Network Working Group	E. Davies
Internet-Draft	Folly Consulting
Intended status: Historic	A. Doria
Expires: August 20, 2009	LTU
	February 16, 2009

TOC

Analysis of Inter-Domain Routing Requirements and History
draft-irtf-routing-history-10.txt

Status of this Memo

Copyright Notice

Abstract

This document analyses the state of the Internet domain-based routing system, concentrating on Inter-Domain Routing (IDR) and also considering the relationship between inter-domain and intra-domain routing. The analysis is carried out with respect to RFC 1126 and other IDR requirements and design efforts looking at the routing system as it appeared to be in 2001 with editorial additions reflecting developments up to 2006. It is the companion document to "A Set of Possible

Requirements for a Future Routing Architecture" [I-D.irtf-routing-reqs] (Doria, A., Davies, E., and F. Kastenholz, “A Set of Possible Requirements for a Future Routing Architecture,” February 2009.), which is a discussion of requirements for the future routing architecture, addressing systems developments and future routing protocols. This document summarizes discussions held several years ago by members of the IRTF Routing Research Group (IRTF RRG) and other interested parties. The document is published with the support of the IRTF RRG as a record of the work completed at that time, but with the understanding that it does not necessarily represent either the latest technical understanding or the technical concensus of the research group at the date of publication.

> [Note to RFC Editor: Please replace the reference in the abstract with a non-reference quoting the RFC number of the companion document when it is allocated, i.e., '(RFC xxxx)' and remove this note.]

Table of Contents

---

1. Provenance of this Document

TOC

In 2001, the IRTF Routing Research Group (IRTF RRG) chairs, Abha Ahuja and Sean Doran, decided to establish a sub-group to look at requirements for inter-domain routing (IDR). A group of well known routing experts was assembled to develop requirements for a new routing architecture. Their mandate was to approach the problem starting from a blank sheet. This group was free to take any approach, including a revolutionary approach, in developing requirements for solving the problems they saw in inter-domain routing. Their eventual approach documented requirements for a complete future routing and addressing architecture rather than just the requirements for IDR.
Simultaneously, an independent effort was started in Sweden with a similar goal. A team, calling itself Babylon, with participation from vendors, service providers, and academia, assembled to understand the history of inter-domain routing, to research the problems seen by the service providers, and to develop a proposal of requirements for a follow-on to the current routing architecture. This group's approach required an evolutionary approach starting from current routing architecture and practice. In other words the group limited itself to developing an evolutionary strategy and consequently assumed that the architecture would probably remain domain-based. The Babylon group was later folded into the IRTF RRG as Sub-group B to distinguish it from the original RRG Sub-group A.
This document, which was a part of Sub-group B's output, provides a snapshot of the current state of Inter-Domain Routing (IDR) at the time of original writing (2001) with some minor updates to take into account developments since that date, bringing it up to date in 2006. The development of the new requirements set is then motivated by an analysis of the problems that IDR has been encountering in the recent

past. This document is intended as a counterpart to the Routing Requirements document ("A Set of Possible Requirements for a Future Routing Architecture") which documents the requirements for future routing systems as captured separately by the IRTF RRG Sub-groups A and B [I-D.irtf-routing-reqs] (Doria, A., Davies, E., and F. Kastenholz, “A Set of Possible Requirements for a Future Routing Architecture,” February 2009.).
The IRTF RRG supported publication of this document as a historical record of the work completed on the understanding that it does not necessarily represent either the latest technical understanding or the technical consensus of the research group at the time of publication. The document has had substantial review by members of the Babylon team, members of the IRTF RRG and others over the years.

---

2. Introduction

TOC

For the greater part of its existence the Internet has used a domain-oriented routing system whereby the routers and other nodes making up the infrastructure are partitioned into a set of administrative domains, primarily along ownership lines. Individual routing domains (also known as Autonomous Systems (ASs)), which maybe a subset of an adminitrative domain, are made up of a finite, connected set of nodes (at least in normal operation). Each routing domain is subject to a coherent set of routing and other policies managed by a single administrative authority. The domains are interlinked to form the greater Internet producing a very large network: in practice, we have to treat this network as if it were infinite in extent as there is no central knowledge about the whole network of domains. An early presentation of the concept of routing domains can be found Paul Francis' OSI routing architecture paper from 1987 [Tsuchiya87] (Tsuchiya, P., “An Architecture for Network-Layer Routing in OSI,” 1987.) (Paul Francis was formerly known as Paul Tsuchiya).
The domain concept and domain-oriented routing has become so fundamental to Internet routing thinking that it is generally taken as an axiom these days and not even defined again (c.f., [NewArch03] (Clark, D., Sollins, K., Wroclawski, J., Katabi, D., Kulik, J., Yang, X., Braden, R., Faber, T., Falk, A., Pingali, V., Handley, M., and N. Chiappa, “New Arch: Future Generation Internet Architecture,” December 2003.)). The issues discussed in the present document notwithstanding, it has proved to be a robust and successful architectural concept that brings with it the possibility of using different routing mechanisms and protocols within the domains (intra-domain) and between the domains (inter-domain). This is an attractive division, because intra-domain protocols can exploit the well-known finite scope of the domain and the mutual trust engendered by shared ownership to give a high degree of control to the domain

administrators, whereas inter-domain routing lives in an essentially infinite region featuring a climate of distrust built on a multtude of competitive commercial agreements and driven by less-than-fully public policies from each component domain. Of course, like any other assumption that has been around for a very long time, the domain concept should be reevaluated to make sure that it is still helping!
It is generally accepted that there are major shortcomings in the inter-domain routing of the Internet today and that these may result in severe routing problems within an unspecified period of time. Remedying these shortcomings will require extensive research to tie down the exact failure modes that lead to these shortcomings and identify the best techniques to remedy the situation. Comparatively, intra-domain routing works satisfactorily, and issues with intra-domain routing are mainly associated with the interface between intra- and inter-domain routing.

> Reviewer's Note: Even in 2001, there was a wide difference of opinion across the community regarding the shortcomings of interdomain routing. In the years between writing and publication, further analysis, changes in operational practice, alterations to the demands made on inter-domain routing, modifications made to BGP and a recognition of the difficulty of finding a replacement may have altered the views of some members of the community.

Changes in the nature and quality of the services that users want from the Internet are difficult to provide within the current framework, as they impose requirements never foreseen by the original architects of the Internet routing system.
The kind of radical changes that have to be accommodated are epitomized by the advent of IPv6 and the application of IP mechanisms to private commercial networks that offer specific service guarantees beyond the best-effort services of the public Internet. Major changes to the inter-domain routing system are inevitable to provide an efficient underpinning for the radically changed and increasingly commercially-based networks that rely on the IP protocol suite.
Current practice stresses the need to separate the concerns of the control plane and the forwarding plane in a router: This document will follow this practice, but we still use the term 'routing' as a global portmanteau to cover all aspects of the system.
This document provides a historical perspective on the current state of inter-domain routing and its relationship to intra-domain routing in Section 3 (Historical Perspective) by revisiting the previous IETF requirements document intended to steer the development of a future routing system. These requirements, which informed the design of the Border Gateway Protocol (BGP) in 1989, are contained in RFC1126 - "Goals and Functional Requirements for Inter-Autonomous System Routing" [RFC1126] (Little, M., “Goals and functional requirements for inter-autonomous system routing,” October 1989.).

Section 3 (Historical Perspective) also looks at some other work on requirements for domain-based routing that was carried out before and after RFC1126 was published. This work fleshes out the historical perspective and provides some additional insights into alternative approaches which may be instructive when building a new set of requirements.
The motivation for change and the inspiration for some of the requirements for new routing architectures derive from the problems attributable to the current domain-based routing system that are being experienced in the Internet today. These will be discussed in Section 5 (Existing problems of BGP and the current Inter-/Intra-Domain Architecture).

2.1. Background

TOC

Today's Internet uses an addressing and routing structure that has developed in an ad hoc, more or less upwards-compatible fashion. The structure has progressed from supporting a non-commercial Internet with a single administrative domain to a solution that is able to control today's multi-domain, federated Internet, carrying traffic between the networks of commercial, governmental and not-for-profit participants. This is not achieved without a great deal of 24/7 vigilance and operational activity by network operators: Internet routing often appears to be running close to the limits of stability. As well as directing traffic to its intended end-point, inter-domain routing mechanisms are expected to implement a host of domain specific routing policies for competing, communicating domains. The result is not ideal, particularly as regards inter-domain routing mechanisms, but it does a pretty fair job at its primary goal of providing any-to-any connectivity to many millions of computers.
Based on a large body of anecdotal evidence, but also on a growing body of experimental evidence [Labovitz02] (Labovitz, C., Ahuja, A., Farnam, J., and A. Bose, “Experimental Measurement of Delayed Convergence,” 2002.) and analytic work on the stability of BGP under certain policy specifications [Griffin99] (Griffin, T. and G. Wilfong, “An Analysis of BGP Convergence Properties,” 1999.), the main Internet inter-domain routing protocol, BGP version 4 (BGP-4), appears to have a number of problems. These problems are discussed in more detail in Section 5 (Existing problems of BGP and the current Inter-/Intra-Domain Architecture). Additionally, the hierarchical nature of the inter-domain routing problem appears to be changing as the connectivity between domains becomes increasingly meshed [RFC3221] (Huston, G., “Commentary on Inter-Domain Routing in the Internet,” December 2001.) which alters some of the scaling and structuring assumptions on which BGP-4 is built. Patches and fix-ups may relieve some of these problems but others may require a new architecture and new protocols.

3. Historical Perspective

TOC

3.1. The Legacy of RFC1126

TOC

RFC 1126 [RFC1126] (Little, M., “Goals and functional requirements for inter-autonomous system routing,” October 1989.) outlined a set of requirements that were intended to guide the development of BGP.

> Editors' Note: When this document was reviewed by Yakov Rekhter, one of the designers of BGP, his view was that "While some people expected a set of requirements outlined in RFC1126 to guide the development of BGP, in reality the development of BGP happened completely independently of RFC1126. In other words, from the point of view of the development of BGP, RFC1126 turned out to be totally irrelevant." On the other hand, it appears that BGP as currently implemented has met a large proportion of these requirements, especially for unicast traffic.

While the network is demonstrably different from what it was in 1989, having

- moved from single to multiple administrative control,
- increased in size by several orders of magnitude, and
- migrated from a fairly tree like connectivity graph to a meshier style,

many of the same requirements remain. As a first step in setting requirements for the future, we need to understand the requirements that were originally set for the current protocols. And in charting a future architecture we must first be sure to do no harm. This means a future domain-based routing system has to support as its base requirement, the level of function that is available today.
The following sections each relate to a requirement, or non-requirement listed in RFC1126. In fact the section names are direct quotes from the document. The discussion of these requirements covers the following areas:

> **Explanation:** Optional interpretation for today's audience of the original intent of the requirement

Relevance:

Is the requirement of RFC1126 still relevant, and to what degree? Should it be understood differently in today's environment?

Current practice: How well is the requirement met by current protocols and practice?

3.1.1. "General Requirements"

TOC

3.1.1.1. "Route to Destination"

TOC

Timely routing to all reachable destinations, including multihoming and multicast.

Relevance: Valid, but requirements for multihoming need further discussion and elucidation. The requirement should include multiple source multicast routing.

Current practice: Multihoming is not efficient and the proposed inter-domain multicast protocol BGMP [RFC3913] (Thaler, D., "Border Gateway Multicast Protocol (BGMP): Protocol Specification," September 2004.) is an add-on to BGP following many of the same strategies but not integrated into the BGP framework.

> Editors' Note: Multicast routing has moved on again since this was originally written. By 2006 BGMP had been effectively superseded. Multicast routing now uses Multiprotocol BGP [RFC4760] (Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4," January 2007.), the Multicast Source Discovery Protocol (MSDP) [RFC3618] (Fenner, B. and D. Meyer, "Multicast Source Discovery Protocol (MSDP)," October 2003.) and Protocol Independent Multicast - Sparse Mode (PIM-SM) [RFC2362] (Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., and V. Jacobson, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification," June 1998.), [RFC4601] (Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol

Specification (Revised)," August 2006.), especially the Source Specific Multicast (SSM) subset.

3.1.1.2. "Routing is Assured"

TOC

This requires that a user be notified within a reasonable time period after persistent attempts, about inability to provide a service.

Relevance: Valid

Current practice: There are ICMP messages for this, but in many cases they are not used, either because of fears about creating message storms or uncertainty about whether the end system can do anything useful with the resulting information. IPv6 implementations may be able to make better use of the information as they may have alternative addresses that could be used to exploit an alternative routing.

3.1.1.3. "Large System"

TOC

The architecture was designed to accommodate the growth of the Internet.

Relevance: Valid. Properties of Internet topology might be an issue for future scalability (topology varies from very sparse to quite dense at present). Instead of setting out to accomodate growth in a specific time period, indefinite growth should be accommodated. On the other hand, such growth has to be accommodated without making the protocols too expensive - trade-offs may be necessary.

Current practice: Scalability of the current protocols will not be sufficient under the current rate of growth. There are problems with BGP convergence for large dense topologies, problems with the slow speed of routing information propagation between routers in transit domains through the intra-domain protocol for example when a failure requires traffic to be redirected to an alternative exit point from the domain (see Section 5.9 (Interaction between Inter-Domain Routing and Intra-Domain Routing), limited support for hierarchy, etc.

3.1.1.4. "Autonomous Operation"

TOC

This requirement encapsulates the need for administrative domains ("Autonomous Systems" - AS) to be able to operate autonomously as regards setting routing policy:

> **Relevance:** Valid. There may need to be additional requirements for adjusting policy decisions to the global functionality and for avoiding contradictory policies. This would decrease the possibility of unstable routing behavior.
> There is a need for handling various degrees of trust in autonomous operations, ranging from no trust (e.g., between separate ISPs) to very high trust where the domains have a common goal of optimizing their mutual policies.
>
> Policies for intra-domain operations should in some cases be revealed, using suitable abstractions.
>
> **Current practice:** Policy management is in the control of network managers, as required, but there is little support for handling policies at an abstract level for a domain.
> Cooperating administrative entities decide about the extent of cooperation independently. This can lead to inconsistent, and potentially incompatible routing policies being applied in notionally cooperating domains. As discussed in Sections (Convergence and Recovery Issues), (Non-locality of Effects of Instability and Misconfiguration)and (Policy Issues),lack of coordination combined with global range of effects of BGP policies results in occasional disruption of Internet routing over an area far wider than the domains that are not cooperating effectively.

3.1.1.5. "Distributed System"

TOC

The routing environment is a distributed system. The distributed routing environment supports redundancy and diversity of nodes and links. Both the controlling rule sets, which implement the routing policies, and the places where operational control is applied, through decisions on path selection, are distributed (primarily in the routers).

> **Relevance:** Valid. RFC1126 is very clear that we should not be using centralized solutions, but maybe we need a discussion on trade-offs between common knowledge and distribution (i.e., to allow

for uniform policy routing, e.g., GSM systems are in a sense centralized, but with hierarchies).

Current practice: Routing is very distributed, but lacking the ability to consider optimization over several hops or domains.

> Editors' Note: Also coordinating the implementation of a set of routing policies across a large domain with many routers running BGP is difficult. The policies have to be turned into BGP rules and applied individually to each router, giving opportunities for mismatch and error.

3.1.1.6. "Provide A Credible Environment"

TOC

The routing environment and services should be based upon mechanisms and information that exhibit both integrity and security. That is the routers should always be working with credible data derived through the reliable operation of protocols. Security from unwanted modification and influence is required.

Relevance: Valid.

Current practice: BGP provides a limited mechanism for authentication and security of peering sessions, but this does not guarantee the authenticity or validity of the routing information that is exchanged.
There are certainly security problems with current practice. The Routing Protocol Security Requirements (rpsec) working group has been struggling to agree on a set of requirements for BGP security since early 2002.

> Editors' note: Proposals for authenticating BGP routing information using certificates were under development by the Secure Inter-Domain Routing (sidr) working group from 2006 through 2008.

3.1.1.7. "Be A Managed Entity"

TOC

Requires that the routing system provides adequate information on the state of the network to allow resource, problem and fault management to be carried out effectively and expeditiously. The system must also

provide controls that allow managers to use this information to make informed decisions and use it to control the operation of the routing system.

> **Relevance:** The requirement is reasonable, but we might need to be more specific on what information should be available, e.g., to prevent routing oscillations.
>
> **Current practice:** All policies are determined locally, where they may appear reasonable but there is limited global coordination through the routing policy databases operated by the Internet registries (AfriNIC, APNIC, ARIN, LACNIC, RIPE, etc.). Operators are not required to register their policies; even when policies are registered, it is difficult to check that the actual policies in use in other domains match the declared policies. Therefore, a manager cannot guarantee to design and implement policies that will interoperate with those of other domains to provide stable routing.
>
> > Editors' note: Operators report that management of BGP-based routing remains a function that needs highly-skilled operators and continual attention.

3.1.1.8. "Minimize Required Resources"

TOC

> **Relevance:** Valid, however, the paragraph states that assumptions on significant upgrades shouldn't be made. Although this is reasonable, a new architecture should perhaps be prepared to use upgrades when they occur.
>
> **Current practice:** Most bandwidth is consumed by the exchange of the Network Layer Reachability Information (NLRI). Usage of processing cycles ("Central Processor Usage" - CPU) depends on the stability of the Internet. Both phenomena have a local nature, so there are not scaling problems with bandwidth and CPU usage. Instability of routing increases the consumption of resources in any case. The number of networks in the Internet dominates memory requirements - this is a scaling problem.

3.1.2. "Functional Requirements"

TOC

3.1.2.1. "Route Synthesis Requirements"

TOC

3.1.2.1.1. "Route around failures dynamically"

TOC

> **Relevance:** Valid. Should perhaps be stronger. Only providing a best-effort attempt may not be enough if real-time services are to be provided for. Detection of failures may need to be faster than 100ms to avoid being noticed by end-users.
>
> **Current practice:** Latency of fail-over is too high; sometimes minutes or longer.

3.1.2.1.2. "Provide loop free paths"

TOC

> **Relevance:** Valid. Loops should occur only with negligible probability and duration.
>
> **Current practice:** Both link-state intra-domain routing and BGP inter-domain routing (if correctly configured) are forwarding-loop free after having converged. However, convergence time for BGP can be very long and poorly designed routing policies may result in a number of BGP speakers engaging in a cyclic pattern of advertisements and withdrawals which never converges to a stable result [RFC3345] (McPherson, D., Gill, V., Walton, D., and A. Retana, “Border Gateway Protocol (BGP) Persistent Route Oscillation Condition,” August 2002.). Part of the reason for long convergence times is the non-locality of the effects of changes in BGP advertisements (see Section 5.3 (Non-locality of Effects of Instability and Misconfiguration)). Modifying the inter-domain routing protocol to make the effects of changes less global, and convergence a more local condition might improve performance, assuming a suitable modification could be developed.

3.1.2.1.3. "Know when a path or destination is unavailable"

TOC

> **Relevance:** Valid to some extent, but there is a trade-off between aggregation and immediate knowledge of reachability. It requires

that routing tables contain enough information to determine that the destination is unknown or a path cannot be constructed to reach it.

Current practice: Knowledge about lost reachability propagates slowly through the networks due to slow convergence for route withdrawals.

3.1.2.1.4. "Provide paths sensitive to administrative policies"

TOC

Relevance: Valid. Policy control of routing has become increasingly important as the Internet has turned into a business.

Current practice: Supported to some extent. Policies can only be applied locally in an AS and not globally. Policy information supplied has a very small probability of affecting policies in other ASs. Furthermore, only static policies are supported; between static policies and policies dependent upon volatile events of great celerity there should exist events that routing should be aware of. Lastly, there is no support for policies other than route-properties (such as AS-origin, AS-path, destination prefix, MED-values etc).

> Editors' note: Subsequent to the original issue of this document mechanisms which acknowledge the business relationships of operators have been developed such as the NOPEER community attribute [RFC3765] (Huston, G., “NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control,” April 2004.). However the level of usage of this attribute is apparently not very great.

3.1.2.1.5. "Provide paths sensitive to user policies"

TOC

Relevance: Valid to some extent, as they may conflict with the policies of the network administrator. It is likely that this requirement will be met by means of different bit transport services offered by an operator, but at the cost of adequate provisioning, authentication and policing when utilizing the

service.

Current practice: Not supported in normal routing. Can be accomplished to some extent with loose source routing, resulting in inefficient forwarding in the routers. The various attempts to introduce Quality of Service (QoS - e.g., Integrated Services and Differentiated Services (DiffServ)) can also be seen as means to support this requirement but they have met with limited success in terms of providing alternate routes as opposed to providing improved service on the standard route.

> Editor's Note: From the standpoint of a later time, it would probably be more appropriate to say "total failure" rather than "limited success".

3.1.2.1.6. "Provide paths which characterize user quality-of-service requirements"

TOC

Relevance: Valid to some extent, as they may conflict with the policies of the operator. It is likely that this requirement will be met by means of different bit transport services offered by an operator, but at the cost of adequate provisioning, authentication and policing when utilizing the service. It has become clear that offering to provide a particular QoS to any arbitrary destination from a particular source is generally impossible: QoS except in very 'soft' forms such as overall long term average packet delay, is generally associated with connection oriented routing.

Current practice: Creating routes with specified QoS is not generally possible at present.

3.1.2.1.7. "Provide autonomy between inter- and intra-autonomous system route synthesis"

TOC

Relevance: Inter- and intra-domain routing should stay independent, but one should notice that this to some extent contradicts the previous three requirements. There is a trade-off between abstraction and optimality.

Current practice: Inter-domain routing is performed independently of intra-domain routing. Intra-domain routing is however,

> especially in transit domains, very interrelated with inter-domain routing.

3.1.2.2. "Forwarding Requirements" TOC

3.1.2.2.1. "Decouple inter- and intra-autonomous system forwarding decisions" TOC

> **Relevance:** Valid.
>
> **Current practice:** As explained in Section 3.1.2.1.7 ("Provide autonomy between inter- and intra-autonomous system route synthesis"), intra-domain forwarding in transit domains is dependent on inter-domain forwarding decisions.

3.1.2.2.2. "Do not forward datagrams deemed administratively inappropriate" TOC

> **Relevance:** Valid, and increasingly important in the context of enforcing policies correctly expressed through routing advertisements but flouted by rogue peers which send traffic for which a route has not been advertised. On the other hand, packets that have been misrouted due to transient routing problems perhaps should be forwarded to reach the destination, although along an unexpected path.
>
> **Current practice:** At stub domains (i.e., domains that do not provide any transit service for any other domains but that connect directly to one or more transit domains) there is packet filtering, e.g., to catch source address spoofing on outgoing traffic or to filter out unwanted incoming traffic. Filtering can in particular reject traffic (such as unauthorized transit traffic) that has been sent to a domain even when it has not advertised a route for such traffic on a given interface. The growing class of 'middle boxes' (midboxes, e.g., Network Address Translators - NATs) is quite likely to apply administrative rules that will prevent forwarding of packets. Note that security policies may deliberately hide administrative denials. In the

backbone, intentional packet dropping based on policies is not common.

3.1.2.2.3. "Do not forward datagrams to failed resources"

TOC

Relevance: Unclear, although it is clearly desirable to minimise waste of forwarding resources by discarding datagrams which cannot be delivered at the earliest opportunity. There is a trade-off between scalability and keeping track of unreachable resources. The requirement effectively imposes a requirement on adjacent nodes to monitor for failures and take steps to cause rerouting at the earliest opportinity if a failure is detected. However, packets that are already in flight or queued on a failed link cannot generally be rescued.

Current practice: Routing protocols use both internal adjacency management sub-protocols (e.g. Hello protocols) and information from equipment and lower layer link watchdogs to keep track of failures in routers and connecting links. Failures will eventually result in the routing protocol reconfiguring the routing to avoid (if possible) a failed resource, but this is generally very slow (30s or more). In the meantime datagrams may well be forwarded to failed resources. In general terms, end hosts and some non-router middle boxes do not participate in these notifications and failures of such boxes will not affect the routing system.

3.1.2.2.4. "Forward datagram according to its characteristics"

TOC

Relevance: Valid. This is necessary in enabling differentiation in the network, based on QoS, precedence, policy or security.

Current practice: Ingress and egress filtering can be done based on policy. Some networks discriminate on the basis of requested QoS.

3.1.2.3. "Information Requirements"

TOC

3.1.2.3.1. "Provide a distributed and descriptive information base"

TOC

> **Relevance:** Valid, however an alternative arrangement of information bases, possibly with an element of centralization for the domain (as mentioned in Section 3.1.1.5 ("Distributed System")) might offer some advantages through the ability to optimize across the domain and respond more quickly to changes and failures.
>
> **Current practice:** The information base is distributed, but it is unclear whether it supports all necessary routing functionality.

3.1.2.3.2. "Determine resource availability"

TOC

> **Relevance:** Valid. It should be possible to determine the availability and levels of availability of any resource (such as bandwidth) needed to carry out routing. This prevents needing to discover unavailability through failure. Resource location and discovery is arguably a separate concern that could be addressed outside the core routing requirements.
>
> **Current practice:** Resource availability is predominantly handled outside of the routing system.

3.1.2.3.3. "Restrain transmission utilization"

TOC

> **Relevance:** Valid. However certain requirements in the control plane, such as fast detection of faults may be worth consumption of more resources. Similarly, simplicity of implementation may make it cheaper to 'back haul' traffic to central locations to minimise the cost of routing if bandwidth is cheaper than processing.
>
> **Current practice:** BGP messages probably do not ordinarily consume excessive resources, but might during erroneous conditions. In the data plane, the near universal adoption of shortest path protocols could be considered to result in minimization of transmission utilization.

3.1.2.3.4. "Allow limited information exchange"

TOC

> **Relevance:** Valid. But perhaps routing could be improved if certain information (especially policies) could be available either globally or at least for a wider defined locality.
>
> > Editors' note: Limited information exchange would be potentially compatible with a more local form of convergence than BGP tries to achieve today. Limited information exchange is potentially incompatible with global convergence.
>
> **Current practice:** Policies are used to determine which reachability information is exported but neighbors receiving the information are not generally aware of the policies that resulted in this export.

3.1.2.4. "Environmental Requirements"

TOC

3.1.2.4.1. "Support a packet-switching environment"

TOC

> **Relevance:** Valid but routing system should, perhaps, not be limited to this exclusively.
>
> **Current practice:** Supported.

3.1.2.4.2. "Accommodate a connection-less oriented user transport service"

TOC

> **Relevance:** Valid, but routing system should, perhaps, not be limited to this exclusively.
>
> **Current practice:** Accommodated.

3.1.2.4.3. "Accommodate 10K autonomous systems and 100K networks"

TOC

> **Relevance:** No longer valid. Needs to be increased potentially indefinitely. It is extremely difficult to foresee the future size expansion of the Internet so that the Utopian solution would be to achieve an Internet whose architecture is scale invariant. Regrettably, this may not be achievable without introducing undesirable complexity and a suitable trade off between complexity and scalability is likely to be necessary.
>
> **Current Practice:** Supported but perhaps reaching its limit. Since the original version of this document was written in 2001, the number of ASs advertised has grown from around 8000 to 20000, and almost 35000 AS numbers have been allocated by the regional registries [Huston05] (Huston, G., “Exploring Autonomous System Numbers,” August 2005.). If this growth continues the original 16 bit AS space in BGP-4 will be exhausted in less than 5 years. Planning for an extended AS space is now an urgent requirement.

3.1.2.4.4. "Allow for arbitrary interconnection of autonomous systems"

TOC

> **Relevance:** Valid. However perhaps not all interconnections should be accessible globally.
>
> **Current practice:** BGP-4 allows for arbitrary interconnections.

3.1.2.5. "General Objectives"

TOC

3.1.2.5.1. "Provide routing services in a timely manner"

TOC

> **Relevance:** Valid, as stated before. It might be acceptable for a more complex service to take longer to deliver, but it still has to meet the application's requirements - routing has to be at the service of the end-to-end principle.

> Editors' note: Delays in setting up connections due to network functions such as NAT boxes are becoming increasingly problematic. The routing system should try to keep any routing delay to a minimum.

Current practice: More or less, with the exception of convergence and fault robustness.

3.1.2.5.2. "Minimize constraints on systems with limited resources"

TOC

Relevance: Valid

Current practice: Systems with limited resources are typically stub domains that advertise very little information.

3.1.2.5.3. "Minimize impact of dissimilarities between autonomous systems"

TOC

Relevance: Important. This requirement is critical to a future architecture. In a domain-based routing environment where the internal properties of domains may differ radically, it will be important to be sure that these dissimilarities are minimized at the borders.

Current: practice: For the most part this capability is not really required in today's networks since the intra-domain attributes are broadly similar across domains.

3.1.2.5.4. "Accommodate the addressing schemes and protocol mechanisms of the autonomous systems"

TOC

Relevance: Important, probably more so than when RFC1126 was originally developed because of the potential deployment of IPv6, wider usage of MPLS and the increasing usage of VPNs.

Current practice:

Only one global addressing scheme is supported in most autonomous systems but the availability of IPv6 services is steadily increasing. Some global backbones support IPv6 routing and forwarding.

3.1.2.5.5. "Must be implementable by network vendors"

TOC

Relevance: Valid, but note that what can be implemented today is different from what was possible when RFC1126 was written: a future domain-based routing architecture should not be unreasonably constrained by past limitations.

Current practice: BGP was implemented and meets a large proportion of the original requirements.

3.1.3. "Non-Goals"

TOC

RFC1126 also included a section discussing non-goals. This section discusses the extent to which these are still non-goals. It also considers whether the fact that they were non-goals adversely affects today's IDR system.

3.1.3.1. "Ubiquity"

TOC

The authors of RFC 1126 were explicitly saying that IP and its inter-domain routing system need not be deployed in every AS, and a participant should not necessarily expect to be able to reach a given AS, possibly because of routing policies. In a sense this 'non-goal' has effectively been achieved by the Internet and IP protocols. This requirement reflects a different world view where there was serious competition for network protocols, which is really no longer the case. Ubiquitous deployment of inter-domain routing in particular has been

achieved and must not be undone by any proposed future domain-based routing architecture. On the other hand:

*ubiquitous connectivity cannot be reached in a policy sensitive environment and should not be an aim,

-Editor's Note: It has been pointed out that this statement could be interpreted as being contrary to the Internet mission of providing universal connectivity. The fact that limits to connectivity will be added as operational requirements in a policy sensitive environment should not imply that a future domain-based routing architecture contains intrinsic limits on connectivity.

*it must not be required that the same routing mechanisms are used throughout provided that they can interoperate appropriately

*the information needed to control routing in a part of the network should not necessarily be ubiquitously available and it must be possible for an operator to hide commercially sensitive information that is not needed outside a domain.

*the introduction of IPv6 reintroduces an element of diversity into the world of network protocols but the similarities of IPv4 and IPv6 as regards routing and forwarding make this event less likely to drive an immediate diversification in routing systems. The potential for further growth in the size of the network enabled by IPv6 is very likely to require changes in the future: whether this results in the replacement of one de facto ubiquitous system with another remains to be seen but cannot be a requirement - it will have to interoperate with BGP during the transition..

Relevance: De facto essential for a future domain-based routing architecture, but what is required is ubiquity of the routing system rather than ubiquity of connectivity and it must be capable of a gradual takeover through interoperation with the existing system.

Current practice: De facto ubiquity achieved.

3.1.3.2. "Congestion control"

TOC

Relevance: It is not clear if this non-goal was to be applied to routing or forwarding. It is definitely a non-goal to adapt the choice of route when there is transient congestion. However, to

add support for congestion avoidance (e.g., Explicit Congestion Notification (ECN) and ICMP messages) in the forwarding process would be a useful addition. There is also extensive work going on in traffic engineering which should result in congestion avoidance through routing as well as in forwarding.

Current practice: Some ICMP messages (e.g., source quench) exist to deal with congestion control but these are not generally used as they either make the problem worse or there is no mechanism to reflect the message into the application which is providing the source.

3.1.3.3. "Load splitting"

TOC

Relevance: This should neither be a non-goal, nor an explicit goal. It might be desirable in some cases and should be considered as an optional architectural feature.

Current practice: Can be implemented by exporting different prefixes on different links, but this requires manual configuration and does not consider actual load.

> Editors' Note: This configuration is carried out extensively as of 2006 and has been a significant factor in routing table bloat. If this need is a real operational requirement, as it seems to be for multihomed or otherwise richly connected sites, it will be necessary to reclassify this as a real and important goal.

3.1.3.4. "Maximizing the utilization of resources"

TOC

Relevance: Valid. Cost-efficiency should be striven for; we note that maximizing resource utilization does not always lead to greatest cost-efficiency.

Current practice: Not currently part of the system, though often a 'hacked in' feature done with manual configuration.

TOC

3.1.3.5. "Schedule to deadline service"

This non-goal was put in place to ensure that the IDR did not have to meet real time deadline goals such as might apply to Constant Bit Rate (CBR) real time services in ATM.

> **Relevance:** The hard form of deadline services is still a non-goal for the future domain-based routing architecture but overall delay bounds are much more of the essence than was the case when RFC1126 was written.
>
> **Current practice:** Service providers are now offering overall probabilistic delay bounds on traffic contracts. To implement these contracts there is a requirement for a rather looser form of delay sensitive routing.

3.1.3.6. "Non-interference policies of resource utilization"

TOC

The requirement in RFC1126 is somewhat opaque, but appears to imply that what we would today call QoS routing is a non-goal and that routing would not seek to control the elastic characteristics of Internet traffic whereby a TCP connection can seek to utilize all the spare bandwidth on a route, possibly to the detriment of other connections sharing the route or crossing it.

> **Relevance:** Open Issue. It is not clear whether dynamic QoS routing can or should be implemented. Such a system would seek to control the admission and routing of traffic depending on current or recent resource utilization. This would be particularly problematic where traffic crosses an ownership boundary because of the need for potentially commercially sensitive information to be made available outside the ownership boundary.
>
> **Current practice:** Routing does not consider dynamic resource availability. Forwarding can support service differentiation.

3.2. ISO OSI IDRP, BGP and the Development of Policy Routing

TOC

During the decade before the widespread success of the World Wide Web, ISO was developing the communications architecture and protocol suite Open Systems Interconnection (OSI). For a considerable part of this time OSI was seen as a possible competitor for and even a replacement

for the IP suite as this basis for the Internet. The technical developments of the two protocols were quite heavily interrelated with each providing ideas and even components that were adapted into the other suite.

During the early stages of the development of OSI, the IP suite was still mainly in use on the ARPANET and the relatively small scale first phase NSFnet. This was effectively a single administrative domain with a simple tree structured network in a three level hierarchy connected to a single logical exchange point (the NSFnet backbone). In the second half of the 1980s the NSFNET was starting on the growth and transformation that would lead to today's Internet. It was becoming clear that the backbone routing protocol, the Exterior Gateway Protocol (EGP) [RFC0904] (Mills, D., “Exterior Gateway Protocol formal specification,” April 1984.), was not going to cope even with the limited expansion being planned. EGP is an "all informed" protocol which needed to know the identities of all gateways and this was no longer reasonable. With the increasing complexity of the NSFnet and the linkage of the NSFnet network to other networks there was a desire for policy-based routing which would allow administrators to manage the flow of packets between networks. The first version of the Border Gateway Protocol (BGP-1) [RFC1105] (Lougheed, K. and J. Rekhter, “Border Gateway Protocol (BGP),” June 1989.) was developed as a replacement for EGP with policy capabilities - a stopgap EGP version 3 had been created as an interim measure while BGP was developed. BGP was designed to work on a hierarchically structured network, such as the original NSFNET, but could also work on networks that were at least partially non-hierarchical where there were links between ASs at the same level in the hierarchy (we would now call these 'peering arrangements') although the protocol made a distinction between different kinds of links (links are classified as upwards, downwards or sideways). ASs themselves were a 'fix' for the complexity that developed in the three tier structure of the NSFnet.

Meanwhile the OSI architects, led by Lyman Chapin, were developing a much more general architecture for large scale networks. They had recognized that no one node, especially an end-system (host) could or should attempt to remember routes from "here" to "anywhere" - this sounds obvious today but was not so obvious 20 years ago. They were also considering hierarchical networks with independently administered domains - a model already well entrenched in the public switched telephone network. This led to a vision of a network with multiple independent administrative domains with an arbitrary interconnection graph and a hierarchy of routing functionality. This architecture was fairly well established by 1987 [Tsuchiya87] (Tsuchiya, P., “An Architecture for Network-Layer Routing in OSI,” 1987.). The architecture initially envisaged a three level routing functionality hierarchy in which each layer had significantly different characteristics:

1. **End-system to Intermediate system routing (host to router)**, in which the principal functions are discovery and redirection.

2. **Intra-domain intermediate system to intermediate system routing (router to router)**, in which "best" routes between end-systems in a single administrative domain are computed and used. A single algorithm and routing protocol would be used throughout any one domain.

3. **Inter-domain intermediate-system to intermediate system routing (router to router)**, in which routes between routing domains within administrative domains are computed (routing is considered separately between administrative domains and routing domains).

Level 3 of this hierarchy was still somewhat fuzzy. Tsuchiya says:

> The last two components, Inter-Domain and Inter-Administration routing, are less clear-cut. It is not obvious what should be standardized with respect to these two components of routing. For example, for Inter-Domain routing, what can be expected from the Domains? By asking Domains to provide some kind of external behavior, we limit their autonomy. If we expect nothing of their external behavior, then routing functionality will be minimal.
>
> Across administrations, it is not known how much trust there will be. In fact, the definition of trust itself can only be determined by the two or more administrations involved.
>
> Fundamentally, the problem with Inter-Domain and Inter-Administration routing is that autonomy and mistrust are both antithetical to routing. Accomplishing either will involve a number of tradeoffs which will require more knowledge about the environments within which they will operate.

Further refinement of the model occurred over the next couple of years and a more fully formed view is given by Huitema and Dabbous in 1989 [Huitema90] (Huitema, C. and W. Dabbous, “Routeing protocols development in the OSI architecture,” 1990.). By this stage work on the original IS-IS link state protocol, originated by the Digital Equipment Corporation (DEC), was fairly advanced and was close to becoming a Draft International Standard. IS-IS is of course a major component of intra-domain routing today and inspired the development of the Open

Shortest Path First (OSPF) family. However, Huitema and Dabbous were not able to give any indication of protocol work for Level 3. There are hints of possible use of centralized route servers.
In the meantime, the NSFnet consortium and the IETF had been struggling with the rapid growth of the NSFnet. It had been clear since fairly early on that EGP was not suitable for handling the expanding network and the race was on to find a replacement. There had been some intent to include a metric in EGP to facilitate routing decisions, but no agreement could be reached on how to define the metric. The lack of trust was seen as one of the main reasons that EGP could not establish a globally acceptable routing metric: again this seems to be a clearly futile aim from this distance in time! Consequently EGP became effectively a rudimentary path-vector protocol which linked gateways with Autonomous Systems. It was totally reliant on the tree structured network to avoid routing loops and the all informed nature of EGP meant that update packets became very large. BGP version 1 [RFC1105] (Lougheed, K. and J. Rekhter, “Border Gateway Protocol (BGP),” June 1989.) was standardized in 1989 but had been in development for some time before this and had already seen action in production networks prior to standardization. BGP was the first real path-vector routing protocol and was intended to relieve some of the scaling problems as well as providing policy-based routing. Routes were described as paths along a 'vector' of ASs without any associated cost metric. This way of describing routes was explicitly intended to allow detection of routing loops. It was assumed that the intra-domain routing system was loop-free with the implication that the total routing system would be loop-free if there were no loops in the AS path. Note that there were no theoretical underpinnings for this work and it traded freedom from routing loops for guaranteed convergence.
Also the NSFnet was a government funded research and education network. Commercial companies which were partners in some of the projects were using the NSFnet for their research activities but it was becoming clear that these companies also needed networks for commercial traffic. NSFnet had put in place "acceptable use" policies which were intended to limit the use of the network. However there was little or no technology to support the legal framework.
Practical experience, IETF IAB discussion (centred in the Internet Architecture Task Force) and the OSI theoretical work were by now coming to the same conclusions:

*Networks were going to be composed out of multiple administrative domains (the federated network),

*The connections between these domains would be an arbitrary graph and certainly not a tree,

*The administrative domains would wish to establish distinctive, independent routing policies through the graph of Autonomous Systems, and

*Administrative Domains would have a degree of distrust of each other which would mean that policies would remain opaque.

These views were reflected by Susan Hares' (working for Merit Networks at that time) contribution to the Internet Architecture (INARC) workshop in 1989, summarized in the report of the workshop [INARC89] (Mills, D., Ed. and M. Davis, Ed., “Internet Architecture Workshop: Future of the Internet System Architecture and TCP/IP Protocols - Report,” 1990.):

> The rich interconnectivity within the Internet causes routing problems today. However, the presenter believes the problem is not the high degree of interconnection, but the routing protocols and models upon which these protocols are based. Rich interconnectivity can provide redundancy which can help packets moving even through periods of outages. Our model of interdomain routing needs to change. The model of autonomous confederations and autonomous systems [RFC0975] (Mills, D., “Autonomous confederations,” February 1986.) no longer fits the reality of many regional networks. The ISO models of administrative domain and routing domains better fit the current Internet’s routing structure.
>
> With the first NSFNET backbone, NSF assumed that the Internet would be used as a production network for research traffic. We cannot stop these networks for a month and install all new routing protocols. The Internet will need to evolve its changes to networking protocols while still continuing to serve its users. This reality colors how plans are made to change routing protocols.

It is also interesting to note that the difficulties of organising a transition were recognized at this stage and have not been seriously explored or resolved since.
Policies would primarily be interested in controlling which traffic should be allowed to transit a domain (to satisfy commercial constraints or acceptable use policies) thereby controlling which traffic uses the resources of the domain. The solution adopted by both the IETF and OSI was a form of distance vector hop-by-hop routing with explicit policy terms. The reasoning for this choice can be found in Breslau and Estrin's 1990 paper [Breslau90] (Breslau, L. and D. Estrin, “An Architecture for Network-Layer Routing in OSI,” 1990.) (implicitly - because some other alternatives are given such as a link state with policy suggestion which, with hindsight, would have even greater problems than BGP on a global scale network). Traditional distance vector protocols exchanged routing information in the form of a destination and a metric. The new protocols explicitly associated policy expressions with the route by including either a list of the

source ASs that are permitted to use the route described in the routing update, and/or a list of all ASs traversed along the advertised route. Parallel protocol developments were already in progress by the time this paper was published: BGP version 2 [RFC1163] (Lougheed, K. and Y. Rekhter, “Border Gateway Protocol (BGP),” June 1990.) in the IETF and the Inter-Domain Routing Protocol (IDRP) [ISO10747] (ISO/IEC, “Protocol for Exchange of Inter-Domain Routeing Information among Intermediate Systems to support Forwarding of ISO 8473 PDUs,” 1993.) which would be the Level 3 routing protocol for the OSI architecture. IDRP was developed under the aegis of the ANSI XS3.3 working group led by Lyman Chapin and Charles Kunzinger. The two protocols were very similar in basic design but IDRP has some extra features, some of which have been incorporated into later versions of BGP; others may yet be so and still others may be seen to be inappropriate. Breslau and Estrin summarize the design of IDRP as follows:

> IDRP attempts to solve the looping and convergence problems inherent in distance vector routing by including full AD [Administrative Domain - essentially the equivalent of what are now called ASs] path information in routing updates. Each routing update includes the set of ADs that must be traversed in order to reach the specified destination. In this way, routes that contain AD loops can be avoided.
>
> IDRP updates also contain additional information relevant to policy constraints. For instance, these updates can specify what other ADs are allowed to receive the information described in the update. In this way, IDRP is able to express source specific policies. The IDRP protocol also provides the structure for the addition of other types of policy related information in routing updates. For example, User Class Identifiers (UCI) could also be included as policy attributes in routing updates.
>
> Using the policy route attributes IDRP provides the framework for expressing more fine grained policy in routing decisions. However, because it uses hop-by-hop distance vector routing, it only allows a single route to each destination per-QOS to be advertised. As the policy attributes associated with routes become more fine grained, advertised routes will be applicable to fewer sources. This implies a need for multiple routes to be advertised for each destination in order to increase the probability that sources have acceptable routes available to them. This effectively replicates the routing table per forwarding entity for each QoS, UCI, source combination that might appear in a packet. Consequently, we claim that this approach does not scale well as policies become more fine grained, i.e., source or UCI specific policies.

Over the next three or four years successive versions of BGP (BGP-2 [RFC1163] (Lougheed, K. and Y. Rekhter, “Border Gateway Protocol (BGP),” June 1990.), BGP-3 [RFC1267] (Lougheed, K. and Y. Rekhter, “Border Gateway Protocol 3 (BGP-3),” October 1991.) and BGP-4 [RFC1771] (Rekhter, Y. and T. Li, “A Border Gateway Protocol 4 (BGP-4),” March 1995.)) were deployed to cope with the growing and by now commercialized Internet. From BGP-2 onwards, BGP made no assumptions about an overall structure of interconnections allowing it to cope with today's dense web of interconnections between ASs. BGP version 4 was developed to handle the change from classful to classless addressing. For most of this time IDRP was being developed in parallel, and both protocols were implemented in the Merit gatedaemon routing protocol suite. During this time there was a movement within the IETF which saw BGP as a stopgap measure to be used until the more sophisticated IDRP could be adapted to run over IP instead of the OSI connectionless protocol CLNP. However, unlike its intra-domain counterpart IS-IS which has stood the test of time, and indeed proved to be more flexible than OSPF, IDRP was ultimately not adopted by the market. By the time the NSFnet backbone was decommissioned in 1995, BGP-4 was the inter-domain routing protocol of choice and OSI's star was already beginning to wane. IDRP is now little remembered.

A more complete account of the capabilities of IDRP can be found in chapter 14 of David Piscitello and Lyman Chapin's book 'Open Systems Networking: TCP/IP and OSI' which is now readable on the Internet [Chapin94] (Piscitello, D. and A. Chapin, “Open Systems Networking: TCP/IP & OSI,” 1994.).

IDRP also contained quite extensive means for securing routing exchanges much of it based on X.509 certificates for each router and public/private key encryption of routing updates.

Some of the capabilities of IDRP which might yet appear in a future version of BGP include the ability to manage routes with explicit QoS classes, and the concept of domain confederations (somewhat different from the confederation mechanism in today's BGP) as an extra level in the hierarchy of routing.

3.3. Nimrod Requirements

TOC

Nimrod as expressed by Noel Chiappa in his early document, "A New IP Routing and Addressing Architecture" [Chiappa91] (Chiappa, N., “A New IP Routing and Addressing Architecture,” 1991.) and later in the NIMROD Working Group documents [RFC1753] (Chiappa, J., “IPng Technical Requirements Of the Nimrod Routing and Addressing Architecture,” December 1994.) and [RFC1992] (Castineyra, I., Chiappa, N., and M. Steenstrup, “The Nimrod Routing Architecture,” August 1996.) established a number of requirements that need to be considered by any

new routing architecture. The Nimrod requirements took RFC1126 as a starting point and went further.
The three goals of Nimrod, quoted from [RFC1992] (Castineyra, I., Chiappa, N., and M. Steenstrup, "The Nimrod Routing Architecture," August 1996.), were as follows:

1. To support a dynamic internetwork of *arbitrary size* (our emphasis) by providing mechanisms to control the amount of routing information that must be known throughout an internetwork.

2. To provide service-specific routing in the presence of multiple constraints imposed by service providers and users.

3. To admit incremental deployment throughout an internetwork.

It is certain that these goals should be considered requirements for any new domain-based routing architecture.

*As discussed in other sections of this document the rate of growth of the amount of information needed to maintain the routing system is such that the system may not be able to scale up as the Internet expands as foreseen. And yet, as the services and constraints upon those services grow there is a need for more information to be maintained by the routing system. One of the key terms in the first requirements is 'control'. While increasing amounts of information need to be known and maintained in the Internet, the amounts and kinds of information that are distributed can be controlled. This goal should be reflected in the requirements for the future domain-based architecture.

*If anything, the demand for specific services in the Internet has grown since 1996 when the Nimrod architecture was published. Additionally the kinds of constraints that service providers need to impose upon their networks and that services need to impose upon the routing have also increased. Any changes made to the network in the last half-decade have not significantly improved this situation.

*The ability to incrementally deploy any new routing architecture within the Internet is still an absolute necessity. It is impossible to imagine that a new routing architecture could supplant the current architecture on a flag day.

At one point in time Nimrod, with its addressing and routing architectures was seen as a candidate for IPng. History shows that it was not accepted as the IPng, having been ruled out of the selection process by the IESG in 1994 on the grounds that it was 'too much of a research effort' [RFC1752] (Bradner, S. and A. Mankin, "The Recommendation for the IP Next Generation Protocol," January 1995.),

although input for the requirements of IPng was explicitly solicited from Chiappa [RFC1753] (Chiappa, J., “IPng Technical Requirements Of the Nimrod Routing and Addressing Architecture,” December 1994.).
Instead IPv6 has been put forth as the IPng. Without entering a discussion of the relative merits of IPv6 versus Nimrod, it is apparent that IPv6, while it may solve many problems, does not solve the critical routing problems in the Internet today. In fact in some sense it exacerbates them by adding a requirement for support of two Internet protocols and their respective addressing methods. In many ways the addition of IPv6 to the mix of methods in today's Internet only points to the fact that the goals, as set forth by the Nimrod team, remain as necessary goals.
There is another sense in which study of Nimrod and its architecture may be important to deriving a future domain-based routing architecture. Nimrod can be said to have two derivatives:

- Multi-Protocol Label Switching (MPLS) in that it took the notion of forwarding along well known paths

- Private Network-Node Interface (PNNI) in that it took the notion of abstracting topological information and using that information to create connections for traffic.

It is important to note, that whilst MPLS and PNNI borrowed ideas from Nimrod, neither of them can be said to be an implementation of this architecture.

3.4. PNNI

TOC

The Private Network-Node Interface (PNNI) routing protocol was developed under the ATM Forum's auspices as a hierarchical route determination protocol for ATM, a connection oriented architecture. It is reputed to have developed several of its methods from a study of the Nimrod architecture. What can be gained from an analysis of what did and did not succeed in PNNI?
The PNNI protocol includes the assumption that all peer groups are willing to cooperate, and that the entire network is under the same top administration. Are there limitations that stem from this 'world node' presupposition? As discussed in [RFC3221] (Huston, G., “Commentary on Inter-Domain Routing in the Internet,” December 2001.), the Internet is no longer a clean hierarchy and there is a lot of resistance to having any sort of 'ultimate authority' controlling or even brokering communication.
PNNI is the first deployed example of a routing protocol that uses abstract map exchange (as opposed to distance vector or link state mechanisms) for inter-domain routing information exchange. One consequence of this is that domains need not all use the same mechanism

for map creation. What were the results of this abstraction and source based route calculation mechanism?
Since the authors of this document do not have experience running a PNNI network, the comments above are from a theoretical perspective. Further research on these issues based on operational experience is required.

---

4. Recent Research Work

TOC

4.1. Developments in Internet Connectivity

TOC

The work commissioned from Geoff Huston by the Internet Architecture Board [RFC3221] (Huston, G., “Commentary on Inter-Domain Routing in the Internet,” December 2001.) draws a number of conclusions from analysis of BGP routing tables and routing registry databases:

*The connectivity between provider ASs is becoming more like a dense mesh than the tree structure that was commonly assumed to be commonplace a couple of years ago. This has been driven by the increasing amounts charged for peering and transit traffic by global service providers. Local direct peering and Internet exchanges are becoming steadily more common as the cost of local fibre connections drops.

*End user sites are increasingly resorting to multi-homing onto two or more service providers as a way of improving resiliency. This has a knock-on effect of spectacularly fast depletion of the available pool of AS numbers as end user sites require public AS numbers to become multi-homed and corresponding increase in the number of prefixes advertised in BGP.

*Multi-homed sites are using advertisement of longer prefixes in BGP as a means of traffic engineering to load spread across their multiple external connections with further impact on the size of the BGP tables.

*Operational practices are not uniform, and in some cases lack of knowledge or training is leading to instability and/or excessive advertisement of routes by incorrectly configured BGP speakers.

*All these factors are quickly negating the advantages in limiting the expansion of BGP routing tables that were gained by the introduction of CIDR and consequent prefix aggregation in BGP. It

is also now impossible for IPv6 to realize the world view in which the default free zone would be limited to perhaps 10,000 prefixes.

*The typical 'width' of the Internet in AS hops is now around five, and much less in many cases.

These conclusions have a considerable impact on the requirements for the future domain-based routing architecture:

*Topological hierarchy (e.g. mandating a tree structured connectivity) cannot be relied upon to deliver scalability of a large Internet routing system

*Aggregation cannot be relied upon to constrain the size of routing tables for an all-informed routing system

---

4.2. DARPA NewArch Project

TOC

DARPA funded a project to think about a new architecture for future generation Internet, called NewArch (http://www.isi.edu/newarch/). Work started in the first half of 2000 and the main project finished in 2003 [NewArch03] (Clark, D., Sollins, K., Wroclawski, J., Katabi, D., Kulik, J., Yang, X., Braden, R., Faber, T., Falk, A., Pingali, V., Handley, M., and N. Chiappa, “New Arch: Future Generation Internet Architecture,” December 2003.).
The main development is to conclude that as the Internet becomes mainstream infrastructure, fewer and fewer of the requirements are truly global but may apply with different force or not at all in certain parts of the network. This (it is claimed) makes the compilation of a single, ordered list of requirements deeply problematic. Instead we may have to produce multiple requirement sets with support for differing requirement importance at different times and in different places. This 'meta-requirement' significantly impacts architectural design.
Potential new technical requirements identified so far include:

*Commercial environment concerns such as richer inter-provider policy controls and support for a variety of payment models

*Trustworthiness

*Ubiquitous mobility

*Policy driven self-organisation ('deep auto configuration')

*Extreme short-timescale resource variability

*Capacity allocation mechanisms

*Speed, propagation delay and delay/bandwidth product issues

Non-technical or political 'requirements' include:

*Legal and Policy drivers such as

-Privacy and free/anonymous speech

-Intellectual property concerns

-Encryption export controls

-Law enforcement surveillance regulations

-Charging and taxation issues

*Reconciling national variations and consistent operation in a world wide infrastructure

The conclusions of the work are now summarized in the final report [NewArch03] (Clark, D., Sollins, K., Wroclawski, J., Katabi, D., Kulik, J., Yang, X., Braden, R., Faber, T., Falk, A., Pingali, V., Handley, M., and N. Chiappa, “New Arch: Future Generation Internet Architecture,” December 2003.).

4.2.1. Defending the End-to-End Principle

TOC

One of the participants in DARPA NewArch work (Dave Clark) with one of his associates has also published a very interesting paper analyzing the impact of some of the new requirements identified in NewArch (see Section 4.2 (DARPA NewArch Project)) on the end-to-end principle that has guided the development of the Internet to date [Blumenthal01] (Blumenthal, M. and D. Clark, “Rethinking the design of the Internet: The end to end arguments vs,” May 2001.). Their primary conclusion is that the loss of trust between the users at the ends of end to end has the most fundamental effect on the Internet. This is clear in the context of the routing system, where operators are unwilling to reveal the inner workings of their networks for commercial reasons. Similarly, trusted third parties and their avatars (mainly mid-boxes of one sort or another) have a major impact on the end-to-end principles and the routing mechanisms that went with them. Overall, the end to end principles should be defended so far as is possible - some changes are already too deeply embedded to make it possible to go back to full

trust and openness - at least partly as a means of staving off the day when the network will ossify into an unchangeable form and function (much as the telephone network has done). The hope is that by that time a new Internet will appear to offer a context for unfettered innovation.

---

5. Existing problems of BGP and the current Inter-/Intra-Domain Architecture

TOC

Although most of the people who have to work with BGP today believe it to be a useful, working protocol, discussions have brought to light a number of areas where BGP or the relationship between BGP and the intra-domain routing protocols in use today could be improved. BGP-4 has been and continues to be extended since it was originally introduced in [RFC1771] (Rekhter, Y. and T. Li, “A Border Gateway Protocol 4 (BGP-4),” March 1995.) and the protocol as deployed has been documented in [RFC4271] (Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” January 2006.). This section is, to a large extent, a wish list for the future domain-based routing architecture based on those areas where BGP is seen to be lacking, rather than simply a list of problems with BGP. The shortcomings of today's inter-domain routing system have also been extensively surveyed in 'Architectural Requirements for Inter-Domain Routing in the Internet' [RFC3221] (Huston, G., “Commentary on Inter-Domain Routing in the Internet,” December 2001.), particularly with respect to its stability and the problems produced by explosions in the size of the Internet.

5.1. BGP and Auto-aggregation

TOC

The initial stability followed by linear growth rates of the number of routing objects (prefixes) that was achieved by the introduction of CIDR around 1994, has now been once again been replaced by near-exponential growth of number of routing objects. The granularity of many of the objects advertised in the default free zone is very small (prefix length of 22 or longer): This granularity appears to be a by-product of attempts to perform precision traffic engineering related to increasing levels of multi-homing. At present there is no mechanism in BGP that would allow an AS to aggregate such prefixes without advance knowledge of their existence, even if it was possible to deduce automatically that they could be aggregated. Achieving satisfactory auto-aggregation would also significantly reduce the non-locality problems associated with instability in peripheral ASs.

On the other hand, it may be that alterations to the connectivity of the net as described in [RFC3221] (Huston, G., “Commentary on Inter-Domain Routing in the Internet,” December 2001.) and Section 2.5.1 may limit the usefulness of auto-aggregation.

5.2. Convergence and Recovery Issues

TOC

BGP today is a stable protocol under most circumstances but this has been achieved at the expense of making the convergence time of the inter-domain routing system very slow under some conditions. This has a detrimental effect on the recovery of the network from failures.
The timers that control the behavior of BGP are typically set to values in the region of several tens of seconds to a few minutes, which constrains the responsiveness of BGP to failure conditions.
In the early days of deployment of BGP, poor network stability and router software problems lead to storms of withdrawals closely followed by re-advertisements of many prefices. To control the load on routing software imposed by these "route flaps", route flap damping was introduced into BGP. Most operators have now implemented a degree of route flap damping in their deployments of BGP. This restricts the number of times that the routing tables will be rebuilt even if a route is going up and down very frequently. Unfortunately, route flap damping responds to multiple flaps by increasing the route suppression time exponentially, which can result in some parts of the Internet being unreachable for hours at a time.
There is evidence ([RFC3221] (Huston, G., “Commentary on Inter-Domain Routing in the Internet,” December 2001.) and measurements by some of the Sub-group B members[Jiang02] (Jiang, Y., Doria, A., Olsson, D., and F. Pettersson, “Inter-domain Routing Stability Measurement,” 2002.)) that in today's network route flap is disproportionately associated with the fine grain prefices (length 22 or longer) associated with traffic engineering at the periphery of the network. Auto-aggregation as previously discussed would tend to mask such instability and prevent it being propagated across the whole network. Another question that needs to be studied is the continuing need for an architecture that requires global convergence. Some of our studies (unpublished) show that, in some localities at least, the network never actually reaches stability; i.e., it never really globally converges. Can a global, and beyond, network be designed with the requirement of global convergence?

TOC

5.3. Non-locality of Effects of Instability and Misconfiguration

There have been a number of instances, some of which are well documented, of a mistake in BGP configuration in a single peripheral AS propagating across the whole Internet and resulting in misrouting of most of the traffic in the Internet.
Similarly, a single route flap in a single peripheral AS can require route table recalculation across the entire Internet.
This non-locality of effects is highly undesirable, and it would be a considerable improvement if such effects were naturally limited to a small area of the network around the problem. This is another argument for an architecture that does not require global convergence.

5.4. Multihoming Issues

TOC

As discussed previously, the increasing use of multi-homing as a robustness technique by peripheral networks requires that multiple routes have to be advertised for such domains. These routes must not be aggregated close in to the multi-homed domain as this would defeat the traffic engineering implied by multi-homing and currently cannot be aggregated further away from the multi-homed domain due to the lack of auto-aggregation capabilities. Consequentially the default free zone routing table is growing exponentially, as it was before CIDR.
The longest prefix match routing technique introduced by CIDR, and implemented in BGP-4, when combined with provider address allocation is an obstacle to effective multi-homing if load sharing across the multiple links is required. If an AS has been allocated its addresses from an upstream provider, the upstream provider can aggregate those addresses with those of other customers and need only advertise a single prefix for a range of customers. But, if the customer AS is also connected to another provider, the second provider is not able to aggregate the customer addresses because they are not taken from his allocation, and will therefore have to announce a more specific route to the customer AS. The longest match rule will then direct all traffic through the second provider, which is not as required.

Example:

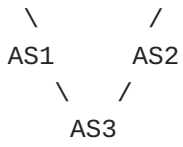

Figure 1: Address Aggregation

In Figure 1 (Address Aggregation) AS3 has received its addresses from AS1, which means AS1 can aggregate. But if AS3 wants its traffic to be seen equally both ways, AS3 is forced to announce both the aggregate and the more specific route to AS2.
This problem has induced many ASs to apply for their own address allocation even though they could have been allocated from an upstream provider further exacerbating the default free zone route table size explosion. This problem also interferes with the desire of many providers in the default free zone to route only prefixes that are equal to or shorter than 20 or 19 bits.
Note that some problems which are referred to as multihoming issues are not, and should not be, solvable through the routing system (e.g., where a TCP load distributor is needed), and multihoming is not a panacea for the general problem of robustness in a routing system [I-D.berkowitz-multireq] (Berkowitz, H. and D. Krioukov, “To Be Multihomed: Requirements and Definitions,” 2001.).

> Editors' Note: A more recent analysis of multihoming can be found in [RFC4116] (Abley, J., Lindqvist, K., Davies, E., Black, B., and V. Gill, “IPv4 Multihoming Practices and Limitations,” July 2005.).

5.5. AS-number exhaustion

TOC

The domain identifier or AS-number is a 16-bit number. When this paper was originally written in 2001, allocation of AS-numbers was increasing 51% a year [RFC3221] (Huston, G., “Commentary on Inter-Domain Routing in the Internet,” December 2001.) and exhaustion by 2005 was predicted. According to some recent work again by Huston [Huston05] (Huston, G., “Exploring Autonomous System Numbers,” August 2005.), the rate of increase dropped off after the business downturn but as of July 2005, well over half the available AS numbers (39000 out of 64510) had been allocated by IANA and around 20000 were visible in the global BGP routing tables. A year later these figures had grown to 42000 (April 2006) and 23000 (August 2006) respectively and the rate of allocation is currently about 3500 per year. Depending on the curve fitting model used to predict when exhaustion will occur, the pool will run out somewhere between 2010 and 2013. There appear to be other factors at work in this rate of increase beyond an increase in the number of ISPs in business, although there is a fair degree of correlation between these numbers. AS numbers are now used for a number of purposes beyond that of identifying large routing domains: multihomed sites acquire an AS number in order to express routing preferences to their various providers and AS numbers are used part of the addressing mechanism for

MPLS/BGP-based virtual private networks (VPNs) [RFC4364] (Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)," February 2006.). The IETF has had a proposal under development for over four years to increase the available range of AS-numbers to 32 bits [RFC4893] (Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space," May 2007.). Much of the slowness in development is due to the deployment challenge during transition. Because of the difficulties of transition, deployment needs to start well in advance of actual exhaustion so that the network as a whole is ready for the new capability when it is needed. This implies that standardisation needs to be complete and implementations available at least well in advance of expected exhaustion so that deployment of upgrades that can handle the longer AS numbers should be starting around 2008 to give a reasonable expectation that the change has been rolled out across a large fraction of the Internet by the time exhaustion occurs.

> Editors' Note: The RIRs are planning to move to assignment of the longer AS numbers by default on 1 January 2009, but there are concerns that significant numbers of routers will not have been upgraded by then.

5.6. Partitioned ASs

TOC

Tricks with discontinuous ASs are used by operators, for example, to implement anycast. Discontinuous ASs may also come into being by chance if a multi-homed domain becomes partitioned as a result of a fault and part of the domain can access the Internet through each connection. It may be desirable to make support for this kind of situation more transparent than it is at present.

5.7. Load Sharing

TOC

Load splitting or sharing was not a goal of the original designers of BGP and it is now a problem for today's network designers and managers. Trying to fool BGP into load sharing between several links is a constantly recurring exercise for most operators today.

TOC

5.8. Hold down issues

As with the interval between 'hello' messages in OSPF, the typical size and defined granularity (seconds to tens of seconds) of the 'keep-alive' time negotiated at start-up for each BGP connection constrains the responsiveness of BGP to link failures.
The recommended values and the available lower limit for this timer were set to limit the overhead caused by keep-alive messages when link bandwidths were typically much lower than today. Analysis and experiment ([I-D.alaettinoglu-isis-convergence] (Alaettinoglu, C., Jacobson, V., and H. Yu, “Towards Milli-Second IGP Convergence,” Nov 2000.), [I-D.sandiick-flip] (Sandick, H., Squire, M., Cain, B., Duncan, I., and B. Haberman, “Fast LIveness Protocol (FLIP),” Feb 2000.) and [RFC4204] (Lang, J., “Link Management Protocol (LMP),” October 2005.)) indicate that faster links could sustain a much higher rate of keep-alive messages without significantly impacting normal data traffic. This would improve responsiveness to link and node failures but with a corresponding increase in the risk of instability, if the error characteristics of the link are not taken properly into account when setting the keep-alive interval.

> Editors' Note: A 'fast' liveness protocol has been standardized as [I-D.ietf-bfd-base] (Katz, D. and D. Ward, “Bidirectional Forwarding Detection,” January 2010.).

An additional problem with the hold-down mechanism in BGP is the amount of information that has to be exchanged to re-establish the database of route advertisements on each side of the link when it is re-established after a failure. Currently any failure, however brief forces a full exchange which could perhaps be constrained by retaining some state across limited time failures and using revision control, transaction and replication techniques to resynchronise the databases. Various techniques have been implemented to try to reduce this problem but they have not yet been standardised.

5.9. Interaction between Inter-Domain Routing and Intra-Domain Routing

TOC

Today, many operators' backbone routers run both I-BGP and an intra-domain protocol to maintain the routes that reach between the borders of the domain. Exporting routes from BGP into the intra-domain protocol in use and bringing them back up to BGP is not recommended [RFC2791] (Yu, J., “Scalable Routing Design Principles,” July 2000.), but it is still necessary for all backbone routers to run both protocols. BGP is used to find the egress point and intra-domain protocol to find the

path (next hop router) to the egress point across the domain. This is not only a management problem but may also create other problems:

*BGP is a path vector protocol (i.e., a protocol that uses distance metrics possibly overridden by policy metrics), whereas most intra-domain protocols are link state protocols. As such BGP is not optimised for convergence speed although distance vector algorithms generally require less processing power. Incidentally, more efficient distance vector algorithms are available such as [Xu97] (Xu, Z., Dai, S., and J. Garcia-Luna-Aceves, “A More Efficient Distance Vector Routing Algorithm,” Nov 1997.).

*The metrics used in BGP and the intra-domain protocol are rarely comparable or combinable. Whilst there are arguments that the optimizations inside a domain may be different from those for end-to-end paths, there are occasions, such as calculating the 'topologically nearest' server when computable or combinable metrics would be of assistance.

*The policies that can be implemented using BGP are designed for control of traffic exchange between operators, not for controlling paths within a domain. Policies for BGP are most conveniently expressed in Routing Policy Support Language (RPSL) [RFC2622] (Alaettinoglu, C., Villamizar, C., Gerich, E., Kessens, D., Meyer, D., Bates, T., Karrenberg, D., and M. Terpstra, “Routing Policy Specification Language (RPSL),” June 1999.) and this could be extended if thought desirable to include additional policy information.

*If the NEXT HOP destination for a set of BGP routes becomes inaccessible because of intra-domain protocol problems, the routes using the vanished next hop have to be invalidated at the next available UPDATE. Subsequently, if the next hop route reappears, this would normally lead to the BGP speaker requesting a full table from its neighbour(s). Current implementations may attempt to circumvent the effects of intra-domain protocol route flap by caching the invalid routes for a period in case the next hop is restored through the 'graceful restart' mechanism.

-Editors' Note: This was standardized as [RFC4724] (Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, “Graceful Restart Mechanism for BGP,” January 2007.).

*Synchronization between intra-domain and inter-domain routing information is a problem as long as we use different protocols for intra-domain and inter-domain routing, which will most

probably be the case even in the future because of the differing requirements in the two situations. Some sort of synchronization between those two protocols would be useful. In the RFC 'IS-IS Transient Blackhole Avoidance' [RFC3277] (McPherson, D., "Intermediate System to Intermediate System (IS-IS) Transient Blackhole Avoidance," April 2002.), the intra-domain protocol side of the story is covered (there is an equivalent discussion for OSPF).

*Synchronizing in BGP means waiting for the intra-domain protocol to know about the same networks as the inter-domain protocol, which can take a significant period of time and slows down the convergence of BGP by adding the intra-domain protocol convergence time into each cycle. In general operators no longer attempt full synchronization in order to avoid this problem (in general, redistributing the entire BGP routing feed into the local intra-domain protocol is unnecessary and undesirable but where a domain has multiple exits to peers and other non-customer networks, changes in BGP routing that affect the exit taken by traffic require corresponding re-routing in the intra-domain routing).

5.10. Policy Issues

TOC

There are several classes of issues with current BGP policy:

*Policy is installed in an ad-hoc manner in each autonomous system. There isn't a method for ensuring that the policy installed in one router is coherent with policies installed in other routers.

*As described in Griffin [Griffin99] (Griffin, T. and G. Wilfong, "An Analysis of BGP Convergence Properties," 1999.) and in McPherson [RFC3345] (McPherson, D., Gill, V., Walton, D., and A. Retana, "Border Gateway Protocol (BGP) Persistent Route Oscillation Condition," August 2002.) it is possible to create policies for ASs, and instantiate them in routers, that will cause BGP to fail to converge in certain types of topology

*There is no available network model for describing policy in a coherent manner.

Policy management is extremely complex and mostly done without the aid of any automated procedures. The extreme complexity means that a highly qualified specialist is required for policy management of border routers. The training of these specialists is quite lengthy and needs

to involve long periods of hands-on experience. There is, therefore, a shortage of qualified staff for installing and maintaining the routing policies. Because of the overall complexity of BGP, policy management tends to be only a relatively small topic within a complete BGP training course and specialised policy management training courses are not generally available.

5.11. Security Issues

TOC

While many of the issues with BGP security have been traced either to implementation issues or to operational issues, BGP is vulnerable to Distributed Denial of Service (DDoS) attacks. Additionally routers can be used as unwitting forwarders in DDoS attacks on other systems. Though DDoS attacks can be fought in a variety of ways, mostly using filtering methods, it takes constant vigilance. There is nothing in the current architecture or in the protocols that serves to protect the forwarders from these attacks.

> Editors' Note: Since the original draft was written, the issue of inter-domain routing security has been studied in much greater depth. The rpsec working group has gone into the security issues in great detail [RFC4593] (Barbir, A., Murphy, S., and Y. Yang, "Generic Threats to Routing Protocols," October 2006.) and readers should refer to that work to understand the security issues.

5.12. Support of MPLS and VPNS

TOC

Recently BGP has been modified to function as a signaling protocol for MPLS and for VPNs [RFC4364] (Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)," February 2006.). Some people see this over-loading of the BGP protocol as a boon whilst others see it as a problem. While it was certainly convenient as a vehicle for vendors to deliver extra functionality for to their products, it has exacerbated some of the performance and complexity issues of BGP. Two important problems are, the additional state that must be retained and refreshed to support VPN (Virtual Private Network) tunnels and that BGP does not provide end-to-end notification making it difficult to confirm that all necessary state has been installed or updated.
It is an open question whether VPN signaling protocols should remain separate from the route determination protocols.

5.13. IPv4 / IPv6 Ships in the Night

TOC

The fact that service providers need to maintain two completely separate networks, one for IPv4 and one for IPv6, has been a real hindrance to the introduction of IPv6. When IPv6 does get widely deployed it will do so without causing the disappearance of IPv4. This means that unless something is done, service providers would need to maintain the two networks in perpetuity (at least on the foreshortened timescale which the Internet world uses).

It is possible to use a single set of BGP speakers with multiprotocol extensions [RFC4760] (Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4," January 2007.) to exchange information about both IPv4 and IPv6 routes between domains, but the use of TCP as the transport protocol for the information exchange results in an asymmetry when choosing to use one of TCP over IPv4 or TCP over IPv6. Successful information exchange confirms one of IPv4 or IPv6 reachability between the speakers but not the other, making it possible that reachability is being advertised for a protocol for which it is not present.

Also, current implementations do not allow a route to be advertised for both IPv4 and IPv6 in the same UPDATE message, because it is not possible to explicitly link the reachability information for an address family to the corresponding next hop information. This could be improved, but currently results in independent UPDATEs being exchanged for each address family.

5.14. Existing Tools to Support Effective Deployment of Inter-Domain Routing

TOC

The tools available to network operators to assist in configuring and maintaining effective inter-domain routing in line with their defined policies are limited, and almost entirely passive.

- There are no tools to facilitate the planning of the routing of a domain (either intra- or inter-domain); there are a limited number of display tools that will visualize the routing once it has been configured.

- There are no tools to assist in converting business policy specifications into the Routing Policy Specification Language (RPSL) language (see Section 5.14.1 (Routing Policy Specification Language RPSL (RFC 2622, 2650) and RIPE NCC Database (RIPE 157))); there are limited tools to convert the RPSL into BGP commands and to check, post-facto, that the proposed policies are consistent with the policies in adjacent domains (always provided that these have been revealed and accurately documented).

*There are no tools to monitor BGP route changes in real time and warn the operator about policy inconsistencies and/or instabilities.

The following section summarises the tools that are available to assist with the use of RPSL. Note they are all batch mode tools used off-line from a real network. These tools will provide checks for skilled inter-domain routing configurers but limited assistance for the novice.

5.14.1. Routing Policy Specification Language RPSL (RFC 2622, 2650) and RIPE NCC Database (RIPE 157)

TOC

Routing Policy Specification Language (RPSL) [RFC2622] (Alaettinoglu, C., Villamizar, C., Gerich, E., Kessens, D., Meyer, D., Bates, T., Karrenberg, D., and M. Terpstra, "Routing Policy Specification Language (RPSL)," June 1999.) enables a network operator to describe routes, routers and autonomous systems ASs that are connected to the local AS. Using the RPSL language (see [RFC2650] (Meyer, D., Schmitz, J., Orange, C., Prior, M., and C. Alaettinoglu, "Using RPSL in Practice," August 1999.)) a distributed database is created to describe routing policies in the Internet as described by each AS independently. The database can be used to check the consistency of routing policies stored in the database.
Tools exist ([IRRToolSet] (Internet Systems Consortium, "Internet Routing Registry Toolset Project," 2006.)) that can use the database to (among other things)

*Flag when two neighboring network operators specify conflicting or inconsistent routing information exchanges with each other and also detect global inconsistencies where possible;

*Extract all AS-paths between two networks that are allowed by routing policy from the routing policy database; display the connectivity a given network has according to current policies.

The database queries enable a partial static solution to the convergence problem. They analyze routing policies of a very limited part of Internet and verify that they do not contain conflicts that could lead to protocol divergence. The static analysis of convergence of the entire system has exponential time complexity, so approximation algorithms would have to be used.
The toolset also allows router configurations to be generated from RPSL specifications.

> Editors' Note: The "Internet Routing Registry Toolset" was originally developed by the University of Southern California's Information Sciences Institute (ISI) between 1997 and 2001 as the

> "Routing Arbiter ToolSet" (RAToolSet) project. The toolset is no longer developed by ISI but is used worldwide, so after a period of improvement by RIPE NCC it has now been transferred to the Internet Systems Consortium (ISC) for ongoing maintenance as a public resource.

---

6. Security Considerations

TOC

As this is an informational draft on the history of requirements in IDR and on the problems facing the current Internet IDR architecture, it does not as such create any security problems. On the other hand, some of the problems with today's Internet routing architecture do create security problems and these have been discussed in the text above.

---

7. IANA Considerations

TOC

This document does not request any actions by IANA.
RFC Editor: Please remove this section before publication.

---

8. Acknowledgments

TOC

The draft is derived from work originally produced by Babylon. Babylon was a loose association of individuals from academia, service providers and vendors whose goal was to discuss issues in Internet routing with the intention of finding solutions for those problems.
The individual members who contributed materially to this draft are: Anders Bergsten, Howard Berkowitz, Malin Carlzon, Lenka Carr Motyckova, Elwyn Davies, Avri Doria, Pierre Fransson, Yong Jiang, Dmitri Krioukov, Tove Madsen, Olle Pers, and Olov Schelen.
Thanks also go to the members of Babylon and others who did substantial reviews of this material. Specifically we would like to acknowledge the helpful comments and suggestions of the following individuals: Loa Andersson, Tomas Ahlstrom, Erik Aman, Thomas Eriksson, Niklas Borg, Nigel Bragg, Thomas Chmara, Krister Edlund, Owe Grafford, Susan Hares, Torbjorn Lundberg, David McGrew, Jasminko Mulahusic, Florian-Daniel Otel, Bernhard Stockman, Tom Worster, and Roberto Zamparo.
In addition, the authors are indebted to the folks who wrote all the references we have consulted in putting this paper together. This includes not only the references explicitly listed below, but also

those who contributed to the mailing lists we have been participating in for years.
Finally, it is the editors who are responsible for any lack of clarity, any errors, glaring omissions or misunderstandings.

9. Informative References

TOC

[Blumenthal01]	Blumenthal, M. and D. Clark, "Rethinking the design of the Internet: The end to end arguments vs," the brave new world , May 2001.
[Breslau90]	Breslau, L. and D. Estrin, "An Architecture for Network-Layer Routing in OSI," Proceedings of the ACM symposium on Communications architectures & protocols , 1990.
[Chapin94]	Piscitello, D. and A. Chapin, "Open Systems Networking: TCP/IP & OSI," Addison-Wesley Copyright assigned to authors, 1994.
[Chiappa91]	Chiappa, N., "A New IP Routing and Addressing Architecture," draft-chiappa-routing-01.txt (work in progress), 1991.
[Griffin99]	Griffin, T. and G. Wilfong, "An Analysis of BGP Convergence Properties," Association for Computing Machinery Proceedings of SIGCOMM '99, 1999.
[Huitema90]	Huitema, C. and W. Dabbous, "Routeing protocols development in the OSI architecture," Proceedings of ISCIS V Turkey, 1990.
[Huston05]	Huston, G., "Exploring Autonomous System Numbers," The ISP Column , August 2005.
[I-D.alaettinoglu-isis-convergence]	Alaettinoglu, C., Jacobson, V., and H. Yu, "Towards Milli-Second IGP Convergence," draft-alaettinoglu-isis-convergence-00 (work in progress), Nov 2000.
[I-D.berkowitz-multireq]	Berkowitz, H. and D. Krioukov, "To Be Multihomed: Requirements and Definitions," draft-berkowitz-multireq-02 (work in progress), 2001.
[I-D.ietf-bfd-base]	Katz, D. and D. Ward, "Bidirectional Forwarding Detection," draft-ietf-bfd-base-11 (work in progress), January 2010 (TXT).
[I-D.irtf-routing-reqs]	Doria, A., Davies, E., and F. Kastenholz, "A Set of Possible Requirements for a Future Routing Architecture," draft-irtf-routing-reqs-11 (work in progress), February 2009 (TXT).
[I-D.sandiick-flip]	Sandick, H., Squire, M., Cain, B., Duncan, I., and B. Haberman, "Fast LIveness Protocol (FLIP)," draft-sandiick-flip-00 (work in progress), Feb 2000.
[INARC89]	Mills, D., Ed. and M. Davis, Ed., "Internet Architecture Workshop: Future of the Internet System Architecture and TCP/IP Protocols - Report," Internet Architecture Task Force INARC, 1990.

[IRRToolSet]	Internet Systems Consortium, “Internet Routing Registry Toolset Project,” IRR Tool Set Website, 2006.
[ISO10747]	ISO/IEC, “Protocol for Exchange of Inter-Domain Routeing Information among Intermediate Systems to support Forwarding of ISO 8473 PDUs,” International Standard 10747 , 1993.
[Jiang02]	Jiang, Y., Doria, A., Olsson, D., and F. Pettersson, “Inter-domain Routing Stability Measurement,” , 2002.
[Labovitz02]	Labovitz, C., Ahuja, A., Farnam, J., and A. Bose, “Experimental Measurement of Delayed Convergence,” NANOG , 2002.
[NewArch03]	Clark, D., Sollins, K., Wroclawski, J., Katabi, D., Kulik, J., Yang, X., Braden, R., Faber, T., Falk, A., Pingali, V., Handley, M., and N. Chiappa, “New Arch: Future Generation Internet Architecture,” December 2003.
[RFC0904]	Mills, D., “Exterior Gateway Protocol formal specification,” RFC 904, April 1984 (TXT).
[RFC0975]	Mills, D., “Autonomous confederations,” RFC 975, February 1986 (TXT).
[RFC1105]	Lougheed, K. and J. Rekhter, “Border Gateway Protocol (BGP),” RFC 1105, June 1989 (TXT).
[RFC1126]	Little, M., “Goals and functional requirements for inter-autonomous system routing,” RFC 1126, October 1989 (TXT).
[RFC1163]	Lougheed, K. and Y. Rekhter, “Border Gateway Protocol (BGP),” RFC 1163, June 1990 (TXT).
[RFC1267]	Lougheed, K. and Y. Rekhter, “Border Gateway Protocol 3 (BGP-3),” RFC 1267, October 1991 (TXT).
[RFC1752]	Bradner, S. and A. Mankin, “The Recommendation for the IP Next Generation Protocol,” RFC 1752, January 1995 (TXT).
[RFC1753]	Chiappa, J., “IPng Technical Requirements Of the Nimrod Routing and Addressing Architecture,” RFC 1753, December 1994 (TXT).
[RFC1771]	Rekhter, Y. and T. Li, “A Border Gateway Protocol 4 (BGP-4),” RFC 1771, March 1995 (TXT).
[RFC1992]	Castineyra, I., Chiappa, N., and M. Steenstrup, “The Nimrod Routing Architecture,” RFC 1992, August 1996 (TXT).
[RFC2362]	Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., and V. Jacobson, “Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification,” RFC 2362, June 1998 (TXT, HTML, XML).

[RFC2622]	Alaettinoglu, C., Villamizar, C., Gerich, E., Kessens, D., Meyer, D., Bates, T., Karrenberg, D., and M. Terpstra, “Routing Policy Specification Language (RPSL),” RFC 2622, June 1999 (TXT).
[RFC2650]	Meyer, D., Schmitz, J., Orange, C., Prior, M., and C. Alaettinoglu, “Using RPSL in Practice,” RFC 2650, August 1999 (TXT).
[RFC2791]	Yu, J., “Scalable Routing Design Principles,” RFC 2791, July 2000 (TXT).
[RFC3221]	Huston, G., “Commentary on Inter-Domain Routing in the Internet,” RFC 3221, December 2001 (TXT).
[RFC3277]	McPherson, D., “Intermediate System to Intermediate System (IS-IS) Transient Blackhole Avoidance,” RFC 3277, April 2002 (TXT).
[RFC3345]	McPherson, D., Gill, V., Walton, D., and A. Retana, “Border Gateway Protocol (BGP) Persistent Route Oscillation Condition,” RFC 3345, August 2002 (TXT).
[RFC3618]	Fenner, B. and D. Meyer, “Multicast Source Discovery Protocol (MSDP),” RFC 3618, October 2003 (TXT).
[RFC3765]	Huston, G., “NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control,” RFC 3765, April 2004 (TXT).
[RFC3913]	Thaler, D., “Border Gateway Multicast Protocol (BGMP): Protocol Specification,” RFC 3913, September 2004 (TXT).
[RFC4116]	Abley, J., Lindqvist, K., Davies, E., Black, B., and V. Gill, “IPv4 Multihoming Practices and Limitations,” RFC 4116, July 2005 (TXT).
[RFC4204]	Lang, J., “Link Management Protocol (LMP),” RFC 4204, October 2005 (TXT).
[RFC4271]	Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” RFC 4271, January 2006 (TXT).
[RFC4364]	Rosen, E. and Y. Rekhter, “BGP/MPLS IP Virtual Private Networks (VPNs),” RFC 4364, February 2006 (TXT).
[RFC4593]	Barbir, A., Murphy, S., and Y. Yang, “Generic Threats to Routing Protocols,” RFC 4593, October 2006 (TXT).
[RFC4601]	Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, “Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised),” RFC 4601, August 2006 (TXT, PDF).
[RFC4724]	

	Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, “Graceful Restart Mechanism for BGP,” RFC 4724, January 2007 (TXT).
[RFC4760]	Bates, T., Chandra, R., Katz, D., and Y. Rekhter, “Multiprotocol Extensions for BGP-4,” RFC 4760, January 2007 (TXT).
[RFC4893]	Vohra, Q. and E. Chen, “BGP Support for Four-octet AS Number Space,” RFC 4893, May 2007 (TXT).
[Tsuchiya87]	Tsuchiya, P., “An Architecture for Network-Layer Routing in OSI,” Proceedings of the ACM workshop on Frontiers in computer communications technology , 1987.
[Xu97]	Xu, Z., Dai, S., and J. Garcia-Luna-Aceves, “A More Efficient Distance Vector Routing Algorithm,” Proc IEEE MILCOM 97, Monterey, California, Nov 1997.

Authors' Addresses

TOC

	Elwyn B. Davies
	Folly Consulting
	Soham, Cambs
	UK
Phone:	+44 7889 488 335
Email:	elwynd@dial.pipex.com
	Avri Doria
	LTU
	Lulea, 971 87
	Sweden
Phone:	+1 401 663 5024
Email:	avri@acm.org