高速切替手法の検討

(BFD for BGP編)

鈴木昭徳
NTTコミュニケーションズ

本発表は、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。

JANOG17での発表において

● インターネットの使われ方の変化
- リアルタイム系トラフィック
  例)　甲子園、競馬、ゴルフ中継のストリーミング配信
  Skypeの広がり、専用ハンドセット
- 生活インフラとしての使われ方
  例)　金融サービス（銀行・株式投資など）
  ネット通販

変わるもの

変わらないもの

● 「ルーティングプロトコルの断検知は長いよね」(リンクダウンしない場合)
- BGPのデフォルト90-180秒　…変わってないよね?
- OSPFのデフォルト40秒　…良いの?

● BFDに注目し、ラボ環境にて検証
- 断検知時間の測定(BFD/Keepaliveを短く)
- 実アプリケーションへの影響(音声、VOD)
- 切替による影響を軽減

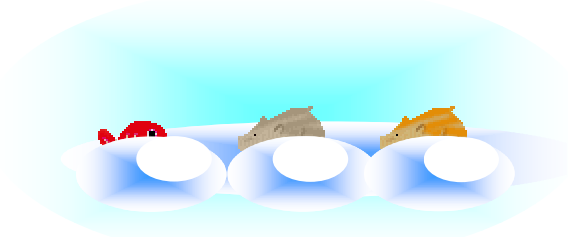
デフォルトのままで、
運用していきますか?

JANOG19での発表は？

- **C社ルータだけでなく、J社も**
  - BFD for BGPの相互接続など
  - Keepaliveインターバルを出来るだけ短くしてみよう

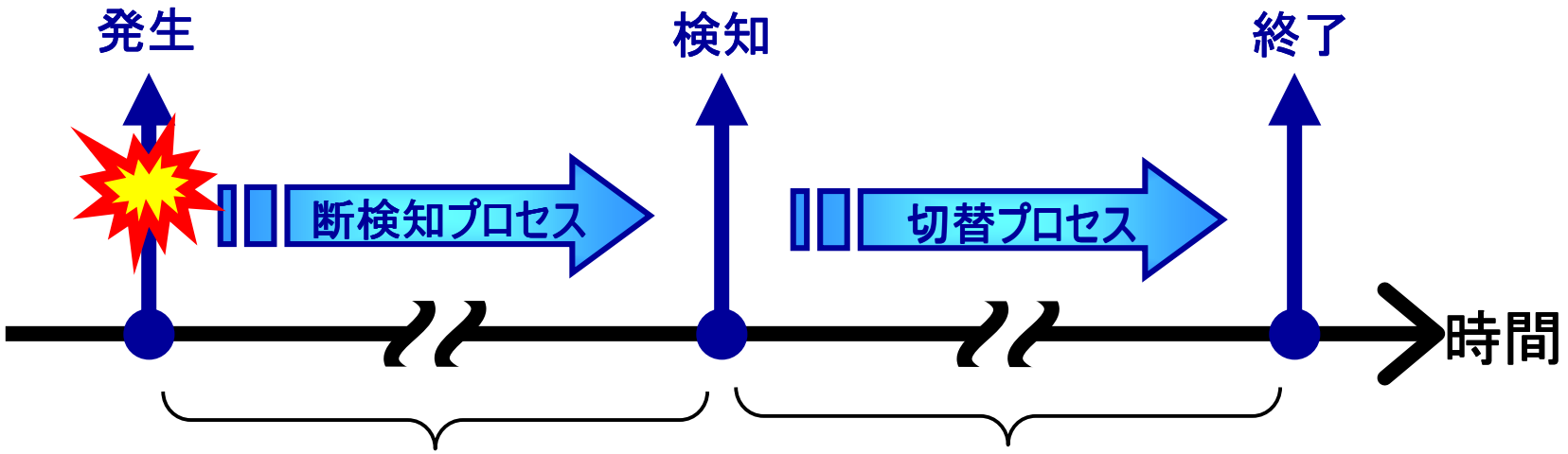
- **実環境下での検証**
  - 商用IXで、且つグローバルASで
  - リアルタイム映像の切替

- **その他**
  - Keepaliveのデフォルト運用の実態
  - 色々わかった事



障害発生から切替終了までの流れ

● 障害発生から切替終了までの流れ(概要)を整理したい



断検知時間

- リンクダウン
- 警報受信
- ホールドタイムの失効
- ガードタイム

短い

長い

切替時間(装置内部処理)

- 検索
- 再計算
- 更新
- 受信端/送信端切替

装置の方式・実装に依存

断検知を早くしたい流れ及び検知

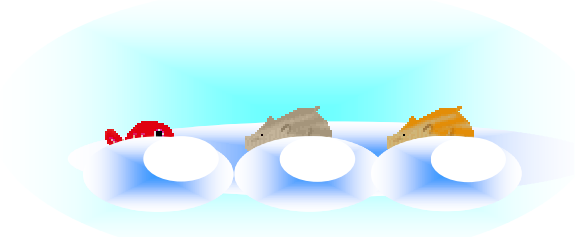
新しく考えられている方式・機能について触れたい

● **BFDの機器への実装**
- C社J社ルータへの実装(OSPF、ISIS、1-hop BGP、Static、RSVPなど)
- 一部で運用している方もいるらしい

今回ココに注目

● **MPLS**
- FFD(Fast Failure Detection) ITU-T Y.1711
- 既存CV(Connectivity Verification)は、秒単位の断検知

● **Ether-OAM(※障害の検知)**
- CC(Continuity Check) ITU-T Y.1731 ※そもそもOAMが無かった
- やっとVLAN単位で断検知が出来る



BFDとは

- 高速に**断検知**し、ルーティングプロトコルに**通知**するだけ
- **Hello**と**同じ**正常性 **Check**
- **BFD**パケットの**ミリsec単位**の送受信
- **IPレイヤ**上で動作
- 送受信間隔の交渉は、**高い値**を選択
- **C-plane/D-plane分離**の考え

インターバル/ホールドタイムを設定
Keepaliveと同様、Echoを返すモードもあるが、実機には未実装

だから、リンクフリープロトコルフリー

BGPは低い値を選択

次に触れます

BFDのC-plane/D-planeの分離（BGP編）

● BGPさん1人の場合（BFD無し）

C-plane＋D-plane（RE内）

Keepaliveを高速化すると、プロトコルさんが苦しいかも...?

● BGPさんとBFDさんの場合

C-plane（RE内）/D-plane（LC内）

C-plane/D-plane役割分担
1つの筐体の中であるが、機能を分離し、
プロセス処理の負荷を分離。

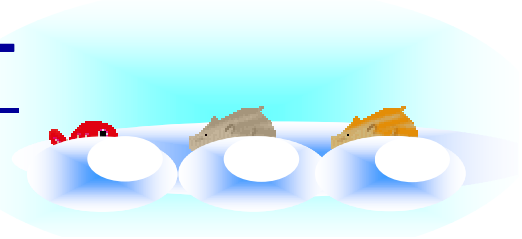
RE: Routing-Engine
LC: Line Card

特にBGPのKeepaliveは、応答を返さないといけないので、Keepaliveインターバルを短くすると落としたりするような話も...

あるISPさんのIXポイントでの Keepalive Hold-Time

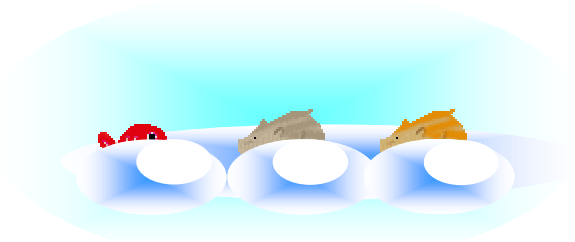
BGP インターバル/ホールドタイム（秒）	ピア数/比率（合計：188ピア） 2006年10月
60/180	142ピア/75.5%
30/90	43ピア/22.9%
13/40	1ピア/0.5%
5/15	1ピア/0.5%
0/0（Keepalive無し）	1ピア/0.5%

ほとんど皆さんDefaultです



断検知時間の検証

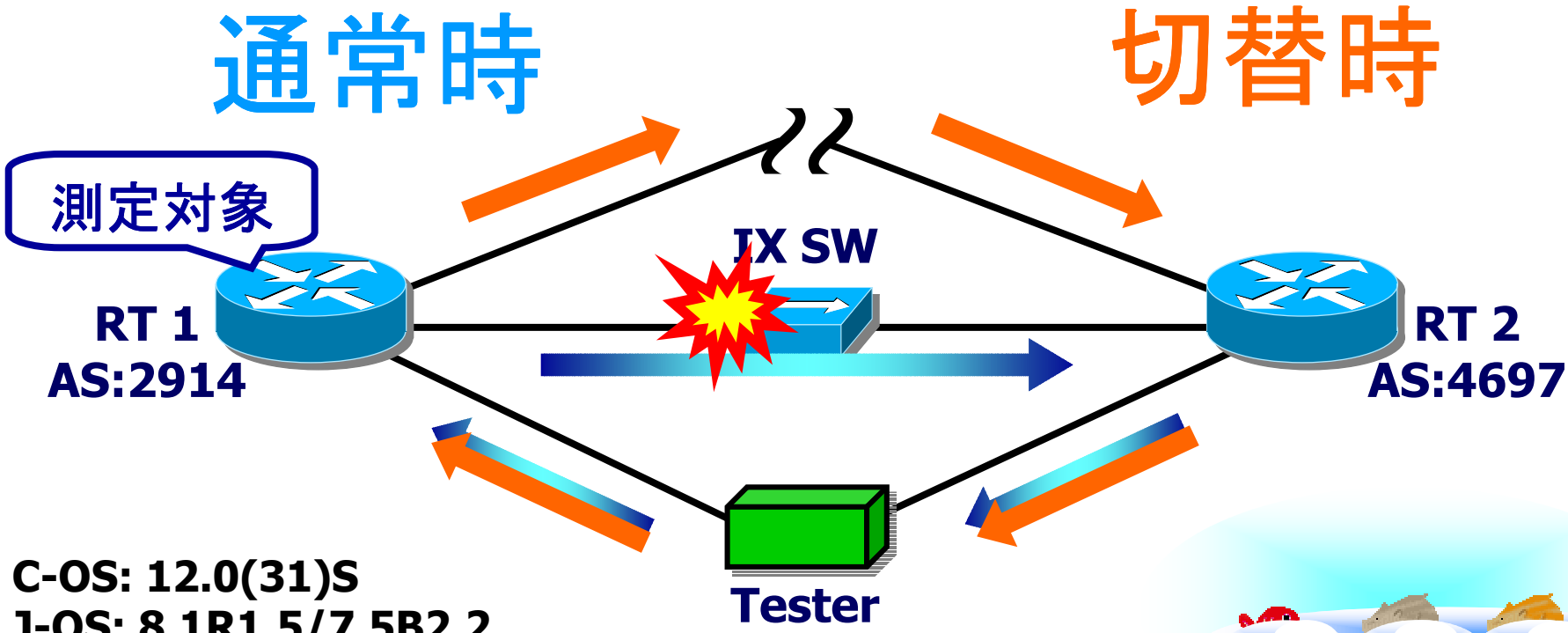
- C社ルータ、J社ルータでBFD for BGPによる計測
- C社ルータ、J社ルータでKeepaliveを短くして計測
- 実際の商用IXへ接続
- グローバルASを使用



実験環境（断検知時間の測定）

● 実際のIXで、グローバルASで検証しました。

“BFD for BGP”、”Keepalive Intervalを短く”
パケットロス値から断検知時間（≒断時間）を測定
コンバージェンスの影響を抑えるため、必要最低限の経路数



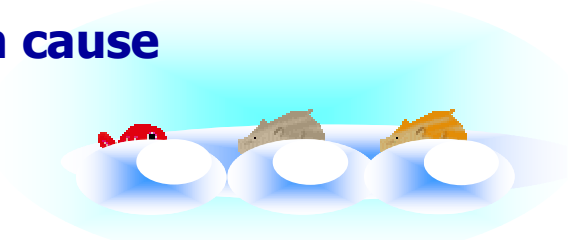
C-OS: 12.0(31)S
J-OS: 8.1R1.5/7.5B2.2

BFD for BGPの検証

- Cルータは、値のバラツキが見られた
- Jルータは、超高速だとBGPピアが不安定に
- C/Jの相互接続は、動作自体は正常且つ安定

Interval Time x Multiplier	Cisco（5回平均） Down Time (msec)	Juniper（5回平均） Down Time (msec)
10msec x 3	580.5	BGPピア不安定のため測定不可
20msec x 3	877.9	
30msec x 3	1022.3	
50msec x 3	727.3	158.4
100msec x 3	686.7	279.7
300msec x 3	687.4	749.5

● Junosマニュアルより
「Specifying an interval smaller than 300ms can cause undesired BFD flapping. 」



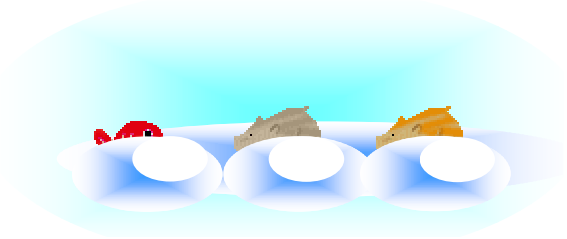
Keepaliveインターバルを短く

- 動作自体は安定（C./J.ともHold Time20秒以下はアラート）
- J.は、明示的に設定投入しないと、Hold Time20秒以下にネゴシエーションされない。（J-OS:7.5）
- J-OS: 8.1RではそもそもHold Time20秒以下に設定できず。

Keepalive Interval / Hold Time(sec)	Cisco（5回平均）	Juniper（5回平均）
	Down Time (sec)	Down Time (sec)
1 / 3 sec	4.58	設定不可
2 / 6 sec	7.66	5.38

※　厳密にはコンバージェンスタイムも含まれるため、切替時間自体は本質的ではない。機器のパフォーマンスによる。

C-OS: 12.0(31)S
J-OS: 7.5B2.2

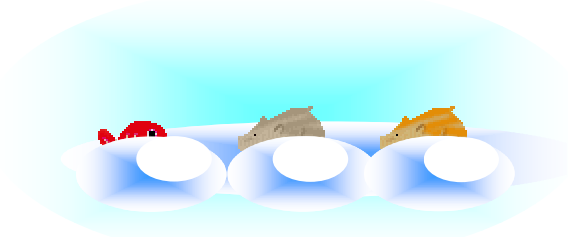


その他の検証

- フルルートのコンバージェンスタイムの計測
  断検知後の内部切替処理時間

- リアルタイム系音声
  Skype（PC、専用ハンドセット）

- リアルタイム系映像（**Frontiers社製　Hdx1000)**
  バッファ無しの**IP**伝送
  デモ映像

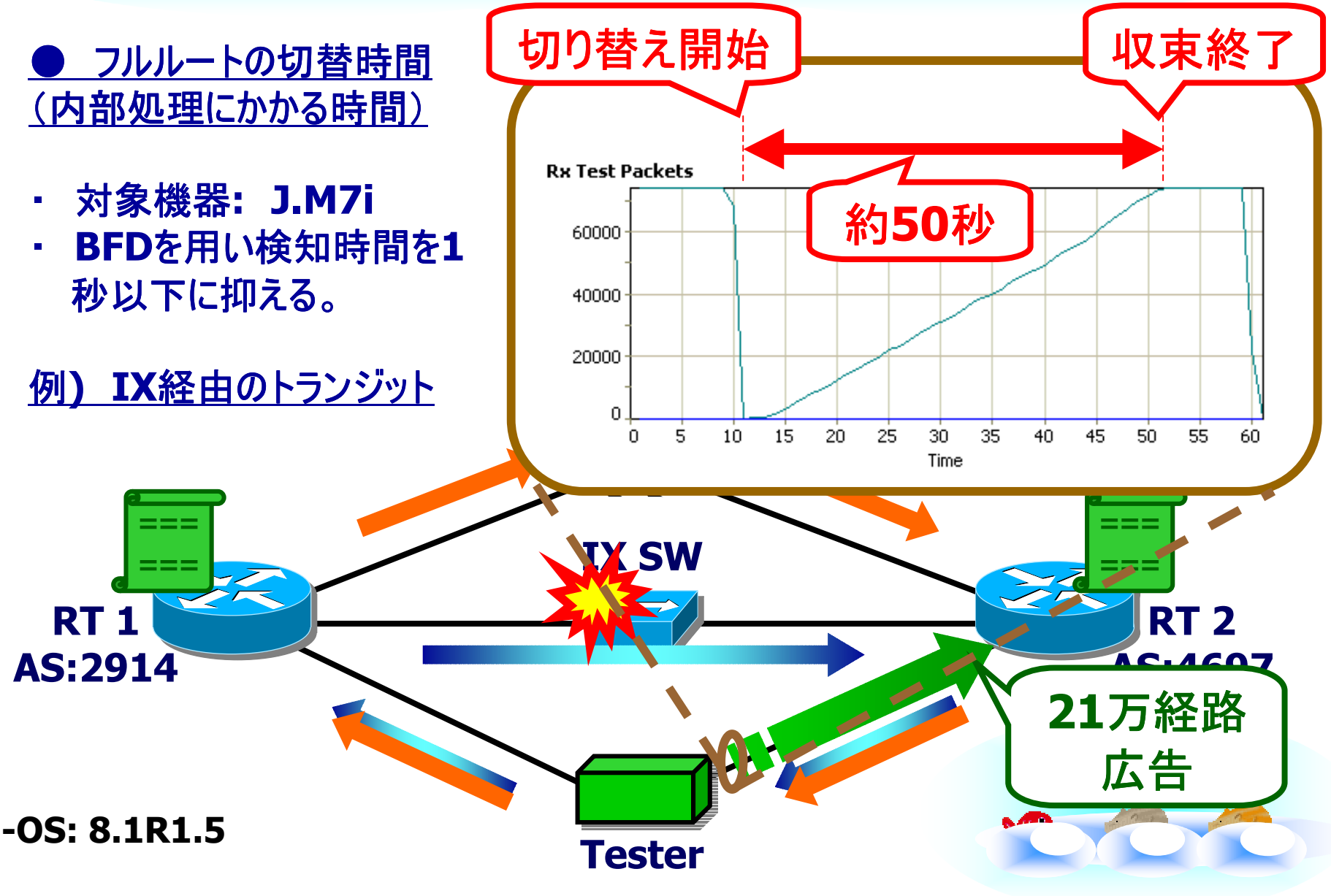


コンバージェンス時間の測定

● フルルートの切替時間
（内部処理にかかる時間）

- 対象機器: J.M7i
- BFDを用い検知時間を1秒以下に抑える。

例) IX経由のトランジット



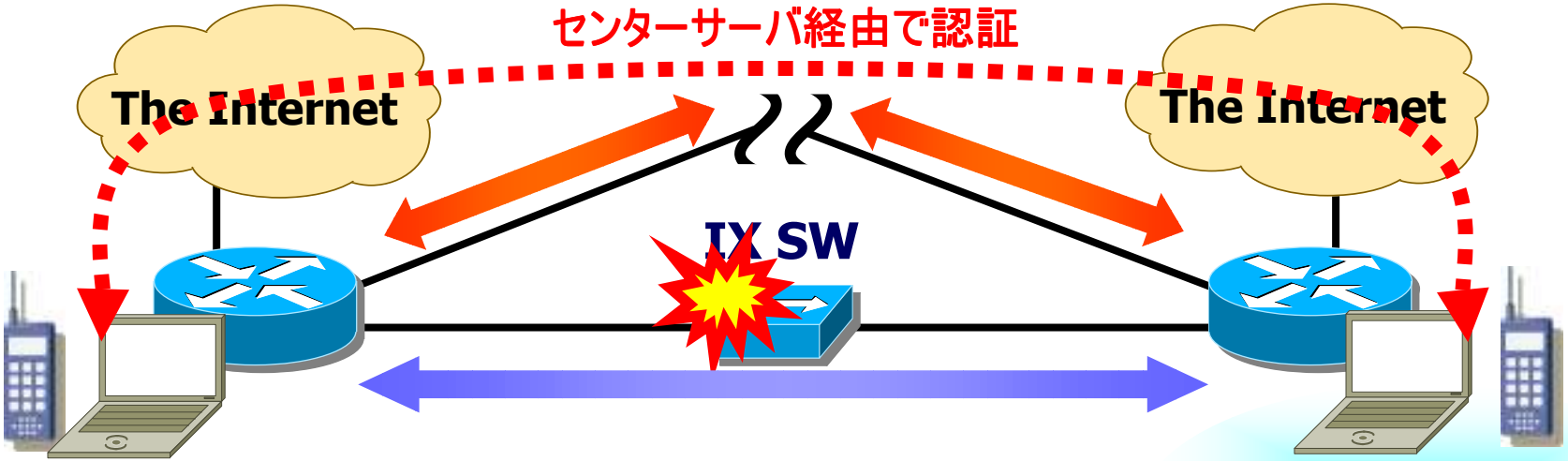
J-OS: 8.1R1.5

リアルタイム音声通信の検証

● Skype通話中の切替（PC、専用ハンドセット）

プライベートアドレス使用時、
スーパーノード経由で、バックアップする。
An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol（学術論文/ネタ元）

- 当然切替にかかる時間分切れる。
- 30秒程度でSkypeセッションが切断。
- プライベートアドレスを使うと、BGPが切替わらなくても10秒ほどで、切替わる!!?



C-OS: 12.0(31)S

リアルタイム映像の切替

● IP伝送装置からmpeg2動画を配信、バッファ無しのモニターで受信

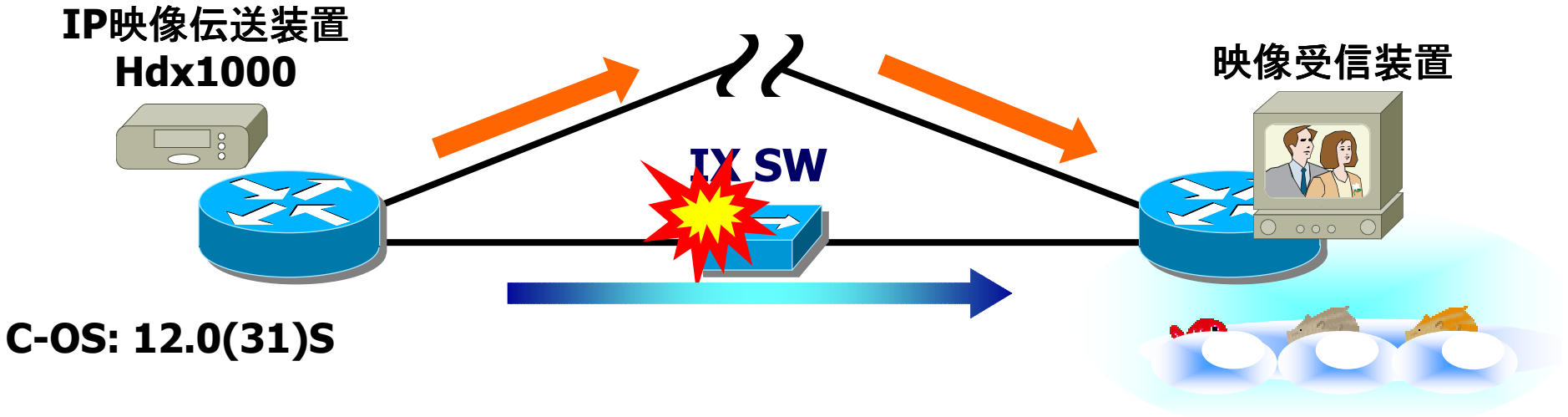
・ 断時間分、映像も途切れる。(BFDの場合、一瞬固まって、また動く)

映像デモ

BGP（Keepalive/Hold Time: 60/180sec)　#1　#2

BGP（Keepalive/Hold Time: 1/3sec)

BFD有り（Interval/Hold Time: 100/300msec)



C-OS: 12.0(31)S

リアルタイム映像について

● 地デジのIP同時再送信の流れ

- 著作権法改正により、IPによる伝送が、同時再送信に限り、権利処理が簡略化
- IPによるリアルタイム映像伝送が一気に広がる可能性

図1　IPマルチキャストを使って放送を同時再送信する際の権利処理が楽に　2006年12月15日に成立した改正著作権法によって実現。地上波放送のIP再送信の開始に向けて大きく前進した。

●放送を同時再送信したい場合に必要な権利処理

	CATVなどの有線放送	IPマルチキャスト放送（自動公衆送信）
著作権者（脚本家など）に対して	許諾が必要	許諾が必要
著作隣接権者（歌手や俳優などの実演家，レコード製作者）に対して	許諾は不要	許諾が必要
〈2007年1月11日〉から	報酬請求権を付与（使用料の請求があったときに支払う）	許諾は不要に（補償金を支払う義務）

CATV並みの権利処理で済むように

CATV：ケーブルテレビ

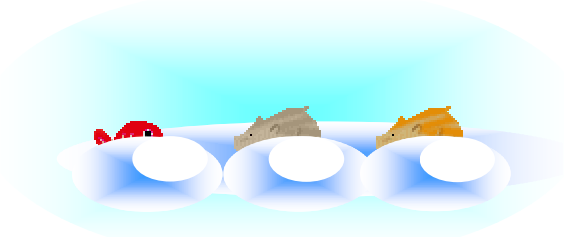
日経コミュニケーション　No. 478号 1.15/2007

すぐに、The Internetに流れ込んでくるという話にはならないが、IP装置によるネットワークへのより高い信頼性が求められてくるのでは?

色々わかった事

● 今回、検証したり、調査してわかった事

・ BFDのパラメータ変更は、BGPをリセットしない	C./J.
・ BFD無し状態で、BFDを新規設定時、BGPリセットなし	C./J
・ BFD状態で、BFDの設定削除、BGPダウン、その後アップ C.はダウンしない	J.
・ BFDパケット送信間隔30msec程で不安定	J.
・ BFDパケット送信間隔5msec程で不安定	C.
・ BFDパケット送信間隔40msec程で設定変更時に不安定	J.
・ BGP Hold Timeが20秒以下にならない	J.8.1R
・ BGP Hold Timeを20秒以下にするには、明示的に設定	J.(8.1R以外)
・ Skypeは、スーパーノード経由で自立的に切り替える	Skype Phone
・ BGPは、Default運用が多い	
・ BFD Authentication未実装　BFD DOS?	C./J.



まとめ

- 今回実環境において、**BFD**の動作も、**BGP Keepalive Interval**を短くした場合も、正常な動作を確認できました。
- 切替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、ネットワークとしてのサービス性向上を図ることができると考えられる。
- 社会的流れが変化（地デジの同時**IP**伝送、著作権法改正など)しているのに、今まで変わらない切替品質でよいのでしょうか**?**