社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＱＡＩ－22（９）
2007／1／19

高信頼ドメイン間経路制御に向けたBGP接続手法の提案

渡里 雅史† 屏 雄一郎† 長谷川輝之† 阿野 茂浩† 山崎 克之††

† 株式会社KDDI研究所 〒356-8502 埼玉県ふじみ野市大原2-1-15
†† 長岡技術科学大学 〒940-2188 新潟県長岡市上富岡町1603-1
E-mail: †watari@kddilabs.jp

あらまし 近年、インターネットの継続的な成長に伴い、ネットワークのロバスト性や信頼性向上を目的とした研究が多く見受けられるようになった。インターネット経路制御アーキテクチャは、端末およびネットワークそのものが動かないことを前提に設計されたことから、端末をはじめネットワークを単位とするモビリティ技術の登場は、スケーラビリティや通信経路の非効率性に関する問題をより深刻化させると同時に、インターネットアーキテクチャの見直しを加速させた。本稿では、これまでに提案されたモビリティ関連のルーティング技術に関する問題点を分析した後、高信頼ドメイン間経路制御の実現に向け、大規模ネットワークの障害回復を実現するBGPピアリング手法について提案する。また、提案方式のBGP拡張に関わる詳細設計ならびに既存ネットワークへの適用性について述べる。
キーワード Border Gateway Protocol (BGP)、ドメイン間経路制御、ロバスト性、耐故障性

A Proposal for Flexible BGP Peering Method towards Dependable Inter-Domain Networking

Masafumi WATARI†, Yuichiro HEI†, Teruyuki HASEGAWA†, Shigehiro ANO†, and Katsuyuki YAMAZAKI††

† KDDI R&D Laboratories Inc. Ohara 2-1-15, Fujimisho-shi, Saitama, 356-8502 Japan
†† Nagaoka University of Technology Kamitomioka 1603-1, Nagaoka-shi, Niigata, 940-2188 Japan
E-mail: †watari@kddilabs.jp

Abstract As the Internet continues to grow, there has been much research activity on providing a more dependable and robust networking. As the Internet routing architecture was designed based on the assumption that terminals and networks are fixed, the notion of network mobility has also brought needs to reexamine the current Internet architecture with concerns on scalability and optimality associated with mobility. This paper describes the current routing protocols related to mobility and discusses, from the IP mobility essence, on ways for providing reliability and dependability to the inter-domain routing. We also present a novel peering method for BGP speakers that would allow ASes to automatically and dynamically restore from failures. The detail extensions made to BGP and its applicability to the current BGP networks are additionally presented.
Key words Border Gateway Protocol (BGP), inter-domain routing, robustness, fault tolerance

1. Introduction

Recently, high quality and high performance have become great concerns for users running disruption sensitive applications over the Internet. On the other hand, continuous and rapid replacement of conventional Public Switched Telephone Network (PSTN)-based voice service to Voice over IP (VoIP) have made the Internet become part of a vital lifeline, where reliability and dependability are considered more important. Unfortunately, however, IP networks today are frequently disconnected and isolated due to failures from operational errors, hardware and software problems, and those of natural disasters. Such an event leading to instability of routes is a critical issue for a vital lifeline, as it not only degrades the IP performance, but also, in the worse case, causes unavailability of the IP service.

As the Internet continues to evolve, providing fault tolerance has become an important key step towards a reliable and dependable inter-networking. Unfortunately, the Internet lacks fundamentally the ability for each individual IP network to autonomously and dynamically reconfigure, reform, and recover networks upon failure. The current Internet is neither incapable nor suited for providing such feature, as the architecture was designed with an assumption that networks do not change topology once configured. The design also included a policy of keeping the core routing system simple and giving the intelligence to end hosts. However, the innovation of network technologies have created the notion of network mobility, where intelligence is given to routers and movement transparency is provided to subnets. Though the architecture has limited applicability, an essence similar to IP mobility shall give a useful input and thoughts in providing mobility to core routing systems.

In this paper, we focus on one of the key components of the Internet, the Border Gateway Protocol (BGP) [1], and discuss, from the IP mobility essence, on ways for providing reliability and dependability to the inter-domain routing. As an application of the discussion, we present a novel peering method for BGP speakers that would allow Autonomous Systems (AS) to automatically and dynamically restore from failures. The remaining sections are organized as follows. Section 2 describes the current work in IP mobility and IP multihoming from the fault tolerance perspective. Section 3 discusses mobility support that covers larger networks for disaster tolerant networking. Section 4 presents the approach taken towards restoration of ASes with detail extensions made to BGP. Finally, we conclude in Section 5.

2. Related Work

2.1 IP Mobility Support Overview

Mobility protocols in general provides ways for nodes to be reachable through an unique identifier regardless of its attachment point to the Internet, and additionally provides ways to preserve established connections during handoffs. For IP mobility, the feature is provided at the IP layer, allowing nodes to remain reachable through an unchanging IP address.

Originally, IP mobility was introduced to relax the use of the IP address as both a network locator and as an identifier of today's Internet architecture, and provide a so-called ubiquitous access allowing users to access and be accessed from the Internet. However, in terms of providing service continuity during handoffs, the technology is also capable of offering robustness and fault tolerance in the event of failures. IP mobility support for subnets can be said more effective than those for hosts, as much more hosts and services can benefit from fault tolerance. Though IP mobility support is currently limited to hosts and subnets, extending the idea for further providing support of larger networks, such as an AS, may realize dependability and reliability to inter-domain networking.

表 1 Different Levels of Mobility Support

Feature	Reference
Host Mobility	Mobility Support for IPv4 (RFC 3344) [2]
	Mobility Support in IPv6 (RFC 3775) [3]
	HIP: Host Identity Protocol (RFC 4423) [4]
	Network-baed Local Mobility Management [5]
	IPv6 Site Multihoming (Shim6) [8]
Subnet Mobility	Network Mobility Support (RFC 3963) [6]
	Connexion by Boeing [7]
AS Mobility ?	Support for larger networks?

In the following section, we describe some of the recent work presented or standardized at the Internet Engineering Task Force (IETF), for understanding the architectural commonalities and differences of IP mobility support and to further discuss applicability for larger networks. Table 1 shows the different level of mobility support. Note that the Shim protocol is not precisely a mobility protocol, but provides a similar feature through multihoming.

2.2 IP Mobility Support Technology

There has been much research in the area of IP mobility support for moving hosts and standardization efforts at the IETF, namely Mobile IP [2,3], HIP [4], and Network-based Local Mobility Management [5]. Additionally, the IETF has also studied and recently standardized an IP mobility protocol for moving subnets, called Network Mobility (NEMO) [6]. The protocol is much like as that of Mobile IPv6, only that the router serving as a gateway between the moving network and the Internet provides movement transparency to the nodes located behind.

For many of these protocols, the network architecture involves an agent located at the infrastructure for forwarding packets. Depending on the protocol, such agent is called a home agent, a mapping agent, an anchor router, or a rendezvous server. In any case, forwarding of packets is achieved in an overlay-like fashion over the IP routing infrastructure using, for example, IP-in-IP tunneling. Such an approach is basically unavoidable with the current Internet architecture, as the IP address is used as both an identifier and a locator.

On the other hand, the Connexion by Boeing [7] demonstrated an interesting approach through its commercial flights for providing global network mobility. The network architecture greatly differs from existing approaches in that the Border Gateway Protocol (BGP) is used for mobility management. The architecture introduces BGP route-

servers as ground stations for announcements and withdrawals of routes associated with the aircraft's mobility. For example, when an aircraft is near ground station "A," routes are announced via ground station "A" with an origin AS number of that of ground station "A." Once the aircraft is near ground station "B," routes are withdrawn from ground station "A" and re-announced via ground station "B" with a new origin AS number of that of ground station "B." Though routes converge within about a minute unfortunately, however, various communities have raised concerns on the impact of the excessive announcements and withdrawals give to the global routing system. As the Internet continues to show a trend in growth on the global routing table, such an instability event, or a flap, is considered critical as they delay route convergence, consume network resources, and bring concerns on scalability.

A related approach, not precisely mobility, but host multihoming, is the Shim protocol [8], proposed and standardized at the SHIM6 WG. The Shim protocol separates the IP address, namely the locator, from the identifier used and presented to the upper layer protocols and applications. The mapping of the locator and the identifier is achieved through introduction of a new sub-layer between the IP layer and the transport layer, called the shim layer. Though the protocol is limited only for hosts and does not provide mobility, the use of multiple locators for a given identifier simultaneously offers high availability and reliability. The protocol also provides mechanism for failure detection and recovery for a faster restoration. From the architectural point of view, the Shim protocol does not rely on an agent at the infrastructure, however, as a tradeoff, peers must also be Shim aware for hosts to benefit from the Shim feature.

3. Proposal of AS Mobility as Fault Tolerant Networking

If one could design the Internet architecture from scratch, one may consider extending IP mobility support to provide fault tolerance for larger networks, such as an AS. As the goal of IP mobility support is to provide a permanent IP address regardless of the node's attachment point to the Internet, AS mobility allows ASes the ability to peer with other ASes on-demand and still maintain reachability. If the architecture could also relax the dependency with other nodes, all routers would eventually become mobile, providing robustness to the network as a whole.

ASes can benefit today from AS mobility as a way to restore its network during disasters. This is rather important in terms of prompt recovery of vital lifelines. The earthquake that hit the Niigata Prefecture on October of 2004 isolated a village for hours as massive amount of incoming

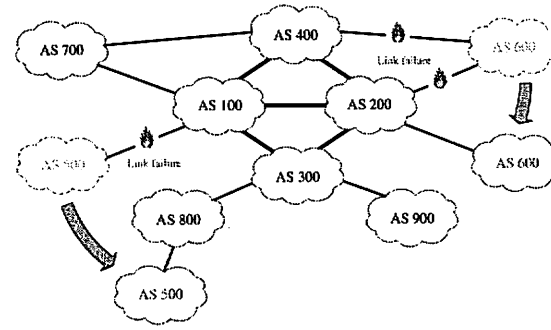

図 1 Overview of AS Restoration

calls to this area heavily congested the network [9]. Operators were unable to reach the sites for recovery, as the roads were either heavily damaged or blocked by landslides. Though the network luckily survived from a complete failure, however, this tragedy demonstrated the need for networks to autonomously and dynamically reconfigure, reform, and recover networks upon failure.

In terms of providing fault tolerance, multihoming is an alternative to AS mobility. For example, many AS composing the Internet today peer with multiple different ASes simultaneously to form a multihomed topology. In the event of a failure, the alternative paths and routes are used as backups. If the availability of these routes can be ensured, permanent Internet connectivity are provided to IP networks. Unfortunately, however, the availability of such pre-configured routes can not be ensured as failure may also affect backup routes.

4. Autonomous System Restoration: System Design

As an attempt to provide dependable inter-domain networking, this paper focuses on one of the key components of the Internet, the Border Gateway Protocol (BGP), and presents a method that allows a network to automatically and dynamically recover from failures, called AS restoration. Dedicated BGP speakers within an AS are provided the ability to autonomously and dynamically reconfigure, reform, and recover networks upon failure.

This section discusses the approaches taken towards AS restoration and additionally describes the overview of the protocol with detail extensions made to the peering procedure of BGP.

4.1 Approaches and Discussions

AS restoration allows for an AS to restore connectivity through another AS upon failure. AS restoration is different from multihoming in that the route used for restoration is not pre-configured, but discovered on-demand. The overview is presented in Figure 1. For example, in the event of a disas-

ter, BGP speakers of isolated ASes may rely on the wireless links such as nearby hot spots or satellite links to discover other peers for restoration. One reason for not multihoming is to save operational cost. Unless otherwise noted, a failure described in this paper represents the unavailability of the layer 3 routes.

In restoring ASes, one can inject routes through the discovered peer and propagate its routes via BGP updates. One can also establishment a tunnel with another node where routes are aggregated and announced. In terms of minimal impact on the global Internet routing system, the tunneling approach is more beneficial, however, the availability of the route to the node is not ensured after failures.

As disaster tolerant networking is particularly needed in today's networks, ease of deployment is also an important factor in designing solutions. For example, solutions that require modifications and replacements of all routers are less deployable as to solutions that require only the dedicated routers to implement a specific capability. As a side note, such a requirement is also listed in [10], a document listing the requirements for a new inter-domain routing architecture.

In summary, we have considered the following items as requirements for a disaster tolerant networking.

- The restoration of ASes should be achieved in the IP routing layer and should not rely on upper layer protocols such as tunneling. For concerns on the affect on the global routing table, the solution is dedicated for lifeline recovery after disaster, thus the restoration should not occur as frequently.
- The restoration process should not rely on the existence of pre-configured routes such as multihoming.
- The solution should be easily deployable to meet today's demands.

4.2 Protocol Overview

BGP requires that all BGP speakers within an AS must be fully meshed, which is known to cause a serious scaling problem for large ISPs. AS Confederation for BGP [11] has thus been defined to relax the full mesh requirement, dividing each AS into multiple sets of smaller ASes, called Member-ASes. The AS restoration discussed in this paper targets restoration for these Member-ASes.

A dedicated BGP speaker of each Member-AS periodically checks for its connectivity with other neighbor members of the confederation. If the BGP speaker finds its AS being isolated from its confederation (i.e. connectivity to neighbor ASes is lost), the BGP speaker then attempts to restore connectivity through discovery of another peer within a reachable distance, either that of the same confederation or a different confederation. If the discovered peer belongs

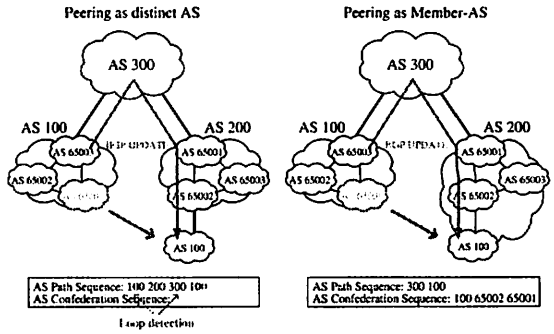

図 2 Migration to peer confederation

to the same confederation, a session is then established dynamically between the two BGP speakers.

If the peer belongs to a different confederation, the BGP speaker translates its AS number along with its AS Confederation Identifier to migrate as a new Member-AS of the peer confederation. Such a migration to the peer confederation is necessary for successfully receiving update messages from the previous AS, as shown in Figure 2. For example, if the member of AS 100 attempts to restore connectivity as a distinct AS of AS 200, the AS Path sequence of the UPDATE message would already include its own AS number, causing loop detection. Thus, the migrating Member-AS must peer as a member of a peer confederation. Once the connectivity through the original confederation recovers, the BGP speaker silently disconnects the established sessions and migrates back to the original confederation. The details are described in the following sections.

4.3 Discovery of Neighbor BGP Speakers

One of the key issues in disaster recovery is the ability for networks to recover from failures promptly. Currently, failure recovery in BGP is realized either by configuring multiple session between two BGP speakers or through multihoming. If existing sessions fail, BGP does not provide any mechanism for peering with other undiscovered BGP speakers.

In the event of a failure, the BGP speaker attempts to discover other neighbor BGP speakers. In performing the discovery, we extend the Neighbor Discovery Protocol (NDP) for IPv6 [12] together with IPv6 Stateless Address Autoconfiguration [13] for generating a link-local address, ensured unique on the link, and additionally for learning the local topology. Figure 3 shows the overview of the discovery procedure. An extension is made to the Neighbor Advertisement (NA) message of the NDP to carry the parameters necessary for establishing BGP sessions, which are, Member-AS Number, AS Confederation Identifier, and IPv4 address of the interface. Figure 4 shows the newly defined option format for use with the NA message.

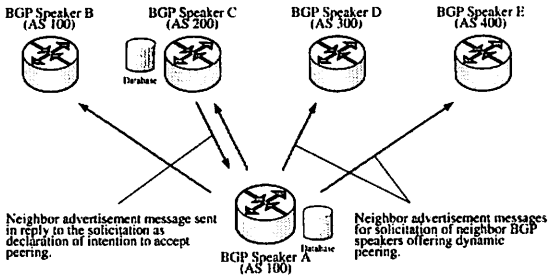

図 3 Neighbor BGP Speaker Discovery

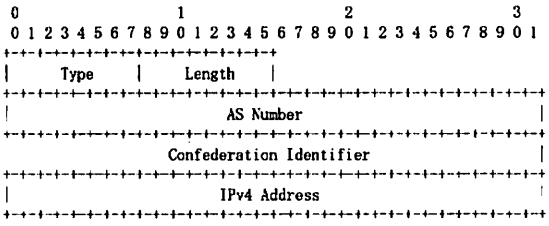

図 4 BGP Option Format for Neighbor Advertisement

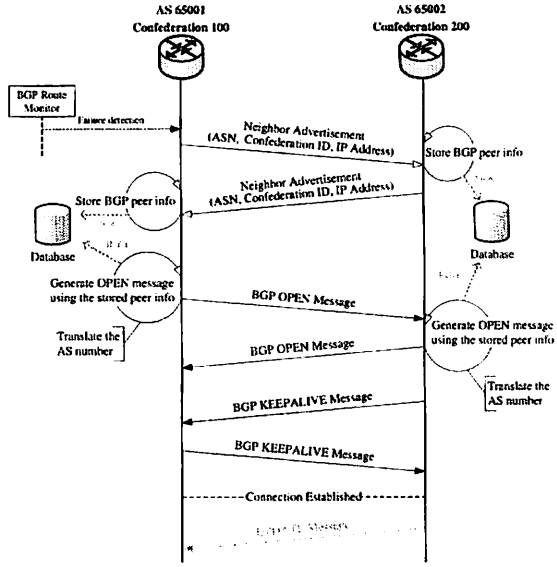

図 5 Sequence of messages for AS Restoration

The message is sent with a newly defined Acknowledgment flag as an indication to request the peer to also reply back with a NA message. Each BGP speaker should store the parameters of the received messages in its database for use later with the BGP OPEN message and additionally to manage multiple peers simultaneously. If multiple BGP speakers are discovered, the BGP speaker may peer with all discovered BGP speakers or may choose a BGP speaker based on local policies.

4.4 Dynamic Peering with Discovered Neighbors

The peering procedure of BGP consists of exchanging a set of OPEN messages and a KEEPALIVE messages over a TCP connection between two BGP speakers. The OPEN message contains a field for carrying the AS number of the BGP speaker and optional parameters, called BGP Capabilities [14], listing the capabilities supported by the speaker.

As discussed earlier, each Member-AS must migrate to the peer confederation for restoration to avoid routing loops. For a Member-AS to migrate to a different AS Confederation, the dedicated BGP speaker must make sure that the Member-AS number does not cause conflicts with the remote Member-AS numbers. As the Member-AS numbers are meant to be visible only within a confederation, other confederations may also use the same Member-AS numbers. As a result, exposing the numbers may result in duplication between confederations. Therefore, the two BGP speakers must use their AS Confederation Identifier for the AS number field of the OPEN message. Additionally, a new capability, called Dynamic Peering Capability, is defined and used in the OPEN messages to request the peer to also send OPEN messages in order to establish connection. The capability code of this message is to be defined with the option length of 0.

Figure 5 shows the sequence of messages for peering. When the BGP speaker of AS 65001 detects itself being isolated from neighboring ASes, the BGP speaker sends a BGP OPEN message with the Dynamic Peering Capability. The message is sent to the candidate listed in the database constructed via neighbor discovery. Once the BGP speaker of AS 65002 receives this message, it verifies the parameters and replies back with an OPEN message. Once both BGP speakers receives a KEEPALIVE message from the peer, the session is successfully established.

4.5 Exchanging BGP Updates and Path Attribute Modification Rules

The primary function of BGP is to exchange network reachability information through UPDATE messages over the established session. Each UPDATE message includes the sequence of which the message traversed, used to avoid routing loops. The sequence is identified as AS_SEQUENCE of the AS_PATH attribute. Messages exchanged within a confederation are identified as AS_CONFEDERATION_SEQUENCE, which is an attribute visible only within the confederation.

For the Member-AS to become part of the remote confederation, the AS_PATH modification rules described in [11] needs to be replaced to avoid routing loops. An example of the path attribute modification is illustrated in Figure 6. In the figure, BGP speakers A and E, and BGP speakers B and C belong to the same confederation, respectively. BGP

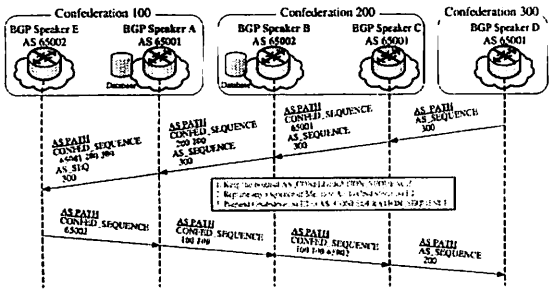

図 6 Sequence of Path Attribute Modification

speakers A and B have the capability for AS restoration, while others are unmodified BGP speakers.

When the BGP speaker A attempts to restore its AS through BGP speaker B, the BGP speaker B replaces, for all messages from C, the entry in the AS_CONFEDERATION_SEQUENCE to its AS Confederation Identifier of 200 and additionally prepends another 200 for its own AS. BGP Speaker A receiving this message prepends its Member-AS number of 65001 to the same sequence. For the UPDATE message originating from BGP speaker E, the BGP speaker A performs the same operation of that of BGP speaker B. As the AS numbers presented in the path attributes are mainly used for loop detection and for counting the number of AS traversed, replacement of the AS numbers allows restoration of AS.

4.6 Failure and Recovery Detection

Failure detection and recovery can be achieved in various ways. Though the detail method is left out of scope, one can monitor the availability of the paths to neighboring ASes. Additionally, for BGP speakers to dynamically establish and terminate sessions following the detection, the BGP finite state machine is extended as shown in Figure 7.

5. Open Discussions and Conclusion

This paper focused the Border Gateway Protocol and discussed, from the IP mobility essence, on ways for providing reliability and dependability to the inter-domain routing. As an application of the discussion, we have presented the notion of AS mobility and described a novel peering method for BGP speakers that allows AS restoration, effective for disaster tolerant networks. The detail extensions made to BGP and NDP were additionally presented.

As a way to verify our proposal, we plan to implement the proposal on Linux system and examine inter-operability with vendor routers. In terms of security threats on the peering process, we believe that the local link of these BGP routers is administratively secured and that it can be further ensured using Secure Neighbor Discovery (SEND) [15]. These issues

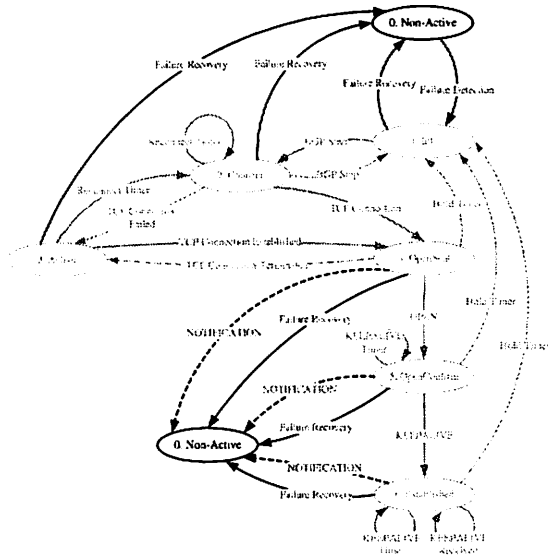

図 7 Changes to BGP Finite State Machine

are, however, open for further study.

文 献

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

[2] C. Perkins, "IP Mobility Support for IPv4," RFC 3344, August 2002.

[3] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," RFC 3775, June 2004.

[4] R. Moskowitz and P. Nikander, "Host Identity Protocol (HIP) Architecture," RFC 4423, May 2006.

[5] J. Kempf, "Problem Statement for Network-based Localized Mobility Management." Internet-Draft (work in progress), draft-ietf-netlmm-nohost-ps-05.txt, September 2006.

[6] V. Devarapalli, R. Wakikawa, A. Petrescu, and P. Thubert, "Network Mobility Basic Support," RFC 3963, January 2005.

[7] A. Dul, "Global IP Network Mobility using Border Gateway Protocol (BGP)," Technical Report, The Boeing Company, March 2006.

[8] P. Savola, "IPv6 Site Multihoming Using a Host-based Shim Layer," IEEE International Conference on Networking, April 2006.

[9] J. Nakazawa and K. Takahashi, "Policy and Planning for Ensuring Information and Communication Networks / Services in Disasters," The Journal of the IEICE, Vol.89, No.9, September 2006.

[10] A. Doria, E. Davies, and F. Kastenholz,"Requirements for Inter-Domain Routing," Internet-Draft (work in progress), draft-irtf-routing-reqs-06.txt, October 2006.

[11] P. Traina, D. McPherson, and J. Scudder, "Autonomous System Confederations for BGP," Internet Draft (work in progress), draft-ietf-idr-rfc3065bis-05.txt, October 2005.

[12] T. Narten, E. Nordmark, and W. Simpson, "Neighbor Discovery for IPv6," RFC 2461, December 1998.

[13] S. Thomson and T. Narten, "IPv6 Stateless Address Autoconfiguration," RFC 2462, December 1998.

[14] R. Chandra and J. Scudder, "Capabilities Advertisement with BGP-4," RFC 3392, November 2002.

[15] J. Arkko, J. Kempf, B. Zill, and P. Nikander, "Secure Neighbor Discovery," RFC 3971, March 2005.