Analyzing the Internet's BGP Routing Table

July 2003

Geoff Huston

The Internet continues along a path of seemingly inexorable growth, at a rate that has, at a minimum, doubled in size each year. How big it needs to be to meet future demands remains an area of somewhat vague speculation. Of more direct interest is the question of whether the basic elements of the Internet can be extended to meet such levels of future demand, whatever they may be. To rephrase this question, are there inherent limitations in the technology of the Internet- or its architecture of deployment-that may impact the continued growth of the Internet to meet ever-expanding levels of demand?

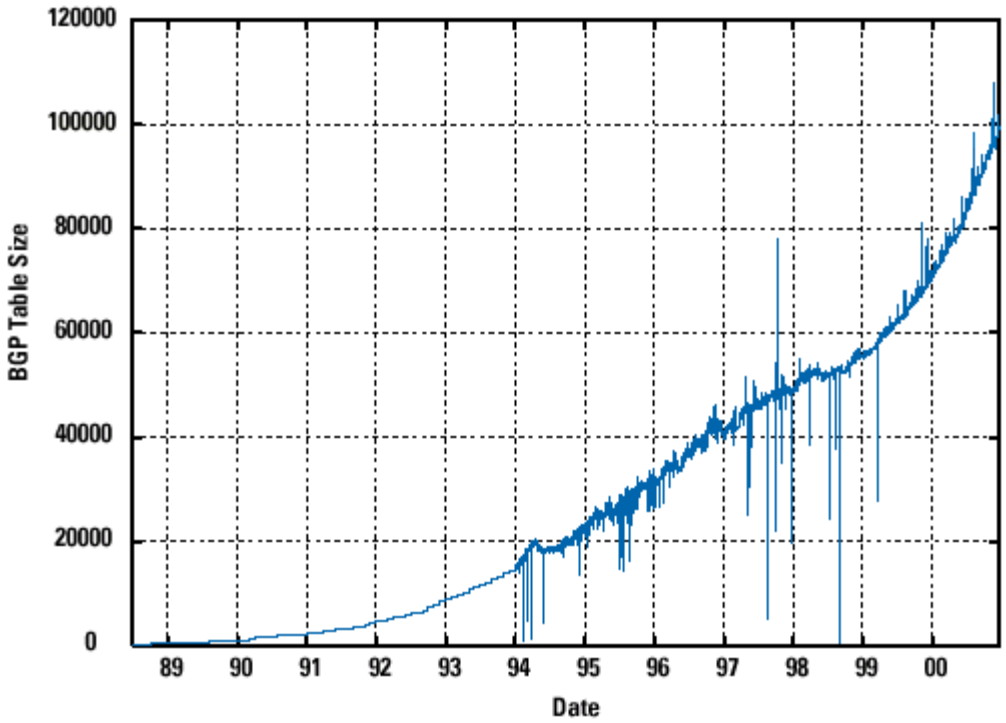
Numerous potential areas can be searched for such limitations, including the capacity of transmission systems, the switching capacity of routers, the continued availability of addresses, and the capability of the routing system to produce a stable view of the overall topology of the network. This article examines the Internet routing system and the longer-term growth trends that are visible within this system.

The structure of the global Internet can be likened to a loose coalition of semi-autonomous constituent networks. Each of these networks operates with its own policies, prices, services, and customers. Each network makes independent decisions about where and how to secure the supply of various components that are needed to create the network service. The cement that binds these networks into a cohesive whole is the use of a common address space and a common view of routing. Integrity of routing within each constituent network, or *Autonomous System* (AS), is maintained through the use of an interior routing protocol (or *Interior Gateway Protocol* , or IGP). The collection of these networks is joined into one large routing domain through the use of an inter-network routing protocol (or *Exterior Gateway Protocol* , or EGP).

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 [1] . As more devices connect to the Internet, they consume addresses, and the associated function of maintaining reachability information for these addresses implies ever-larger routing tables. The work in studying the limitations of the 32-bit IPv4 address space produced many outcomes, including the specification of IPv6, as well as the refinement of techniques of *Network Address Translation* (NAT) 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, because the routing space is the cross product of the complexity of the topology of the network, multiplied by the number of autonomous domains of connectivity policy multiplied by the base size of a routing-table entry. When a network 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 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 *Border Gateway Protocol* (BGP) routing table that is visible at that point.

Measurements of the size of the routing table were somewhat sporadic in the beginning, and many measurements were taken at approximately monthly intervals from 1988 until 1992 at Merit [2] . 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 at the start of that year. This measurement technique was adopted by the author in 1997, using a measurement point located at the edge of AS 1221 in Australia, again using an hourly interval for the measurement [6] . The result of these efforts is that we now have a detailed view of the dynamics of the Internet routing-table growth that spans 13 years (Figure 1).

Figure 1: BGP Table Growth 1988-2000



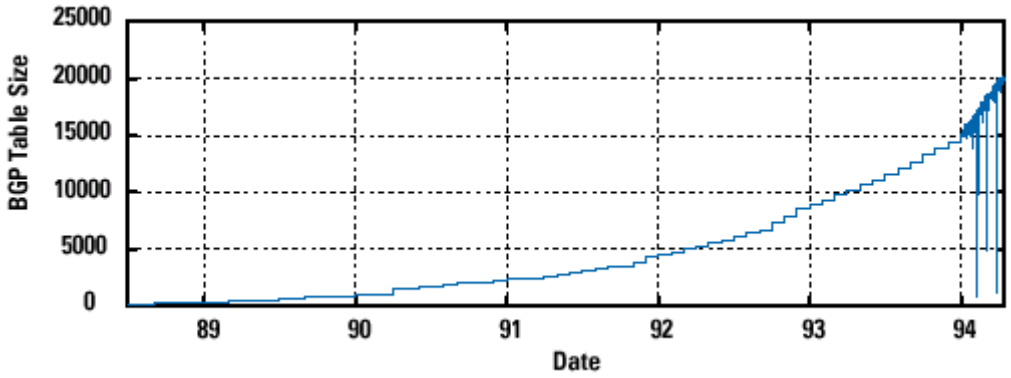
BGP Table Growth

At a gross level, there appear to be four distinct phases of growth visible in this data.

Pre-CIDR Growth

The initial characteristics of the routing-table size from 1988 until April 1994 show definite characteristics of exponential growth (Figure 2). Much of this growth can be attributed to the growth in deployment of the historical Class C address space (/24 address prefixes). Unchecked, this growth would have lead to saturation of the BGP routing tables in nondefault routers within a few years. Estimates of the time at which this would have happened vary somewhat, but the overall observation was that the growth rates were exceeding the growth in hardware and software capability of the deployed network at that time.

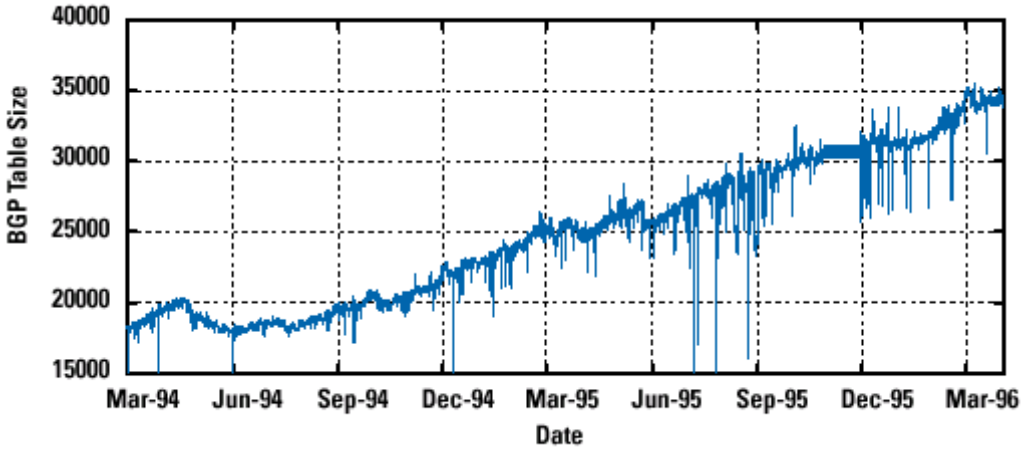
Figure 2: BGP Table Growth 1988-1994



CIDR Deployment

The response from the engineering community was the introduction of routing software 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. A con certed effort was undertaken in 1994 and 1995 to deploy *Classless Interdomain Routing* (CIDR), based on encouraging deployment of the CIDR-capable version of the BGP protocol, BGP4. The effects of this ef fort are visible in the routing table (Figure 3). Interestingly enough, the efforts of the *Internet Engineering Task Force* (IETF) CIDR Deploy ment Working Group are visible in the table, with downward movements in the size of the routing table following each IETF meeting.

Figure 3: BGP Table Growth 1994-1995

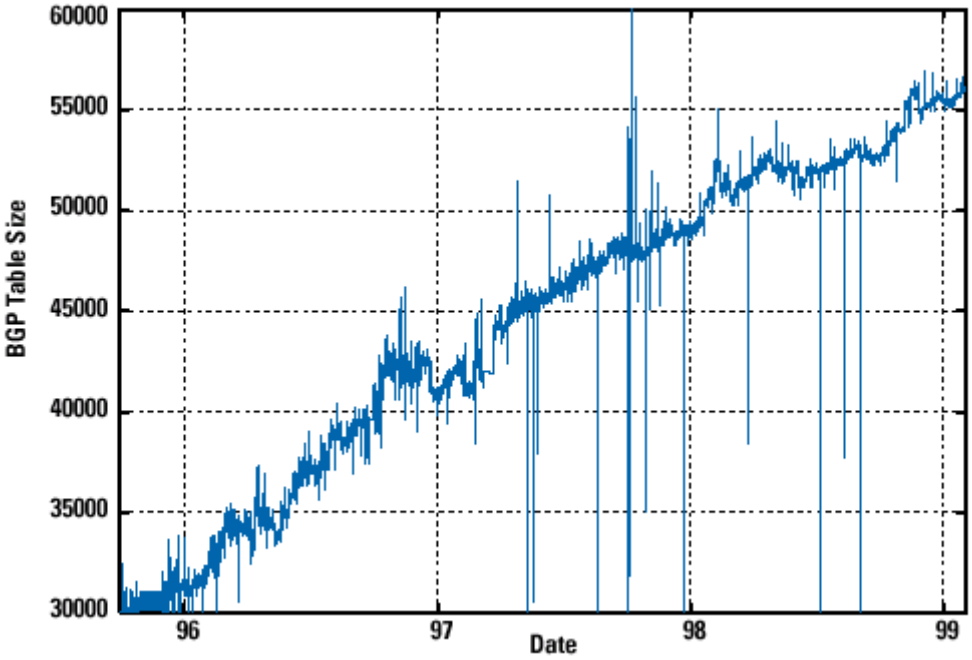


The intention of CIDR was one of supporting an address architecture termed "provider address aggregation," where a network provider is al located an address block from the address registry, and announces this entire block into the exterior routing domain. Customers of the pro- vider use a suballocation from this address block, and these smaller routing elements are 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 approximately 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.

CIDR Growth

For the next four years until the start of 1998, CIDR proved remarkably effective in damping unconstrained growth in the BGP routing table. While other metrics of Internet size grew exponentially during this period, the BGP table grew at a linear rate, adding about 10,000 entries per year. (Figure 4). Growth in 1997 and 1998 was even lower than this linear rate. Although the reasons behind this are somewhat speculative, it is relevant to note that this period saw intense aggregation within the *Internet Service Provider* (ISP) industry, and in many cases this aggregation was accompanied by large-scale renumbering to fit within provider-based aggregated address blocks. During this period, credit for this trend also must be given to Tony Bates, whose weekly reports of the state of the BGP address table, including listings of further potential for route aggregation, provided considerable incentive to many providers to improve their levels of route aggregation [4].

Figure 4: BGP Table Growth 1995-1998



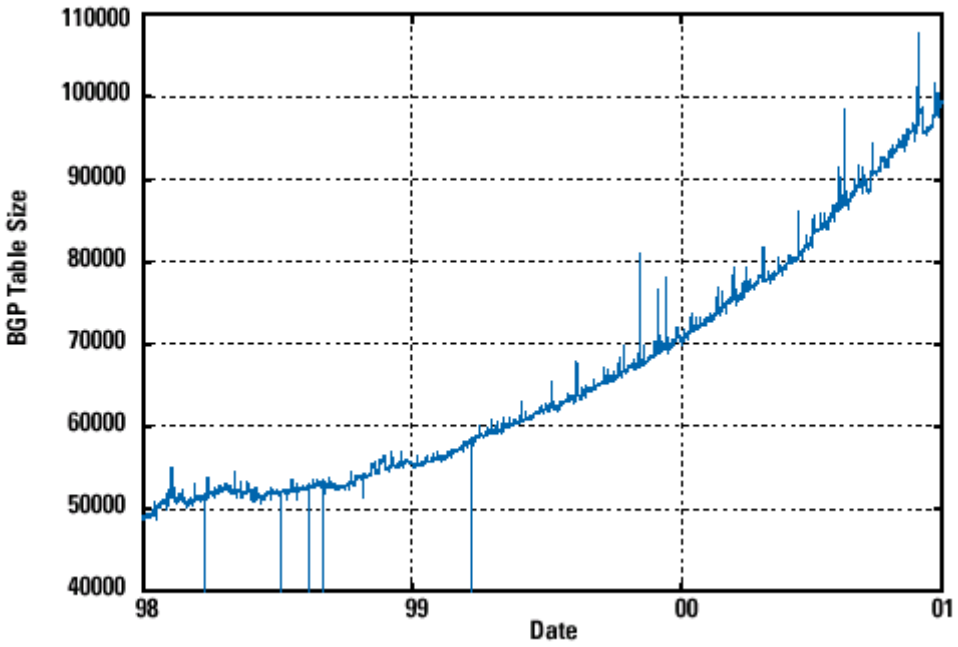
A close examination of the table reveals a greater level of stability in the routing system at this time. The short-term (hourly) variation in the number of announced routes decreased, both as a percentage of the number of announced routes and in absolute terms. One of the other benefits of using large aggregate address blocks is that an 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 at which an aggregate route is used in place of a collection of more specific routes. This, coupled with widespread adoption of BGP route flap damping, has been every effective in reducing the short-term instability in the routing space. It has been observed that whereas the absolute size of the BGP routing table is one factor in scaling, another is the processing load imposed by continually updating the routing table in response to individual route withdrawals and announcements. The encouraging picture from this table is that the levels of such dynamic instability in the network have been reduced considerably by a combination of route flap damping and CIDR.

Current Growth

In late 1998, the trend of growth in the BGP table size changed radi cally, and the growth for the past two years is again showing all the signs of a reestablishment of exponential growth. It appears that CIDR has been unable to keep pace with the levels of growth of the Internet.

(Figure 5). Once again the concern is that this level of growth, if sus tained, will outstrip the capability of hardware, or current capability of the BGP routing protocol, or possibly both.

Figure 5: BGP Table Growth 1998-2000



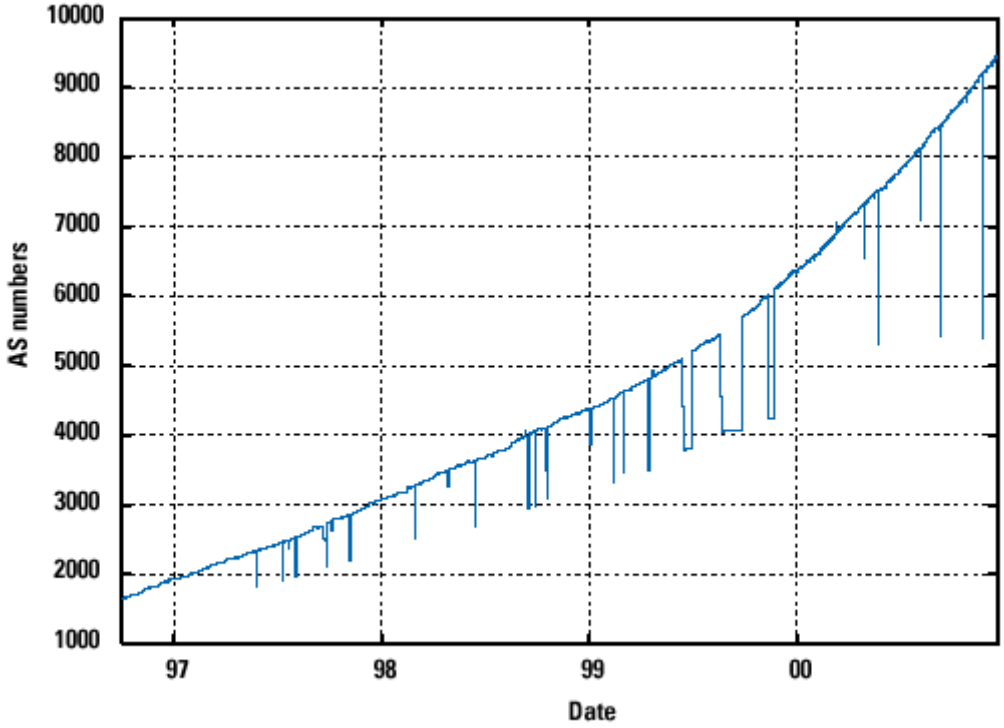
Related Measurements Derived from BGP Table

The level of analysis of the BGP routing table has been extended in an effort to identify the reasons for this resumption of exponential growth. Current analysis includes measuring the number of ASs in the routing system, and the number of distinct AS paths, the range of addresses spanned by the table, and the average span of each routing entry.

AS Number Consumption

Each network that is multihomed within the topology of the Internet and wishes to express a distinct external routing policy must use an AS to associate its advertised addresses with such a policy. In general, each network is associated with a single AS, and the number of ASs in the default-free routing table tracks the number of entities that have unique routing policies. There are some exceptions to this, including large glo bal transit providers with varying regional policies, where multiple ASs are associated with a single network, but such exceptions are relatively uncommon. The trend of AS number deployment over the past four years is also exponential (Figure 6). The growth in the number of ASs can be correlated with the growth in the amount of address space spanned by the BGP routing table. At the end of 2000, the span of ad vertised addresses is growing at an annual rate of 7 percent, while the number of ASs is growing by 51 percent. Each AS is, on average advertising smaller address ranges. This points to increasingly finer levels of routing detail being announced into the global routing domain, a trend that causes some level of concern.

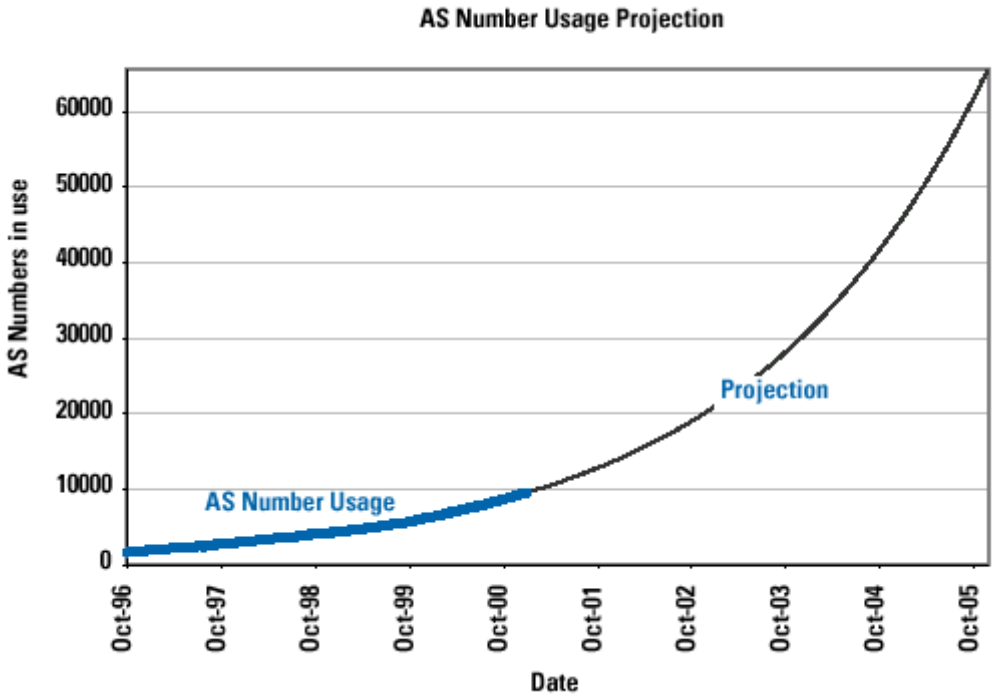
Figure 6: AS Number Deployment



This is a likely result of an increasingly dense interconnection mesh, where an increasing number of networks are moving from a single-homed connection into multihoming and peering. The spur for this may well be the declining unit costs of communications bearer services.

If this rate of growth continues, the 16-bit AS number set will be ex-hausted by late 2005 (Figure 7). Work is under way within the IETF to modify the BGP protocol to carry AS numbers in a 32-bit field [5] . Al-though 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.

Figure 7: AS Number Projections

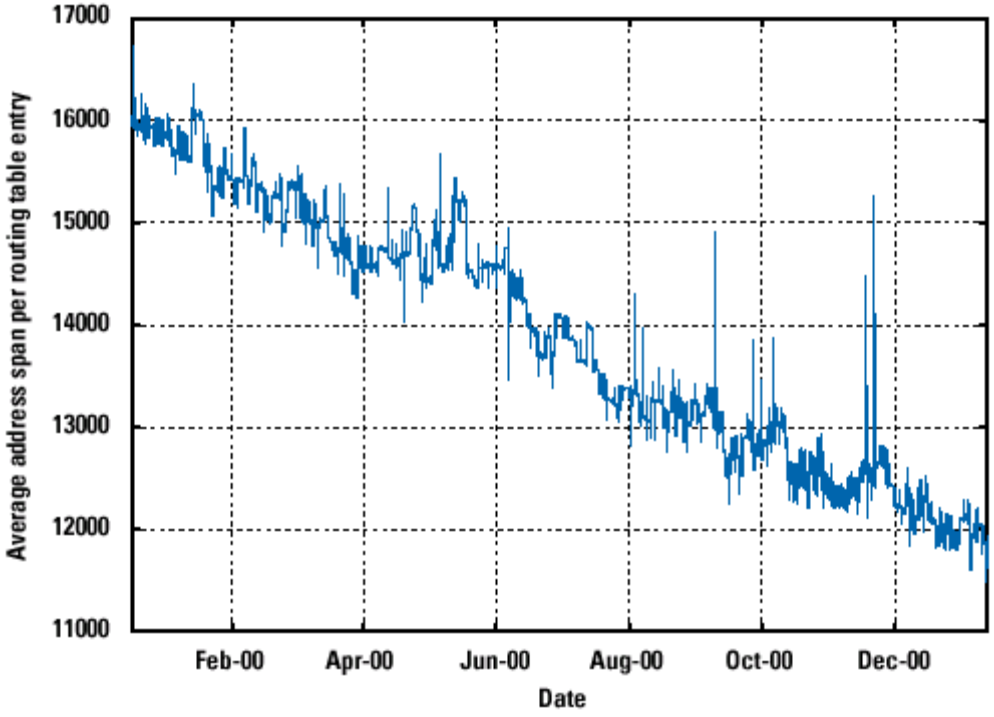


Average Prefix Length of Advertisements

The intent of CIDR aggregation was to support the use of large aggre gate address announcements in the BGP routing table. To check whether this is still the case, researchers have tracked the average span of each BGP announcement for the past 12 months. The data indicates a de cline in the average span of a BGP advertisement from 16,000 individual addresses in November 1999 to 12,100 in December 2000 (Figure 8). This corresponds to an increase in the average prefix length from /18.03 to /18.44. Separate observations of the average prefix length used to route traffic in operation networks in late 2000 indicate an average length of 18.1 [8] . Again, this trend is cause for concern because it implies the increasing spread of traffic over greater numbers of increasingly finer forwarding-table entries. This, in turn, has implications for the design of high-speed core routers, particularly when extensive use is made of cached forwarding entries within the switching subsystem.

One potential scenario is that the size of the advertisement continues to decrease. With the widespread use of address translation gateway sys tems, such as NAT, and the continued concern over the finite nature of the IPv4 address pool, this is certainly a highly likely scenario. Projec tions of the average prefix length of advertisements using current trends in the number of BGP table entries and the total address span adver tised in the BGP table indicate a lengthening of the average prefix length of advertisements by 1 bit length every 29 months. This has implica tions in the lookup algorithms used in routing design, depending on the space/time trade-offs used in the lookup algorithm design. This trend implies that either lookups need to search deeper through the prefix chain to find the necessary forwarding entry, requiring faster memory subsystems to perform each lookup, or the lookup table needs to be both larger and more sparsely populated, increasing the requirements for high-speed memory within the router forwarding subsystem.

Figure 8: Average Span of BGP Advertisement



Prefix Length Distribution

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

An extensive effort was introduced in the mid-1990s to move away from extensive use of the Class C space and to encourage providers to advertise larger address blocks. This has been reinforced by the address registries who have used provider allocation blocks of /19 and, more re-cently, /20. These measures were introduced when there were approxi-mately 20,000 to 30,000 entries in the BGP table. It is interesting to note that five years later, of the 96,000 entries in the routing table, about 53,000 entries have a /24 prefix. In absolute terms, the /24 prefix set is the fastest-growing prefix set in the entire BGP table.

The routing entries of these smaller address blocks also show a much higher level of change on an hourly basis. Although a large number of BGP routing points perform route flap damping, 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 number of these small prefixes is 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 recomputations that occur in the wake of announcements and withdrawals. Withdrawals are of particular concern because of 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 about 2 minutes to propagate a route withdrawal across the BGP domain [7] . 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 about 4000 entries; in relative terms, this is one of the fastest-growing prefix sets. The number of entries corresponding to very small address blocks (smaller than a /24), although small in number as a proportion of the to tal 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 an nouncing a very small address block with a distinct routing policy would have no particular beneficial outcome, because 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 lim ited 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 announcement of very small address blocks is likely to continue.

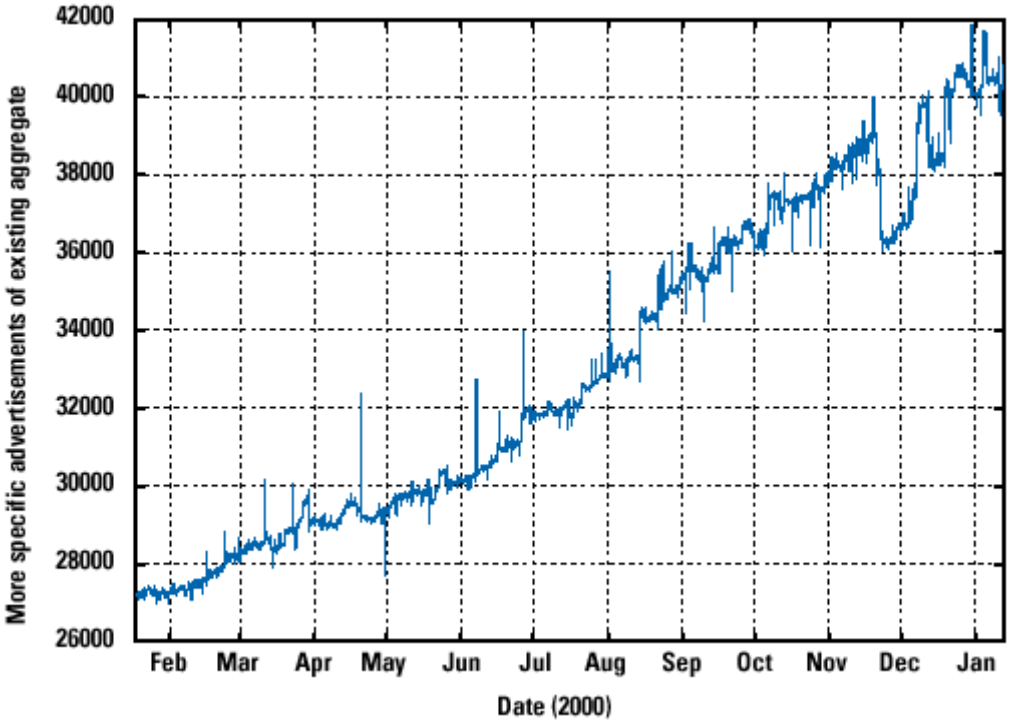
Aggregation and Holes

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 not to promulgate a different connectivity policy, but to perform some rudimentary form of load balancing and mutual backup for multihomed 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 pol icy 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.

Although it is difficult to understand whether the use of such specific advertisements was intended to be an exception to a more general rule or that it was not intended to be within the original intent of CIDR de ployment, there appears to be very widespread use of this mechanism within the routing table. Approximately 37,500 advertisements, or 37 percent of the routing table, is being used to punch policy holes in exist ing aggregate announcements (Figure 9). Of these, the overall majority of about 30,000 routes use distinct AS paths, so that once more we are seeing a consequence of finer levels of granularity of connection policy in a densely interconnected space.

Although 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.

Figure 9: More Specific Advertisements

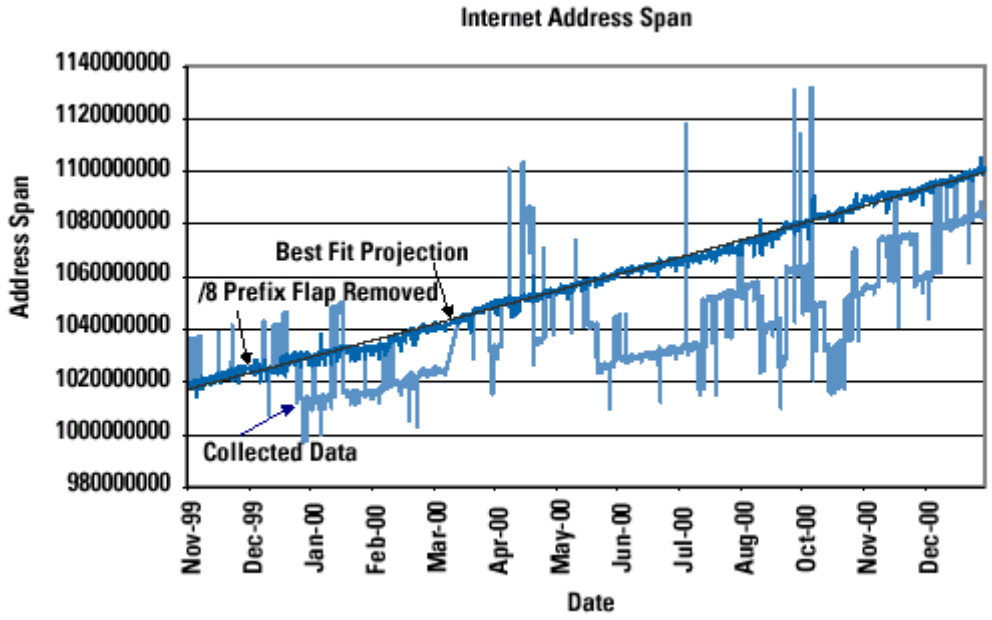


Address Consumption

A decade ago there were two major concerns over scaling of the Internet, and of the two, the consumption of address space was considered to be the more immediate and compelling threat to the continued viability of the network to sustain growth.

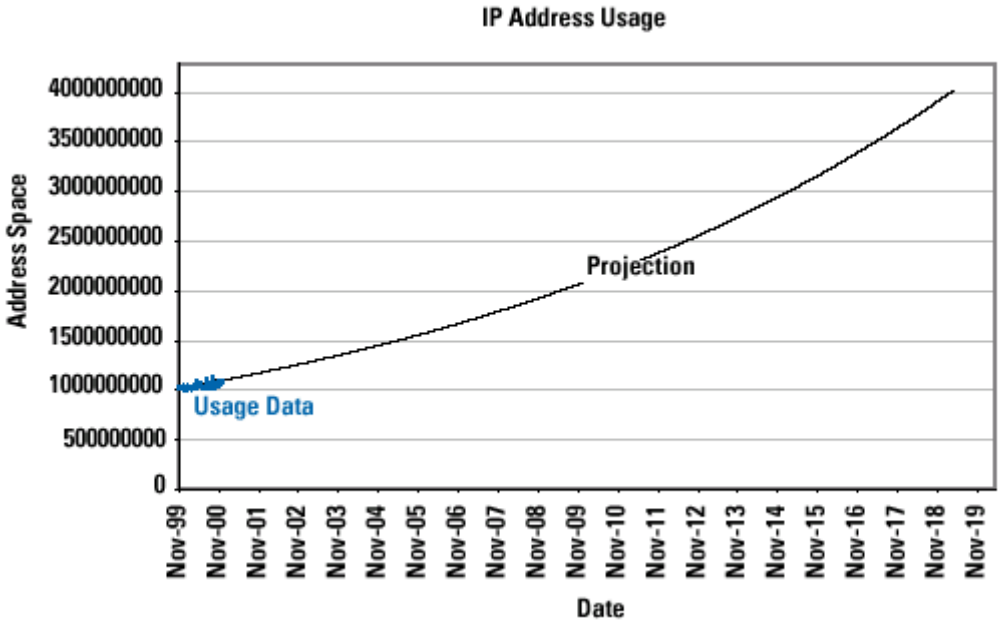
Within the scope of this exercise, it has been possible to track the total span of address space covered by BGP routing advertisements. Over the period from November 1999 until December 2000, the span of address space has grown from 1.02 billion addresses to 1.06 billion. However, numerous /8 prefixes are periodically announced and withdrawn from the BGP table, and if the effects of these prefixes are removed, the final value of addresses spanned by the table is approximately 1.09 billion addresses (Figure 10).

Figure 10: Total Address Space



This is an annual growth rate of a little less than 7 percent, and at that rate of address deployment, the IP Version 4 address space will be able to support another 19 years of such growth (Figure 11). Compared to the 42-percent growth in the number of routing advertisements, it would appear that much of the growth of the Internet in terms of growth in the number of connected devices is occurring behind various forms of NATs. In terms of solving the perceived finite nature of the ad dress space identified just under a decade ago, the Internet appears so far to have embraced the approach of using NATs, irrespective of their various perceived functional shortcomings [3] . This observation also sup ports the observed increase of smaller address fragments supporting distinct policies in the BGP table, because such small address blocks en compass arbitrarily large networks located behind one or more NAT gateways.

Figure 11: Address Space Projection



Anomalies

A common space such as the inter-provider domain is not actively man aged by any single entity, and various anomalies appear in the routing table from time to time. One notable event occurred in late 1997, when some large prefixes were deconstructed into a massive set of /24 prefixes and this set was inad vertently passed into the inter-provider BGP domain. The BGP table graphs show a sudden upswing in the number of routing table entries from 50,000 entries to about 78,000 entries. It could have been higher, except that a commonly used routing hardware platform at the time ran into table memory exhaustion at that number of table entries, and fur ther promulgation of additional routing entries ceased. Numerous other anomalies also exist in the table, including the presence of a /31 prefix and several hundred /32 prefixes.

Although many of these anomalies can be attributed to configuration errors of various forms, the underlying observation is that there are no universally used strong filters on what can broadcast into the BGP routing space. Considering the distributed nature of this table and the critical role that it plays in supporting the global Internet, this can be considered a significant current vulnerability. One potential response is to make more use of authentication measures. A validity check could be a precondition to accepting any route advertisement, allowing the receiver of the advertisement a means to check that the origin AS intended to advertise this route. This would create greater resiliency against inadvertent leaks of large sets of advertisements into the broader interdomain space. It would also improve the resiliency of the BGP domain against some forms of deliberate attack.

Conclusions

There are strong parallels between the BGP routing space and the condition commonly referred to as "The Tragedy Of The Commons." The BGP routing space is simultaneously everyone's problem, because it impacts the stability and viability of the entire Internet, and no one's problem, in that no single entity can be considered to manage this common resource.

In other common resource domains, when the value of the resource is placed under threat because of damaging exploitative practices, the most typical form of corrective action is through the imposition of a consistent set of policies and practices intended to achieve a particular outcome. The vehicle for such an imposition of policies and practices is most commonly that of

regulatory fiat. In a globally distributed space such as the BGP table, it is a challenging task to identify the source and authority of such potential regulatory activity.

Multihomed Small Networks

It would appear that one of the major drivers of the recent growth of the BGP table is that of small networks multihoming with numerous peers and numerous upstream providers. In the appropriate environment where numerous networks are in relatively close proximity, using peer relationships can reduce total connectivity costs, as compared to using a single upstream service provider. Equally significantly, multihoming with numerous upstream providers is seen as a means of improving the overall availability of the service. In essence, multihoming is seen as an acceptable substitute for upstream service resiliency.

This has a potential side effect: When multihoming 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 incentive to spend the additional money required to engineer resiliency into the network. The actions of the multihomed network clients then become self-fulfilling.

One way to characterize this behavior is that service resiliency in the In ternet 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.

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. The deployment model of a sin gle-homed network with a single upstream provider is rapidly being supplanted by a model of extensive interconnection at the edges of the Internet. The underlying deployment model assumed by CIDR as sumed a different structure, more akin to a strict hierarchy of supply providers. The business imperatives driving this denser mesh of inter connection in the Internet are irresistible, and the casualty in this case is the CIDR-induced dampened growth of the BGP routing table.

Traffic Engineering via Routing

Further driving this growth in the routing table is the use of selective ad vertisement of smaller prefixes along different paths in an effort to undertake traffic engineering within a multihomed environment. Al though considerable effort is being undertaken to develop traffic engineering tools within a single network using Multiprotocol Label Switching (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, further exacerbating the growth and stability pressures being placed on the BGP routing domain.

The effects of CIDR on the growth of the BGP table have been out standing, not only because of their initial impact in turning exponential growth into a linear growth trend, but also because CIDR was effective for far longer than could have been reasonably expected in hindsight. The current growth factors at play in the BGP table are not easily sus ceptible to another round of CIDR deployment pressure within the operator community. It may well be time to consider how to manage a BGP routing table that has millions of small entries, rather than the ex pectation of tens of thousands of larger entries.

We started this journey over ten years ago when considering the scaling properties of addressing and routing. It is perhaps fitting that we tie the two concepts back together again as we consider the future of the BGP inter-provider routing space. The observation that the BGP growth pressures are largely due to an uptake in multihoming and the associated advertisement of discrete connectivity policies by increasingly smaller networks at the edge of the network has a corollary for address allocation policy. In such a ubiquitous environment of multihomed networks, we will also need to review how address blocks are allocated to network providers, because the concept of provider-based address allocation that assumes a relatively strict hierarchical supply structure is becoming less and less relevant in today?s Internet.

References

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

[2] V. Fuller, T. Li, J. Yu, and K. Varadhan, "Supernetting: an Address Assignment and Aggregation Strategy," RFC 1338, June 1992.

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

[4] T. Bates, "The CIDR Report," updated weekly at: http://www.employees.org/~tbates/cidr-report.html

[5] E. Chen, Y. Rekhter, "BGP Support for Four-Octet AS Number Space," work in progress, currently published as an Internet Draft: draft-chen-as4bytes-00.txt, November 2000.

[6] "BGP Table Report" updated hourly at http://www.telstra.net/ops/bgp

[7] C. Labovitz, A. Ahuja, "The Impact of Internet Policy and Topology on Delayed Routing Convergence?Update to This Work," ISMA Winter 2000 Workshop, CAIDA, December 2000.

[8] Peter Lothberg, personal communication.

GEOFF HUSTON holds a B.Sc. and a M.Sc. from the Australian National University. He has been closely involved with the development of the Internet for the past decade, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. Huston is currently the Chief Scientist in the Internet area for Telstra. He is also a member of the Internet Architecture Board, and is the Secretary of the Internet Society Board of Trustees. He is author of The ISP Survival Guide, ISBN 0-471-31499-4, Internet Performance Survival Guide: QoS Strategies for Multiservice Networks, ISBN 0471-378089, and coauthor of Quality of Service: Delivering QoS on the Internet and in Corporate Networks, ISBN 0-471-24358- 2, a collaboration with Paul Ferguson. All three books are published by John Wiley & Sons. E-mail: gih@telstra.net