



BFD deployment 状況と BGP convergence の実際

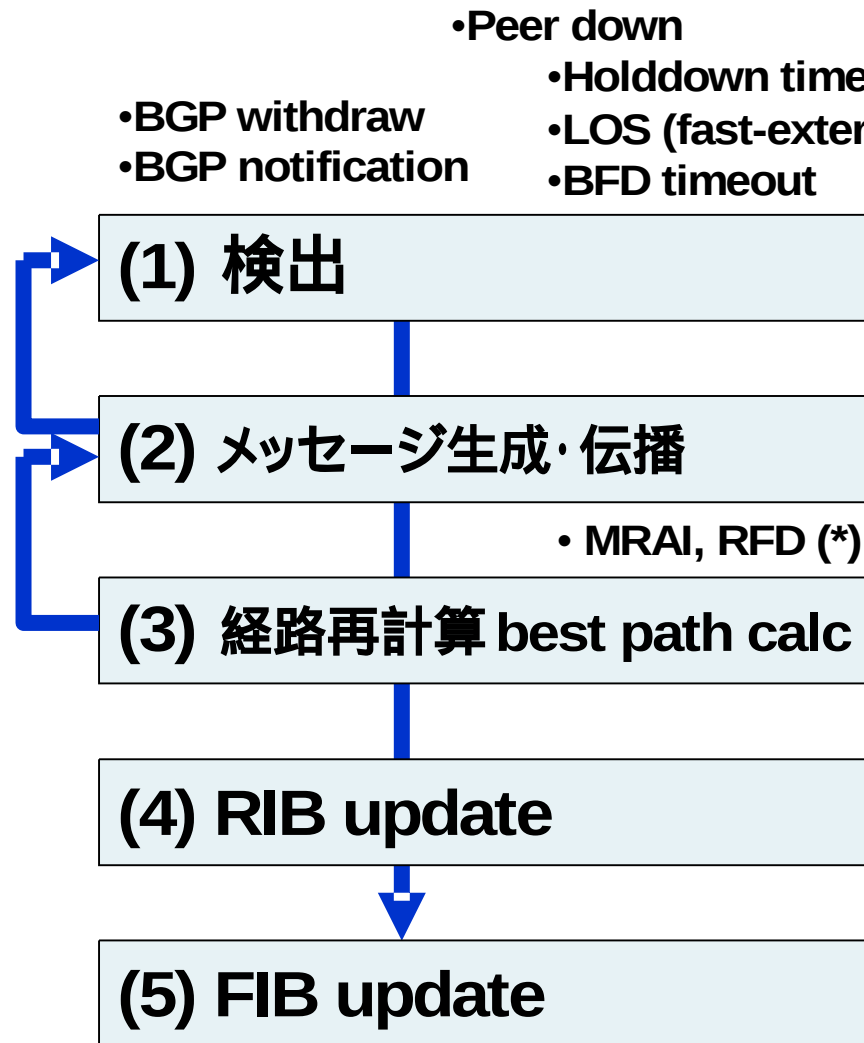
河野 美也 Miya Kohno, mkohno@cisco.com

The InternetのeBGP peeringで

... BFDを使用している例を調査してみた。

- 実際のdeployment事例はほとんど皆無に近い(?)
 - 顧客(CE)側が未サポート、相手(Upstream)側が未サポート
 - インターネットでは、stabilityが第一命題
RIPE-229ではfast-external-falloverさえやめろといっている...
 - これまでも、keepalive/holddownを然程短縮していない
但し、5sec/15secに短縮している例はあった
 - インターネットにおけるBFDの適用はまだ充分検証されたとはいえない
でも、状況は刻々と変わっている

BGP convergence



- **Convergence**

(1) à (2) à (3) à (4) à (5)

(*)

MRAI : Min route advertisement interval

RFD : Route Flap Dampening

BFD 考慮事項 – local matter

- L1. 下位レイヤとの組み合わせ
- L2. Link Aggregationとの組み合わせ
- L3. uRPFとの組み合わせ
- L4. IPv6
- L5. Graceful Restartとの組み合わせ

L1. 下位レイヤとの組み合わせ

- 下位レイヤのアーキテクチャとResiliency

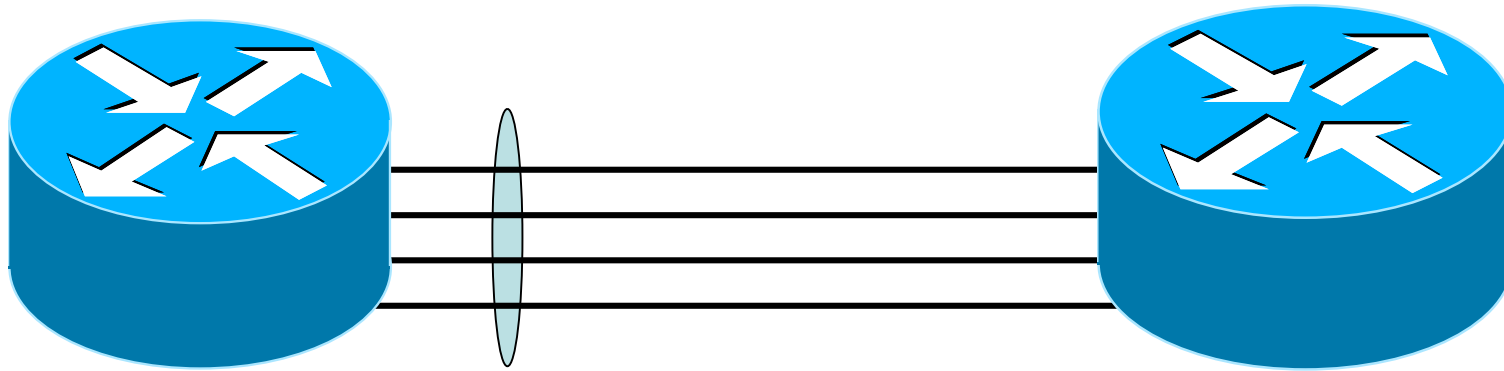
- 光伝送装置 (Optical Protection)
- Optical Cross Connect (Optical Protection)
- Ethernet Switch (STP/RSTP, VSRP/ESRP)
- UDLD
- Link Aggregation, LACP
- Ethernet Ring (MRP/EAPS, 802.17-RPR)
- 802.1ad Provider Bridge, 802.1ah Mac-in-Mac, EoE, VPLS
...それ以前に、現在はRouted PortしかBFDをサポートしていないかも、

- 警報転送

- Autonegotiation
- LFS (Link Failure Signal)
- E-LMI
- Ethernet OAM (802.1ag, Y.1731)

- BFD検出時間 > 下位レイヤの収束時間
- Alarmによる検出とBFDによる検出との棲み分け

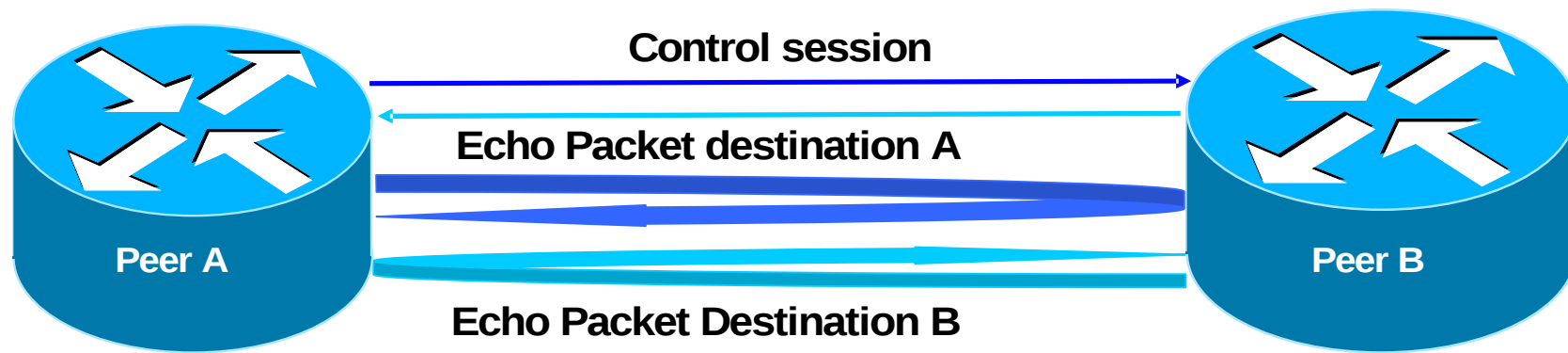
L2.Link Aggregationとの組み合わせ



- BFDは各component linkでなく、bundleされたlogical linkに対し、動作する。

- 各Component link毎の高速障害検出には使用できない。

L3.uRPFとの組み合わせ



- uRPF が設定されていると、echo packetを廃棄してしまうので注意。

L4.IPv6

draft-ietf-bfd-v4v6-1hopより

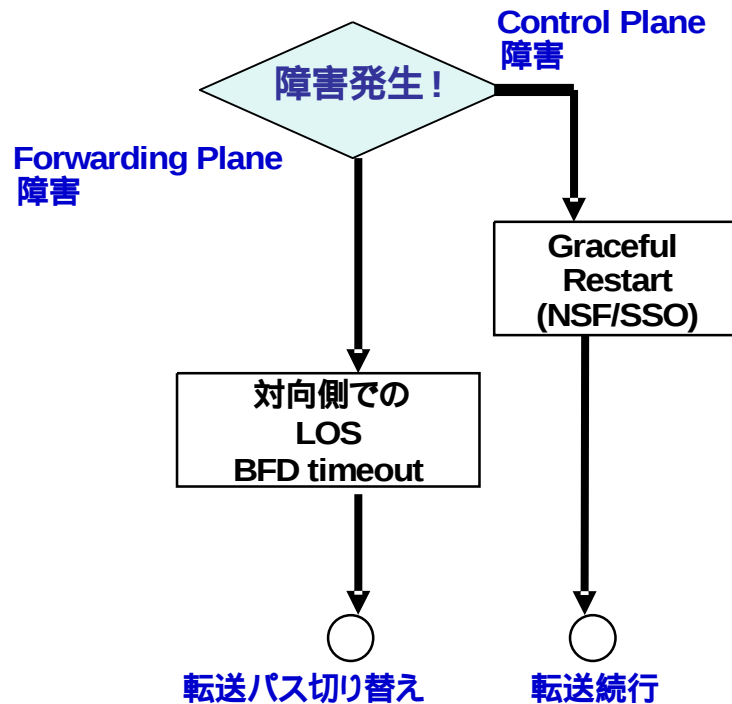
4.2. BFD for IPv6

In the case of IPv6, BFD Control packets MUST be transmitted in UDP packets with destination port 3784, within an IPv6 packet.

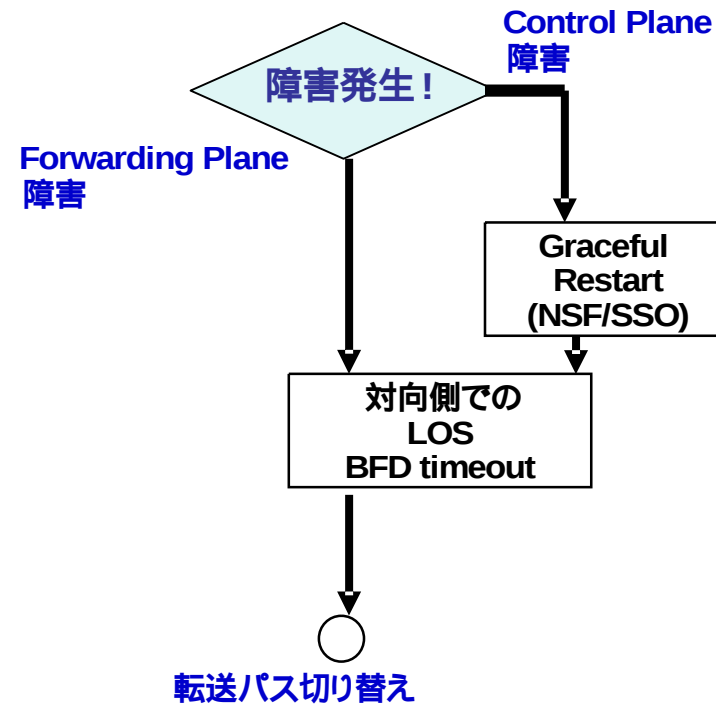
- IPv6の検出のために、IPv6 BFDのセッションが必要。
- 未サポート!! (2007年3月現在)
- EEM/TCL ?!

L5.Graceful Restartとの組み合わせ

distributed BFD



centralized BFD



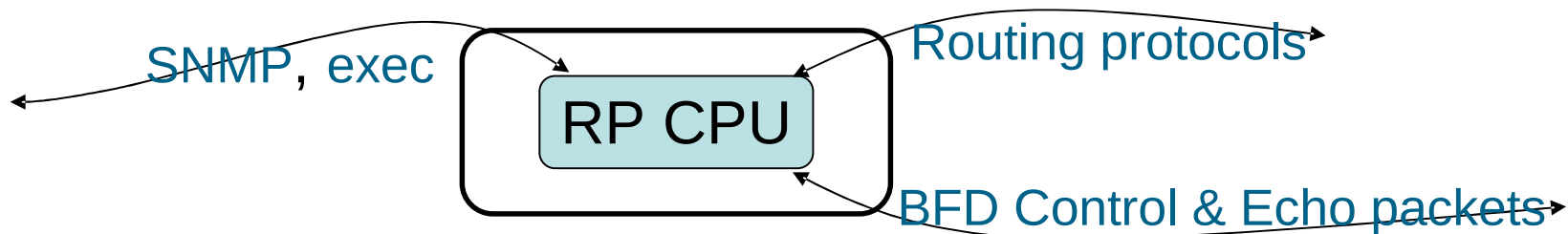
- BFDの実装によっては、graceful-restartにより転送を保持している際に切替ってしまう可能性がある。

BFD 実装形態

- Centralized
- Distributed
- Semi-dedicated

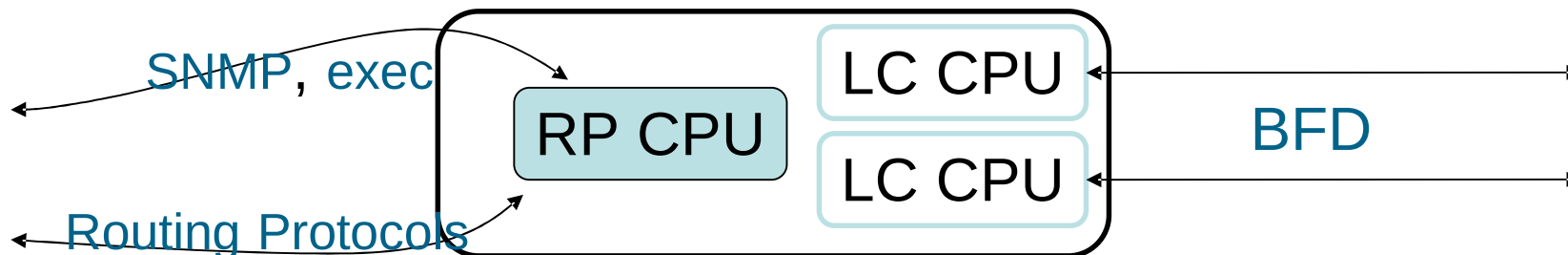
Centralized

- 共通のCPUで、すべてのControl Plane (+Data Plane) 処理を実行する。
- 負荷競合の可能性
 - BFD echo 処理
 - BGP UPDATE 処理
 - IGP SPF
- Graceful Restartとの共存問題
- Scalability / Performance



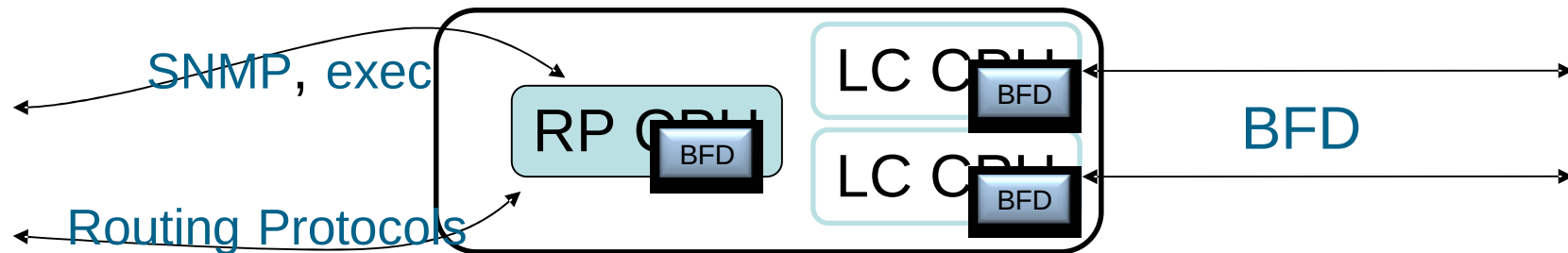
Distributed

- BFDセッション処理は、LC CPUにオフロードされている。
- RP 処理には影響されない。
- Control PlaneとForwarding Planeの分離可能
- RP <-> LC間IPC bottleneck ?
- RP 上のBFD clientへの障害通知伝播の遅延可能性
- Multi-hop, multipath, link-bundlingは考慮が必要



Semi-dedicated

- BFD session処理が、専用もしくは準専用HW上に実装される。
- Scalability実現のために、さらにDistributed構成をとる 場合もある。
- 実現できればこれが理想的(?)。



BFD 考慮事項 – global matter

G1. Route Oscillation ?

G2. PICとの組み合わせ

G1. Route Oscillation ?

一般的に、Fast Detection-ConvergenceとStability/Scalabilityは拮抗する。

→ 共存させるために:

- Carrier delay
for DOWN event, UP event
- Exponential Backoff
for IGP spf delay, LSA generation
- BGP RFD, MRAI (*)
- Interface dampening
- ...

(*) しかし、arbitraryなdampeningは、却って害がある。Local isolationが必要。
Is RFD harmful ?!!

<http://www.ripe.net/ripe/docs/routeflap-damping.html>

G1. Route Oscillation ?

Routing Stability、およびInstabilityによる影響は、永年の研究課題

- T. G. Griffin and G. Wilfong, .An analysis of BGP convergence properties,. *ACM SIGCOMM*, September 1999.
- C. Labovitz, R. Malan, and F. Jahanian, .Internet routing stability, *IEEE/ACM Trans. Networking*, pp. 515.528, October 1998.
- C. Labovitz, R. Malan, and F. Jahanian, .Origins of pathological Internet routing instability,. *IEEE INFOCOM*, 1999.
- C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, .Delayed Internet routing convergence,. *IEEE/ACM Trans. Networking*, vol. 9, pp. 293.306, June 2001.
- C. Labovitz, A. Ahuja, and F. Jahanian, .Experimental study of Internet stability and wide-area network failures,. *Fault-Tolerant Computing Symposium*, June 1999.
- W. Fang and L. Peterson, .Inter-AS traffic patterns and their implications,. *IEEE Global Internet*, December 1999.
- A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, .Deriving traffic demands for operational IP networks: Methodology and experience,. *IEEE/ACM Trans. Networking*, vol. 9, June 2001.
- J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, BGP Routing Stability of Popular Destinations, imw02
- R. Bush, T. Griffin, Z. Mao, E. Purpus, Happy Packets – Some Initial Results, ripe48, Sept.2004
- <http://www.ietf.org/internet-drafts/draft-irtf-routing-reqs-07.txt>
-

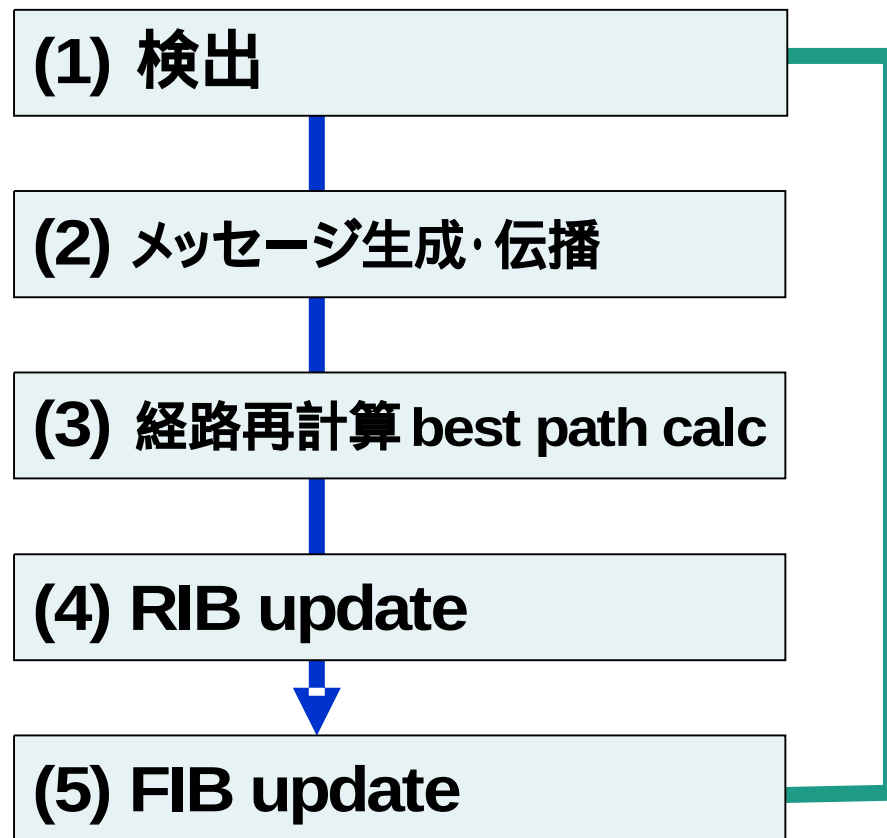
決定論的なことは言えない。しかも、時間が経つにつれて、前提条件となる状況も変化している。

**適用形態によっては、stability阻害要因になる可能性もある。
Interface Dampeningのような技術と組み合わせ、Flapを局所化させる、等。**

BGP PIC

(Prefix Independent Convergence)

•Nexthop Invalid

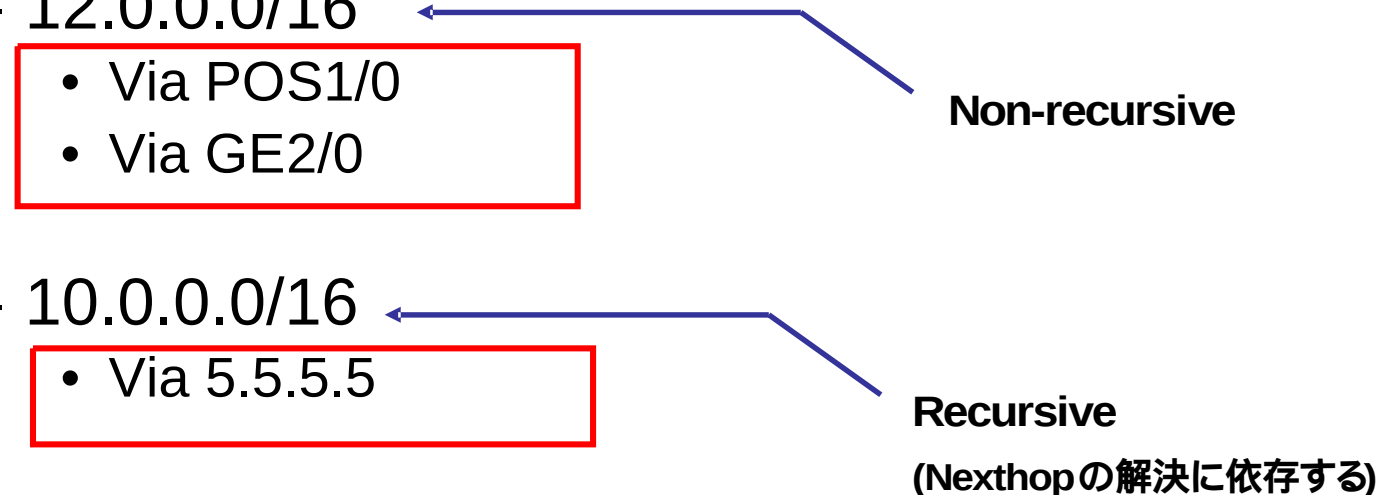


- **PIC**
(1) à (5)
- **Convergence**
(1) à (2) à (3) à (4) à (5)

BGP PIC - Basic Terminology

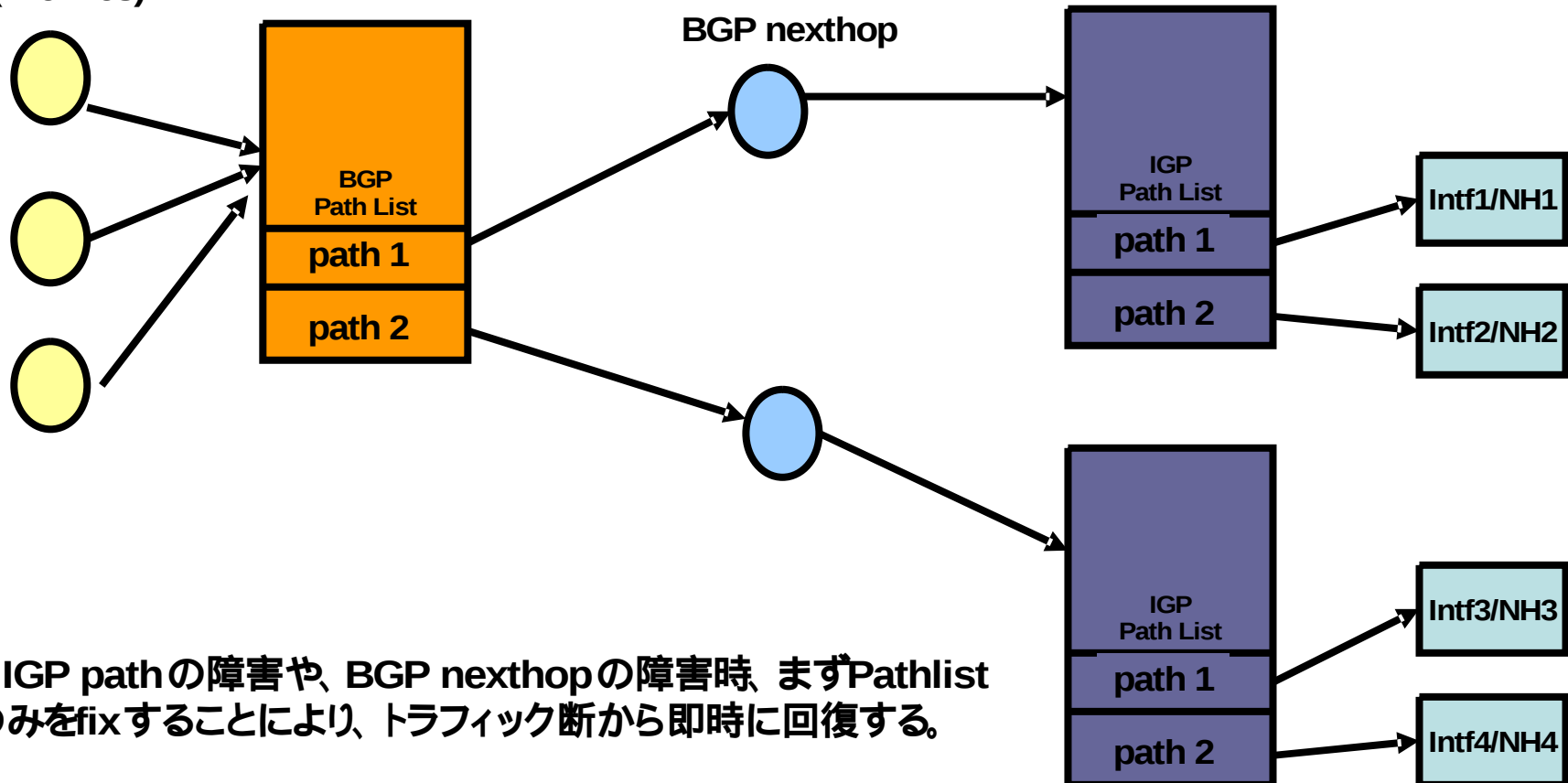
- **Prefix** – Routing Protocol から学んだ経路
 - 12.0.0.0/16
 - **Pathlist** – Routing Protocol から学んだNexthop のリスト
 - 12.0.0.0/16
 - Via POS1/0
 - Via GE2/0

Non-recursive
 - 10.0.0.0/16
 - Via 5.5.5.5

Recursive
(Nexthopの解決に依存する)
- 

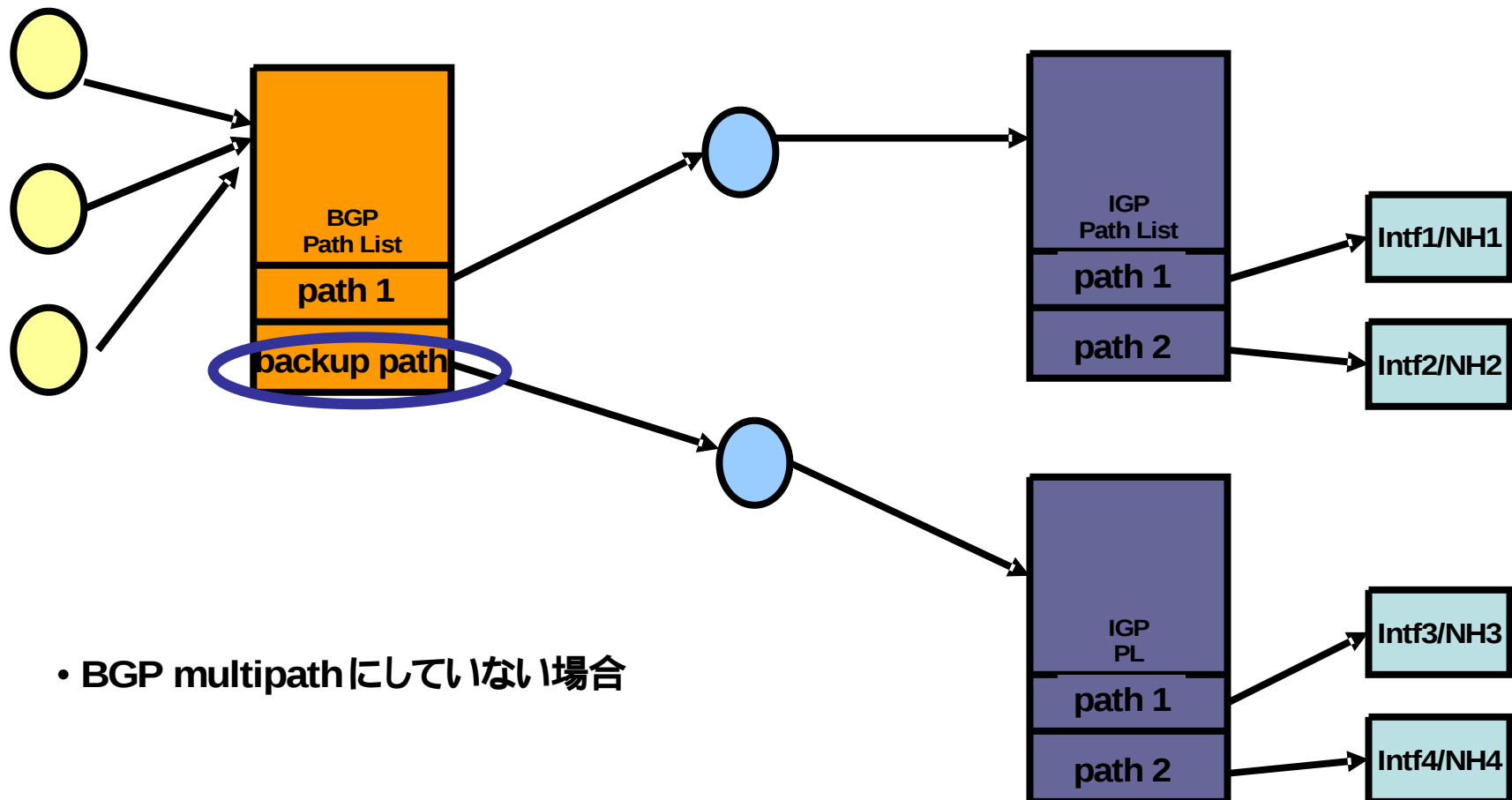
BGP-PICとFIB 構造

FIB Table
Entries
(Prefixes)



- IGP pathの障害や、BGP nexthopの障害時、まずPathlistのみをfixすることにより、トラフィック断から即時に回復する。
- その後、通常のBGP convergenceが行われる。

BGP-PIC single path ?



Single path時の問題

- Problem:
 - BGP は一つのbest pathしか広告しない。
 - 別のnexthopから広告されても、RRが介在している場合は、RRが一つのbest pathを選んでしまう。
- Solution (?!)
 - Add-path draft

ADD-PATH

- あるprefixに対する複数のpathを、元のpathを暗黙的にreplaceすることなく、広報させることを許容するための機構。
- 他のprefixと区別するため、path identifierを付加している。

```
+-----+
| Path Identifier (4 octets) |
+-----+
| Length (1 octet)         |
+-----+
| Prefix (variable)        |
+-----+
```

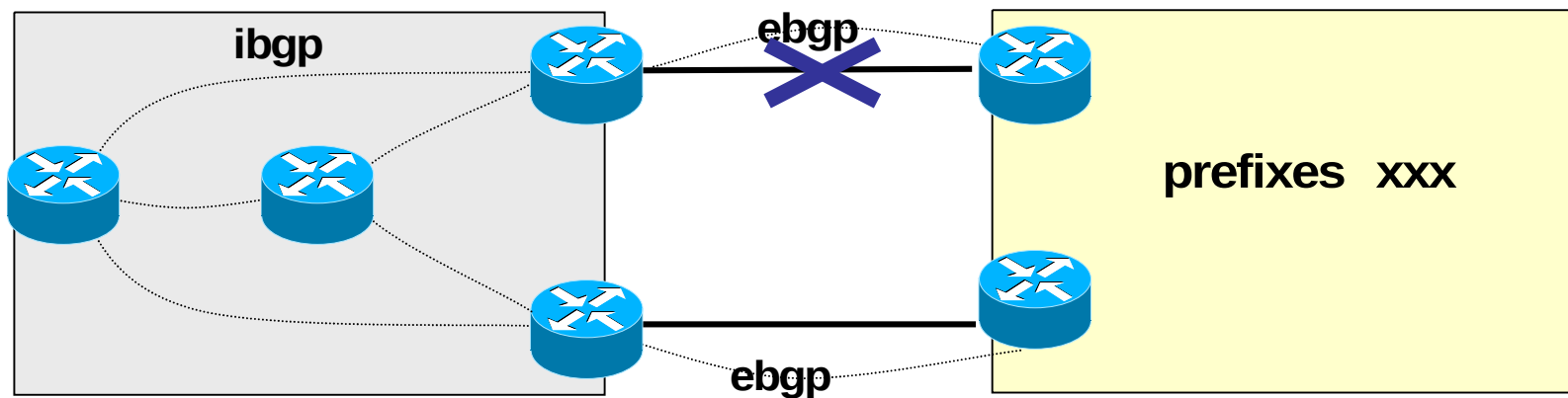
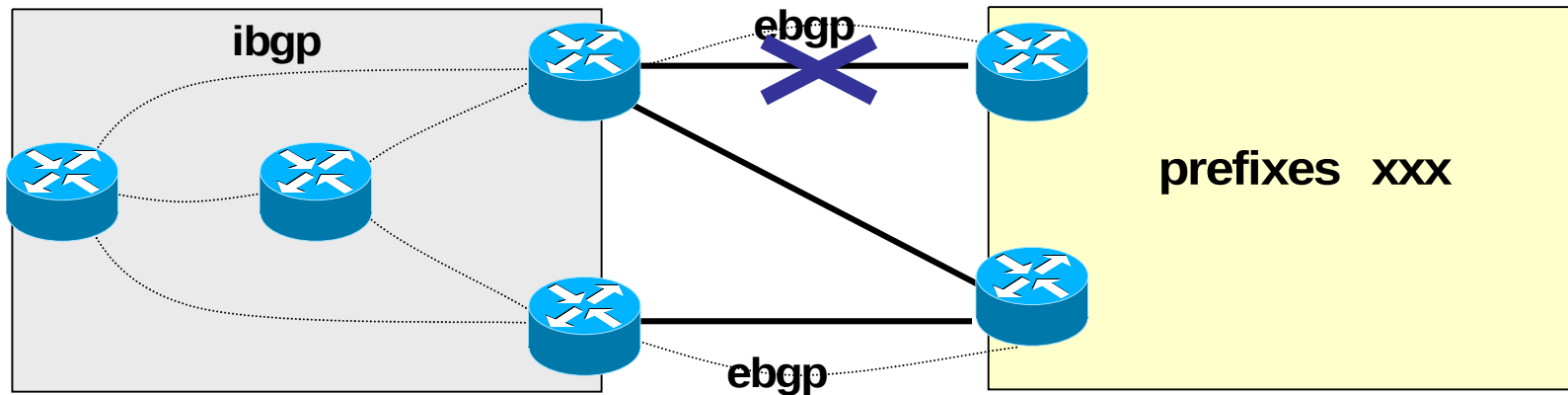
```
+-----+
| Path Identifier (4 octets) |
+-----+
| Length (1 octet)         |
+-----+
| Label (3 octets)         |
+-----+
.....
+-----+
| Prefix (variable)        |
+-----+
```

(for rfc3107)

ADD-PATH – 動作

- New capability: Add-path
- 複数PATHを受けられるということをcapabilityにて広報する。
- 特定のAFI/SAFIをcapabilityで指定することにより、そのAFI/SAFIに対して複数PATHを送信することを示す。
- Convergence の改善と複数のPATHを広報することの負荷とのtrade-off ??

G2. PICとの組み合わせ



おまけ - BFD and MPLS

- BFD for MPLS-FRR trigger (*1)
or
- BFD for MPLS-LSP liveness (*2)

