CISCO SYSTEMS

High Avaialabilityの設計と実装 ― 理想と現状

MPLS Japan 2004

2 Nov. 2004, Miya Kohno (mkohno@cisco.com)

High Availabilityを実現する技術

High Availabilityを実現するための、多岐に渡る技術が開発されている。

Fast Detection

- POS trigger control
- Link Pass Through
- OAM/VCCV
- Fast Hello, RSVP Hello
- BFD …

L1/L2 protection

- SONET/SDH APS
- DPT/RPR
- Ethernet Ring
- Link Bundling
  - Ethernet Channel
  - POS Channel

L3 Fast Convergence

- IGP Fast Convergence
- BGP Fast Convergence
- HSRP/GLBP/FHSRP…
- Load Balancing

MPLS-TE Protection

- TE/Fast Reroute
  - Link/Node Protection
  - Path Protection

Fault Tolerant Architecture

- NSF
- SSO/NSR
- ISSU
- MK/Process Preemption…

Resiliency

- Graceful Restart
- Graceful Shutdown, …

では、それらをfull動員させれば、High Availabilityを達成できるのか。

- … それがあまり簡単ではない。☹

Need to be careful…

1. 時定数の問題

各要素技術が、それぞれ異なる時定数を持つ。

- SONET/SDH APS --- 50m秒
- Ethernet LOS検出 --- driver仕様に依存。(数10m秒~ 数100m秒)
- Ethernet Ring --- 方式により異なる。(数10m秒~ 数10秒)
- RPR --- 50m秒
- TE/FRR --- 検出ロジック(POS alarm or rsvp-hello) およびProtection手法に依存
- Link Bundling --- 条件、設定に依存 (数秒～数10秒)
- Fast Hello --- 条件、設定に依存 (1秒～)

- IGP Convergence --- 条件、設定に依存 (数100m秒~)
- BGP Convergence --- 条件、設定に依存 (数秒 ~)

→ 各要素の時定数を考慮した設計をしないと、非効率、干渉を起こす。

Nレイヤの検出時間 ＞ N-1レイヤの検出時間

BGPの再計算契機 ＞ IGPの収束時間 (BGP nexthopをIGPで解決している場合)

Wait for BGP.. (BGPルータが再起動した際の、IGPが先に収束してしまうことによる不整合を防ぐ)

Need to be careful…

2. *Hysteresis, Dampening, Exponential Backoff*

- ネットワークは、系のつながり
- ある系の故障が他に影響を及ぼし、Cascade Failure, Meltdownを引き起こす危険性がある。

→ 障害検知および障害への反応は、単に早ければよい、というものではない。

→ 適切なhysteresis、Dampening、Exponential Backoff Algorithmにより系を不安定状態から保護する必要がある。

Need to be careful…

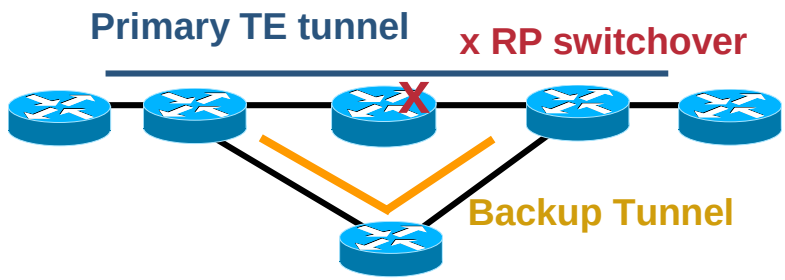
3. 組み合わせの問題

- Fast Convergence/Protectionは、相手の故障を一刻も早く検知し、一刻も早く他ノードに通知し、一刻も早く切り替える。
- NSF/Graceful Restartでは、相手が故障しても切り替えず、そのまま転送を保つ。

→ 拮抗する

（例）FRR- Node Protecition. RPR+を定義していたために、却って収束が遅くなった例。

Redundancy Mode	Traffic Loss Time
Non-redundancy (single RP)	4 msec
RPR	140 msec
RPR +	7300 msec



Need to be careful…

3. 組み合わせの問題

回線障害は、Fast Convergence/Protection
Control Plane障害は、可能であれば、NSF/GR。

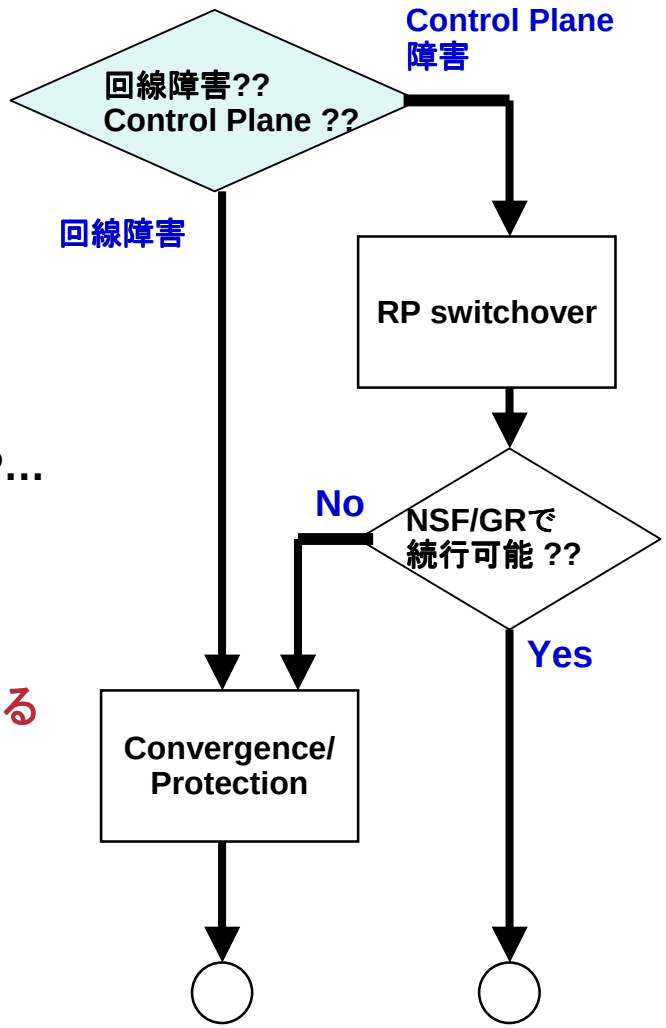
→ 全てのProtocol/Stateが維持されるか

IGP, BGP, LDP, targeted LDP, RSVP-TE, HSRP…
結構いろいろ使っていたりする。

どれかが切れると不整合の原因になる。

→ keep alive等Control Planeで回線障害を検出する場合

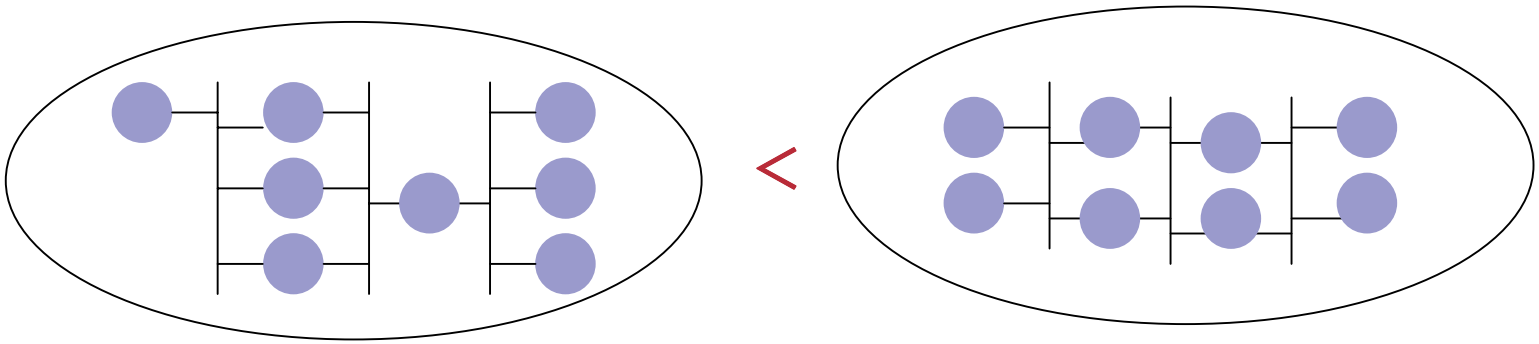
判別がつかない！！
BFDは大丈夫？ → 実装による



Need to be careful…

4. バランスの問題

- HAは、冗長を基本とするため、コスト高になりがちだが、バランスが悪いと、コストをかけただけの効果が得られない。



- 各Component（ ）のAvailabilityを90%とした場合、

  右の系のAvailability ＝ 96.1%

  左の系のAvailability ＝ 80.8%

  •かけたコスト（ の数）は同じなのに…

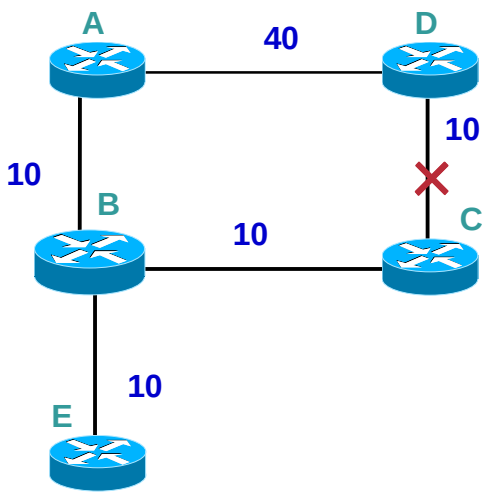
Need to be careful…

4. バランスの問題

- 少なくとも同一機能要素の中では、ヴェンダー、SW versionを統一したり、CPU性能を統一したり、利用するHA技術要素のconsistencyを保つ必要がある。

  → 収束速度が異なると、Micro Loopとなり、却って効率悪化の可能性がある。

  → 切り替えポリシーが異なると、不整合が起こる。



- BのみFast ConvergenceのTuning(SPF delay timer, LSA throttle timer, etc.)が行なわれていると仮定する。
- E->DのTrafficに注目。Best Pathは、E->B->C->D。
- C-D間回線障害が起こると、Bは直ちにSPF計算を行いAにTrafficを送るが、Aはまだ収束しておらず、BにTrafficを送り返す。(→ これは、いずれにせよ回線が切れているのでまぁよい。)
- C-E回線が復旧する。
- Bは直ちにCにTrafficを送るが、Cはまだ収束しておらず、BにTrafficを送り返す。→ これは問題。

まとめ

- Networkは”Complex System”
- High Availabilityを実現するためには、

  各系の挙動や仕組みを把握し、

  - 時定数
  - Hysteresis, Dampening, Exponential Backoff
  - 組み合わせ
  - バランス

  ……に充分注意した上で、

ネットワーク全体からのTop-Down設計を行なう必要がある。

Disclaimer

- このプレゼンテーションは、Network設計のSEの立場から書きました。

- ヴェンダーとしては、当然乍ら、
  - High Availability機能の実装をすること
  - バグを無くすこと
  - Standardizationにコミットすること

  …が至上命題と認識しています。

なのであまりいぢめないでね。