

Revisiting the Impact of Traffic Engineering Techniques on the Internet's Routing Table

Pedro A. Aranda Gutiérrez

University of Paderborn, Germany
paaguti@hotmail.com

Abstract. This paper studies the effect of simple Traffic Engineering techniques on the size of the Internet's default free routing table. Current best practises for traffic balancing in the Internet are based in disaggregating prefixes that cause an increase in size of the Internet's core routing table. An algorithm to show the impact of these techniques on the growth of the routing table is proposed. This algorithm is applied on routing tables between January 2001 and December 2009 and the results are discussed. Finally an alternative architecture is proposed, which allows Traffic Engineering while keeping the Internet routing table size optimised.

Keywords: Routing protocols, Network Operations, Network management, Network monitoring.

1 Introduction

IP routing protocols control the exchange of network layer reachability information between nodes in an IP network. This information, also known as routing information, is used to build the routing table. IP addresses in a router are grouped into ranges, which are known as prefixes. The packet forwarding process in IP nodes, which computes the outgoing port for an incoming IP packet, is controlled by the routing table and uses longest mask prefix matching on the destination address to compute the output port. The two basic concepts which have to be understood are longest mask prefix matching and the Route Decision Process. Longest mask prefix matching implies that a router will always prefer the most specific routing information installed in the routing table to reach a given IP address. The Route Decision Process (RDP) is specific for each routing protocol used by the router and defines the way routing information received from neighbours is treated and when routes are installed in the routing table.

The Internet is an IP infrastructure which is divided into different independent and interconnected domains, which are known as Autonomous Systems (ASes) [12]. Each AS is assigned addressing space in the form of one or more prefixes by the Internet Routing Registries (IRRs) of the region it belongs to. It is only allowed to advertise routing information for the addressing space it has been assigned. While there exist different routing protocols to control the routing information exchange within an AS, BGP-4 [21] is the only routing

protocol which controls the routing information between ASes. ASes which are interconnected are said to be peering. Internet Service Providers (ISPs) control the traffic distribution on their peering links, because they need to assure that no traffic is lost when a link or border router fails, or that traffic levels are such, that the network meets Quality of Service (QoS) criteria, etc. Internet Service Providers have developed different strategies to cope with all these requirements. In order to assure resilient connectivity, most ISPs are *multi-homed*. i.e. have more than one upstream connection, sometimes to more than one upstream provider.

The rest of this paper is structured as follows: Section 2 discusses how ISPs or larger sites with more than one access to the Internet implement Traffic Engineering with BGP-4 and focuses on two techniques used to balance the inbound traffic of an AS, which can be considered current best practises. Section 3 presents the Internet routing table compression algorithm used to estimate the impact of the use of current best practises for Traffic Engineering in the Internet. The algorithm is then applied on data from the RIPE's Routing Repositories to quantify the impact of Traffic Engineering (TE) techniques on the current Internet routing table. Section 4 presents an alternative to control the growth of the Internet routing table while allowing for TE solutions and shows the impact on the Internet's core routing table if it had been applied between January 2001 and December 2009. Finally, Section 6 presents the conclusion.

2 BGP-4 and Traffic Engineering

Internet Service Providers organise their interconnection through peering agreements, which include the definition of technical and economical conditions under which they exchange traffic. The technical definitions include addressing space which is made mutually accessible, the mechanisms to route traffic through the interconnection links and Internet Protocol layer parameters like round trip delay and tolerated levels of packet loss and acceptable traffic levels for the in- and outbound links. Service Level Agreements (SLAs) introduce an additional incentive for mechanisms to control the in- and outbound traffic of a network and, thus, for the implementation of TE techniques. However, BGP-4 as the inter-domain routing protocol of the Internet lacks real TE capabilities. Despite this, routing configurations targeting simple load balancing between independent links to a major upstream ISP to load sharing between several upstream ISPs have been deployed from the early days of the commercial Internet. These are documented in vendor manuals [5], [20] and books about BGP-4 [14], [11]. In order to arrive as close as possible to the desired traffic distributions, attributes in the routing advertisements are manipulated in order to influence the routing decision process. Two examples are shown in Section 2.1. These configurations are considered current best practises [4], [14].

Controlling the inbound traffic of an AS implies influencing the routing decisions of other ASes. As Griffin and Wilfong have demonstrated in [10], predicting BGP-4 behaviour is impossible. BGP-4 is not always guaranteed to converge to

one single solution in the presence of policies. Since the effect of configuration changes cannot be predicted, arriving at traffic conditions that comply to the SLAs signed between an AS and its peers is an iterative process of Trial and Error based on deploying a certain routing configuration, assessing its quality by the traffic distribution it creates in the inter-provider links, refining the configuration and reassessing. This process aims at an ideal traffic distribution with respect to some objective, e.g. minimisation of peering costs, uniform traffic distribution, etc.

In order to achieve the best approximation to the ideal traffic distribution, ISPs fraction their addressing space. G. Huston [13] recognises the use of this technique and examines its impact on the routing tables of the Default Free Routing area of the Internet. But, as Section 3 shows, the impact is greater than Huston's graphics imply.

2.1 Current Best Practises to Control the Inbound Traffic

Prefixes are sets of contiguous IP addresses, designated by a base address (B) and a mask (B_M). In the case of IPv4 the base address and the mask are unsigned 32-bit integers and in the case of IPv6, they are 128-bit long integers. The mask B_M has its M most significant bits set to 1 and the rest set to 0.

Definition 1. B/M represents a valid prefix \mathcal{P} if and only if $B \wedge B_M = B$.

Definition 2. Let \mathcal{P} be the set of addresses represented by prefix B/M and A be an IP address of the same family as B then:

$$A \in \mathcal{P} \Leftrightarrow A \wedge B_M = B$$

Definition 3. Subnetting: Let \mathcal{P} be the set of addresses represented by prefix B/M . \mathcal{P} can be divided in two subsets, known as sub-networks $\mathcal{P}_1, \mathcal{P}_2$ that contain the same amount of IP addresses. This implies that the mask length will be incremented by 1. The sub-networks will be denoted as $\mathcal{P}_1 = B/M + 1$ and $\mathcal{P}_2 = B_1/M + 1$ in the rest of this paper.

Multihoming to one provider. Figure 1(a) shows an AS which is providing Internet access to another AS through two connections. The downstream ISP owns prefix B/M and has subnetted it into two subnetworks $B/M + 1$ and $B_1/M + 1$. It advertises the three prefixes and uses an agreed marker (a.k.a. community) to signal the preference. The upstream ISP translates this marker to his local preference and uses it to filter the advertisements to *the Internet*. Internally, due to longest prefix matching, the ISP will use the sub-networks $B/M + 1$ and $B_1/M + 1$ to direct the traffic towards its client. This configuration also assures redundancy: in case one link fails, the full routing information is available through the other. Figure 1(a) shows the ideal situation, where the sub-networks advertised for TE purposes do not progress to the Internet. In real life, as shown in Section 3, most of the TE sub-networks progress to the Internet.

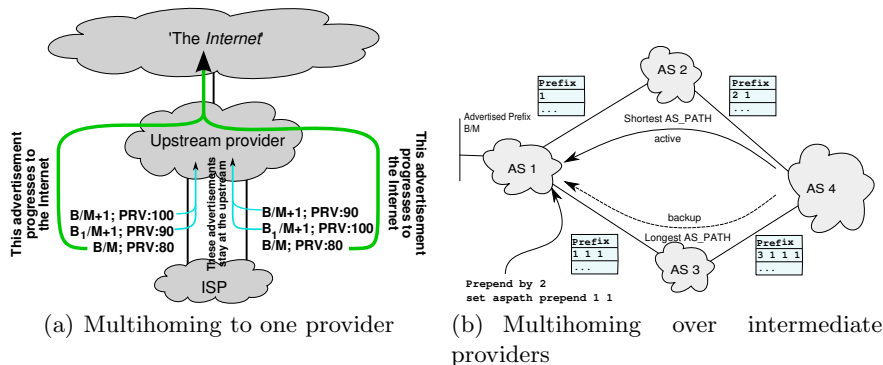


Fig. 1. Current Practises in Multihoming: fractioning the assigned addressing space to balance the incoming traffic

Multihoming to more than one provider. The technique described in the previous section only works for direct client-provider peerings, because communities are not guaranteed to progress beyond the first AS and the local preference is only valid within the AS where it is defined. The only attribute which is guaranteed to progress through Autonomous Systems is the Autonomous System Path (AS_PATH). The number of hops it contains is one of the first variables used by the RDP to calculate the best path. On each interdomain border, BGP-4 speakers prepend their Autonomous System Number (ASN) at the beginning of the AS_PATH attribute when exchanging routing information with speakers outside their AS. Additionally, the AS_PATH attribute can be manipulated by AS_PATH Prepending, a technique which consists in prepending the ASN more than once. AS_PATH Prepending is the result of routing policies programmed in routers and cannot happen by protocol interaction due to loop protection mechanisms.

Figure 1(b) shows how the Autonomous System AS1 is signalling to AS4 to prefer the path through AS2 over the path through AS3 to send traffic to the prefix B/M . As in the case shown in Figure 1(a), this configuration assures a main path and an alternative in case the main path fails. And as in that case, the prefix B/M assigned to the ISP might well be one of the sub-networks $B/M+1$ or $B_1/M+1$. As shown further on, most downstream providers also use AS_PATH Prepending instead of communities when multi-homing to one provider, in order to gain some independence from the upstream providers’ policies.

3 Assessing the Room for Optimisation in the Internet’s Core Routing Table

The routing table in Internet core routers is a data structure which holds basically two types of data: reachability information and its attributes. The reachability information of an entry is the prefix. The attributes include the next-hop.

This data structure is also known as the Routing Information Base (RIB). The IP packet switching process in a router is controlled by the Forwarding Information Base (FIB). The FIB maps prefixes to their output interface and is generated by combining the RIBs of all active routing processes.

Prefixes are assigned in the public addressing space as per RFC 1930 [12] to Autonomous Systems needing public IP addresses by the different Regional Internet Registries. Prefixes from the public addressing space are biunivocally linked to their Autonomous System Number (ASN) at any given moment in time. In order to detect the configurations presented in Section 2.1 and eliminate sub-netting, the Internet routing table is modelled as a directed graph. The root of the graph is the router the routing table was extracted from and the Autonomous Systems (ASes) are the vertices of the graph. The leaves of the graph are $\{AS, Prefix\}$ pairs that represent the address allocations made by the different Internet Routing Registries (IRRs) to the ASes in their regions.

3.1 A Routing Table Compression Algorithm for the Internet

Algorithm 1 shows the proposed compression algorithm. It is not intended to be applied directly in routers, but rather helps assessing the overhead introduced by current best practises to implement multi-homing over different providers, load balancing, precaution against prefix-hijacking, etc. Therefore, optimisations were not sought and computational time analysis was not performed. As discussed in the conclusion, the aforementioned techniques can be ported to the new proposed architecture or implemented with alternative technologies. The algorithm is applied until no further optimisations can be introduced in the routing table. The concepts of sub- and supernetting are interpreted restrictively, in the sense that prefixes are associated to the AS that originated them and to the Autonomous System Path (AS_PATH) they are received through and sub- or supernetting is only allowed when both prefixes belong to the same AS and are reached through the same sequence of Autonomous Systems. The algorithm uses the following functions to check for possible optimisations:

- `nextAggregation(prefix)` decrements the prefix mask length by one
- `IsFeasible(prefix)` checks whether the prefix is correct as per Definition 1.
- `Contains(prefix1, prefix2)` checks that both prefixes are originated by the same Autonomous System and that *prefix₂* is completely contained in *prefix₁* and that both prefixes are reached following the same AS_PATH.

Algorithm 1 does not affect the reachability of hosts in the Internet. The BGP-4 routing table of a router is a directed graph. The root of the graph is the router itself and each leaf contains a prefix that can be reached from the root paired with its AS. The other nodes of the graph represent the ASes traversed by a packet on his way to a given prefix. This graph has two types of edges; the regular edges connecting two nodes and the irregular edges connecting a node with itself, which are discarded by the algorithm. Algorithm 1 will only merge two paths of the directed graph if they share all nodes expect the leafs and if the

Data: An Internet routing table as an array of $\{prefix, AS_PATH\}$ pairs

Result: The Internet table with one level of optimisation and a flag indicating whether the routing table was modified or not.

```

changed ← false;
foreach index = 0 to length(InetTable) - 2 do
  this_Prefix ← InetTable[index];
  next_Prefix ← InetTable[index + 1];
  aggregateThis = nextAggregation(this_Prefix);
  if IsFeasible(aggregateThis) then
    if Contains(aggregateThis, nextPrefix) then
      /* remove next_Prefix from the Internet table */
      removeFromTable(InetTable[index + 1]);
      /* replace this_Prefix with the aggregation */
      InetTable[index] ← aggregateThis;
      /* signal that the table has changed */
      changed ← true;
    end
  end
end
return changed

```

Algorithm 1. Routing table compression algorithm

leafs refer to prefixes that can be aggregated and are assigned to the same AS. The algorithm respects the address allocations made by the IRRs and the paths followed by packets at AS level and thus produces equivalent routing tables, in the sense that packets will arrive to their assigned destinations.

3.2 Status Quo

Algorithm 1 was applied on the routing table contributed by collecting device 203.119.76.3 to the RRC00 repository on the 5th of September, 2009. The initial routing table size was 287,414 routes. After 10 iterations, the algorithm was not able to reduce the table further. The resulting routing table had a size of 185,334 routes. Speaking in relative terms, the table could be reduced by 35.5% without losing connectivity. Figure 2 shows the reduction achieved on the first 6 iterations. It is worth to be noted that the first iteration reduces the routing table size by 22.4%. The improvement in the routing table size is mainly obtained in the first 4 iterations, where a reduction in size by approx. 35% is achieved. Figure 2 shows the compression ratio and the amount of routes suppressed after each iteration. The proposed algorithm suppresses more than 100.000 routes after the third iteration.

4 An Alternative Traffic Engineering Architecture for IP

The previous section shows that there is a significant amount of disaggregated routing information in the Internet's routing table, which is scoped. Thus, for

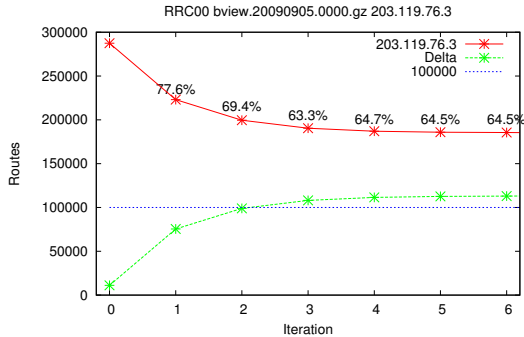


Fig. 2. Reduction achieved on the first 6 iterations for a default free routing table from September 2009

example, the target of the disaggregated information in the configuration shown in Figure 1(a) is the Provider, while in Figure 1(b), the target is a distant AS, which sometimes is not known to the provider. T. Li proposed a new BGP-4 attribute to limit the scope of an advertisement by the number of hops in 2007 in an Internet draft [17], which was later abandoned. In this section, a new architecture with techniques to enforce scopes to BGP-4 advertisements is proposed.

Since the Internet has become a basic infrastructure on which many critical applications rely, any proposal to enhance it must provide smooth migration mechanisms. This is one of the reasons, why the migration to IPv6 is taking so long. The architecture proposed in this paper is backward compatible, non-mandatory and might be adopted incrementally in different regions in the Internet. It is based on the principle that information which is essential for routing purposes should be kept in the Internet's routing table, while the remaining routing information can be migrated to a parallel routing table, which is managed by the Internet Service Providers (ISPs) involved in a certain Traffic Engineering (TE) configuration.

Figure 3 shows the relationship between the different routing tables and the forwarding table in a TE enabled router. The support for the best aggregations would be implemented by the left side of the figure. These components are part of the current router architectures. Support for interdomain TE routes is added on the right side of the figure and is basically a replica of the current BGP-4 implementation. In order to keep the information exchange for the main routing table isolated from the information exchange for the TE routing table isolated, the use of Multiprotocol Extensions for BGP-4 (MP-BGP) is proposed. MP-BGP is widely used when routers exchange other routing information in addition to pure IPv4 routing information (e.g. IPv6 in RFC 2545 [18]), and when partial IPv4 routing information needs to be kept isolated from the main IPv4 routing table (e.g. IPv4 VPN in RFC 2917 [19]).

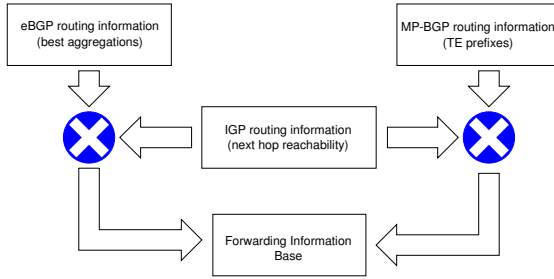


Fig. 3. A high level Traffic Engineering routing architecture

4.1 Advantages

One of the main advantages of the proposed architecture is that the routing setups needed for Traffic Engineering are kept local at the providers which are concerned by them and do not trickle into the Global Internet routing table. Figure 4 shows a routing configuration with unintended side effects which would be avoided with the proposed architecture. Additionally, as shown in Figure 5(a), it would mean going back in time to mid-2006 with regard to routing table sizes, with a routing table size reduction of approx 33% or 100.000 routes (see Figure 5(b)).

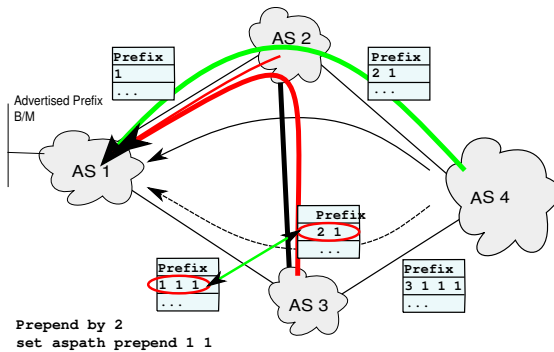


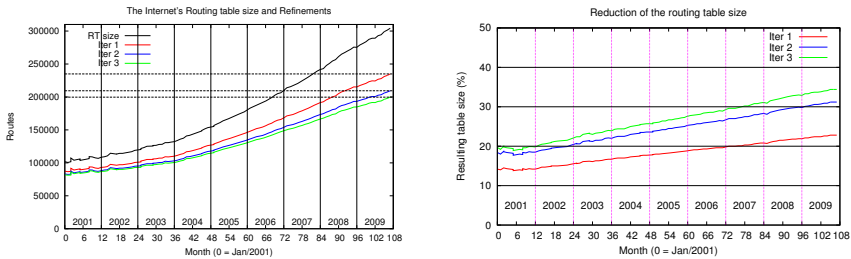
Fig. 4. Unintended side effects of Traffic Engineering using AS_PATH Prepending

There should not be a significant impact on the ability of transit ISPs to implement traffic balancing based on their clients' advertisements. When multi-homing to one provider, as presented in Figure 1(a), the upstream ISP is expected to have enough clients to arrive at a near optimum traffic distribution using the clients' best aggregations. For the case presented in Figure 1(b), AS1 and AS4 have to exchange the MP-BGP information. This can easily be done using multi-hop eBGP configurations (see [11], [14], etc.). AS2 and AS3 do not experience the hassle of having to take into account AS1's routing configurations. Figure 4

shows how, with current configurations, the traffic from a directly connected AS (AS3) can be diverted to another AS (AS2) by the TE routing configurations of AS1.

4.2 Historical Evolution

In order to study the historical evolution of the Internet’s default free routing table, data from the RRC00 repository were used. The first routing table dump files of the months between January 2001 and December 2009 were used and the route collector (IP = 203.119.76.3) was selected because it contributed data to all analysed table dump files.



(a) Routing table size on the first 3 iterations (b) Reduction achieved on the first 3 iterations

Fig. 5. Evolution between January of 2001 and December of 2009

Figure 5(a) shows the evolution of the collector’s routing table size and of the resulting routing table after the first three iterations. The horizontal dotted lines mark the size of the December, 2009 routing table size after the first three iterations. They show that after the first iteration, the resulting routing table has the size of the unreduced Internet routing table of beginning of 2008. After the second and third iterations, the resulting size is that of the Internet’s routing table of 2007.

One of the measures which can be used to study the trend in the use of subnetting is the efficiency of iteration i defined as $\rho_i = 1 - \frac{N_i}{N_0}$, where N_i is the number of routes in the resulting routing table and N_0 is the initial routing table size. ρ_i is the percentage of routes which could be effectively removed. Figure 5(b) shows that this percentage has been growing steadily since the end of 2001. The relationship of this measure with the “natural” fragmentation of the Internet’s routing table is quite unsettling. The process by which the IRRs assign addresses to ISPs implies some level of fragmentation of the routing table. But if ISPs were not using subnetting for Traffic Engineering purposes, the reduction algorithm would not be able to reduce the routing table: Between successful addressing space allocations to a specific ISP, IRRs continue allocating addressing space to other ISPs. The probability that an AS gets adjacent and aggregatable addressing space in two consecutive applications is almost zero.

This confirms the effectiveness of the compression algorithm in suppressing subnetting introduced by the ASes with configurations akin to Figure 1(a) while preserving connectivity.

5 Related Work

The algorithm proposed in this paper shows that there is room for optimisation in the Internet's routing table, which would result in improved scalability and manageability. Fall et al. [7] have recently studied the impact of the size of the Internet's routing table on cost and CO_2 footprint and conclude that Moore's law will be able to cope with the growth of the routing table size. The architecture proposed in this paper reduces the complexity of the routing table and of the dynamics which can be linked to TE techniques, with the objective of reducing the OPEX of ISPs.

Other work related with the compression of the Internet's core routing table includes the virtual aggregation proposal ViAggre [3]. Its main drawback is, as the authors recognise, that they manipulate the routing tables and the results might divert the traffic to different paths, even at the Autonomous System level. Freedman et al. also study the aggregation level of the Internet's routing table in [8]. They conclude that geographic dispersion of IP prefixes reduces the level of compression which can be achieved when looking for the best aggregations in the Internet's routing table. The work described in the present paper shows that, limiting aggregation prefixes which share the same sequence of ASes produces significant savings. The aggregated prefixes fall under the multi-homing scenario depicted in Figure 1(a) and are likely to be geographically adjacent. This adjacency should be dealt with within the different Internet Service Providers and not affect the Internet's core routing table.

Suri et al. [22] study the compression of routing tables which take into account the source and destination address fields in IP packets. This kind of routing tables is significant for edge devices of the Internet like access routers. The approach of this paper differs in two main aspects from the approach proposed by Suri: firstly, this paper concentrates on core devices in the Internet, where routing is done based on the destination address only. Secondly, while Suri's approach can be implemented in routing devices, the algorithm presented in this paper is intended for assessing the amount of optimisation which is achievable.

Routers have evolved to complex systems with multiple routing protocols building concurrently the routing table [9, 11, 14]. These devices have a FIB and multiple RIBs. Draves et al. [6] propose an algorithm to compress the routing table which is better applicable to the FIB of a router than to the RIB. This is so, because they only take into account the destination address and the next hop. While their proposed algorithm provides a highly efficient tree to determine the next hop, essential information regarding the path followed by the packets is lost. Therefore, Draves' algorithm is incompatible with BGP-4.

Some level of de-aggregation in the Internet's routing table is used to protect ASes against prefix hijacking [15, 1]. In this scope, the INTERSECTION project

[2] proposes alternative techniques to detect and provide remedy in case of prefix hijacking. In general, prefix hijacking should not be remedied by introducing additional, more specific routes to the Internet's table but by filtering: if an AS advertises a prefix it does not own, the upstream providers should filter it out and notify the offending AS. Systems like INTERSECTION allow for quick detection of suspicious prefixes. PHAS [16] is aligned with the INTERSECTION approach.

6 Conclusion and Further Work

This paper shows that ISPs massively use subnetting techniques as part of their Traffic Engineering (TE) implementations by using an algorithm to find best aggregations in the Internet's routing table, which achieves 33% optimisation rates on current routing tables. This algorithm, however, is not intended to be applied directly on routers. This paper also shows that the trend in the current Internet is to increase the use of TE techniques and therefore to decrease the optimisation of the routing table. To remedy this and render the infrastructure more controllable, an alternative TE architecture for IP networks is proposed.

Further work on this architecture includes modelling and simulating this architecture, and studying its impact on proposed IPv6 transition mechanisms and on the architecture of the IPv6 Internet in the long run.

Acknowledgement

This work was made possible by the extensive BGP-4 data collections of the RIPE's Routing Repository.

References

1. YouTube Hijacking: A RIPE NCC RIS case study, <http://www.ripe.net/news/study-youtube-hijacking.html>
2. INTERSECTION (INfrastructure for heTEroogeneous, Resilient, SEcure, Complex, Tightly Inter-Operating Networks) (January 2008), <http://www.intersection-project.eu/> (last visit June 25, 2010)
3. Ballani, H., Francis, P., Cao, T., Wang, J.: Making routers last longer with ViAggre. In: NSDI 2009: Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation, pp. 453–466. USENIX Association, Berkeley (2009)
4. Green, B.R., Smith, P.: CISCO - ISP Essentials. Cisco Press (September 2002)
5. Cisco Systems Inc. Interworking technology handbook
6. Draves, R., King, C., Venkatachary, S., Zill, B.D.: Constructing optimal ip routing tables. In: Proc. IEEE INFOCOM, pp. 88–97 (1999)
7. Fall, K., Iannaccone, G., Ratnasamy, S., Godfrey, P.B.: Routing Tables: Is Smaller Really Much Better? In: Proceedings of Hotnets 2009. ACM, New York (2009)
8. Freedman, M.J., Vutukuru, M., Feamster, N., Balakrishnan, H.: Geographic locality of ip prefixes. In: IMC (2005)

9. Gredler, H., Goralski, W.: The Complete IS-IS Routing Protocol. In: Computer Science. Springer, London (2005)
10. Griffin, T.G., Wilfong, G.: An analysis of BGP convergence properties. In: Proc. of SIGCOMM 1999, pp. 277–288. ACM Press, New York (1999)
11. Halabi, S.: Internet Routing Architectures, 2nd edn. Cisco Press (2000)
12. Hawkinson, J., Bates, T.: Guidelines for creation, selection, and registration of an Autonomous System (AS). RFC 1930 (Best Current Practice) (March 1996)
13. Huston, G.: Analysing the Internet BGP Routing Table. The Internet Protocol Journal 4(1) (2001)
14. van Beijnum, I.: BGP - Building Reliable Networks with the Border Gateway Protocol. O'Reilly, Sebastopol (2002)
15. Kapela, A., Pilisov, A.: Stealing the Internet. DefCon August 16 (2008) (last visit, July 17, 2009)
16. Lad, M., Massey, D., Pei, D., Wu, Y., Zhang, B., Zhang, L.: Phas: A prefix hijack alert system (2006)
17. Li, T., Fernando, R., Abley, J.: The AS_PATHLIMIT Path Attribute (2001), <http://tools.ietf.org/html/draft-ietf-idr-as-pathlimit-03> (last visit: January 17, 2010)
18. Marques, P., Dupont, F.: Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing. RFC 2545 (Proposed Standard) (March 1999)
19. Muthukrishnan, K., Malis, A.: A Core MPLS IP VPN Architecture. RFC 2917 (Informational) (September 2000)
20. Networks, J.: Examine BGP Routes and Route Selection in Juniper routers (last visit December 12, 2009)
21. Rekhter, Y., Li, T., Hares, S.: A Border Gateway Protocol 4 (BGP-4). RFC 4271 (Draft Standard) (January 2006)
22. Suri, S., Sandholm, T., Warkhede, P.: Compressing two-dimensional routing tables. Algorithmica 25, 287–300 (2003)