
■ Interdomain Routing Security Workshop 9
■ 日時: 2006年6月23日 15:00～18:00
■ 会場: Cisco Systems 赤坂オフィス

■ネットワークの高速切り替え手法の検討
NTTコミュニケーションズ 鈴木昭徳さん

□アジェンダ

- 自己紹介
- きっかけ
- 切り替えステップ
- BFD
- コンバージェンス
- Ether-OAM

□自己紹介

（少し前）AS7515/2914 設計・構築・運用
トラブルシューティングとか・・・

（今） 現場を離れ、技術戦略系（国内）お仕事
新技術の目利き
MPLS, GMPLS 技術の新網の玉込め

□きっかけ
L2 スイッチがあると、故障時にリンクが落ちないので、すぐには切り替わらない
-> 特にIX

このような場合にはルーティングプロトコルの Hello に依存していた。
デフォルト運用だと、1分半とか3分待たないと落ちているのがわからない
-> Hold Timer の Expire 待ち

□切り替えステップ
"切り替え時間"＝"断検知時間"＋"再計算時間"

断検知時間へのアプローチ
Hello のチューニング
再計算時間短縮へのアプローチ
IP Fast Reroute (?)
BGP Convergence Optimization / Cisco の GRIP

□BFD
Bidirectional Forwarding Detection
1秒以下で **障害をすること検知のみ** を目的
Hello や KeepAliveと同じ正常性チェック
BFD パケットの msec 単位の送受信
IPレイヤ上で動作
IGP, BGP, RSVPなどに断情報を通知

min-TX-interval
自身が許容する最小の送信間隔(msec)
min-RX-interval
自身が許容する最小の受信間隔(msec)
multiplier(ホールドタイム)
受信間隔に対する乗数(整数値)

送信側が 10msec, 受信側が 100msec を期待している場合、
ネゴシエーションの結果 100msec となる

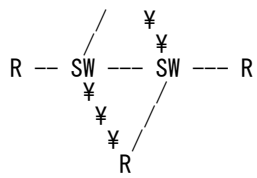
Q: デフォルト値はある?
A: デフォルトはない。機器の仕様上明記しないと動かない

IXを想定し、BGPで検証してみた
- "Keepaliveを最小" vs "BFDを用いた場合"

BFD
10msec/3回で落ちる設定でも問題なし
ばらつきがある
1 msec で multiplier 1 回という設定もできた。
さすがにそれだとばたつくが、10msec ならラボ環境では問題なし

- ばらつき…精度?
- コンバージェンスのぶれ?

R
/ ¥



KeepAlive の場合、
　keepalive を短くすると不安定になると警告されるが、問題なし
　Inter: 1秒
　Hold: 3秒

　切り替わり時間(平均) 2.887

　ISP ではないが、NTT-Com 様 L2MPLS 網では IS-IS で Interval 値 2 秒、
　Expire 値は 6 秒

　元々10秒/30秒だったが、サイレント故障(SONETの口は生きており、PPPの
　レイヤまでは生きているが故障している)を検知するために短めにしている
　オーバーレイの機器よりも短くしたいようだ

BFDの実装状況

C社:
　EIGRP
　OSPF (Juniperと相互接続確認)
　BGP
　IS-IS
　※いずれもv4
J社:
　OSPF
　IS-IS
　Static
　RSVP
　LDP
　※いずれもv4
　BGPがほしい

□コンバージェンスへのアプローチ
　実はあまり有効な案がない

　IP Fast Reroute (複雑...)
　　→トポロジによる
　　→IGPベース
　BGP Convergence Optimization
　　40%短縮、Black Box?

□Ether-OAM ってどう？

Y.1731 や 802.1ag
-> VLAN 単位で次のことができる
　CC - Continuity Check
　　BFDに近い。定期的にチェック
　LB - Loopback
　AIS - Alarm Indication Signal
　LT - Link Trace
　LM - Latency Management
　DM - Delay management など

パフォーマンスチェックする標準機能も検討されている
お客さんが ONU の先を落としてしまった場合
　ONU が AIS を投げ、配下の故障だとわかる

□まとめ
　BGPのインターバルを短くするとflapすると警告されるが、
　実際に1秒,3秒でも問題ないことをラボ環境で検証
　コンバージェンス時間短縮のアプローチは、網構成に依存したり、
　ベンダ各社の日々の努力に頼っている。増え続けるBGP経路などの
　状況を考えると標準的な手法が欲しい。

□おまけ
　「IX環境の高度化」というテーマでやっている
　- 高速切り替え
　- C/Dプレーンの分離など
興味がある方は a.suzuki@ntt.com まで

Pre-Computation Sec. Routing table
　事前に2つのRT、メモリが倍

□質疑応答

Q: 最後のやつの利点は?
A: あらかじめ計算しておき、FIBを切り替える
Foundryさんで似たようなことをしていると聞いた事がある
Forwarding tableが変わってない

Q: Forwarding tableってパスっと変わる?
A: XR 3.4 から変わる
マルチパスならロードバランスパスがある状態なので、ぱすっと変わるようになる
Route Reflector の場合、ADD_PATH をつけることで Equal cost でない場合も対応予定

メモリ、ルックアップ、コンバージェンスと3つ別の方向になっているlookupが遅くなるとかの弊害も考えられる

Q: IXでやろうと考えていると、BFDを投げたとしても相手が反応しなければならない。IXって一人一人やらなければならないから、スイッチが仲介することはできないのか。IXの中のスイッチがいて、それをチェックすることはできないか。
BFDコーディネーション的役割。

A: 今の主な実装は、Ethernet は静かに死ぬため報告しない
Q: IXでやっているとことはないのではないか

あるポートが死んでいると、反対側が検知できなくて死ぬ。俺は死ぬとEtherのレベルかIPのレベルかわかる、などの機能があると良いのではないか

BGPを切るときはshutdownを叩いてから切りましょう。

Q: shutdownしてから落とす人？
A: 挙手 ... 半数ぐらい
数年前まではそのまま落としている
ゲームとか90秒切れたらまずい

Q: shutdownするとき、peer-group でまとめて落とせるか？
A: できる、peer-group の一個だけもできる
最近は neighbor bgp で shutdown を叩ける
re-activate できる
障害の時にはどうにもならない

Q: 短くして使っている人は?
A: デフォルトはいじっている。
Cisco : 60, 180 なので 30, 90 にしている
逆に長くしてくれ、といわれたこともある
アジアと繋がっている国際回線上で、アジアの対向上のルータが強くなかったためそうしていた

コンバージェンスのところだが、BGP の fast convergence とかを google で捜すと資料が出てくる。path MTU discovery とか、input の hold queue とか、経路問題でも効いてくる

iBGP は長くてもいい。OSPF 的に reroute してくれればいい

iBGP だと 576 になるらしく、TCP Path MTU discovery をいれると、コンバージェンスが良くなる

Q: msec を eBGP を検知する必要はある？
A: きっとない? お客さんが監視しているひとが多い
回線サービスではないが、1秒以下は障害と見なさないという仕様になっているらしい

MPLS なら MPLS のレイヤで、IS-IS は最後の沽券のような位置づけセカンダリパスを設けておいて、何百msec 待つ

Q: BFD 投げて相手が BFD 未対応の場合には？
A: 落ちっぱなし
想像でいうとlink up になっているのではないか

Q: その場合、BFD は open しつづける?
A: BFD が success してから open しつづけるのではないか
down を検知するところしかやっていないと、上がりにくくなるのではないか

Q: BFD 対応は Juniper と Cisco だけ?
A: 他は future release にキーワードが載っているだけで、
対応したとかは効かない

Q: 10秒ぐらいはどうか
A: 10秒だとBFDはいらないのではないかと言われるかも

Q: ボーダルータが Cisco の人同士でやってみる?
A: Juniper が対応していないからちょっとなあ...
Juniper に BFD いれろと要求した際、あまり要求がないといわれた。皆で要求していただけるとうれしい

JUNOS はいま7.6。8.1で載るといわれた。年内? 次は8になるようだ動くようであればいくつか事例を作り、IX上でやってみるのはどうか

Q: ちなみにIXの収束時間は、IXのスイッチ間が切れたときの収束にかかる時間はどのぐらいか? IX内部を接続している回線が収束するまでBFDで検知せずに待った方がインパクトは少ないのか?
802.1wだと2, 3秒かかる気がする
A: dix-ieはRSTPを使っているという噂
JPIXは手動切り替え(STPで切り替えるよりは)
JPNAPはRapidが動いている

1秒未満はやめて欲しい
早いと光パッチパネルで数十msecだが、ワーストケースだと1秒
ワーストケースだとBFDで切れてしまうかもしれない

Rapidを使うと収束は1秒ぐらい
msecでやられると違う
用途でやると推奨値によって違うのではないか

早く検知するなら、BGPのexpire timeを1秒にすればいいのではないか。
障害でどういうのが多いかというと、IXの中が切れるということもあるが、相手のルータが落ちるケースが一番多い

Q: BGPはチューニングをつめるとどこまでいく?
A: keepaliveは1秒、hold downは3秒
両端で違う値を使うことはできるが、短い値になる
ネゴシエーションで決まる
BGPのチューニングで不満が高まってくる頃にはBFDの実装も普及してきている?

Q: IX側としては、極端に短くしてほしくない? そんなのは関与しないのか?
A: IX側の事情は、間のリンクが切れているから検知を長くしてねということではないと思う (直感では)

Q: 300msec なら検知せず、1秒待った方がいいのではないか
A: Cの最長は10000msec(10秒)
ただし、4000msec以上に設定すると、値をそろえても上がらなかったBFDはお互いに長い方にあわせる

収束は数秒で収まったほうがいい
お客さんに推奨するかは考えていない

実際にやっているかは別とし、Keep-Alive は2秒にしている…と議事録に書いておくとか(笑)

以前実験の際には細かくしすぎて、ばたばたしてしまうことがあったCのデフォルトは長すぎる。180秒は長すぎる
ばたつくという話を考えると、皆Ciscoというわけはない。
世の中で一番遅いBGPは7000シリーズ
一番短いので5秒・15秒という設定はある

Q: 国際展開しているところは?
A: あまりない

Q: BFD は設定は neighbor 毎?
A: Cisco の BFD の設定は neighbor 単位でできる

```
router bgp zzz
        neighbor x.x.x.x fall-over bfd
```

fallover bfd という設定がグローバルにあれば全部に適用できるのだが…

Q: 相互接続は?
A: 10台ぐらいルーター集めてやってみるといいのではないか
いまのところはC社しかない
今年中にはJも対応する
引き続き少しずつしらべていく
皆でおためし実験会ができるといいな

■4Octet-ASの実態と憂鬱

まほろば工房　近藤邦昭さん

□2 Octet-ASの寿命 (64000を使い切るまで)
　　...2014年ぐらい
　　exponentialでいくと2011年,2012年ぐらい

□APNIC での対策
2007年1月から
　　希望者へ4-octet ASの配布開始
2009年1月から
　　デフォルトで4-octet ASの配布を開始
　　希望者には2-octet ASを配布
2010年1月まで
　　2-octet,4-octetの区別を行わず配布
　　つまり,4-octetしか配布しない

Q: 2-octet と 4-octet の区別がないから、状況次第で前の方から割り振りで
　軌道修正できる余地もあるのではないか?
A: あまり考えていないのではないか。そうであればスケジュールを後ろに倒す
　のではないか

Q: APNICのミーティングでは反対意見はなし
　こういうのを出して試行することで、4-octet AS を早く復旧させたいと、
　という意図があるのではないか？
　…かくして4-octetの刺客はやってくる

Internet-Draft は出ている
2-octet であっても NEW_AS_PATH があれば対応できる

□どんなん？
4-octet から 2-octet 空間に open する場合
BGP の capability check して、AS_TRANS (AS23456) で open

知らないといわれたら、NEW_AS_PATH をつっこむ
AS_PATH 属性の中に、NEW_AS_PATH 属性をつけて投げる
　　本当にわからないなら捨てる
　　→4-octet AS は知っていることを前提にしているので、
　　AS PATH の情報がなくなる

2-octet から 4-octet に出るときも同様
　　NEW_AS_PATH を送る
　　AS_PATH だけしか送らないと、loop detection ができない

ID の中で 2-octet のルータは、NEW_AS_PATH に対応してねと書いてある
…でもそうなら 4-octet に対応する

消失すると、4-Octet 空間は、AS_TRANS の山に prepend されているように見
える
-> いろんなASが同じoriginに見えてしまう

トラフィックエンジニアリングは当然できない

□サポート状況
JUNOSe 4.1.0 以降サポート
IOS XR 3.4
ソフトウェアルータ

TWNIC OPM での資料がよく書かれている

□というわけで
　　- ループしても大丈夫?
　　- みんな一斉に 4-octet になれないよね？
　　- 自分のASの到達性に不安を感じませんか？

トランジットプロバイダが 4-octet AS を使い出したときにインパクトが多い

Q: 4-octet が 2-octet を受けて、4-octet に渡しても問題ないが、
A: 誰が最初に始めるか? Geoff? 彼は今ネットワークを持っていない
　音頭取りしながら少しずついれていかないと...。まず使ってみたい。

Q: そういう表明をしているのはAPNICだけか?
A: 他のRIRも足並みをそろえてくる
　NIRはそれでやるしかない
　今一番 ASN のばらまき方が激しいのは ARIN
　APNICも増えている。ARINも収まったのではないか。

数年前にUSのキャリアと話したとき、興味はないとのことだった
そのうちだけれども、4-octet capable にするリスクをおかすモチベーション
はない

Q: 4-octet をもらった人が困る?
A: 世の中にほとんど 2-octet しかみえない中で、4-octetをもらった人が
いると、どの AS_TRANS に向かっているかわからない

loop detection ができるかが最大のネック
経路が存在しなければただのunreach
prefixが違ってoriginが違うと、別の経路になる

4-octet AS に移行するという意欲を感じない
持っている人はモチベーションがない - 自分の経路が届けばいい
セキュリティホール修正の際に4-octet対応コードをいれるとか

ぶっちゃけベースでいうと、IXの中のシステムも対応しなければいけない
　　管理システムが 2-octet ベース
　　netflow, sflow のコレクタが全滅
　　顧客システムでの対応が必要

Q: Transit ISP に とりあえず4-octetでつなぎたいと言ってみるのは?
A: 4-octetに対応している数台のルータで収容…

Q: レジストリによるチェックは？
A: ASN を出したあと、流しているかチェックしている
　流す範囲を限定しているなど、こういう風に流していると説明が必要
　JPNIC は 4-octet AS を試す必要がある
　IOS XRしか対応していないとなるとJPNICは買えないだろう

Q: Interopでやるのはどうか? 全体のShownetのASを4-octetでやるは?
A: 今年は内部でASを分けた
　Routing Registry のシステムが対応するのは無理ではないか
　　APNIC は RIPE whois を利用しているが、無理ではないか
　　以前から RIPE に変えるとはいっていたが、ARIN の whois も
　　RIPE に変わった

Q: Zebra は対応する?
A: メンテナンスが止まっている。Quagga だったら可能性はある
　ZebOS ベースでインプリされる可能性はある。
　順次ソフトウェアルータは対応していくだろう

口質疑応答
BGP はエンタープライズでローカルで投げる運用をしているところがある
最大シェアは YAMAHA
拠点数がグローバルにつながらないが 65535 だと足りない
-> 4-octet ならいいのかもしれない

無茶な使い方ではあるが、あまり表にでてこない
相互接続する必要がなければ勝手に使ってよい

IPアドレスに関しては、プライベートアドレスが足りない、ということを
レジストリに示すとグローバルから割り当てるというルールもある

4-octet AS の相互接続検証については、DistIX が興味をもってくれるのでは
ないか

■xSPルータにおいて設定を推奨するフィルタの項目について(IPv6版)
KDDI 石原清輝さん、KDDI 向井将さん、DTI 馬渡将隆さん

今後、IPv6 の文章は、IPv4 の文章同様、JANOG Comment にしようと思っている

Q: Last-Call 手続きにはいって OK か?
A: OK

IPv6 のルータをオペレーションしている人はいるのか? - 数名挙手
人間のリソースがあるのなら英語化などをしていきたい

- Team Cymru
背景: 前回の IRS で IPv6 の bogon リストがないという話が挙がった

IPv6 のリストはあることはある
　　ベータ版; ここ三年ぐらいメンテナンスしていない
　　これをスターティングポイントにして手伝ってくれる人がいれば
　　welcomeとのこと

まずはドキュメントの英文化をすすめる
　　Team Cymru に協力する会を作るのはどうか?
　　IPv4 については nspsec-jp が管理している

Q: IANA のデータと何が違うのか?
A: Allocation を見ている。IANA は RIR レベルの割り当てを掲載

allocation されているかどうかを見てフィルタしているところもある
ワーキンググループを作成してはどうか

■次回のIRS

ネタが少ない
-> Security を抜いて Inter-Domain Routing にする

例えば..

- IXにおける増速時における作業のtypicalなやりかた、
- BGPを複数本張っている場合のばらしかた

単純に multipath にし、ルータにばらけさせている
向こう側が multipath にしないと、(in 側が) 偏ってしまうケースもある
ルータのマルチパスのアルゴリズムによる
こっちから送るプレフィックスをルータ毎に分けたり…

高田さんは資料を作る時間がない…

Q: 資料なしでプレゼンするのは?
A: 資料なしでは難しい

全く資料を用意しないわけではなく、PPT 2,3 枚程度の資料を
用意し、ログを取るとtipsが溜まっていく

□ネタ
- 最近の netflow のトレンド
- peer watcherの裏側を探る
- 紫が最近出した BlackDiamond の小さいやつってどうよ

□Link Aggrigation 横並び対決:
JPIX, JPNAP, Equinix, BBIX
-> Link Aggrigation やってる、やってない
IX がサービスを開始するにあたり、どの程度相互接続の裏を取ったか
が聞きたい

□その次 : 相性問題話
100M だと平気だが、300M だとぱたつくリンク (Foundry - Juniper)
Juniperの特定のPICがダメ

□匿名座談会
音声変換・磨りガラス越しにやる?
困ったことを持ち寄る会にするのはどうか?

■次回の日程
2006年9月22日(金)
会場: Cisco Systems 赤坂オフィス

ネタ1: netflow, sflow
ネタ2: 高田さん独演会

以上。