==========
IRS Workshop 8 Meeting Summary
日時 2006/04/21 15:00-18:00
場所 NEC本社ビル 241
==========

■ タイトル : OSPF の無停止メンテナンスへの挑戦
発表者　: NTT データ 吉野 誠吾
発表時間 : 15:00-16:30

IRS のターゲットとしているセキュリティには直接関係ないが、安定してうごかないと、セキュリティもないだろうということで、取り上げさせて頂く。
最初は OSPF のおさらい。時間があれば、OSPFに関して覚えておきたいことなどに触れる。

o OSPF の DRインターフェース断が他の通信に与える影響
JANOG ML でやりとりしていた DRインターフェース断が他の通信に与える影響を取り上げる。

1. 問題の概要
   どんな問題だったのか？
2. OSPF
   2.を理解すると、3はすぐ理解できる。
3. 問題発生のメカニズム
4. 解決策、回避策
   4. はMLでもらったコメントを整理。
5. ルータの実装改善提案
   最後に、OSPFもしくはルータの実装をいじらないとだめだが、改善提案について。

MLでは話題にあがらなかった話しも１つあるので、それも後程触れる。

◎ 1. 問題の概要

OSPF を用いたネットワークで DR のリンクが落ちると、LAN 上の他のルータの通信も止まってしまうこともある。OSPFの仕様上の問題。どのメーカのルータでも起こり得る。

o 1. 問題の概要：図で示すと

たとえばメンテナンスで、RT21 を作業したい場合、interface を上から shutdown していこうとした時、メンテナンスするので、traffic を予め通さないようにするために、

コストを上げて RT22 経由に移動させる。そこで、安心して RT21 を shutdown すると RT22 経由のトラフィックも止まってしまう。これは問題ではないか？

o 1. 問題の概要：発生のトリガー

メンテナンスについて中心に話していたが、機器障害でも同じ。

o 1. 問題の概要：二重化されていても
二重化されていないからいけないのでは？と指摘されたが、たとえば NWB1 で束ねている場合。二重していてもRT21 で落すと、DR が２つとも RT21 になっていると、すべて止まってしまう。

- DRが１台のルータにあつまる可能性
  => ルータプライオリティを設定している場合は大丈夫だろうが、Router ID の大小で決ってしまう。
  => メンテナンスによって、１台に集まってしまうこともある。

◎ 2.OSPF: OSPFとは

問題発生のメカニズムを理解するには OSPF に対する知識が必要。

- ospf をどう表現するのか？

リンクという言葉があるが、この概念が重要。

※)ルータに隣接して書かれている数字：コスト値

ここでは、

2-1 3つの Link type (p-p, transit, stub)
2-2 Router LSA と Network LSA
2-3 Link type と LSA の関係
2-4 LSA の flush(消去): MaxAge

の 4つについて説明していく。

リンクタイプについては、Virtual Link もあるが、それは使わないだろうと思って、上記でまとめさせて頂いている。
また、LSA が無効になる時の動作も、ここでおさらいしている。

o 2-1 3つの Link type (p-p, trasit, stub)

1.point-to-point

ダイレクトにつながっている場合。router link と呼んだ方がよいということを書いているのは後程理由(※2)を述べる。

2. transit network
LAN のネットワークに複数台ぶら下がっている場合(今回話題になっていた箇所)

3. stub network
ルータがネットワークに１台しかつながっていない場合。この部分を stub network と呼ぶ。無理矢理図で書くと、ここで挙げたような図になる。

重要:
右下もtransit 。隣接ルータが１台でもあると transit。MLでうまくいかなかったという話をしたが、これは stub, trainsit の違いによる。

※2) link
回線に見えるがルータから見た接続関係と捉えた方がわかりやすい。表を参照のこと。

- p-p
point-to-point だと Link-ID に隣接ルータの Router IDが入る。

- transit network
Link-IDには DRのインターフェースIP

- stub network
IP prefix が Link ID に入っている。

補足としては、IP subnet mask について。p2p routing table をつくるには、別途情報が必要。
Link type が 3つあるのは理解しずらかったので、しつこく説明して見た。

Q. Cisco だとピントこない。コンフィグ的には？
redistribute は stub. Passive だと transit になる？
Router LSA を見ると、それで変わってくる。それがどれに bind されているの？

A. external の話はない。すべて intra-area の話。
network コマンドで記述された範囲の情報。
redistribute static とかしたものは含まれていない。あとで router LSA の話もしますので、その時にまた質問してください。

o 2-1 numbered p-p は stub が必要

OSPFの解説本に書いているかと思うが、/30 とか振っている場合、Router LSA では2つのリンクの情報として表現される。

```
     10.0.1.5/30          10.0.1.6/30
[RT]-----------------------------[RT2]
```

Link Data に prefix のマスク情報が入っている。つまり stub の情報が必要になっている。

o 2-2 Router LSA と Network LSA

Link Type の話から一旦離れる。Router LSA と NEtwork LSA の話。
実際のDBの断片をLSAとして扱っている。RFC では 5つ定義されているが、エリア内のトポロジーであれば２つあれば表現できるので、このドキュメントでは 2つ説明している。

o 2-2 Router LSA

Router LSAってなんだっけ？
=> Router につながるリンクの一覧。図の緑で囲っているルータを例にとると 4つ作られる。

o 2-2 Network LSA
もう 1種類。Network LSAについて
trainsit ネットワークの場合。複数ルータが接続される可能性があるので、ネットワークにつながっているルータの一覧。

o 2-2 Router LSA のカバー範囲
Router LSA の情報だけで、どれだけのトポロジーが描ける？
=> traisit 以外。

o 2-2 Network LSA のカバー範囲
Network LSA の情報だけで、どれだけのトポロジーが描ける？
=> transit LSA のみ。

o 2-2 Router LSA と Network LSAで
Area 内すべてをカバーできる仕組みになっている。ここまでがLSAの説明。

o Link Type とLSAの関係

- Router 1台で ethernet(L2 とか)につながっている場合を考える。Stub であれば Router LSA のみで表現できる。そこにもう１台つながっていると、transit netwrok に変わってしまう。で、1DRが選ばれるが、それがNetwork LSA を生成する。

この状態で、Link が切れると、また Link type が stub にかわってしまう。この時、network LSA が削除される。

o LSA の flush(消去): MaxAge
routing プロトコルは分散プロトコルの 1種だとおもうが、OSPFには削除するためのメッセージやコマンドがない。
このため、MAxAge に 3600(秒)をいれたものを送信することで flush する。
ルータは 3600 秒たつと、LSAに変化がなくても、再送することになっている。大抵のOSPFの解説本はこの話題について触れられている。

要はここで、MaxAge をいれたものを送信することで flush することを確認したい。

今回の問題を説明するうえで必要なことは以上。

Q. MaxAge をいれることを flush ですよね？3600 という値が決め打ちで、コマンド代わりになっている？
A. 私も美しくない実装だと思うが、そうです。

Network type という概念がある。Broadcast, non-braodcast, p2p multicast .. とはちがう概念 => p-p

ここまで理解してもらえれていると、今回発生した問題は理解しやすい。

◎ 3. 問題発生のメカニズム。

前提：RT21 が NWB の DR になっていること

R21 において、NWB に接続している側の interface を shutdown すると、NWB がなくなったように RT21には見える。このため、RT21 は MaxAge をいれて、NWC 経由でルータに広報。他のルータは NWB がないという前提で routing table を再生成する。
RT3 からすると、NWBより上のルータが存在しないように見える。
BDR が DR に昇格するまで、通信が切れてしまう。

o 3. 問題発生のメカニズム：発生要因
なにがポイントだった？
1. Network LSA を DRルータしか生成できない。DRへ依存しすぎる。
2. 冗長構成で、裏のネットワーク経由で情報がつたわったのが問題。
3. これから落すよといったメッセージを伝える方法がない。

◎ 4. 解決策、回避策

問題を発生させてたくない！

1. OSPF をあきらめる。
2.ルータ間を直接接続する(back-to-back)

間に L2 スイッチが入っていても、1対1 ならOK。石井さんがVLAN で分けているから大丈夫とおっしゃっていたのはここからだろうと思う。

ISISに切替えるのも、トポロジーを返るのもなかなか難しい。新規に構築する際には考えた方がよい。

- 緩和策

タイマーを短くすればよいのでは？
=> １秒未満にしても、検出後ルーティングテーブルの再度生成を行なうのに時間がかかる場合もある。検出時間＝ダウン時間ではない

- 回避策
無停止にしたい！

残念ながら完璧な手法がない。しかし、こういう実装にすればいいんじゃというのが見えて来た。
それには 8つある。

o 4. 回避策案1 : 1経路ずつメンテナンスする

一応、全断は避けられる。ただし、コストをかえてまわらなくてはならないので、オペレーションが複雑になる。

o 4. 回避策案2 : 電源を落す

いきなり電源おとす。こうすれば、MaxAge 投げる暇もない。
=> 多分大丈夫。ただし、ルータは高度化しているので、今後電源スイッチを見て動作する実装がでてこないともかぎらない。

そもそも Juniper はおとしちゃだめとも言われているので、すべてのルータで行なえるわけではない。

o 4. 回避策案3 : DRを譲る

DRをゆずってからおちれば？
これが一番妥当に思えるが、可能な実装と、不可な実装があった。ただ、DRじゃなくなったときに flush してしまう実装がある。flush してしまうと、問題回避にならない(IOSだと flush しない. Jupier だと flush してしまう(2006年4月現在？)

Q. DRの priority を 0にすると、すぐに代わっちゃう？
A. 受け取った時、DRを選びなおすことをやる。いろんなタイミングでやるらしく、hello

パケットのインターバルまつ必要あるかもしれない。なお、0 にした瞬間に hello が

出るかどうかまでは確認していない。
なお、Jupniper だと変えた瞬間 flush してしまう。どう回避すべきかはRFCには書い

てない。

o 4. 回避策案4: DeadInterval を長く変更
永見さんから頂いた指摘。意図を確認し忘れたが、おそらくこういうことだとおもう

結局は DeadInterval を長く変更しても、回避できなかった。Timer値を変えると、Adjacency
がすべて切れてしまう。つまり RT21 にとって Stub に変わってしまう。その瞬間 Network
LSA が flush してしまう。

o 4. 回避策案5: Passive インタフェースに変更

いとうさんから話で passive interface すれば？　という話もいただいたが、これもstub に変わってしまうので flush してしまう。

o 4. 回避策案6: (floating) static で補完
最もですが、ISP 的に static を書いて矛盾なく redistribute をするのは、かなり現

実的には難しいかと思っている。他に手がなければ……

o 4. 回避策案7: Cisco graceful shutdown

よいコマンドだと思うが、この問題には対応できない。

Comment: トポリジを勘違いしてました by 河野さん

o 4. 回避策案8: point-to-multipoint mode

飯田さんのblogに書いてあった手法(※3)

p-2-multipoint mode にすればOK。実は Ethernet でも動く。ATM や FRなど、broadcast
がないリンクで、動かすための機能。ルータ間がすべて point-to-point 扱いになる。

問題の回避が可能は可能だが、いくつか制限事項(※4)がある。

- blog の話
「指名ルータをなくすには」
用途が違うので、Cisco とJuniper でタイマーが異なってしまう(Jupiper は timer が 10sec
のまま) Juniper は LAN用にサポートされているわけではないので、neighbor が自動検出され

Comment: interface によって、Cisco でもneihghbor を書かないとだめなものがある。
原因は不明だが点…(経験上) by 近藤

Comment: RT21 からみると、RT1 と RT22 がつながっているように見えない。ネットワーク構成
によっては、RT1 経由になるので、負荷がかかるので、気を付ける必要がある。
RT21 と RT22 の横っちょに、ネットワークがあって、その間で通信する場合など
=> すべてに p2m のリンクを張れば回避可能な問題

※3) http://tiida.cocolog-nifty.com/netblog/2004/10/7_p2mp.html

※4) Cisco からみると、transit に見えなくなる。Stub network で投げようとしているが、
/32 になってしまう(/24だが)
つまり、さっきのLAN上にサーバが置いてあると、通信できなくなる。

◎ 5. ルータの実装改善提案

1. p-2-m mode の broadcast network への適用
なんとかこれをLANで動くように標準化すればよいんではないかと思う。
おそらくここにあげた５つの点で行けるかと思う。
これを実装してもらえると、回避は可能だろうと思う。

2. DR をgraceful にゆずる。
flush をしないだけの話なのだが、古い情報が残るのは大丈夫か。

Comment: RT1 が flush を受け取ることが問題。受け取ったとしても、うまく処理すれば？
=> 他のインタフェースから受け取った場合など、すこし難しそう

Comment: Router priority をいちいち低くしなければいけないけど……

Comment: ISISとOSPFの大きな違い。detamistic??(書き取れず)

Comment: 復旧させても DR は、元にもどらない。

Comment: DR専用ルータおけば？

Commnet: p-p と stub だけ。transit network は不要になる。DRが不要になる。その分、
Router の負荷が高くなる。

Comment: Switch を置けば……(書き取れず)

Comment: OSPF をどんどんISIS化しようしている？

Comment: graceful restart を拡張しようという動きもある(後者の方法)。もうすこし綺麗
にしょうという話。ISIS化というのが、ちょっとある。p2m はISISにはない。

Q. p2m はいらないの？＞ISIS
A. NBLA はべつのトラブルがあるので……

Q. ISIS 化とは？
A. 私の考え。multipritocol family 。大きな選択をした。

Q. OSPF の MaxAgeを可変にできるようにするのは？
A. OSPFはincrimental なので可変にするとわかりずらいのでは？ISISは減算なので。

Comment: MaxAge は delete だけではなく、LSAがどっちが新しいかを判定するのに使ってい
るらしいので、おそらくMaxAge はいじれない。

Q. MPLS にするのは？
A. 余計なオペレーションが入るのでは？
A. でもOSPFを気にしなくてもよい。
A. MPLS自体のルーティングは気を使う必要があるよね。
=> 同じ問題が起きるけど、MPLS 的には起きないので、実影響はない。

Q. 同じ問題に遭遇したのは？
A. あるよ。DRだったことを気付いてなかった時が。３０秒まてば元にもどっちゃう。
   まぁインターネットだから。うちで起きなくても、他で起こしているだろうね。
   たしかに、身も蓋もないが、実際にある。

Q. Intra で考えたら問題だよね？
A. Static? ISISでやればいいじゃん。

Q. ISIS ならOK？
A. べつの問題はおきるかもしれないが、DR問題は起きない。

Comment: Enhanced EIGRPをつかえば

Comment: 最近は切れると厳しい

Comment: 昔は、EIGRPを使っていたが、最近はOSPFを使っている。

Comment: 差せば使える利点を取ろうとしているんだから、もしこれがいやなら、他
         で避けるべきでは

その他コメント多数(書き取れず)

「1経路ずつメンテナスする」
 ＝＞ これが一番いいのでは？

(floating) static で補完
1. static じゃなくて、default じゃだめ？そのルータにとっては、全経路あるのと一緒
。
  => 賛成。
  => Internet に抜けるというネットワークなら辛くなるのでは？

Comment: RT1 で 外へ出て行くとして、default をかいてあったとしたら、下側を真面目

         に経路を書かないとだめだね。

         やっぱり

         「1経路ずつメンテナスする」

          がよさそう。

         もどし忘れてたら、はまるが。

Q. Operational RFC とかにしないの？
  きっと誰かが 英語にしてくれるよ。

Q. ダウン検出について ietf にながれている。それがあれば解決するの？
A. 議論が迷走してる。

Comment: 大抵 Juniper のOSPF がおかしい

Q. 次のアクションを決めよう。
A. とりあえずベンダに投げようと思っている。

flush しなきゃいけないときは、flush する必要がある。ループしているところに flush

しているのが問題。

Q. Cisco の中で動きがある？
A. ある。

Comment: メーカというよりプロトコル問題

Comment: DR の思想に無理があるのでは……

Comment: DRを移った状態にしていると、DRが一台に集中するとはまる。

Q. IETF に投げてみる？
A. 似た問題が進行している。埋もれちゃうともったいない。

◎ 今後
  まずステータスアップデートする。

■ タイトル : IPv6 フィルタ
  発表者　 : KDDI 向井
  発表時間 : 16:30 - 17:00

前回の議論を元に、文書化した。別途、Appendix(6boneの停止についてなど)の追記を行
なう。IPv4 版もある。
当初は、peer と trainsit では、別々の文書にしていたが、１つにまとめてほしいとい
うリクエストがあったので、すこし内容としては冗長だが、一緒にした。
いままでの議論は盛り込まれていることと思う。

o 概要について
IPv6の運用をしているISPが少ないが、いまはこうだろいという思いもありつつ、概要を
書いている。

ターゲットは、もちろんIPv6ネットワーク。話題としては旬な、Outbankd port 25 blocking
などは、現状では視野に入れていない。この文書では、それぞれ最低限やりましょうとい
うことを記述している。

あとで webに載せるので、細かい点はつっこみをください。専門用語部分は IPv4 と同様。

◎ トランジット接続

o 経路Ingress フィルタ
ここが前回盛り上がった部分。/56 という話題ではあるが、どこまで accept するかと
いったことはなかなか判断できない。 or longer と記述しておきたい。
ドキュメントに書いちゃうと、やっちゃう人が出て来る。 exact に書くか、例で終らせ
るか。いまは後者。

Q. 基本的には /32 しか流れないんだけど、/48 が流れるので、こうしましょうという話？
A. puching がでてきたら、ISP的には救済方法として必要になる。それをきっちゃってよい
かというと……現実的には /48 は流れてて、/52 とかもある。

C: コメントとして、現状どういう経路があるか欲しい。このドキュメントを読んで入って
来ると、到達性がなくなるかも。

A. なんとも言えない。
未割り当てのアドレスは現状でかい。IPv6 の Bogon List は信頼できそうなのはない。
作りましょうというプロジェクトはどこかにあると思うが、みんなが見れる状態のもの
はない。 リファレンスとしてだせるものは現状ない。
ないのに、参照しろというのはまずそう。削る？
がんばって書く？
Bogon Route Server はまだない。

o Ingress の AS-PATHフィルタ
IPv4 と同様。たまに異常に長いものもある。トンネルの山？だれか追いかけた人がある？

Q. 50hop の根拠は？
A. 前回の会場からの挙手。IPv4でも同じ値を書いてある。

o Max-Prefix-Limits について
IPv4 で使っている人も多いと思う。

いまの段階で考慮されることを記述。

◎ Peer
peerの場合も考慮するところは変わらない。
peer の相手から広告をすると通知があった prefix のみをaccept ぐらいか。

o Ingress の AS-PATHフィルタ
ASパスの更新部分

非対象ルーティングも考慮すると、すべきでないこともあるが、余裕があれば、souce address check

o ルータ自身へのアクセス
o Ingress のパケットフィルタ
eBGP の部分

Q. これだと 179 port だけを許容すると書いて大丈夫か。
A. 自分がセッションを張りに行く場合と、どっちかが 179 。TCPを理解していない人までサポート範囲に入れる？

o Router のフィルタ
  System Protection ACL とは書いてあるが、自分のRPを保護するための技術なので、あれば設定してもよいだろう。

o IPアドレス割り振りリスト
Q. まとまっているところないの？IPv4は Team Cymru の方がまとまっている
A. Team Cymru output をまっている状況。叩き台を作って、やってよというのが、よさそう。

Q. Router に対する packet filter は、順番をまちがえると DADが動作しなさそう。
A. ICMP を基本的に受け入れると書いてあるので大丈夫だと思うが、触れた方がよいかもしれない。

Q. 8.2 との文書との差分は？
A. 中身については、ほぼ変わらない。
   packet filter には触れていないので、その点については、こっちにしか記述がない。
   英語にすると良さそう。

o 8-1-1 IPv6 BGP filter recommendation

Q. こまかいのを止めるというのは、ちょっとピントこない
A. もめている。ヨーロッパの影響がある。結構可変している。

Q. Black hole filteringという話は IPv6 ではないの？
A. 動作は IPv6でも同様。

Comment: 1年から２年後に考慮しなおせばよさそう。いま無理して書いても……

Q. これを英語にするの？
A. 当面はよさそう。まだドラブルを経験していないので。
   危なそうなので、先に決めようのが日本人、ドラブってからやるのが欧州人？
   ただ、こういうのがあるよというのはつたえておいた方がよい。

◎ 今後
Appendix 6bone と 6to4 について記述します。それをやればラストコールかな。
スケジュール的には、5月末ぐらいまでに書くのがよさそう。

■ タイトル : [ルーティングセキュリティに関連して
　　～IRRとzebraのInteraction実装のお披露目あるかも～]
  発表者　: NTT-C 吉田
  発表時間 : 17:00 - 18:00

・BGP経路情報の信頼性向上
　　IRRとzebraのinteraction

・経路情報の信頼性/信憑性
　　認証機構は働いていないので性善説
　　経路ハイジャックが実際に観測されている
　　more specificに流される
　　　さらにmore specificで対抗する。。。
　　認証機構の必要性を検討しないと。

・Inter-ASにおける経路の脆弱性「何を確認するか？」
　　受信した経路のOriginは正しいか？
　　まずはOrigin ASの認証/確認から
　　　S-BGP / soBGP / psBGP
　　　　PKIを使った経路配信メカニズム
　　　　実装はあるがdeployにはほど遠い
　　　IRRの利用
　　　　IRRのDB情報と経路情報の整合性を確認
　　　　IETFでPKIの認証基盤にも対応可能なよう調整中
　　　　　ROA情報をIRR objectへ、
　　　　　それをもとにroute object等の登録を
　　　　　するような基盤の確立

・IRR+zebra実装概要
　　BGP経路情報を受信する際に、IRR DBへ問い合わせ、
　　検索結果を経路制御に自動的に反映させる

　　誤った経路情報を受信しないことが可能

Adj-RIBs-Inに入る前にIRR DBに問合せ、
BGP経路情報のOrigin/prefixと
IRRのRoute Objectの整合性を確認。
irr status codeに基づき、
Local-RIB、Adj-RIBs-Outに反映

BGPとIRR情報の整合性度合いを 7つのstatus-codeで表現

Command : show ip bgp irr-status [1-7]
          show ip bgp irr-cache
          show ip bgp ※irr status-codeも表示
          clear ip bgp irr-cache A.B.C.D/M
          route-mapでirr-statusを指定

ログ出力
iBGPへの適用

利用方法(1)
        transit経路受信する箱(IRR+Zebra)
        transit経路をIRR+Zebraに注入、nexthopを書き換えて伝播

利用方法(2)
        JPIRRをmirrorしているIRR DBとZebraが連携
        transit経路中にハイジャックを検地したら
        細かい経路を被害者のASに広告(community付加)
        該当community経路をIGPに広告(nexthop変更)
                globalにも即座に広告
        Root Serverと同じ概念

利用するためには正しいDB情報での運用が重要。

IRRとのProtocol動作は？
        Zebra側で複数queryするとか

Future Plan
        実装
                問い合わせ先DBの冗長化
                専用protocolの実装 --- IRRdの改修
                        自身でDBを保持する？
                IPv6対応
                DB情報の修正によるstatus-code再検査機能
        標準化
                他ベンダへの実装のお願い
        オペレーションの連携
                JPIRRに登録されている日本の経路を確認し
                不可解な経路の検出する仕組み

codeのライセンスは？
        主張しないつもり。GPL？

利用方法(1)にはpeer以外にtransit経路もらえるのか？
        transit買ってればもらえるでしょ。
        それとも付加サービス？

IRR+Zebraが落ちたら影響度は？
        不正経路がcommunity広告されないだけ

IRR+Zebraの経路広告を信じるかどうかは各AS次第

各々のtransit経路をみてあげる
        極論すればIprefix規模のISPでも利用可能

ハイジャック経路対策で自AS内経路はrejectしている

Telecom-ISACに参加できないところは同様のサービスに
期待するしかない？

transit ASが下位ASに利用を要請(？)

JPIXのマルチラテラルなサービスに適用するとか

IRR+Zebraは、現在eBGP Multihop運用している

各ASはnexthop-selfせずにIRRに経路広告
        特定communityが付加されていたらnexthop-self

IRR-Zebraが不正動作したら細かい経路は増大する恐れ
さらに細かくちぎるレベルは？
        単純に半分にするだけ。

最後には？？？
結局自ASから広告しないのでは。。。

広告するのは別として、検知するところまでが現実的か。

破綻しないようにせめて日本国内経路のみで。
JPIRR+Zebra
