Using BGP-4 to Migrate to a Future Internet

Pedro A. Aranda Gutiérrez[1], Petteri Pöyhönen[2],
Luis Enrique Izaguirre Gamir[1], and Francisco Huertas Ferrer[1]

[1] Telefónica, Investigación y Desarrollo,
Emilio Vargas, 6, Madrid Spain
[2] Nokia Siemens Networks,
Linnoitustie 6, 00260 Espoo Finland

Abstract. The Internet has evolved to become one of the most critical communication infrastructures in the planet. And yet, some of its underlying concepts and protocols do not provide the adequate level of reliability for such an essential role in global communications. The inter-domain routing protocol of the Internet, Border Gateway Protocol (BGP-4), is being used with varying degree of success for tasks for which it was not originally designed, such as Traffic Engineering. This paper presents a rationalised view of the different functions implemented by routing nowadays and proposes the use of Autonomous System Compartments. The Autonomous System (AS) Compartments imply a new routing hierarchy over the traditional BGP-4 routing, where specific functionalities like Traffic Engineering can be better controlled and additional routing incentives can be introduced. The FP-7 project 4WARD is working on new communication paradigms for the Future Internet and AS Compartments are a choice to contain the Generic Path (GP) concept developed by it. In order to provide inter-domain capabilities and a migration tool to connect GP islands, the multiprotocol mechanism of the BGP-4 routing is used. This paper presents the AS Compartment concept and the integration of Generic Paths in it, as well as an implementation of the GP-BGP concept for the J-Sim simulator (JSIM) environment.

Keywords: Autonomous Systems, Inter-Domain Routing, Compartments, Traffic Engineering.

1 Introduction

The Internet is being perceived as a commodity nowadays. On the other hand it truly is a critical communication infrastructure. Many services are used and contents are distributed over the Internet. Users are stationary when using their PCs from home and/or office and moving users when using their mobile devices. Especially with the adoption of new 3G radio access networks like High-Speed Downlink Packet Access (HSDPA) and wider adoption of mobile broadband, the difference between these two types of users in terms of access bandwidth is vanishing. For service providers this means a potentially larger amount of users and for the Internet naturally this means higher traffic transfer demands. While operators and other service providers introduce new exciting value added services,

K. Pentikousis et al. (Eds.): MONAMI 2010, LNICST 68, pp. 14–25, 2011.

this also requires a better communication infrastructure, especially in terms of availability and stability. When using the Internet to provide IP backbone connectivity between mobile operators, we face many challenges to ensure a similar stability compared to the GPRS Roaming Exchange (GRX) [1] service quality level. A global use of the Internet typically also involves the use of BGP-4 [2] based core routing. This core network provides a fairly resilient routing, but it is a well-known behaviour that it also could take relatively long time, i.e., tens or hundreds of seconds, until the routing system restores its stable state after a routing incident. Such an incident can occur due to a configuration error, network maintenance, (physical) link failure and so on. Routing system stability is perhaps one of the main challenges and is something that should be taken into account while considering "better than best effort" end-to-end services.

Despite the BGP-4 protocol and routing being well-defined, there are different deployment practices, which are derived from the need for traffic engineering in order to comply with peering agreements. Not all BGP-4 route attributes are used in all network domains in a consistent way. An example of this is the Multi-Exit Discriminator (MED) attribute [3], which provides a mechanism for an AS to indicate to adjacent ASs the optimal inbound link (e.g. in the case of multi-homing). Another example is the AS Path (AS_PATH) attribute. [4] further explains differences on the routing policy deployment and how they affect to the BGP-4 routing due to the diversity in processing BGP-4 messages.

The FP-7 4WARD project [5] has taken a Clean Slate approach to the Future Internet, exploring new insights in multi-access and resource management. One of the solutions which have emerged from this effort is the GP concept [6], as a new paradigm to support rich and flexible communication schemes. As all new technologies, the adoption cannot be expected to be instantaneous and unanimous across the Internet. The most reasonable scenario is a gradual adoption by smaller user groups, resulting in networking "islands" which need to be interconnected. This has been the case of IPv6 with the 6bone [7] and other migration mechanisms [8,9].

This paper presents a framework to interconnect GP islands using the Autonomous System Compartment concept and multiprotocol extensions to BGP-4. The rest of paper is structured as follows. Section 2 describes a new concept called AS Compartment and explains how this concept is used in inter-domain networking environment. Section 3 discusses the roles of traffic engineerig in the AS Compartments and describes a high level logical architecture. Section 4 presents the simulation experiments carried out to evaluate how a BGP-4 multiprotocol extension functions and performs on top of BGP-4. Section 5 provides conclusions and finally, section 6 outlines the related work.

2 AS Compartments

In order to improve BGP-4 routing resilience and to support more flexible ways of doing Traffic Engineering (TE), a new concept called AS Compartment is proposed. The concept introduces a new routing "hierarchy" on top of the BGP-4 routing system and instead of relying on the standard BGP-4 convergence, a fast re-routing is supported at the AS Compartment level. The AS Compartment routing complements the BGP-4 routing by also taking into account different end-to-end incentives and could use for instance multi-path routing between AS Compartments. In case of a multi-path routing, the authors of [10,11] define new multi-path routing protocols designed for inter-domain environment that could be used to implement the multi-path routing support in AS Compartments. An AS Compartment could be either a single AS or it can include a set of ASs. And therefore, an AS Compartment could represent one or more Autonomous System Numbers (ASNs) and could be configured for instance to use AS Confederations [12]. If an AS Compartment consists of multiple ASs, then it is assumed that each border gateway hosting a *path* end point has connectivity to each other inside the AS Compartment. The AS Compartments provide means to introduce a limited control over the BGP-4 infra without modifying the basic BGP-4 protocol and to separate traffic engineering and routing functions from each other. So one of the main challenges is to ensure that for certain type of traffic that is passing through the BGP-4 routing system the perceived connection quality is sufficient only with introducing additional functionality in a selected set of Autonomous Systems as illustrated in Figure 1.

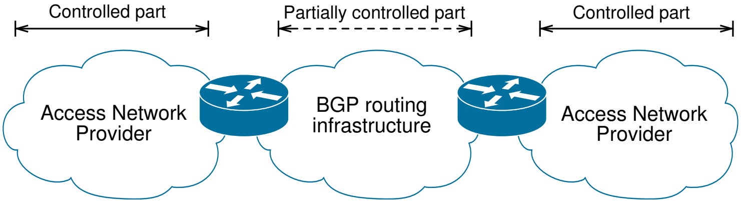


Fig. 1. An end-to-end connectivity over Internet core

An AS Compartment does not necessarily mean that QoS-aware routing is supported. Thus a simple form of AS Compartment could be a *best effort* network with some advanced routing and traffic engineering functionalities implemented. AS Compartments are interconnected by *paths* that are used to route traffic between AS Compartments. A *path* could be for instance IP tunnel having BGP-4 routable IP addresses as end points. This tunnel is terminated at border gateways located at the AS Compartments.

Figure 2 shows an example of two AS Compartments connected by 3 *paths* that are transported over normal BGP-4 routing implemented by AS topology. *Path 1* and *Path 2* are used to transfer traffic to the "remote" AS Compartment and *Path 3* is used to transfer traffic from the "remote" AS Compartment. The selection between *Path 1* and *Path 2* is done based on the underlying BGP-4 routing info. For instance, if *AS-n* indicates in its route advertisements with the MED attribute that one link should be preferred over another, then this is taken into account when selecting an active path. Additionally, this selection process could use any available BGP-4 routing information as long as BGP-4

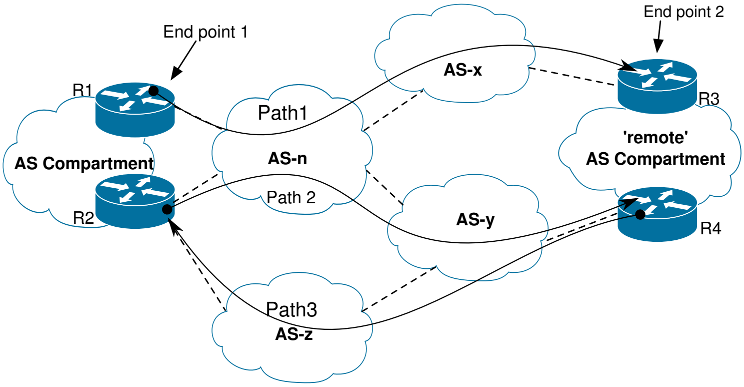


Fig. 2. An example of AS Compartments connected over the BGP routing

routing practices are honoured. In other words, AS Compartment routes are *paths* between Compartments and these routes exist only when corresponding BGP-4 routes are also present.

Concerning different types of relations between ASs, there are some exceptions we should consider. For instance, Tier1 and Tier2 ASs could have peering relations with other same level ASs. These peering agreements are not used for transit traffic due to the lack of economic incentives, thus the main motivation of their use is to minimize transit costs as well as minimize end-to-end latencies [13]. However, in some situations, it might be preferred to override this peering policy. For instance, to make it possible that two (or more) single homed ASs could create a multi-homed AS Compartment. This implies that the authorities creating an AS Compartment over a peering relation have common incentive(s) to allow the use of their peering link(s) for the selected transit traffic.

3 Traffic Engineering in AS Compartments

In order to implement new policies and to avoid contradicting routing policies between ASs, TE functions are separated from routing functions and a uniform TE process over the ASs is defined. Also, all ASs belonging into the same AS Compartment should contribute and implement "atomic" policies towards this AS Compartment. The objective of traffic engineering is to distribute traffic at AS peering points in such a way that they comply with the Service Level Agreements (SLAs) signed by each AS with its peers, assuming that SLAs are convertible from one AS to another one of the AS Compartment. Ideally, both input and output traffics should be controlled. In the current Internet, controlling the output traffic can be implemented internally in the AS, but controlling the input traffic can only be achieved by controlling the routing preferences in other ASs.

Basically, there are three kinds of attributes in the BGP-4 routing decision depending on their scope; 1) router local, 2) AS local, and 3) global [14]. The routing decision process during which the best path is computed is the well-defined process taking also into account the local policies. So for the inbound traffic, an AS can tweak the route attributes to be announced in hope of influencing a neighbour AS best path selection. For the outbound traffic, there are more powerful means available like the attributes representing local policies like the *local preference* attribute. There are also other attributes, conditions and local policies influencing to the routing decision like route type ("customer", "provider" or "peer"), an internal (Interior Gateway Protocol (IGP)) topology, the BGP-4 community attribute, and so on. Since the AS Compartments are operating on top of BGP-4 routing, they can coordinate how to handle both inbound and outbound traffic in order to comply additional routing incentives without making this visible at the BGP-4 level.

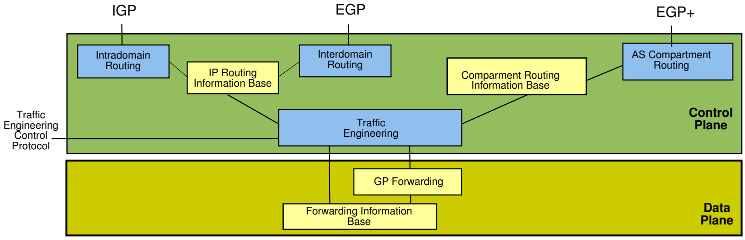

IGP
EGP
EGP+
Intradomain Routing
IP Routing Information Base
Interdomain Routing
Comparment Routing Information Base
AS Compartment Routing
Traffic Engineering Control Protocol
Traffic Engineering
Control Plane
GP Forwarding
Forwarding Information Base
Data Plane

Fig. 3. Extracting the TE functionality from the extended routing framework

Figure 3 shows a high level architecture for node integrating traditional IP routing with augmented GP routing and TE functions. The control plane integrates today's IP intra- and inter-domain routing functions, GP routing functions and a separate TE block.

4 Towards a Practical Implementation

One way to interconnect different GP islands over traditional IP based infrastructures is to use IP tunnel [15] having BGP routable addresses as end points. This tunnel is terminated at border gateways located at GP islands. Inter-GP island routing is established by defining the GP Network Layer protocol that defines the so-called GP-BGP-4, a new routing hierarchy that enables GP island to exchange their routes over the traditional BGP-4.

To enable BGP-4 to support routing for multiple Network Layer protocols, Multiprotocol Extensions for BGP-4 [16] adds the ability to associate a particular Network Layer protocol (e.g., IPv6, IPX, L3VPN, etc.) with the next hop information and Network Layer Reachability Information (NLRI). GP-BGP uses the multiprotocol extensions capabilities to exchange GP islands routes and achieve, together with IP tunneling, the inter-GP islands routing.

In order to provide a proof of concept and a first evaluation of the applicability of the proposed solution, the Generic Path-BGP-4 extensions have been implemented for a proof of concept on the JSIM [17]. Figure 4 shows a basic simulation environment with three nodes, the two in the edges running a GP-BGP

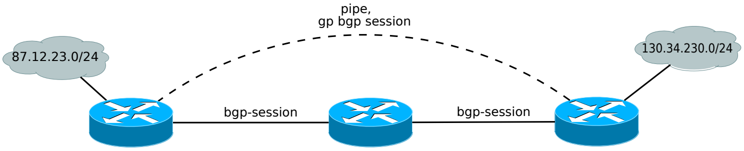


Fig. 4. GP-BGP Basic Setup

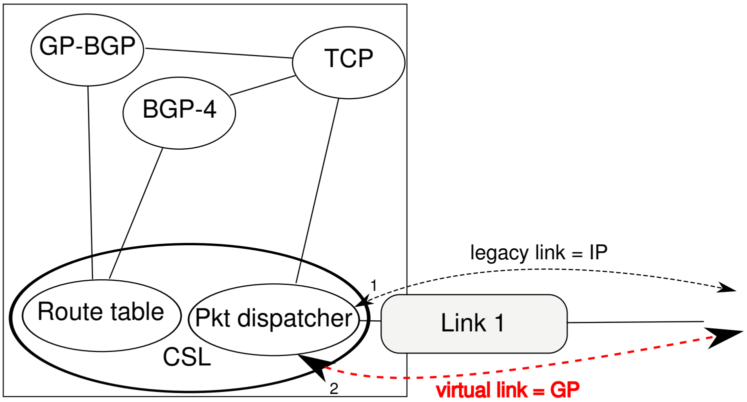


Fig. 5. Node structure in the simulation environment

session over an IP tunnel that basically follow the logical functionality decomposed in Figure 5. In order to obtain early results, routing compartments are simulated as IPv4 prefixes.

4.1 Fallback Scenario

There is however a situation where there is no loss of connectivity when a specific link or router failure happens, and this is when the AS Compartment concept is fully exploited. Figure 6 shows the simulation environment in that case. It shows a topology where the border gateways of the routing Compartments (CTs) - simulated by prefixes 1.0.0.0/8, 2.0.0.0/8, 6.0.0.0/8 and 7.0.0.0/8- run a full mesh of GP-BGP sessions. Dashed lines show GP-BGP sessions, while continuous lines show IP-BGP sessions. As mentioned before and in order to obtain early results, routing compartments were simulated as IPv4 prefixes.

The following listing shows the routing status for Router1 after the BGP sessions both IP and GP exchange their routes:

```
192.168.1.2 -> direct link
192.168.1.3 -> direct link
192.168.1.6 -> 192.168.1.3 (1003 1004 1006)
192.168.1.7 -> 192.168.1.3 (1003 1004 1007)
10.0.1.2 -> virtual connection (192.168.1.2)
10.0.1.6 -> virtual connection (192.168.1.6)
10.0.1.7 -> virtual connection (192.168.1.7)
```

where the virtual connection refers to the IP tunnel to be used for reaching the other CT routes. The IP tunnels generated are shown in Figure 7(a).

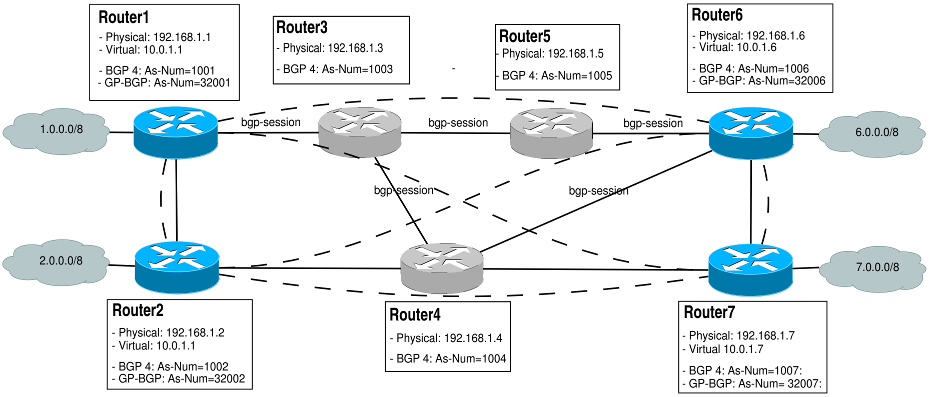
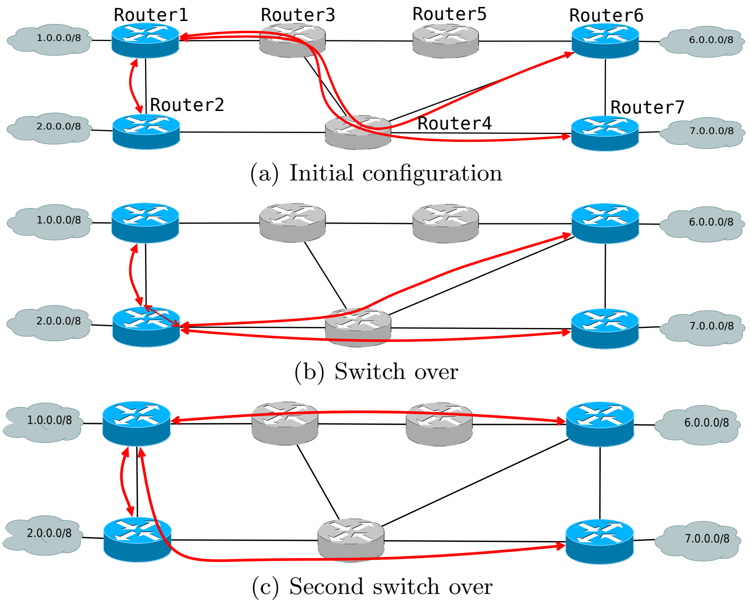


Fig. 6. Connection of AS Compartments



(a) Initial configuration

(b) Switch over

(c) Second switch over

Fig. 7. Evolution of tunnels in Router1

At instant t=1000s, the link between Router3 and Router4 fails so, when Router3 stops receiving the IP-BGP keep-alive messages, it resets its BGP Finite State Machine with Router4:

```
1084.8517681890637 bgp-id=3232235779, peer=(1004;192.168.1.4/32), ESTABLISHED=(HoldTimerExp)=>IDLE
```

Keep-alive message interval is set to T=30s while the time to tear down a session when not receiving keep-alives is 3*T for both GP-BGP and IP-BGP- Router3 withdraws the routes previously exchanged with Router4 announcing that withdraws back to Router1, which modifies its routing table accordingly:

```
1089.8523388557303 192.168.1.7 -> 192.168.1.2 (1002 1004 1007)
1089.8529095223969 192.168.1.6 -> 192.168.1.2 (1002 1004 1006)
1119.8523281890637 192.168.1.6 -> 192.168.1.3 (1003 1005 1006)
```

The route towards 192.168.1.6 and 192.168.1.7 traverses now Router2 instead of Router3. Note that the route towards 192.168.1.6 bounces back to Router3 once Router3 inform Router1 about the new path across AS 1005 instead of AS 1004.

On the meantime, even before the IP-BGP routing reorders its routing, the GP-BGP in Router1 reroutes its traffic through Router2 when it does not receive the appropriate keep-alive messages from Router6 and Router7:

```
1075.3373485086665 bgp-id=167772417, peer=(32006;10.0.1.6/32),ESTABLISHED=(HoldTimerExp)=>IDLE
1089.0377309912103 bgp-id=167772417, peer=(32007;10.0.1.7/32),ESTABLISHED=(HoldTimerExp)=>IDLE

1075.3373485086665 6.0.0.0/8 -> 10.0.1.2 (32002 32006)
1089.0377309912103 7.0.0.0/8 -> 10.0.1.2 (32002 32007)
```

With this routing change on the GP-BGP level, traffic towards 6.0.0.0/8 and 7.0.0.0/8 from 1.0.0.0/8 will be sent towards Router2 using the previously setup tunnel for encapsulating traffic towards 2.0.0.0/8 from Router1. Router2 then decapsulates the traffic and encapsulates it again using as well the previously setup tunnels for encapsulating traffic towards 6.0.0.0/8 and 7.0.0.0/8 from Router2. This behaviour is shown in Figure 7(b).

It is important to note that traffic is forwarded to 6.0.0.0/8 and 7.0.0.0/8 via Router2 using two consecutive tunnels before IP-BGP routing reacts from the link failure and updates its routes:

```
time routing recovers towards 6.0.0.0/8 is  setup again via GP-BGP = 1075.3373485086665 s
time routing recovers towards 7.0.0.0/8 is setup again via GP-BGP = 1089.0377309912103 s
time routing recovers towards 6.0.0.0/8 and 7.0.0.0/8 via IP-BGP = 1089.8523388557303 s
```

Once Router1 re-establishes the GP-BGP session with Router6 and Router7,

```
1099.0435123245438 bgp-id=167772417,peer=(32007;10.0.1.7/32),OPENCONFIRM=(RecvKeepAlive)=>ESTABLISHED
1106.342276508667 bgp-id=167772417,peer=(32006;10.0.1.6/32),OPENCONFIRM=(RecvKeepAlive)=>ESTABLISHED
```

and the routes towards 6.0.0.0/8 and 7.0.0.0/8 are announced again to Router1, Router1 selects these paths:

```
1104.0466803245442 7.0.0.0/8 -> 10.0.1.7 (32007)
1141.3457858420006 6.0.0.0/8 -> 10.0.1.6 (32006)
```

falling back to the previous situation where direct and unique tunnels where used to forward traffic between the different CT routes. This behaviour is shown in Figure 7(c).

It is important to note that even the tunnels are the same as in the initial situation, they traverse different ASs as specified by the underlying IP-BGP routing protocol.

Finally, as the link from Router3 and Router4 recovers at t=2000s the IP-BGP converges back to the initial situation of Figure 7(a). This recovery is totally transparent to GP-BGP protocol as no sessions states are changed and no routing announcements are done: the only changes are on the ASs the tunnels traverse but not on any of the GP-BGP signalling.

4.2 Simulation Results

In order to check the scenario shown in Figure 6, 1000 simulations were performed to see how the unpredictable conditions such as time delays, packet losses, etc. affect these results. Figure 8 compares the density functions of IP traffic recovery time in the IP-BGP case versus the GP-BGP case.

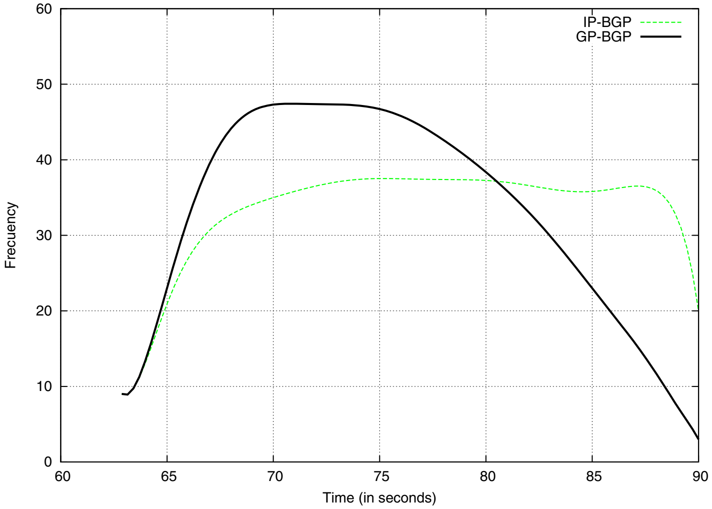


Fig. 8. Density function of traffic recovery time

It is important to note that IP traffic recovery time in the GP-BGP case is always slightly better than IP-BGP case. This is the case because the IP traffic for the GP-BGP routes can be recovered in two ways; the first one when GP-BGP decides to re-route the traffic through the new path, and the second one when BGP-4 detects the failure and chooses a new alternative route to forward the traffic. IP traffic for the GP-BGP case take advantage from these two detection and recovery algorithms: its recovery time is the minimum between the two. Also, this gain can be improved by decreasing the GP-BGP time between keep-alive messages, so failures can be detected earlier; however, there is a tradeoff with the extra signalling overhead it introduces, therefore extra care is needed when setting this parameter.

5 Conclusion

The 4WARD project has investigated different approaches to provide enhanced services which are not possible nowadays over the Internet. One of the cornerstones for such novel networking proposals is the Generic Path concept. But, in order for any Future Internet technology to be deployed, a migration path which takes into account pre-existing network technologies is needed. Experience shows that adoption of a new technology starts with small isolated sections which need interconnection in order to flourish. In this paper we have presented the AS Compartment concept and supporting BGP extensions which will be helpful for such migration scenarios. We show through simulations, that the AS Compartment concept is not only applicable in Future Internet migration scenarios, but might also be used in the near future to enhance the resilience of the current Internet.

6 Related Work

One of the main challenges in the area of BGP-4 routing is scalability in terms of a size of Routing Information Base (RIB) and Forwarding Information Base (FIB) entries. There are many reasons why the sizes of these tables have increased, but maybe most significant reasons are address prefix de-aggregation and use of routing policies based on various reasons. There are many BGP-4 improvement proposals to improve the scalability. [18] proposes the method according to which a set of topologically co-located Internet Service Providers (ISPs) could agree to share a network prefix(es) and aggregate the common prefix(es). The authors of [19] discovered that many routable prefixes share same AS path and in order to optimize the space usage prefixes are divided into atoms that are then routed and advertised instead of prefixes. This would reduce both FIB and RIB sizes in the Default-Free Zone (DFZ) and therefore also potentially improve convergence. [20] analyzes the current BGP-4 routing including both interior and exterior BGP-4. Based on the analysis, the authors proposed a new enhanced BGP-4 protocol called the *atomic BGP* that could be also deployed incrementally. This protocol makes an AS to use non-contradicting routing policies, i.e., all routers inside the AS make route selection and dissemination in the same way. As a result of this, an AS can be seen as a single node to the outside. The *atomic BGP* can lead to a simpler network management which could mean less routing errors and misconfigurations resulting "false" BGP-4 updates and convergence.

Current practices to implement traffic engineering in BGP-4 routing [21] have to be re-examined. They are also relevant while designing how multi-path routing is setup and determine the kind of benefits which can be derived from their use. For instance, if the main motivation to use multi-pathing is to improve resilience, then it becomes essential to try minimizing a number of common BGP-4 links to be used by multi-path flows linked to a single end-to-end session to maximize resilience in case of BGP-4 routing failure.

Another issue of the current BGP-4 routing system is its relatively slow convergence from routing errors/updates. BGP-4 convergence can be divided into two phases, 1) failure detection and 2) path exploration. Once the routing failure has been detected in an AS, there is typically a predefined delay until BGP-4 updates are sent to notify other ASs. The default value of the Minimum Router Announcement Interval (MRAI) are 5 seconds for interior and 30 seconds for exterior routing. BGP-4 system behaviour has been widely analysed based on BGP-4 traffic samples collected by Oregon Routeviews [22] and the RIPE RIS [23] projects. Thus, the analysis in [24] clearly shows that approximately 36% of monitored update sequences took longer than 60 seconds to complete. The authors of [25] have studied how Voice over IP (VoIP) calls and their quality degradation correlate with the BGP-4 updates in the core network. The study shows that BGP-4 has a similar negative impact on the VoIP quality as network congestion.

The impact, though, is not extremely severe for some application types and use scenarios. For instance, Delay Tolerant Network (DTN) based applications as well as elastic traffic could tolerate connection breaks quite well. On the other hand, in any real-time application the situation would suffer a greater impact, since typically relatively long connection breaks are not transparent to the end-users.

7 EU Disclaimer

This paper describes work undertaken in the 4WARD project, which is part of the EU IST programme. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the 4WARD project. All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors view.

References

1. GSM Association. Official document: Ir.34 - inter-service provider ip backbone guidelines, v. 4.4 (June 2008), http://www.gsmworld.com/documents/ir3444.pdf
2. Rekhter, Y., Li, T., Hares, S.: A Border Gateway Protocol 4 (BGP-4). RFC 4271 (January 2006)
3. McPherson, D., Gill, V.: BGP MED Considerations. RFC4451 (March 2006)
4. Scholl, T.: Best Practices for Network Interconnections. Presentation on NANOG43 (June 2008)
5. The 4WARD Project, http://www.4ward-project.eu/ (last visit April 27, 2010)
6. The 4WARD Project. D-5.1 Architecture of a Generic Path (2009) (last visit April 27, 2010)
7. 6bone, http://en.wikipedia.org/wiki/6bone (last visit April 27, 2010)
8. Carpenter, B., Moore, K.: RFC3056 - Connection of IPv6 Domains via IPv4 Clouds (2001), http://www.ietf.org/rfc/rfc3056.txt (last visit April 27, 2010)
9. Making the Transition From IPv4 to IPv6 (Reference), http://docsun.cites.uiuc.edu/sun_docs/C/solaris_9/SUNWaadm/IPV6ADMIN/p21.html (last visit April 27, 2010)
10. Fujinoki, H.: Multi-path BGP (MBGP): A solution for improving network bandwidth utilization and defense against link failures in inter-domain routing. In: 16th IEEE International Conference on Networks, ICON (2008); ISSN: 1556-6463, Print ISBN: 978-1-4244-3805-1

11. Xu, W., Rexford, J.: MIRO: Multi-path Interdomain ROuting. ACM SIGCOMM Computer Communication Review archive 36, 171–182 (2006); ISSN:0146-4833
12. Traina, P., McPherson, D., Scudder, J.: Autonomous System Confederations for BGP. RFC5065 (August. 2007)
13. William, B.: Norton. Internet Service Providers and Peering (2000)
14. Caesar, M., Rexford, J.: BGP routing policies in ISP networks. IEEE Network 19, 5–11 (2005)
15. Perkins, C.: IP Encapsulation within IP. RFC 2003 (October 1996)
16. Bates, T., Chandra, R., Katz, D., Rekhter, Y.: Multiprotocol Extensions for BGP-4. RFC4760 (January 2007)
17. `http://sites.google.com/site/jsimofficial`
18. Internet Service Provider Address Coalitions (ISPACs). IETF Internet draft
19. Verkaik, P., Broido, A., Hyun, Y., Claffy, K.C.: Atomised Routing. Presentation in RIPE45 meeting, `http://www.nlnet.nl/project/atombr/20030512-atoms-ripe45.pdf`
20. Zhang-Shen, R.: Atomic Routing Theory: Making an AS Route Like a Single Node. Invited talk in NSF FIND Routing Workshop (2008)
21. Halabi, B.: Internet routing architectures, 2nd edn. (2000)
22. `http://www.routeviews.org/`
23. `http://www.ripe.net/projects/ris/rawdata.html`
24. `http://www.potaroo.net/ispcol/2007-06/dampbgp.html`
25. Kushman, N., Kandula, S., Katabi, D.: Can You Hear Me Now?! It Must Be BGP. ACM SIGCOMM 37, 75–84 (2007)