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

2008－IOT－2 (10)
2008／7／25

OSPFを用いたEBGP接続障害の高速復旧手法に関する提案

渡里 雅史 † 屏 雄一郎 † 阿野 茂浩 † 山崎 克之 ††

† 株式会社 KDDI 研究所 †† 長岡技術科学大学

Border Gateway Protocol 4（BGP）を用いた Autonomous System（AS）間の経路制御では，安定した接続性の確保のため，各 AS は隣接する他の AS と予め複数の代替経路を確保する．しかしながら，実際の EBGP 経路障害では，障害検知から復旧までに秒オーダのダウンタイムを伴うと共に，大量発生する経路制御メッセージがルータの処理能力を圧迫し，ネットワーク全体の安定性を著しく低下させる問題がある．本問題を解決するため，筆者らは，ネットワーク全体の安定性を低下することなく EBGP 接続障害から高速に復旧させる手法を提案する．提案手法は，障害に伴う経路変動の伝播範囲を特定のルータに限定させた上で，経路収束の速い Open Shortest Path First（OSPF）を用いて代替経路を確保することを特徴とする．本稿では，提案手法の詳細設計ならびに汎用 PC を用いたプラットホーム上でのプロトタイプ実装とその評価について述べる．

Proposal of Fast EBGP Link Restoration Method using OSPF

Masafumi WATARI † Yuichiro HEI † Shigehiro ANO † Katsuyuki YAMAZAKI ††

† KDDI R&D Laboratories Inc.

†† Nagaoka University of Technology

The Border Gateway Protocol 4 (BGP) allows an Autonomous System (AS) to establish multiple external BGP connections with other ASes to ensure availability of alternative paths. However, recovery through these connections upon a external BGP link failure causes seconds of downtime. The BGP Update messages produced by the failure also heavily stresses the BGP speakers making the whole AS unstable. This paper presents a novel approach of providing fast restoration for external BGP link failures using OSPF. The method maintains stability of the network during restoration by propagating the failure to limited routers. The detail extensions made to BGP/OSPF and its evaluation through prototype implementation are presented.

1 はじめに

近年，IP 電話をはじめ，オンラインゲーム，株取引，企業向け IP-VPN など，安定性およびリアルタイム性を重視する通信サービスが広く普及しつつある．これらの通信サービスは，IP 転送品質劣化の影響を受け易いため，サービス品質の保証には，IP パケット単位の品質（遅延，ジッタ，ロスなど）などの確保に加え，一定の品質を保つため，安定した接続性の確保が重要である．特にこれらの通信サービスの利用形態には，ユーザ同士またはサーバが異なる管理ドメインに接続する場合が想定されるため，ドメイン間での安定した接続性確保が重要となる．

インターネットにおけるドメイン間の経路制御には，Border Gateway Protocol 4（BGP）[1] が広く利用されている．BGP におけるドメインは Autonomous System（AS）と呼ばれ，各 AS は同一または異なる他の AS と予め複数の External BGP（EBGP）接続を確立することで，より信頼性の高い接続性を確保できる．EBGP 接続障害時には，別の EBGP 接続に切り替えることで復旧する．しかしながら，実際の EBGP 接続障害では，障害発生から復旧まで秒オーダのダウンタイムを伴う．また，大量発生する経路制御メッセージがルータの処理能力を圧迫し [2]，ネットワーク全体の安定性を著しく低下させる問題がある [3]．

そこで，これらの課題を解決するため，本稿では，ネットワーク全体の安定性を低下することなくEBGP接続障害から高速に復旧する手法を提案する．提案手法は，障害に伴う経路変動の伝播範囲を特定のルータに限定させた上で，経路収束の速いOpen Shortest Path First（OSPF）[4]を用いて代替経路を確保することを特徴とする．本稿では，提案手法の詳細設計ならびに汎用PCを用いたプラットホーム上でのプロトタイプ実装とその評価について述べる．

本論文の構成は以下の通りである．2章でEBGP接続障害の現状について説明し，3章で関連研究について述べる．4章で提案する高速切替手法について述べ，5章で詳細設計について説明する．6章で提案方式の動作検証ならびに切替時間に関する評価結果を示し，7章で結論を提示する．

2 EBGP接続障害の現状

EBGP接続障害時における復旧プロセスは，大きく二つに分けられる．一つは，障害発生から障害検知である．BGPでは，通常，Keepaliveメッセージを用いた接続状態の監視とホールドタイマを用いた障害検知を行う．障害検知に掛かる時間は，本メッセージの送信間隔とホールドタイマの値に大きく依存する．多くのBGPの実装では，最小3秒のホールドタイマが指定可能である．しかしながら，リアルタイム性を要するアプリケーションでは許容できないパケットロスを伴うため，近年では，BGPとは独立の接続状態監視プロトコルとして，Bidirectional Forwarding Detection（BFD）[5]技術の標準化が進められている．BFDでは，独自のKeepaliveメッセージを用いた接続状態の監視により，数10ミリ秒から数100ミリ秒での高速な障害検知が可能となるため，今後は標準機能として広く実装されていく可能性がある．

二つ目のプロセスは，障害検知から障害復旧である．BGPでは，障害を検知すると最適経路の再計算を行い，代替経路が存在する場合は，Announceメッセージを用いて広報し，存在しない場合は，Withdrawメッセージを用いて障害経路の取り下げを行う．バックボーンルータでは，フルルートと呼ばれるインターネット全体の経路を扱うため，EBGP接続障害時には多くのメッセージの生成と伝搬が必要となる．図1は，Quagga Routing Software Suite（Quagga）[6]のBGP実装を用いて，Internal BGP（IBGP）スピーカ間でフルルートの受信に必要としたメッセージ数とその累積プレフィックス数の関係を示す．計測は，Route Views Project [7]のroute-views2.oregon-ix.netで観測した242,397[1]のプレフィックスとASパスリストをエミュレーションした上で行った．全プレフィックスの取り下げには238個のWithdrawメッセージを要し，同プレフィックスの広報には42,411個のAnnounceメッセージを要した．また，この時のtcpdumpによる全メッセージの伝播時間は，それぞれ0.809秒と5.169秒であった．

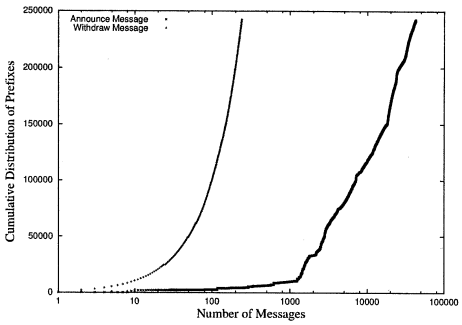

図. 1 フルルート交換時の受信メッセージ数と累積プレフィックス数の関係

これらの値は，BGPの実装やルータの処理性能によって異なるが，経路数の増加に伴い，今後より多くのメッセージ数が必要となることが想定される．特にAS内にRoute Reflector（RR）[8]を設置している場合は，RRへWithdrawメッセージが伝播された後，RRから各クライアントに対してAnnounceメッセージが伝播されるため，メッセージ伝播だけにさらに多くの時間を要する．実際には，RRと接続するクライアント数およびBGPセッション数に応じて処理が増加し，最適経路の再計算と経路表の更新を伴うため，復旧までにより多くの時間を要する．また，これらのメッセージ伝播時は，各ルータの負荷を高めるため，AS内の安定性を一時的に低下させる問題もある．

本稿では，この障害検知から障害復旧までのプロセスに着眼し，上述の問題を解決する新たな復

[1] 2008年3月1日00時00分のRIBデータを使用．
http://archive.routeviews.org/bgpdata/2008.03/RIBS/rib.20080301.0000.bz2

旧手法について提案する．

3　関連研究

EBGP 接続障害時の復旧時間短縮には，交換する Update メッセージ数を減らす方法が考えられる [9, 10]．例えば，フルルートの取り下げであれば，障害したネクストホップの IP アドレスのみを通知し，受信ルータ側で，同ネクストホップに一致するすべての経路を取り下げる方法が考えられる．この場合では，実質 1 つの Withdraw メッセージでフルルートの取り下げが可能となる．しかし，実際のルータでは，メッセージ受信後に BGP Routing Information Base（RIB）の更新，最適経路の再計算，Forwarding Information Base（FIB）の更新など，経路数に応じて多くの処理を伴う．そのため，例えば 1 つの RIB/FIB エントリの更新に 146 μ 秒 [11] を要する場合では，24 万エントリの更新には約 35 秒を要する．このため，障害復旧時間の短縮には，交換するメッセージ数の削減に加え，更新する RIB/FIB エントリ数を減らす工夫が必要となる．

この問題に対して Bonaventure et al. [12] は，FIB 内にネクストホップテーブルを導入し，更新する FIB エントリ数を減らす方法を提案している．本テーブルは，各ネクストホップに対して，予めプライマリとバックアップの二つの出力インタフェースを設定し，プライマリの接続状態に応じてバックアップへの切替えを行う．障害時は，本テーブル内の 1 エントリの更新のみとなるため，障害検知から 50 ミリ秒以内での復旧が可能となる．しかしながら，AS 内に RR が介在する場合では，RR と各クライアントに対して通常の BGP におけるフルルート交換が必要となる．また，プライマリとバックアップへの切り替えは，本テーブル内のフラグに基づいて行われるため，設定可能なバックアップインタフェースが一つに限られる．複数のバックアップに対応するためには，FIB 構造のさらなる拡張，各バックアップの接続性監視，優先順位付けなど，より複雑な仕組みが必要となる．

4　IGP を用いた高速障害復旧手法

EBGP 接続障害は，ネクストホップへの到達性障害であると考えられる．一方，AS におけるネクストホップへの到達性は，Interior Gateway Protocol（IGP）によって確保されるため，ネクストホップへの到達性障害は，IGP における経路障害とも考えられる．このため，IGP により障害箇所を迂回し，ネクストホップへの到達性を再確保することで，EBGP 接続障害の復旧が可能となる．そこで，本稿では，同一 AS へ複数の EBGP 接続がある構成を前提に，EBGP 接続障害からの復旧手法として，IGP を用いてネクストホップへの到達性を再確保する手法を提案する．提案手法は，同じネクストホップへの到達性を復旧させるため，BGP のフルルート交換が必要なく，また，IGP を用いて単一の FIB エントリのみ更新するため，高速かつ安定した障害復旧が可能となる．

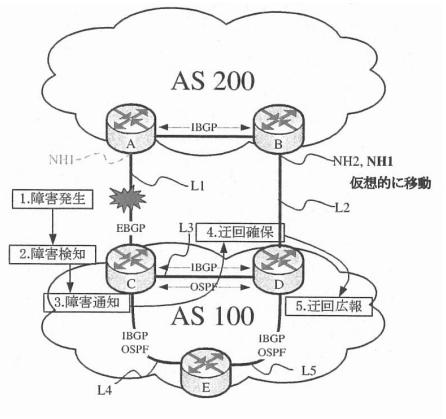

図. 2　提案手法の概要

図 2 に提案手法の概要を示す．AS100 および AS200 における AS 間の接続は，EBGP であり，AS 内は IBGP および IGP として OSPF を想定する．AS100 から AS200 へのネクストホップは，それぞれ NH1 および NH2 とする．ここで，リンク L1 にて障害が発生した場合，通常の BGP では，ルータ C において障害を検知すると共に，AS 内の BGP ルータ D および E に対して Withdraw メッセージを伝播する．一方，提案手法では，ルータ C において障害を検知すると，まず EBGP 接続を有するルータ D に対して，障害を通知する．本通知を受信したルータ D では，同 AS へのリンクである L2 を用いて，仮想的に NH1 への到達性を確保する．すなわち，ルータ C に対して NH1 への到達性を広報すると共に，自身の NH1 へ向けの経路を NH2 に設定する．これにより，ルータ C は障

Router C (Before)

_	Destination	Nexthop
BGP	Prefix X	NH1
OSPF	NH1	Connected (L1)
OSPF	Router D	Connected (L3)
OSPF	NH2	Router D
OSPF	Router E	Connected (L4)

Router D (Before)

_	Destination	Nexthop
BGP	Prefix X	NH2
OSPF	NH1	Router C
OSPF	Router C	Connected (L3)
OSPF	NH2	Connected (L2)
OSPF	Router E	Connected (L5)

Router C (After)

_	Destination	Nexthop
BGP	Prefix X	NH1
OSPF	NH1	***Router D***
OSPF	Router D	Connected (L3)
OSPF	NH2	Router D
OSPF	Router E	Connected (L4)

Router D (After)

_	Destination	Nexthop
BGP	Prefix X	NH2
OSPF	NH1	***Connected (L2)***
OSPF	Router C	Connected (L3)
OSPF	NH2	Connected (L2)
OSPF	Router E	Connected (L5)

図. 3　障害前後における経路表

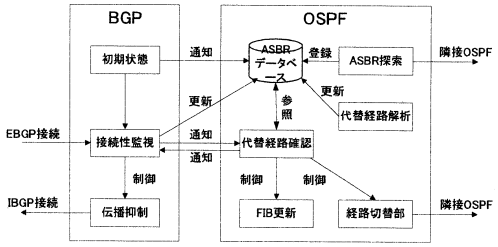

図. 4　機能構成

害したNH1への到達性をルータD経由で確保できると判断し，BGPにおけるWithdrawメッセージの送信を抑制し，OSPFを用いて迂回経路を確保する．

図3に，障害前後におけるルータCおよびDの経路表を示す．ルータCにおける障害前のPrefix XへのネクストホップはNH1であるが，障害後のルータCにおけるネクストホップはルータDとなる．ルータCは，FIBにてNH1をrecursiveに検索することで，宛先ネクストホップを解決する．また，ルータDは，NH1へのネクストホップを自身のリンクであるL2に向けることで，AS200へのパケット転送が可能となる．

5　システム設計

本章では，提案手法に関わる詳細設計について述べる．図4に提案手法の主要機能構成を示す．BGPでは，EBGP接続状態の監視とOSPFへの通知，また，OSPFによる復旧時は，Withdrawメッセージ伝播の抑制を行う．OSPFでは，BGPからの通知に基づき，新たなデータベースの管理と迂回経路の確保，また，隣接OSPFルータとネクストホップ情報の交換を行う．

5.1　代替経路の管理

本稿では，EBGP接続を持つルータをAS Border Router（ASBR）と呼ぶ．ASBRは，EBGP接続障害時の代替経路を確保するため，AS内の各ASBRと予めネクストホップ情報を共有し，ASBRデータベースに管理する．ネクストホップ情報の共有には，ASBRの探索を兼ねるため，OSPFを拡張し，新たに代替経路通知LSAを定義した．各ASBRは，OSPF起動時にネイバに対して本LSAをフラッディングし，EBGP接続のネクストホップアドレスとAS番号を通知する．本LSAを受信したASBRでは，送信元のルータIDと共にこれらの情報をASBRデータベースに格納する．表1に本データベースの例を示す．ASBRデータベースには，使用する代替経路の優先順位を決めるため，プレファレンス値を設けた．

表1　ASBRデータベース

AS番号	ネクストホップ	ルータID	Pref.
200	172.0.1.1	10.0.0.1	20
200	172.0.2.1	10.0.0.2	10
300	172.0.3.1	10.0.0.3	30
300	172.0.4.1	10.0.0.4	10

なお，本LSAは，通常ルータとの互換性を維持しながらAS内の全ASBRに通知するため，Opaque LSA [13] を用いた．通常ルータで特別な処理をすることなく，全ASBRへの通知が可能となる．

5.2　提案手法に関わるBGP拡張

通常のBGPにおけるFinite State Machine（FSM）[1] では，ホールドタイマのタイムアウトに伴い，EstablishedからIdle状態へ遷移し，Withdrawメッセージの伝播を行う．一方，提案手法では，IGPによる障害復旧とするため，本FSMを拡張し，新たにFallback状態を定義した．障害を検知したBGPルータは，Fallback状態に遷移し，OSPFに対してEBGP接続障害の通知を行う．OSPFにて代替経路を確保できない場合は，通常のBGPにおけるIdle状態へと遷移する．OSPFにおいて代替経路が確保できた場合は，Withdrawメッセージ伝播を抑制し，復旧はOSPFに委ねる．また，障害箇所が復旧した場合は，Fallback状態から通常のBGPにおけるOpenSent状態へと遷移し，OSPF

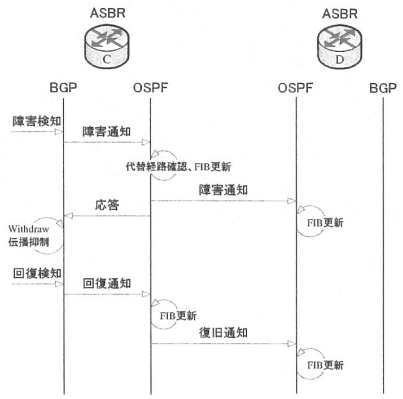

図. 5　メッセージ交換のシーケンス

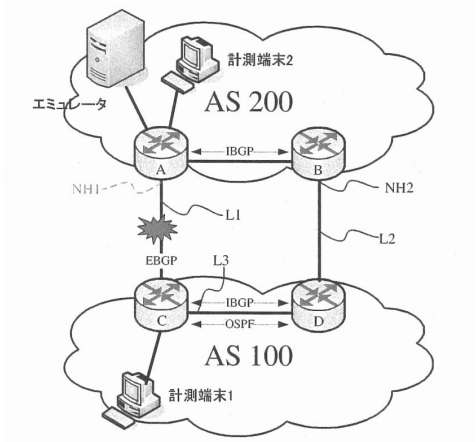

図. 6　実験ネットワーク構成

に対して EBGP 接続復旧の通知を行う．

5.3　提案手法に関わる OSPF 拡張

OSPF では，5.1 節における代替経路の管理に加え，BGP からの通知に基づき，代替経路を確保する．図 5 に障害検知から代替経路への切り替えまで，また，回復検知から経路の切り戻しまでのメッセージ交換シーケンスを示す．

ルータ C の OSPF では，BGP からの障害通知を受けると ASBR データベースを参照し，代替経路の確認を行う．代替経路が存在する場合は，OSPF ネイバ D に対して，EBGP 接続障害を通知するため，迂回経路設定要求 LSA を送信すると共に FIB の更新を行う．本 LSA を受信した OSPF ネイバ D では，迂回経路設定のため FIB を更新する．また，EBGP 接続が回復した場合は，ルータ C の OSPF にて，FIB の更新を行い，OSPF ネイバ D に対して回復を通知する．

6　評価

提案手法の有用性を評価するため，プロトタイプ実装を用いた検証評価を行う．以下，実験環境について説明し，評価結果を示す．

6.1　実験構成

図 6 に実験ネットワーク構成を示す．実験ネットワークは，Intel Quad-Core 2.13GHz の CPU と 2GB のメモリを搭載した汎用 PC 4 台を用いて，Linux Fedra 7 と Quagga [6] の BGP および OSPF を動作させた．このうち，ルータ C および D は，提案手法を実装した Quagga を用いた．また，エミュレータ装置を用いて，2 章と同じ 242,397 のプレフィックスをエミュレーションし，ネットワーク全体へ伝播させた．なお，BGP における Keepalive タイマは 5 秒，ホールドタイマは 15 秒とした．

6.2　実験結果

提案手法の基本動作検証のため，ルータ A および C 間の EBGP 接続においてリンク障害を発生させた．その結果，ルータ C では，BGP ホールドタイマのタイムアウトにより NH1 への接続障害を検知し，ルータ D に対して迂回経路設定要求 LSA を送信すると共に，FIB 内の NH1 への宛先を L1 からルータ D へ更新した．また，ルータ D では，受信した LSA に基づき，FIB 内の NH1 への宛先を自身の L2 へ更新することを確認した．

提案手法の有用性評価のため，エンドエンドにおけるダウンタイムを計測した．計測には ping を 10ms 間隔で計測端末 1 から 2 に対して送信した．プレフィックス間における収束時間の差を調査するため，最初と最後に伝播するプレフィックスを同時に計測した．図 7 と図 8 に通常の BGP と提案手法における計 10 回の計測結果を示す．通常の BGP では，障害発生から最後のプレフィックスの復旧までに平均 2675 個のパケットロス，時間にして約 26.75 秒を要した．また，最初と最後のプレフィックス間の差は平均 11.18 秒であり，単一の RIB/FIB の更新に約 46 μ 秒を必要とした．一方，

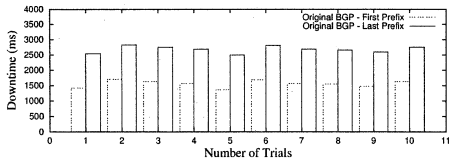

図. 7　通常の BGP におけるダウンタイム

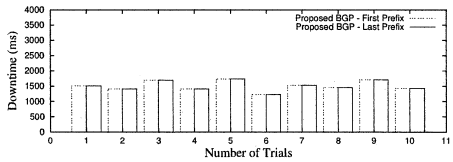

図. 8　提案手法におけるダウンタイム

提案手法では，最後のプレフィックスの復旧時間は平均 15.10 秒であった．また，最初と最後のプレフィックス間での差は 1 パケットであり，提案手法ではプレフィックス数に依存しない障害復旧が可能であることを確認した．なお，回復時はどちらもパケットロスがないことを確認した．

提案手法の切り替え時間を計測するため，ルータ C の障害検知から FIB 更新完了までの時間 T_1，ルータ D の LSA 受信から FIB 更新完了までの時間 T_2 を求めた．また，同様に回復時において，ルータ C の回復検知から FIB 更新完了までの時間 T_1，ルータ D の LSA 受信から FIB 更新完了までの時 T_2 を求めた．表 2 は，それぞれ 10 回の平均を示す．障害時の復旧は，ルータ間での LSA 伝播時間を除き平均 2.3 ミリ秒であった．回復時は，ルータ D の FIB 更新に実装上の問題と考えられる多くの時間を要している．ただし，ルータ C の切り替えのタイミングで復旧するため，上述の通りパケットロスがないことを確認した．

また，FIB 更新に伴うネクストホップの recursive lookup に要した時間を計測した結果，CPU クロック数にして 50 回の平均が 248,680 クロックであり，時間にしておよそ 117 μ 秒と十分に小さいことを確認した．また，2 回目以降は，キャッシュを参照するため，転送するパケット毎に recursive lookup の必要がなく，データトラヒックへの影響が少ないことを確認した．

表 2　経路切り替え時間（ミリ秒）

	T_1	T_2	$T_1 + T_2$
接続障害時	2.1	0.2	2.3
接続回復時	2.1	221.3	223.4

7　今後とまとめ

本稿では，EBGP 接続障害からの復旧時間を短縮するため，IGP を用いて代替経路を確保する新たな手法を提案した．提案手法は，フルルートの交換が必要なく，また，特定のルータ間で制御するため，AS 内の安定性を維持した復旧が可能となる．実インターネットの約 24 万経路をエミュレーションした環境下で，障害検知から 3 ミリ秒以内の切り替えが可能であることを確認した．

今後は，対向 AS 双方への提案手法の導入，また，BFD を導入した場合の障害発生から復旧までの全ダウンタイムを計測し，提案手法の有用性評価を行う．

参考文献

1) Y. Rekhter, T. Li, and S. Hares, "A Border Gateway Protocol 4 (BGP-4)," RFC 4271, January 2006.
2) D. Meyers, L. Zhang, and K. Falls, "Report from the IAB Workshop on Routing and Addressing," RFC 4984, September 2007.
3) L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S. Wu, and L. Zhang, "Observation and analysis of BGP behavior under stress," ACM SIGCOMM Workshop on Internet Measurment, November 2002.
4) J. Moy, "OSPF Version2," RFC 2178, April 1998.
5) D. Katz and D. Ward, "Bidirectional Forwarding Detection," Internet Draft (Work in Progress), March 2008.
6) Quaaga Routing Software Suite http://www.quagga.net/
7) Route Views Project http://www.routeviews.org/
8) T. Bates, E. Chen, and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)," RFC 4456, April 2006.
9) M. Bhatia, J. Halpern, and P. Jakma, "Advertising Multiple NextHop Routes in BGP," Internet Draft (Work in Progress), August 2006.
10) D. Pei, M. Azuma, N. Ngyuyen, J. Chen, D. Massey, and L. Zhang, "BGP-RCN: Improving BGP convergence through Root Cause Notification," Computer Networks and ISDN Systems, June 2005.
11) P. Francois, C. Filsfils, J. Evans, and O Bonaventure,"Achieving sub-second IGP convergence in large IP networks," ACM SIGCOMM Computer Communication Review, July 2005.
12) O. Bonaventure, C. Filsfils, and P. Francois, "Achieving Sub-50 Milliseconds Recovery Upon BGP Peering Link Failures," ACM CoNext'05, October 2005.
13) R. Coltun, "The OSPF Opaque LSA Option," RFC 2178, April 1998.