prefixes received by "edge" ASes after the aggressive aggregation strategy being taken by all "core" ASes. We observed considerable reduction for most edge ASes: the worst case was in AS 19151 which still had 145033 routes and the best case was observed in AS22338 which only had 25901routes left. The discrepancy was caused by the position of the observation points. Some edge ASes who have many peers or customers can still receive numerous routes accordingly, while the others will not have many routes other than the aggregated routes advertised from the "core". We plotted 11 ASes from the 50 distinctive observation ASes in figure 5. The axis X is the "edge" ASes ordered by their remaining RIB/FIB size. The axis Y represents the number of the entries in the FIB of a corresponding AS. We also plotted a CDF of all 50 edge ASes in figure 6, and more than 90% of the edge ASes will have a FIB size of no more than 100000 entries.

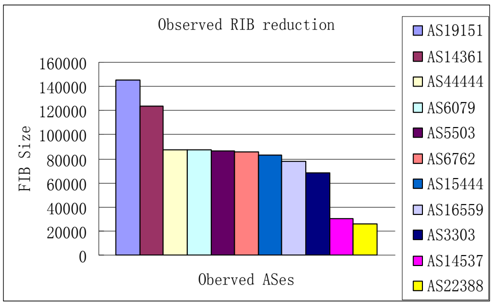

Figure 5. Sampled edge ASes

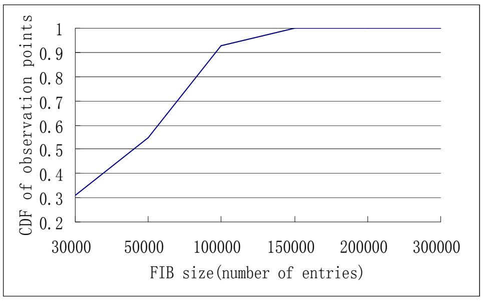

Figure 6. CDF of FIB reduction

We further evaluated the gains against the size of the "core", vary from 0 to 30 "core" ASes. The observed FIB reduction at the edge ASes was plotted in figure 7 in a quartile formation. There would be notable reduction if the "core" has over 15 tier-1 ASes. But the gains varied according to the different locations of "edge" ASes. The discrepancy of gains would shrink as the "core" size extends to 30. If all tier-1 ASes have been included in the "core", generally the gains would converge to an average level (about 60% reduction of the routing table size). At last, we randomly chose 30 ASes other than the tier-1 ASes to construct a "core". Not surprisingly the result was rather bad, and almost no reduction was observed. The results have shown that the tier-1 ASes in the "core" played an important role in this scheme. This observation gives another proof of the "Robust Yet Fragile (RYF)" nature of the Internet.

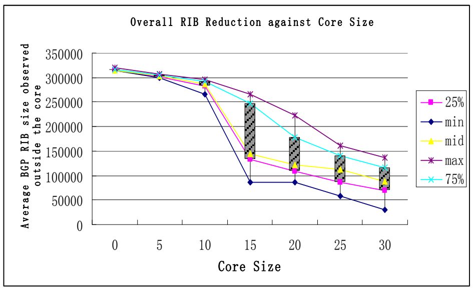

Figure 7. Gains against different "core" sizes

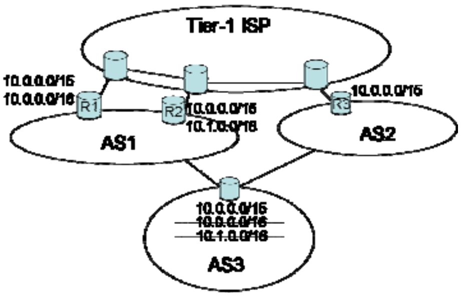

Figure 8. TE solution with BGP community

5. Impacts

The aggressive aggregation strategy constrains the freedom of tier-1 ISPs to announce routes with more specific prefixes. Then what impacts will be induced by this strategy?

5.1. Impacts to core ASes (ISPs)

To tier-1 ISPs, the first concern is the impact on TE: ISPs hope to manipulate inbound traffic by way of de-aggregating routes at different inlets. E.g. in figure 8, a tier-1 ISP wants to advertise “10.0.0.0/15” to Router1, Router2 and Router3, and advertise more specific “10.0.0.0/16” to Router1, “10.1.0.0/16” to Router2 respectively. In this case AS3 will receive all 3 prefixes. Hopefully AS3 will follow the routes guided by the tier-1 ISP, and may forward all packets destined to “10.0.0.0/15” via AS1. How can an ISP make it with the aggressive aggregation strategy? We propose a practical way out in this case: the ISP can announce the more specific prefix within a BGP community. In our example, the ISP can establish a community with AS1 and AS3, and advertise the more specific prefixes (i.e. “10.0.0.0/16” and “10.1.0.0/16”) within the community. Therefore, these trivial prefixes will not be propagated to other irrelevant edge ASes. Basically the traffic engineering in the tier-1 ISP will not be influenced if it can establish appropriate communities with its customers.

5.2. Impact to edge ASes

To the edge ASes, the most concerned impact is the path stretch incurred by our routing scheme. Some of the multihoming “edge” ASes may have more than one access to the “core”, they may prefer non-aggregated routes to facilitate their route choice between different providers or access points to the core. Since their providers (tier-1 ASes) aggregate most “core transit routes”, the edge ASes lost most information of the AS path from the “core” to the destination. In this case, they may choose an access to the core which is not on the shortest path to the destination. The ratio between the suboptimal path and the shortest path is the path stretch. Undoubtedly, the shortest path has the minimum stretch of 1.

In practice, a multihoming customer can enjoy absolute freedom on choosing providers to forward a certain packet. Basically there are 3 strategies: one, “hot potato” principle; two, next-hop policy (cold potato); and three, shortest AS path policy. The “hot potato” principle concerns the intra-domain distance most. A packet will be forwarded to the closest outlet to the provider. The next-hop policy mainly concerns the inter-domain preference. A packet will be relayed to the most preferred providers (ranked by traffic fee, bandwidth or social reason etc.). The shortest “AS path policy” concerns the delay, with an assumption that a shorter AS path will have a smaller round trip delay. A packet will be forward to the provider who has a route to the destination with the shortest AS path. Aggressive aggregation strategy may hide AS path information from the “edge” ASes, therefore the edge AS may choose a provider on the suboptimal path to the destination. In other words, our scheme may incur path stretch in this case.

We will give a conservative estimation of the stretch through pragmatic analysis. When tier-1 AS, *X* and *Y*, advertise aggregated prefixes *p* to their customers, assume that the “edge” AS *E* has no route to *p* except through “core” AS *X* or *Y*. The “edge” AS *E* may choose a route to the nearest tier-1 AS when forwarding a packet destined to *P*. Suppose this time *X* is the nearest tier-1 AS to *E*, but *Y* is closer than *X* to the destination. L_1 denotes the distance between the *E* and *X*. L_1’ denotes the distance between E and Y. L_2 denotes the distance between *X* and the destination. L_2’ denotes the distance between *Y* and the destination. Therefore, we have the following inequations:

$$L_1 \le L_1'$$
$$L_2 \ge L_2'$$

However, within the “core”, the tier-1 ASes are densely connected and most likely apply the “transit-free” policy. Namely a tier-1 AS will not provide transit service to other two of its tier-1 peers. Within the core, AS *X* has the detailed routing information to the destination and will forward the packet directly to the exit core AS. If *Y* is the exit AS, there will be one extra AS hop; otherwise there would be no stretch. With this assumption, L_2 will have no more than one extra hop compared with L_2’. In the worst case, we have the following equations:

$$L_1 = L_1'$$
$$L_2 = L_2'+1$$

Put everything together, the suboptimal path has a length of $L_1 + L_2$. The optimal path has a length of $L_1' + L_2'$. We further assume the "core" transit path has a minimum length of 2(one hop from the edge to the core, another hop from the core to the edge). Then we have the following conjecture:

$$stretch = \frac{L_1 + L_2}{L_1'+L_2'} \le \frac{L_1'+L_2'+1}{L_1'+L_2'} = 1 + \frac{1}{L_1'+L_2'} \le 1 + \frac{1}{2} = 1.5$$

According to our assumption, the length of the "transit core" path is at least 2; therefore the maximum stretch is 1.5. Normally, a transit "core" path can be longer than the averaged inter-AS distance, which is between 3 and 4. As a conservative estimation, the average stretch will be about 1.3. Other inter-AS paths which do not transit the "core" will not have any stretch, because the edge AS knows the best route to the destination via its peers or customers.

We did not simulate the stretch incurred by our routing scheme, because the inter-domain routing practice can be very complicated. In too many cases the AS path is not the shortest one appeared in the RIB. The edge networks (customers) have the freedom to choose the access point to their providers as they like. We believe our estimation has given a credible upper bound of this metric.

6. Feasibility

Our proposal can be relatively easier to be implemented compared with alternative routing schemes, e.g. the LISP[16] or other core/edge separation solutions. Firstly, our proposal will not change the forwarding mechanism, and even will not change the routing functioning in the BGP implementation. Such transit tunneling or address mapping system in LISP is not required. The only requirement is the routing configuration to implement aggregation strategy. Secondly, the subjects of our scheme are confined in a very limited "core". The aggressive aggregation strategy only involves a few tier-1 ASes. Our solution is absolutely transparent to the numerous edge networks and end hosts. Any proxy deployment or host upgrading is not required. Thirdly, our routing scheme does not require a kickoff day. Any "core" ISP may take an incremental deployment within its ASes without the risk of routing loops. Ideally it can be a "default" course of action, if nothing significantly changes in the Internet's architecture in the near future.

The only mandatory is that all tier-1 ISPs should agree on applying the strategy lest unfair competition happens. This proposal requires an intense cooperation among tier-1 ISPs. In some sense, we are devising a cooperative mechanism in the Internet routing system. According to[15], there are strong parallels between the BGP routing space and the condition commonly referred to as the tragedy of the commons. The BGP routing space is simultaneously everyone's problem, as it impacts the stability and viability of the entire Internet, and none's problem in that no single entity can be considered to manage this common resource. One possible solution of the tragedy of the commons may rely on the imposition of a consistent set of policies and practices intended to achieve a particular outcome. The vehicle for such an imposition of policies and practices is most commonly that of regulatory fiat.

One limitation of our scheme is that there is no scalability improvement in the "core". Since all tier-1 ASes still receive de-aggregated fragmental prefixes from their customers and peering tier-1 ASes. They will have a full RIB/FIB as normal size in DFZ. However, if the tier-1 ISPs start seriously feeling the effects of routing table pressure they will probably apply virtual aggregation techniques within the "core". We believe such virtual aggregation techniques as CRIO[9] or Viaggre [10] can be deployed efficiently within a small region. Because virtual aggregation may introduce high stretch and the choice of the locations for a large scale virtual aggregation points can be difficult, these problems may hinder the technique to be applied worldwide. But within the "core" of the Internet, the size is limited (a dozen of tier-1 ASes) and the topology is relatively stable, the virtual aggregation is a very promising technique.

7. Related works

Recently many researches have addressed the Internet routing scalability problems. Most of them admit that the impending pressure will be a critical issue along the path of Internet evolution, although very few [18] may argue that there will be no problem if Moor's law can bless us to the end. We find most relevant proposals try to suppress the routing information within the routing system. No matter by way of route aggregation, hierarchical routing with core/edge separation or more radical "clean slate" routing paradigms for innovative routing schemes [21, 22], most of these schemes are either not practical to be implemented or have inherent limitations.

About BGP route aggregation, BGP CIDR reports [4] publish their statistics data on how route aggregation can reduce Internet routing table sizes. However, the CIDR report only suggests the tier-1 ASes to aggregate their announced prefixes, which means to aggregate the prefixes originated in the tier-1 ASes. Obviously their proposal is different from our routing scheme although our study also referred to their data. We don't require the origin AS to give up de-aggregation since it may facilitate regional traffic engineering. On the other hand we put emphasize on the "filter" functioning of the "core" to prevent the de-aggregated prefixes from proliferating to irrelevant edge ASes.

B.Zhang et. al. in [6] propose a route aggregation approach to aggregate FIB. Z. Uzmi et.al propose similar techniques called MALTA [7] recently, which provides better aggregation but is also more complex. Their techniques aim to reduce the route redundancy in a router through aggregating prefixes with the same forwarding port. However, their proposal will not alleviate the inflation of BGP RIB. With the consequent increasing of BGP update, the computing overhead might be prohibitively high. Our method focuses on the control plane and addresses the root cause of route scalability directly.

Costas Kalogiros in [8] analyzes the incentive of aggressive route aggregation in BGP. However, their analysis is confined in a non-cooperative game played between individual ASes. We try to establish a cooperative mechanism within the "core", which is the most efficient and pragmatic way to solve the tragedy of the commons problem.

Virtual aggregation has been proposed in [9, 10]. This scheme requires additional routing deployment and implementations. Viaggre [10] is designed for intra-domain routing and CRIO [9] may require cooperative deployment among a few ASes. Generally, virtual aggregation can solve regional BGP scalability problem at the cost of high stretch and transmission overhead (since virtual aggregation relies heavily on the tunnel along sub optimal paths). Compared with virtual aggregation techniques, our proposal is much more straightforward and can be applied in the Internet-wide scenario.

There are other inter-domain route aggregation schemes [11,12,13] based on hierarchical routing which either requires renumbering or new addressing mechanism. And basically these schemes are designed for IPv6. Our scheme focuses on the control plane of BGP without any addressing prerequisite.

A most recent research measures the growth of Internet and analyzes the trends in [18]. Their major findings indicate that the growth of routing table and BGP dynamics is due to the growing edge of the Internet. In this light, core/edge separation could be an approach worth exploring to improve routing dynamics.

As we have mentioned in section 6, there are several core/edge separation routing schemes, such as LISP[16] and [19] , also propose to establish routing hierarchy in the Internet routing system. However, the boundary between the core and the edge is not clear in their schemes. In addition, almost all core/edge separation routing schemes require to separate the address space, therefore a mapping system is required, and the forwarding mechanism needs to be changed when a packet has to transit the core.

Of the hierarchical routing, Krioukov has conducted highly relevant theoretical analysis in [17], which includes a pessimistic conclusion that any highly aggregated tree-structured routing scheme will be very inefficient (cause high stretch) on Internet-like topologies. Our proposal suggests a moderate aggregation and takes full advantages of the topology discrepancy between the "core" and the "edge" ASes. In our routing hierarchy, the "root" of the tree is the "core", which is a group of densely inter-connected large scale networks. This structure is more or less similar to the structure of a "fat" tree. The difference between our routing structure and the Fat tree is that the branches and leaves that are not in the "core" may have additional links. So this is not a strictly "tree-structure", but a level structure which matches the feature of the Internet topology very well. Our analysis justifiably claims the stretch will be affordable.

8. Conclusion

Based on our observations and fundamental understanding, we find that a major improvement on routing scalability can be achieved via aggressive route aggregation. Only a small "transit core" in the Internet is the most critical subject to this aggregation strategy. We then propose a scheme to accomplish the scalability improvements. Our simulation confirms that if the 30 tier-1 ASes apply aggressive aggregation before advertise fragmental prefixes to their "downstream" customers, there would be considerable RIB reductions observed from different "edge" networks. The average FIB size would be approximately 100000 entries (about 70% gains). We also analyze the impacts of the aggregation strategy on traffic engineering and other practical concerns to ISPs and their customers. we propose to use BGP community attribute to meet the ISPs' requirement on TE. According to our conservative estimation the possible AS path stretch can be at most 1.5 and averagely no more than 1.3, which is affordable to most customers.

The contributions of our proposal can be listed as following:

- Our scheme provides a pragmatic routing hierarchy in the Internet to improve routing scalability;
- We provide a cooperate mechanism to tackle the tragedy of the commons. And this cooperative community only involves a few members in the "core".
- We confirm the gains of our routing scheme via simulations based on reliable BGP data collected from dozens of distinct observation points in the Internet.

The Internet has evolved into a stage that not all participants stand on the same level. The most significant and well developed pioneers have taken critical vantage positions and should take more responsibilities for the sake of the public good. It is a good attempt to organize cooperative communities across administrative boundaries. We are addressing the scalability problem from the perspective of the BGP routing. Similar profound understandings of the Internet ecosystem may lend a hand to the architectural design of the future Internet.

9. Acknowledgements

This research was Supported by National Science Foundation of China under Grant 61073172, Program for New Century Excellent Talents in University, Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant 200800030034, and National Basic Research Program ("973" Program) of China under Grant 2009CB320501.

10. References

[1] D. Meyer, L. Zhang, "Report from the IAB Workshop on Routing and Addressing". RFC 4984 (Informational), September 2007.

[2] BGP CIDR Report, http://bgp.potaroo.net/.

[3] Routeviews, http://www.routeviews.org.

[4] CAIDA Internet Topology, http://irl.cs.ucla.edu/topology/.

[5] X. Meng, Z. Xu. "IPv4 Address Allocation and the BGP Routing Table Evolution", SIGCOMM Computer Communication Review, vol. 35, no.1, pp. 71-80, 2005.

[6] B. Zhang, L.Wang, "FIB aggregation", Internet Draft. http://tools.ietf.org/html/draft-zhang-fibaggregation-01, 2009.

[7] Zartash Uzmi, Ahsan Tariq, "FIB Aggregation with SMALTA", Internet Draft. http://tools.ietf.org/html/draft-uzmi-smalta-00, 2010.

[8] Costas Kalogiros, Marcelo Bagnulo, "Understanding Incentives for Prefix Aggregation in BGP", In Proceedings of the 2009 workshop on Re-architecting the Internet, pp. 49-54, 2009.

[9] X. Zhang, P. Francis, "Scaling IP Routing with the Core Router-Integrated Overlay", In Proceedings of IEEE Internet Conference on Network Protocols 2006, pp. 147-156, 2006.

[10] H. Ballani, P. Francis, “Making Routers Last Longer with ViAggre”. In Proceedings of NSDI 2009, pp. 453-466, 2009

[11] E. Nordmark, M. Bagnulo. “Shim6: Level 3 Multihoming Shim Protocol for IPv6”, RFC 5533. http://tools.ietf.org/html/rfc5533.

[12] C. VOGT, “Six/One Router: A Scalable and Backwards Compatible Solution for Provider-Independent Addressing”, In Proceedings of MobiArch’08, pp. 13-18, 2008.

[13] X. Yang, D. Clark, “NIRA: A New Inter-Domain Routing Architecture”, IEEE/ACM Transactions on Networking, vol.15, no. 4, pp. 775-788, 2007.

[14] Wikipedia, http://en.wikipedia.org/wiki/Tier_1_network#List_of_Tier_1_networks.

[15] G. Huston, “Analyzing the Internet BGP routing table”, Internet Protocol Journal vol.4, no.1, pp. 2-15 2001.

[16] D. FARINACCI , V. FULLER, “Locator/ID Separation Protocol (LISP)”. Internet Draft, 2009.http://tools.ietf.org/html/draft-farinacci-lisp-12.

[17] D. Krioukov, K. claffy, “On Compact Routing for the Internet”, ACM SIGCOMM Computer Communication Review, vol. 37, no. 3, pp. 41-52, 2007.

[18] L. Cittadini, W. Mühlbauer, “Evolution of Internet Address Space Deaggregation: Myths and Reality”, IEEE Journal on Selected Areas in Communications, vol. 28, no. 8, pp. 1238-1249, 2010.

[19] D. Massey, L. Wang, “A Scalable Routing System Design for Future Internet”, In Proceedings of ACM SIGCOMM Workshop on IPv6, pp. 1-6, 2007.

[20] B. Tian, L. Gao, “On characterizing BGP routing table growth”, Computer Networks, vol.45, no. 1, pp. 45-54, 2004.

[21] Y. Xia, W. Zhang, "Can Longest Prefix Matching Make The Path Length Shorter?", JDCTA: International Journal of Digital Content Technology and its Applications, Vol. 4, No. 8, pp. 172 - 181, 2010

[22] W. Zhang, X. Yin, “Real Aggregation for Reducing Routing Information Base Size”, Journal of Convergence Information Technology, vol.5, no.6, pp. 47-53, 2010.