Title:	Document Version:
Deliverable D1.2 **Requirements from a network operator perspective**	1.3

Project Number:	Project Acronym:	Project Title:
035167	RiNG	Routing in Next Generation

Contractual Delivery Date:	Actual Delivery Date:	Deliverable Type* - Security**:
31/12/2007	15/11/2008	R – PU

* Type: P - Prototype, R - Report, D - Demonstrator, O - Other
** Security Class: PU- Public, PP – Restricted to other programme participants (including the Commission), RE – Restricted to a group defined by the consortium (including the Commission), CO – Confidential, only for members of the consortium (including the Commission)

Responsible:	Organization:	Contributing WP:
Alessandro Capello	Telecom Italia	WP1

Authors (organizations):

All partners.

Abstract:

This document addresses the analysis of routing requirements from a network operator perspective.

The document is focused on a small subset of topics: scalability, traffic engineering, policy management and security. They are considered the most relevant aspects from a network operator point of view.

Each topic is discussed in the first part of the document, aiming at underlining main issues that currently affects the inter-domain routing architecture. The purpose is to summarize the most relevant analysis performed by the research community, giving a clear view of the open problems.

The second part of the document defines a set of metrics to be used to quantify properties and performances of the routing architecture (at least those properties which are quantifiable). Such metrics are expected also to be useful to compare alternative approaches to inter-domain routing, leading the discussion to a more pragmatic approach.

Finally, specific requirements, related to the four mentioned topics, are stated, recurring whenever possible to the metrics previously described. A main objective of this document is to provide an organic analysis of all requirements in order to help the research community to elaborate effective solutions.

Keywords:

Routing, inter-domain, BGP, requirements, network operator.

Revision History

The following table describes the main changes done in the document since its creation.

Revision	Date	Description	Author (Organization)
v0.1	08/05/2007	Document creation – added paragraphs related to current inter-domain routing issues	Alessandro Capello (Telecom Italia)
v0.2	27/11/2007	Document Update – added paragraphs related to metrics and requirements	Alessandro Capello (Telecom Italia)
v0.3	23/12/2007	Document Update – completed the executive summary and the conclusions	Alessandro Capello (Telecom Italia)
v1.0	27/12/2007	Last Revision	Alessandro Capello (Telecom Italia)
v1.1	27/12/2007	Document Revision	Alvaro Vives (Consulintel)
v1.2	31/12/2007	Integration of comments	Alessandro Capello (Telecom Italia)
v1.3	15/11/2008	Final Review	Jordi Palet (Consulintel)

Executive Summary

This document addresses the analysis of routing requirements from a network operator perspective. The distinction between Service Providers and End-Users is fundamental in order to clearly identify the actors involved and to perform a cost-benefits analysis of the requirements expressed by each side.

The document is focused on a small subset of topics: scalability, traffic-engineering, policy management and security. They are considered the most relevant aspects from a network operator's point of view.

Each topic is discussed in the first part of the document, aiming at underlining main issues that currently affects the inter-domain routing architecture. The purpose is to summarize the most relevant analysis performed by the research community, giving a clear view of the open problems.

The second part of the document defines a set of metrics to be used to quantify properties and performances of the routing architecture (at least those properties which are quantifiable). Such metrics are expected also to be useful to compare alternative approaches to inter-domain routing, leading the discussion to a more pragmatic approach.

Finally, specific requirements, related to the four mentioned topics, are stated, recurring whenever possible to the metrics previously described. A main objective of this document is to provide an organic analysis of all requirements in order to help the research community to elaborate effective solutions.

In the future, these requirements, and the issues which they have been derived from, will be usefully integrated or modified according to new achievements in the analysis of Internet behavior as a system. In that sense, this document is a "live document".

From a Service Provider's point of view, a careful cost-benefit analysis needs to be performed to evaluate which are the requirements that should have higher priority. Four main items have been analyzed: scalability, traffic engineering, policy management and security. Issues related to these items are probably the most relevant ones for Service Providers as they have direct impacts on both services and investments.

The performed analysis has shown that major concerns by Service Providers are related to the behavior and the functionalities of the BGP protocol. In that sense, from a network provider perspective, the evolution of the BGP protocol itself is of primary interest.

Any modifications to the BGP protocol raise the problem of the deployability of solutions which affect the routing architecture of the entire Internet. This represents one of the main issues faced by current research activities. Anyway, this problem must not divert research efforts from exploring viable solutions to these issues.

Another important aspect is the knowledge of the Internet as a system. We currently know little about it and, as a consequence, current solutions are more patches which address specific symptoms rather than treatments targeted to the main problems. Hence, any efforts to improve the knowledge of the Internet, through theoretical speculations and through active measurements, should be more actively supported by ISPs.

Table of Contents

1. Introduction *5*

1.1 The BGP routing protocol **5**

1.2 Analysis of Internet topology **7**

2. Analysis of current Inter-domain routing issues *9*

2.1 Scalability issues **9**

2.2 Traffic engineering issues **14**

2.3 Policy management and configuration issues **16**

2.4 Security issues **17**

3. Definition of useful metrics *19*

3.1 Internet topology **19**

3.2 Scalability **19**

3.3 Security **20**

4. Definition of requirements *21*

4.1 Scalability **21**

4.2 Traffic engineering **24**

4.3 Policy management and configuration **26**

4.4 Security **28**

5. Conclusions *29*

6. References *30*

1. INTRODUCTION

This document addresses the analysis of routing requirements from a network operator perspective. The distinction between Internet Service Providers (ISPs) and End-Users is fundamental in order to clearly identify the actors involved and to perform a cost-benefit analysis of the requirements expressed by each side.

The analysis is related to inter-domain routing since it has been identified as one of the main problems concerning the evolution of the Internet. Starting from the Routing and Addressing Workshop [1], held by the Internet Architecture Board (IAB) in October 2006, several initiatives, especially in IETF, were born to address this topic.

The purpose of this document is to collect information regarding the behavior of the inter-domain routing architecture, highlighting the main issues which are currently unresolved, and to define a set of requirements which address these issues. The identification of the main issues and the definition of requirements are performed considering the point of view of an ISP.

The document is focused on a small subset of items: scalability, traffic-engineering, policy management and security. They are considered the most relevant aspects for ISPs.

Each topic is discussed in the first part of the document, aiming at underlining main issues that currently affects the inter-domain routing architecture. The purpose is to summarize the most relevant analysis performed by the research community, giving a clear view of the open problems.

The second part of the document defines a set of metrics to be used to quantify properties and performances of the routing architecture (at least those properties which are quantifiable). Such metrics are expected also to be useful to compare alternative approaches to inter-domain routing, leading the discussion to a more pragmatic approach.

Finally, specific requirements, related to the four mentioned topics, are stated, recurring whenever possible to the metrics previously described. A main objective of this document is to provide an organic analysis of all requirements in order to help the research community to elaborate effective solutions.

1.1 The BGP routing protocol

The Autonomous Systems (ASes) that collectively comprise the Internet are controlled by individual organizations. They vary in size, from large national and multinational networks owned by corporations and governments, to small networks servicing a single business or school. There are three types of ASes: stub, multihomed, and transit. Stub ASes are communication endpoints, with connections to the rest of the Internet only made through a single upstream provider. Multihomed ASes are similar to stub ASes, but possess multiple upstream providers. Transit ASes have connections to multiple ASes and allow traffic to flow through to other ASes, even if the traffic does not originate or terminate within them. These ASes are often ISPs, providing connectivity to the global Internet for their customers. ISPs can form peering relationships with each other, where they mutually forward their customer traffic over common links.

Within an AS, routers communicate with each other through the process of intra-domain routing. This is accomplished using an interior gateway protocol (IGP) such as the Open Shortest Path First protocol (OSPF) or the Intermediate System to Intermediate System protocol (IS-IS). ASes communicate routing information via an external gateway protocol (EGP). The de facto standard EGP in use on the Internet is BGP version 4 [2].

A router running the BGP protocol is known as a BGP speaker. BGP speakers communicate across TCP and become peers or neighbors. By employing it, BGP does not need to provide error correction at the transport layer. Each pair of BGP neighbors maintains a session, over which information is communicated. BGP peers are often directly connected at the IP layer; that is, there are no intermediate nodes between them. This is not necessary for operation, as peers can form a multi-hop session, where an intermediate router that does not run BGP passes protocol messages to the peer. This is a less commonly seen configuration. BGP peers within the same AS (internal peers) communicate via internal BGP (IBGP). External BGP (EBGP) is used between speakers in different ASes (external peers). The routers that communicate using EBGP, which are connected to routers in different ASes, are called border routers.

Each AS originates one or more prefixes representing the addresses assigned to hosts and devices within its network. BGP peers constantly exchange Network Layer Reachability Information (NLRI) — the set of known prefixes and paths for all destinations in the Internet — via UPDATE messages. Each AS advertises the prefixes it is originating to its peers.

Additionally, all ASes update their routing tables based on their neighbors' NLRI, and forward the received information to each of their other neighbors.

This process ensures that all ASes are informed of the reachability of all prefixes. For as long as the session is active, peers use UPDATE messages to inform each other of routing table changes, which include the addition of new routes and withdrawal of old ones.

BGP is a path vector protocol. ASes establish an AS path for each advertised prefix. The paths are vectors of ASes that packets must traverse to reach the originating AS. Path vectors are stored in a routing table and shared with neighbors via BGP. It is ultimately this information that is used to forward individual packets toward their destination.

ASes are assigned an AS number (ASN) in a similar manner, with ICANN being the ultimate authority for delegating numbers. ASNs are used to identify the AS, and can be public or private. Public ASNs appear in BGP path vectors and are globally visible. Private ASNs can be assigned by an ISP to a customer that does not want to administer its own globally visible AS but wants to perform BGP peering with the provider, to gain benefits such as traffic engineering over multiple links.

ASes are not only bound by physical relationships; they are also bound by business or other organizational relationships. When an AS owner serves as a provider to another organization, there are associated contractual agreements involved. Such agreements are often defined by service level agreements (SLAs) which indicate the quality of service that the provider will guarantee. Therefore, for legal and financial reasons, it is necessary to be able to enforce SLAs at the routing policy level. BGP enforces routing policies, such as the ability to forward data only for paying customers through a number of protocol features. Principal among these is the assignment of attribute values in UPDATE messages.

Policies configured in a BGP router allow it to filter the routes received from each of its peers (import policy), filter the routes advertised to its peers (export policy), select routes based on desired criteria, and forward traffic based on those routes. For example, a transit AS may have

several peers. The BGP policy may be configured to only allow routes to transit the network if they come from peers who have signed a contract with the organization allowing transit service. BGP routers can be configured with route preferences, selective destination reporting (i.e., reporting a destination to some neighbors and not others), and rules concerning path editing. Setting policy often involves techniques to bias BGP's route selection algorithm. For example, one of the most significant criteria BGP uses for path selection is the length of an AS path vector. This length can be modified by an organization repeatedly adding it's AS number to a path, in order to discourage its use (a technique known as padding or prepending).

1.2 Analysis of Internet topology

A deep knowledge of the real Internet topology is fundamental in order to properly evaluate the behavior of current inter-domain routing architecture. Routing scalability parameters depend strongly on the characteristics (in particular, the size and the structure) of the underlying network topology.

Another important issue is related to the evolution of network topology: if no model is developed that can predict how the network will evolve, it is nearly impossible to perform analysis of routing behavior in the medium-long term.

For what concerns Internet topology, some remarkable observations have been exploited in well known studies that have revealed how many topology parameters have power-law distributions. In [3], power-laws are proposed as a tool to describe the Internet topology. The power-laws seem to capture concisely the highly skewed distributions of the Internet graph properties. Formulas have been derived that link the exponents of power-laws with graph metrics such as the number of nodes, the number of edges and the average neighborhood size.

The question of how closely the real Internet topology follows power laws remains still open to debate [4]. Nevertheless, the fact that Internet topology, at the granularity of Autonomous Systems (ASs), is a fat-tailed scale-free small-world network is not in dispute.

Both the fat-tail and scale-free properties refer to node degree distribution in a graph, but these two topology characteristics are completely independent:

1. The node degree distribution has a fat tail if there is a noticeable number of high-degree nodes ("hubs") in the graph.

2. The graph node degree distribution is scale-free if it lacks any characteristic scale, which effectively means that the graph has many low-degree nodes.

Node degree distributions in such random graphs are close to a narrow Poisson distribution centered around its average (characteristic scale) and exponentially decreasing at high degrees. The term *scale-free* is often used as a substitute for *power-laws*. Scale-free networks necessarily possess the "small-world" property. Indeed, even a moderate number of 'hubs' guarantees low average shortest path length (or simply the average distance) and low width of the distance distribution. For example, in the AS-level graph of the Internet, the average AS-hop distance is approximately 3.5 and the dispersion is around 1. Approximately 86% of AS pairs are at a distance of 3 or 4 AS hops.

The small-world property and other properties of scale-free networks are drastically affect traditional views, approaches and even methodology of network-related research. Unfortunately, in many cases (including the Internet case) other graph classes with quite different properties

have been traditionally assumed to be sufficiently accurate models for realistic networks. These models have little to do with reality, a gap that demands fundamental reexamination of what it is known about large-scale networks and in particular what depends on their topology.

These properties of the current Internet topology have consequences for what concerns the routing system. In fact, hierarchical routing, as it is used today, is known to not perform well on small-world networks like the Internet. Hierarchical aggregation requires topologies with distances growing quickly (e.g. polinomially) with the number of nodes, like Grids and Trees.

Effectiveness of hierarchical aggregation depends either on the abundance of remote nodes or on strong regularity of tree structures; none of these properties is present in scale-free graphs.

Applying hierarchical routing to an Internet AS-level topology may incur in a ~15-times path length increase [5]. This confirms that the knowledge of the current Internet topology has a fundamental impact on the best routing strategy to be used.

2. ANALYSIS OF CURRENT INTER-DOMAIN ROUTING ISSUES

This chapter contains an analysis of main issues related to inter-domain routing. They can be considered, from the perspective of the current Internet architecture, the actual requirements not (yet) addressed by the deployed inter-domain routing mechanism, based on BGP protocol.

In particular, issues described here are those more relevant from a network operator point of view.

2.1 Scalability issues

Scalability, in this context, refers to the ability of the routing architecture to deal with an increasing number of actors (Autonomous Systems) which actively participate to the global routing mechanism and an increasing number of objects (IP address prefixes) exchanged by those actors.

These numbers have a direct effect on routing platforms (routers) with an impact on hardware and software performances. An increasing number of IP prefixes means that a larger portion of memory is required on routers to store all that information and a higher CPU load is needed to verify the validity of the advertised routes. At the same time, as Internet topology is growing denser, also the number of routing messages tends to increase, stressing CPU performances.

The scalability issue has been the main driver of the discussion concerning the inter-domain routing system, culminated in the IAB Workshop [1]. Nevertheless, the initial attention was on the memory consumption and the risk that technology evolution (especially for what concerns the manufacturing of bigger and faster FIB (Forwarding Information Base) memories) would not be sufficient to sustain the DFZ (Default Free Zone, which corresponds to set of routers that need to store all routes advertised in the Internet) routing table growth (see Figure 1) in a cost effective way. In the last year, the stress on the memory issue has decreased, while concerns regarding the dynamic behavior of BGP have become the main scalability problem.

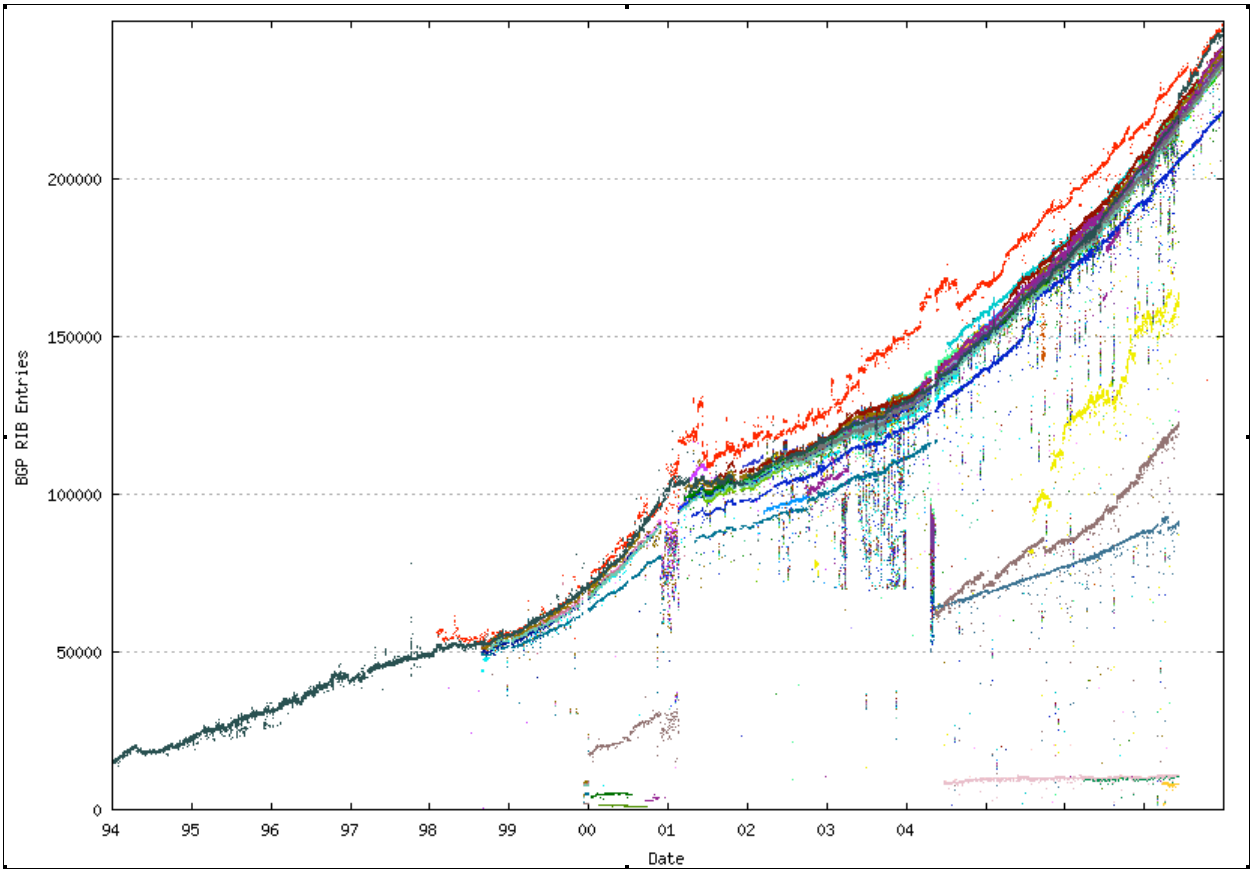

Figure 1: DFZ routing table growth (source: www.potaroo.net)

All routing protocols have to deal with the trade-off between scalability and speed of convergence. BGP makes no exception.

One of the problems closely related to BGP is that, while it has been widely used in the Internet, its behavior in a real-world environment is yet to be fully understood. There have been many BGP performance studies; a number of them focused on BGP's convergence behavior. A famous study by Labovitz and Ahuja [10] categorizes BGP routing events into four basic types: Tdown (a previously reachable destination is withdrawn), Tup (a previously unreachable destination is announced), Tlong (an existing path is replaced by a longer one) and Tshort (an existing path is replaced by a shorter one). They observed that Tup and Tshort events typically converge in a relatively short time period, but Tdown and Tlong events can trigger path explorations and take several minutes or more to converge. As shown by other studies, packet delivery performance is related, but not equivalent, to the routing convergence time. For example, in [11] BGP is used as a case study to examine packet delivery performance in the reaction to simple topological events of a loss of connectivity followed by a recovery (i.e., Tdown followed by Tup). For a given loss of connectivity to the destination, they use extra downtime and false uptime to measure the difference between the actual failure duration and the perceived reachability by the rest of the network provided by the routing protocol. Their results show that, although extra downtime closely matches Tup convergence delay and false uptime closely matches Tdown convergence delay, when failure durations are short, shorter Tdown convergence time may actually lead to reduced packet delivery. They also show that BGP's long MRAI (Minimum Router Advertisement Interval) timer adds substantial delay to routing convergence, suggesting a re-examination of the MRAI timer value in order to further improve packet delivery.

Chang [12] studied the effects of large BGP routing tables on commercial routers. They showed that some routers would reset one or all of the BGP peering sessions when they run out of

memory and then repeat this behavior after they re-establish the BGP sessions. As a result, routing table size would oscillate. When such routers form a chain, the routing table oscillation would propagate. They also studied whether various existing mechanisms can prevent the BGP sessions from failing under large routing table load. This kind of study could be usefully integrated with tests performed on current routers, as BGP implementations have been deeply modified and improved since the study by Chang.

Other studies [13] provide in-depth analysis of BGP's behavior under stressful conditions in real operational environment. Through in-depth BGP log data analysis, they conclude that BGP stood up well during the Nimda worm attack; the majority of the network prefixes exhibited no significant routing instability. Their analysis, however, does reveal several weak points both in the protocol and in its implementation:

- First, it is an evidence that BGP peering does not work well over unstable network connectivity. Even though BGP peering sessions seem relatively stable over good connectivity of short distance, the common setting in today's operational Internet, a global routing protocol must be truly robust and perform well even under adverse conditions.

- Secondly, although Code Red and Nimda attacks mainly affected connectivity at certain edges, whose intermittent reachability rippled through to the rest of the Internet as rapid BGP update exchanges, it is an evidence that, with the current BGP design, a local change can indeed cause a global effect. A truly resilient global routing protocol must keep local changes local in order to scale.

- Finally, although BGP's slow convergence after a failure or route change has not been shown to significantly impact the Internet performance, during the worm attack, the slow convergence's "amplifier" effect made the superfluous BGP updates, such as those due to local connectivity changes, multiple-fold worse.

Memory consumption

Prefix de-aggregation is leading to a significant growth of the DFZ RIB (Routing Information Base), because topological aggregation is currently the only known practical approach to limit its size. The current growth is also raising some concerns about the capacity and the performances of memory subsystems needed to sustain it.

The growth rate has been growing at greater than linear rates for several years and many efforts have been put in by vendors in order to guarantee scalability to their products. As interfaces' line rates are also increasing, the technology needed to build big and fast memories is becoming more and more complex. The subsystem which requires the most sophisticated technology is the FIB, which has to perform look-up operations at rates imposed by the interface speed. Considering a 40 Gb/s interface (STM-256/OC-768) and 40 byte packets, the cycle time of the buffer memory at each port is required to be less than 8 ns. This limit can be even lower when the port speed to a switch fabric is higher than that of the line speed (usually twice). This is to overcome some performance degradation that otherwise arises due to output port contention and the overhead used to carry routing, flow control, and QoS (Quality of Service) information in the packet/cell header. This is still very challenging with current memory technology, especially when the required memory size is very large and cannot be integrated into the ASIC (Application-Specific Integrated Circuit).

The current technology trend seems to favor the adoption of solutions which use parallelism and low latency DRAM (RL-DRAM, Reduced Latency Dynamic Random Access Memory). This

approach is currently able to manage roughly 1-2 million prefixes with 40 Gb/s line rates and is supposed to reach 10 million prefixes at 100 Gb/s line rates in the next decade.

Technology seems to offer guarantees to sustain the current growth of the DFZ routing table and the increasing speed of routers' links. However, the cost of this technology will need to be monitored in order to maintain its level within boundaries acceptable for Service Providers' budget limitations. Moreover, unforeseen increases of the growth of the DFZ RIB could be challenging to tackle. As a consequence, there are growing concerns about how far this process can continue without severe cost burdens and technical limitations which would degrade the reliability of the Internet.

Dynamics and convergence

The growth of the DFZ routing table, apart from imposing requirements on RIB and FIB memory sizes, exposes the core of the Internet to the dynamic nature of the edges. De-aggregation leads to an increased number of BGP UPDATE messages injected into the DFZ. Consequently, additional processing is required to maintain state for the longer prefixes and to update the FIB. Although the size of the RIB is bounded by the given address space size and the number of reachable hosts, the amount of protocol activity required to distribute dynamic topological changes is not and the BGP UPDATE churn the network can experience is essentially unbounded. The analysis reported in [6] shows that the UPDATE churn, as currently measured, is heavy-tailed: a relatively small number of Autonomous Systems (ASes) or prefixes are responsible for a disproportionately large fraction of the UPDATE churn that we observe today.

Issues directly related to the dynamic behavior of the BGP protocol worsen this. In particular, it has been observed that BGP shows a sort of amplification effect on the number of UPDATE messages spread through the network. This is due to the *path hunting* mechanism of BGP, which requires each node to go through a number of sub-optimal states before discovering the optimal path toward a destination (see [7]). Clearly, a high number of UDATE messages means an increased computation load for processors placed on routers' linecards and it has a negative impact on Service Provider's investments, requiring frequent upgrade of installed hardware.

The underlying problem is that BGP's path selection algorithm merely tries all the available paths when it receives a withdraw message, regardless of whether these paths have already been invalidated by the withdraw message. Such exhaustive search not only results in long convergence time, but also produces a significant number of unnecessary BGP updates.

At the same time, path hunting mechanism has a negative effect on convergence times. The speed of convergence is the time it takes for a router's routing processes to reach a new, stable, "solution" after a change which takes place somewhere in the network. In effect, what happens is that the output of the routing calculations stabilizes -- the Nth iteration of the software produces the same results as the N-1th iteration. It is well recognized that BGP is slow in converging to a stable state after modifications that occur in the network. Anyway, the primary goal of a routing protocol is to deliver packets. Other routing protocol performance metrics, such as routing stability and convergence delay, are all important issues, however they should also be considered with respect to maximizing packet delivery. Ideally, a routing protocol should deliver packets to destinations as long as any valid path exists that can reach the destination. In today's Internet, network failures, caused by either planned maintenance or unexpected events, occur frequently [8][9]. The BGP protocol can adapt to failures by converging to a new set of valid paths. However, the routing adjustment may take a long time due to various delays in the propagation of UPDATE messages and the exploration of alternative paths (i.e., the path hunting

previously cited). As a result, the period of destination unreachability can be substantially longer than the time period of actual physical connectivity losses.

In this case, “slow” convergence means times from tens of seconds to several minutes. These times are incompatible with most services offered by Service Providers nowadays. Up to now, services with strict requirements in terms of packet loss, delay and jitter have been limited to single Autonomous System environments. For example, Service Providers usually offer voice and video services only in their domestic networks; multimedia services which span two or more Autonomous Systems are eventually available only for large enterprise customers and are tackled by special projects and ad hoc deployments.

The lack of inter-domain failover due to delayed BGP routing convergence can potentially become one of the key factors contributing to the “gap” between the needs and expectations of today’s data networks. Multi-homed failovers averages several tens of seconds and may trigger oscillations lasting minutes. These delays grow linearly with the addition of new Autonomous Systems to the Internet in the best case, and exponentially in the worst.

In addition, a change can "ripple" back and forth through the system. One change can go through the system, causing some other router to change its advertised connectivity, causing a new change to ripple through. These oscillations can work on for a considerable amount of time before exhausting themselves. It is also possible that these ripples never die out. In this situation the routing and addressing system is unstable and it never converges.

All these issues affect investments done by network operators. As the Internet grows, they have to periodically upgrade routing platforms in order to tackle the increasing impact of routing information. Those investments are not related to a growing business (which would justify them) but are a tax which needs to be paid in order to participate to the global routing mechanism, and guarantee the reachability of the entire Internet.

2.2 Traffic engineering issues

Network operators must have control over the flow of traffic into, out of, and across their networks. Operating a large IP backbone requires continuous attention to the distribution of traffic over the network. Equipment failures and changes in routing policies in neighboring domains can trigger sudden shifts in the flow of traffic. Flash crowds caused by special events and new applications can also cause significant changes in the load on the network. Network failures and traffic fluctuations degrade user performance and lead to inefficient use of network resources.

However, the Border Gateway Protocol does not facilitate common traffic engineering tasks, such as balancing load across multiple links to a neighboring AS or directing traffic to a different neighbor (see [14]). Solving these problems is difficult because the number of possible changes to routing policies is too large to exhaustively test all possibilities, some changes in routing policy can have an unpredictable effect on the flow of traffic, and the BGP decision process implemented by router vendors limits an operator's control over path selection.

BGP provides networks with a limited set of mechanisms to achieve this control (e.g. LOCAL PREFERENCE for outbound control, MED and AS-PATH pre-pending for inbound control). However, these mechanisms only offer ISPs unilateral control over traffic. Unfortunately, unilateral decisions of neighboring networks may have undesirable interactions, and may result in unstable routing [16], poor performance [17], and huge, unpredictable shifts in network traffic volumes [18].

Network operators adapt to changes in the distribution of traffic by adjusting the configuration of the routing protocols running on their routers. Additionally, routing configuration changes are often necessary after deploying new routers and links. In practice, though, most traffic in a large backbone network traverses multiple domains, making inter-domain routing an important part of traffic engineering. The need for inter-domain traffic engineering can be motivated with some examples:

- Congested edge link: The links between domains are common points of congestion in the Internet. Upon detecting an overloaded edge link, an operator can change the interdomain paths to direct some of the traffic to a less congested link.

- Upgraded link capacity: Operators of large IP backbones frequently install new, higher-bandwidth links between domains. Exploiting the additional capacity may require routing changes that divert traffic traveling via other edge links to the new link.

- Violation of peering agreement: An AS pair may have a business arrangement that restricts the amount of traffic they exchange; for example, the outbound and inbound traffic may have to stay within a factor of 1.5. If this ratio is exceeded, an AS may need to direct some traffic to a different neighbor.

The state of the art for inter-domain traffic engineering is extremely primitive. The IETF's Traffic Engineering Working Group, now closed, focused almost exclusively on intra-domain traffic engineering, and noted that inter-domain traffic engineering "is usually applied in a trial-and-error fashion. A systematic approach for inter-domain traffic engineering is yet to be devised" [15]. This means that operators make manual changes in the routing policies without a good understanding of the effects on the flow of traffic or the impact on other domains.

In [14] it is observed that guidelines and good practices can alleviate some problems, like limiting the influence of neighboring domains and reducing the overhead of routing changes.

Nevertheless, BGP doesn't provide enough tools to obtain these results without relying too much on manual configuration.

2.3 Policy management and configuration issues

BGP has had success as a policy-based inter-domain routing protocol. The flexibility with which polices can be specified and enforced has enabled ISPs and other organizations to fine tune their interaction, which has helped to support a more reliable and predictable Internet. Nevertheless, the policy management mechanisms of BGP also have some drawbacks from the Service Providers' point of view.

BGP-speaking routers make routing decisions and propagate routing messages based on local configuration in the hope that the resulting path assignments will provide stable, global connectivity (see [21]). BGP routing involves local decisions made by routers based on flexible policies, which means that BGP's behavior is largely defined by its configuration. Controlling BGP's behavior by manipulating router configuration is a double-edged sword: its flexibility allows operators to use BGP to accomplish a wide range of tasks, but it also means that misconfigurations can, and do, cause significant problems. Even one misconfiguration can cause a complex sequence of errors that adversely affect global connectivity and are difficult to debug.

BGP offers great flexibility for the setting of routing policies. Because BGP's path selection is based on an AS's local preferences, rather than shortest paths, a group of ASes can have preferences that cause BGP to oscillate forever [19]. Even when given stable inputs, BGP might never converge. Griffin et al. [20] showed that, in general, determining whether a set of ASes would experience a policy dispute is an NP-complete problem.

Today, network operators, protocol designers, and researchers typically reason about BGP's behavior by observing the effects of a particular configuration on a live network. The state of the art for router configuration typically involves logging configuration changes and rolling back to a previous version when a problem arises. The lack of a formal reasoning framework means that router configuration is time-consuming. Furthermore, today's routing configuration is based on the manipulation of low-level mechanisms (e.g., access control lists, import and export filters, etc.), which makes routing configuration tedious, error-prone and difficult to reason about.

Currently, at most static constraint checking is used to verify the consistency of router configuration, but it has several shortcomings. First, since static analysis requires checking router configurations against many rules, exhaustively specifying every possible rule is manual and tedious.

Second, certain configuration pitfalls may only be exposed as a result of a specific message pattern or failure scenario and are not evident from static analysis alone. For example, visibility can be violated if a neighboring AS does not send routing advertisements with consistent attributes on all BGP sessions with its neighbor.

Configuration checking can help network operators find errors and mistakes in configuration, but these techniques are only treating the symptoms of a more fundamental problem: today's routing configuration languages are based on low-level mechanisms, rather than operator intent. For example, to specify that routes heard from one AS should not be re-advertised to another, a network operator must correctly define access control lists, import and export policies, and communities across multiple routers.

2.4 Security issues

Inter-domain routing involves thousands of Autonomous Systems, administrated by different organizations, which exchange hundreds of thousands routes through a number of different routing policies, independently set and configured by ASes. In spite of such a complex scenario, the BGP protocol currently offers limited guarantees in terms of security. This is due to the fact that the Internet was designed to enable communication between largely trusted peers. Similarly, the BGP protocol was designed to enable inter-domain routing between trusted networks. However, the success and the growth of the Internet have caused increasing commercial interests, which, in turn, have changed the nature of the Internet itself. Hence, assumptions of trustiness present in the Internet's original design are no longer valid.

Nevertheless, it still appears difficult to improve BGP security: some studies (see [22]) suggest that this lack of security mechanism is due to the fact that inter-domain routing must support complex policies.

Others [23] suggest that there are three primary limiting factors of BGP that lead to the vulnerabilities:

- BGP does not protect the integrity, freshness and origin authentication of messages.
- BGP does not validate an AS's authority to announce reachability information.
- BGP does not ensure the authenticity of the path attributes announced by an AS.

Several threat models regarding BGP have been published; they provide an outline of the different types of attacks that can be brought by means of the protocol itself (see [23] and [24]). Some of these attacks can occur on the BGP session between two ASes; others can impact ASes and network far from the attacking point, by leveraging the fact that BGP is a distributed protocol run by hundreds of thousands of routers and that each AS is indirectly connected to every other AS in the Internet.

Attacks against confidentiality. The routes exchanged by two ASes by means of the BGP protocol can help inferring the business relationship between the two organizations. In some cases, not always, this is considered a violation of confidential trade secrets. The BGP protocol, using TCP as an underlying transport, is subject to attacks that aim at observing the traffic between the two parties (e.g., TCP session hijacking).

Attacks against message integrity. In this case, the attacker is able not only to observe packets between the two ASes but also to modify or delete packets. The objective can be the misbehavior of the BGP protocol, got confused by the modifications to the routing packets (e.g., re-announce of withdrawn routes or withdrawal of valid routes).

Session termination. The objective of this kind of attack is to force the closure of the BGP session, which causes the reset of the session itself. The main effect is a decrease of the stability and availability of the routing system. Moreover, if the number of exchanged routes is large, the reset of the BGP session could have an impact on CPU load.

Fraudulent origin attack. An AS can advertise prefixes which don't belong to it (prefix hijacking), claiming to be the originator AS. This potentially causes the traffic to be routed to the wrong destination AS, which may want to discard it. The same effect can be obtained by

announcing de-aggregated routes belonging to a larger prefix. As routers perform longest prefix matching, the longest (de-aggregated) prefix is always chosen as the preferred path; as a consequence, if the de-aggregated prefix has been injected fraudulently, the malicious AS receives all the traffic destined to that prefix.

Corruption of path information. This type of attack implies the alteration of the information contained in an UPDATE message. All BGP attributes can potentially be altered (e.g., AS-PATH, Multi-exit discriminator, etc.), causing unwanted modifications of the path taken by traffic.

Denial of service attacks. Routers which act as BGP speakers can also be attacked by means of Denial of Service attacks. These can leverage vulnerabilities of the TCP connection (e.g., TCP RESET attack, TCP SYN flood attack, etc.) to cause outages of BGP routers. Due to the distributed nature of the protocol, consequences are even worse, because if a router goes down and then comes up again, the routing table needs to be re-created and prefixes need to be re-announced through the BGP protocol.

Misconfigurations. Besides intentional attacks, BGP is vulnerable to misconfigurations, whose effects are often the same as an attack. As it has been explained in the previous paragraph, BGP is complex to configure and even minor errors can create widespread damages. A previous analysis of BGP misconfigurations ([25]) found that in the course of a day, 0.2-1% of all prefixes in the global routing table are misconfigured. Two forms of misconfigurations were identified: a router exports a route it should have filtered (*export misconfiguration*) or an AS accidentally injects a prefix into the global BGP tables (*origin misconfiguration*). An example of router misconfiguration that led to widespread damage occurred in October 2002 with the ISP WorldCom. Improper filtering rules added to a router caused the routing tables of WorldCom's internal infrastructure to become flooded with external routing data. Faced with this additional burden, the internal routers became overloaded and crashed repeatedly. This caused prefixes and paths advertised by these routers to disappear from routing tables and reappear when the routers came back online. As the routers came back after crashing, they were flooded with the routing table information by their neighbors. The flood of information would again overwhelm the routers and cause them to crash. This process of route flapping served to destabilize not only the surrounding network, but the entire Internet. Prefix de-aggregation can allow adversaries to take over a prefix by advertising a more specific prefix block. This can be due to a malicious attack but also to misconfigurations. The classical example occurred in 1997, when misconfigured routers in the Florida Internet Exchange (AS7007) de-aggregated every prefix in their routing table and started advertising the first /24 block of each of these prefixes as their own. This caused backbone networks throughout North America and Europe to crash, as AS7007 was overwhelmed by traffic and the routes it advertised started flapping. This was not a malicious attack, but a mere error made by the network operators. Consider that a well-planned, targeted, malicious attack on BGP could do very serious harm to the network infrastructure.

The consequences of attacks, both intentional and unintentional, are diverse and can have many ramifications. For example, an individual router is subject to being overloaded with information, knocked offline or taken over by an attacker. An autonomous system can have its traffic blackholed or otherwise misrouted, and packets to or from it can be grossly delayed or dropped altogether. Malfunctioning ASes harm their peers by forcing them to recalculate routes and alter their routing tables. As the previous examples have shown, these events can disrupt international backbone networks and have the potential to bring a large part of the Internet to a standstill. From the individual level of an organization's traffic being stolen to the worldwide scale of IP traffic being globally subverted, the threats against BGP are a matter of grave concern to anybody reliant on the Internet.

3. DEFINITION OF USEFUL METRICS

The analysis of current issues related to inter-domain routing, discussed in the previous chapter, is characterized by an approach which is fundamentally qualitative. This means that it lacks clearly stated parameters to analyze and compare possible solutions to the described problems.

The present chapter tries to overtake these limitations, by defining a set of metrics, or measurable parameters, which should be useful to quantify how many issues are satisfied by solutions being proposed in the research community.

3.1 Internet topology

Inter-domain routing effectiveness strongly depends on Internet topology. Evolution of routing protocols should constantly look at current trends in the development of the Internet. In that sense, it would be useful to summarize the main characteristics of Internet evolution by means of an agreed set of topological metrics.

Number of entities (Autonomous Systems): is the total number of Autonomous Systems present in the Internet.

Mean number of neighboring ASes: is the mean number of ASes which an AS is connected to; it indicates how meshed is the Internet.

Mean number of paths between two end-points (leaf ASes): indicates how many paths are available to connect two end-points in the Internet.

Number of transit relationships: this parameter quantifies the number of transit relationships with respect to the total number of relationships between ASes.

Number of peering relationships: this parameter quantifies the number of peering relationships with respect to the total number of relationships between ASes.

Mean number of AS traversed to reach a destination: this corresponds to the *stretch* parameter used to describe the Internet topology.

3.2 Scalability

Metrics related to scalability issues are probably among the most important, as they allow measuring the current status of suitability of routing hardware with respect to the amount of information which has to be exchanged.

Taking into considerations the metrics defined in the previous chapter, it is also clear that, at present, it is only possible to use today's architectural elements, like Autonomous Systems and IP address prefixes, as quantities for describing requirements. For example, the Internet can be expected to grow to tens or hundreds of thousands of ASes and tens of millions of prefixes.

Clearly, any new architecture may eliminate some current architectural elements (for examples, ASes) and introduce new ones. As a consequence, targets may need to be reformulated in order to use the correct quantities.

Number of prefixes in the DFZ routing table: is the number of IP prefixes stored in the RIB and in the FIB of routers inside the DFZ.

Mean number of paths per prefix: is the number of different paths seen by an AS to reach a destination.

Mean number of prefixes exchanged between two peers: this parameter is useful to understand how big is the DFZ with respect to the entire Internet.

Number of bytes per prefix: determines the efficiency in storing the routing information of a single prefix.

Number of bytes per path: same as above, but with regard to a single path.

Protocol dynamics are another fundamental aspect related to scalability. As a consequence, the availability of a useful set of metrics concerning dynamic behavior of the routing protocol are indeed necessary.

Mean number of update messages received for a single topology change: this parameter refers to the amount of messages generated by the routing protocol and gives an indication of the amplification effect of the protocol itself.

Number of update messages received per second: is a measure of the level of instability of the network and also of the ability of the routing protocol to mask network events.

Number of prefixes contained in a single update message: currently it refers to the packing capability provided by BGP but, more in general, indicates the ability of the routing protocol to put more information in a single message.

Measure of blackholing: the quantity of network traffic blackholed in case of a network change which originates a protocol's re-convergence process.

Measure of convergence time: is the measure of the mean time that the routing protocol requires to reach a new stable state after a network change.

Measure of false uptime and extra downtime: as explained previously, this is the measure of the "inefficiency" of the routing protocol in adopting an available path to route the traffic after a network change.

3.3 Security

Most mechanisms currently proposed to improve security in the inter-domain routing architecture create concerns for ISPs due to impact on convergence and scalability.

Metrics related to security should be able to evaluate the performances of the BGP protocol when the security mechanisms are applied.

In particular, it is interesting to evaluate the CPU load in case of adoption of an encryption mechanism to secure the session between peers, or CPU load and memory occupation when a PKI is used to verify authorization.

4. DEFINITION OF REQUIREMENTS

After having described the main issues related to current inter-domain routing infrastructure and after having outlined a list of metrics for measuring specific properties/performances of the routing system, this chapter contains some requirements which address the issues discussed before by means of the mentioned metrics.

The list of requirements doesn't want to be exhaustive (for more extensive lists of requirements concerning inter-domain routing, see **Error! Reference source not found.** and **Error! Reference source not found.**), but it is focused on the main problems which currently affect inter-domain routing from the point of view of network operators.

4.1 Scalability

Scalability is, in general, one of the most important concerns for ISPs. Also when referring to inter-domain routing, scalability is a primary issue, as it has been described previously. As a consequence, network operators really need a routing architecture which is scalable in a "definitive way". That means that they would really like to not worry about scalability at all and upgrade their networks only to support more traffic (which means more revenues) or to enable new services (which means more revenues, as well). The necessity to upgrade network platforms only to sustain the routing system is currently a big drawback. Network operators can't continue to add more computing power to routers forever. Other strategies must be available.

In this paragraph, scalability requirements are discussed together with performance requirements. This could sound strange, as all routing protocols are based on a trade-off between scalability and dynamic performances (convergence times, etc.). Nevertheless, scalability, in a broader sense, not only refers to the ability to constrain resources while the network grows but also involves the necessity to guarantee adequate performances as the number of network objects increases and the network itself becomes more complex. Hence, both requirements can usefully be discussed together.

Solving the scalability issue has proved to be impossible till now also because we really know little about the dynamics of Internet growth, so we can hardly predict how it will evolve in the future. Moreover, changes to the routing architecture are likely to have an impact on the evolution of the Internet itself, making the problem even more complex.

As a general statement due to this lack of knowledge, the routing architecture should not make any assumptions about the topology of the Internet or presume a particular network structure. The network does not have a "clear" structure; in fact, relationships between Autonomous Systems tend to change as economic factors suggest new and more profitable interconnection models.

It is clear that only a completely new routing system could eventually solve the scalability problem in a definitive way. Moreover, even if such a routing system were found, there would be the issue of deploying it in a cost effective manner and without disrupting the current services and economies based on the Internet.

Bringing back the discussion to the current inter-domain architecture, given the foreseen growth in terms of ASes and IP prefixes (hundreds of thousands of ASes and tens of millions of

prefixes), the routing architecture should have the ability to constrain the increase in load (CPU, memory space and network bandwidth) on any single router to be less than the current growth of hardware capabilities, as described by Moore's Law, doubling every 18 months.

In terms of metrics, memory consumption needs to be reduced: this implies a reduction of the number of prefixes advertised, for example through an improved aggregation strategy, or fewer bytes used to memorize entries in the FIB, for example by adopting more efficient encoding schemes.

Observations about evolution of the Internet also indicate that the degree of meshiness is growing. That means that there are an increasing number of alternative paths to reach Internet destinations. This increase of the interconnectivity level must be managed efficiently by the routing system in order to safeguard routing tables, computation times and routing protocol traffic and to prevent them from growing without bounds.

One cause of meshiness growth depends on the increasing adoption of multi-homing as an interconnection strategy between Autonomous Systems. Among the reasons to multi-home there are reliability, load sharing and performance. Multi-homing should be supported by the routing system without impacting negatively on scalability. One problem of the current routing architecture is that multi-homing leads to bigger BGP routing tables. Hence, the adoption of an efficient way to manage multi-homing could have a positive effect on the size of routing tables, with benefits also for ISPs.

Anyway, there is a lack of reliable measures about the ratio of de-aggregated routes due to multi-homing. This is a gap that should be filled in order to understand the benefits of solutions which address that problem. From an ISP's perspective, the problem of multi-homing is essentially related to the de-aggregation effect: investments directed to solve this problem would be justified only if multi-homing were the primary cause of the growth of the DFZ routing table. Otherwise, multi-homing would be essentially a requirement for End-Users and it should not impose new investments to ISPs.

The speed of convergence is another important issue related to routing dynamics. It is generally considered to be a function of the number of networks and the amount of connections between those networks. As either numbers are growing, the Internet evolution tends to increase convergence times.

In that sense, the first requirement is to guarantee that convergence can be reached by the system. When a single change of any type (link addition or deletion, router failure or restart, etc.) is introduced into a stabilized system, the system should be able to re-stabilize within a bounded time. This requirement is a fairly abstract one as it would be impossible to test in a real network. Even worse, defining a single target for maximum convergence time for the real Internet is absurd.

Anyway, the growth rate of the convergence time should not be related to the growth rate of the Internet as a whole. This implies that the convergence time should not be a function of basic network elements (such as prefixes and links/paths), and that the Internet should be continuously divisible into portions that limit the scope and effect of a change, thereby limiting the number of routers, prefixes, links and so on involved in the new calculations.

Even if requirements on convergence times can be difficult to set, it is legitimate for ISPs to define requirements concerning packet delivery. Performances of packet delivery in case of network changes (i.e., faults, network upgrades, routing policy changes, etc.) should progressively meet requirements comparable to those currently imposed on internal networks:

packet losses should be constrained from tens or hundreds of milliseconds to a few seconds (for more infrequent outages). These values are necessary to let operators extend their services from a simple domestic network scenario to a complex multi-AS scenario (i.e., IPTV service which spans multiple ASes, etc.).

More details about the inter-domain topology would probably help achieving better packet delivery performances.

False uptime and extra downtime should be reduced as much as possible, in order to limit, or possibly avoid, unnecessary packets lost.

In order to improve scalability, the routing architecture should have a mechanism to limit the scope of any one change's visibility and effects. In that case, the number of routers that have to perform calculations in response to a change could be kept small. Routers within that scope would know of the change and recalculate their routing tables based on that change. On the contrary, routers outside of the scope wouldn't see it at all. Anyway, this could be considered a possible solution to scalability rather than a requirement. Nevertheless, the existing hierarchy among ISPs should be better exploited to meet scalability requirements: the “transit service” offered by bigger ISPs should include a sort of filtering of routing instability towards their customer ISPs.

In case of particular stress conditions (for example, see [13]), the protocol should also be able to react to an increasing instability in the network by reducing the update rate. In other words, a feedback mechanism could be introduced in order to regulate the dynamic behavior of the protocol on the basis of the stability (or instability) level of the network.

4.2 Traffic engineering

The ability to perform traffic engineering is critical for ISP. Traffic engineering is, at base, another alternative or extension to path selection mechanisms offered by the routing system. No fundamental changes to the current mechanisms are needed, but the iterative processes involved in traffic engineering could benefit from some additional capabilities and state in the network.

As the network is becoming increasingly complex, with private peering arrangements set up between providers at every level of their hierarchy and even by certain large enterprises, the routing system must facilitate traffic engineering of these peer routes so that traffic can be readily constrained to travel as the network operators desire, allowing optimal use of the available connectivity.

Traffic engineering typically involves a combination of off-line network planning tools and administrative control functions which examine the expected and measured traffic flows and calculate changes to static configurations and policies in the routing system. During normal operations, these configurations control the actual flow of traffic and affect the dynamic path selection mechanisms; the results are measured and fed back to network planning.

The routing system must be able to generate statistical and accounting information in such a way that traffic engineering and network planning tools mentioned above can be used in both real time and off-line planning and management.

The routing system should also support the controlled distribution of traffic over multiple links or paths toward 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. Policies can be used to disseminate the attributes and to classify traffic for the different paths.

The increasing meshiness of the Internet should be exploitable by ISPs through more efficient traffic engineering mechanisms. One of the techniques used by BGP to improve scalability is that each router selects a single best route for each destination and advertises this route to its peers. As a consequence, many available paths are hidden by the routing system and cannot be used for traffic engineering purposes. This trade-off should be improved in order to enable more flexible policies.

ISPs would also benefit from the availability of “high-level” languages to define rules for classifying the traffic and assigning classes of traffic to different paths (or prohibiting it from certain paths). The rules should allow traffic to be classified based upon at least the following attributes:

- IPv6 FlowIDs
- DSCP (Differentiated Service Control Points) values
- source and/or destination prefixes
- source and/or destination transport ports
- random selections at some probability

A mechanism is also needed that allows operators to plan and manage the traffic load on the various paths. To start, this mechanism could be semi-automatic or even manual. Eventually it ought to become fully automatic in order to ease configuration tasks.

When multi-path forwarding is used, options must be available to preserve packet ordering where appropriate (such as for individual TCP connections).

The routing system should be able to support forwarding over multiple parallel paths when available. This feature is required when the offered traffic is known to exceed the available capacity of a single link and when it is advisable to share the load over multiple paths for cost or resiliency reasons.

The routing system should also have mechanisms to allow the traffic to be reallocated back onto a single path when multiple paths are not needed.

Moreover, Service Providers would benefit from the availability of a wider set of metrics related to inter-domain paths and topology. While the current path selection mechanism is entirely based on policies, the improved metrics should also be related to physical properties of the paths in order to add "engineering" parameters to the selection of inter-domain routes. This would have benefits also to guarantee adequate QoS properties to traffic which spans over multiple ASes.

4.3 Policy management and configuration

The basic requirement related to policy management and, more in general, to the configuration of inter-domain routing is to require less manual configuration than today.

The current practice involves the configuration of ingress policies to set relevant attributes to incoming routes and the configuration of egress policies based on the attributes previously set. This means that many routers inside an AS have to work in a coordinated way in order to correctly exploit the configured policies.

Relying only on human inspection to verify to correctness of routing configuration is not sufficient: in order to verify the solidity of routing policies, automatic verification tools are needed. They should also exploit statistical or Artificial Intelligence approaches (i.e., Bayesian networks) to guess what correct configuration looks like. For example, if an AS has several hundred routers and all but a few are configured in a certain way, more likely than not the deviations are mistakes. Applying these techniques to router configuration might identify a variety of errors, from simple mistakes, such as missing statements, to more subtle errors, such as misconfigured policies. Network operators need tools based on systematic verification techniques to ensure that BGP operational behavior is consistent with the intended behavior (i.e., that the network is operating correctly).

BGP filters used to select which routes can be received by peers should be configured and maintained automatically, through the use of reliable repositories of policies (i.e., databases maintained by RIRs, etc.).

ISPs often use defensive configuration (e.g., by actively setting the same MED, origin type, and next-hop on incoming routes) to prevent possible attacks from outside but precisely determining vulnerability requires emulating exactly what happens when its neighbors send it a particular set of routes. Similarly, certain property violations may only arise under specific failure scenarios; for this reason, network operators need tools (i.e., a dynamic simulation and emulation environment) to test configuration robustness.

These tools could be off-line tools used to check the configuration whenever it changes or better they could be integrated in the protocol itself. This would be a sort of auto-checking mechanism of the routing protocol to verify that all the components of a certain routing policy have been configured correctly and that they will perform the expected behavior. The routing protocol should also provide mechanisms to avoid (or, at least, to detect) situations where policy configurations make the system unstable [19].

The protocol should also carry more information to help operators to diagnose problems arising at the inter-domain level (for example, by pinpointing cause and location of a routing change). Currently, this kind of diagnosis can rarely be performed because too much information is hidden by the protocol to guarantee scalability. Again, being the inter-domain routing architecture a complex distributed system, management performed through centralized off-line tools tends to become difficult to be used and too slow (it also has to be scalable as the number of network appliances grows). As a consequence, it would be extremely useful if the protocols could integrate functionalities to ease the management of the system and the diagnosis of problems in order to enable fast reactions.

For what concerns configuration, a more effective approach would be to allow an operator to specify high-level policies without having to worry about the details of how the policy is actually

implemented. Also, the configuration itself should be generated automatically, avoiding issues related to vendor specific syntax. Because checking today's routing configuration requires understanding how to specify operator intent, configuration checking is not only useful but can also improve ISPs' ability to use more robust configurations, with benefits for the entire Internet.

4.4 Security

The Internet is experimenting an increasing trust problem between end users making use of the network, between users and their suppliers of services (i.e., network operators, service providers, etc.), and among the multiplicity of providers. Hence security, in all its aspects, is fundamental in the inter-domain routing architecture.

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

From an ISP perspective, requirements related to security aspects should address the three main issues described in paragraph 2.4:

- Protection of the integrity, freshness and origin authentication of messages.
- Validation of an AS's authority to announce reachability information.
- Ensuring the authenticity of the path attributes announced by an AS.

This may require fundamental changes to the current BGP behavior. For this reason, the deployability of new solutions needs to be carefully evaluated. It's fairly impossible to imagine the deployment of mechanisms which require disruptions of the current architecture.

As a first step, the routing communication needs to be secured. But securing the BGP session, as done today, only secures the exchange of messages from the peering domain, not the content of the information. In other words, it is possible to confirm that the received information is what the neighbor really sent, but there's no way to know whether this information (that originated in some remote domain) is true or not.

The communicating entities must be able to identify who sent and who received the information (authentication) and the communicating entities must be able to verify that the information has not been changed on the way (integrity). ISPs should also have the possibility to verify that the path followed by traffic actually corresponds to the route advertised by peers. This is not only a problem of trustiness but also of policy checking, as discussed in the previous paragraph.

A decision has to be made on whether to rely on chains of trust, or whether it is also needed authentication and integrity of the information end-to-end. This information may include both routes and addresses. There has been interest in having digital signatures on originated routes as well as countersignatures by address authorities to confirm that the originator has authority to advertise the prefix. Even understanding who can confirm the authority is non-trivial, as it might be the provider who delegated the prefix (with a whole chain of authority back to ICANN) or it may be a regional registry.

The routing communication should also be robust against DOS attacks, if necessary by adopting more robust transport protocols.

Requirements which address potential malicious attacks should also be able to tackle problems caused by misconfigurations, in addition to specific requirements stated in the previous paragraph about configuration checking.

5. CONCLUSIONS

This document has presented in a concise way the main issues which currently affect the inter-domain routing architecture from a network provider perspective. Starting from these issues, the analysis defined a set of requirements in order to state what should be the goals of current research trends concerning the evolution of the routing architecture.

In the future, these requirements, and the issues which they have been derived from, will be usefully integrated or modified according to new achievements in the analysis of Internet behavior as a system. In that sense, this document is a "live document".

From a Service Provider's point of view, a careful cost-benefit analysis needs to be performed to evaluate which are the requirements that should have higher priority. Four main items have been analyzed: scalability, traffic engineering, policy management and security. Issues related to these items are probably the most relevant ones for Service Providers as they have direct impacts on both services and investments.

The solutions which target these requirements also need to be evaluated in terms of costs. Service Providers can afford new investments only if they solve relevant issues, thus compensating costs due to inefficiencies. Requirements which don't provide significant benefits for Service Providers should be addressed by solutions which don't require them to invest in new technologies.

The performed analysis has shown that major concerns by Service Providers are related to the behavior and the functionalities of the BGP protocol. In that sense, from a network provider perspective, the evolution of the BGP protocol itself is of primary interest.

Any modifications to the BGP protocol raise the problem of the deployability of solutions which affect the routing architecture of the entire Internet. This represents one of the main issues faced by current research activities. Anyway, this problem must not divert research efforts from exploring viable solutions to these issues.

Another important aspect is the knowledge of the Internet as a system. We currently know little about it and, as a consequence, current solutions are more patches which address specific symptoms rather than treatments targeted to the main problems. Hence, any efforts to improve the knowledge of the Internet, through theoretical speculations and through active measurements, should be more actively supported by ISPs.

6. REFERENCES

[1] D. Meyer, L. Zhang, K. Fall, "Report from the IAB Workshop on Routing and Addressing", IETF RFC 4984, September 2007.

[2] Y. Rekhter, T. Li, S. Hares, "A Border Gateway Protocol 4 (BGP-4)", IETF RFC 4271, January 2006.

[3] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," in ACM SIGCOMM, pp. 251–262, 1999.

[4] A. Broido and k claffy, "Internet topology: Connectivity of IP graphs," in SPIE International Symposium on Convergence of IT and Communication, 2001.

[5] D. Krioukov, K. Fall, and X. Yang, "Compact routing on Internet-like graphs," in IEEE INFOCOM, 2004.

[6] G. Huston, G. Armitage, "Projecting Future IPv4 Router Requirements from Trends in Dynamic BGP Behavior", ATNAC 2006.

[7] G. Huston, T. Li, "BGP Stability Improvements", IETF Internet Draft draft-li-bgp-stability-01, Work in progress, June 2007.

[8] G. Iannaccone, C.-N. Chuah, R. Mortier, S. Bhattacharyya, C. Diot, "Analysis of link failures over an IP backbone", in ACM SIGCOMMInternet Measurement Workshop (IMW), November 2002.

[9] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental study of internet stability and wide-area network failures", in Proceedings of FTCS99, June 1999.

[10] C. Labovitz, A. Ahuja, A. Abose, F. Jahanian, "Delayed internet routing convergence", in Proc. of ACM SIGCOMM, 2000.

[11] D. Pei, L. Wang, D. Massey, S. F. Wu, L. Zhang, "A study of packet delivery performance during routing convergence", in The International Conference on Dependable Systems and Networks (DSN), 2003.

[12] D.-F. Chang, R. Govindan, and J. Heidemann, "An empirical study of router response to large BGP routing table load", Tech. Rep. ISI-TR-2001-552, USC/Information Sciences Institute, December 2001.

[13] L. Wang, X. Zhao, D. Pei, R. Bush, et alii, "Observation and Analysis of BGP Behavior under Stress", Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement 2002.

[14] N. Feamster, J. Borkenhagen, J. Rexford, "Guidelines for interdomain traffic engineering", ACM SIGCOMM Computer Communications Review, October 2003.

[15] D. O. Awduche, A. Chiu, A. Elwalid, I. Widjaja, X. Xiao, "Overview and principles of Internet traffic engineering", IETF RFC 3272, May 2002.

[16] R. Mahajan, D. Wetherall, T. Anderson, “Towards coordinated interdomain traffic engineering”, in HotNets-III, 2004.

[17] N. Spring, R. Mahajan, T. Anderson, “Quantifying the Causes of Internet Path Inflation”, in SIGCOMM, Aug. 2003.

[18] R. Teixeira, T. Griffin, G. Voelker, and A. Shaikh, “Network sensitivity to hot potato disruptions”, in SIGCOMM, 2004.

[19] T. Griffin, G. Huston, “BGP Wedgies”, IETF RFC 4264, November 2005.

[20] T. Griffin, F.B. Shepherd, G. Wilfong, “The stable paths problem and interdomain routing”, IEEE/ACM Trans. Networking 10, 1 (2002), 232–243.

[21] N. Feamster, “Rethinking Routing Configuration: Beyond Stimulus-Response Reasoning”, Workshop on Internet Routing Evolution and Design, October 2003.

[22] N. Feamster, H. Balakrishnan, J. Rexford, “Some Foundational Problems in Interdomain Routing”, 3rd ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets), San Diego, CA, November 2004.

[23] S. Murphy, “BGP Security Vulnerabilities Analysis”, IETF RFC 4272, January 2006.

[24] K. Butler, T. Farley, P. McDaniel, J. Rexford, “A Survey of BGP Security”, Draft version, April 2005.

[25] R. Mahajan, D. Wetherall, T. Anderson, “Understanding BGP misconfiguration”, ACM SIGCOMM 2002, Pittsburgh, PA, USA.

[26] A. Doria, E. Davies, F. Kastenholz, “Requirements for Inter-Domain Routing”, IETF Internet Draft draft-irtf-routing-reqs-08, Work in progress, October 2007.

[27] E. Davies, A. Doria, “Analysis of IDR Requirements and History”, IETF Internet Draft draft-irtf-routing-history-06, Work in progress, October 2007.