IRTF Routing Research   Elwyn Davies, Avri Doria, Nortel Networks
Internet Draft   Malin Carlzon, SUNET
Anders Bergsten, Olle Pers, Yong Jiang, Telia Research
Lenka Carr Motyckova, Pierre Fransson, Olov Schelen
Lulea University of Technology

Document: draft-davies-fdr-reqs-00.txt   February, 2001

Future Domain Routing Requirements
<draft-davies-fdr-reqs-00.txt>

Status of this Memo

Discussion and suggestions for improvement are requested. This document will expire before September, 2001. Distribution of this draft is unlimited.

Copyright Notice

CONTENTS

Abstract

This document sets out a set of requirements which we believe are desirable for the future routing architecture and routing protocols of a successful Internet. This first version is, of necessity, incomplete. It is hoped that this document will be useful as a catalyst to the work that needs to be done in both the IRTF and the IETF.

1. Introduction

It is generally accepted that there are major shortcomings in the inter-domain routing of the Internet today and that these may result in meltdown 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.

Various developments 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 which were never foreseen by the original architects of the Internet routing system.

Taken together with the radically altered and now commercially-based organization of the Internet and the potential advent of IPv6, major changes to the inter-domain routing system are inevitable.

Although inter-domain routing is seen as the major source of problems, the interactions with intra-domain routing and the constraints that confining changes to the inter-domain would impose, means that we should consider the whole area of routing as an integrated system. This is done for 2 reasons:

- Requirements should not presuppose the solution. A continued commitment to the current definitions and split between inter-domain and intra-domain routing would constitute such a presupposition. Therefore the name Future Domain Routing (FDR)is being used in this document,

- As an acknowledgement of how intertwined inter-domain and intra-domain routing are within today’s routing architecture.

Although the meaning of Domain Routing will be developed implicitly throughout the document, a bit of explicit definition of the word ‘domain’ is required. This document uses domain in a very broad sense to mean any collection of systems or domains which come under a common authority that determines the attributes that define, and the policies that control that collection. The use of domain in this context is very similar to the concept of Region that was put forth by John Wroclawski in his Metanet model [10]. The idea includes the notion that within a domain certain system attributes will characterize the behavior of the systems in the domain and that there will be borders between domains. The idea of domain presented does not inherently presuppose that the identifying behaviors between two domains are to be the same. Nor does it presuppose anything about the hierarchical nature of domains. Finally it does not place restrictions on the nature of the attributes that might be used to determine membership in a domain. Since today’s routing domains are a subset of all domains as conceived of in this document, there has been no attempt to create a new term.

This draft makes a start on this process in Section 2 by revisiting the requirements for a future routing system which were last documented in RFC1126 - “Goals and Functional Requirements for Inter-Autonomous System Routing” [4] as a precursor to the design of BGP in 1989. The historical perspective is also fleshed out by looking at some other work that has been carried out since RFC1126 was published. Some of the requirements derive from the problems that are currently being experienced in the Internet today. These will be discussed in Section 3. The environment in which the future domain routing system will have to work is covered in Sections 4 - 6. Specific requirements for a future Domain routing system are discussed in Sections 7 - 10. Inevitably this document is incomplete: Some known Open Issues are discussed in Section 11.

1.1 Background

Today’s Internet uses an addressing and routing structure which has developed in ad hoc, more or less upwards compatible fashion from the essentially single domain, non-commercial Internet to a solution which is handling, albeit not totally satisfactorily, today’s multi-domain, federated, combined commercial and not-for-profit Internet. The result is not ideal, particularly as regards inter-domain routing mechanisms which have to implement a host of domain specific routing policies for competing, communicating domains.

Based a large body of anecdotal evidence, but also on experimental evidence [11] and analytic work on the stability of BGP under certain policy specifications [12], the main Internet inter-domain routing protocol, BGP4, appears to have a number of problems which need to be resolved. Some of these problems may be relieved by patches and fix-ups and some of these problems may require a new architecture and new protocols. The starting point of this work is to step back from the current state, examine how the Internet might develop in the future and derive a new set of requirements for a routing architecture from this work.

The development of the Internet is likely to be driven by a number of changes that will affect the organization and the usage of the network, including:

- Ongoing evolution of the commercial relationships between (connectivity) service providers, leading to changes in the way in which peering between providers is organised and the way in which transit traffic is routed
- Requirements for traffic engineering within and between domains including coping with multiple paths between domains
- Potential addition of a second IP addressing technique through IPv6.

- Incorporation of alternative forwarding techniques such as the pipes supplied by combined MPLS and Optical Lambda environments
- Support for alternative and multiple routing techniques which are better suited to delivering some types of content.

1.2 Goals

This section attempts to answer two questions:

- What are we trying to achieve in a new architecture?
- Why should the Internet community care?

There is a third question which needs to be answered as well, but which, for the present, is mostly not explicitly discussed:

- How will we know when we have succeeded?

1.2.1 Providing a Routing System matched to Domain Organisation

Many of today's routing problems are caused by a routing system which is not well-matched to the organization and policies which it is trying to support. It is our goal to develop a routing architecture where even domain organization which is not envisioned today can be served by a routing architecture that matches its requirements.

We will know when this goal is achieved when the desired policies, rules and organization can be mapped into the routing system in a natural, consistent and simply understood way.

1.2.2 Supporting a range of different communication services

Currently only best-effort datagram connectivity is supported in BGP. With, for example, DiffServ it is possible to construct different services within the network. A number of PDBs has been proposed to the IETF. Also a number of services has been talked about outside the IETF. These services might for example be Virtual Wire [18] and Assured rate [19].

Providers today promise how traffic will be handled in the network, for example delay and packet loss guarantees, and this will probably be even more relevant in the future. Communicating this information (i.e., the service characteristics) in routing protocols is necessary in near future.

Thus, a goal with this architecture is to allow for more information passed between operators and support other services than the best-effort datagram connectivity service.

1.2.3 Scaleable well beyond current predictable needs

Any proposed new FDR system should scale beyond the size and performance we can foresee for the next ten years. The previous IDR proposal has, with some massaging, held up for somewhat over ten years in which time the Internet has grown far beyond the predictions that were made in the requirements that were placed on it originally.

1.2.4 Supporting alternative forwarding mechanisms

With the advent of circuit based technologies (e.g., MPLS [24], G-MPLS [25]) managed by IP routers there are forwarding mechanisms other than the datagram service that need to be supported by the routing architecture.

An explicit goal of this architecture is to support more forwarding mechanisms than just the datagram forwarding.

1.2.5 Supporting separation of topology map and connectivity service

It is envisioned that an organization can support multiple services on top of a single network. These services can, for example, be of different quality, of different type of connectivity, or different protocols (e.g. IPv4 and IPv6). For all these services there may be common domain topology, even though the policies might differ.

Thus, a goal with this architecture is to support separation between creation of an domain (or organization) topology map and service creation.

1.2.6 Achieving full/appropriate separation of concerns between routing and forwarding

The architecture of a router is composed of two main separable parts; control and forwarding. These components, while inter-dependent, perform functions that are largely independent of each other. Control (routing, signaling, and management) is typically done in software while forwarding typically is done with specialized ASICs or network processors.

The nature of an IP based network today is that control and data protocols share the same network and forwarding regime. This may not always be the case in future networks and we should be careful to avoid building this sharing in as an assumption in the FDR.

A goal of this architecture is to support full separation of control and forwarding.

1.2.7 Providing means of using different routing paradigms seamlessly in different areas of the same network

A number of different routing paradigms have been used or researched in addition to conventional shortest path hop-by-hop paradigm that is the current mainstay of the Internet. In particular, differences in underlying transport networks may mean that other kinds of routing are more relevant, and the perceived need for traffic engineering will certainly alter the routing chosen in various domains.

1.2.8 Preventing denial of service and other security attacks

Part of the problem here is that the Internet offers a global, unmoderated connectivity service. Existence of a route to a destination effectively implies that anybody who can get a packet onto the network is entitled to use that route. Whilst there are limitations to this generalization, there is a clear invitation to denial of service attacks. A goal of the FDR system should be to allow traffic to be specifically linked to whole or partial routes so that a destination or link resources can be protected form malicious use.

1.2.9 Providing provable convergence with verifiable policy interaction

It has been shown both analytically by Griffin et al (see [12]) and practically (see [20]) that BGP will not converge stably or is only meta-stable (i.e. will not reconverge in the face of a single failure) when certain types of policy constraint are applied to categories of network topology. The addition of policy to the basic distance vector algorithm invalidates the mathematical proofs that applied to RIP and could be applied to a policy free BGP implementation.

A goal of the FDR should be to achieve mathematically provable convergence of the protocols used which may involve constraining the topologies used and vetting the polices imposed to ensure that they are compatible across domain boundaries.

1.2.10 Providing robustness despite errors and failures

From time to time in the history of the Internet there have been occurrences where global connectivity has been destroyed by people misconfiguring routers. This should never be possible.

A goal of the FDR is to be robust to configuration errors and failures. This should probably involve ensuring that the effects of misconfiguration and failure can be confined to some suitable locality of the failure or misconfiguration: This is not currently the case as both misconfigurations and problems in any AS can, under the right circumstances, have global effects across the entire Internet.

1.2.11 Simplicity in management

With the policy work ([26], [27] and [28]) that has been done at IETF there exists an architecture that standardizes and simplifies management of QoS. This kind of simplicity is needed in a future domain routing architecture and its protocols.

A goal of this architecture is to make configuration and management of inter-domain routing as simple as possible.

2. Historical Perspective

2.1 The legacy of RFC1126

RFC1126 outlined a set of requirements that were to guide the development of BGP. While the network is definitely different then it was in 1989, 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. Which means a future domain routing 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:

Optionally, interpretation for today's audience of the 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?

2.1.1 "General Requirements"

2.1.1.1 "Route to Destination"

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 is an add-on to BGP following many of the same strategies but not integrated into the BGP framework [23].

2.1.1.2 "Routing is Assured"

This requires that a user be notified within a reasonable time period of 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.

2.1.1.3 “Large System

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 a very sparse to a quite dense now). Instead of setting growth in a time-scale, indefinite growth should be accommodated.

Current practice: Scalability of the protocols will not be sufficient under the current rate of growth . There are problems with BGP convergence for large dense topologies, problems with routing information propagation between routers in transit domain, limited support for hierarchy, etc.

2.1.1.4 “Autonomous Operation”

Relevance: Valid. There may need to be additional requirements for adjusting policy decisions to the global functionality and to avoid contradictory policies would decrease a possibility of unstable routing behavior.
There should also be a separate requirement for handling various degrees of trust in autonomous operation, 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, to a global routing mechanism?

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.

2.1.1.5 “Distributed System”

The routing environment is a distributed system. The distributed routing environment supports redundancy and diversity of nodes and links. Both data and operations are distributed.

Relevance: Valid. RFC1126 is very clear that we should not be using centralized solutions, but maybe we need a requirement on trade-offs between common knowledge and distribution (e.g., to allow for uniform policy routing) (e.g., GSM systems are in a sense centralized (but with
hierarchies) and they work) This requirement should not rule out certain solutions that are needed to meet other requirements in this document.

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

2.1.1.6 “Provide A Credible Environment”

Routing mechanism information must be integral and secure (credible data, reliable operation). Security from unwanted modification and influence is required.

Relevance: Valid.

Current practice: BGP provides a mechanism for authentication and security. There are however security problems with current practice.

2.1.1.7 “Be A Managed Entity”

Requires that a manager should get enough information on a state of network so that (s)he could make informed decisions.

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 no global coordination, and therefore a manager cannot make a globally consistent decision.

2.1.1.8 “Minimize Required Resources”

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 NLRI. Usage of 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. Memory requirements are dominated by the number of networks in the Internet – this is a scaling problem.

2.1.2 “Functional Requirements”

2.1.2.1 “Route Synthesis Requirements”

2.1.2.1.1 “Route around failures dynamically”

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. Detections may need to be faster than 100ms to avoid being noticed by end-users.

Current practice: latency of fail-over is too high (minutes).

2.1.2.1.2 “Provide loop free paths”

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 routing-loops may occur due to bad routing policies.

2.1.2.1.3 "Know when a path or destination is unavailable"

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.

2.1.2.1.4 "Provide paths sensitive to administrative policies"

Relevance: Valid. Policy control of routing is of increasingly importance as the Internet has turned into business.

Current practice: Supported to some extent. Policies are only possible to apply locally in an AS and not globally. At least there is very small possibilities to affect policies in other AS's. 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).

2.1.2.1.5 "Provide paths sensitive to user policies"

Relevance: Valid to some extent, as it may contradict with the policies of the network administrator.

Current practice: not supported in normal routing. Can be accomplished to some extent with lose source routing, resulting in inefficient forwarding in the routers.

2.1.2.1.6 "Provide paths which characterize user quality-of-service requirements"

Relevance: Valid to some extent, as it may contradict the policies of the operator

Current practice: Creating routes with specified QoS is not possible now.

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

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

2.1.2.2 “Forwarding Requirements”

2.1.2.2.1 “Decouple inter- and intra-autonomous system forwarding decisions”

Relevance: Valid.

Current practice: As explained in 2.1.2.1.7, intra-domain forwarding in transit domains is codependent with inter-domain forwarding decisions.

2.1.2.2.2 “Do not forward datagrams deemed administratively inappropriate”

Relevance: Valid, however 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 there is packet filtering, e.g., to catch source address spoofing on outgoing traffic or to filter out unwanted incoming traffic. In the backbone, intentional packet dropping based on policies is not common.

2.1.2.2.3 “Do not forward datagrams to failed resources”

Relevance: blurry to some extent. There is a trade-off between scalability and keeping track of unreachable resources. The closer to a failing resource, the stronger reason for that the failure should be known.

Current practice: routing protocols keep track of failing routers, but not other resources (e.g., end-hosts switches etc.)

2.1.2.2.4 “Forward datagram according to its characteristics”

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

Current practice: ingress and egress filtering can be done on policy.

2.1.2.3 “Information Requirements”

2.1.2.3.1 “Provide a distributed and descriptive information base”

Relevance: Valid, however hierarchical IBs might provide more possibilities.

Current practice: IBs are distributed, not sure whether they support all provided routing functionality.

2.1.2.3.2 “Determine resource availability”

Relevance: Valid. It should be possible for reource availablity and levels of resource availability to be determined. This prevents needing to discover unavailabity through failure.

2.1.2.3.3 “Restrain transmission utilization”

Relevance: Valid. However certain requirements, as fast detection of faults may be worth consumption of more resources.

Current practice: BGP messages probably do not ordinarily consume excessive resources, but might during erroneous conditions.

2.1.2.3.4 “Allow limited information exchange”

Relevance: Valid. But perhaps routing could be improved if certain information could be globally available.

Current practice: Policies are used to determine which reachability information that is exported.

2.1.2.4 “Environmental Requirements”

2.1.2.4.1 “Support a packet-switching environment”

Relevance: Valid but should not be exclusive.

Current practice: supported.

2.1.2.4.2 “Accommodate a connection-less oriented user transport service”

Relevance: Valid, but should not be exclusive.

Current practice: accommodated.

2.1.2.4.3 “Accommodate 10K autonomous systems and 100K networks”

Relevance: No longer valid. Needs to be increased substantially.

Current Practice: Yes but perhaps reaching the limit.

2.1.2.4.4 “Allow for arbitrary interconnection of autonomous systems”

Relevance: Valid. However perhaps not all interconnections should be used globally.

Current practice: BGP-4 allows for arbitrary interconnections.

2.1.2.5 “General Objectives”

2.1.2.5.1 “Provide routing services in a timely manner”

Relevance: Valid, stated before. The more complex a service is the longer it should be allowed to take, linearly, polynomially, exponentially (NP-complete problems?)

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

2.1.2.5.2 “Minimize constraints on systems with limited resources”

Relevance: Valid

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

2.1.2.5.3 “Minimize impact of dissimilarities between autonomous systems”

Relevance: Important. This requirement is critical to a future architecture. In a domain 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 isn’t required in today’s networks since the intra-domain attribute are nearly identical to start with.

2.1.2.5.4 “Accommodate the addressing schemes and protocol mechanisms of the autonomous systems”

Relevance: Important

Current practice: Largely only one global addressing scheme is supported in most autonomous systems.

2.1.2.5.5 “Must be implementable by network vendors”´

Requirement: Valid

Current practice: BGP was implemented;

2.1.3 “Non-Goals”

RFC1126 also included a section discussing non goals. To what extent are these still non goals? Does the fact that they were non-goals adversely affect today’s IDR system?

2.1.3.1 “Ubiquity”

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 FDR. On the other hand, ubiquitous connectivity cannot be reached in policy sensitive environment and should not be an aim.

Relevance: De facto essential for a FDR but ensure that we mean ubiquity of the routing system rather than ubiquity of connectivity.
Current practice: de facto ubiquity achieved.

2.1.3.2 “Congestion control”

Relevance: Not clear if they mean routing or forwarding. It is definitely a non-goal to adapt the choice of route at transient congestion. However, to add support for congestion avoidance (e.g., ECN and ICMP messages) in the forwarding process would be OK.
Current practice: There exists some ICMP-messages (source quench) but these are not used.

2.1.3.3 “Load splitting”

Relevance: This should not be a non-goal, or an explicit goal. It might be desirable in some cases.
Current practice: Can be implemented by exporting different prefixes on different links, but this requires manual configuration and does not consider actual load.

2.1.3.4 “Maximizing the utilization of resources”

Relevance: Valid. Cost-efficiency should be strived for, maximizing resource utilization does not always lead to greatest cost-efficiency.

2.1.3.5 “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 CBR services in ATM.

Relevance: The hard form of deadline services is still a non-goal for the FDR 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.

2.1.3.6 “Non-interference policies of resource utilization”

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.

Current practice: Routing does not consider dynamic resource availability.
Forwarding can support service differentiation.

2.2 Nimrod Requirements

Nimrod as expressed by Noel Chiappa in his early document, “A New IP Routing and Addressing Architecture” and later in the NIMROD Working Group documents RFC 1753 and RFC1992 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 goals of Nimrod, quoted from RFC1992, were as follows:

1. To support a dynamic internetwork of arbitrary size 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 remain as requirements for any new domain routing architecture.

- As discussed in other sections of this document the amount of information needed to maintain the routing system is growing at a rate that does not scale. 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 on information that are distributed can be controlled.
  This goal will be reflected in the requirements for the future domain 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. There have been no changes to the network in the last half decade that have improved this situation any.
- This is still a absolute necessity. It is impossible to imagine that a new routing architecture could supplant the current architecture on a flag day. Instead any new

architecture will need to be able to incrementally deploy within the Internet.

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. The reason offered are various. 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 requirements 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 FDR. Nimrod can be said to have two derivatives:

- MPLS in that it took the notion of forwarding along well known paths
- 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.

2.3 PNNI

PNNI 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 it methods from a study of the Nimrod architecture. What can be gained from an analysis of what 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?

Additionally PNNI is not designed to support a single standardised "SPF" algorithm that must be present in all routers. Instead it relies on the entry node to compute a constraint-based path. It also relies on topological maps that presented an abstracted view of one network to another. 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. Information on these issues, and any other relevant issues, is solicited from those who do have such operational experience.

3. Existing problems of BGP and the current EGP/IGP Architecture

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 IGPs in use today could be improved. This section is, to a large extent, a wish list for the FDR 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' [13], particularly with respect to its stability and the problems produced by explosions in the size of the Internet.

3.1 BGP and Auto-aggregation

The stability and later linear growth rates of the number of routing objects (prefices) 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 DFZ 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 prefices 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.

3.2 Convergence and Recovery Issues

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, the effect of route flap damping is exponential in its behavior which can result in some parts of the Internet being inaccessible for hours at a time.

There is evidence ([13] and our own measurements) 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.

3.3 Non-locality of Effects of Instability and Misconfiguration

There have been a number of instances, some of which are well-documented (e.g. The April 1997 incident when misconfiguration of BGP at a small company in Virginia, USA, turned the company into a traffic magnet for much of the traffic in the Internet resulting in global problems until it was fixed) 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, 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.

3.4 Multihoming Issues

As discussed previously, the increasing use of multi-homing as a robustness technique by peripheral ASs 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 DFZ routing table is growing exponentially again.

The longest prefix match routing technique introduced by CIDR, and implemented in BGP4, 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:
AS3 has received its addresses from AS1, which means AS1 can Aggregate. But if AS3 want its traffic to be seen equally both ways, AS1 is forced to announce both the aggregate and the more specific route to AS3.

```
                 \       /
                 AS1      AS2
                    \   /
                     AS3
```

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 DFZ route table size explosion. This problem also interferes with the desire of many providers in the DFZ to route only prefixes which are equal to or shorter than 20 or 19 bits.

3.5 AS-number exhaustion

The domain identifier or AS-number is a 16-bit number. Allocation of AS-numbers is currently increasing 51% p.a. [13] with the result that exhaustion is likely around 2005. The IETF is currently studying proposals to increase the available range of AS-numbers to 32 bits, but this will present a deployment challenge during transition.

3.6 Partitioned AS’s

BGP is unable to handle an AS which has been split into two or more unconnected pieces. One school of opinion is that this is appropriate behaviour and should not be changed: The view is that responsibility for maintaining connectivity within the AS should belong solely to the administrators of the domain.   On the other hand, improving the robustness of the FDR may necessitate solving this problem, particularly as multi-homing becomes increasingly prevalent.

3.7 Load Sharing

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. Traffic engineering extensions to the FDR which will facilitate load sharing are essential.

3.8 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 ‘hold down’ 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 ([14], [15]) indicate that faster links could sustain a much higher rate of keep-alive messages without significantly impacting normal data traffic. This would improve BGP’s 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 hold-down interval.

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 re-synchonise the databases. Proprietary techniques have been implemented to try to reduce this problem.

3.9 Interaction between Inter domain routing and intra domain routing

Today, many operators' backbone routers run both I-BGP and an IGP maintain the routes that reach between the borders of the domain. Exporting routes from BGP into IGP and bringing them back up to BGP is not recommended [29], but it is still necessary for all backbone routers to run both protocols. BGP is used to find the egress point and IGP 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 distance vector protocol, as compared with most IGPs which are link state protocols, and as such it is not optimised for convergence.

- The metrics used in BGP and the IGP are rarely comparable or combinable.

- Policy control in BGP is designed for simple policies between operators, not for controlling paths within a domain.

- If all paths between two border routers have been lost, and this is known by the IGP this may not always be used in BGP. Instead the border router may wait until the logical connection between the borders has been lost, and first at this point declare the destinations as unreachable.

- Synchronization between IGP and EGP is a problem as long as we use different protocols for IGP and EGP, 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. The draft 'OSPF Transient Blackhole Avoidance' [22], the IGP side of the story is covered.

- Synchronizing in BGP means waiting for the IGP to know about the same networks as the EGP, which can take a significant period of time and slows down the convergence of BGP by adding the IGP convergence time into each cycle.

3.10 Policy Issues

There are several classes of issue 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 [12] and in McPherson [20] it is possible to install policies in routers that will cause routing loops and will never 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 highly qualified specialist are 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.

3.11 Security Issues

While many of the issues with BPG security have been traced either to implementation issues or to operational issues, BGP is vulnerable to 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, most filtering methods, it is takes constant vigilance. There is nothing in the current architecture or in the protocols that serves to protect the forwarders from these attacks.

3.12 Support of MPLS and VPNS

Recently BGP has been modified to function as a signalling protocol for MPLS and for VPNs [16]. This over-loading of the BGP protocol is seen as a boon by some and as a problem by others. 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.

An ISP that is providing VPN service needs to distribute VPN specific state to the provider edge (PE) nodes involved in each VPN (core nodes, i.e. ISP's nodes that are not PE nodes, do not need the VPN specific state). Specifically, each PE node participating in VPN X must distribute a VPN Tunnel Object to every other PE node in VPN X . The VPN Tunnel Object includes the originating PE's Router ID, the VPN's identifier X, a VPN Tunnel identifier, e.g. a label, and either the VPN destinations that are reachable using that tunnel or the virtual Router ID of a VPN specific virtual router that is reachable via the tunnel.

A PE node must distribute VPN Tunnel Objects pertaining to VPN X through the ISPs network to every other PE

Nodes participating in VPN X. In one proposal, an ISPs IBGP system is used for this distribution. The proposal requires scaleability in the number of PEs, VPNs and therefore VPN Tunnel Objects and so recommends the use of Route Reflectors within the IBGP system. In this application, BGP fails to meet the applications requirements in several ways: for example, delivery of the VPN Tunnel Objects to the appropriate PE Nodes is unreliable (a RR cannot guarantee propagation of BGP routes) and no confirmation of delivery is given. Since BGP has no notion of end-to-end messages, reliability and acknowledgements will not be possible. Additionally, the RRs are burdened with storing the locally irrelevant VPN Tunnel Objects’ data in their RIBs. The RRs’ RIB sizes then adversely affects processing of IBGP updates containing the VPN Tunnel Objects. In a final, typically BGP example, these two problems multiply each other: for reduced unreliability, a PE may attach to two different RRs which leads to a four times increase in RR RIB sizes and the number of updates a RR must process.

In creating the future domain routing architecture, serious consideration must be given to the argument that VPN signalling protocols should remain separate from the route determination protocols.

3.13 IPv4 / IPv6 Ships in the Night

The fact that service providers would need to maintain two completely separate networks; one for IPv4 and one for IPv6 has been a real hindrance to the introduction of IPv6. Even if IPv6 does get 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.

3.14 Existing Tools to Support Effective Deployment of Inter-Domain Routing

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

Routing Policy Specification Language RPSL enables a network operator to describe routes, routers and autonomous systems ASs that are connected to the local AS.

Using the RPSL language 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 (RIPE 81, 181, 103) that can be applied on the database to answer requests of the form, e.g.

- 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 which 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 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.

4. Expected Users

This section addresses the question of the target audience of the FDR both in terms of organizations that might own networks which would use FDR and the human users who will have to interact with the FDR>

4.1 Organisations

The organizations that own networks connected to the Internet have become much more diverse since RFC1126 [4] was published. In particular a major part of the network is now owned by commercial service provider organizations in the business of making profits from carrying data traffic.

4.1.1 Commercial Service Providers

The routing system must take into account their desires for commercial secrecy and security, as well as allowing them to organize their business as flexibly as possible.

Service providers will often wish to conceal the details of the network from other connected networks. So far as is possible, the routing system should not require the service providers to expose more details of the topology and capability of their networks than is strictly necessary.

Many service providers will also offer contracts to their customers in the form of Service Level Agreements (SLAs) and the routing system must allow the providers to support these SLAs through traffic engineering and load balancing as well as multihoming allowing them to achieve the degree of resilience and robustness that they need.

Service providers can be categorized as

- Global Service Providers (GSPs) with networks which have a global reach. Such providers may and usually will wish to constrain traffic between their customers to run entirely on their networks. Such providers will interchange traffic at multiple peering points with other GSPs and need extensive policy-based controls to control the interchange of traffic. Peering may be through the use of dedicated private lines between the partners or increasingly through Internet Exchange Points.
- National Service Providers (NSPs)which are similar to GSPs but typically cover one country. Such organizations may operate as a federation which provides similar reach to a GSP and may wish to be able to steer traffic preferentially to other federation members to achieve global reach.
- Local Internet Service Providers (ISPs) operate regionally and will typically purchase transit capacity from NSPs or GSPs to provide global connectivity, but may also peer with neighbouring ISPs.

The routing system should be sufficiently flexible to accommodate the continually changing business relationships of the providers.

4.1.2 Enterprises

The leaves of the network domain graph are in many cases networks supporting a single enterprise. Such networks cover an enormous range of complexity with some multi-national companies owning networks which rival the complexity and reach of a GSP whereas many fall into the Small Office-Home Office (SOHO) category. The routing system should allow simple and robust configuration and operation for the SOHO category, whilst effectively supporting the larger enterprise.

Enterprises are particularly likely to lack the capability to configure and manage a complex routing system and every effort should be made to provide simple configuration and operation for such networks.

Enterprises will also wish to be able to change their service provider with ease.

Enterprises will wish to be able to multihome to one or more providers to provide robustness.

4.1.3 Domestic Networks

Increasingly domestic networks are likely to be ‘always on’ and will resemble SOHO enterprises networks with no special requirements of the routing system.

In the meantime, the routing system must support dial-up users.

4.1.4 Internet Exchange Points

Peering of service providers, academic networks and larger enterprises is increasingly happening at specific Internet Exchange Points where many networks are linked together in a relatively small physical area. The resources of the exchange may be owned by a broker or jointly by the connecting networks. The routing systems should support such exchange points without requiring the exchange point to either operate as a superior entity with every connected network logically inferior to it or requiring the exchange point to be a member of one (or all) connected networks.

4.1.5 Content Providers

Content providers are at one level a special class of enterprise, but the desire to deliver content efficiently means that a content provider may provide multiple replicated origin servers or caches across a network. The routing system should facilitate delivering content from the most efficient location.

4.2 Human Users

This section covers the most important human users of the FDR and their expected interactions with the system.

4.2.1 Network Planners

The routing system should allow them to plan and implement a network which can be proved to be stable and will meet their traffic engineering requirements.

4.2.2 Network Operators

The routing system should, so far as is possible, be simple to configure and operate, behave in a predictable, stable fashion and deliver appropriate statistics and events to allow the network to be managed and upgraded in an efficient and timely fashion.

4.2.3 Mobile End Users

The routing system must support mobile end users.

5. Mandated Constraints

While many of the requirement to which the protocol must respond are technical, some aren't. These mandated constraints are those that are determined by conditions of the world around us. Understanding these requirements requires and analysis of the world in which these systems will be deployed,. The constraints include those that are determined by:

- Environmental factors.
- Geography
- Political boundaries and considerations
- Technological factors such as the prevalence of different levels of technology in the developed world as opposed to in the developing or undeveloped world.

5.1 The Federated Environment

The graph of the Internet network with routers and other control boxes at the nodes and communication links along the edges is today partitioned administratively into a large number of disjoint domains, known as Autonomous Systems (ASs).

A common administration may have responsibility for one or more domains which may or may not be adjacent in the graph.

Commercial and policy constraints affecting the routing system will typically be exercised at the boundaries of these domains where traffic is exchanged between domains.

The perceived need for commercial confidentiality will seek to minimise the information transferred across these boundaries, leading to requirements for aggregated information, abstracted maps of connectivity exported from domains and mistrust of supplied information.

One possible extension to the requirements would be to require the protocols to provide the ability for groups of peering domains to be treated as a (super-)domain. These domains would have a common administrative policy.

5.2 Working with different sorts of network

The diverse Layer 2 networks over which the layer 3 routing system is implemented have typically been operated totally independently from the layer 3 network. Consideration needs to be given to the degree and nature of leakage of information between the layers that is desirable. In particular, the desire for robustness through diverse routing implies knowledge of the underlying networks to be able to guarantee the robustness.

Mobile access networks may also impose extra requirements on Layer 3 routing.

5.3 Delivering Diversity

The routing system is operating at Layer 3 in the network. To achieve robustness and resilience at this layer requires that where multiple diverse routes are employed as part of delivering the resilience, the routing system at Layer 3 needs to be assured that the Layer 2 and lower routes are really diverse. The ‘diamond problem’ is the simplest form of this problem – layer 3 provider attempting to provide diversity buys layer 2 services from two separate providers who in turn buy wayleaves from the same provider:

```
          Layer 3 service
           /           \
          /             \
      Layer 2         Layer 2
    Provider A      Provider B
          \             /
           \           /
           Trench provider
```

Now when the backhoe cuts the trench, the Layer 3 provider has no resilience unless he had taken special steps to verify that the trench wasn’t common. The routing system should facilitate avoidance of this kind of trap.

5.4 When will the new solution be required?

There is a full range of opinion on this subject. An informal survey indicates that the range varies from 2 years to 6 years. And while there are those, possibly outliers, who think there is no need for a new routing architecture as well as those who think a new architecture was need years ago, the median seems to lie at around 4 years. As in all projections of the future this is largely not provable.

6. Assumptions

The assumptions so far in the work to derive the requirements for the Future Routing Domain have been:

The assumptions so far in the work to derive the requirements for the Future Routing Domain have been:

1. The number of hosts today is somewhere in the area of 100 Million. With dial in and NATs this is likely to turn into up to 500 Million users (see [30]). In a number of years, with wireless accesses and different ‘gizmos‘ attaching to the Internet, we are likely to see a couple of Billion ‘users‘ on the Internet. The number of globally addressable hosts is very much dependent on how common NATs will be in the future.
2. NATs exist and we cannot assume that NATs will cease being a presence in the networks.
3. The number of operators in the Internet will probably not grow very much, as there is a likelihood that operators will tend to merge. However, as Internet-connectivity expands to new countries, new operators will emerge and then merge again.
4. Today, there are around 9,500 AS’s with a growth rate of around 51% per annum [13]. With current use of AS’s (for e.g., multi-homing) the number of AS’s grow to 70,000 within 3 - 5 years.
5. In contrast to the number of operators, the number of domains is likely to grow significantly. Today, each operator has different domains within an AS, but this also shows in SLAs and policies internal to the operator. Making this globally visible would create a number of domains 10-100 times the amount of ASs, i.e., between 100,000 and 1,000,000.
6. With more and more capacity at the edge of the network the IP network will expand. Today there are operators with several thousands of routers, but this is likely to be increased. A domain will probably contain tens of thousands of routers.
7. The speed of connections in the (fixed) access will technically be (almost) unconstrained. However, the cost for the links will not be negligible so that the apparent speed will be effectively bounded. Within a number of years some will have Gigabit-speed in the access.
8. At the same time, the bandwidth of wireless access still has a strict upper-bound. Within the foreseeable future each user will only have a tiny amount of resources available compared to fixed accesses (10kbps to 2Mbps with only a few achieving the higher figure).
9. Assumptions 7 and 8 taken together suggest a span of bandwidth between 10 kbps to 1000 Mbps.
10. The speed in the backbone has grown rapidly, and there is no evidence that the growth will stop in the coming years. Terabit-speed is likely to be the minimum backbone speed in a couple of years.
11. There have been discussions as to whether Moore's law will continue to hold for processor speed. If Moore's law does not hold, then communication circuits might play a more important role in the future. Also, optical routing is based on circuit

technology which is the main reason for taking ´circuits´ into account when designing an FDR.

12. However, the datagram model still remains the fundamental model for the Internet.
13. The number of peering points in the network is likely to grow, as multi-homing becomes important. Also traffic will become more locally distributed, which will drive the demand for local peering.
14. The FDR will achieve the same degree of ubiquity as the current Internet and IP routing.

7. Functional Requirements

This section includes a detailed discussion of new requirements for a future domain routing architecture. As discussed in section 2.1 a new architecture must build upon the requirements for past routing architecture. For that reason, the requirements discussed in section 2.1 are not repeated here. In case where the requirement has changed significantly, was omitted from the discussions in RFC1126 or were treated as non-goals in RFC1126 but may now be significant, it will be discussed in further detail I this section.

7.1 Topology

7.1.1 The same topology information should support different path selection ideas:

The same topology information need to provide a more flexible spectrum of path selection methods that we might expect to find in a future Internet, including, amongst others, both distributed techniques such as hop by hop, shortest path, local optimization constraint-based, class of service, source address routing, and destination address routing as well as the centralized, global optimization constraint-based ‘traffic engineering’ type (Open constraints should be allowed). Allowing different path selection techniques to be used will produce a much more predictable and comprehensible result than the ‘clever tricks’ which are currently needed to achieve the same results. Traffic engineering functions need to be combined.

Routers need to know the domain topology. BGP today operates with a policy database, but does not provide a link state database for the connectivity of each AS – the extent to which this is feasible or desirable needs to be investigated.

7.1.2 Separation between the routing information topology from the data transport topology.

The controlling network should be logically separate from the controlled network. Physically, the two functional "planes" can reside in the same nodes and share the same links, but this is not the only possibility. Other options can also be feasible, and may sometimes be necessary. An example is a pure circuit switch (that cannot see individual IP packets), combined with an external controller. Another example may be where there are multiple links between two routers, and all the links are used for data forwarding, but only one is used for carrying the routing session.

7.2 Distribution

7.2.1 Distribution mechanisms

The important requirement is that every entity gets the information it needs in a fast, reliable, and trusted way.

Possible distribution mechanisms for routing information exchange may be for example full mesh, route reflections, flooding, and multicast.

The current I-BGP seems to have unnecessary limitations in this respect, where a router requires full mesh to obtain all available routes. Route reflection avoids the need of full meshes but loses information since the route reflector chooses the best route for all the other routers. This best route might be different if all routers do the selection themselves in a full mesh.

7.2.2 Path advertisement

The inter-domain routing system must be able to advertise more kinds of information than just connectivity and AS path. The FDR should support the Service Level Specifications (SLSs) that are being developed under the Differentiated Services imprimatur.

Examples of such additional information can be:

- QoS information

To allow an ISP to sell predictable end-to-end QoS service to any destination, the routing system should have information about the end-to-end QoS. This means that the routing system should be able to support different paths for different DSCP’s or TOS-values. The outing system should also be able to carry information about the expected (or actually, promised) characteristics of the entire path and also the price for the service. (If such information is exchanged at all between network operators today, it is through bilateral management interfaces, and not through the routing protocols.)

This would allow for the operator to optimise the choice of path based on a price/performance trade-off.

It is possible that providing dynamic QoS information to control routing is not scalable, and an alternative would be to use static class-of-service information such as is suggested in the Differentiated Services work.

- security information

Security characteristics of other ASs (in the path or in the map) can allow the routing entity to choose routing decision based on some political reasons. The information itself is assumed to be so secure that you can trust it.

- usage and cost information

This can be used for billing and traffic engineering purpose. In order to support cost based routing policies for customers (ie peer ISPs), information such as "traffic on this link or path costs XXX USD per Gigabyte" needs to be advertised, so that the customer can choose a cheap or an expensive route from an economic perspective.

- monitored performance

Some performance information such as delay and drop frequency can be carried. (This is may only be suitable inside a domain.). This should support at least the kind of delay bound contractual terms that are currently being offered by service providers.

7.2.3 Stability of Routing Information

7.2.3.1 Avoiding Routing Oscillations

The FDR should seek to minimize oscillations in route advertisements.

7.2.3.2 Providing Loop Free Routing and Forwarding

In line with the separation of concerns of routing and forwarding, the distribution of routing information should be, so far as is possible, loop-free, and the forwarding information created from this routing information should also seek to minimize loops in the data forwarding paths.

7.3 Addressing

7.3.1 Support mix of IPv4 and IPv6 addresses and other types of addresses too

The routing system must support a mix of different kinds of addresses, including at least IPv4 and IPv6 addresses, and preferably various types of non-IP addresses too. For instance networks like SDH/SONET and WDM may prefer to use non-IP addresses.

7.3.2 Support for domain renumbering

The routing system must support renumbering (when a new prefix is given to an old network, and the change is known in advance).

7.3.3 Multicast and Anycast

The routing system must support multicast addressing, both within a domain and across multiple domains. It also needs to support anycast addressing within a domain, and inter-domain anycast addressing should preferably not be excluded.

7.3.4 Address scoping

The routing system must support scoping of addresses, for each of the unicast, multicast, and anycast types.

For unicast address scoping as of IPv6, there seems to be no special problems with respect to routing. Inter-domain routing handles only global addresses, while intra-domain routing also needs to be aware of site-local addresses. Link-local addresses are never routed at all.

For scoping in a more general sense, and for scoping of multicast and anycast addresses, more study may be needed to identify the requirements.

7.3.5 Mobility Support

The routing system shall support end system mobility (and movability, and portability, whatever the differences may be).

We observe that the existing solutions based on re-numbering and/or tunneling are designed to work with the current routing, so they do not add any new requirements to future routing. But the requirement is general, and future solutions may not be restricted to the ones we have today.

7.4 Management Requirements

7.4.1 Simple policy management

- Less manual configuration than today
- Operators/providers want easy handling, but cannot afford to lose control.
  - All the information should be available
  - But should not be visible except for when desired.
- Advertise policy (not only the result of policy)
- Policy conflict Resolution

(e g one would like to have one default behavior, and possibilities to choose other options. But much of this depends on implementation, and not on the protocols)

7.5 Mathematical Provability

The protocol is required to be resistant to bad routing policy decisions made by operators. Tools are needed to check compatibility of routing policies. Routing policies are compatible if their global interaction does not cause divergence (collection of ASes exchange routing messages indefinitely never entering a stable state). Tools must be provided to make routing system convergent. A routing system is convergent if after an exchange of routing information, routing tables reach a stable state that does not change until routing policies change.

To achieve the above mentioned goals a mechanism is needed to publish and communicate policies so that operational coordination and fault isolation is possible. Tools are required that verify stable properties routing system in specified parts of Internet. The tools should be efficient (fast) and have a broad scope of operation (check large portions of Internet).

Tools analyzing routing policies can be applied statically or (preferably) dynamically. Dynamic solution requires tools that can be used for run time checking for a source of oscillations that arise from policy conflicts. Research is needed to prove that there is an efficient solution to the dynamic checking of oscillations.

7.6 Traffic Engineering

7.6.1 Load Balancing (ECMP/OMP)

The routing system shall support the controlled distribution over multiple links or paths, of traffic towards the same destination. This applies to domains with two or more connections to the same neighbor domain, and to domains with connections to more than one neighbor domain. Load balancing can be both static and dynamic.

(*the next part is correct, but is it needed?*)
In intra-domain routing, the metric needs to contain more properties of the link such as delay, loss and utilization, to construct multiple paths and split load.

7.6.2 Peering support

The FDR must support peer–level connectivity as well as purely hierarchical inter-domain connections. The network is becoming increasingly complex with private peering arrangements set up between providers at every level of the hierarchy of service providers and even by certain large enterprises, in the form of dedicated extranets.

The FDR must facilitate traffic engineering of these peer routes so that the network operators can make optimal use of the available connectivity.

7.7 Multi-homing support

An FDR protocol must support multi-homing, i.e. support an AS to peer with several other domains.

As soon as a domain is multi-homed its prefixes are generally hard to aggregate as they are advertised further away from the multihomed domain, even if a domain is allotted a group of prefixes by a provider domain. As described above, multi-homing is leading to explosion of the size of the routing tables in the DFZ.

The rapid growth of the size of the routing tables has to be solved by one means or another. This may be achieved by forcing domains to aggregate more, by a form of auto-aggregation or by looking at a new routing architecture.

7.7.1 Support for NATs

One of our assumptions is that NATs are here to stay. The FDR should seek to work with NATs to aid in bi-directional connectivity through the NAT without compromising the additional opacity and privacy which the NAT offers. This problem is closely analogous to the abstraction problem which is already under discussion for the interchange of routing information between domains.

7.8 Statistics support

Both the routing and forwarding parts of the FDR must maintain statistical information about the performance of their functions. This may be an extended version of the MIBs provided for IP forwarding, BGP and the relevant IGP.

8. Performance Requirements

Over the past several years, the perfomance of the routing system has frequently been discussed. Some of the questions being asked include:

- How fast does an AS converge? How fast must domains converge?
- How big are the Areas, the ASs? How big should domains be?
- How much or how little data may be transferred in a routing message?
- How much state can be stored and processed in route control processors.
- Measures of network availability
- Measure of network reliability
- Global and Local measures of network Stability
- Capacity Measurement

In many cases there has been very little data or statistical evidence for many of the performance claims being made. In recent years several efforts have been initiated to gather data and do the analyses required to make scientific assessments of the performance issues and requirements. In order to complete this section of the requirements analysis, the data and analyses from these studies needs to be gathered and collated into this document. This work has been started but has yet to be completed.

9. Backwards compatibility (cutover) and Maintainability

This area poses a dilemma. On one hand it is an absolute requirements that introduction of FDR not require any flag days. The network currently in place has to keep running at least as well as it does now while the new network is being brought in around it.

However, at the same time, it is also an absolute requirement that the new architecture not be limited by the restrictions that plague today’s network. Thos restrictions cannot be allow to become permanent baggage on the new architecture. If they do, the effort to create a new system will come to naught.

These two requirements have significance not only for the transition strategy, but for the architecture itself since the determine that it must be possible for an internet such as today’s BGP controlled network, or one of its ASs, can exist as a domain within the FDR.

10. Security Requirements

It must be possible to secure the routing communication: the communicating entities shall be able to identify who sent and who received the information (authentication), and verify that the information has not been changed on the way (integrity).

Security is more important in inter-domain routing where the operator has no control to the other domains, and less serious in intra-domain routing since all the links and the nodes are under the administration of the operator and can be expected to share a trust relationship.

The routing communication mechanism shall be robust against denial-of-service attacks.

Should we also require:

- that no one else but the intended recipient can access (privacy) or understand (confidentiality) the information?
- possibility to verify that all the information has been received (non-repudiation)?

Is there a need to separate security of routing from security of forwarding?

Securing the BGP session, as done today, only secures the exchange of messages from the peering AS, not the content of the information. In other words, we can confirm that the information we got is what our neighbor really sent us, but we do not know if this information (that originated in some remote AS) is true or not.

Is it enough to rely on chains of trust (we trust our peers who trust their peers who..), or do we also need authentication and integrity of the information end-to-end?

The FDR should seek to cooperate with the security policies of firewalls whenever possible. This is likely to involve further requirements for abstraction of information, as the firewall is seeking to minimize interchange of information which could lead to a security breach.

11. Open Issues

This section covers issues that need to considered and resolved in deciding on a future domain routing architecture. While they can't be described as requirements, they do affect the types of solution that are acceptable. The discussions included below are very open-ended.

11.1 System Modeling

It is still a new assumption that object modeling of a system is an essential first step to creating a new system. Frequently the effort to object model becomes an end in itself and does not lead to system creation. But there is a balance and a lot that can be discovered in an ongoing effort to model a system such as the future domain routing system.

It is recommended that this process be included in the requirements. It should not, however be a gating event to all other work.

Some of the most important realizations will occur during the process of determining the following:
- Object classification
- Relationships and containment
- Roles and Rules

11.2 Advantages and Disadvantages of having the same protocols for EGP and IGP

Inter-domain and intra-domain routing have different targets and business assumptions. An IGP figures out how each node in the network gets to every other node in the network in the most optimal way. In this context the word *optimal* refers to the cost of the path measured by metrics associated with each link in the network. The area of network infrastructure (primarily routers) over which an IGP runs is typically under the same technical and administrative control, and it defines the boundary of an AS (Autonomous System). The purpose of an EGP is to allow two different ASs to exchange routing information so that data traffic can be forwarded across the AS border. Because an AS border router both separates and attaches two different areas of technical and administrative control, the specifications and implementations of EGPs include mechanisms for doing policy routing, meaning that control can be exerted over which routing information crosses the border between two ASs. EGPs contain features that are like metrics in IGPs, but unlike IGPs, the function of an EGP is not necessarily to optimize the path that data traffic takes through a backbone. Having different protocols for EGP and IGP reflects this difference.

However, there is increasing demand in IGP to do policy routing. The shortest path may not be the best path in the light of the policies. Network operators need to have more flexibility in choosing routes for reasons such as load balancing. This means both inter-domain routing and intra-domain routing are for the same purpose of choosing the best route according to operators’ own policies. Having the same protocol will emphasize the need to do policy control in IGP. This especially important since the current IBGP is actually for intra-domain routing.

This comment touches on the fact that the level of manual control (policy) is much larger in EGP. Why is this so?

EGP:
- Manifests business relations to peers, providers and customers.
- Borders to resources outside of our control. We don’t trust others to behave well when configuring routing. These resources are also often be less stable (eg customer access).
- Network size extremely large. This gives many updates which means we need to have a simple calculation of paths. It also gives an extremely large amount of information (due to the network size)

which gives the need for aggregation. Also we need policy to protect our network from receiving bad announcements causing our egress traffic to take the "wrong" way and to avoid sending bad announcements attracting the "wrong" traffic.

IGP:

- The network resources are under our control and we trust ourself to behave well (in a sense defined by ourselves) when configuring routing.
- The network resources of today are fairly stable in a backbone network.
- The size of the network is limited. So, the domain is fairly stable which gives a limited number of updates. Limited number of updates gives the option of using processor intensive automation (distributed link state routing). This gives us fast and easy to manage dynamic routing. BUT stability and visibility issues still constrain us from going further down the path of policy routing.

11.2.1 The necessity to clearly identify all identities related to routing

As in all other fields, the words used to refer to concepts and to describe operations about routing are important. Rather than describe concepts using terms that are inaccurate or rarely used in the real world of networking, an effort is necessitated to use the correct words. Many networking terms are used casually, and the result is a partial or incorrect understanding of the underlying concept. Entities such as nodes, interfaces, sub-networks, tunnels, and the grouping concepts such as ASs, domains, areas, and regions, need to be clearly identified and defined to avoid mixing from each other. And even if they are all identified by IP numbers, the routing entities should know what kind of entities they are.

There is also a need to separate identifiers (what or who) from locators (where) from routes (how to reach). One of the problems with the current BGP is if there is a topology change, the amount of information circulated is a function of the number of IP prefixes being routed. This is a common problem for a distance vector protocol. If the topology information is properly separated from addressing information in a state map, then when a link between two ASs goes down, this is the only information which needs to be advertised, instead of advertising the inability to reach some network prefixes. This example shows the need to separate end node identifiers from routing information.

11.2.2 Map distribution and/or route Distribution

11.2.2.1 Map Abstraction

If every detail is advertised throughout the Internet, there will be a lot of information. Scalable solutions requires abstraction.

- If we summarise too much, some information will be lost on the way.
- If we summarize too little, then more information then required is available contributing to scaling limitations.
- One can allow more summarisation, if there also is a mechanism to query for more details within policy limits.
- The basic requirement is not that the information shall be advertised, but that the information shall be available to those who need it. (We should not presuppose a solution where advertising is the only possible mechanism.)

11.2.3 Robustness and redundancy:

The routing association between two domains should survive even if some individual connection between two ASBR routers goes down.

The "session" should operate between logical "routing entities" on each domain side, and not necessarily be bound to individual routers or IP addresses. Such a logical entity can be physically distributed over multiple network elements. Or it can reside in a single router, which would default to the current situation.

11.2.4 Hierarchy

A more flexible hierarchy with more levels and recursive groupings in both upward and downward directions allows more structured routing. So that no single level will get too big for routers to handle.

Note that groupings can look different depending on which aspect we use to define them. A DiffServ area, a MPLS domain, a trusted domain, a QoS area, a multicast domain, etc, do not always coincide. And neither are they strict hierarchical subsets of each other. The basic distinction at each level is "this grouping versus everything outside".

Each AS is still independent, and forms the basis for policy decisions. However, is there a need for a higher level aggregation which is above AS? If yes, who will be responsible for this level? Can a network make policy decisions on such aggregated ASs without seeing the individual ASs?

11.3 Introduction of new control mechanisms

Is it be possible to apply a control theory framework, and analyze the stability of the control system of the whole network domain, for e g speed and the frequency response, and then use the results from that analysis to set the timers and other protocol parameters.

11.4 Robustness

Is solution to the Byzantine Generals problem a requirement? What are some of the other network robustness issues that must be resolved.

11.5 VPN Support

Today BGP is also used for VPN and other stuff for example as described in RFC2547

Internet routing and VPN routing have different purposes, and most often exchange different information between different devices. Most Internet routers do not need to know any VPN specific information. The concepts should be clearly separated.

But when it comes to the mechanisms, VPN routing can share the same protocol as ordinary Internet routing, it can use a separate instance of the same protocol, or it can use a different protocol. All variants are possible and have their own merits.

For example, all the AS Border Routers within one AS participate in a full-mesh I-BGP process for distributing external IP routes. At the same time a separate "VPN-routing" protocol can be operating between all the PE routers of some "VPN provider". These PE routers can be located in different ASs, and some of them may also be ASBRs.

11.6 End to End Reliability

The existing Internet architecture neither requires or provides end-to-end reliability of control information dissemination. For example, in distributing VPN information there is, however, a requirement for end to end reliability of control information, i.e. the ends of the VPN established need to have a acknowledgement of the success in setting up the VPN. While it is not necessarily the function of a routing architecture to provide end-to-end reliability for this kind of purpose, we must be clear that end-to-end reliability becomes a requirement if the network has to support such reliable control signalling. There may be other requirements that derive from requiring the FDR to support reliable control signaling.

12. Acknowledgements

The authors would like to acknowledge the helpful comments and suggestions of the following individuals: Loa Anderson, Tomas Ahlström, Niklas Borg, Nigel Bragg, Krister Edlund, Owe Grafford, Torbjörn Lundberg, Jasminko Mulahusic, Bernhard Stockman, Henrik Villför, Tom Worster, 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 reference explicitly listed below, but those who contributed to the mailing lists we have been participating in for years.

13. References

[1] Clark, D., "Policy Routing in Internet Protocols", RFC 1102, May 1989.

[2] Estrin, D., "Requirements for Policy Based Routing in the Research Internet", RFC 1125, November 1989.

[3] Steenstrup, M,. “An Architecture for Inter-Domain Policy Routing”, RFC 1478, June 1993

[4] Little, M., "Goals and Functional Requirements for Inter-Autonomous System Routing", RFC 1126, July 1989.

[5] Perlman, R., “Interconnections Second Edition”, 1999, Addison Wesley Longman, Inc.

[6] Perlman, R., "Network Layer Protocols with Byzantine Robust-ness", Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, August 1988.

[7] Castineyra, I., Chiappa, N., Steenstrup, M., “the Nimrod Routing Architecture”, RFC1992, Aug 1996

[8] Chiappa, N., “IPng Technical Requirements of the Nimrod Routing and Addressing Architecture”, RFC 1753, Dec 1994

[9] Chiappa, N., “A New IP Routing and Addressing Architecture”

[10] Wroclowski, J., The Metanet White Paper - Workshop on Research Directions for the Next Generation Internet, 1995

[11] Labovitz, C., Ahuja, A., Farnam J., Bose, A., Experimental Measurement of Delayed Convergence, NANOG

[12] Griffin, T.G., Wilfong, G., An Analysis of BGP Convergence Properties, SIGCOMM 1999

[13] Huston, G., Architectural Requirements for Inter-Domain Routing in the Internet, Internet Draft – draft-iab-bgparch-00, Feb 2001, Work in Progress

[14] Alaettinoglu, C., Jacobson, V. and Yu, H, , Towards Milli-Second IGP Convergence, Internet Draft - draft-alaettinoglu-isis-convergence-00, Nov 2000 Work in Progress

[15] Sandick, H., Squire, M., Cain, B., Duncan, I., Haberman, B., Fast LIveness Protocol (FLIP), Internet Draft - draft-sandiick-flip-00, Feb 2000, Work in Progress

[16] Rosen, E. and Rekhter, Y., BGP/MPLS VPNs, RFC2547, March 1999

[17] Clark, D., Chapin, L., Cerf, V., Braden, R., Hobby, R., “towards the Future Internet Architecture”, RFC1287, December 1991

[18] Jacobson, V., Nichols, K. and Poduri, K., The ‘Virtual Wire’ Behavior Aggregate, Internet Draft – draft-ietf-diffserv-pdb-vw-00, July 2000, Work in Progress

[19] Seddigh, N., Nandy, B., and Heinanen, J., An Assured Rate Per-Domain Behaviour for Differentiated Services, Internet Draft - draft-ietf-diffserv-pdb-ar-00, Feb 2001, Work in Progress

[20] McPherson, D., Gill, V., Walton, D. and Retana, A., “BGP Persistent Route Oscillation Condition”, Internet Draft - draft-mcpherson-bgp-route-oscillation-00, Dec 2000, Work in Progress

[21] Hain, T, “Architectural Implications of NAT”, RFC 2993, November 2000

[22] McPherson, D. and Przygienda, T., OSPF Transient Blackhole Avoidance, Internet Draft - draft-mcpherson-ospf-transient-00, July 2000 Work In Progress

[23] Thaler, D., Estrin, D. and Meyer, D. (editors), Border Gateway Multicast Protocol (BGMP): Protocol Specification, Internet Draft - draft-ietf-bgmp-spec-02, Nov 2000 Work in progress

[24] Rosen E. Et al., Multiprotocol Label Switching Architecture, RFC 3031

[25] Ashwood-Smith P. Et al., Generalized MPLS - Signaling Functional Description, Internet Draft – draft-ietf-mpls-generalized-signaling-01.txt, Work in progress

[26] IETF Resource Allocation Protocol working group, http://www.ietf.org/html.charters/rap-charter.html

[27] IETF Configuration management with SNMP working group, http://www.ietf.org/html.charters/snmpconf-charter.html

[28] IETF Policy working group, http://www.ietf.org/html.charters/policy-charter.html

[29] Yu J., Scalable Routing Design Principles, RFC 2791

[30] Telcordia Technologies Netsizer web site http://www.netsizer.com/

14. Author’s Addresses

Elwyn Davies
Nortel Networks
London Road
Harlow, Essex CM17 9NA, UK
Phone: +44-1279-405498
Email: elwynd@nortelnetworks.com

Avri Doria
Nortel Networks
600 Technology Park Drive
Billerica, MA
Phone: +1 978 288 6627
Email: avri@nortelnetworks.com

Malin Carlzon
Royal Institute of Technology
Network Operating Centre
KTHNOC
SE-100 44
Stockholm, Sweden
Phone: +46 70 269 6519
Email: malin@sunet.se

Anders Bergsten
Telia Research AB
Aurorum 6
S-977 75 Lulea, SWEDEN
Phone: +46 920 754 50
Email: anders.p.bergsten@telia.se

Olle Pers
Telia Research AB
Stockholm, SWEDEN
Phone: +46 8 713 8182
Email: olle.k.pers@telia.se

Yong Jiang
Telia Research AB
123 86 Farsta SWEDEN
Phone: +46 8 713 8125
Email: yong.b.jiang@telia.se

Lenka Carr Motyckova
Div. of Computer
Lulea University of Technology
S-971 87
Lulea, SWEDEN
Phone: (+46) 920 91769
Email: lenka@sm.luth.se

Pierre Fransson
Div. of Computer
Lulea University of Technology
S-971 87
Lulea, SWEDEN
Phone: (+46) 70 646 0384
Email: pierre@cdt.luth.se

Olov Schelen
Div. of Computer
Lulea University of Technology
S-971 87
Lulea, SWEDEN
Phone: (+46) 70 536 2030
Email: Olov.Schelen@cdt.luth.se