Scaling Inter-Domain Routing—A View Forward

December 2001
Geoff Huston

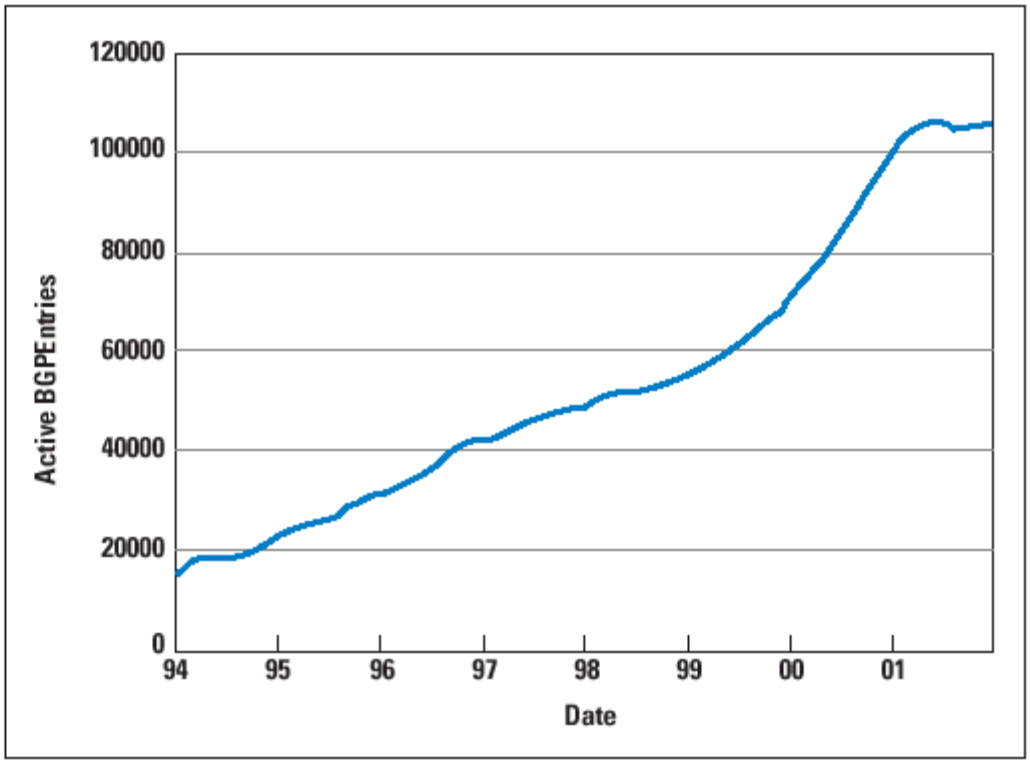
In the previous IPJ article, "Analyzing the Internet BGP Routing Table," (Vol. 4, No. 1, March 2001) we looked at the characteristics of the growth of the routing table in recent years. The motivation for this work is to observe aspects of the Internet routing table in order to understand the evolving structure of the Internet and thereby attempt to predict some future requirements for routing technology for the Internet.

The conclusions drawn in the previous article included the observation that multihomed small networks appeared to be a major contributor to growth of the Internet routing system. It also observed that there was a trend toward a denser mesh of inter-Autonomous System connectivity within the Internet. At the same time there has been an increase of various forms of policy-based constraints imposed upon this connectivity mesh, probably associated with a desire to undertake various forms of inter-domain traffic engineering through manipulation of the flow of routing information.

Taken together, these observations indicate that numerous strong growth pressures are being exerted simultaneously on the inter-domain routing space. Not only is the network itself growing in size, but also the internal interconnectivity of the network is becoming more densely meshed. The routing systems that are used to maintain a description of the network connectivity are being confronted with having to manipulate smaller route objects that describe finer levels of network detail. This is coupled with lengthening lists of qualifying attributes that are associated with each route object. The question naturally arises as to whether the ***Border Gateway Protocol*** (BGP) and the platforms used to support BGP in the Internet today can continue to scale at a pace that matches the growth in demands that are being placed upon it.

The encouraging news is that there appears to be no immediate cause for concern regarding the capability of BGP to continue to support the load of routing the Internet. The processor and memory capacity in current router platforms is easily capable of supporting the load associated with various forms of operational deployment models, and the protocol itself is not in imminent danger of causing network failure through any internal limitation within the protocol itself. Also, numerous network operators have exercised a higher level of care as to how advertisements are passed into the Internet domain space and, as a result, the growth rates for the routing table over 2001 shows a significant slowdown over the rates of the previous two years (Figure 1).

Figure 1: BGP Table Size 1994–2001



However, the observed trends in inter-domain routing of an increasingly detailed and highly qualified view of a more densely interconnected and still-growing network provide adequate grounds to examine the longer term routing requirements. It is useful, therefore, to pose the question as to whether we can continue to make incremental changes to the BGP protocol and routing platforms, or whether the pace of growth will, at some point in time, mandate the adoption of a routing architecture that is better attuned to the evolving requirements of the Internet.

This article does not describe the operation of an existing protocol, nor does it describe any current operational practice. Instead it examines those aspects of inter-domain routing that are essential to today's Internet, and the approaches that may be of value when considering the evolution of the Internet inter-domain routing architecture. With this approach, the article illustrates one of the initial phases in any technology development effort; that of an examination of various requirements that could or should be addressed by the technology.

Attributes of an Inter-Domain Routing Architecture

Let's start by looking at those aspects of the inter-domain routing environment that could be considered a base set of attributes for any inter-domain routing protocol.

Accuracy

For a routing system to be of any value, it should accurately reflect the forwarding state of the network. Every routing point is required to have a consistent view of the routing system in order to avoid forwarding loops and black holes (points where there is no relevant forwarding information and the packet must be discarded). Local changes in underlying physical network, or changes in the policy configuration of the network at any point, should cause the routing system to compute a new distributed routing state that accurately reflects the changes.

This requirement for accuracy and consistency is not, strictly speaking, a requirement that every node in a routing system has global knowledge, nor a requirement that all nodes have

precisely the same scope of information. In other words, a routing system that detects and avoids routing loops and inconsistent black holes does not necessarily need to use routing systems that rely on uniform distribution of global knowledge frameworks.

Scalability

Scalability can be expressed in many ways, including the number of routing entries, or prefixes, carried within the protocol, the number of discrete routing entities within the inter-domain routing space, the number of discrete connectivity policies associated with these routing entries, and the number of protocols supported by the protocol. Scalability also needs to encompass the dynamic nature of the network, including the number of routing updates per unit of time, time to converge to a coherent view of the connectivity of the network following changes, and the time taken for updates to routing information to be incorporated into the network forwarding state. In expressing this ongoing requirement for scalability in the routing architecture, there is an assumption that we will continue to see an Internet that is composed of a large number of providers, and that these providers will continue to increase the density of their interconnection.

The growth trends in the inter-domain routing space do not appear to have well-defined upper limits, so placing bounds on various aspects of the routing environment is impractical. The only practical way to describe this attribute is that it is essential to use a routing architecture that is scalable to a level well beyond the metrics of today's Internet.

In the absence of specific upper bounds to quantify this family of requirements, the best we conclude here is that at present we are working in an inter-domain environment that manipulates some 10^5 distinct routing entries, and at any single point of interconnection there may be of the order of 10^6 routing protocol elements being passed between routing domains. Experience in scaling transmission systems for the Internet indicates that an improvement of a single order of magnitude in the capacity of a technology has a relatively short useful lifetime. It would, therefore, be reasonable to consider that a useful attribute is to be able to operate in an environment that is between two to three orders of magnitude larger than today's system.

Policy Expressiveness

Routing protocols perform two basic tasks: first, determining if there is at least one viable path between one point in the network and another, and secondly, where there is more than one such path, determining the "best" such path to use. In the case of interior routing protocols, "best" is determined by the use of administratively assigned per-link metrics, and a "best" path is one that minimizes the sum of these link metrics.

In the case of the inter-domain routing protocols, no such uniformly interpreted metric exists, and "best" is expressed as a preference using network paths that yield an optimal price and performance outcome for each domain.

The underlying issue here is that the inter-domain routing system must straddle a collection of heterogeneous networks, and each network has a unique set of objectives and constraints that reflect the ingress, egress, and transit routing policies of a network. Ingress routing policies reflect how a network learns information, and which learned routes have precedence when selecting a routing entry from a set of equivalent routes. In a unicast environment, exercising control over how routes are learned by a domain has a direct influence over which paths are taken by traffic leaving the domain. Egress policies reflect how a domain announces routes to its adjacent neighbours. A domain may, for example, wish to announce a preferential route to a particular neighbour, or indicate a preference that the route not be forwarded beyond the adjacent neighbour. In a unicast environment, egress routing policies have a bearing on which paths are used for traffic to reach the domain. Transit routing policies control how the routes learned from an adjacent domain are advertised to other adjacent domains. If a domain is a transit provider for another domain, then a typical scenario for the transit provider would be to

announce all learned routes to all other connected domains. For a multi-homed transit customer, routes learned from one transit provider would normally not be announced to any other transit provider.

This requirement for policy expressiveness implies that the inter-domain routing protocol should be able to attach various attributes to protocol objects, allowing a domain to communicate its preferences relating to handling of the route object to remote domains.

Robust Predictable Operational Characteristics

A routing system should operate in such a way that it achieves predictable outcomes. The inference here is that under identical initial conditions a routing system should always converge to the same routing state, and that with knowledge of the rules of operation of the protocol and the characteristics of the initial environment, an observer can predict what this state will be. Predictability also implies stability of the routing environment, such that a routing state should remain constant for as long as the environment itself remains constant.

The routing protocol should operate in a way that tends to damp propagation of dynamic changes to the routing system rather than amplify such changes. This implies that minor variations in the state of the network 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 behaviour of the system.

The routing system should have robust convergence properties. A change in the physical configuration or policy environment in any part of the network causes a distributed computation of the routing state. Convergence implies that this distributed computation reaches a conclusion at some point. The requirement for a robust convergence property implies that the distributed computation should always halt, that the halting point be reached quickly, and the system should avoid generating transitory incorrect intermediate routing states. The interpretation of "quickly" in this context is variable. Currently, this value for BGP convergence time is of the order of tens to hundreds of seconds. In order to support increasingly time-critical applications, there appears to be an emerging requirement to reduce the median convergence time for the inter-domain routing protocol to a small number of seconds.

Efficiency

The routing system should be efficient, in that the amount of network resources, in terms of bandwidth and processing capacity of the network switching elements, should not be disproportionately large. This is an area of trade-off in that 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 continuously maintain an accurate view of the connectivity of the network, but at a cost of higher overhead. It is necessary to pass enough information across the system to allow each routing element to have a sufficiently accurate view of the network, yet ensure that the total routing overhead is low.

Evolving Requirements of Inter-Domain Routing

Layered on top of the base set of routing requirements listed above are a second set of requirements that can be seen as reflecting current directions in the deployed Internet, and are not necessarily well integrated into the existing routing architecture.

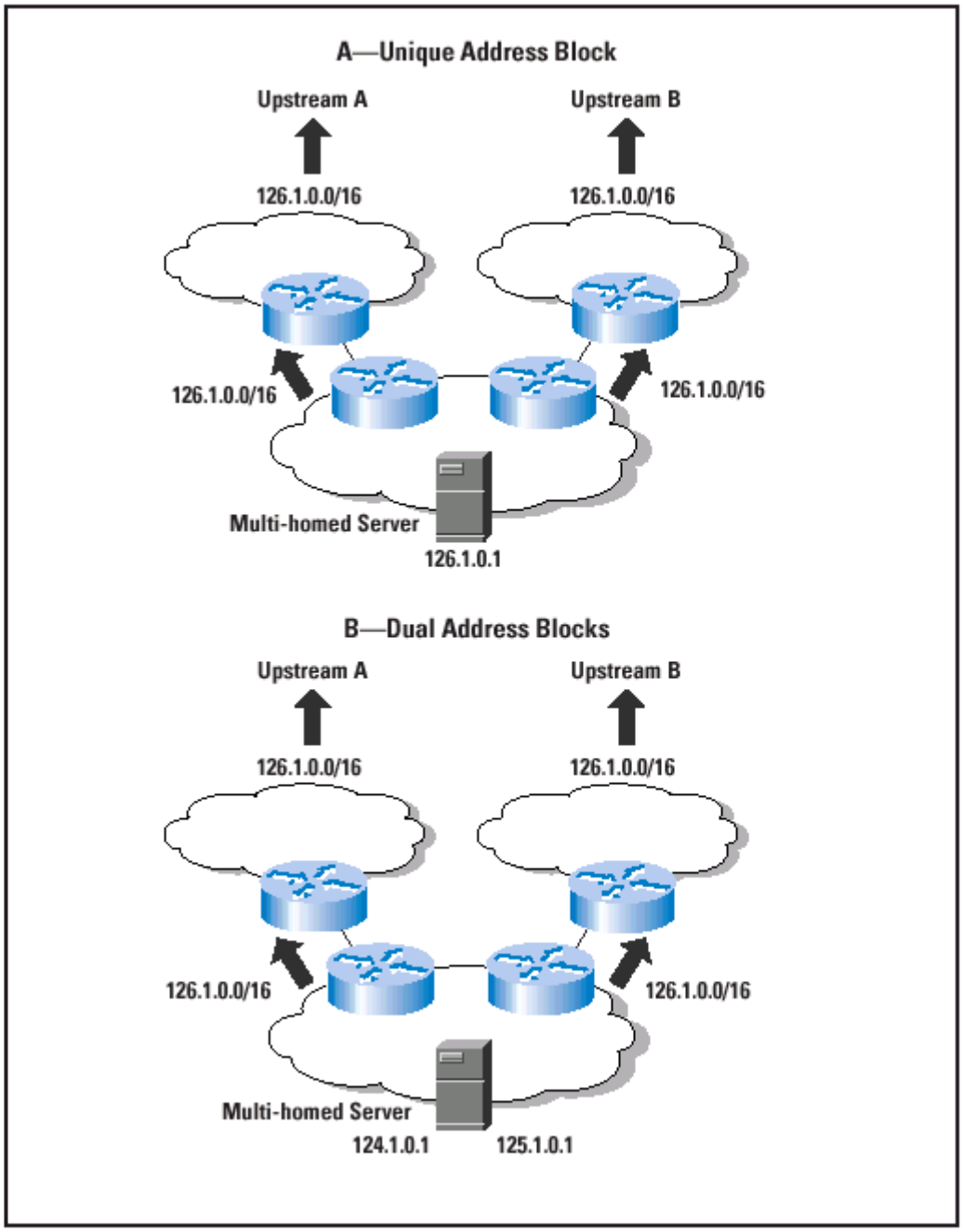
Multi-Homing of Edge Networks

Multi-homing refers to the practice of using more than one upstream transit provider. The common motivation for such a configuration is that if service from one transit provider fails, the customer can use the other provider as a means of service restoration. It may also allow some form of traffic balancing across multiple services. With careful use of route policies, the customer can direct traffic to each provider to minimize delay and loss, achieving some improved application performance.

The issue presented by multi-homing is that the multi-homed network is now not wholly contained within a service hierarchy of any particular provider. This implies that routing information describing reachability to the multi-homed customer cannot readily be aggregated into any single provider's routing advertisements, and the usual outcome is that the multi-homed customer must independently announce its reachability to each transit provider, who in turn must propagate this information across the routing system.

The evolving requirement here is one that must be able to integrate the demands of an increasing use of multi-homing into the overall network design. Two basic forms of approach can be used here—one is to use a single address block across the customer network and announce this block to all transit providers as an unaggregatable routing advertisement into the inter-domain routing system, and the other is to use multiple address blocks drawn from each provider's address block, and use either host-based software or some form of dynamic address translation within the network in order to use a source address drawn from a particular provider's block for each network transaction (Figure 2). The second approach is not widely used, and for the immediate future the requirement for multi-homing is normally addressed by using unique address blocks for the multi-homed network that are not part of any provider's aggregated address blocks. The consequence of this is that widespread use of multi-homing as a means of service resiliency will continue to have an impact on the inter-domain routing system.

Figure 2: Routing Approaches to Multi-Homing



Inter-Domain Traffic Engineering

In an increasingly densely interconnected network, selecting and using just one path between two points is not an optimal outcome of a routing architecture. Of more importance is the ability to identify a larger set of viable paths between these points and distribute the associated traffic flows in such a way that each individual transaction uses a single path, but the total set of flows is distributed across the set of paths.

To achieve this outcome, more information must be placed into the routing system, allowing a route originator to describe the policy-based preferences of which sets of paths should be preferred for traffic destined to the route originator, allowing a transit service operator to add information regarding current preferences associated with using particular transit paths, and allowing the traffic originator the ability to use local traffic egress policies to reach the destination. These traffic engineering-related preferences are not necessarily represented by

static values of routing attributes. One of the requirements of traffic engineering is to allow the network to dynamically respond to shifting traffic load patterns, and this implies that there is a component of dynamic information update that is associated with such traffic engineering-related aspects of the routing system.

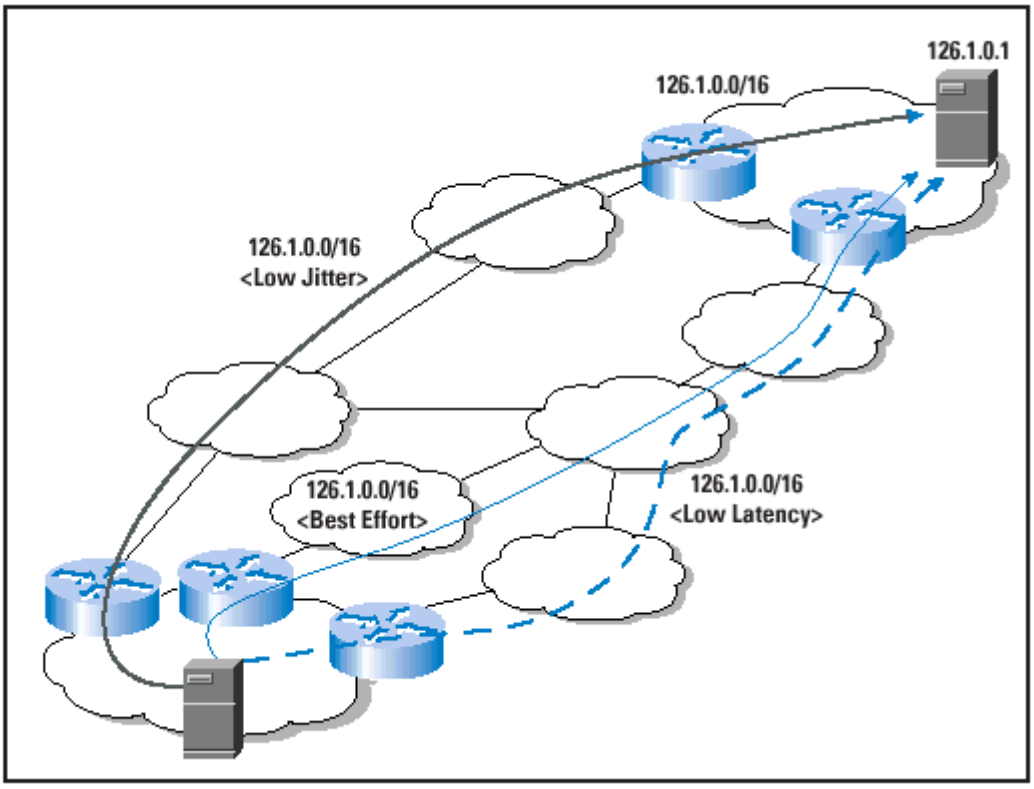
At an abstract level, this greater volume of routing information is needed in order to address the dual role of the routing system as both an inter-domain connectivity maintenance protocol and as a traffic-engineering tool.

Inter-Domain Quality of Service

Quality of Service (QoS) is a term that encompasses a wide variety of mechanisms. In the case of routing, the term is used to describe the process of modifying the normal routing response of associating a single forwarding action with a destination address prefix in such a way that there may be numerous forwarding decisions for a particular address prefix. Each forwarding decision is associated with a particular service response, so that a "best-effort" path to a particular destination address may differ from a "low-latency" path, which in turn may differ from a "high-bandwidth" path, and so on.

As with inter-domain traffic engineering, this requirement is one which would be expected to place greater volumes of information into the routing domain. At an abstract level this requirement can be seen as the association of a service quality attribute with an address prefix, and passing the paired entity into the routing domain as a single routing object. The inference is that multiple quality attributes associated with a path to a particular prefix would require the routing system to independently manipulate multiple route objects, because it would be reasonable to anticipate that the routing system would select different paths to reach the same address prefix if different QoS service attributes were used as a path qualifier (Figure 3).

Figure 3: Inter-Domain Routing with QoS



Approaches to Inter-Domain Routing

Let's now take this set of requirements and attempt to match them to various approaches to routing protocols.

Routing is a distributed computation wherein each element of the computation set must reach an outcome that is consistent with all other computations undertaken by other members of the set. There are two major approaches to this form of distributed computation, namely ***serial*** or ***parallel*** computation. Serial computation involves each element of the set undertaking a local computation and then passing the outcomes of this computation to its adjacent elements. This approach is used in various forms of distance-vector routing protocols where each routing node computes a local set of selected paths, and then propagates the set of reachable prefixes and the associated path metric to its neighbours. Parallel computation involves rapid flooding of the current state of connectivity within the set to all elements, and all set elements simultaneously compute forwarding decisions using the same base connectivity data. This approach is used in various forms of link-state routing protocols, where the protocol uses a flooding technique to rapidly propagate updated link-status information and then relies on each routing node to perform a local path selection computation for each reachable address prefix. Is one of these approaches substantially better suited than the other to the inter-domain routing environment?

Open or Closed Routing Policies

One of the key issues behind consideration of this topic is that of the role of ***local policy*** . Using a distance-vector protocol, a routing domain gathers selected path information from its neighbours, applies local policy to this information, and then distributes this updated information in the form of selected paths to its neighbour domains.

In this model the nature of the local policy applied to the routing information is not necessarily visible to the domain neighbours, 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 readvertised to each neighbour.

In link-state protocols, a routing domain effectively broadcasts its local domain 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 inter-domain 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 neighbour, and which routes will be advertised to each neighbour 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 favour 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.

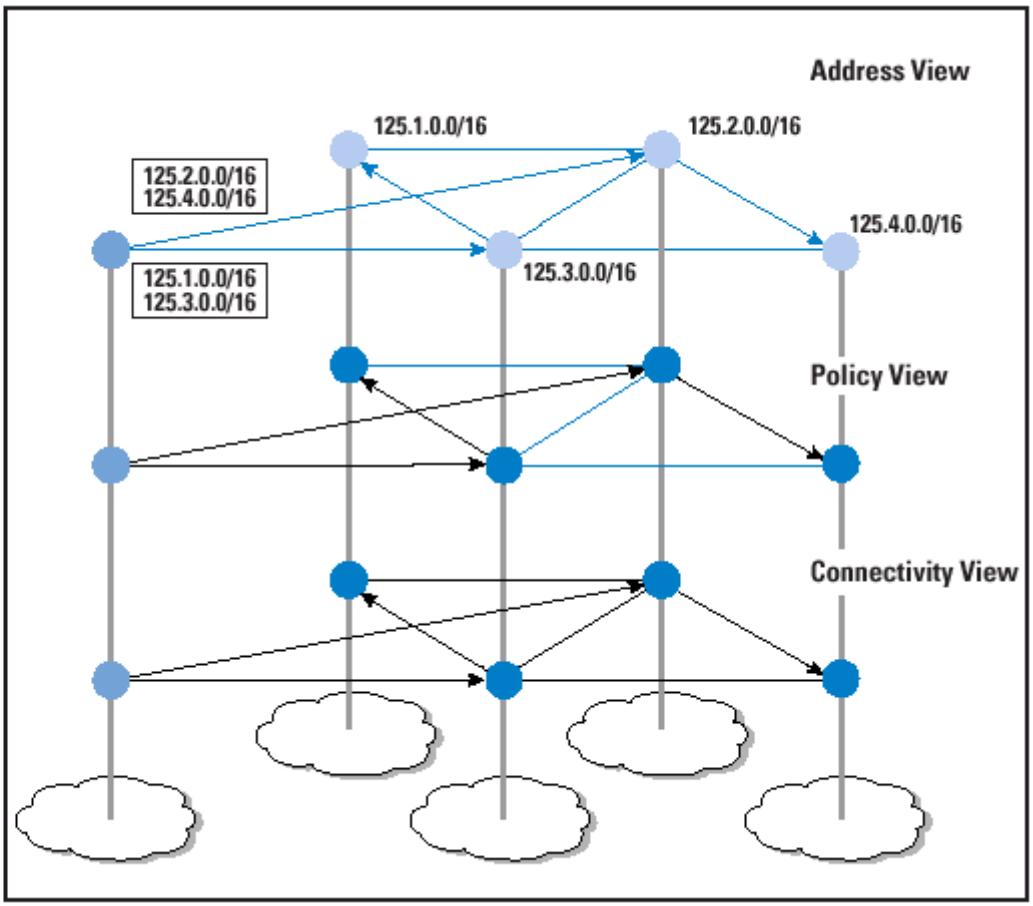
Separation of Functions

The inter-domain routing function undertakes many roles simultaneously. First, it maintains the current view of inter-domain connectivity. Any changes in the adjacency of a domain are reflected in a distributed update computation that determines if the adjacency change implies a change in path selection and in address reachability. Secondly, it maintains the set of currently reachable address prefixes. And finally, the protocol binds the first two functions together by associating each prefix with a path through the inter-domain space.

This association uses a policy framework to allow each domain to select a path that optimizes local policy constraints within the bounds of existing constraints applied by other domains. This policy may be related to traffic-engineering objectives, QoS requirements, local cost optimization, or related operational or business objectives.

An alternative approach to inter-domain routing is to separate the functions of connectivity maintenance, address reachability, and policy negotiation. As an example of this approach, a connectivity protocol can be used to identify all viable paths between a source and a destination domain. A policy negotiation protocol can be used to ensure that there are a consistent sequence of per-domain forwarding decisions that will pass traffic from the source domain to the destination domain. An address reachability protocol can be used to associate a collection of address prefixes with each destination domains. This framework is illustrated in Figure 4.

Figure 4: A Multi-Tiered Approach to Inter-Domain Routing



Address Prefixes and Autonomous System Numbers

One observation about the current inter-domain routing system is that it uses a view of the network based on computing the optimal path to each address prefix. This view is translated

into an inter-domain routing protocol that uses the address prefix as the basic protocol element and attaches various attributes to each address prefix as they are passed through the network.

As of late 2001, the routing system had some 100,000 distinct address prefixes and 11,500 origin domains. This implies that each origin domain is responsible for an average of 8 to 9 address prefixes. If each domain advertised its prefixes with a consistent policy, then each address prefix would be advertised with identical attributes. If the routing protocol were to be inverted such that the routing domain identifier, or ***Autonomous System*** number, were the basic routing object and the set of prefixes and associated common set of route attributes were attributes of the Autonomous System object, then the number of routing objects would be reduced by the same factor of between 8 and 9.

The motivation in this form of approach is that seeking clear hierarchical structure in the address space as deployed is no longer feasible, and that no further scaling advantage can be obtained by various forms of address aggregation within the routing system. This approach replaces this address-based hierarchy with a two-level hierarchy of routing domains. Within a routing domain, routing is undertaken using the address prefix. Between routing domains, routing is undertaken using domain identifiers and associated sets of domain attributes.

Although this approach appears to offer some advantage in creating a routing domain, one-tenth of the size of the address prefix-based routing domain, it is interesting to note that since late 1996 the average number of address prefixes per Autonomous System has fallen from 25 to the current value of 9. In other words, the number of distinct routing domains is growing at a faster rate than the number of routed address prefixes. While the adoption of a domain-based routing protocol offers some short-term advantages in scaling, the longer-term prospects are not so attractive, given these relative growth rates.

Routing Hierarchies of Information

The scaling properties of an inter-domain routing protocol are related on the ability of the protocol to remove certain specific items of information from the routing domain at the point where it ceases to have any differentiating impact. For example, it is important for a routing protocol to carry information that a particular domain has multiple adjacencies and that there are a number of policies associated with each adjacency, and propagate this information to all local domains. At a suitably distant point in the network, the forwarding decision remains the same regardless of the set of local adjacencies, and propagation of the detail of the local environment to points where the information ceases to have any distinguishing outcome is unproductive.

From this perspective, scaling the routing system is not a case of determining what information can be added into the routing domain, but instead it's a case of determining how much information can be removed from the routing domain, and how quickly.

One way of removing information is through the use of hierarchies. Within a hierarchical structure, a set of objects with similar properties are aggregated into a single object with a set of common properties. One way to perform such aggregation is by increasing the amount of information contained in each aggregate route object. For example, if single route objects are to be used that encompass a set of address prefixes and a collection of Autonomous Systems, then it would be necessary to define additional attributes within the route object to further qualify the policies associated with the object in terms of specific prefixes, specific Autonomous Systems, and specific policy semantics that may be considered as policy exceptions to the overall aggregate. This approach would allow aggregation of routing information to occur at any point in the network, allowing the aggregator to create a compound object with a common set of attributes, and a set of additional attributes that apply to a particular subset of the aggregate.

Another approach to using hierarchies 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 when the logical scope of the object is 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. This approach places control of aggregation with the route object originator, allowing the originator to specify the extent to which a specific route object should be propagated before being subsumed into an aggregate object.

It is not entirely clear that the approach of exploiting hierarchies in an address space is the most appropriate response to scaling pressures. Viewed from a more general perspective, scaling of the routing system requires the systematic removal of information from the routing domain. The way this is achieved is by attempting to align the structure of deployment with some structural property of the syntax of the protocol elements that are being used as routing objects. Information can then be eliminated through systematic aggregation of the routing objects at locations within the routing space that correspond to those points in the topology of the network where topology aggregation is occurring. The maintenance of this tight coupling of the structure of the deployed network to the structure of the identifier space is the highest cost of this approach. Alterations to the topology of the network through the relocation or reconfiguration of networks requires renumbering of the protocol element if hierarchical aggregation is to be maintained. If the address space is the basis of routing, as at present, then this becomes a large-scale exercise of renumbering networks that in turn implies an often prohibitively disruptive and expensive exercise of renumbering collections of host systems and associated services.

One view of this is that the connectivity properties of the Internet are already sufficiently meshed that there is no readily identifiable hierarchical structure, and that this trend is becoming more pronounced, not less. In that case, the most appropriate course of action may be to re-examine the routing domain and select some other attribute as the basis of the routing computation that does not have the same population, complexity, and growth characteristics as address prefixes, and base the routing computation on this attribute. One such alternative approach is to consider Autonomous System numbers as routing "atoms" where the routing system converges to select an Autonomous System path to a destination Autonomous System, and then uses this information to add the associated set of prefixes originated by this Autonomous System, and next-hop forwarding decision to reach this Autonomous System into the local forwarding table.

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.

The rationale for extending BGP would be to increase the number of commonly supported transitive route attributes, and, potentially, allow a richer syntax for attribute definition which in turn would allow the protocol to use a richer set of semantic definitions in order to express more complex routing policies.

This direction may sound like a step backward, in that it proposes an increase in the complexity of the route objects carried by the protocol and potentially increases the amount of local processing capability required to generate and receive routing updates. However, this can be offset by potential benefits that are realizable through the greater expressive capability for the policy attributes associated with route objects. It can allow a route originator an ability to specify the scope of propagation of the route object, rather than assuming that propagation will be global. The attributes can also describe intended service outcomes 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 behaviour 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. Use of a parallel- processing approach to the distributed computation of routing, such as that used in the link-state protocols, can offer the benefits of faster convergence times and avoidance of unstable transient routing states. On the other hand, link-state protocols present issues relating to policy opaqueness, as described above. Another major issue with such an approach is the need to address the efficiency of inter-domain link state flooding.

The inter-domain space would need some further levels of imposed structure similar to intra-domain areas in order to ensure that individual link updates are rapidly propagated across the relevant subset of the network. The use of such an area structure may well imply the need for an additional set of operator relationships, such as mutual transit. Such inter- domain relationships may prove challenging to adapt to existing operator practices.

Another approach could be based on the adoption of a multi-layer approach of separate protocols for separate functions, as described above. A base inter-domain connectivity protocol could potentially be based on a variant of a link-state protocol, using the rapid convergence properties of such protocols to maintain a coherent view of the current state of connectivity within the network. The overlay of a policy protocol would be intended as a signalling mechanism to allow each domain to make local forwarding decisions that are consistent with those adopted by adjacent domains, thereby maintaining a collection of coherent inter-domain paths from source to destination. Traffic engineering can also be envisaged as an overlay mechanism, allowing a source to make a forwarding decision that selects a path to the destination where the characteristics of the path optimize the desired service outcomes.

Directions for Further Activity

Although short-term actions based on providing various incentives for network operators to remove redundant or inefficiently grouped entries from the BGP routing table may exist, such actions are short-term palliative measures, and will not provide long-term answers to the need for a scalable inter-domain routing protocol. One approach to the longer term requirements may be to preserve many of the attributes of the current BGP protocol, while refining 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 administrative boundaries of information summarization, as well as retaining the approach of associating each prefix advertisement with an originating Autonomous System. The concept of policy opaqueness would also be retained in such an approach, implying that each Autonomous System accepts a set of route advertisements, applies local policy constraints, and readvertises 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 numerous 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.

Another approach is to consider the feasibility of decoupling the requirements of inter-domain connectivity management with the applications of policy constraints and the issues of sender- and 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

Autonomous System, and use a further overlay to support inter-domain traffic-engineering requirements. The underlying assumption of such an approach is that if the functional role of inter-domain routing is divided into distinct components, each component will have superior scaling and convergence properties which in turn will 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 the identity of a network, its location relative to other networks, and maintenance of a feasible path set 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 numerous distinct scalable components rather than using a single monolithic routing protocol.

Further Reading

[0] Huston, G., "Analyzing the Internet BGP Routing Table," ***The Internet Protocol Journal*** , Vol. 4, No. 1, March 2001.
www.cisco.com/warp/public/759/ipj_4-1/ipj_4-1_bgp.html

[1] Huitema, C., ***Routing in the Internet, 2nd Edition*** , ISBN 0130226475, Prentice Hall, January 2000. *A good introduction to the general topic of IP routing.*

[2] Rekhter, Y., and Li T., "A Border Gateway Protocol 4 (BGP-4)," RFC 1771, March 1995. *The base specification of BGP 4. This document is currently being updated by the IETF. The state of this work in progress as of November 2001 is documented as an Internet Draft, draft-ietf-idr-bgp4-15.txt*

[3] Elwyn Davies et al., "Future Domain Routing Requirements," work in progress, July 2001. *This work is currently documented as an Internet Draft, draft-davies-fdr-reqs-01.txt. It contains a review of an earlier effort in enumerating routing requirements ("Goals and Functional Requirements for Inter-Autonomous System Routing," RFC 1126, October 1989), as well as a commentary on a proposed set of current routing requirements.*

GEOFF HUSTON holds a B.Sc. and 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. He is currently the Chief Scientist in the Internet area for Telstra, a member of the Internet Architecture Board, and is the Secretary of the APNIC Executive Committee. 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