Network Working Group
Request for Comments: 3221
Category: Informational

G. Huston
Internet Architecture Board

December 2001

Commentary on Inter-Domain Routing in the Internet

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Copyright Notice

Abstract

This document examines the various longer term trends visible within the characteristics of the Internet's BGP table and identifies a number of operational practices and protocol factors that contribute to these trends. The potential impacts of these practices and protocol properties on the scaling properties of the inter-domain routing space are examined.

This document is the outcome of a collaborative exercise on the part of the Internet Architecture Board.

Table of Contents

1. Introduction

This document examines the various longer term trends visible within the characteristics of the Internet's BGP table and identifies a number of operational practices and protocol factors that contribute to these trends. The potential impacts of these practices and protocol properties on the scaling properties of the inter-domain routing space are examined.

These impacts include the potential for exhaustion of the existing Autonomous System number space, increasing convergence times for selection of stable alternate paths following withdrawal of route announcements, the stability of table entries, and the average prefix length of entries in the BGP table. The larger long term issue is that of an increasingly denser inter-connectivity mesh between ASes, causing a finer degree of granularity of inter-domain policy and finer levels of control to undertake inter-domain traffic engineering.

Various approaches to a refinement of the inter-domain routing protocol and associated operating practices that may provide superior scaling properties are identified as an area for further investigation.

This document is the outcome of a collaborative exercise on the part of the Internet Architecture Board.

2. Network Scaling and Inter-Domain Routing

Are there inherent scaling limitations in the technology of the Internet or its architecture of deployment that may impact on the ability of the Internet to meet escalating levels of demand? There are a number of potential areas to search for such limitations. These include the capacity of transmission systems, packet switching capacity, the continued availability of protocol addresses, and the capability of the routing system to produce a stable view of the overall topology of the network. In this study we will look at this latter capability with the objective of identifying some aspects of the scaling properties of the Internet's routing system.

The basic structure of the Internet is a collection of networks, or Autonomous Systems (ASes) that are interconnected to form a connected domain. Each AS uses an interior routing system to maintain a coherent view of the topology within the AS, and uses an exterior routing system to maintain adjacency information with neighboring ASes to create a view of the connectivity of the entire system.

This network-wide connectivity is described in the routing table used by the BGP4 protocol (referred to as the Routing Information Base, or RIB). Each entry in the table refers to a distinct route. The attributes of the route, together with local policy constraints, are used to determine the best path from the local AS to the AS that is originating the route. Determining the 'best path' in this case is determining which routing advertisement and associated next hop address is the most preferred by the local AS. Within each local BGP-speaking router this preferred route is then loaded into the local RIB (Loc-RIB). This information is coupled with information obtained from the local instance of the interior routing protocol to form a Forwarding Information Base (or FIB), for use by the local router's forwarding engine.

The BGP routing system is not aware of finer level of topology of the network on a link-by-link basis within the local AS or within any remote AS. From this perspective BGP can be seen as an inter-AS connectivity maintenance protocol, as distinct from a link-level topology management protocol, and the BGP routing table can be viewed as a description of the current connectivity of the Internet using an AS as the basic element of connectivity computation.

There is an associated dimension of policy determination within the routing table. If an AS advertises a route to a neighboring AS, the local AS is offering to accept traffic from the neighboring AS which is ultimately destined to addresses described by the advertised routing entry. If the local AS does not originate the route, then the inference is that the local AS is willing to undertake the role of transit provider for this traffic on behalf of some third party. Similarly, an AS may or may not choose to accept a route from a neighbor. Accepting a route implies that under some circumstances, as determined by the local route selection parameters, the local AS will use the neighboring AS to reach addresses spanned by the route. The BGP routing domain is intended to maintain a coherent view of the connectivity of the inter-AS domain, where connectivity is expressed as a preference for 'shortest paths' to reach any destination address as modulated by the connectivity policies expressed by each AS, and coherence is expressed as a global constraint that none of the paths contains loops or dead ends. The elements of the BGP routing domain are routing entries, expressed as a span of addresses. All addresses advertised within each routing entry share a common origin AS and a common connectivity policy. The total size of the BGP table is therefore a metric of the number of distinct routes within the Internet, where each route describes a contiguous set of addresses that share a common origin AS and a common reachability policy.

When the scaling properties of the Internet were studied in the early 1990s two critical factors identified in the study were, not surprisingly, routing and addressing [2]. As more devices connect to the Internet they consume addresses, and the associated function of maintaining reachability information for these addresses, with an assumption of an associated growth in the number of distinct provider networks and the number of distinct connectivity policies, implies ever larger routing tables. The work in studying the limitations of the 32 bit IPv4 address space produced a number of outcomes, including the specification of IPv6 [3], as well as the refinement of techniques of network address translation [4] intended to allow some degree of transparent interaction between two networks using different address realms. Growth in the routing system is not directly addressed by these approaches, as the routing space is the cross product of the complexity of the inter-AS topology of the network, multiplied by the number of distinct connectivity policies multiplied by the degree of fragmentation of the address space. For example, use of NAT may reduce the pressure on the number of public addresses required by a single connected network, but it does not necessarily imply that the network's connectivity policies can be subsumed within the aggregated policy of a single upstream provider.

When an AS advertises a block of addresses into the exterior routing space this entry is generally carried across the entire exterior routing domain of the Internet. To measure the common characteristics of the global routing table, it is necessary to establish a point in the default-free part of the exterior routing domain and examine the BGP routing table that is visible at that point.

3. Measurements of the total size of the BGP Table

Measurements of the size of the routing table were somewhat sporadic to start, and a number of measurements were taken at approximate monthly intervals from 1988 until 1992 by Merit [5]. This effort was resumed in 1994 by Erik-Jan Bos at Surfnet in the Netherlands, who commenced measuring the size of the BGP table at hourly intervals in 1994. This measurement technique was adopted by the author in 1997, using a measurement point located at the edge of AS 1221 at Telstra in Australia, again using an hourly interval for the measurement. The initial measurements were of the number of routing entries contained within the set of selected best paths. These measurements were expanded to include the number of AS numbers, number of AS paths, and a set of measurements relating to the prefix size of routing table entries.

This data contains a view of the dynamics of the Internet's routing table growth that spans some 13 years in total and includes a very detailed view spanning the most recent seven years [6]. Looking at just the total size of the BGP routing table over this period, it is possible to identify four distinct phases of inter-AS routing practice in the Internet.

3.1 Pre-CIDR Growth

The initial characteristics of the routing table size from 1988 until April 1994 show definite characteristics of exponential growth. If continued unchecked, this growth would have lead to saturation of the available BGP routing table space in the non-default routers of the time within a small number of years.

Estimates of the time at which this would've happened varied somewhat from study to study, but the overall general theme of these observations was that the growth rates of the BGP routing table were exceeding the growth in hardware and software capability of the deployed network, and that at some point in the mid-1990's, the BGP table size would have grown to the point where it was larger than the capabilities of available equipment to support.

3.2 CIDR Deployment

The response from the engineering community was the introduction of a hierarchy into the inter-domain routing system. The intent of the hierarchical routing structure was to allow a provider to merge the routing entries for its customers into a single routing entry that spanned its entire customer base. The practical aspects of this change was the introduction of routing protocols that dispensed with the requirement for the Class A, B and C address delineation, replacing this scheme with a routing system that carried an address prefix and an associated prefix length. This approached was termed Classless Inter-Domain Routing (CIDR) [5].

A concerted effort was undertaken in 1994 and 1995 to deploy CIDR routing in the Internet, based on encouraging deployment of the CIDR-capable version of the BGP protocol, BGP4 [7].

The intention of CIDR was one of hierarchical provider address aggregation, where a network provider was allocated an address block from an address registry, and the provider announced this entire block into the exterior routing domain as a single entry with a single routing policy. Customers of the provider were encouraged to use a sub-allocation from the provider's address block, and these smaller routing elements were aggregated by the provider and not directly passed into the exterior routing domain. During 1994 the

size of the routing table remained relatively constant at some 20,000 entries as the growth in the number of providers announcing address blocks was matched by a corresponding reduction in the number of address announcements as a result of CIDR aggregation.

3.3 CIDR Growth

For the next four years until the start of 1998, CIDR proved effective in damping unconstrained growth in the BGP routing table. During this period, the BGP table grew at an approximate linear rate, adding some 10,000 entries per year.

A close examination of the table reveals a greater level of stability in the routing system at this time. The short term (hourly) variation in the number of announced routes reduced, both as a percentage of the number of announced routes, and also in absolute terms. One of the other benefits of using large aggregate address blocks is that instability at the edge of the network is not immediately propagated into the routing core. The instability at the last hop is absorbed at the point where an aggregate route is used in place of a collection of more specific routes. This, coupled with widespread adoption of BGP route flap damping, was very effective in reducing the short term instability in the routing space during this period.

3.4 Current Growth

In late 1998 the trend of growth in the BGP table size changed radically, and the growth for the period 1998 - 2000 is again showing all the signs of a re-establishment of a growth trend with strong correlation to an exponential growth model. This change in the growth trend appears to indicate that pressure to use hierarchical address allocations and CIDR has been unable to keep pace with the levels of growth of the Internet, and some additional factors that impact the growth in the BGP table size have become more prominent in the Internet. This has lead to a growth pattern in the total size of the BGP table that has more in common with a compound growth model than a linear model. A good fit of the data for the period from January 1999 until December 2000 is a compound growth model of 42% growth per year.

An initial observation is that this growth pattern points to some weakening of the hierarchical model of connectivity and routing within the Internet. To identify the characteristics of this recent trend it is necessary to look at a number of related characteristics of the routing table.

BGP table size data for the first half of 2001 shows different trends at various measurement points in the Internet. Some measurement points where the local AS has a relative larger number of more specific routes show a steady state for the first half of 2001 with no appreciable growth, while other measurement points where the local AS has had a lower number of more specific routes initially show a continuation of table size growth. There are a number of commonly observed discontinuities in the data for 2001, corresponding to events where a significant number of more specific entries have been replaced by an encompassing aggregate prefix.

4. Related Measurements derived from BGP Table

The level of analysis of the BGP routing table has been extended in an effort to identify the factors contributing to this growth, and to determine whether this leads to some limiting factors in the potential size of the routing space. Analysis includes measuring the number of ASes in the routing system, and the number of distinct AS paths, the range of addresses spanned by the table and average span of each routing entry.

4.1 AS Number Consumption

Each network that is multi-homed within the topology of the Internet and wishes to express a distinct external routing policy must use a unique AS number to associate its advertised addresses with such a policy. In general, each network is associated with a single AS, and the number of ASes in the default-free routing table tracks the number of entities that have unique routing policies. There are some exceptions to this, including large global transit providers with varying regional policies, where multiple ASes are associated with a single network, but such exceptions are relatively uncommon.

The number of unique ASes present in the BGP table has been tracked since late 1996, and the trend of AS number deployment over the past four years is also one that matches a compound growth model with a growth rate of 51% per year. As of the start of May 2001 there were some 10,700 ASes visible in the BGP table. At a continued rate of growth of 51% p.a., the 16 bit AS number space will be fully deployed by August 2005. Work is underway within the IETF to modify the BGP protocol to carry AS numbers in a 32-bit field. [8] While the protocol modifications are relatively straightforward, the major responsibility rests with the operations community to devise a transition plan that will allow gradual transition into this larger AS number space.

4.2 Address Consumption

It is also possible to track the total amount of address space advertised within the BGP routing table. At the start of 2001 the routing table encompassed 1,081,131,733 addresses, or some 25.17% of the total IPv4 address space, or 25.4% of the usable unicast public address space. By September 2001 this has growth to 1,123,124,472 addresses, or some 26% of the IPv4 address space. This has grown from 1,019,484,655 addresses in November 1999. However, there are a number of /8 prefixes that are periodically announced and withdrawn from the BGP table, and if the effects of these prefixes is removed, a compound growth model against the previous 12 months of data of this metric yields a best fit model of growth of 7% per year in the total number of addresses spanned by the routing table.

Compared to the 42% growth in the number of routing advertisements, the growth in the amount of address space advertised is far lower. One possible explanation is that much of the growth of the Internet in terms of growth in the number of connected devices is occurring behind various forms of NAT gateways. In terms of solving the perceived finite nature of the address space identified just under a decade ago, this explanation would tend to indicate that the Internet appears so far to have embraced the approach of using NATs, irrespective of their various perceived functional shortcomings. [9] This explanation also supports the observation of smaller address fragments supporting distinct policies in the BGP table, as such small address blocks may encompass arbitrarily large networks located behind one or more NAT gateways. There are alternative explanations of this difference between the growth of the table and the growth of address space, including a trend towards discrete exterior routing policies being applied to finer address blocks.

4.3 Granularity of Table Entries

The intent of CIDR aggregation was to support the use of large aggregate address announcements in the BGP routing table. To confirm whether this is still the case the average span of each BGP announcement has been tracked for the past 12 months. The data indicates a decline in the average span of a BGP advertisement from 16,000 individual addresses in November 1999 to 12,100 in December 2000. As of September 2001 this span has been further reduced to an average 10,700 individual addresses per routing entry. This corresponds to an increase in the average prefix length from /18.03 to /18.44 by December 2000 and a /18.6 by September 2001. Separate observations of the average prefix length used to route traffic in operation networks in late 2000 indicate an average length of 18.1 [11]. This trend towards finer-grained entries in the routing table is potentially cause for concern, as it implies the increasing spread

of traffic over greater numbers of increasingly smaller forwarding table entries. This, in turn, has implications for the design of high speed core routers, particularly when extensive use is made of a small number of very high speed cached forwarding entries within the switching subsystem of a router's design.

A similar observation can be made regarding the number of addresses advertised per AS. In December 1999 each AS advertised an average of 161,900 addresses (equivalent to a prefix length /14.69, and in January 2001 this average has fallen to 115,800 addresses, an equivalent prefix length of /15.18.

This points to increasingly finer levels of routing detail being announced into the global routing domain. This, in turn, supports the observation that the efficiencies of hierarchical routing structures are no longer being fully realized within the deployed Internet. Instead, increasingly finer levels of routing detail are being announced globally in the BGP tables. The most likely cause of this trend of finer levels of routing granularity is an increasingly dense interconnection mesh, where more networks are moving from a single-homed connection with hierarchical addressing and routing into multi-homed connections without any hierarchical structure. The spur for this increasingly dense connectivity mesh in the Internet may well be the declining unit costs of communications bearer services coupled with a common perception that richer sets of adjacencies yields greater levels of service resilience.

4.4 Prefix Length Distribution

In addition to looking at the average prefix length, the analysis of the BGP table also includes an examination of the number of advertisements of each prefix length.

An extensive program commenced in the mid-nineties to move away from intense use of the Class C space and to encourage providers to advertise larger address blocks, as part of the CIDR effort. This has been reinforced by the address registries who have used provider allocation blocks that correspond to a prefix length of /19 and, more recently, /20.

These measures were introduced in the mid-90's when there were some 20,000 - 30,000 entries in the BGP table. Some six years later in April 2001 it is interesting to note that of the 108,000 entries in the routing table, some 59,000 entries have a /24 prefix. In absolute terms the /24 prefix set is the fastest growing set in the BGP routing table. The routing entries of these smaller address blocks also show a much higher level of change on an hourly basis. While a large number of BGP routing points perform route flap

damping, nevertheless there is still a very high level of announcements and withdrawals of these entries in this particular area of the routing table when viewed using a perspective of route updates per prefix length. Given that the numbers of these small prefixes are growing rapidly, there is cause for some concern that the total level of BGP flux, in terms of the number of announcements and withdrawals per second may be increasing, despite the pressures from flap damping. This concern is coupled with the observation that, in terms of BGP stability under scaling pressure, it is not the absolute size of the BGP table that is of prime importance, but the rate of dynamic path re-computations that occur in the wake of announcements and withdrawals. Withdrawals are of particular concern due to the number of transient intermediate states that the BGP distance vector algorithm explores in processing a withdrawal. Current experimental observations indicate a typical convergence time of some 2 minutes to propagate a route withdrawal across the BGP domain. [10]

An increase in the density of the BGP mesh, coupled with an increase in the rate of such dynamic changes, does have serious implications in maintaining the overall stability of the BGP system as it continues to grow. The registry allocation policies also have had some impact on the routing table prefix distribution. The original registry practice was to use a minimum allocation unit of a /19, and the 10,000 prefix entries in the /17 to /19 range are a consequence of this policy decision. More recently, the allocation policy now allows for a minimum allocation unit of a /20 prefix, and the /20 prefix is used by some 4,300 entries as of January 2001, and in relative terms is one of the fastest growing prefix sets. The number of entries corresponding to very small address blocks (smaller than a /24), while small in number as a proportion of the total BGP routing table, is the fastest growing in relative terms. The number of /25 through /32 prefixes in the routing table is growing faster, in terms of percentage change, than any other area of the routing table. If prefix length filtering were in widespread use, the practice of announcing a very small address block with a distinct routing policy would have no particular beneficial outcome, as the address block would not be passed throughout the global BGP routing domain and the propagation of the associated policy would be limited in scope. The growth of the number of these small address blocks, and the diversity of AS paths associated with these routing entries, points to a relatively limited use of prefix length filtering in today’s Internet. In the absence of any corrective pressure in the form of widespread adoption of prefix length filtering, the very rapid growth of global announcements of very small address blocks is likely to continue. In percentage terms, the set of prefixes spanning /25 to /32 show the largest growth rates.

4.5 Aggregation and Holes

With the CIDR routing structure it is possible to advertise a more specific prefix of an existing aggregate. The purpose of this more specific announcement is to punch a 'hole' in the policy of the larger aggregate announcement, creating a different policy for the specifically referenced address prefix.

Another use of this mechanism is to perform a rudimentary form of load balancing and mutual backup for multi-homed networks. In this model a network may advertise the same aggregate advertisement along each connection, but then advertise a set of specific advertisements for each connection, altering the specific advertisements such that the load on each connection is approximately balanced. The two forms of holes can be readily discerned in the routing table - while the approach of policy differentiation uses an AS path that is different from the aggregate advertisement, the load balancing and mutual backup configuration uses the same As path for both the aggregate and the specific advertisements. While it is difficult to understand whether the use of such more specific advertisements was intended to be an exception to a more general rule or not within the original intent of CIDR deployment, there appears to be very widespread use of this mechanism within the routing table. Some 59,000 advertisements, or 55% of the total number of routing table entries, are being used to punch policy holes in existing aggregate announcements. Of these the overall majority of some 42,000 routes use distinct AS paths, so that it does appear that this is evidence of finer levels of granularity of connection policy in a densely interconnected space. While long term data is not available for the relative level of such advertisements as a proportion of the full routing table, the growth level does strongly indicate that policy differentiation at a fine level within existing provider aggregates is a significant driver of overall table growth.

5. Current State of inter-AS routing in the Internet

The resumption of compound growth trends within the BGP table, and the associated aspects of finer granularity of routing entries within the table form adequate grounds for consideration of potential refinements to the Internet's exterior routing protocols and potential refinements to current operating practices of inter-AS connectivity. With the exception of the 16 bit AS number space, there is no particular finite limit to any aspect of the BGP table. The motivation for such activity is that a long term pattern of continued growth at current rates may once again pose a potential condition where the capacity of the available processors may be exceeded by some aspect of the Internet routing table.

5.1 A denser interconnectivity mesh

The decreasing unit cost of communications bearers in many part of the Internet is creating a rapidly expanding market in exchange points and other forms of inter-provider peering. A model of extensive interconnection at the edges of the Internet is rapidly supplanting the deployment model of a single-homed network with a single upstream provider. The underlying deployment model of CIDR was that of a single-homed network, allowing for a strict hierarchy of supply providers. The business imperatives driving this denser mesh of interconnection in the Internet are substantial, and the casualty in this case is the CIDR-induced dampened growth of the BGP routing table.

5.2 Multi-Homed small networks and service resiliency

It would appear that one of the major drivers of the recent growth of the BGP table is that of small networks, advertised as a /24 prefix entry in the routing table, multi-homing with a number of peers and upstream providers. In the appropriate environment where there are a number of networks in relatively close proximity, using peer relationships can reduce total connectivity costs, as compared to using a single upstream service provider. Equally significantly, multi-homing with a number of upstream providers is seen as a means of improving the overall availability of the service. In essence, multi-homing is seen as an acceptable substitute for upstream service resiliency. This has a potential side effect that when multi-homing is seen as a preferable substitute for upstream provider resiliency, the upstream provider cannot command a price premium for proving resiliency as an attribute of the provided service, and therefore has little economic incentive to spend the additional money required to engineer resiliency into the network. The actions of the network's multi-homed clients then become self-fulfilling. One way to characterize this behavior is that service resiliency in the Internet is becoming the responsibility of the customer, not the service provider.

In such an environment resiliency still exists, but rather than being a function of the bearer or switching subsystem, resiliency is provided through the function of the BGP routing system. The question is not whether this is feasible or desirable in the individual case, but whether the BGP routing system can scale adequately to continue to undertake this role.

5.3 Traffic Engineering via Routing

Further driving this growth in the routing table is the use of selective advertisement of smaller prefixes along different paths in an effort to undertake traffic engineering within a multi-homed environment. While there is considerable effort being undertaken to develop traffic engineering tools within a single network using MPLS as the base flow management tool, inter-provider tools to achieve similar outcomes are considerably more complex when using such switching techniques.

At this stage the only tool being used for inter-provider traffic engineering is that of the BGP routing table. Such use of BGP appears to place additional fine-grained prefixes into the routing table. This action further exacerbates the growth and stability pressures being placed on the BGP routing domain.

5.4 Lack of Common Operational Practices

There is considerable evidence of a lack of uniformity of operational practices within the inter-domain routing space. This includes the use and setting of prefix filters, the use and setting of route damping parameters and level of verification undertaken on BGP advertisements by both the advertiser and the recipient. There is some extent of 'noise' in the routing table where advertisements appear to be propagated well beyond their intended domain of applicability, and also where withdrawals and advertisements are not being adequately damped close to the origin of the route flap. This diversity of operating practices also extends to policies of accepting advertisements that are more specific advertisements of existing provider blocks.

5.5 CIDR and Hierarchical Routing

The current growth factors at play in the BGP table are not easily susceptible to another round of CIDR deployment pressure within the operator community. The denser interconnectivity mesh, the increasing use of multi-homing with smaller address prefixes, the extension of the use of BGP to perform roles related to inter-domain traffic engineering and the lack of common operating practices all point to a continuation of the trend of growth in the total size of the BGP routing table, with this growth most apparent with advertisements of smaller address blocks, and an increasing trend for these small advertisements to be punching a connectivity policy 'hole' in an existing provider aggregate advertisement.

It may be appropriate to consider how to operate an Internet with a BGP routing table that has millions of small entries, rather than the expectation of a hierarchical routing space with at most tens of thousands of larger entries in the global routing table.

6. Future Requirements for the Exterior Routing System

It is beyond the scope of this document to define a scalable inter-domain routing environment and associated routing protocols and operating practices. A more modest goal is to look at the attributes of routing systems as understood and identify those aspects of such systems that may be applicable to the inter-domain environment as a potential set of requirements for inter-domain routing tools.

6.1 Scalability

The overall intent is scalability of the routing environment. Scalability can be expressed in many dimensions, including number of discrete network layer reachability entries, number of discrete route policy entries, level of dynamic change over a unit of time of these entries, time to converge to a coherent view of the connectivity of the network following changes, and so on.

The basic objective behind this expressed requirement for scalability is that the most likely near to medium trend in the structure of the Internet is a continuation in the pattern of dense interconnectivity between a large number of discrete network entities, and little impetus behind hierarchical aggregating structures. It is not an objective to place any particular metrics on scalability within this examination of requirements, aside from indicating that a prudent view would encompass a scale of connectivity in the inter-domain space that is at least two orders of magnitude larger than comparable metrics of the current environment.

6.2 Stability and Predictability

Any routing system should behave in a stable and predictable fashion. What is inferred from the predictability requirement is the behavior that under identical environmental conditions the routing system should converge to the same state. Stability implies that the routing state should be maintained for as long as the environmental conditions remain constant. Stability also implies a qualitative property that minor variations in the network’s state should not cause large scale instability across the entire network while a new stable routing state is reached. Instead, routing changes should be propagated only as far as necessary to reach a new stable state, so that the global requirement for stability implies some degree of locality in the behavior of the system.

6.3 Convergence

Any routing system should have adequate convergence properties. By adequate it is implied that within a finite time following a change in the external environment, the routing system will have reached a shared common description of the network's topology that accurately describes the current state of the network and is stable. In this case finite time implies a time limit that is bounded by some upper limit, and this upper limit reflects the requirements of the routing system. In the case of the Internet this convergence time is currently of the order of hundreds of seconds as an upper bound on convergence. This long convergence time is perceived as having a negative impact on various applications, particularly those that are time critical. A more useful upper bound for convergence is of the order of seconds or lower if it is desired to support a broad range of application classes.

It is not a requirement to be able to undertake full convergence of the inter-domain routing system in the sub-second timescale.

6.4 Routing Overhead

The greater the amount of information passed within the routing system, and the greater the frequency of such information exchanges, the greater the level of expectation that the routing system can maintain an accurate view of the connectivity of the network. Equally, the greater the amount of information passed within the routing system, and the higher the frequency of information exchange, the higher the level of overhead consumed by operation of the routing system. There is an element of design compromise in a routing system to pass enough information across the system to allow each routing element to have adequate local information to reach a coherent local view of the network, yet ensure that the total routing overhead is low.

7. Architectural approaches to a scalable Exterior Routing Protocol

This document does not attempt to define an inter-domain routing protocol that possess all the attributes as listed above, but a number of architectural considerations can be identified that would form an integral part of the protocol design process.

7.1 Policy opaqueness vs. policy transparency

The two major approaches to routing protocols are distance vector and link state.

In the distance vector protocol a routing node gathers information from its neighbors, applies local policy to this information and then distributes this updated information to its neighbors. In this model the nature of the local policy applied to the routing information is not necessarily visible to the node's neighbors, and the process of converting received route advertisements into advertised route advertisements uses a local policy process whose policy rules are not visible externally. This scenario can be described as 'policy opaque'. The side effect of such an environment is that a third party cannot remotely compute which routes a network may accept and which may be re-advertised to each neighbor.

In link state protocols a routing node effectively broadcasts its local adjacencies, and the policies it has with respect to these adjacencies, to all nodes within the link state domain. Every node can perform an identical computation upon this set of adjacencies and associated policies in order to compute the local forwarding table. The essential attribute of this environment is that the routing node has to announce its routing policies, in order to allow a remote node to compute which routes will be accepted from which neighbor, and which routes will be advertised to each neighbor and what, if any, attributes are placed on the advertisement. Within an interior routing domain the local policies are in effect metrics of each link and these polices can be announced within the routing domain without any consequent impact.

In the exterior routing domain it is not the case that interconnection policies between networks are always fully transparent. Various permutations of supplier / customer relationships and peering relationships have associated policy qualifications that are not publicly announced for business competitive reasons. The current diversity of interconnection arrangements appears to be predicated on policy opaqueness, and to mandate a change to a model of open interconnection policies may be contrary to operational business imperatives.

An inter-domain routing tool should be able to support models of interconnection where the policy associated with the interconnection is not visible to any third party. If the architectural choice is a constrained one between distance vector and link state, then this consideration would appear to favor the continued use of a distance vector approach to inter-domain routing. This choice, in turn, has implications on the convergence properties and stability of the inter-domain routing environment. If there is a broader spectrum of choice, the considerations of policy-opaqueness would still apply.

7.2 The number of routing objects

The current issues with the trend behaviors of the BGP space can be coarsely summarized as the growth in the number of distinct routing objects, the increased level of dynamic behaviors of these objects (in the form of announcements and withdrawals).

This entails evaluating possible measures that can address the growth rate in the number of objects in the inter-domain routing table, and separately examining measures that can reduce the level of dynamic change in the routing table. The current routing architecture defines a basic unit of a route object as an originating AS number and an address prefix.

In looking at the growth rate in the number of route objects, the salient observation is that the number of route objects is the byproduct of the density of the interconnection mesh and the number of discrete points where policy is imposed of route objects. One approach to reduce the growth in the number of objects is to allow each object to describe larger segments of infrastructure. Such an approach could use a single route object to describe a set of address prefixes, or a collection of ASs, or a combination of the two. The most direct form of extension would be to preserve the assumption that each routing object represents an indivisible policy entity. However, given that one of the drivers of the increasing number of route objects is a proliferation of discrete route objects, it is not immediately apparent that this form of aggregation will prove capable in addressing the growth in the number of route objects.

If single route objects are to be used that encompass a set of address prefixes and a collection of ASs, then it appears necessary to define additional attributes within the route object to further qualify the policies associated with the object in terms of specific prefixes, specific ASs and specific policy semantics that may be considered as policy exceptions to the overall aggregate

Another approach to reduce the number of route objects is to reduce the scope of advertisement of each routing object, allowing the object to be removed and proxy aggregated into some larger object once the logical scope of the object has been reached. This approach would entail the addition of route attributes that could be used to define the circumstances where a specific route object would be subsumed by an aggregate route object without impacting the policy objectives associated with the original set of advertisements.

7.3 Inter-domain Traffic Engineering

Attempting to place greater levels of detail into route objects is intended to address the dual role of the current BGP system as both an inter-domain connectivity maintenance protocol and as an implicit traffic engineering tool.

In the current environment, advertisement of more specific prefixes with unique policy but with the same origin AS is often intended to create a traffic engineering response, where incoming traffic to an AS may be balanced across multiple paths. The outcome is that the control of the relative profile of load is placed with the originating AS. The way this is achieved is by using limited knowledge of the remote AS's route selection policy to explicitly limit the number of egress choices available to a remote AS. The most common route selection policy is the preference for more specific prefixes over larger address blocks. By advertising specific prefixes along specific neighbor AS connections with specific route attributes, traffic destined to these addresses is passed through the selected transit paths. This limitation of choice allows the originating AS to override the potential policy choices of all other ASs, imposing its traffic import policies at a higher level than the remote AS's egress policies.

An alternative approach is the use of a class of traffic engineering attributes that are attached to an aggregate route object. The intent of such attributes is to direct each remote AS to respond to the route object in a manner that equates to the current response to more specific advertisements, but without the need to advertise specific prefix route objects. However, even this approach uses route objects to communicate traffic engineering policy, and the same risk remains that the route table is used to carry fine-detailed traffic path policies.

An alternative direction is to separate the functions of connectivity maintenance and traffic engineering, using the routing protocol to identify a number of viable paths from a source AS to a destination AS, and use a distinct collection of traffic engineering tools to allow a traffic source AS to make egress path selections that match the desired traffic service profile for the traffic.

There is one critical difference between traffic engineering approaches as used in intra-domain environments and the current inter-domain operating practices. Whereas the intra-domain environment uses the ingress network element to make the appropriate path choice to the egress point, the inter domain traffic engineering has the opposite intent, where a downstream AS (or egress point) is attempting to influence the path choice of an upstream AS (or ingress

point). If explicit traffic engineering were undertaken within the inter-domain space, it is highly likely that the current structure would be altered. Instead of the downstream element attempting to constrain the path choices of an upstream element, a probable approach is the downstream element placing a number of advisory constraints on the upstream elements, and the upstream elements using a combination of these advisory constraints, dynamic information relating to path service characteristics and local policies to make an egress choice.

From the perspective of the inter-domain routing environment, such measures offer the potential to remove the advertisement of specific routes for traffic engineering purposes. However, there is a need to adding traffic engineering information into advertised route blocks, requiring the definition of the syntax and semantics of traffic engineering attributes that can be attached to route objects.

7.4 Hierarchical Routing Models

The CIDR routing model assumed a hierarchy of providers, where at each level in the hierarchy the routing policies and address space of networks at the lower level of hierarchy were subsumed by the next level up (or 'upstream') provider. The connectivity policy assumed by this model is also a hierarchical model, where horizontal connections within a single level of the hierarchy are not visible beyond the networks of the two parties.

A number of external factors are increasing the density of interconnection including decreasing unit costs of communications services and the increasing use of exchange points to augment point-to-point connectivity models with point-to-multi-point facilities.

The outcome of these external factors is a significant reduction in the hierarchical nature of the inter-domain space. Such a trend can be viewed with concern given the common approach of using hierarchies as a tool for scaling routing systems. BGP falls within this approach, and relies on hierarchies in the address space to contain the number of independently routing objects. The outcomes of this characteristic of the Internet in terms of the routing space is the increasing number of distinct route policies that are associated with each multi-homed network within the Internet.

One way to limit the proliferation of such policies across the entire inter-domain space is to associate attributes to such advertisements that specify the conditions whereby a remote transit AS may proxy-aggregate this route object with other route objects.

7.5 Extend or Replace BGP

A final consideration is to consider whether these requirements can best be met by an approach of a set of upward-compatible extensions to BGP, or by a replacement to BGP. No recommendation is made here, and this is a topic requiring further investigation.

The general approach in extending BGP appears to lie in increasing the number of supported transitive route attributes, allowing the route originator greater control in specifying the scope of propagation of the route and the intended outcome in terms of policy and traffic engineering. It may also be necessary to allow BGP sessions to negotiate additional functionality intended to improve the convergence behavior of the protocol. Whether such changes can produce a scalable and useful outcome in terms of inter-domain routing remains, at this stage, an open question.

An alternative approach is that of a replacement protocol, and such an approach may well be based on the adoption of a link-state behavior. The issues of policy opaqueness and link-state protocols have been described above. The other major issue with such an approach is the need to limit the extent of link state flooding, where the inter-domain space would need some further levels of imposed structure similar to intra-domain areas. Such structure may well imply the need for an additional set of operator inter-relationships such as mutual transit, and this may prove challenging to adapt to existing practices.

The potential sets of actions include more than extend or replace the BGP protocol. A third approach is to continue to use BGP as the basic means of propagating route objects and their associated AS paths and other attributes, and use one or more overlay protocols to support inter-domain traffic engineering and other forms of inter-domain policy negotiation. This approach would appear to offer a means of transition for the large installed base currently using BGP4 as their inter-domain routing protocol, placing additional functionality in the overlay protocols while leaving the basic functionality of BGP4 intact. The resultant inter-dependencies between BGP and the overlay protocols would require very careful attention, as this would be the most critical aspect of such an approach.

8. Directions for Further Activity

While there may exist short term actions based on providing various incentives for network operators to remove redundant or inefficiently grouped entries from the BGP routing table, such actions are short term palliative measures, and will not provide long term answers to the need to a scalable inter-domain routing protocol.

One potential short term protocol refinement is to allow a set of grouped advertisements to be aggregated into a single route advertisement. This form of proxy aggregation would take a set of bit-wise aligned routing entries with matching route attributes, and under certain well identified circumstances, aggregate these routing entries into a single re-advertised aggregate routing entry. This technique removes information from the routing system, and some care must be taken to define a set of proxy aggregation conditions that do not materially alter the flow of traffic, or the ability of originating ASes to announce routing policy.

A further refinement to this approach is to consider the definition of the syntax and semantics of a number of additional route attributes. Such attributes could define the extent to which specific route advertisements should be propagated in the inter-domain space, allowing the advertisement to be subsumed by a larger aggregate advertisement at the boundary of this domain. This could be used to form part of the preconditions of automated proxy aggregation of specific routes, and also limit the extent to which announcement and withdrawals are propagated across the routing domain.

It is unclear that such measures would result in substantial longer term changes to the scaling and convergence properties of BGP4. Taking the requirement set enumerated in section 6 of this document, one approach to the longer term requirements may be to preserve a number of attributes of the current BGP protocol, while refine other aspects of the protocol to improve its scaling and convergence properties. A minimal set of alterations could retain the Autonomous System concept to allow for boundaries of information summarization, as well as retaining the approach of associating each prefix advertisement with an originating AS. The concept of policy opaqueness would also be retained in such an approach, implying that each AS accepts a set of route advertisements, applies local policy constraints, and re-advertises those advertisements permitted by the local policy constraints. It could be feasible to consider alterations to the distance vector path selection algorithm, particularly as it relates to intermediate states during processing of a route withdrawal. It is also feasible to consider the use of compound route attributes, allowing a route object to include an

aggregate route, and a number of specifics of the aggregate route, and attach attributes that may apply to the aggregate or a specific address prefix. Such route attributes could be used to support multi-homing and inter-domain traffic engineering mechanisms. The overall intent of this approach is to address the major requirements in the inter-domain routing space without using an increasing set of globally propagated specific route objects.

A potential applied research topic is to consider the feasibility of de-coupling the requirements of inter-domain connectivity management with the applications of policy constraints and the issues of sender- and/or receiver-managed traffic engineering requirements. Such an approach may use a link-state protocol as a means of maintaining a consistent view of the topology of inter-domain network, and then use some form of overlay protocol to negotiate policy requirements of each AS, and use a further overlay to support inter-domain traffic engineering requirements. The underlying assumption of such an approach is that by dividing up the functional role of inter-domain routing into distinct components each component will have superior scaling and convergence properties which in turn to result in superior properties for the entire routing system. Obviously, this assumption requires some testing.

Research topics with potential longer term application include the approach of drawing a distinction between a network's identity, a network's location relative to other networks, and a feasible path between a source and destination network that satisfies various policy and traffic engineering constraints. Again the intent of such an approach would be to divide the current routing function into a number of distinct scalable components.

9. Security Considerations

Any adopted inter-domain routing protocol needs to be secure against disruption. Disruption comes from two primary sources:

- Accidental misconfiguration
- Malicious attacks

Given past experience with routing protocols, both can be significant sources of harm.

Given that it is not reasonable to guarantee the security of all the routers involved in the global Internet inter-domain routing system, there is also every reason to believe that malicious attacks may come from peer routers, in addition to coming from external sources.

A protocol design should therefore consider how to minimize the damage to the overall routing computation that can be caused by a single or small set of misbehaving routers.

The routing system itself needs to be resilient against accidental or malicious advertisements of a route object by a route server not entitled to generate such an advertisement. This implies several things, including the need for cryptographic validation of announcements, cryptographic protection of various critical routing messages and an accurate and trusted database of routing assignments via which authorization can be checked.

10. References

[1] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

[2] Clark, D., Chapin, L., Cerf, V., Braden, R. and R. Hobby, "Towards the Future Internet Architecture", RFC 1287, December 1991.

[3] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification, RFC 2460, December 1998.

[4] Srisuresh, P. and K. Egevang, "Traditional IP Network Address Translator (Traditional NAT)", RFC 3022, January 2001.

[5] Fuller, V., Li, T., Yu, J. and K. Varadhan, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", RFC 1519, September 1993.

[6] Huston, G., "The BGP Routing Table", The Internet Protocol Journal, vol. 4, No. 1, March 2001.

[7] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[8] Vohara, Q. and E. Chen, "BGP support for four-octet AS number space", Work in Progress.

[9] Hain, T., "Architectural Implications of NAT", RFC 2993, November 2000.

[10] Labovitz, C., Ahuja, A., Bose, A. and J. Jahanian, "Delayed Internet Routing Convergence", Proceedings ACM SIGCOMM 2000, August 2000.

[11] Lothberg, P., personal communication, December 2000.

11. Acknowledgements

This document is the outcome of a collaborative effort of the IAB, and the editor acknowledges the contributions of the members of the IAB in the preparation of the document. The contributions of John Leslie, Thomas Narten and Abha Ahuja in reviewing this document are also acknowledged.

12. Author

Internet Architecture Board
Email: iab@ietf.org

Geoff Huston
Telstra
5/490 Northbourne Ave
Dickson ACT 2602
Australia

EMail: gih@telstra.net

13. Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.