LL-007

イベントに基づく BGP トラフィックの解析
Analysis of Event-driven BGP Traffic Behavior

福田 健介† 廣津 登志夫† 明石 修† 菅原 俊治‡
Kensuke Fukuda Toshio Hirotsu Osamu Akashi Toshiharu Sugawara

1. はじめに

Border Gateway Protocol (BGP) 4[1] は現在のインターネットにおけるデファクトスタンダードのインタードメインルーティングプロトコルである．インタードメインルーティングの観点から見ると，インターネットは，ルーティングの単位となる Autonomous system (AS) がヘテロに接続したものと考えることができる．現在の AS 数は 6000 以上であり，AS 間を流れる経路数は 10 万のオーダーとなる．経路の信頼性を確保するために，AS は他の AS に対して複数のリンクを持つことも多く，そのトポロジーはスケールフリーと呼ばれるネットワーク構造となっている [2]．このように自律分散かつ複雑なネットワークを正確に効率よく経路制御することがインタードメインルーティングプロトコルで求められている．

BGP はパスベクタアルゴリズムに基づいており，AS はネットワークトポロジーやポリシーが変化した時のみ，隣接する AS に対してアップデートメッセージ (Advertise(ad) および Withdrawal(wd)) を送信する．Advertise は宛先への経路が新しく発見された場合，および，現在利用可能ではなかった経路が利用可能になった場合に送付される．それに対して，Withdrawal はその宛先への経路が消えたことを意味する．AS はアップデートメッセージを受け取ると，その内容にしたがって自 AS のルーティングテーブルを再計算し，必要があれば，他のピアリングしているルータ (隣接 AS の BGP ルータ) にアップデートメッセージを送信する．つまり，ある原因で生じた経路のダイナミクス (イベント) は，アップデートメッセージで表現される．

インタードメインルーティングが現実のインターネットの到達可能性や基本的な性能に強く影響を及ぼすことから，BGP トラフィックのダイナミクス (とりわけ安定性) の解析は，近年重要な研究テーマの一つとなっている [3, 4]．歴史的には，インターネットバックボーンは個々のリンクやルータの障害に対しては，比較的ロバストであると考えられてきた．しかしながら，最近の BGP トラフィックの測定および解析によって，アルゴリズム，実装，運用方法等の理由によって，必ずしもロバストではないことが示されている [3, 4]．

しかしながら，システムが自律的に動作していること，経路数が多いこと等の理由により，大量のデータから実際に何が起きているのかを把握することは難しい．そこで，本稿では，[4] らが用いているイベントの定義を拡張し，100 日に渡る BGP アップデートメッセージデータを用いて，経路ごとのマクロな BGP ダイナミクスを統計的手法で解析する．本論文で示す主たる解析結果は以下のとおりである．

†NTT 未来ねっと研究所
‡NTT コミュニケーション科学基礎研究所

表 1: 観測 BGP メッセージフォーマット

type	prefix	AS path
ad	10.0.0.0/19	(AS_1, AS_2, AS_3, AS_4)
wd	192.168.10.0/24	

1. 各イベントは互いに必ずしも同期しているわけではない．
2. ASpath 長は振動する傾向にあるが，対称ではなく伸びやすい傾向にある．
3. 経路の再計算を伴わない Advertise メッセージが，全イベントの 16%を占める．
4. 断線時間の分布は，非常に不安定な経路ではダンプニングの効果で 20〜30 分程度の特徴的なサイズを持つが，それ以外の経路では裾の長い分布になりやすい．
5. イベントが生じる経路には強い局所性が存在する．すなわち，多くのメッセージは少数の経路にのみ関係している．
6. 経路が初めて広報されてから消滅するまでのイベントパターンを見ることで，経路の特性が抽出可能．

2. データセット

2.1 測定データ

本研究では，米国東海岸における Tier1 ISP にピアリングしたルータで観測した，フルルートの BGP メッセージをデータとして使用した．測定期間は 2003 年 8 月よりおよそ 100 日間である．1 つの BGP アップデートメッセージには複数の ad/wd メッセージが含まれている．オリジナルのメッセージから記録した ad データは <Time, Prefix, ASpath> からなり，それぞれ，メッセージの到着時刻，広報される経路の CIDR アドレス (Prefix)，Prefix までの AS パスを表す．また，wd データは，到着時刻，消去される経路の CIDR アドレスの組 <Time, Prefix> からなる (表 1 参照)．

BGP は通信に TCP を使用し，かつアップデートメッセージに複数のメッセージが含まれることから，あまり細かい時間粒度でメッセージのタイムスタンプをつけることに意味がない．そのため，観測では各メッセージに対してタイムスタンプは 1 秒ごとに付加している．

2.2 イベントの定義

個々の経路に関するイベントは，ad/wd メッセージの組み合わせで表現される．本研究では，[4] で扱われているイベントを拡張し，合計 10 個のイベントを定義する．

- Ad: ある Prefix に関する ad メッセージが来た場合に，ルーティングテーブルに，その Prefix に関する経路がなく，かつ，その Prefix を包含する Prefix[§] が存在しない場合．これは新たな経路の通知に対応する．
- Ex: ある Prefix に関する ad メッセージが来た場合に，ルーティングテーブルに，その Prefix に関する経路は存在しないが，その Prefix を包含する Prefix が存在する場合．より詳細な経路が通知された場合に対応する．
- Wa: ある Prefix に関する wd メッセージが来た場合に，ルーティングテーブルに，Prefix およびその Prefix を包含する Prefix が存在する場合．これは，指定された Prefix に関する経路はなくなるが，代わりに包含する Prefix に対する経路が選択されることに対応する．ただし，包含する Prefix が存在しているからと言って，必ずしもその Prefix への到達性が保証されているわけではないことに注意が必要である．
- Wd: ある Prefix に関する wd メッセージが来た場合に，ルーティングテーブルに，その Prefix を包含する Prefix も存在しない場合．これは明らかにその Prefix に対する到達性が失われたことを意味する．
- Lp: ある Prefix に関する ad メッセージが来た場合に，同じ Prefix の経路が存在するが，AS パス長が増加した場合．経路の切断等により，遠回りな経路が広報されたことを意味する．
- Sp: ある Prefix に関する ad メッセージが来た場合に，同じ Prefix の経路が存在するが，AS パス長が減少した場合．Lp の逆で，経路の復帰等により近い経路が広報されたことを意味する．
- Dp: ある Prefix に関する ad メッセージが来た場合に，同じ Prefix の経路が存在し，AS パス長が同じであるが，中間の AS 番号が異なる場合．
- Lvp: ある Prefix に関する ad メッセージが来た場合に，同じ Prefix の経路が存在し，AS パス長が増加し，かつ，同じ AS 番号がプリペンドされた場合．これは，マルチホームされている 2 つの AS 間でリンクが切り替わった場合等に相当する．
- Svp: ある Prefix に関する ad メッセージが来た場合に，同じ Prefix の経路が存在するが，AS パス長が減少し，かつ，同じ AS 番号が減った場合．これは，Lvp の逆のパターンである．
- Sm: ある Prefix に関する ad メッセージが来た場合に，同じ Prefix の経路が存在し，かつ，AS パスが全く同じ場合．これは経路制御する上では何の影響も及ぼさない．

図 1 はイベント数の各時間ごとの推移を示したものである．グラフのビンサイズは 0.5 日であり，期間は 40 日分である．この観測データを見ると，$t = 70, 92, 95$ 付近で Wd, Ad, Ex イベントが突出して起きていることから，ピアリングしているルータの近傍もしくは主要な幹線リンクで大きな障害があったことがわかる．しかし，AS パスの増減に関するイベントが，このような断線イベントとはあまり類似性が見られないことから，ピアリングしているルータ自身の障害ではないことが予想される．また経路変更イベント (Lp, Sp)(Lvp, Svp) は似たような振る舞いを示していることが見てとれる．これは，ある Prefix に関して，Lp と Sp もしくは Lvp と Svp が対になっており，比較的短い時間間隔で経路のスイッチが起きていることを意味している．つまり，Wd イベントを伴わずに，経路上の AS で障害が生じて，最短経路が次善の長いパスを持つ経路に切り替わり，そして復帰することに対応する．しかし，厳密には両者は一致しておらず，全般的に Sp より Lp の数が多いことを考慮すると，経路が延びる方向には働きやすいものの短縮される方向には働きにくいことを示唆している．さらに，興味深い点としては，Sm イベントは他のイベントとの類似性がほとんど見られずに，頻繁にバースト的な振る舞いが見られることである．

[§]例えば Prefix10.0.1.0/24 に対して 10.0.0.0/8 のような Prefix

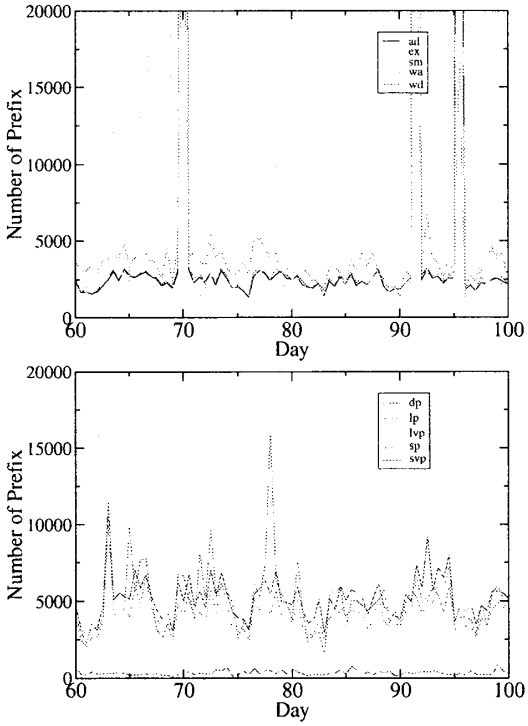

図 1: イベント時系列

3. 解析結果

3.1 各イベントの出現頻度

表 2 に測定時間内に観測された各イベントの割合を示す．表中の 10min は測定開始後，最初の 10 分間のイベント数を表し，rest はそれ以降に生じたイベント数を表す．10min では，ピアリングしているルータからのフルルートが到着するため，インターネット上の全ての経路数に相当する Ad イベントおよび Ex イベントが発生する．この間に，観測ルータではルーティングテーブルの計算が行われる．これらのイベントは測定のために必要

表 2: 各イベントの出現頻度

event	10min	rest (%)
Ad	83863	2409830 (17.6)
Ex	38802	1541636 (11.2)
Wa	36	1068882 (7.8)
Wd	27	2877933(21.0)
Lp	551	1290909 (9.4)
Sp	944	1030851 (7.5)
Dp	861	1185365 (8.6)
Lvp	1	71022 (0.5)
Svp	1	69232 (0.5)
Sm	1563	2174781 (15.9)

な処理であり，実際の経路制御のダイナミクスを反映しているわけではない．そのため，以降の解析では，最初の 10 分間のイベントを除いたデータを用いる．rest のデータを見ると，狭義の意味で到達性が失われるイベント Wd は全イベント中の 21%程度の頻度で起きており，必ずしも到達性が広く実現されているわけではないことがわかる．また，同種のメッセージが到達するイベント Sm も 16%程度で生じている．プロトコルの仕様としては経路に変更がない場合には，同種のメッセージが生成されることはほとんどない¶ことから，この値はかなりの高い割合であることがわかる．

3.2 断線時間の分布

次に，各 Prefix での断線時間に着目する．断線時間は，実際に該当 Prefix への到達性がなくなった期間に相当する．イベントでは，ある Prefix に関して Wd イベントが生じてから Ad もしくは Ex イベントが生じるまでの時間と定義される．直感的には，断線時間にはルータの (再) 起動に要する時間のスケールと，大規模な故障に伴う時間のスケールがあると考えられる．もし，断線時間がランダムであると仮定するのであれば，その存在分布は指数的な減衰を持つことが予想される．

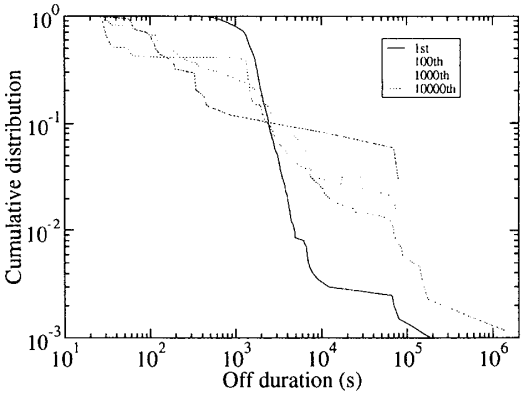

図 2: Prefix ごとの断線時間の分布

図 2 に断線時間の累積分布を示す．プロットはそれぞれ Wd イベント数の順位の異なる Prefix を表しているが，経路の安定性によって分布型が大きく異なっていることがわかる．1st(最も不安定な経路) では，1000 秒未満の短い断線時間はほとんど存在せず，10^3 から 10^4 秒のあたりに集中しており，一定時間の間断線するイベントが繰り返し生じていることを意味する．これは，ルートダンピング [5] の影響であると考えられる．ルートダンピングは，頻繁に断線するパスに関して，しきい値回数以上の断線が生じると一定時間 ad メッセージの送信を抑制する機能である．この機能によって頻繁に生じる経路計算の負荷が軽減される．つまり，当該 Prefix の断線時間は単純な故障ではなく，設定もしくはリンク自身に問題があると示唆される．

逆に，1000th(1000 番目に不安定な経路) のようなプロトコルによる制約がないリンクでは，分布は長い裾を持った分布となる．すなわち大部分の断線は 10^3 未満の短いタイムスケールで生じることが多いが，上限サイズが存在しないため，確率的には 10^6 秒にもおよぶ大きな障害が生じている．

さらに，100th では，2 つの特徴的なサイズが読み取れるが，30 秒付近のジャンプは，ルータの再起動によるものと考えられる．

3.3 イベント発生の局所性

各種イベントが全ての Prefix に対して均等に生じているのか，もしくは強い局所性があるかを調べるために，各 Prefix ごとのイベント生成数を解析した．Prefix あたりのイベント数累積分布を図 3 に示す．全てのイベントに関する分布 (図中の all) はベキ的減衰 $P(N) \propto N^{-\gamma}$, $\gamma \approx 1.5$ で特徴づけられる．つまり大多数の Prefix でのイベント数は少数であるものの，少数の Prefix に関するイベントは多数存在する．例えば，全 Prefix のうち約 90%では，それぞれ 50 回程度のイベント数しか生じていないが，10000 回程度のイベントが生じる Prefix も存在している．言い換えれば，イベントが生じる経路には非常に強い局所性が存在する．これは，経路の変動が空間的にランダムに生じていないことを意味する．

同様に，データをイベントごとにブレイクダウンした結果を見ると，どのイベントに関しても短いタイムスケールを除いて，広くベキ的減衰を満たしている．さらに興味深いことに，その指数の値はどのイベントでもほぼ同じである．これは，ネットワーク上でランダムに各イベントが起きている可能性を否定している．

3.4 イベントパターンのダイナミクス

Prefix ごとに Ad イベントが生じてから Wd イベントが生じるまでを一つのダイナミクスと捉える．果たして，Wd イベントが多い Prefix と少ない Prefix では，ダイナミクスに典型的なパターンが存在するのだろうか？図 4 は，prefix ごとの Wd イベントの順位と，その Prefix に現れたイベントのユニークなパターン数をプロットしたものである．Wd イベント数 (実線) に比べてパターン数が少なければ，経路は安定したパターンで揺らいでいると言える．1〜20 位までは多数のパターンが検出されているが，30〜100 位では単純なパターン (2 種類) しか現れない．これは，単に経路が断線しやすい場合にも，その原因および断線までの過程が全く異なることを意味している．また，1〜7 位および 30 位〜110 位はそれぞれ同じ ASpath を共有しているため，ユニークなパターン数は同数となった．さらに 100 位以降では色々なパターン数が混在している．表 3, 4 に，Wd イベントが最も数が多い Prefix(1st)，100 番目 (100th)，1000 番目 (1000th)，および 10000 番目 (10000th) の Prefix に関するイベントパターンとその分布を示す．1st および 1000th では多種のパターンが観測されているが，一度代替経路が選択された後に (元の経路に復帰せず) 経路が失われるパターン (ad:lp:wd) が多いことがわかる．また，100th や 10000th では代替経路が提供されていないリンクでの断線が原因であり，トポロジー的に改善の余地があることを示唆している．

¶可能性としては，属性情報が異なるものが到着しているケースおよび，wd メッセージが到着してすぐに ad メッセージが到着した場合，上流 AS のルータが wd メッセージを観測ルータに送信することなく，ad メッセージのみを送信することによって引き起こされるケースが考えられる

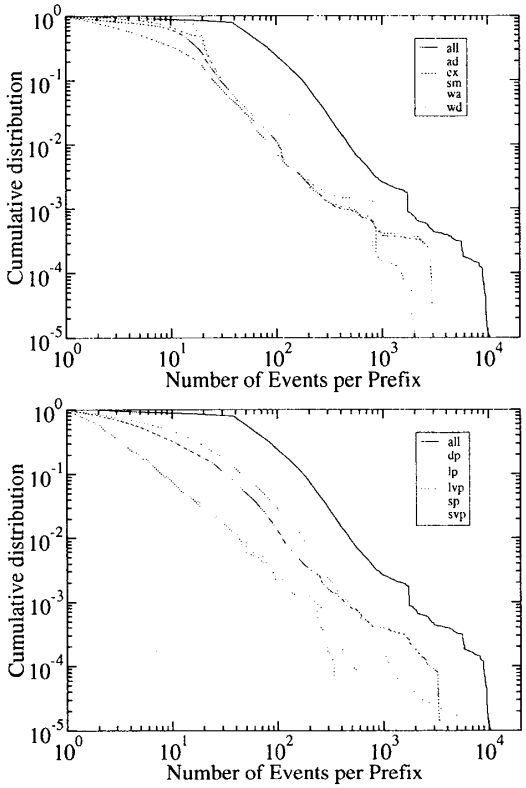

図 3: イベント発生の局所性

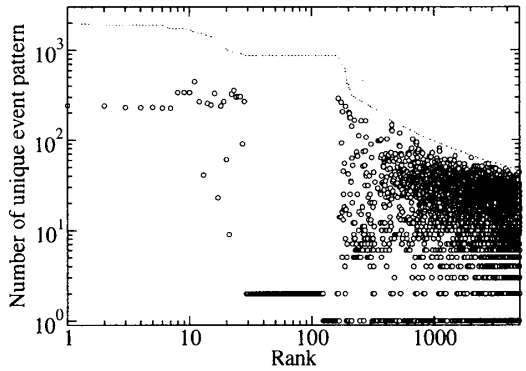

図 4: ユニークなイベントパターンの分布．実線はその順位での Wd イベントの総数を示す．

4. まとめ

本研究では，100 日に渡る BGP アップデートメッセージから再構築されたイベント時系列に着目して，BGP ダイナミクスの統計性を解析した．主たる知見は，(1) イベント時系列は必ずしも互いに同期しているわけではない，(2)ASpath 長は振動する傾向にあるが伸びる方向に働きやすい，(3) 経路計算に直接影響を及ぼさない重複した ad メッセージの到着数は全体の 15%超を占める，(4) イベントが生じやすい経路にはベキ的減衰で特徴づけられる強い局所性が存在する，(5) 断線時間の分布およびそのイベントパターンを見ることで経路の特徴が抽出可能，である．

表 3: イベントパターンの内訳．最後の行はユニークなイベントパターン数を表す

path(1st)	%	path(100th)	%
ad:wd	34.9	ad:wd	99.7
ad:lp:wd	11.4	ad:sm:wd	0.3
ad:sp:lp:wd	5.1	-	-
ad:sp:wd	3.8	-	-
ad:lp:sp:lp:wd	3.7	-	-
ad:lp:sp:wd	3.6	-	-
ad:sm:wd	3.3	-	-
242		2	

表 4: イベントパターンの内訳 (cont)

path(1000th)	%	path(10000th)	%
ad:wd	63.5	ad:wd	100
ad:lp:wd	9.4	-	-
ad:lp:sp:wd	6.3	-	-
ad:sp:wd	4.2	-	-
ad:sm2:wd	3.1	-	-
ad:lp:dp:wd	2.1	-	-
ad:svp:sm5:wd	1.0	-	-
17		1	

断線時間の分布およびイベントパターンを組み合わせることで，障害の分類および経路の特性をより詳細に知ることが可能となる．今後は，(1) 得られた知見をルール化しエージェントベースのネットワーク診断システム [6] への応用，(2) 時間・空間的な BGP トラフィック統計性の解析手法 [7] を組み合わせた多地点情報を用いた経路の特徴づけを行う予定である．

参考文献

[1] Y. Rekhter, and T. Li. "Border Gateway Protocol 4, " *RFC 1771*, Jul., 1995.

[2] M. Faloutsos, P. Faloutsos, and C. Faloutsos. "On power-law relationship of the internet topology," *Proc. of ACM SIGCOMM 99*, pp.251–262, (1999).

[3] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. "Delayed Internet routing convergence," *Proc. of ACM SIGCOMM 2000*, pp.175–187, (2000).

[4] L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S. F. Wu, and L. Zhang. "Observation and Analysis of BGP Behavior under Stress," *Proc. of ACM SIGCOMM Measurement Workshop 2002*, pp.183–195, (2002).

[5] C. Villamizar, R. Chandra, R. Govindan. "BGP Route Flap Damping, " *RFC 2439*, Nov., 1998.

[6] O. Akashi, T. Sugawara, K. Murakami, M. Maruyama, and K. Koyanagi. "Agent System for Inter-AS Routing Error Diagnosis," *IEEE Internet Computing*, vol.6, pp.78–82, (2002).

[7] K. Fukuda, T. Hirotsu, O. Akashi, and T. Sugawara, "Time and Space Correlation in BGP Message Flows," (submitted).