ネットワークにおける
高速切り替え手法の検討

今までのホールドタイムでいいんでしょうか？

鈴木昭徳
NTTコミュニケーションズ
JANOG17

この発表は，独立行政法人通信情報研究機構『インターネット中枢機能のセキュティ強化に関する研究開発』の一環として行われました。

最近のインターネットって

- リアルタイムトラフィック
  - 音声（Skypeとか・・・）
    - Skypeの企業での使用
    - 専用ハンドセットの販売
    - グループウェアとの連携
  - IPTV
    - GyaOさん600万登録者の達成
    - コンテンツ配信についての法整備

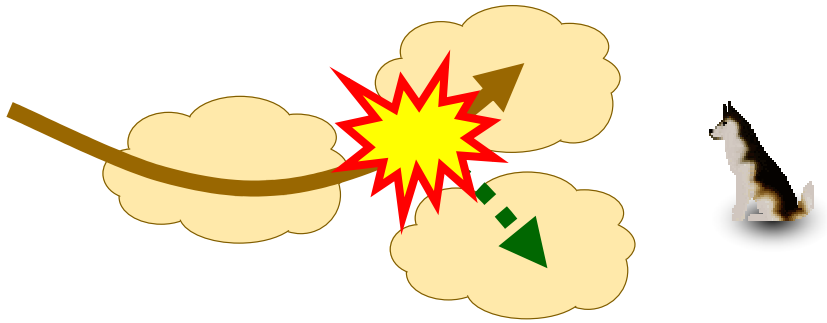
サイボウズOffice，
アリエル・プロジェクトA，
ライブドア・サイバーフォン
とか
ビジネスモデルが出来上が
りつつあり，関係団体も意
識が変わってきてる。

- ミッションクリティカルなサービス
  - ネット証券
  - 銀行サービス
  - ネットShopping（年末商戦時とか）

1秒の切断も許さ
れない！

インフラとしての安心感？

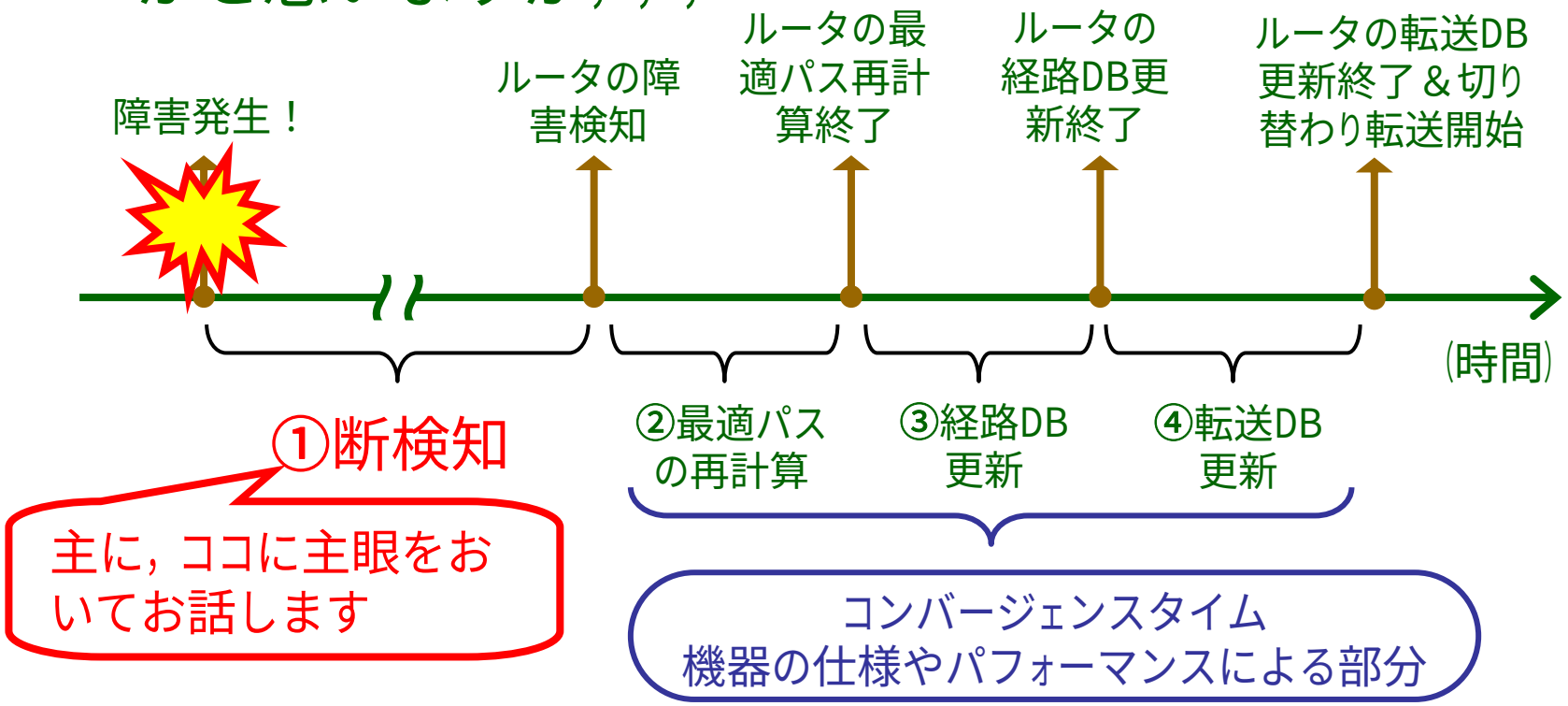
- インターネット独自の要素
  - たくさんの通信事業者 (ISP) の寄せ集め
  - DOSアタックなどの脅威
  - *故障時に早く・ちゃんと切り替わる？*
  - とかとか。。。



今回は、この “切り替え” について，考えてみました。

ルータの通信の切り替え過程

- ハードウェアの実装や，仕様によって若干違いはあるかと思いますが，，，



インターフェースでリンクダウンを検知すれば，
①は，一瞬ですが，そうでないときもありますよね。

リンクダウンが検知できない時

- ルータ間にEtherスイッチがある時
  - AS間・・・　IXに接続してのBGP接続
  - AS内・・・　トラフィック分散の為に使用
    リンクダウン転送機能なしのMC

そういうときには，通信断を検知できないの？

IGPや，BGPには”正常性確認機構”
が実装されていますよね。

HellotoかKeepaliveとかって，Default運用？

- リンクダウンを検知できない時は，　Helloや，Keepaliveのホールドタイムを失効（Expire）するまで，障害を認識できません。
  - AS間・・・　BGPのKeepalive
  - AS内・・・　IGPのHello

でも、結局ルータのDefaultのまま使ってませんか？（たぶん結構多いのでは？）
BGPピアに関しては、対向ルータの機種がなんとなく推測（※）出来たりとか。。。
C社ルータ・・・180秒，
J社ルータ・・・９０秒

（※）Keepaliveのタイマーネゴシエーション

出来るだけ短くしてみよう

今回は、BGPを例にとって考えてみました。

- 色々ご意見おありかと思いますが。。。

- そもそも，どこまで短く出来るんだろう？

※BGPのKeepaliveに関して

- C社ルータ･･･　インターバル：　1秒，　ホールド：　3秒
- J社ルータ･･･　インターバル：　2秒，　ホールド：　6秒

両ルータとも「ホールドタイムを20秒以下にすると，Flapするからよくないよ」というAlertが出ましたが，設定できるんだから試してみました。

ここで、ちょっとおさらい

- BGPのKeepaliveインターバル、ホールドタイムは、値の短いほうにネゴシエーションされます。

インターバル：60秒
ホールドタイム：180秒
がいいなー。

インターバル：30秒
ホールドタイム：90秒
がいいなー。

じゃあ、
インターバル：30秒
ホールドタイム：90秒
にします。

インターバル：30秒
ホールドタイム：90秒
ボクは、そのまま。

※　インターバルがホールドタイムの1/3になってますが、
RFC1771では1/3が推奨されています。
J社ルータに関しては、ホールドタイムを明示的に設定するのみで、
インターバルタイムは、自動的に1/3の間隔で、送信されます。

そりゃ，ちゃんと動くけど，，，　あれ？

組み合わせで，最短時間での設定

1）　C社ルータ同士　　・・・　OK

2）　J社ルータ同士　　・・・　OK

3）　C社ルータとJ社ルータ　　・・・　あれ？ (次のスライドで)

※C社ルータの最小値・・・

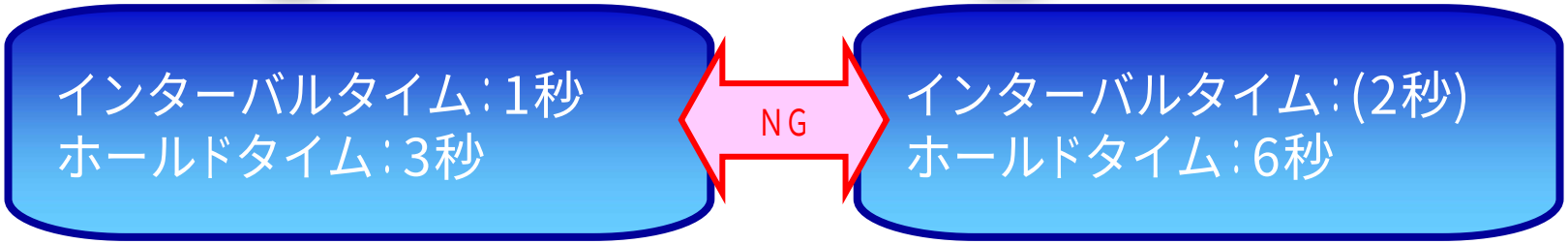
インターバルタイム：1秒，　ホールドタイム：3秒

※J社ルータの最小値・・・

インターバルタイム：2秒，　ホールドタイム：6秒

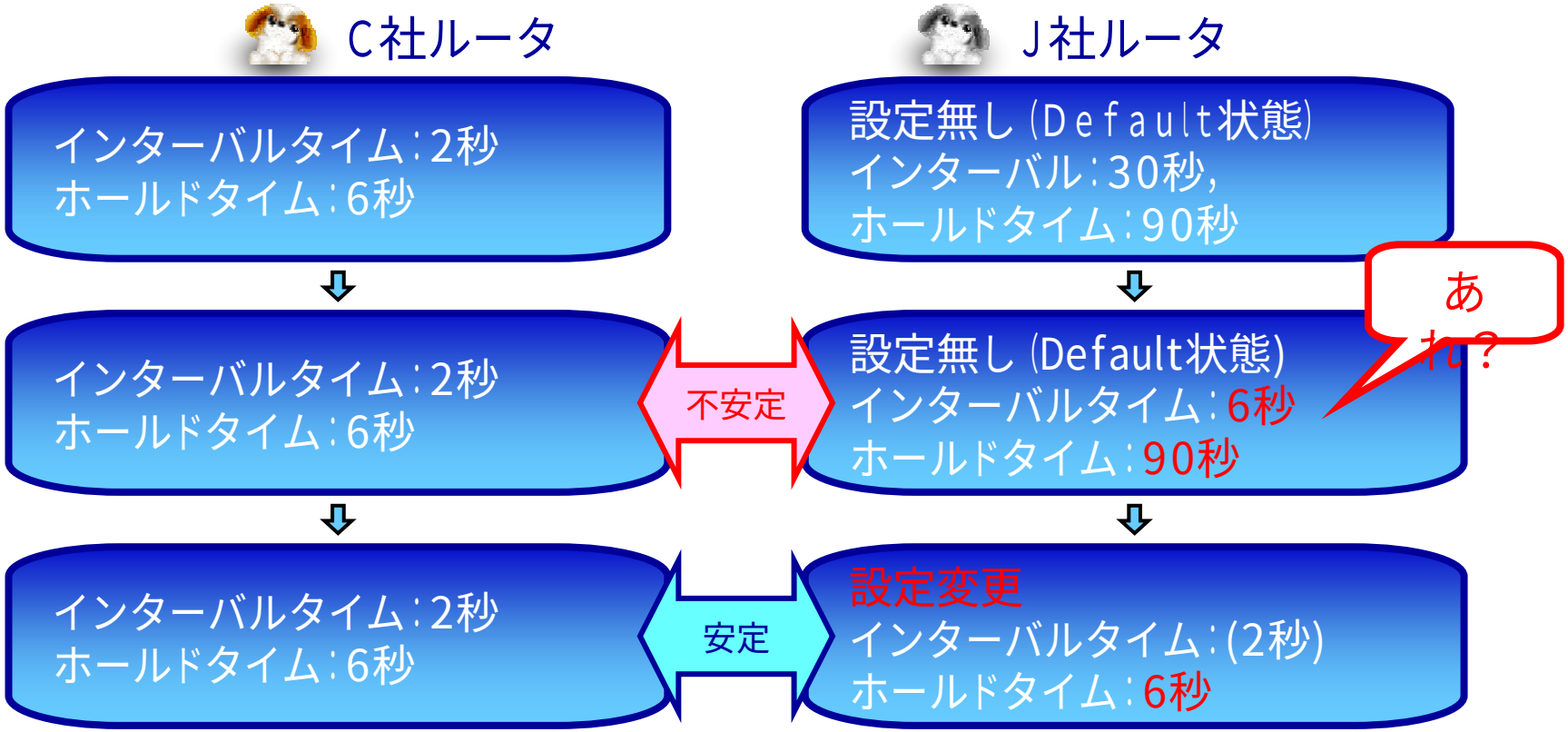
C社ルータ

J社ルータ



細かな所では，癖があるみたいです

- C社とJ社の最短時間での組み合わせ



J社ルータは，明示的に設定しないと、インターバルタイムが6秒以下にならないようです。

・・・意外と，癖がありました。

1秒以下の切り替わりって，できないの？

- “ＢＦＤ”ってのがあるみたい。。。
  (Bidirectional Forwarding Detection)
  draft-ietf-bfd-base-04.txt， draft-ietf-bfd-mpls-02.txt
  draft-ietf-bfd-v4v6-1hop-04.txt， draft-ietf-bfd-multihop-03.txt
  draft-ietf-bfd-mib-02.txt， draft-ietf-bfd-generic-01.txt
  - IP網内での障害を1秒以下で検知することのみを目指したプロトコルで，仕組みはIGPのHelloや，BGPのKeep aliveと同じ。しかし，それを高速に行う。BFDコントロールパケット(24bytes)という。
  - IPレイヤ上で動作。IGPやBGPのみならず，MPLSやGREやなどの導通確認を可能にすることを目指している。
  - C社製，J社製ルータにv4で，一部のプロトコルのみ実装済み。Asyncモードのみ。Demandモードは未実装。Echoファンクションも未実装。
  - インターバルタイムのネゴシエーションは，値の高いほうを選択する。

BFDの設定パラメータ

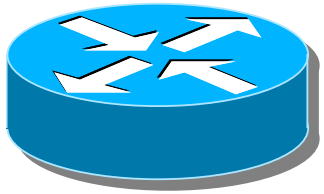
- 3つの設定パラメータがあります。

1) min-TX-interval
   最小の送信を許容するインターバルタイム (msec)
2) min-RX-interval
   最小の受信を許容するインターバルタイム (msec)
3) multiplier(ホールドタイム)
   受信するインターバルタイムに対する乗数 (整数値)

対向とのネゴシエーション

ホールドタイム

対向のTX-interval(ネゴ後)と，
自身のmultiplier値の乗算



送信インターバル

自身のmin-TX-intervalと，
対向のmin-RX-intervalとで，
比較して，値の大きいほう

BFDの設定　C社

インターフェースのところで設定

```
interface GigabitEthernet 1/0
 ip address x.x.x.x y.y.y.y
 bfd interval 50 min_rx 50 multiplier 3

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

router ospf zzz
 network x.x.x.x y.y.y.y area 0
 bfd all interfaces
```

BGPの場合

OSPFの場合

BFDの設定　J社

```
[protocols ospf area 0]

interface ge-1/0/0.0 {
    bfd-liveness-detection {
        minimum-interval 50;
        minimum-receive-interval 100;
        minimum-transmit-interval 200;
        multiplier 5;
    }
}
```

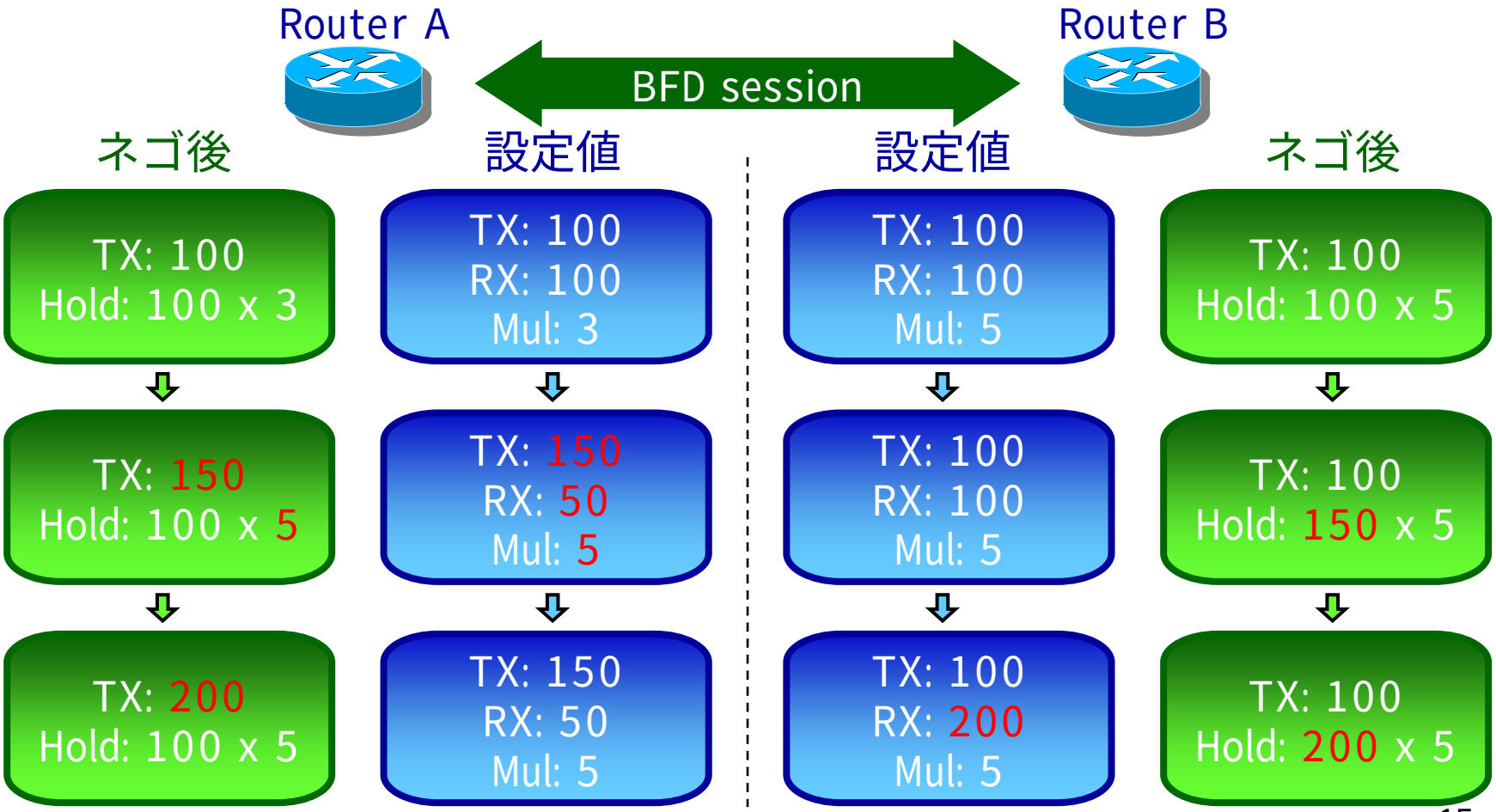
プロトコルズのところで設定

送信，受信を一括で設定する場合

送信，受信を別々に設定する場合

BFDのタイマーネゴシエーションについて

- BFDのタイマーネゴシエーションは，値の高いほう
  Flapへの心配や、機器のパフォーマンスを考慮した設計？

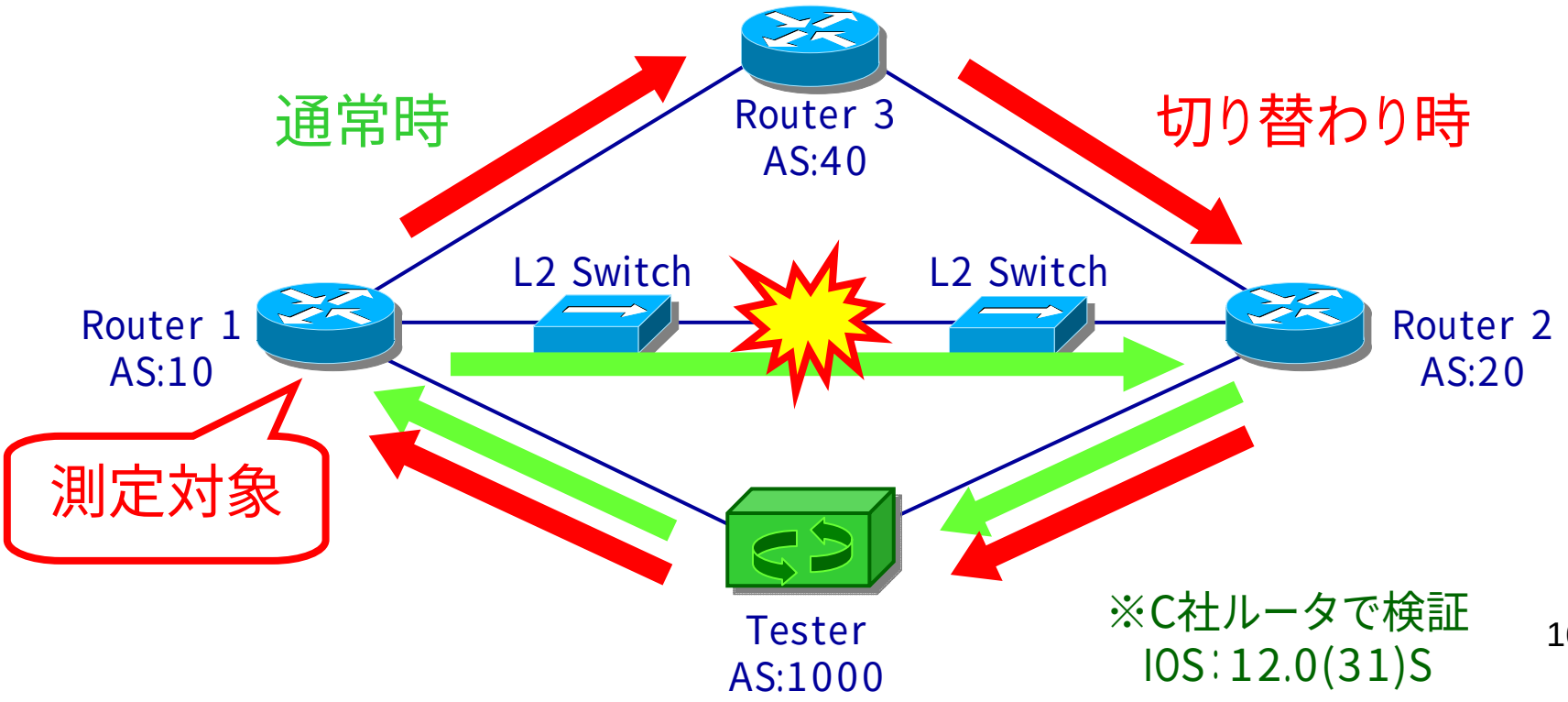


※C社ルータで確認。

切り替わり時間の測定

”Keepaliveを最短にした場合”と，“BFDを用いた場合”について，検証してみました。

- 検証環境
  - パケットロス値により切り替わり時間を測定
  - コンバージェンス時間の影響を少なくするために，必要な経路のみ広告



※C社ルータで検証
IOS：12.0(31)S

測定結果 ～BFDの場合～

- 測定結果
  - 10msec間隔/3回でも，問題なく動作する。
  - 多少ばらつきあり。ルータの精度？ コンバージェンスタイムの影響？
  - インターバルタイムが大きいと，障害を発生するタイミングが影響。

multiplier値	3回		5回		10回	
Interval時間(m秒)	切り替わり時間(秒)					
	実測値	平均値	実測値	平均値	実測値	平均値
10	0.923	0.544	1.099	0.694	0.262	0.460
	0.318		0.534		0.816	
	0.392		0.449		0.302	
50	0.923	1.039	0.908	0.717	1.001	0.820
	1.191		0.568		0.697	
	1.003		0.674		0.763	
100	0.353	0.843	0.856	0.907	1.761	1.741
	1.248		0.933		1.590	
	0.928		0.933		1.873	
500	1.079	1.020	3.305	2.854	5.426	5.053
	1.267		2.709		4.526	
	0.713		2.547		5.206	
1000	3.016	2.869	5.428	5.054	8.702	9.643
	2.318		5.299		10.693	
	3.272		4.434		9.534	

測定結果　～BFDなし～

BGPのKeepaliveの最小値にて測定しました。

- 測定結果
  - BGPのKeepaliveを短くしても，不安定になることなく，問題なく動作。
  - インターバルタイムを大きいと，障害を発生するタイミングが影響。

Interval時間(秒)	切り替わり時間(秒)	
	実測値	平均値
BGP Inter：1秒 Hold：3秒	2.872	2.887
	3.022	
	2.766	

その他の検証

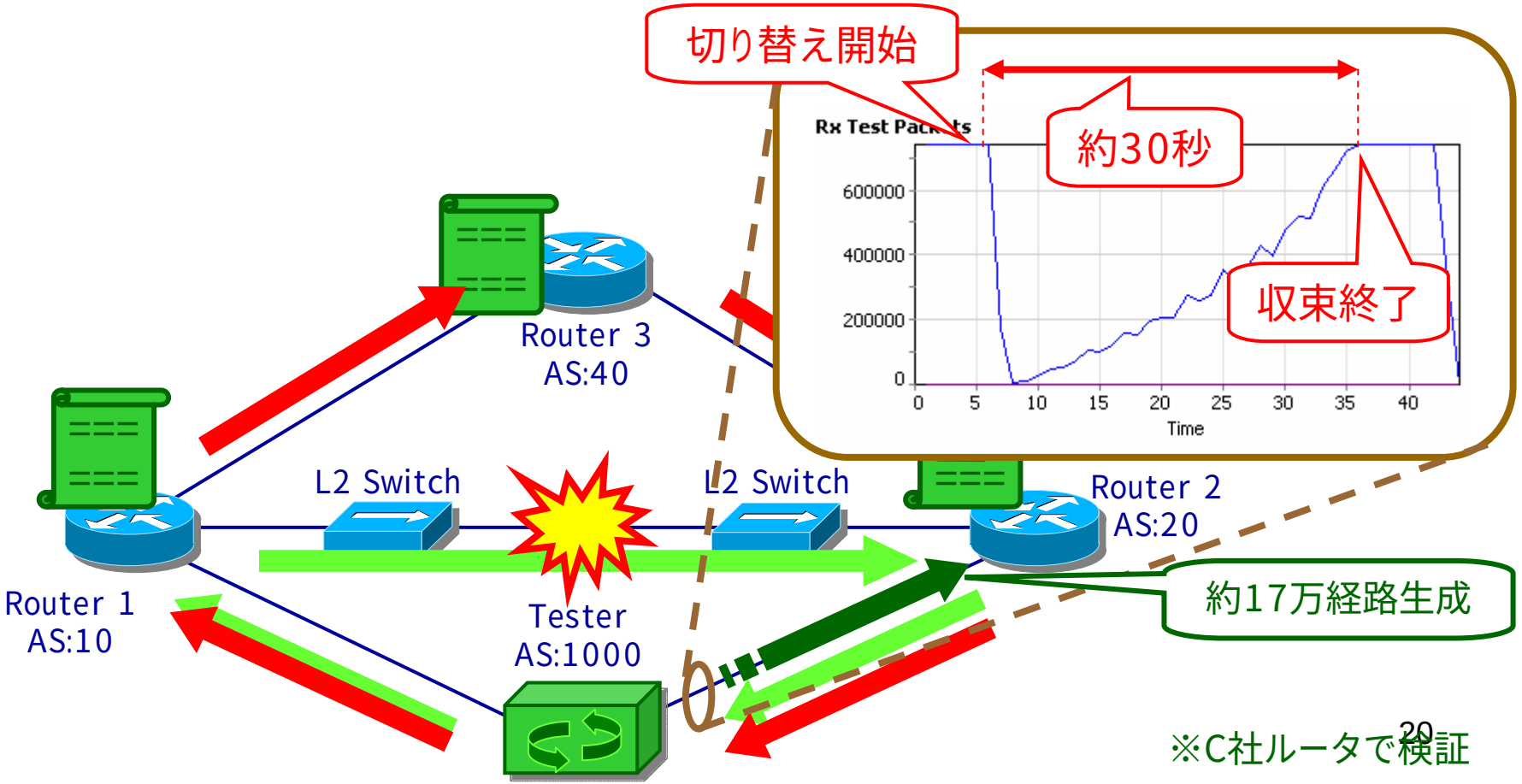
1） 経路数を増やして，コンバージェンスタイムの計測
- 1-1 Full-routeの更新を同時に実行

2） C社ルータとJ社ルータの相互接続
- 2-1 BFD on OSPF

3） リアルタイム系アプリケーションを流してみました
- 3-1 音声
- 3-2 IPTV

1-1　Full-routeの更新を同時に実行

- 動作的には，ちゃんと動きますね。
  - “切り替わり時間” ＝ “Keepaliveや，BFDでの障害検知時間” ＋”Full-routeのコンバージェンス時間”
  - コンバージェンスに約30秒。機器のパフォーマンスによりますね。



※C社ルータで検証

2-1　BFD on OSPF　C社とJ社ルータで

J社ルータは，BFD on BGPが未サポートなので，OSPFでインターオペイラビリティを確認しました。

- あっけないくらいに，ちゃんと動きますね。
  - タイマーネゴシエーションも正常に動作。
  - 大規模なネットワークでは、未検証。

- サポート状況

C社

- EIGRP
- OSPF
- BGP
- IS-IS

※　いずれもv4のみ


J社

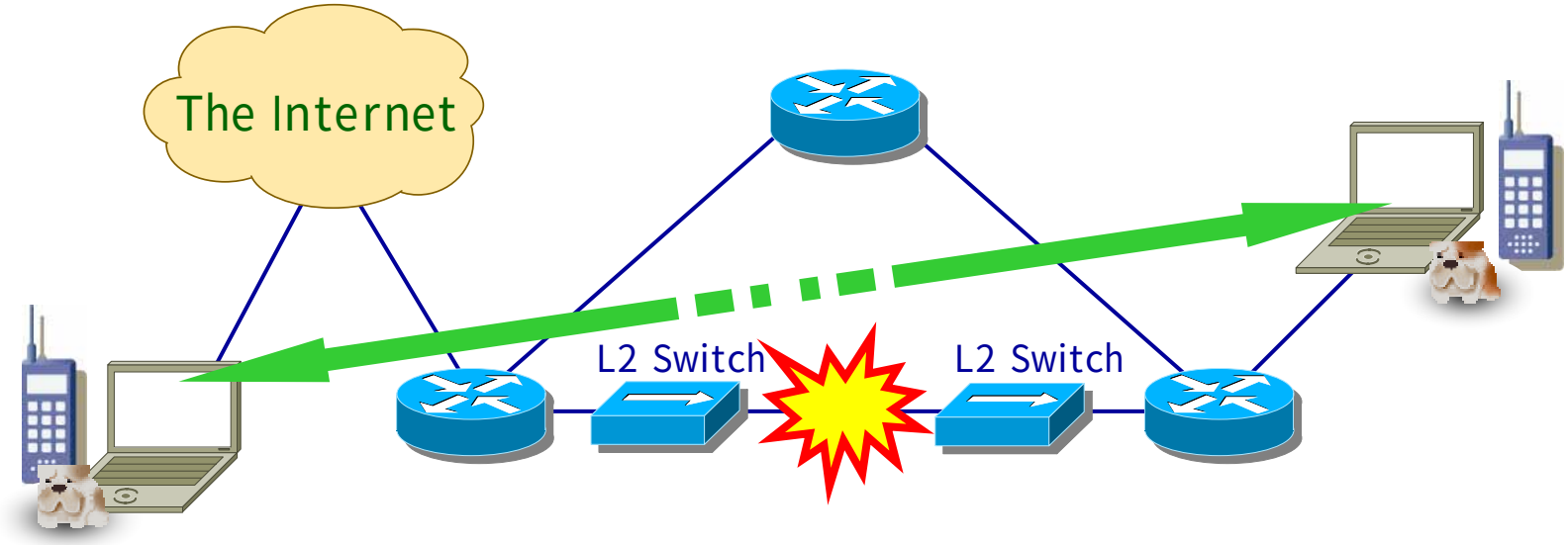
- OSPF
- IS-IS
- Static

※　いずれもv4のみ

3-1 リアルタイム系アプリケーション 音声

Skypeでの通話中に切替えてみました。

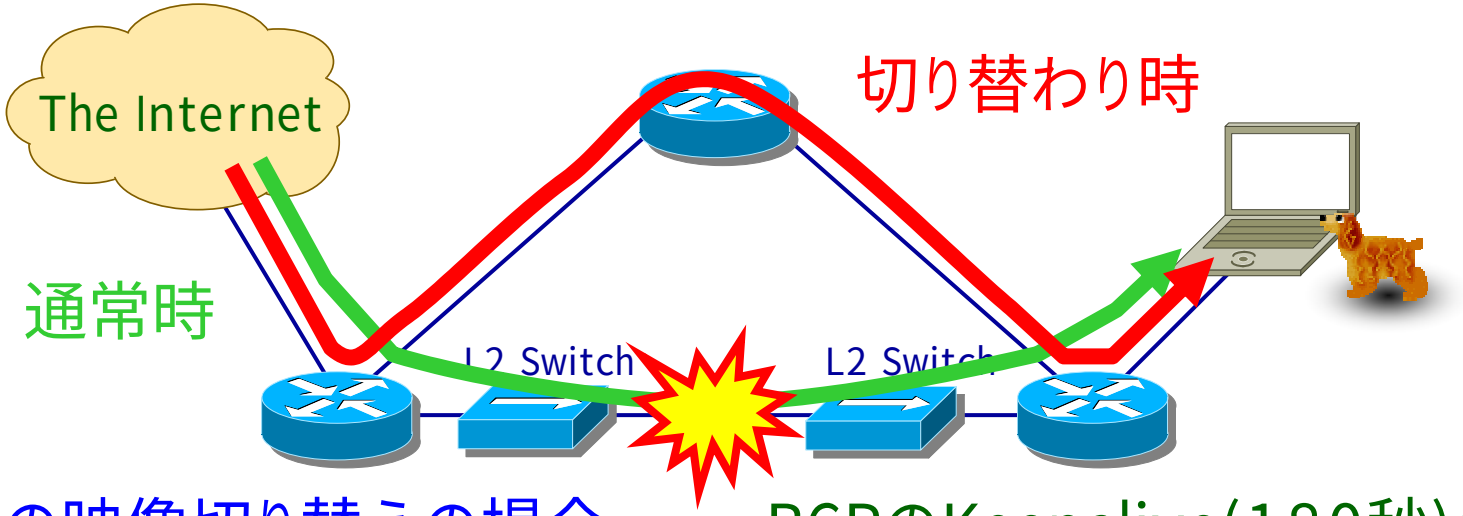
- 切り替わりに要する時間分，当然音声は切れます。ですが、BFDを使用した場合は，1秒程度の断で，意思疎通に問題はありませんでした。



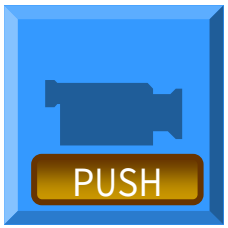
3-1 リアルタイム系アプリケーション IPTV

IPTV視聴中に切替えてみました。

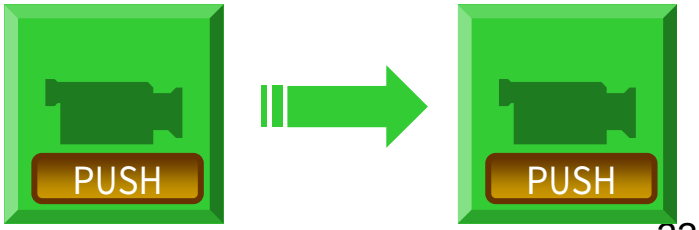
- BFDでの切り替えは一瞬。トラフィック転送は継続される。
- また，予想以上にIPTVのアプリケーションのつくり込みが良かったです。・・・なかなか，切れません。



BFDでの映像切り替えの場合



BGPのKeepalive(180秒)の場合



まとめ

- BFDの場合の切り替えも，Keepalive値を最小にした場合の切り替えも，Lab内で実機での正常な動作を確認できました。
- 切り替え時間の短縮化の効果は，使用するアプリケーションによって様々だが，少なくとも数十秒ー百数十秒の障害検知時間より，サービス性向上を図ることができると考えられる。
- 適用範囲としては，AS内ではEtherスイッチを適用している箇所や，AS間ではIXでのBGP接続のような場面で，ホールドタイムの見直しや，BFDの適用が考えられるのでは。

ずっと，Defaultのままで運用しますか？

また，私どもの所では，実網を用い，IX環境下で，引き続き実験を行っていこうとしています。
ご興味がある方は，是非お声をかけてください。