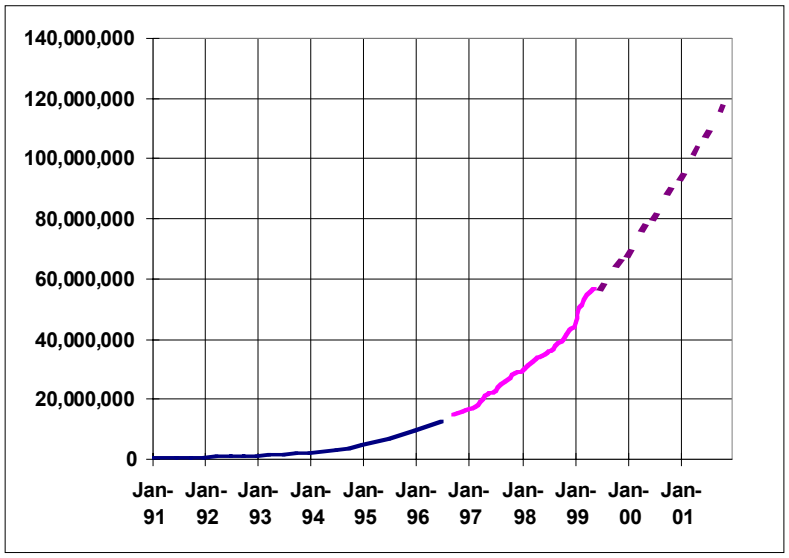
Growing Pains: From CIDR to IPv6

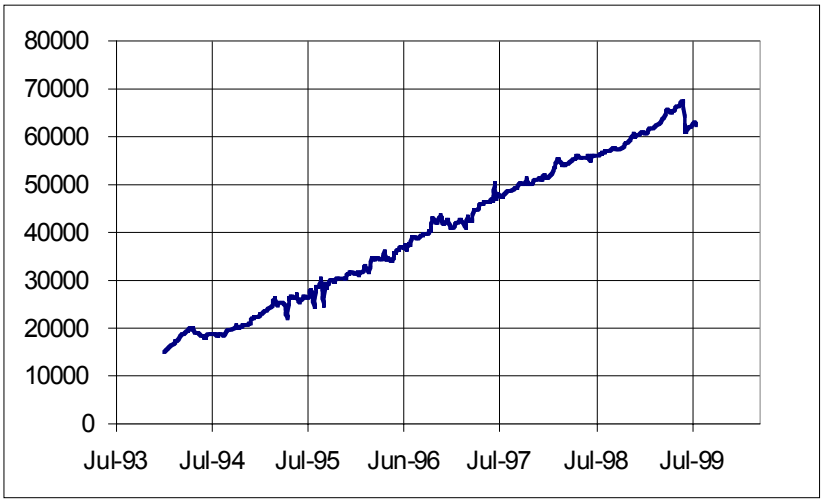
The Internet is growing, and it is growing fast—the number of connected hosts is doubling almost every year, while the volume of traffic is doubling every 6 to 10 months. This growth has been sustained for several years, and all measures indicate that it may well continue at the same rate for another five years, at least. Internet providers must invest continuously to build up network capacity, but they also have to cope with a side effect of the growth, the strains it places on the routing fabric.

10.1 Addresses, Networks, and Routing Tables

Nobody knows exactly how large the Internet is, how many Internet users there are worldwide. The very high growth rate that it has enjoyed makes all measurement attempts approximate at best, since the numbers may well change during the measure. In my opinion, one of the most significant measures is the number of hosts registered in the domain name system. This number is not directly related to the number of users: for example, there are 18 million AOL users, but AOL declares fewer than 2 million addresses. However, the size of the address space is a good indication of the size of the routing problem, because the routing system must guarantee that all the registered addresses will be reachable. (The numbers between 1991 and July 1996 come from the DNS studies conducted by Network Wizards; those between September 1996 and June 1999 from the Netsizer service, www.netsizer.com; and the numbers between 1999 and 2001 are interpolations.)



Thanks to CIDR and to network aggregation, the number of routes that have to be propagated by BGP does not grow quite as fast as the number of hosts. As shown in the following figure, the number of routes grows almost linearly over time, while the number of hosts grows exponentially. But the growth is still very fast. There were almost 60,000 routes declared in the BGP routing tables in June 1999. (The chart is drawn from numbers collected by Erik-Jan Bos and Geoff Huston and published at http://www.telstra.net/ops/bgptable.html.)



The growth in routes and networks has some very direct consequences on the load of the system. Obviously, routers need to manage ever larger tables. But the memory effect is probably not quite as important as the traffic effect: more networks mean more routing traffic.

10.1.1 Routing Instabilities

In a communication to the 1997 ACM/SIGCOMM conference[1], Craig Labovitz et al. reported their finding after the analysis of nine months of BGP traffic. In short, they found many more updates than would have been required for a normal operation of the network. The Internet, at the time, included about 1300 autonomous systems, 1500 paths, and 45,000 prefixes. Most paths remain available for relatively long periods, as shown in Paxson's study [2]; thus, a normal figure would have been somewhere between 10,000 and 100,000 updates per day. Instead, Labovitz et al. observed 3 to 6 million updates per day. This had a dramatic impact on Internet service, as routers spent precious computing resources handling these unnecessary updates, not to mention the traffic overhead. Moreover, in some routers, updating the routing tables required flushing a "route cache" in the route processors, which resulted in temporary slowdown. Too frequent updates could, in fact, result in loss of service.

Most of these updates were pathological, redundant messages that did not reflect changes in the network topology. A table presented by Labovitz et al. [1] shows a partial list of updates on February 1, 1997. The actual provider names have been replaced by anonymous letters, and the updates are categorized as either route announcements (i.e., actual path updates) or route withdrawals. The "unique" column lists the number of updates that would have been necessary if withdrawals and announcements were not duplicated in unnecessary, redundant messages.

Network	Announce	Withdrawn	Unique
A	1,127	23,276	4,344
B	0	36,776	8,424
C	32	10	12
D	63	171	28
E	1,350	1,351	8
F	11	86,417	12,435
G	2	61,780	10,659
H	21,197	77,931	14,030
I	259	2,479,023	14,112
J	2,335	1,363	853

The very large number of duplicate withdrawals was clearly due to bugs. In fact, the majority of the announcements resulted from hardware and software bugs in several popular routers. Faulty hardware could, for example, detect a connectivity loss when in fact the line was only temporarily unavailable for a few milliseconds. Faulty software would fail to keep proper track of previously sent withdrawals and would repeat them "just in case." These problems have since been corrected, but other, more systematic problems remain. In a consecutive study, Craig Labovitz and his partners studied the behavior of three "backbone" ISPs between November 1997 and November 1998 [3]. They were able to look both at the OSPF and BGP level.

The OSPF reports showed that outages were due to the expected mix of scheduled maintenance, power outages, fiber and circuit cuts, and hardware and software problems. The majority of outages were detected in the customer networks that the ISP served, rather than in the backbone itself: backbone routers, for example, would typically be located in operation centers

equipped with back-up power, and would not be subject to power failures. Another characteristic was the random distribution of the outages, which did not exhibit any specific frequencies.

On the other hand, BGP outages, which are characterized by the unavailability of interdomain paths, showed a very clear periodic behavior, including a component with a period of 7 days, and another with a period of 24 hours. In fact, the frequency of updates was, by and large, correlated with the volume of traffic, and would increase during the typical busy hours. This behavior contradicts the theoretical notion that updates would be triggered by the loss of links, by hardware or software failures in the routers. If that was true, the behavior of BGP would have mimicked that of OSPF!

The most likely culprit for the outages observed in November 1998 is network congestion. As the web traffic continues to grow, network links are very often congested. This means that the BGP packets, which are carried by TCP connections, may easily be delayed when the TCP connection "backs off" to ease congestion at a bottleneck link. This back-off may result in the failure to transmit keep-alive messages at the specified interval, which leads to the categorization of a border router as "down" when it is merely experiencing congestion. This bad diagnostic can have a snowball effect. First, the routes that were transiting through the "down" router will be replaced or removed, creating a spike of BGP traffic. Soon after that, the "down" router will try to reestablish the BGP connection, and will send a complete update, listing all available paths. This exchange of traffic will contribute to the congestion, and also to the overload of downstream routers that receive the sequences of updates. The downstream routers may in turn become congested and fail to send keep-alive messages at the specified interval, amplifying the initial problem.

The congestion is more likely to affect internal BGP connections, because these connections transit through several interior routers, while external BGP connections are typically through a single hop. A possible cure deployed by some routers is to increase the priority or "precedence" of the IP packets that carry BGP traffic, ensuring that they will be transmitted normally even during congestion. That cure will ease some of the problems reported by Labovitz et al. [3], but would still leave an open vulnerability. If routers start to receive too many updates, if they cannot get enough processing resources, they may take too long to send the keep-alive messages and trigger the same kind of cumulative failures.

10.1.2 Controlling the Rate of the Advertisements

Interdomain routing is, in effect, a distributed application performed by the routers of tens of thousands of Internet Service Providers. Just as the OSPF flooding process distributes link state information to all the routers in an area, the BGP update process distributes reachability information to all border routers. Indeed, the BGP update is not a full flooding process. Updates that only affect a secondary path, one which has not been retained by an intermediate domain, do not get propagated. But, on the other hand, a multihomed AS that can choose between several paths toward a given destination will receive as many copies of each update that the destination triggers. Local updates are propagated at least once to all border routers in the Internet.

If each of the about 60,000 entries in the routing tables would only be updated once a week, each router receives about 6 updates per minute, a reasonable proposition. But the experi-

ence shows that a small fraction of the routes contribute an inordinate amount of updates. This phenomenon, informally known as "route flap," can be caused by software or hardware bugs, by the interaction between BGP and network congestion described in the previous section, or by local decisions.

Whatever the cause of a route flap, it is necessary to mitigate its effects. If a misbehaving router sends too many updates at too short intervals, its neighbors that try to process all the updates will exhaust their computing resource and may fall into a congested state that triggers further instabilities. A solution, proposed in RFC-2439 [4], is to limit the rate at which updates are accepted for any given path. Limiting the rate obviously diminishes the responsiveness of the protocol and should be done with caution. In practice, one should ensure:

- that if a path remained stable for a long time and just changed once, the update will be propagated quickly;
- that if a path oscillates rapidly, updates will be suppressed and the path will be declared unreachable;
- that if an alternate path exists, the oscillating path will be damped more aggressively than if it was the sole path to the destination; and
- that damping does not affect the internal BGP connections, as it is important that all routers within an AS keep a coherent view and advertise the same path for any given destination.

BGP-4, as specified in RFC-1771 [5], contains a partial provision against route flaps. Border routers are expected to wait some minimal delay between consecutive announcements and another, shorter delay between successive route originations. The standard interval values are set to 30 seconds for the advertisement interval, to 15 seconds for the origination interval. These intervals achieve one important result: routes that just changed are more likely than others to change again at a brief interval. Waiting for 15 or 30 seconds allows routers to "pack" consecutive updates, thus diminishing the global load. But such delays are far too short to solve the damping problem. Let's suppose that 20% of the 60,000 interdomain paths advertised in June 1999 exhibit some instability. A 30-second interval still means more than 400 advertisements per second, and many more for multihomed networks. This is already too much for some routers!

The basic idea of RFC-2439 is to associate to each path a "figure of merit." The figure of merit is computed as an exponentially smoothed average of the number of updates received in the recent path. Specifically, the following formula is used:

- On reception of an update, increase the figure of merit by one.
- At regular intervals, multiply the figure of merit by a coefficient K, $0 < K < 1$. The actual value of the coefficient will be lower if the path is available and is the only path to the destination than if it is not available, or if alternate paths exist.

Paths that remained stable for a long time will have a figure of merit near zero, while paths that exhibit high instability will have a higher figure. Routers would then refrain from advertising, or using paths whose figures of merit exceed a threshold value—a value of 2 is suggested in RFC-2439 [4].

10.1.3 Controlling the Source of the Advertisements

We already observed the similarity between OSPF flooding and BGP updates. As in OSPF, information that originates from any border router gets propagated to the whole set of routers in the Internet, which will then base their route computations on this information. Needless to say, inaccurate information results in inaccurate routing. A classic example of failure occurs when a leaf network misconfigures its addresses. The perennial "network legend" used to be that some brainy but careless MIT alumni would use the MIT class-A address in their network, then connect to their ISP. The ISP would pick up the address, advertise it to the whole world through BGP, and subsequently receive traffic for the MIT class-A address that will, essentially, go down a black hole.

Such examples may well be apocryphal, but a real-life incident did occur on April 25, 1997. A small Virginia ISP misconfigured one of its border routers. That router would essentially receive all the path announcements from its neighbor, remove the AS_PATH information, and replace it with a single element, as if the destinations were directly accessible through that ISP. It then reannounced the modified routes to its neighbors. As a consequence, many networks sent their entire Internet traffic to the Virginia ISP, effectively disconnecting the Internet for up to two hours.

The only ISPs that did not suffer from the incident were the ones that exercised caution and refused to believe or accept the modified path. In fact, similar incidents had been happening before, albeit at a lower scale, and work was already underway to provide a method for systematically controlling the veracity, or at least the plausibility, of route announcement. To be precise, three methods have been proposed. One possible way to check whether routes are acceptable is to compare them to a registration database, into which each ISP would document the network that it serves and the autonomous systems with which it peers. The Internet Route Registry (IRR) was set up to that effect. It continues and incorporates an earlier effort, the Route Arbiter Database (RADB) that was set up by Merit for the NSF. The IRR, is in fact, a collection of databases maintained by regional networks and organizations such as Merit for the original RADB, Réseaux IP Européens (RIPE) for the European networks, Cable & Wireless, ANS, Level3, and Bell Canada. A complete listing of Internet Registries is available on the web [6]. All entries follow a common format, specified in RIPE report 181 [7]. Entries in the database describe networks, the routes that link networks and autonomous systems, autonomous systems and their readiness to accept routes from other systems. Network operators that want to use the IRR services will get copies of the databases, in order to collate the various entries. They can then check the plausibility of announcements by checking the correlation between the AS_PATH information and the advertised routes. The last domain in the AS_PATH, or at least one domain in the AS_SET component of the AS_PATH, should match one of the ASs declared in a route entry for the reachable prefix. One could then, in theory, verify that the path itself is plausible by checking that each component of the AS_SEQUENCE has declared its readiness to accept routes from the other components.

There are, however, two practical problems with the IRR approach. One is the declarative nature of the databases. Domain managers have to feed their information into the regional bases. Some will do it, but many won't, or won't do it in time. This leaves holes in the data set, routes for which no declaration is registered. Paths that lead to nonregistered routes cannot be verified.

This also causes some of the information to be obsolete. If obsolete information contradicts a valid path, that path may be rejected, leading to an unnecessary lack of connectivity.

A possible way to alleviate the load of the registration procedures would be to use the domain name system instead of a special purpose database. The domain name system is, in fact, a database distributed over thousands of Internet servers. It is already used to provide information about addresses, and it would be possible to provide information about prefixes as well. In a proposition submitted to the IETF, Tony Bates, Randy Bush, Tony Li, and Yakov Rekhter propose to create a special DNS tree to hold domain information. The information on a prefix such as 10.1.128/20 may be encoded under a DNS name such as 128/20.1.10.bgp.in-addr.arpa. The routers that receive a path could use the DNS to retrieve the AS records associated to the advertised prefixes, and to check that the path only contains prefixes that can legitimately be announced by the neighboring AS. Using the DNS has two advantages:

- Because the DNS is a distributed system, the information could be managed directly by the owner of the prefix.
- Because the DNS information can be secured through electronic signatures [8], the router can check the source and the veracity of the information.

But a similar or better level of security can be achieved without having to query the DNS each time a border router receives a path announcement. In a contribution to the IETF, Sandra Murphy analyzes the security of BGP [9] and proposes a set of possible improvements:

- The originating AS could certify the validity of the announcement by attaching a digital signature attribute to the path information. The digital signature would attest to the identity of the AS that inserted the network information. Certificates could be also attached to attest that the AS is authorized to announce this network.
- If aggregation occurs, the aggregator could insert a digital signature attesting to the origin of the aggregation, and possibly a certificate attesting to its right to aggregate.
- A digital signature could be inserted by each relay, attesting to the veracity of the path.

These concepts form the basis of a "secure BGP" that has still, however, to be developed. The potential problem here is the overhead of signatures, in terms of both transmission volume and computation overhead. To be reasonably secure, key lengths of a thousand bits should be used. In 1999, exponentiation of such numbers still requires a significant amount of time, maybe one tenth of a second. This would seriously limit the rate of update transmissions. However, the potential benefit is very large, and the performance of computers increases every year. We may see such solutions become routine in a few years.

10.2 The Structure of Interconnections

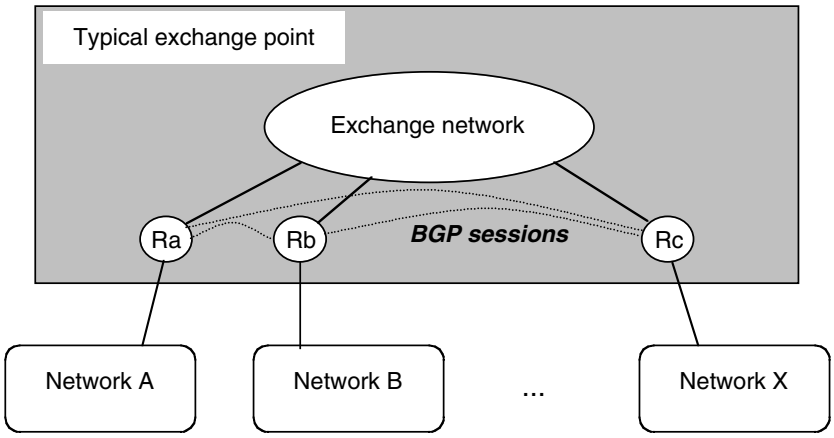
The Internet may appear to be thriving, but the system is experiencing tensions. The initial interconnection structure, based on public exchange points, is being questioned by large operators for a mix of business and technical reasons. The business reasons have, in fact, pushed a small number of providers to develop very large networks that cover whole continents, and that may well

evolve to cover the whole world. Managing these networks and their interconnections poses specific problems.

10.2.1 From Public Exchanges to Private Peering

The original NSFnet plans called for a set of five Network Access Points (NAP) through which regional networks would connect to the NSFnet and to alternative commercial networks. The exchanges would complement an existing set of interconnection points such as the Commercial Internet Exchange (CIX) and the two Federal Internet Exchanges (FIX), FIX-East and FIX-West. Similar exchanges would be developed in the other countries, allowing for local connections of Internet providers.

An exchange typically is a computer center in which each participating network provider brings a router. These routers are connected to a "neutral" interconnection system provided by the exchange. The interconnection normally is not provided by a router, because of policy considerations. An exchange is a place where multiple providers come and interconnect, but these interconnections still result in most cases from bilateral contracts between every two pairs of connected networks. Each of these "peering agreements" is materialized by at least one BGP session between the routers belonging to the network. If the exchange network was materialized as a single router, this router would route all packets bound to a given destination to a single path, regardless of peering contracts. In fact, an exchange built around a high-speed router would act as a transit provider. The BGP sessions would be established between each network and the exchange's router.



In the most common setup, the exchange network is provided using "layer 2" switching technology. That technology must provide enough bandwidth to accommodate the aggregate traffic of the participating network interconnections. Various technologies have been used over time:

- The first exchanges were build around a 100-Mbps FDDI ring. That technology was adequate when the network connections ran at T1 or E1 speed—i.e., 1.5 or 2 Mbps. It is inadequate now.

- ☞ Several exchanges were built using an ATM switch or a network of ATM switches. That technology is capable of providing OC3 or maybe OC12 connections, and proved hard to develop initially, as ATM switches did not have adequately sized buffers, and were lacking adequate queue management techniques. These problems have now been fixed.
- ☞ Other exchanges have been built using switched Ethernet technology, at speeds such as 10 Mbps, 100 Mbps, or 1 Gbps.

In the future, we may see exchanges using even more advanced technology, such as optical fibers and Wave Division Multiplexing. Exchanges played an important role in the development of the Internet, allowing hundreds of networks to connect in a single place. Several exchanges developed additional services, such as access to route information registries, name servers, and cache servers. However, the whole notion of exchanges is being questioned by the largest Internet providers, who attack it for a mix of business and technology reasons and who prefer to use private interconnections.

The business reason is rather easy to explain. When they join an exchange, network providers sign a contract that, in effect, forces them to exchange traffic with all other exchange members. This is very advantageous for small providers, who, by getting access to an exchange point, get access, in effect, to the whole Internet, without having to pay an interconnection fee to larger providers. But it is equally disadvantageous for the large providers, who feel their networks provide the bulk of the Internet's transit capacity, and who would normally charge for their services. However, this problem could be solved by slightly modifying the exchange connection rules. Just because a connection takes place in an exchange point does not mean that it has to be free. The technical problems, on the other hand, cannot be dealt with by simply rewriting a contract. Exchanges, by their very nature, become focal points for congestion. The congestion can happen at the routing level, at the access level, and eventually in the exchange itself.

The routing problem may become very hard if a single router has to manage the peering with hundreds of partners. The memory requirement increases, because BGP requires that routers maintain a complete copy of the announcement received from each peer. The rate of announcements also increases with the number of peers, possibly leading to saturation of the local processing resource. As fledging networks connect to the exchange, some inexperienced operators will fatally make mistakes, which can easily generate abusive messaging and processing loads.

The access link between the exchange and each network is shared by sources in all peering networks. If one of the source networks sends an inordinate amount of traffic, the access link will become congested. This congestion will then be shared by all the peering networks. It is theoretically possible to use techniques such as weighted fair queuing to diminish the effect of this problem; we will present these techniques in chapter 14. But these techniques are hard to deploy at high speed. Moreover, the network access router has only indirect control of the amount of traffic that it will accept from the peering partners. BGP allows routers to limit the range of destinations that they announce to partners, following the policy specified in the peering agreements. However, a misconfigured peer can send traffic for destinations that were not announced. Routers are typically not configured to check the source of the packets. If a destination was announced to peer A but not to peer B, that destination is in effect usable by peer B as well. The potential for errors, and possibly also for abuse, is quite high.

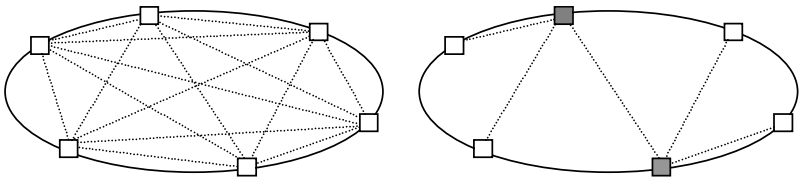
We explained above that exchanges typically are built around the fastest local network technology that is available when they are created. However, as the traffic increases at a very high rate, the exchange network needs to be continually upgraded, and access routers need to be often replaced by more powerful models. This is, by nature, a discontinuous process. The consequences, if the exchange networks become congested, are very scary. Thousands of networks will be congested, and a large fraction of the Internet will experience poor service.

Private interconnections, in practice, do not suffer from the same problems. Because they are used only by a pair of networks, there will be only one BGP connection to maintain, typically with another well-operated, large network. The capacity can be provisioned to match the planned traffic between the two networks. Moreover, if the traffic increases, it is possible to multiply the number of private connections in order to split their loads, while multiplying the number of exchange points would be a complicated process.

10.2.2 Managing a Large AS

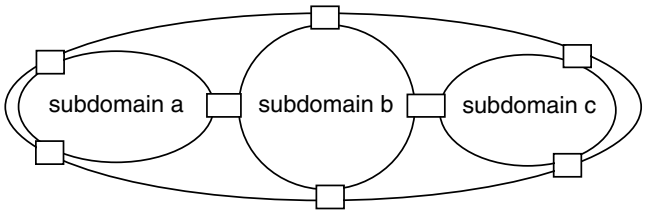
Large networks that cover a whole continent need to manage a large number of interconnections, and may thus include a very large number of border routers. A value of more than 100 would not be surprising. Maintaining such a large number of routers increases the burden of internal BGP connections. The first problem is that BGP, in theory, requires a full mesh of interconnections, and the second problem is that the frequency of updates will increase with the number of connections.

Maintaining a complete mesh of connections means that each time a new border router is added to the network, all other border routers will have to be configured to establish a BGP session with that border router. The memory and process requirements of each border router will be proportional to the number of border routers in the whole network, which will clearly pose a scaling problem. RFC-1966 [11] proposes to reduce this burden by introducing route reflectors, routers that act as relays for routing updates. Routers are grouped in clusters that only communicate with the other internal BGP routers through a small number of reflectors. The reflectors communicate with the other internal BGP routers, or with other reflectors. In order to avoid loops that could be caused by misconfiguration, RFC-1966 introduces two new BGP attributes, an ORIGINATOR_ID (code 9), identifying the routers that created the path, and a CLUSTER_LIST that identifies all clusters that the internal path traversed. Clusters are identified by the 32-bit identifier of their main reflector. Paths shall never be advertised back to their originator within a cluster, and clusters should not accept paths from other clusters if their identifier is already present in the router list.



I-BGP connections, with and without reflectors

RFC-1965 [10] proposes another way to organize the routing within a large domain. It introduces the concept of "confederation" that was first developed in the Inter-Domain Routing Protocol (IDRP) [12], a protocol developed to play the same role of BGP in the ISO/CLNP framework. For some time, many interdomain routing experts believed that IDRP could one day replace BGP. This did not happen, but many of the IDPR concepts, such as confederation, are being retrofitted into BGP.



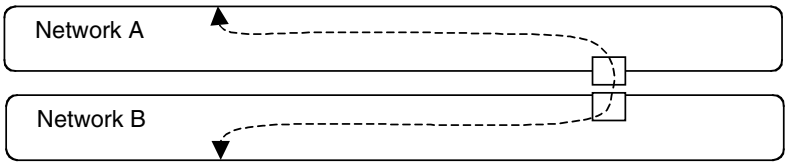
AS as confederation of subdomains

The basic idea of confederation is to consider a large domain as a collection of smaller domains. This naturally brings the advantages of clustering, since the full mesh of IGP only has to be maintained within a subdomain, or between the internal border routers of the subdomains. Another advantage is that it is possible to perform address aggregation within a subdomain. Given proper address allocation, local customers of a subdomain can be represented by a single address prefix in other subdomains, thus diminishing the size of the internal routing tables. Again, care must be taken to avoid the possible effects of misconfiguration. RFC-1965 proposes to modify the structure of the AS_PATH attribute, adding an AS_CONFED_SET and an AS_CONFED_SEQUENCE component to represent the traversal of domains, in addition to the AS_SET and AS_SEQUENCE that represent the traversal of ASs.

In order to deal with alternate paths, large domains also require consistent parameterization of the LOCAL_PREFERENCE attribute. For example, if a customer network is multihomed, it can be reached through the direct path, i.e., through its normal connection to the local ISP, through a back-up connection, through another ISP, or perhaps through another "customer-supplied" path. This requires manual configuration of the local preference, a task that providers would like to automate. In order to solve this problem, RFC-1997 proposed to add a COMMUNITY attribute to BGP [13, 14].

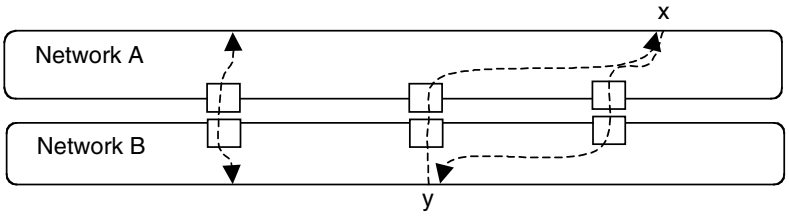
10.2.3 Interconnection Between Large ASs

Suppose that networks A and B cover the same geographical area. Connecting these two networks through a single peering point would not be very efficient, as all the traffic would have to be routed back and forth through that point. Suppose, for example, the two networks both serve North America and peer in Washington, DC. Traffic between two users located in the same city, say Los Angeles, will have to be hauled back and forth through Washington, DC!



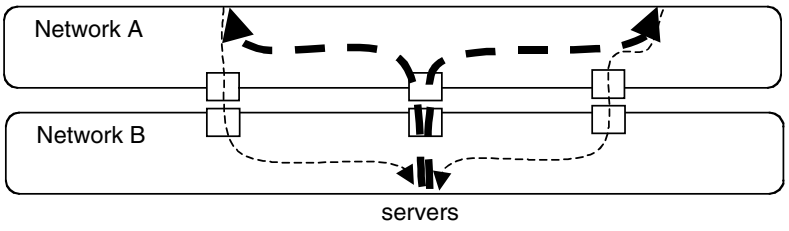
A single point of interconnection is not very efficient

A simple improvement is to establish multiple peering points between the two networks. However, in a classic BGP configuration, all the routers that belong to the same AS will advertise the same routes. The internal configuration of each network is hidden through address aggregation. As a consequence, the interior routing protocols in network A can only consider that all gateways to network B are at an equal distance to all destinations announced by B, and vice versa. Suppose that the networks A and B, instead of peering only in Washington, DC, now also peer in Los Angeles and Chicago. The problem we mentioned in the previous paragraph is solved: a connection between two users located in Los Angeles will only use the Los Angeles gateway. But consider a TCP connection between *x*, connected to network A in Washington, DC, and *y*, connected to network B in Chicago. The packets sent from *x* to *y* will be routed by the IGP of network A toward the nearest point of access to B, Washington, DC, and will move from there to *y*. In the reverse direction, the packets sent by *y* toward *x* will be routed through the Chicago access point to A, and from there to *x*. This asymmetric routing is often called "hot potato routing."



Advertising the same information leads to hot potato routing

Whether hot potato routing is good or bad depends on the nature of the interconnected networks. In the case of two large transit networks with balanced interconnections, the consequences probably are not very important, as asymmetries in one direction will be compensated by asymmetries in another direction. The consequences may, however, be quite drastic if one of the networks specializes in "web hosting" services, providing high-speed connection to a set of web servers, often concentrated in a single computer center.



Asymmetric routing and server farms

Web servers are characterized by a highly asymmetric traffic pattern. They receive short commands and respond with large quantities of text, images, and video. Asymmetric routing guarantees that, even if the network provider took pains to provide high-speed connections and a large number of interconnection points, the outgoing traffic will be concentrated on a single gateway. The solution to this problem requires cooperation from network A, which will have to program its border routers to announce multiple paths, with properly configured MULTI_EXIT_DISC parameters.

10.3 Routing Table Aggregation and Address Allocation

The growth of the Internet routing table can only be achieved if addresses are aggregated—that is, if multiple long prefixes can be represented by a single short prefix. Address aggregation also concurs to a reduction in the number of updates propagated through the Internet, as the path leading to the aggregate will tend to be more stable than the path to individual components. But routing table aggregation can be achieved only if the addresses are assigned in a coordinated fashion. In this section, we will present this coordination and the other problems that need to be solved, such as the relationship with the providers and the need, at some stage, to engage in address renumbering.

10.3.1 Coordinated Address Allocation

The aggregation of addresses by common prefixes is facilitated by the regional scope of the current address assignment authorities. One of the first actions of the coordinators has been to lay out an addressing plan by continents, allocating to these continents a range of class-C addresses [17]:

Multiregional	192.0.0.0–193.255.255.255
Europe	194.0.0.0–195.255.255.255
Others	196.0.0.0–197.255.255.255
North America	198.0.0.0–199.255.255.255
Central/South America	200.0.0.0–201.255.255.255
Pacific Rim	202.0.0.0–203.255.255.255
Others	204.0.0.0–205.255.255.255
Others	206.0.0.0–207.255.255.255

The continent-level authorities then delegate a fraction of their share to lower-level authorities. For example, the European authority is the "network coordination center" of RIPE; it delegates slices of its spaces to national authorities, such as FRNIC in France. Other regional authorities are the American Registries for Internet Numbers (ARIN) and the Asia Pacific Network Information Center (APNIC).

If the plan is followed, we can hope that all the European networks will be represented by exactly one entry in the routing tables of other continents—that is, a 7-bit prefix corresponding to all IP addresses included between 194.0.0.0 and 196.0.0.0. But there are still some debates

going on about, among other things, the "provider" or "geographic" nature of the addresses. In fact, as the Internet grows, the registries have started to allocate addresses in new ranges. The typical process is for IANA to delegate an address range to a regional registry, which then subdelegates it through either national registries or network providers.

10.3.2 Provider and Clients

Until 1992, the Internet network numbers had no relation at all to the network's topology. This is both an advantage, as it provides flexibility, and an inconvenience, as it is the main cause for the explosion of the routing table. A coordinated assignment strategy will obviously have to remove some of the flexibility in order to deflate the tables, but the real question is, how much? There are two responses to this question, called "provider addressing" and "geographical addressing."

The Internet's topology bears little relation to geography. It is expressed in terms of links, routers, and connections. All of these are laid down by Internet providers. Two organizations that are neighbors in a given city may be either very near or very far apart in terms of Internet distance. They will be very near if they subscribe to the same provider, quite far apart if they don't. In the latter case, the packets will have to travel to the point where the two providers interconnect, which may be in the same city, in the same region, or maybe only near some national "Internet interconnection." There were examples, in June 1994 in France, of providers interconnected only through a European gateway. If a client of A sent a packet to a client of B, the packet would go from Paris to Amsterdam and back!

The only strategy that guarantees all the benefits of "routing table aggregation" is thus called "provider allocation" [4]. Each Internet provider receives a slice of the address space and "sells" the network numbers to its clients. As a result, all of the clients' addresses share the same prefix and can be aggregated as one single entry in the other providers' routing tables.

But suppose now that an organization wishes to switch from provider A to provider B. If it keeps the address that was assigned by A, it obliges provider B to announce this "exception" to all other providers. Failure to do so would result in all the traffic being directed to the old provider, A. All of a sudden, we have added an individual entry for the organization in all these routing tables, and the benefits of aggregation have been lost. We may, perhaps, tolerate this disorder during an interim period, but the organization will eventually have to change its Internet address for a new one allocated by B.

We will see in the next paragraph that changing an Internet address is not always very easy. If changing providers means that we will have to renumber, we may think twice and compare the benefits of a slightly lower tariff or a slightly better service with the administrative cost of address renumbering. This may have a freezing effect on competition, something many of us are quite reluctant to accept.

The "provider addressing" strategy has a logical consequence—addresses are really owned by the provider. This, in a way, locks the client in the provider's arms and is a big departure from the current situation, where addresses are owned by the clients. Today, because I own my address, I need only to sign a contract to change providers. I may have several simultaneous subscriptions to different providers and select the best provider for any particular destination; I may even

change the routing as a function of the time of the day if the providers have a service that varies with time. I cannot do the same with provider addressing.

Geographical addressing is an alternative to provider addressing that leaves the ownership of the addresses to the clients, yet still allows some aggregation. Address slices are allocated to regions or cities; the clients receive addresses within these regions. The providers will have to maintain detailed tables for the cities they serve, but we will probably have a good aggregation at the higher level of the hierarchy—by region, by country, or by continent. Many experts assert, however, that the geographical addressing would not provide sufficient aggregation; the "detailed tables" could be so large that the providers would not be capable of handling them.

10.3.3 Will We Need to Renumber?

The provider addressing strategy just described results in address aggregation only if the users that move from one provider to another do change their addresses. With the current technology, changing a machine's address rates from relatively easy to almost undoable [19].

Changing your address means that you will replace the old value in all the places where this value is stored. In theory, if one uses state-of-the-art technology, there are only two such places: the master file of the BOOTP or DHCP server and the database of the DNS server. If one updates the BOOTP or DHCP file, the address will be updated the next time the host is initialized; if one updates the DNS, all our partners will get the new address the next time they query the DNS.

Things, however, are not so easy. Consider first the host "reinitialization." It may be standard practice in some places to power down your personal computer when you leave the office, but that is not always the case. My workstation, for example, has been up for the entire last month. There are many programs that can take advantage of the night, such as automatic backup of the hard disk. It is also a good time to receive mail from your friends from other continents. And many people do work at night, through modem connections. All these hosts that stayed up will have to be rebooted explicitly. Then it is often the case that these hosts, which are rarely rebooted, do not use BOOTP or DHCP. We will have to configure their addresses manually.

On most systems, it is almost impossible to change the IP address without rebooting. This is very annoying, as it implies that the servers running on these systems will have to be stopped. In some cases, doing a clean stop may be a very lengthy operation, as one has to stop the system according to very precise procedures in order to guarantee a clean restart. Even if one could change the IP address without rebooting, one will not be able to keep the TCP connections; the TCP context is identified by a combination of TCP ports and IP addresses.

Until all hosts have migrated to the new address, we will have a mix of old and new addresses on the local subnet. This is annoying, though not fatal. The routers will have to be programmed to announce these two networks. The most annoying side effect is that two hosts that do not belong to the same IP subnet may need to send their packets to a router to communicate. Hopefully, this situation will not last very long.

Once we have accomplished this transition and updated the DNS server, we might think we are finished, but this is only "almost true." For one thing, the DNS uses "replication" and "caching" to enhance reliability and response times. This means that there are copies of the old

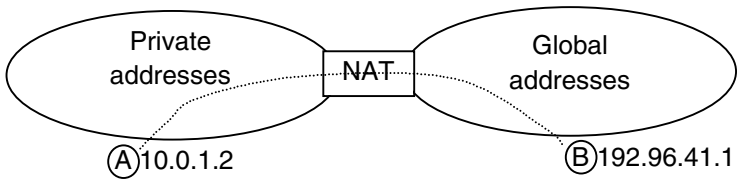
addresses in the "secondary servers" and in an unpredictable number of caches all over the Internet. We can probably trigger an update of the secondary servers; in any case, they will get the new values during their next scheduled update. But we will have to wait for the caches to time-out, which can take a few days, if we have not taken the precaution to initialize the DNS records with short TTL values.

Then there is the problem of all those applications that use Internet addresses directly, that think it is smart to write down your address in a file somewhere so that they don't need to use the name server. This is a thoroughly bad idea, but software engineers are people, and people are human. They have bad ideas, and it is quite difficult to drag them away from using them. The only thing one can hope is that they will learn. Maybe by reading this book.

In summary, renumbering is difficult today, network managers with hands-on experience will tell you. This does not bode well for a strategy that would rely on frequent renumbering for routing table aggregation, but time will tell. It may well be the case that someone will come out with a very clean implementation that combines the DNS and the DHCP server, which entirely parameterizes the old and new addresses, which allows us to keep both addresses for a transition period so that we don't have to reboot or break the TCP connections. They may even allow us to keep two addresses forever, so that one can buy Internet services from multiple providers. Who knows? Until now, the Internet has survived because smart people came out with smart solutions. As the size of the network increases, the supply of smart people is certainly not diminishing!

10.3.4 The NAT Alternative

The Network Address Translation technology was developed for networks that don't want to be fully connected to the Internet, and that accept having all their outgoing traffic "rewritten" by an address translation gateway. When the NAT technology is used, the "partially connected" network uses "private addresses"—that is, addresses that are not routable in the Internet. A set of such addresses has been set aside in RFC-1597 [20] and RFC-1918 [22]. One of the network addresses that was so set aside was "network 10," the class-A network that was initially allocated to the Arpanet and that was not used anymore after its decommissioning.



The reserved addresses cannot be used in the global Internet, which means that host B of our example will never be able to send a packet toward the private address 10.0.1.2 of host A. To communicate, these two hosts must rely on the address translation service of the NAT gateway. The communication will typically proceed as follows:

1. A decides to contact B and sends a packet from source address 10.0.1.2 (private) toward the destination address 192.96.41.1.
2. The NAT intercepts the packet and translates the address 10.0.1.2 into a global address, such as 192.1.2.3, taken out of a pool managed by the NAT. The packet is then relayed toward B.
3. B receives the packet and replies by sending a packet from 192.96.41.1 toward 192.1.2.3.
4. The NAT intercepts the packet, realizes that the address 192.1.2.3 has been reserved as a translation of 10.0.1.2, and performs the reverse translation.
5. A receives the packet. Successive packets will undergo the same translation, as long as the NAT remembers that 10.0.1.2 maps to 192.1.2.3.

The translation, however, is not quite so simple as just managing a cache of addresses and rewriting headers. First, one must deal with the TCP and UDP checksums, which protect not only the data but also the source and destination addresses. The checksum must be updated to reflect the address translation. Then, the NAT must deal with applications that directly manage addresses. FTP, for example, routinely sends over its control connection the addresses and ports that will be used for its data connections. This means that the NAT has to recognize the applications that are being used, and that it shall in some cases translate parts of the packet content in addition to the packet header.

On the surface, NATs may look attractive. The network can use a large address space, and will not normally need to renumber, even if it changes providers. But these arguments are very shortsighted, as explained in RFC-1627. At the time of e-commerce, corporations often have to establish business relations with their customers and providers, which is not facilitated by the use of private addresses. If two organizations are using the same private addresses, any direct interconnection is going to be awkward, at best. Then, if two corporations that used the same private addresses merge, the network administrators will be confronted with a serious nightmare.

The worst effect of NAT is on network security. In fact, NAT vendors often say that the use of private address space provides some level of security, because machines are not readily accessible to the outside. This is, at best, a half truth. As soon as a privately addressed host has contacted an outside host, the NAT establishes an address mapping that can then be used to send packets inside the private network. The security will at that point rely on the scanning of inappropriate packets by the NAT box, not on the fact that the addresses are private. This scanning of packets is exactly the same as performed by ordinary firewalls, and the NAT will not provide a better level of security than these ordinary firewalls. But firewalls also allow hosts to build up additional security by using end-to-end encryption, while NATs don't. NATs are incompatible with IP security, because IP security protects the packet header as well as the packet content—one of the services of IP security is to verify the origin of the packet, and NATs are essentially lying about this origin. Then, NATs have to examine the content of the packet to track addresses, and they cannot do so if the packets are encrypted.

I would never recommend relying on the NAT technology. Just do the paperwork and get unique addresses.

10.4 Is IPv6 the Solution?

The next generation Internet Protocol, IPv6, was specifically designed to ease the Internet's growing pains. The larger addresses would enable more hosts to connect to the Internet, and the addresses would be structured hierarchically in order to lower the size of the routing tables. Many ISPs have been reluctant to embrace IPv6, even if this did not slow down the evolution of BGP-4, which has become multiprotocol capable in order to support IPv6 and possibly other protocols.

10.4.1 A Larger Address Space Is a Two-edged Sword

The main advantage of IPv6, its larger addresses, is viewed with some suspicion by the practitioners of interdomain routing. They note that the growth of the routing table has only been controlled through harsh discipline imposed by the address registries, and that IPv6 has the potential of loosening that discipline. If addresses are plentiful, then it will be more difficult for the registries to invoke scarcity and to restrain users.

These fears are not necessarily well founded. The IETF has proceeded with great care to define a "provider-based" address structure for IPv6. The aggregatable structure allows for several levels of hierarchy, such as link, site, intermediate provider, and top-level provider. Only 13 bits have been allocated to the top level of the hierarchy, which should, in theory, limit the global routing tables to 8,192 entries, at least until someone decides to open up that field.

The real test, indeed, will only come from experience, to see whether the users can accept provider-based addressing. IPv6 contains much better support for this than IPv4. Automatic address configuration should make renumbering easy, especially if it is combined with automatic update of routing prefixes in routers. The support of multiple simultaneous addresses should release the need for multihoming, one of the main sources of entropy in the routing tables. In between, one can only prepare for the deployment of IPv6.

10.4.2 Extending BGP for IPv6

BGP-4, as defined in RFC-1771 [23], only supports the old IP protocol, IPv4. For this reason, the IETF initially thought to base IPv6 interdomain routing on IDRP [12], which was thought easier to extend to a multiprotocol environment. This initial project was however abandoned, mainly because there is much more experience available today with BGP than with IDRP. Transitioning to a new IP is hard enough without having to learn yet another routing protocol. The IETF eventually developed a "multiprotocol" version of BGP-4. This version, documented in RFC-2283 [24], will be able to support several protocols in addition to IPv4. It assumes that all BGP routers will be IPv4 capable, which will certainly be true at least during a transition phase. It also assumes that the decomposition in Autonomous Systems will be maintained irrespective of the protocol that is being used, i.e., that the AS number will not depend on whether the AS runs IPv4, IPv6, both, or even other protocols.

To develop multiprotocol capabilities, BGP-4 has to be extended. BGP-4 uses IPv4 addresses in three places:

- ☞ The network layer reachability information (NLRI) found in path announcements and route withdrawals is expressed as a set of IPv4 prefixes.
- ☞ The aggregator attribute contains an AS number and the IPv4 address of the aggregating router.
- ☞ The next-hop attribute contains the IPv4 address of the router that serves the path.

One may observe that in the aggregator attribute, the IPv4 address is used solely as an identifier. As long as BGP-4 routers are IPv4-capable and have unique IPv4 addresses, there is no need to change the syntax of that attribute. One may also observe that the next hop attribute is always used in conjunction with a list of reachable domains. In order to forward IPv6 packets, one needs to know the IPv6 address of the next hop, and the same will be true for any address family.

RFC-2283 extends BGP by defining two new path attributes: the Multiprotocol Reachable NLRI (MP_REACH_NLRI, code 14) and Multiprotocol Unreachable NLRI (MP_UNREACH_NLRI, code 15). The MP_REACH_NLRI attribute contains a list of one or many address family reachability data sets, each composed of an address family identifier followed by next-hop information and a list of prefixes. The MP_UNREACH_NLRI address family unreachability data sets, each composed of an address family identifier and a list of prefixes.

In each of these attributes, the address family is identified by a 16-bit address family, registered with the IANA, and by a one-octet "subsequent address family" qualifier. The following table lists the address family numbers that were defined by the IANA in July 1999. In addition to IPv4 and IPv6, the list includes several proprietary protocols that are often used in corporate networks, as well as several ITU-defined address plans.

Number	Address Family Description
0	Reserved
1	IP (IP version 4)
2	IP6 (IP version 6)
3	NSAP
4	HDLC (8-bit multidrop)
5	BBN 1822
6	802 (includes all 802 media plus Ethernet "canonical format")
7	E.163 (telephone numbers)
8	E.164 (SMDS, Frame Relay, ATM)
9	F.69 (Telex)
10	X.121 (X.25, Frame Relay)
11	IPX
12	Appletalk
13	Decnet IV
14	Banyan Vines
15	E.164 with NSAP format subaddress (UNI-3.1)
65535	Reserved

The subsequent address family code further qualifies the type of reachability information carried in the attribute, defining whether the addresses are used for unicast routing (code 1), for multicast forwarding (code 2), or for both (code 3).

Each prefix information is composed of a one-octet length field that provides the length of the prefix in bits, followed by the minimal number of octets necessary to encode the prefix.

The next-hop information is composed of the next-hop address, encoded as a one-octet length field followed by a variable length value, and of a variable number of "subnetwork point of attachment addresses" (SNPA). In pure Internet technology, one would have called these fields "subnetwork addresses." They are, for example, the Ethernet addresses of the next hop. The SNPA acronym is carried over from the OSI terminology used in IDRP. The presence of these subnetwork addresses is optional. In IPv6, these addresses can be obtained through the next-hop discovery protocol.

Using the BGP-4 multiprotocol extension to carry IPv6 extension is fairly straightforward. The only problem that was debated was whether the next-hop information should carry a global address or a local scope address. Global addresses can be used in all conditions but are subject to renumbering, while local addresses can only be used within one link (link local) or within a site (site local) but will not be affected by renumbering. For some time, the working group debated the use of site local addresses for the internal BGP connections. The idea was not adopted, because there is no clear relation between a "site" and an autonomous system. What was adopted was a possibility to encode either a global address or the combination of a global address and a link local address in the next-hop information, so that routers that happen to be on the same link can use the link local address and not be affected by renumbering operations. The use of BGP-4 for IPv6 is detailed in RFC-2545 [25].

10.5 Waiting for the New IP

Despite spectacular incidents, the interdomain routing architecture has been able to cope with the growth of the Internet. The whole process relies heavily on manual configuration of routing parameters in "BGP configuration tables," which requires very skilled operators. As the Internet continues to grow, as we transition toward a multiprotocol environment, BGP probably will continue to evolve and new parameters will be defined. It will certainly be many years before we see a complete transition to IPv6, and many more years before a new routing technology appears.

References

1. C. Labovitz, G. R. Malan, and F. Jahanian, "Internet Routing Instability," Proceedings of ACM/SIGCOMM 1997, *Computer Communication Review*, Vol. 27, No. 4, October 1997.
2. V. Paxson, "End to End Routing Behavior in the Internet," Proceedings of ACM/SIGCOMM 1996, *Computer Communication Review*, October 1996.
3. C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental Study of Internet Stability and Wide-Area Backbone Failures," *Proceedings of IEEE/INFOCOM '99*, March 1999.
4. C. Villamizar, R. Chandra, and R. Govindan, "BGP Route Flap Damping," RFC-2439, November 1998.
5. Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," RFC-1771, March 1995.
6. Listing of Internet Routing Registries, http://www.merit.edu/list-of-routing-registries.html

7. T. Bates, E. Gerich, L. Joncheray, J-M. Jouanigot, D. Karrenberg, M. Terpstra, and J. Yu, "Representation of IP Routing Policies in a Routing Registry," RIPE report RIPE-181, October 1994. Also, RFC-1786, March 1995.
8. D. Eastlake, "Domain Name System Security Extensions," RFC-2535, March 1999.
9. S. Murphy, "BGP Security Analysis," work in progress, June 1999.
10. P. Traina, "Autonomous System Confederations for BGP," RFC-1965, June 1996.
11. T. Bates and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh IBGP," RFC-1966, June 1996.
12. C. Kunzinger, Editor, "Inter-Domain Routing Protocol," ISO/IEC 10747, October 1993.
13. R. Chandra and P. Traina, "BGP Communities Attribute," RFC-1997, August 1996.
14. E. Chen and T. Bates, "An Application of the BGP Community Attribute in Multi-Home Routing," RFC-1998, August 1996.
15. V. Fuller, T. Li, J. I. Yu, and K. Varadhan, "Supernetting: An Address Assignment and Aggregation Strategy," RFC-1338, June 1992.
16. V. Fuller, T. Li, J. I. Yu, and K. Varadhan, "Classless Inter-Domain Routing (CIDR): An Address Assignment and Aggregation Strategy," RFC-1519, September 1993.
17. E. Gerich, "Guidelines for Management of IP Address Space," RFC-1466, May 1993.
18. Y. Rekhter and T. Li, "An Architecture for IP Address Allocation with CIDR," RFC-1518, September 1993.
19. H. Berkowitz, P. Ferguson, W. Leland, and P. Nesser, "Enterprise Renumbering: Experience and Information Solicitation," RFC-1916, February 1996.
20. Y. Rekhter, B. Moskowitz, D. Karrenberg, and G. de Groot, "Address Allocation for Private Internets," RFC-1597, March 1994.
21. E. Lear, E. Fair, D. Crocker, and T. Kessler, "Network 10 Considered Harmful (Some Practices Shouldn't Be Codified)," RFC-1627, June 1994.
22. Y. Rekhter, B. Moskowitz, D. Karrenberg, G. de Groot, and E. Lear, "Address Allocation for Private Internets," RFC-1918, February 1996.
23. Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," RFC-1771, March 1995.
24. T. Bates and R. Chandra, "Multiprotocol Extensions for BGP-4," RFC-2283, February 1998.
25. P. Marques and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing," RFC-2545, March 1999.

Page 250 to be blank