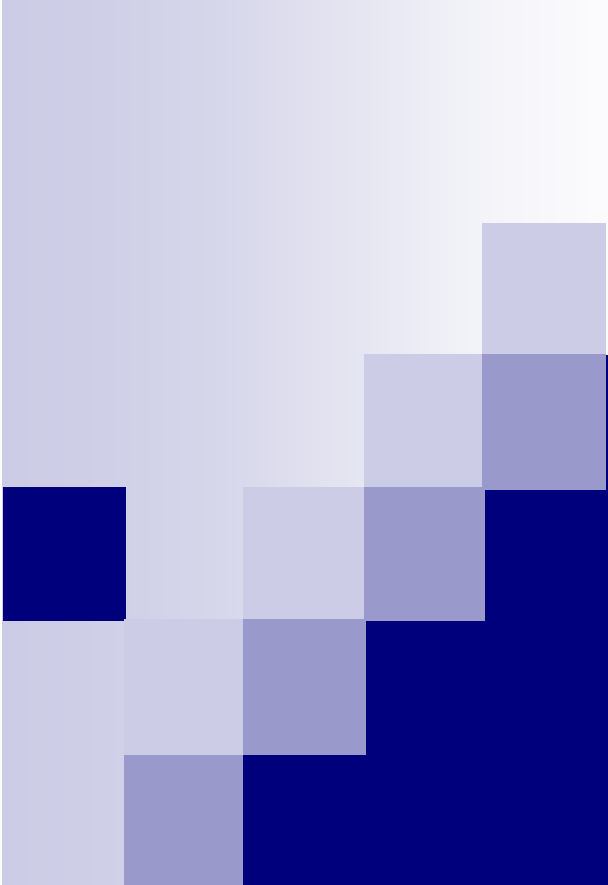




BGP Holdtime短縮の 運用

さくらインターネット(株)

技術部 大久保修一

ohkubo@sakura.ad.jp



今日のAgenda

- はじめに
- BGPのHoldtime短縮について
- ルータの検証方法について
- 実際の設定
- 現在の運用状況
- まとめ

これまでの高速切替に関する議論

- JANOG17 (2006/1/20)
 - ネットワークにおける高速切り替え方式の検討 (鈴木さん)
- IRS9 (2006/6/23)
 - IXの高速切り替えに関する提案 (鈴木さん)
- 第10回JPNAPユーザ会 (2006/9/20)
 - BFDの実装状況とその検証結果 (鈴木さん)
 - Cisco社ルータのBFDの取組み/実装状況 (河野さん)
- JANOG19 (2007/1/25)
 - 高速切替手法の検討(BFD for BGP編) (鈴木さん)
- IRS12 (2007/3/9)
 - Fast Detection, Fast Convergence
 - JANOG19 review and recent BFD standardization/implementation (鈴木さん)
 - From the IX point of view (三宅さん)
 - BFD deployment around the world (河野さん)
 - Y.1731 vs 802.1ag (石田さん)

たくさんありまし
た

他にもたくさんあると思います。



また高速切替ネタですか？ ^^;

- 既に議論し尽されている感じもありますが。。。
- 今回は、BGPのHoldtime短縮に特化した内容です。
- BFD、Ether OAMの話は出てきません m(__)m
- 弊社(さくらインターネット)で実施したBGP Holdtime短縮の運用について紹介いたします。

なぜHoldtime短縮か？

- Inter ASでの通信断にもお客様が気づく
- オンラインゲームサービス
 - ゲーマーは遅延に「とても敏感」に反応する。
- リアルタイム系アプリケーションの普及
 - ストリーミング、IP電話、株取引
- Inter ASでも断時間に対する要求はシビア
 - 数分間の断は致命的に。。。
 - せめて、TCPのコネクションがダウンしないレベルに
 - 10～20秒程度にはしたい
- BFDはどうか？
 - ルーターが対応していない ;;
 - ピア先も対応していないと
- BGPのHoldtime短縮で対応
 - タイマーは、Keepalive 5秒/Holdtime15秒を目標とする



BGPのHoldtime、短縮していいの？

■ 議論の分かれるところ

- Internet経路の安定性
- Inter ASでは、高速切替は必要ない
- してはいけない！！
- BFDを使え！
- 某商用IXでは、Keepalive 60秒、Holdtime 180秒を推奨

■ (某商用IX)「接続の手引き」より引用

(2) BGP のセッション保持時間

BGP のセッション保持時間は以下の値を推奨しております。

セッション保持時間 **180秒**, キープアライブ **60秒**

(CISCOルータ、RiverStoneルータではデフォルト設定です)

(Juniperルータでは変更する必要があります)

└─ デフォルト値(30秒/90秒)を変更してまでも
 タイマーを延ばせと。。。



個人的な考え

- 断検出はオフロードされた専用チップで処理されるべき(BFDなど)
 - ルーティングレベル(BGP)で高速断検出するべきではない
- Inter ASでは、1秒以下での切り替りは必要ないだろう
- IX内の障害切替え時間よりも短くすべきではない
 - IX内の収束が1～2秒程度であれば、BGPの切替は3～4秒程度
- 理想的には。。。
 - BGPのHoldtimeはデフォルト運用。
 - BFDをminimum-interval 1秒、multiplier 3くらいで動作させる。
- あくまでも、BGPのHoldtime短縮は、BFD実装までの繋ぎとする



Holdtime短縮の適用箇所

- Internet eXchange経由のPublic Peer
 - Switched network → リンクダウン伝わらず → 短縮する
- トランジット
 - 直接接続 → リンクダウンが伝わる → 必要なし
- Private Peer
 - 直接接続 → リンクダウンが伝わる → 必要なし
- BGPカスタマ向け
 - リクエストがあれば対応する
- iBGPピア
 - そもそもピアダウン発生を想定していない

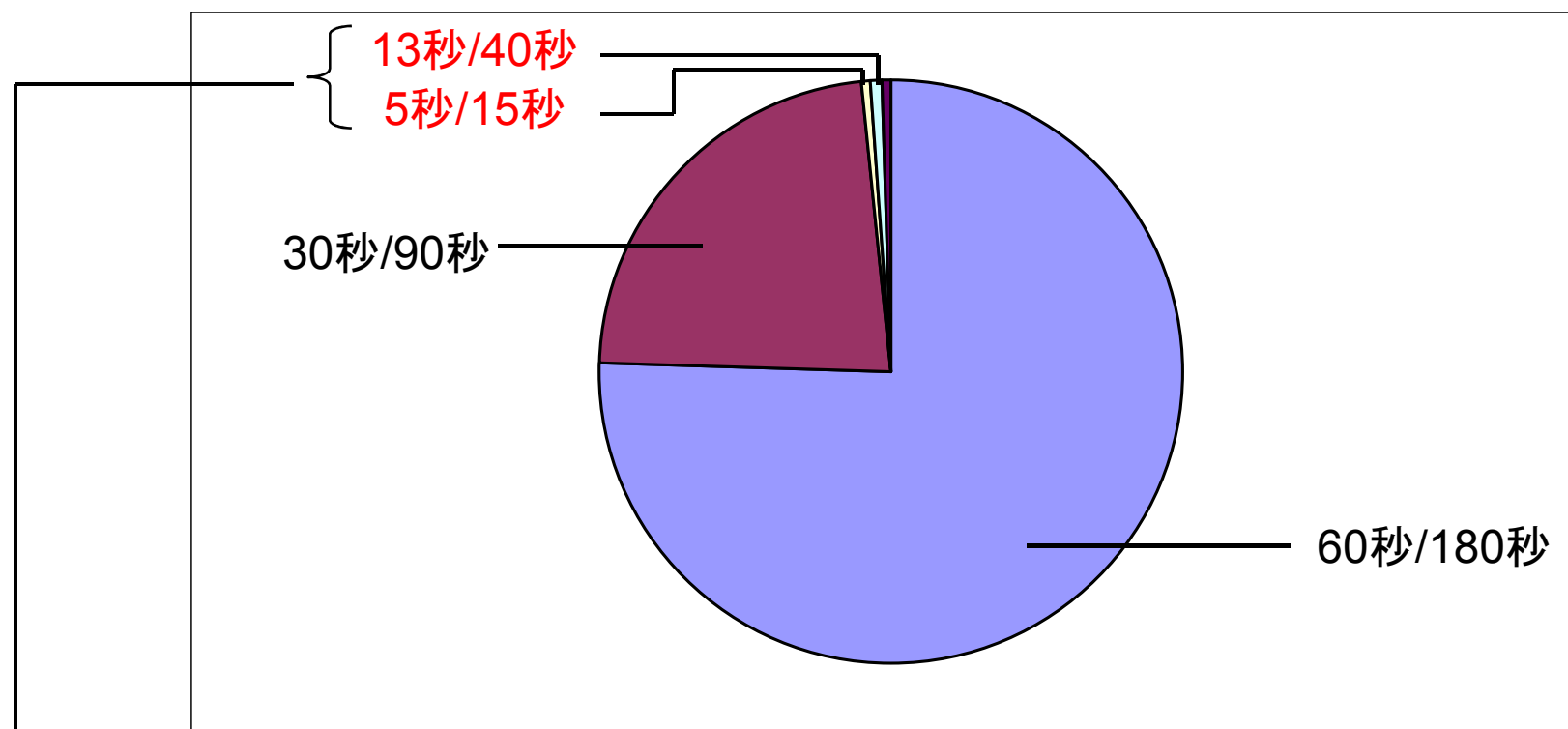


直接接続の注意事項

- (デフォルトで)リンクダウンしてもeBGPピアが即時にダウンしないルーター
 - Foundry
 - 下記コマンドで、即時にダウンするようになる
 - `router bgp`
 - `fast-external-falover`
 - ALAXALA、Force10は問題なし
- 直接接続でもリンクダウンが伝わらない
 - Link State転送できないM/Cが入っている
 - AutoNegotiation Disable状態で、片方向リンクダウン

皆さん、短縮してますか？

- 弊社のPublicピアの状況を調べてみました。



短縮されていたのは、2ピアだけでした。 ※2006年10月時点



ルーターの負荷は大丈夫か？

- 最低でも自分のところのルーターの動作確認はしておかないと。。
- 要件
 - 5秒/15秒で100ピア安定運用できること
 - トランジットフルルート → 2～3ピア
 - 数十から数千経路のFreePeer → たくさん
 - (通常ありうる)他の処理によるCPU負荷増大時に、ピアダウンが発生しないこと



5秒/15秒で100ピア安定運用できること

試験項目

- 1秒/3秒に設定したピアを100ピア張る
- 定常時のCPU負荷を測定
- 他の処理に影響が発生しないか？
- 長期(数日間)に渡ってピアダウンしないか？

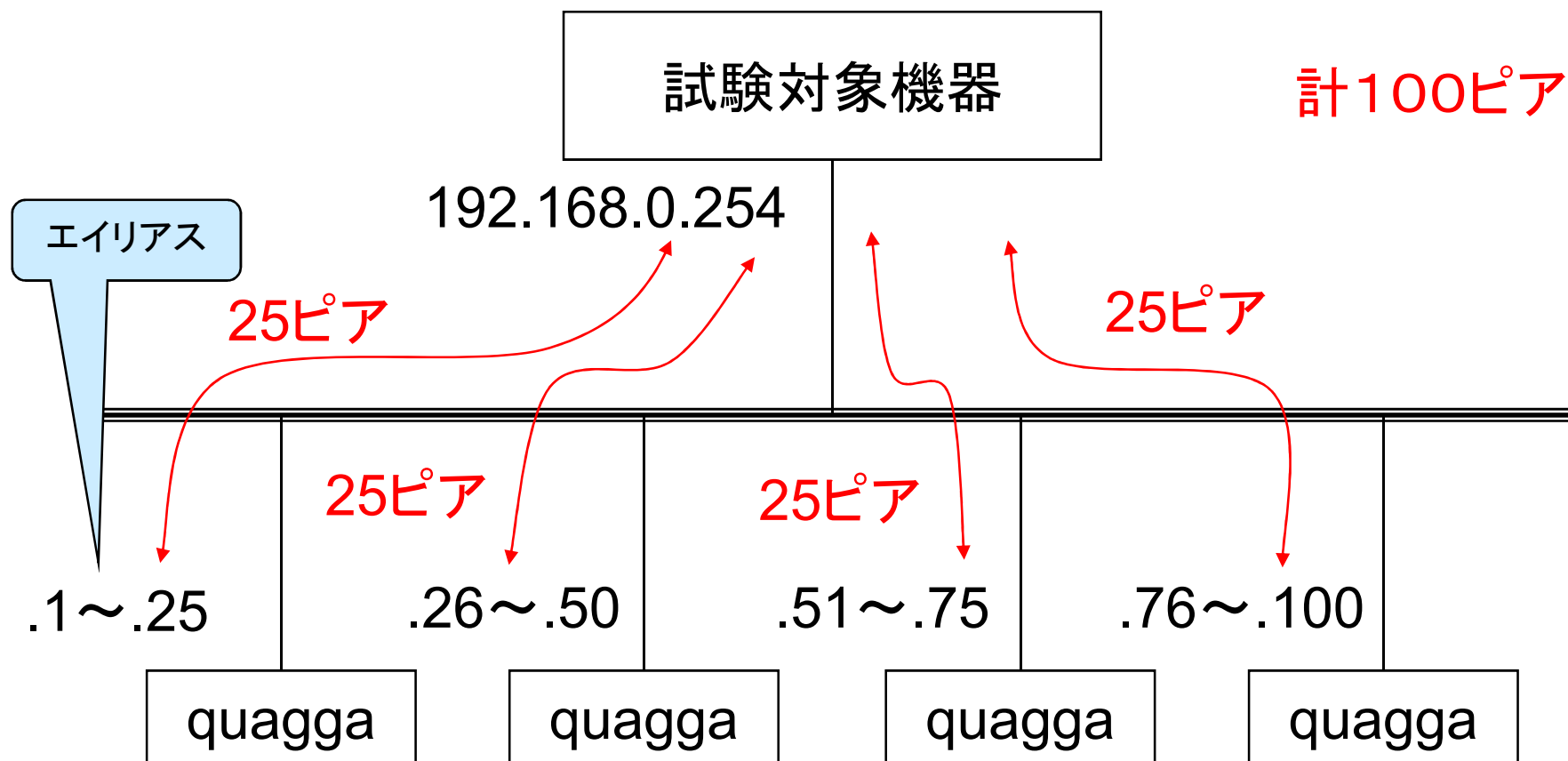


CPU負荷増大時にピアダウンが発生しないこと

試験項目

- 100ピア張っている最中に新規にBESTになるフルルートを受信する
- 転送時にTTLが0となるパケットを大量に送付
- ルータ宛てDoSアタック (ICMP flood、UDP flood、TCP syn attackなど)
- ARP未解決な宛先のパケットを大量に送付
- ルータに対してsnmpwalk
- フルルートを受信しているピアのroute-mapを書き換えて
clear ip bgp nei xxx soft in
- いくつかの負荷のかかるshowコマンド
 - show ip bgp route detail (pager offの状態)
 - show tech

試験構成



1台につき、25プロセスずつbgpdを起動

Quaggaのbgpdをたくさん起動する

- 通常、bgpdは179番ポートでwildcard bindしてしまう。
 - 1つのサーバで1プロセスしか動作できない。
- 179番ポートを変更するのは厳しい
- 指定IPアドレスのみで待ち受けるように、パッチをあてました。(-Iオプション)
 - 起動例
bgpd -d -I 192.168.0.1 -A 192.168.0.1 -i
/var/run/quagga/192.168.0.1.pid -f bgpd-1.conf

Quaggaのbgpdをたくさん起動する

```
quagga@test15 ~% ps xww
```

PID	TT	STAT	TIME	COMMAND
62687	??	Ss	0:17.51	/usr/local/sbin/bgpd -l 192.168.0.1 -A 192.168.0.1
62689	??	Ss	0:17.70	/usr/local/sbin/bgpd -l 192.168.0.2 -A 192.168.0.2
62691	??	Ss	0:16.86	/usr/local/sbin/bgpd -l 192.168.0.3 -A 192.168.0.3
62693	??	Ss	0:16.26	/usr/local/sbin/bgpd -l 192.168.0.4 -A 192.168.0.4
62695	??	Ss	0:15.86	/usr/local/sbin/bgpd -l 192.168.0.5 -A 192.168.0.5
62697	??	Ss	0:18.00	/usr/local/sbin/bgpd -l 192.168.0.6 -A 192.168.0.6
62699	??	Ss	0:17.02	/usr/local/sbin/bgpd -l 192.168.0.7 -A 192.168.0.7
62701	??	Ss	0:16.45	/usr/local/sbin/bgpd -l 192.168.0.8 -A 192.168.0.8
62703	??	Ss	0:16.07	/usr/local/sbin/bgpd -l 192.168.0.9 -A 192.168.0.9

<snip>



試験結果

- (Foundry RXシリーズ) 1秒/3秒 100ピアでCPU負荷は1%未満
 - 5秒/15秒 100ピアでは問題ないだろう
- (某国産ルータ) 1秒/3秒でピアダウンしやすい
sho ip bgp -F とかでピアダウン発生
5秒/15秒だと問題ない
- マルチスレッドになっていないBGPデーモンは不安定
 - Quaggaなど
 - BGPルーティングの処理中にKeepAliveの処理ができないと、相手側でHoldtime Expiredが発生する
- Holdtime短縮しなくても、179番ポートへのSyn Attackには弱い
 - 別途パケットフィルタで対応



実環境で短縮設定を実施

- 影響の少なそうなところから順次1週間おきに実施
- 2006/10/11 JPIX大阪
- 2006/10/18 DIX-IE
- 2006/10/25 NSPIXP3
- 2006/11/01 JPIX東京
- 2006/11/08 JPNAP東京
- ご協力ありがとうございました m(__)m



設定方法(Foundry RXシリーズ)

•設定

```
router bgp
  neighbor xx.xx.xx.xx timers keep-alive 5 hold-time 15
```

```
clear ip bgp nei xx.xx.xx.xx
```

※ KeepaliveとHoldtimeを個別に指定する

•確認

```
#sho ip bgp nei xx.xx.xx.xx
1  IP Address: xx.xx.xx.xx, Remote AS: xxx (EBGP)
   Description: xxxxxxxx
   State: ESTABLISHED, Time: 227d3h21m6s, KeepAliveTime: 5, HoldTime: 15
      KeepAliveTimer Expire in 3 seconds, HoldTimer Expire in 13 seconds
```



設定方法(アラクサラAXシリーズ)

•設定

```
xxxx(config)# show
bgp yes
externalpeeras xxx
peer xx.xx.xx.xx
ipv4-uni
refresh
holdtime 15
```

※ 設定はHoldtimeのみ。
Keepaliveは自動的に1/3の秒数に
設定される。

•確認

```
xxxx# sho ip bgp neighbors xx.xx.xx.xx
BGP Peer: xx.xx.xx.xx , Remote AS: xxx
Remote Router ID: xx.xx.xx.xx
BGP Status: Established HoldTime: 15
Last Keep Alive Sent: 11:16:18 Last Keep Alive Received: 11:16:20
```



ピアがあがらない;;

- 短縮設定後にピアが復旧しない。。。
全部、Juniperでした。
- JUNOS 8.2以降??
20秒未満のHoldtimeを受け付けない。
下記のエラーでピアアップしません。

Notification Message (3), length: 21, OPEN Message Error (2),
subcode Unacceptable Hold Time (6)

- ピア先でメンテナンス後に復旧しない
 - JUNOSのバージョンアップ
 - Cisco→Juniperへのリプレイス
- タイマーを10秒/30秒に変更して対応



JUNOSの隠しコマンド

- 隠しコマンドを入力することで、20秒未満の Holdtime動作が可能らしいですが、、、

```
neighbor x.x.x.x {  
    minimum-hold-time 15;
```

- 実環境で使うのは、やはり厳しいようで、、、



微妙におかしい

- タイマーが5秒/30秒になる？

- ルータのバグか・・・？ クリアすると正常に(10秒/30秒)

```
#sho ip bgp nei xx.xx.xx.xx
```

```
1 IP Address: xx.xx.xx.xx, Remote AS: xxxx (EBGP)
```

```
State: ESTABLISHED, Time: 150d1h34m13s,
```

```
KeepAliveTime: 5, HoldTime: 30
```

```
KeepAliveTimer Expire in 4 seconds, HoldTimer Expire in 29 seconds
```

- 5秒/15秒で設定し、こちらは5秒/15秒と表示されるが、
対向では6秒/20秒と表示されるらしい。

- 対向がJuniper

```
Peer ID: xx.xx.xx.xx Local ID: xx.xx.xx.xx Active Holdtime: 20
```

```
Keepalive Interval: 6 Peer index: xx
```

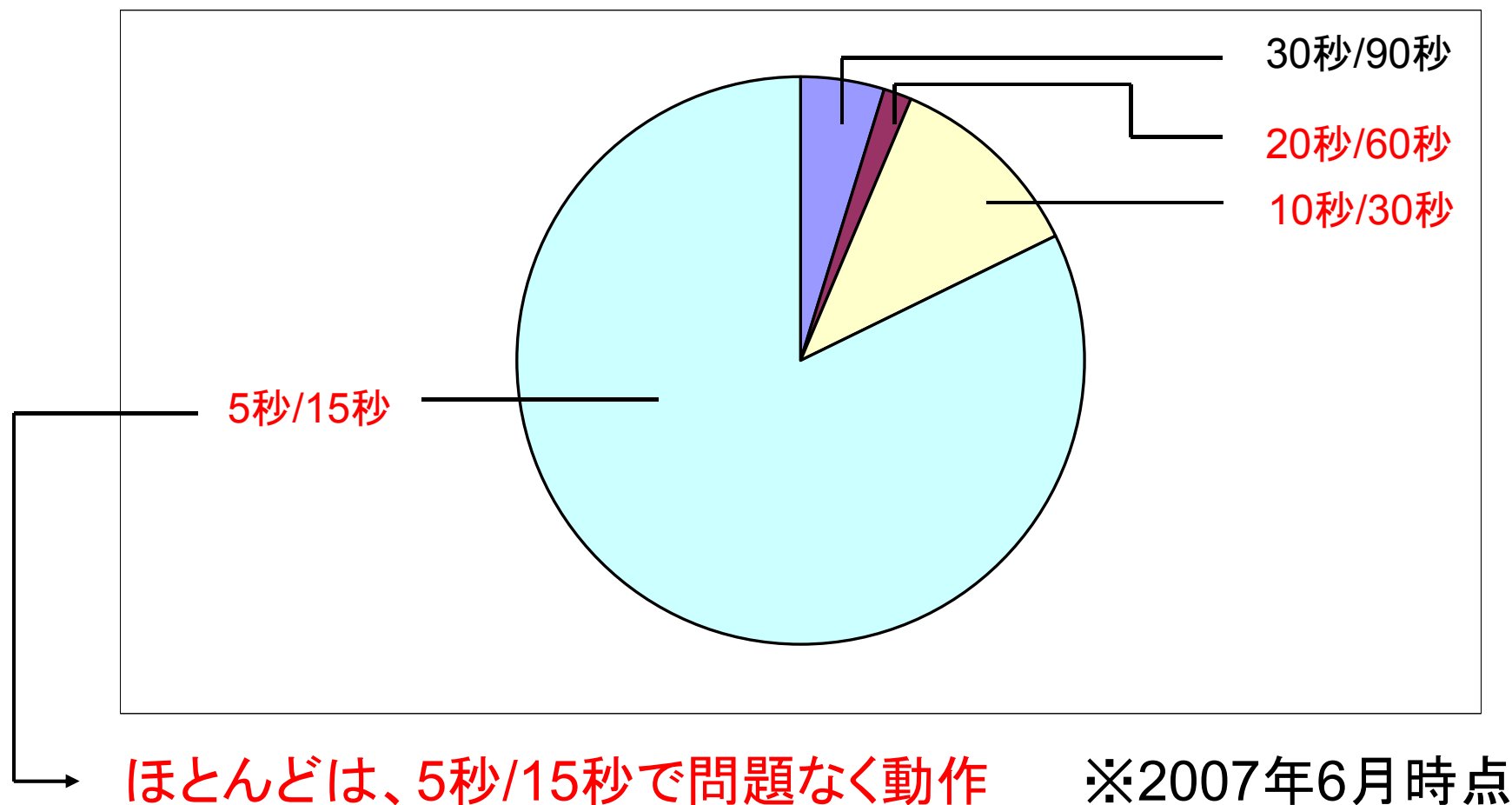


なんとなく落ちやすくなったピア

- 最初は5秒/15秒でスタート
- 原因不明の(メンテナンス等は除く) Holdtime expiredにて生じたピアダウンが3回以上
→10秒/30秒に変更
- それでもだめな場合
→20秒/60秒に変更
- それでもだめな場合
→30秒/90秒に変更
- ※タイマー値変更の場合は、明示的にクリアせず、次回ダウンしたときに反映されるように。

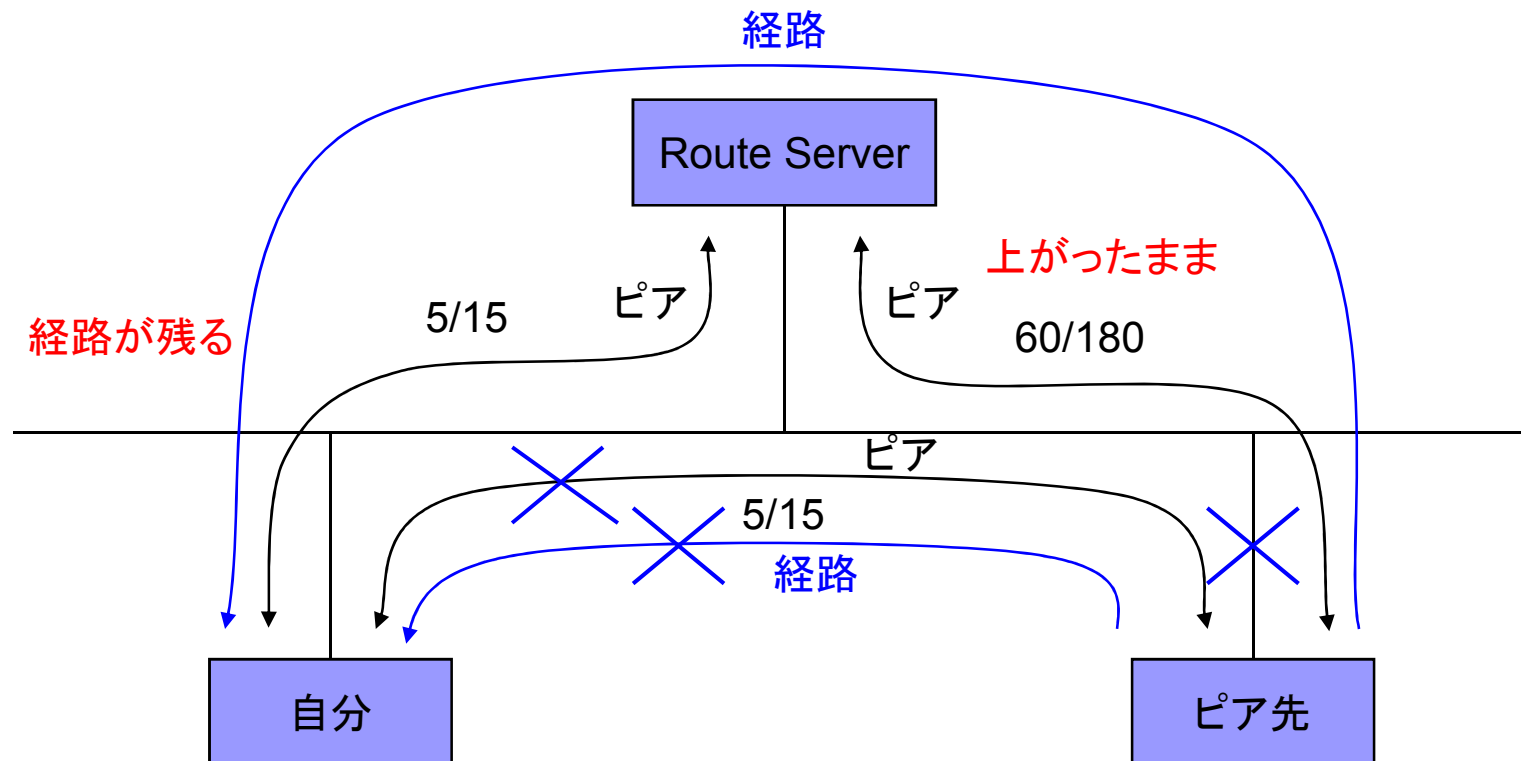
現在のHoldtime設定状況

- 弊社のPublicピアの状況を調べてみました。



某商用IXのRoute Server

- ピア先とRS間のHoldtimeが長いと、直接ピアを短縮しても意味がない。





まとめ

- 短縮設定には、それなりの手間がかかる;;
- 通信断→eBGPピアダウン→iBGPの収束→エンド
ホスト間の通信の復旧の時間は現状未検証
- 設定後、InterASにおける障害で問い合わせがきた
ことは、今のところない。
- 皆さんも是非！！
 - ただし、、、
 - Holdtimeは小さい方で動作します。
 - 変更しても良いか、ピア先に了承を得る必要があります。



JANOG20の宣伝 ☺

- JANOG20のプログラム委員やっています。
- JANOG20 2日目(7月12日) 11:55～12:40
- 「高速な障害復旧に必要な思いやり」
林 秀明さん (古河ネットワークソリューション株式会社)
松嶋 聡さん (ソフトバンクテレコム株式会社)
- BFDではない高速切替手法の発表です。
- ぜひ、北海道帯広へ！