

## Real Aggregation for Reducing Routing Information Base Size

Weiguo Zhang, Xia Yin, Jianping Wu, Wei Zhang  
Dept. Computer Science and Technology, THU  
{zwg, yxia, wjp, zwei}@csnet1.cs.tsinghua.edu.cn

Shuangming Huang  
Network Management Center, CESEC  
palthsm@sina.com

doi: 10.4156/jcit.vol5.issue6.4

### Abstract

*It is commonly recognized that the Internet routing and addressing architecture is facing challenges in scalability, multihoming, and inter-domain traffic engineering. In this paper we study real aggregation (RA), which is a routing compression method by suppressing redundant RA prefixes of Routing Information Base (RIB). RA method can be used as a short-term solution because it does not require changes to routing protocols or router hardware and meets the need of incremental deployment. We design and implement different algorithms and evaluate their performance for full RIB and routing updates. The results show that RA method can reduce the RIB table size about 50% and naturally reduce the size of Forwarding Information Base (FIB). Furthermore, we propose two new measures for the stability of routing and apply them to the RIB collected from Route Views. The results show that these qualitative measures are rather effective for the stability of routing.*

**Keywords:** RIB, FIB, Prefix, Aggregation, Routing Scalability

### 1. Introduction

The rapid growth of the Internet during the past few years has led to increased concerns about the scalability of the underlying routing infrastructure. The most immediate concern is the rapid BGP routing table inflation and increasing update churns, which introduce large amount of state in the control plane and bring undue burdens to Internet Service Providers (ISPs). Thus the primary task is to improve the scalability of routing infrastructure.

Several solutions have been proposed by some pioneer research organizations such as IRTF RRG<sup>[1]</sup> and IETF GROW<sup>[2]</sup> working groups. Their proposals can be divided into long-term and short-term solutions. Core-edge separation schemes such as LISP<sup>[3]</sup>, APT<sup>[4]</sup>, Ivip<sup>[5]</sup> and TRRP<sup>[6]</sup> are typical long-term solutions. These long-term solutions seem to address the root causes of the routing table growth. However, these solutions adopt indirect ways to resolve the scalability of routing, which may invite new scalability problem such as the scalability problem of mapping system. Therefore, these fundamental changes to the Internet routing architecture and protocols will take a long time to realize. Furthermore, incremental deployment is one of the main barriers for long-term solutions. Short-term solutions such as Virtual Aggregation<sup>[7]</sup> and FIB Aggregation<sup>[8]</sup> are not going to substitute the long-term architectural solutions because it does not address the root causes of the routing scalability problem. Instead, they mainly aim to reduce the size of the FIB. These approaches are more practical because it can be accomplished by updating software or reconfigure the routers. Short-term solutions can co-exist and be complementary with any long-term solutions.

This paper investigates the feasibility of RA, which is a real aggregation method by suppressing redundant covered prefixes in the RIB. All prefixes in the RIB can be divided into two main categories: RA prefixes and non-RA prefixes. RA prefixes are defined as prefixes covered by any other prefix with the same next hop (except 0.0.0.0). Non-RA prefixes are defined as prefixes not covered by any other prefix with the same next hop. RA can be implemented only by updating software at a router and its incremental deployment is also viable.

Conventional methods to reduce routing table size are implemented either by aggregating multiple prefixes into a numerically representative single prefix, or by filtering “abnormal” prefixes, such as prefixes greater than 24, private IP addresses, unallocated addresses, multicast addresses and reserved

---

This research is under the support of National Basic Research Program of China (grant No. 2009 CB3 20502), National Key Technology R&D Program of China (grant No. 2008BAH37B03).

addresses etc. However, prefix filtering fails to eliminate redundant logical prefixes. Prefixes aggregation is also less appealing to be adopted, due in part to the slightest gain of aggregation, and due in part to the manipulation of multihoming and traffic engineering to split the prefix. Different from conventional ways, RA does not create any new prefix and only eliminate the existing RA prefixes. Furthermore, RA has little impact on multihoming and traffic engineering because RA only suppresses redundant routing information. Though RA might influence the routing decision of the remote Autonomous Systems (ASes), it is likely to be adopted by individual ISPs. Firstly, it is an inevitable trade-off between scalability and fine granularity routing operations. Secondly, fine-grained route policies can be reset by complex and effective BGP policies.

In this paper we perform a systematic analysis and evaluation of RA method. We first introduce the notion RA prefix and analyze the distribution of RA prefixes. We recognize that RA prefixes are unequally distributed and can be divided into some “levels”, each greatly attributing to the growing of RIB. Thus we design and implement RA method by different RA algorithms. To statically evaluate the full RIB with RA method, we give two full RA algorithms and calculate the RIB reduction size and computing overhead based on the RIB provided by RouteViews and Ripe. To handle dynamic routing changes, we provide another two RA update algorithms. Our simulation shows that RA can reduce RIB size about 50% and naturally reduce the FIB size. Finally, we propose two new measures for the stability of routing and apply them to the RIB collected from Route Views. The results show that these qualitative measures are rather effective for the stability of routing. Overall, since RA method can provide significant reduction in RIB size and support incremental deployment, it seems to be an ideal short-term solution.

This rest of the paper is organized as follows. In Section 2, RA prefixes are introduced briefly. In Section 3, RA method is presented, including design goals, methodology and algorithms. In Section 4, full RA algorithms and RA update algorithms are evaluated. In Section 5, two new qualitative measures for the stability of routing first are proposed, and then experiments and evaluation are performed. Finally, in Section 5, the conclusion is made.

## 2. RA Prefixes

Conventionally, IP prefixes in the BGP RIB are sorted into covering prefixes and covered prefixes<sup>[9]</sup>. This classification is based on one primary factor: numerical aggregation. Though these studies based on conventional classification can characterize prefix length distribution and analyze the global routing table growth, but they cannot provide effective techniques to reduce the size of RIB.

Different from previous work, we focus on not only the numerical aggregation, but also the same next-hop. Our new classification based on the numerical aggregation and the same next-hop aims to provide an effective scalability technique to reduce the size of RIB. According our classification, all prefixes in the RIB can be classified into RA prefixes and non-RA prefixes.

We use the RouteViews routing tables collected from 2001 to 2009 to calculate the ratio of the RA prefixes to the original RIB. From figure 1 we can estimate that about 49%-51% of the BGP table entries are RA prefixes and the rest are non-RA prefixes. Though RA prefixes and non-RA prefixes are by no means unchangeable, this ratio has been fairly stable over our study period. The causes of the existing RA prefixes may attribute to the factors such as multihoming, failure to aggregate, load balancing and address fragmentation.

### **Definition 1: level-m prefix**

Prefixes can be divided into some “levels”. Level-1 prefix is defined as a prefix not covered by any other prefixes. Recursively, a level-m prefix is covered by 1 to (m-1) level prefixes.

### **Definition 2: level-n RA prefix**

Level-1 RA prefix is defined as a prefix not covered by any other RA prefixes. Recursively, a level-n RA prefix is covered by 1 to (n-1) level RA prefixes.

As shown in the figure 1, level-5 RA prefix can be found in the RIB, which means too many redundant prefixes. RA prefixes are unequally distributed, the same as the prefixes allocated by IANA. Figure 2 shows the uneven distribution of RA prefixes. We can make the following observations: (1) Some RA prefixes can take up to 99% of whole prefixes while some prefixes

do not contain RA prefixes at all. (2) The lower the level, the higher the ratio of RA coverage in the RIB. Level-1 RA prefixes account for more than 75% and Level-5 RA prefixes take only less than 1% of all RA prefixes.

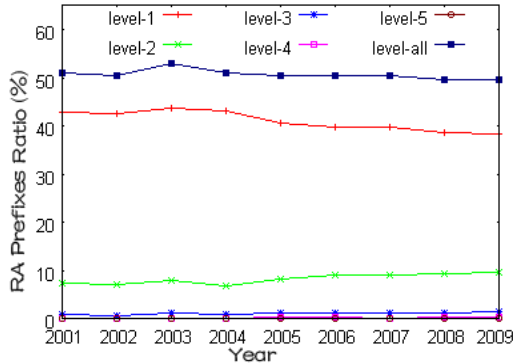


Figure 1. The ratio of RA prefixes to RIB

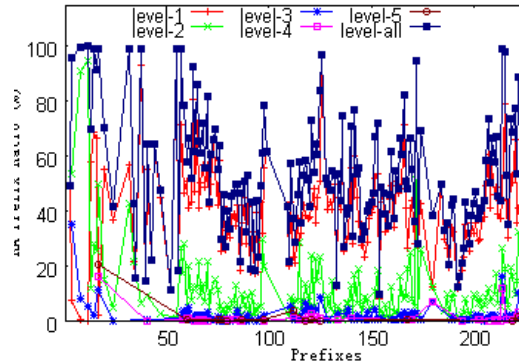


Figure 2. RA prefixes distribution

### 3. RA Design

#### 3.1. Design Goal

Different from conventional aggregation ways, RA method never introduces any new entry to the RIB. RA method aims to eliminate RA prefixes and shrink both the RIB and the FIB.

RA method also aims for deployability and hence is guided by three major design goals: 1) No changes to routing protocols; 2) effectively eliminate RA prefixes in the full RIB; 3) efficiently update an aggregated RIB upon a routing change. To achieve these goals, we design and implement different algorithms for full RIB and routing updates. These algorithms use a patricia trie to store IP prefixes and next-hop information for RIB. For a network device that uses other data structure to implement RIB, these algorithms still apply.

#### 3.2. Methodology

The term RIB stands for Routing Information Base (database) and refers to a routing table. Each RIB entry contains the destination IP prefix and associated route information. The Border Gateway Protocol (BGP) is the de facto exterior gateway protocol deployed on the Internet and maintains full AS path and many other attributes for each prefix in RIB. BGP RIB consists of Adj-RIB-In, Loc-RIB and Adj-RIB-Out [10]. Adj-RIB-In and Adj-RIB-Out store incoming and outgoing network advertisements received from other BGP speakers. Loc-RIB is the core database that stores routes that have been selected by this BGP device and are considered valid to it.

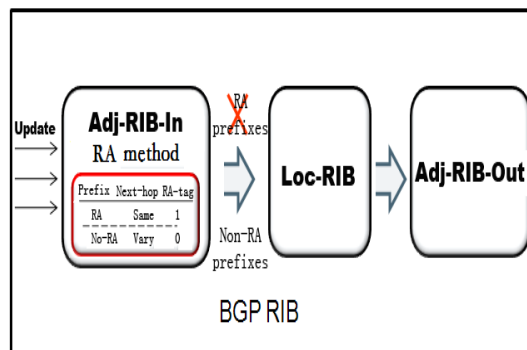


Figure 3. RA method implementation

To effectively eliminate RA prefixes in the RIB and meet the needs of multihoming and traffic engineer, RA method should be implemented in Adj-RIB-In for BGP RIB (Figure 3). The main reasons are as follow. Firstly, because RA method is accomplished in the front part of the BGP RIB, it can maximally simplify RA implementation. Secondly, local routing policies such as multihoming and traffic engineer will not be influenced by RA method. Thirdly, when an update message withdraws a prefix, all RA prefixes covered by the withdrawn prefix needs to be recovered. Adj-RIB-In contains unprocessed routing information that has been advertised to the local BGP speaker by its peers. Therefore, RA method can tag these RA prefixes in the Adj-RIB-In for future recovery.

### 3.3. RA Algorithms

To eliminate RA prefixes in the RIB, RA method should satisfy the requirement of routing correctness: every destination has a non-NULL next-hop and RA prefixes exist in the RIB. With these requirements satisfied, RA algorithms can work well.

We use a patricia trie-based RIB to store IP prefixes and their next-hop information. Compared with FIB aggregation, RA algorithm is more complicated: for the same prefix RA has to maintain different next-hop while FIB aggregation faces only one next-hop. Therefore, our RA algorithm adds new linked list to store different next-hop information. Though our RA algorithms use patricia trie, the general techniques can be applied by other network device.

**Full RA Algorithms:** There are two RA algorithms for the full RIB. A relatively simple algorithm, called SFR (Simple Full RA algorithm) is implemented by traversing the tree recursively from the root node in postorder. When it arrives at a node with a prefix, it compares all next-hops of this prefix with its immediate ancestor prefix. The immediate ancestor prefix is found by looking up the nearest ancestor node. The other algorithm is relatively complex algorithm, called CFR (Complex Full RA algorithm). CFR also traverses recursively the tree in postorder. When it arrives at a node with a prefix, the CFR tries to compare all next-hops of this prefix with its all ancestor prefixes. The SFR algorithm is simple but not complete to eliminate all RA prefixes in the RIB while the CFR is on the contrary.

**RA Update Algorithms:** Because of the routing dynamics, routing table updates occur from time to time. Though full RA algorithm can ideally eliminate RA prefixes, it will incur excessive computation overhead. Therefore, we need to tradeoff efficiency with overhead. We design and implement two incremental update algorithms: SRU (Simple RA Update algorithm) and CRU (Complex RA Update algorithm). When adding new prefixes, two algorithms are the same. The new prefix node compare the next-hop with not only its immediate ancestor prefix, but also all the nearest prefix descendants. However, when handling a withdrawn prefix, e.g. prefix A, SRU simply deletes A and restores the next-hop for each of A's nearest descendants. While CRU must compare the next-hop of A's immediate ancestor with that of A's each descendant before delete A.

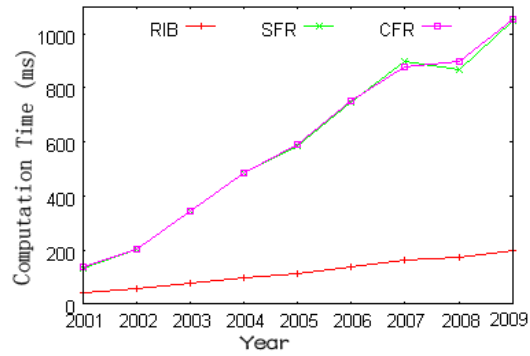
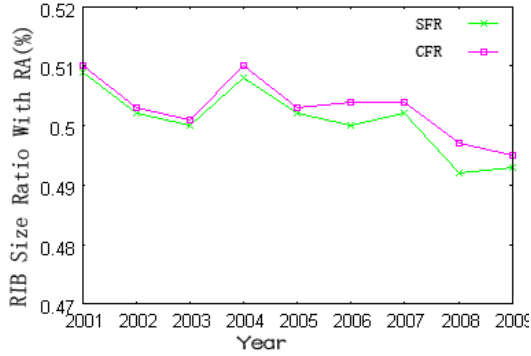
## 4. RA Evaluation

All RA algorithms are implemented in C and the evaluation is based on a Linux machine. By using public routing tables from RouteViews and Ripe, we evaluate RA algorithms for the RIB size reduction and computing overhead.

### 4.1. RIB Size Reduction and Overhead

We apply full RA algorithms to RouteViews routing tables from 2001 to 2009. We calculate the ratio between the FIB size with RA method and the original RIB size. Figure 4 shows, with our full RA algorithms, the RIB size can be reduced about 50%.

Figure 5 shows the computing time in the case of the original RIB, SFR and CFR. Computing overhead is no more than a few hundred milliseconds. However, SFR and CFR consume more and more computing time than the original RIB as the Internet moves on. The main reason is that the meshed connectivity of multihomed nodes increases and RA method has to deal with more and more next-hop information. Compared with SFR, CFR can eliminate RA prefixes thoroughly but only consumes slightly more time than SFR.



**Figure 4.** RIB compression ratio with RA method **Figure 5.** Computing times for full RA algorithms

## 4.2. Incremental Update Handling

We evaluate our incremental update algorithms by using one month (December 2009) of BGP routing updates collected by RouteViews. The result is presented in Table 1. Over 0.32bn routing updates were observed during this month. There were about half part of the RIB changed for SRU and CRU. Update handling algorithm has minimal impact on the processing time for route updates because each route update took at most 3-4 $\mu$ s to process.

In order to verify the validity of our incremental update algorithms, we design and implement a comparison of full RA algorithm and incremental update algorithm. First we calculate the size of RA prefixes for the RIB collected by Route Views at 31 December 2009 and the size is around 6078676. Furthermore, we do the same work at 30 November 2009 and continue applying our incremental update algorithms until 31 December 2009. Finally, we get the size of RA prefixes, which is around 6038972. Because the error ratio is only 0.65%, it is feasible for our incremental update algorithms.

RA is particularly appealing because RA can effectively improve the routing stability. About half of routing updates can be removed from the RIB by RA algorithms. Therefore, this routing updates will not be announced to peers.

## 4.3. Comparison with Other Methods

RA technique is based on RIB aggregation with the same next-hop. Table 2 shows a comparison of Virtual Aggregation (VA), Route Filtering (RF) [11], FIB Aggregation (FA) and RA. Firstly, RA and RF can reduce the size of RIB and FIB while VA and FA only reduce the size of FIB. Secondly, RA and RF have global impacts on other network-wide operations while VA and FA's impacts are limited within single local ISP, AS or router. Thirdly, RA and FA can be done by software update at a router while VA and RF are simple by a "configuration-only" approach. Fourthly, VA imposes stretch on traffic and other methods increase computational overhead. Lastly, because VA and FA only reduce the size of FIB and RF only filters "abnormal" prefixes, they are not ideal solutions for the routing scalability problem. RA can reduce the size of RIB and FIB by eliminating redundant prefixes, but the ratio of RIB aggregation is about 50%, RA is slightly better than the former three methods.

**Table 1.** Results of incremental update algorithms

	<b>RIB</b>	<b>SRU</b>	<b>CRU</b>
No. of changes(bn)	0.32	0.17	0.15
Total Proc. Time(s)	586	1535	1629
Avg. Proc. Time( $\mu$ s)	1.84	4.82	5.12
No. of Tire Nodes Affected	1.00	1.01	1.01

**Table 2.** Comparison with other methods

	<b>VA</b>	<b>RF</b>	<b>FA</b>	<b>RA</b>
RIB/FIB	FIB	RIB/FIB	FIB	RIB/FIB
Range	ISP/AS	Router	Router	Router
Impact	Local	Global	Local	Global
Way	Command configure		Software update	
Fault	Path stretch	Filter time	Lookup time	
Scalability	Weak	Weak	Weak	Middle

## 5. Scalability Measurement

### 5.1. Aggregatable-Prefix-Distance and RA-Distance

The size and growth rate of RIB and FIB are two conventional criteria to measure the scalability of routing. Though these two measures can visually reveal the routing scalability problem, they are quantitative indexes. To make further research in the routing scalability problem, we propose two new qualitative indexes: Aggregatable-Prefix-Distance and RA-Distance.

**Definition 3: Aggregatable-Prefix-Distance (APD)**

APD is defined as the level difference between two numerically aggregated prefix nodes in a patricia trie-based RIB. APD is denoted by the sign  $D_{AP}$ .

**Definition 4: RA-Distance (RD)**

RD is defined as the level difference between two numerically aggregated RA prefixes nodes in a patricia trie-based RIB. RD is denoted by the sign  $D_{RA}$ .

Since APD and RD can significantly reveal the aggregation ability for the RIB, they can be served as qualitative measures for the scalability of routing. The smaller the value of APD and RD, the better the scalability of routing. An ideal scalable routing architecture, average APD and RD for the RIB should be zero or near-zero.

### 5.2. Experiment and Evaluation

We measured the average DAP and DRA by using the RIB collected from RouteViews from 2001 to 2009. The results are presented in Figure 6. First of all, the average value for DAP and DRA is far greater than zero and the trend moves in an ascending path, which implies that the routing scalability problems are more and more serious. Moreover, the similarities in curve development between DAP and DRA are so striking. This illuminates that strong correlation between DAP and DRA. RA prefixes can be used to effectively analyze the features of the routing scalability problems. Eventually, the average value for DAP and DRA ranges from 1.1 to 1.4, which means level-1 prefix and RA prefix are in the majority. This can explain why the results for our two RA algorithms show not much difference.

Figure 7 shows the change of maximal DAP and DRA. For one thing, maximal DAP and DRA are very large. DAP reaches at least 10 and DRA is no less than 5. This shows that the aggregation ability for the current RIB is bad. For another, maximal DAP is far greater than DRA. This shows that RA method can only remove those prefixes that have the same next-hop information with their covering prefixes. However there still exist considerable non-aggregatable prefixes with different next-hop information in the RIB. Therefore, RA method can partially alleviate routing scalability problem and be used as an effective short-term solution. A substantial solution of the routing scalability problems can only be found in an architectural reform.

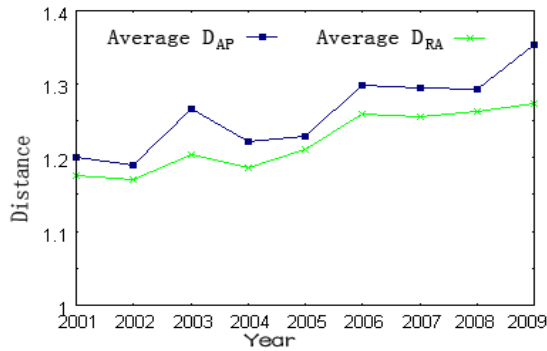


Figure 6. Average D<sub>AP</sub> and D<sub>RA</sub>

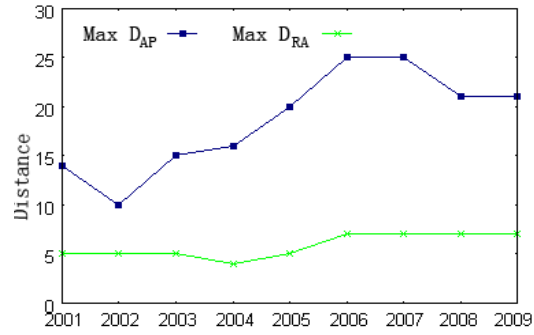


Figure 7. Maximal D<sub>AP</sub> and D<sub>RA</sub>

## 6. Conclusion and Future work

RA method aims to reduce the size of RIB by eliminating redundant RA prefixes. It does not require changes to routing protocols or router hardware and meets the need of incremental deployment so that RA method seems to be an ideal short-term solution. By implementing and evaluating RA algorithms, we find that our algorithms can reduce the RIB size about 50% and naturally reduce the FIB size. Furthermore, we propose two new measures for the stability of routing and apply them to the RIB collected from Route Views. The results show that these qualitative measures are rather effective for the stability of routing.

Whether fundamental changes are necessary for the routing scalability problem is an open issue. After all, hierarchical routing is the only known way to address scalability problem so far. From the results of RA prefixes distribution (see Figure 1), we can find that RA method can get a better aggregation ratio if there are numerous “large” prefixes with short prefix length in the RIB. Therefore, how to introduce much more “large” prefixes is critical to RA method. Maybe RA with good IP address allocation policy can effectively address the routing scalability problem on the premise of meeting the need of multihoming and traffic engineer. For the future work, we plan to study IP address allocation policy and propose a long-term solution based on RA and effective IP address allocation policy.

## 7. References

- [1] IRTF Routing Research Group. <http://www.irtf.org/charter?gtype=rg&group=rrg>.
- [2] IETF Global Routing Operations (GROW). <http://www.ietf.org/dyn/wg/charter/growcharter.html>.
- [3] D. Farinacci, V. Fuller, and D. Oran, “Locator/ID Separation Protocol (LISP)”, draft-farinacci-lisp-00.txt, 2007.
- [4] D. Jen, M. Meisel, D. Massey, L. Wang, B. Zhang and L. Zhang, “APT: A Practical Tunneling Architecture for Routing Scalability”, Technical Report 080004, UCLA, 2008.
- [5] R. Whittle, “Ivip (Internet Vastly Improved Plumbing) Architecture”, Draft whittle-ivip-arch-02, August 2008.
- [6] W. Herrin. “Tunneling Route Reduction Protocol (TRRP)”, <http://bill.herrin.us/network/trrp.html>.
- [7] H. Ballani, P. Francis, C. Tuan, and J. Wang, “Making Routers Last Longer with ViAggre”, In Proceeding(s) of NSDI, 2009.
- [8] B. Zhang, L. Wang, X. Zhao, Y. Liu, and L. Zhang. “FIB Aggregation”, draft-zhang-fibaggregation-02.txt, October 2009.
- [9] X. Meng, Z. Xu, B. Zhang, G. Huston, S. Lu, L. Zhang, “IPv4 address allocation and the evolution of the BGP routing table”, in Proceeding(s) of ACM SIGCOMM, pp. 71-80, 2005
- [10] Y. Rekhter, T. Li, and S. Hares, “A Border Gateway Protocol (BGP-4)”, RFC 4271, 2006.
- [11] S. Bellovin, R. Bush, T. Griffin and J. Rexford, “Slowing routing table growth by filtering based on address allocation policies”, Unpublished, 2001.