OSPFのDRインタフェース断が他の通信に与える影響

ーJanog IRS8 発表用ー

2006／4／21

株式会社NTTデータ
吉野　誠吾

OSPFのDRインタフェース断が他の通信に与える影響

1. 問題の概要
2. OSPF
3. 問題発生のメカニズム
4. 解決策、回避策
5. ルータの実装改善提案

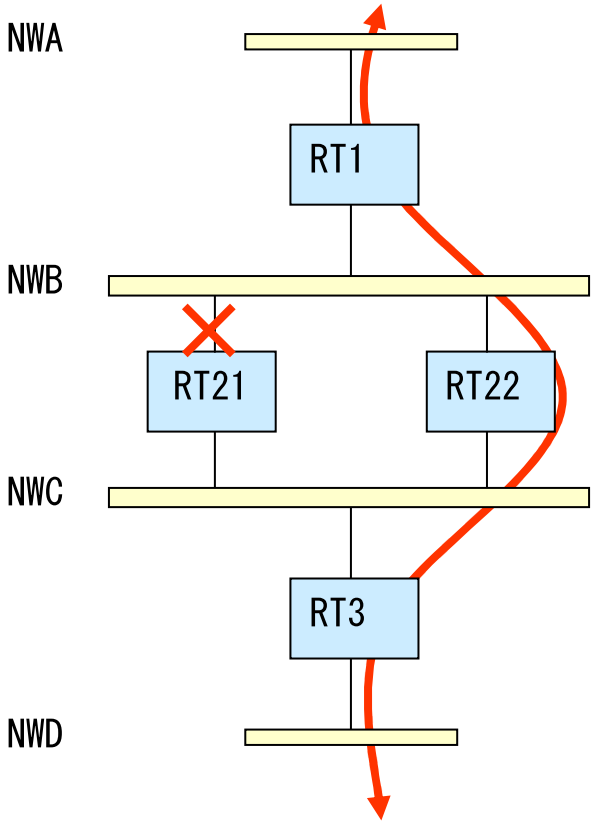
1. 問題の概要

OSPF を用いたネットワークで、
DR のインタフェースのリンクが落ちる（落とす）と、
LAN 上の他ルータ間の通信も止まってしまう場合がある、

という問題について述べる。

本問題は OSPF の仕様上の問題であるため、全ての OSPF ルータで発生し得る。

1. 問題の概要:図で示すと



RT21 が NWB の DR となっている状態。
RT21 のインタフェースをメンテナンスのため
shutdown(故障で down でも同じ)してしまうと、
矢印の通信も止まってしまう。

実通信に関係のない装置の停止が影響する。

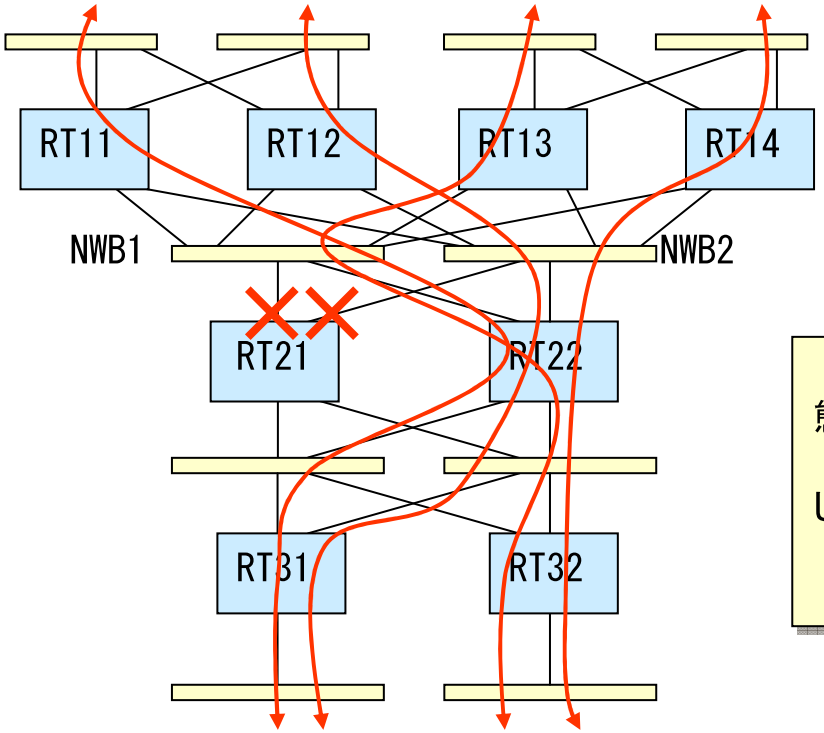
1. 問題の概要：発生のトリガー

本問題は、以下のどちらのトリガーでも発生する。

1. 機器故障　　：インタフェースのハード障害による down
2. メンテナンス：メンテナンスによる shutdown

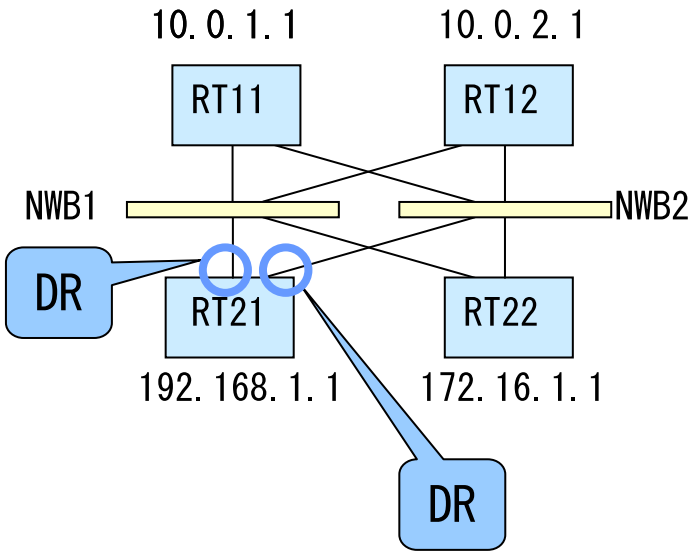
特に 2. のケースの方がより問題と認識されるケースが多いものと考える。

1. 問題の概要:二重化されていても(1/2)



RT21 が NWB1、NWB2 の両方の DR となっている状態。
メンテナンスのためインタフェースを続けて shutdown してしまうと、
NWB1、NWB2 を経由する通信が全て止まってしまう。
結果、全断となる。

1. 問題の概要:二重化されていても(2/2)



Router Priority が同じ値の場合(設定していないような場合)、DR は Router ID 値の大きいルータのインタフェースとなる。

このため、左図のような環境において、2 つのネットワークの DR が 1 台のルータに集中してしまう可能性は高くなる。

結果、先に説明したように全断となってしまう状況はできやすくなる。

1. 問題の概要：本ドキュメントは

本ドキュメントは、問題が発生するメカニズムを解説し、Janog ML の議論でいただいた現状取りうる解決策と回避策をまとめたもの。
また、機器やプロトコルの実装変更による解決策を提案する。

OSPF について、説明上必要な事項については解説するが、OSPF の基本は理解されていることを前提とした説明となっている。

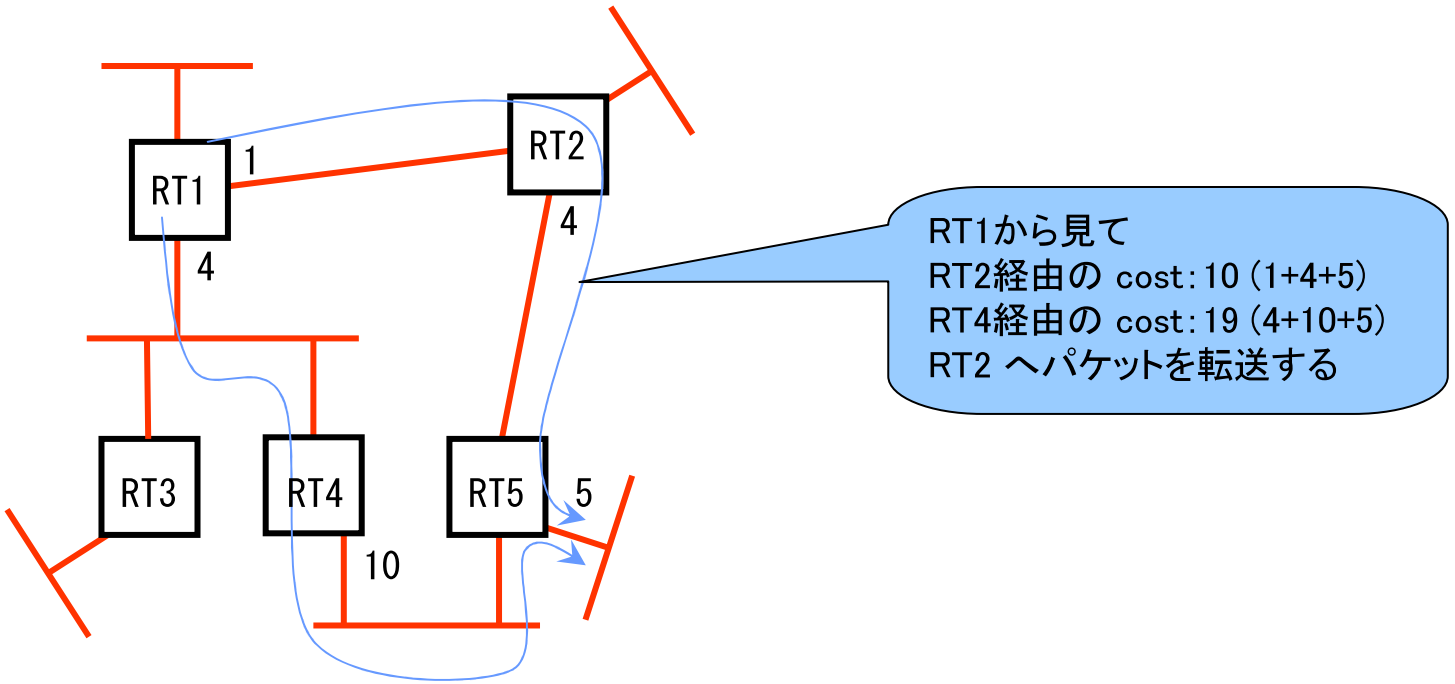
Janog：JApan Network Operators' Group（http://www.janog.gr.jp/）

OSPFのDRインタフェース断が他の通信に与える影響

1. 問題の概要
2. OSPF
3. 問題発生のメカニズム
4. 解決策、回避策
5. ルータの実装改善提案

2. OSPF：OSPF とは・・

Router－Router 間の接続を Link で表現する。
Link には、通過するのに必要な cost 値が付与される。
OSPF は Router が目的の IP network に関して cost 値の総和が最小となる経路を選択するプロトコル。

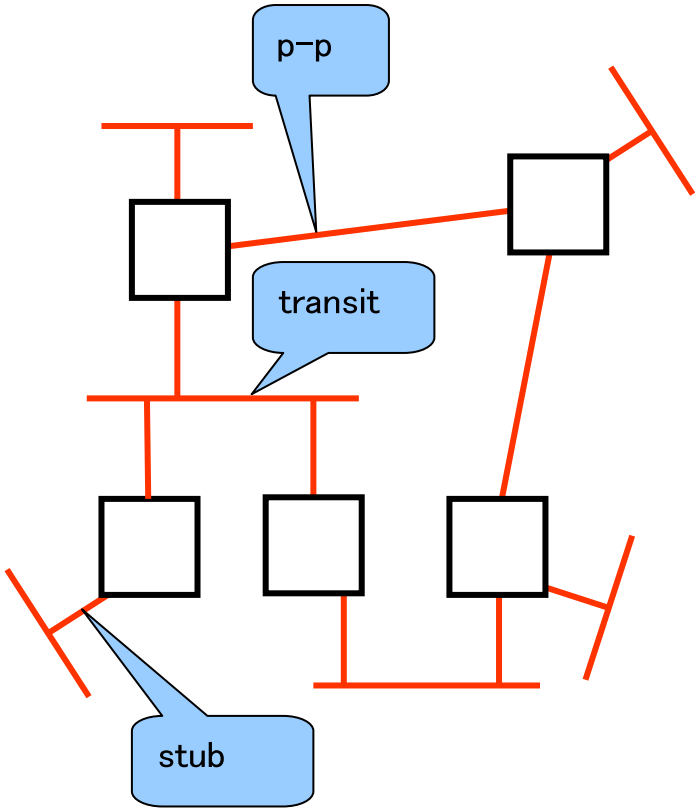


2. OSPF：OSPF とは・・

２－１　3つの Link type (p-p、transit、stub)
２－２　Router LSA と Network LSA
２－３　Link type と LSA の関係
２－４　LSA の flush(消去)：MaxAge

2－1　3つの Link type (p-p、transit 、stub)

Router と Router を結ぶ Link には、(virtual link を除くと)以下の3種類の type がある。



1. p-p(point-to-point)
　他の Router との接続を意味する。
　router link と呼んだ方が分かりやすいと思われる。
2. transit network
　複数の Router が接続する IP network との接続を意味する。
3. stub network
　他に Router が接続していない IP network との接続を意味する。
　直訳は切株。木構造のネットワークの端っこで、これ以上先(OSPF の場合 Router)がないという意味。

重要:
　同じ LAN インタフェースでも(Adjacent な)対向 Router が 1 台も、いないと stub、いると transit となる。

2－1　3つの Link type（p-p、transit 、stub）

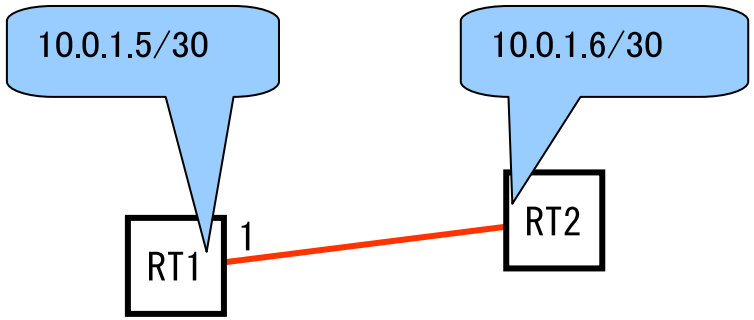
前のスライドの図では link = 回線のように見えるが、Router から見た接続関係と捉えた方がより分かりやすいと考える。
そのため、p-p は router link と呼んだ方が分かりやすいと考える。

Link type	何との接続か？	Link ID に入る値	Link Data に入る値	IP subnet mask の情報は？
p-p	隣接 Router	隣接 Router ID	Router のインタフェース IP（unnumbered の場合はMIB-Ⅱ ifIndex）	含まない
transit network	Network LSA	DR のインタフェース IP	Router のインタフェース IP	
stub network	IP network	IP prefix	IP address mask	含む

ルーティングテーブルを作るためにはさらに別の情報が必要。
p-p の場合は、別途 stub network の情報を付与（後述）する。
transit の場合は、Network LSA の中に情報を含んでいる。

2－1　numbered p-p は stub が必要

p-p(router) の Link 情報には IP address mask に関する情報は含まない。



Router LSA には、IP アドレスが付与されている物理回線に対して 2 つの Link 情報を含む

■RT1 の Router LSA の例

Link 数：2

Link ID：RT2 の Router ID

　Link Data：10.0.1.5

　Link type：p-p

　metric：1

Link ID：10.0.1.4

　Link Data：255.255.255.252

　Link type：stub network

　metric：1

２－２　Router LSA と Network LSA

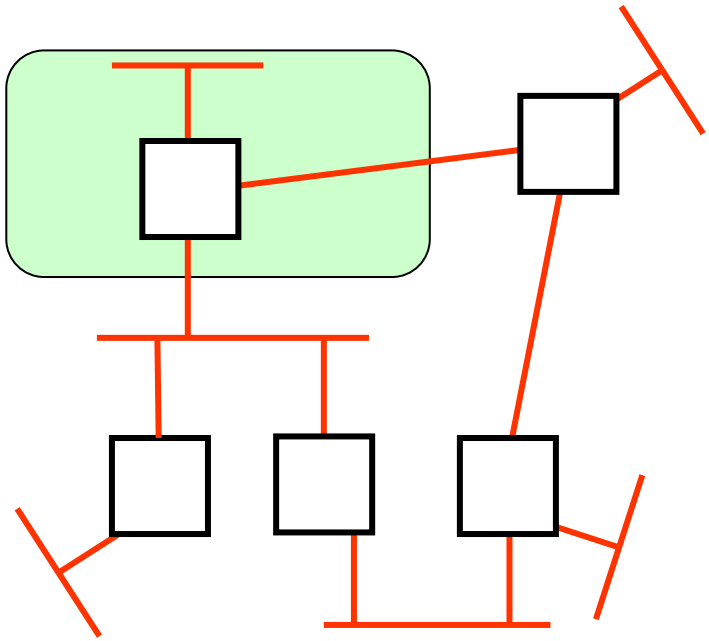
OSPF では Router と Link の情報を LSA という単位で構成する。
RFC では 5 つの LS type が定義されているが、Area 内のトポロジーは、

1. Router LSA
2. Network LSA

の 2 つで表現できる。

2－2　Router LSA

Router につながる Link の一覧。
各 Router が 1 つ生成する。

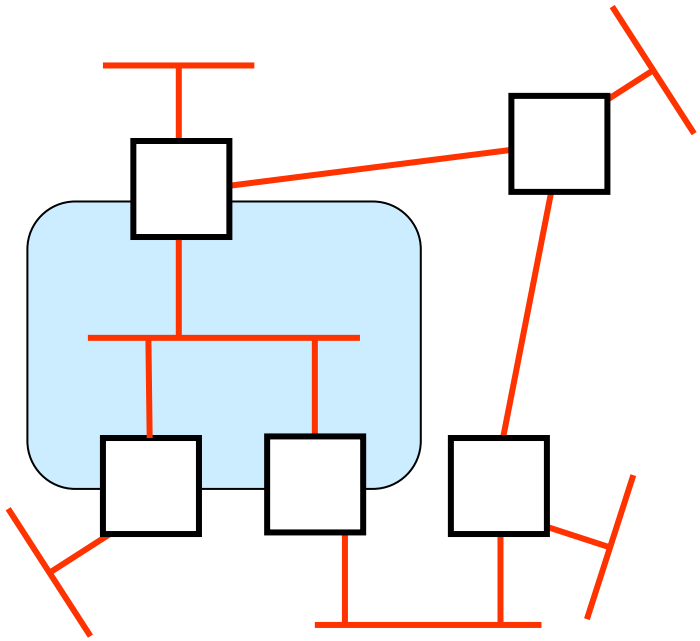


Router LSA に含む Link の数：n
Link 1 の情報
Link 2 の情報
　　|
Link n の情報

Link の情報とは、
1. Link ID
2. Link Data
3. Link type (p-p、transit、stub)
4. metric (cost のこと)

2－2　Network LSA

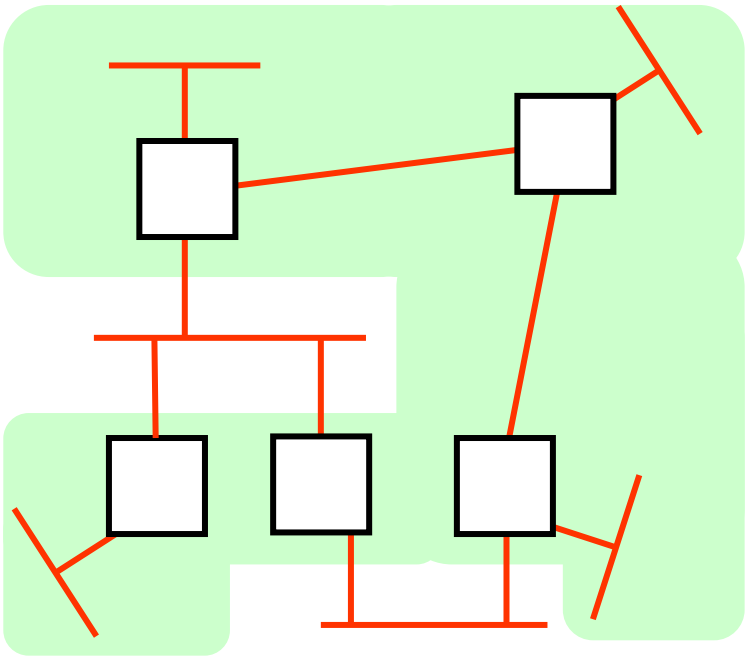
Network（link type=transit）につながる Router の一覧。
各 Network の DR Router のみが 1 つ生成する。



Network mask
Router 1 の インタフェース IP address
Router 2 の インタフェース IP address
　　|
Router n の インタフェース IP address
（n はパケット長から計算して求める）

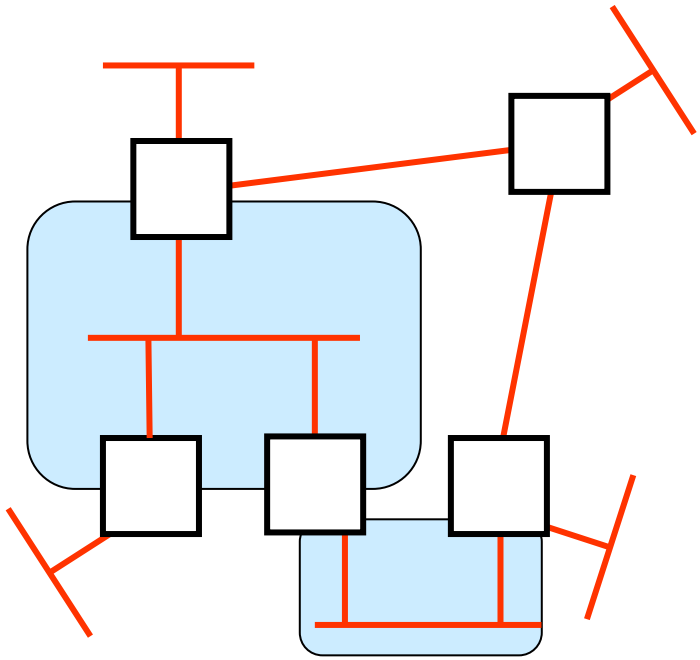
２－２　Router LSA のカバー範囲

Router 自身と、Link type が transit 以外の部分



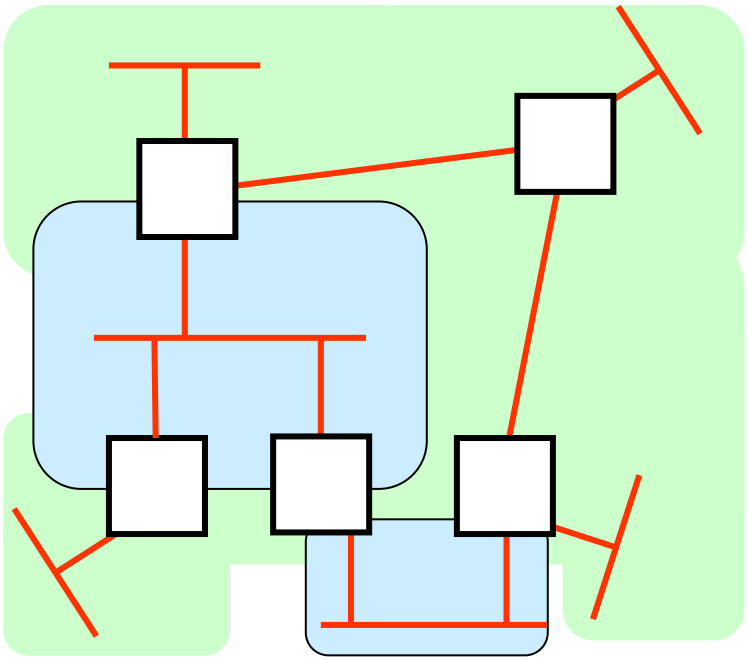
2－2　Network LSA のカバー範囲

Link type が transit の部分のみ。

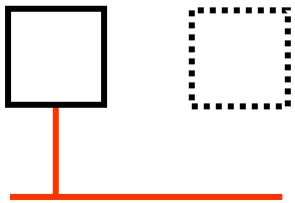


２－２　Router LSA と Network LSA で

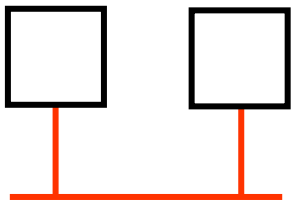
Area 内全てをカバーする



２－３　Link type と LSA の関係



stub



transit

Router LSA のみ

それぞれの Router が Router LSA を生成
DR に選ばれた Router が Network LSA を生成

重要！
Adjacency が切れてしまうと、link type は stub に変化する。
この結果、
1. Router LSA の修正と再送信
2. Network LSA の flush（削除のための送信）

が必要となる

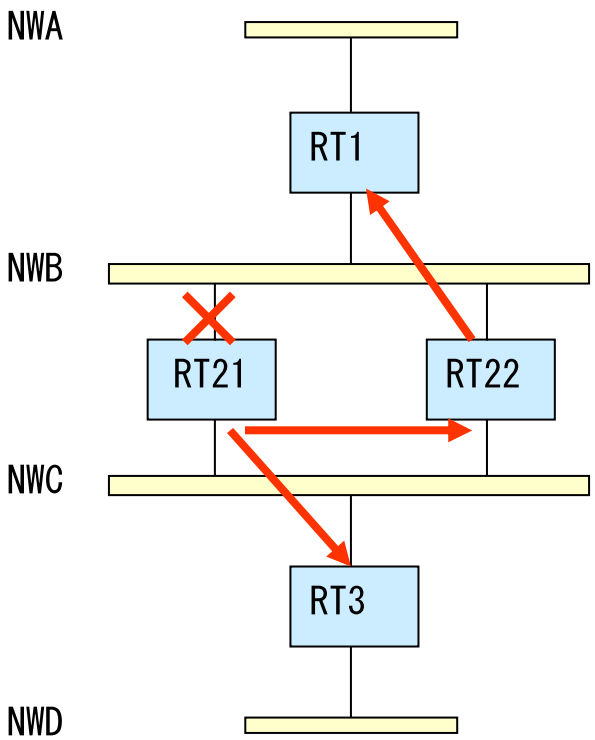
２－４　LSA の flush(消去) : MaxAge

　OSPF には削除するためのメッセージやコマンドはない。
　LSA が不要になった場合、ネットワーク上から削除するには、LSA の LS Age フィールド（単位は秒）に MaxAge（3600秒）を設定して送信することによって伝える。
　何も変化がなく 3600 秒経過してもネットワークから消えていく。このため、LSA に変化がなくても、LSA を生成した Router は、LSRefreshTime（30分）経過すると再送信する仕様となっている。

OSPFのDRインタフェース断が他の通信に与える影響

1. 問題の概要
2. OSPF
3. 問題発生のメカニズム
4. 解決策、回避策
5. ルータの実装改善提案

3. 問題発生のメカニズム



1. RT21 が NWB の DR となっているとする。
2. NWB の Network LSA は RT21 が生成している
3. RT21 の NWB との接続が切れると Network LSA を flush（削除）する
4. MaxAge の Network LSA は他の全てのルータに伝わってしまう（図の矢印）
5. 全てのルータは NWB が存在しなくなったものとしてルーティングテーブルの再計算を行う
6. 結果 NWB を経由した通信はできなくなる。
7. NWB 上の RT1 もしくは RT22 は DeadInterval（デフォルト 40 秒）後に RT21 との neighbor が切れたことを認識し、BDR が DR に昇格する
8. 昇格した DR が新たな NWB の Network LSA を生成し、全てのルータに伝え、ルーティングテーブルを再計算して復旧する。

3. 問題発生のメカニズム：発生要因

要因としては以下のような点が挙げられる。

1. Network LSA を DR ルータしか生成できない（DR への依存）
2. 冗長化構成のため、裏から削除メッセージ（MaxAge）が全体に伝わる
3. （メンテナンス時に）DR ルータが他のルータに、「今から落ちる」と伝える術がない

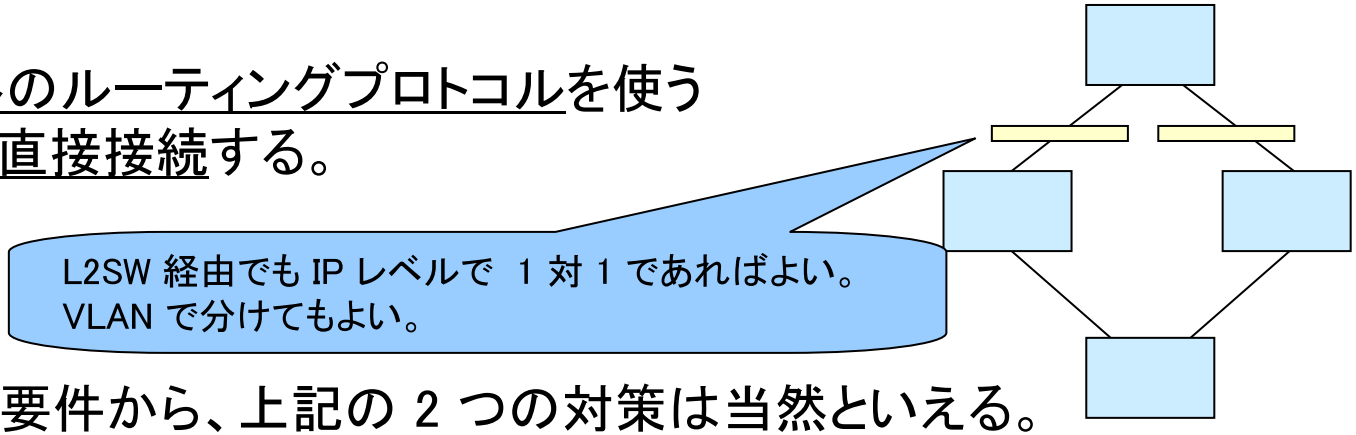
OSPFのDRインタフェース断が他の通信に与える影響

1. 問題の概要
2. OSPF
3. 問題発生のメカニズム
4. 解決策、回避策
5. ルータの実装改善提案

4. 解決策、回避策

解決策:
　現状では、障害時、メンテナンス時の両方において本問題の発生を抑制するには、以下のどちらかしかないようである。

1. OSPF 以外のルーティングプロトコルを使う
2. ルータ間を直接接続する。



本問題の発生要件から、上記の 2 つの対策は当然といえる。
　しかし、現実的には既存のネットワークにおいて、OSPF 以外のプロトコルへの切替や使用は難しいと考えるし、トポロジーを変更するのも簡単ではないと考える。
　新規に構築する際には考慮したい内容と言える。

4. 解決策、回避策

緩和策:

デフォルトの 40 秒程度止まるのが長い、という場合、タイマー値の変更や BFD などの早期検出機能を用いて、検出時間を短縮するという対策である。

ただし、ルータの実装においては、検出後ルーティングテーブルの変更までに数秒かかる物もあるため、検出時間≠停止時間であることに注意を要する。

本ドキュメントでは、そもそもの停止をなくす目的で議論、提案するものであるので、本対策(fast convergence)についての詳細は省略する。

4. 解決策、回避策

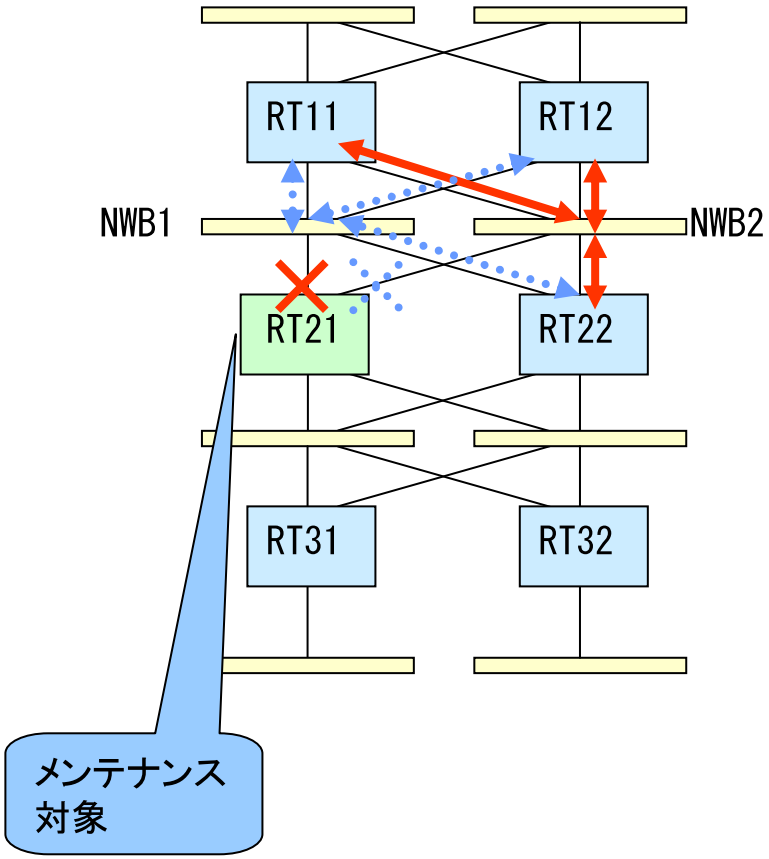
回避策:

次に、機器故障時はやむを得ないとして、メンテナンス時の本問題発生の抑制や事象の緩和(停止時間の短縮)について Janog ML 等で検討した内容を整理したい。

残念ながら完璧な手法は見つかっていないが、本問題に直面した方が同様の検討に時間を要することを避けるため整理するものである。

また、OSPF 仕様の問題点や動作の再確認ができ、この検討がルータへの改善提案を考えるきっかけともなっている。

4. 回避策案1:1経路ずつメンテナンスする

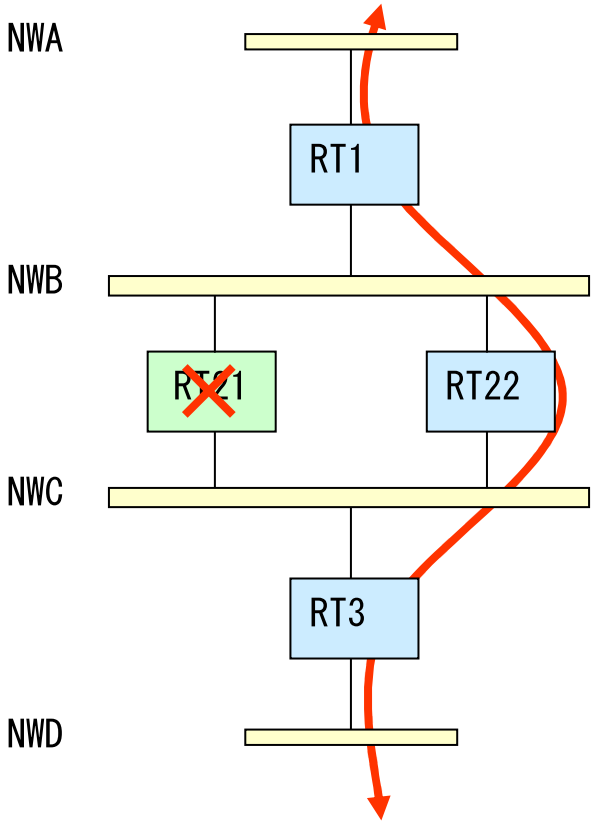


考え方:メンテナンス対象のルータに加えて、shutdownするインタフェースの L2SW もトラフィックが経由しない状態(左図の NWB2 経由)として作業を行う。
　作業完了後、通信の経路を変えて(左図の NWB1 経由)、同様の作業を行う。

本問題の回避:可能

説明:
　冗長経路が必須である。
　また、作業が複雑で工数が多くなる。

4. 回避策案2:電源を落とす

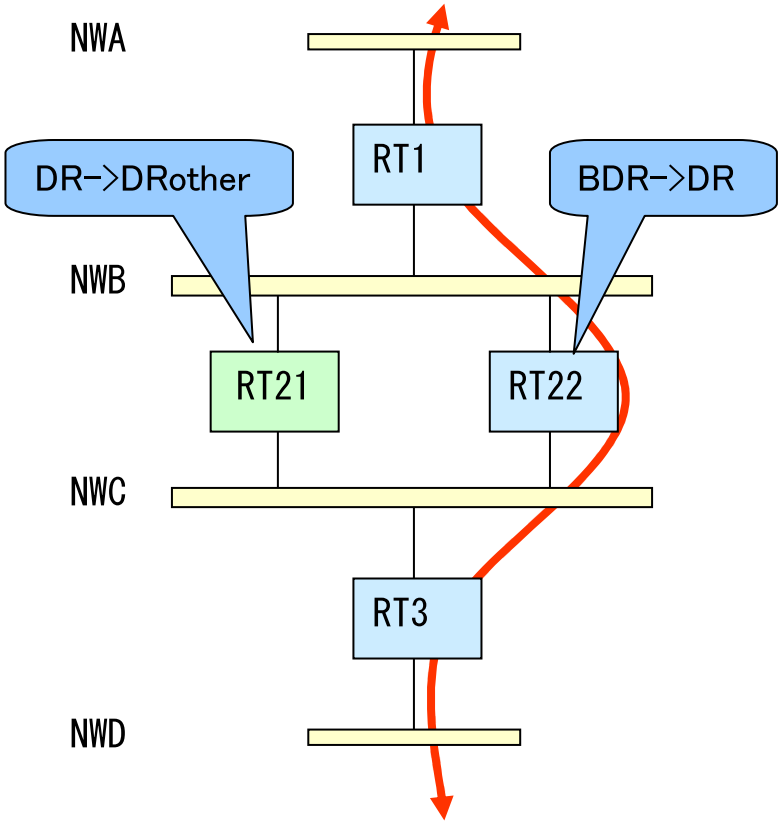


考え方:作業対象のルータを電源をいきなり落とし、ネットワークから切り離して作業を行う。

本問題の回避:可能

説明:
　電源を落としてはいけない仕様の装置も存在するので、全ての場合に適用できるわけではない。
　電源スイッチが押された際に、Network LSA を flush する仕様の装置が(今後も)ないとは限らない。(再起動のコマンドでは、このような実装が存在する)

4. 回避策案3:DRを譲る

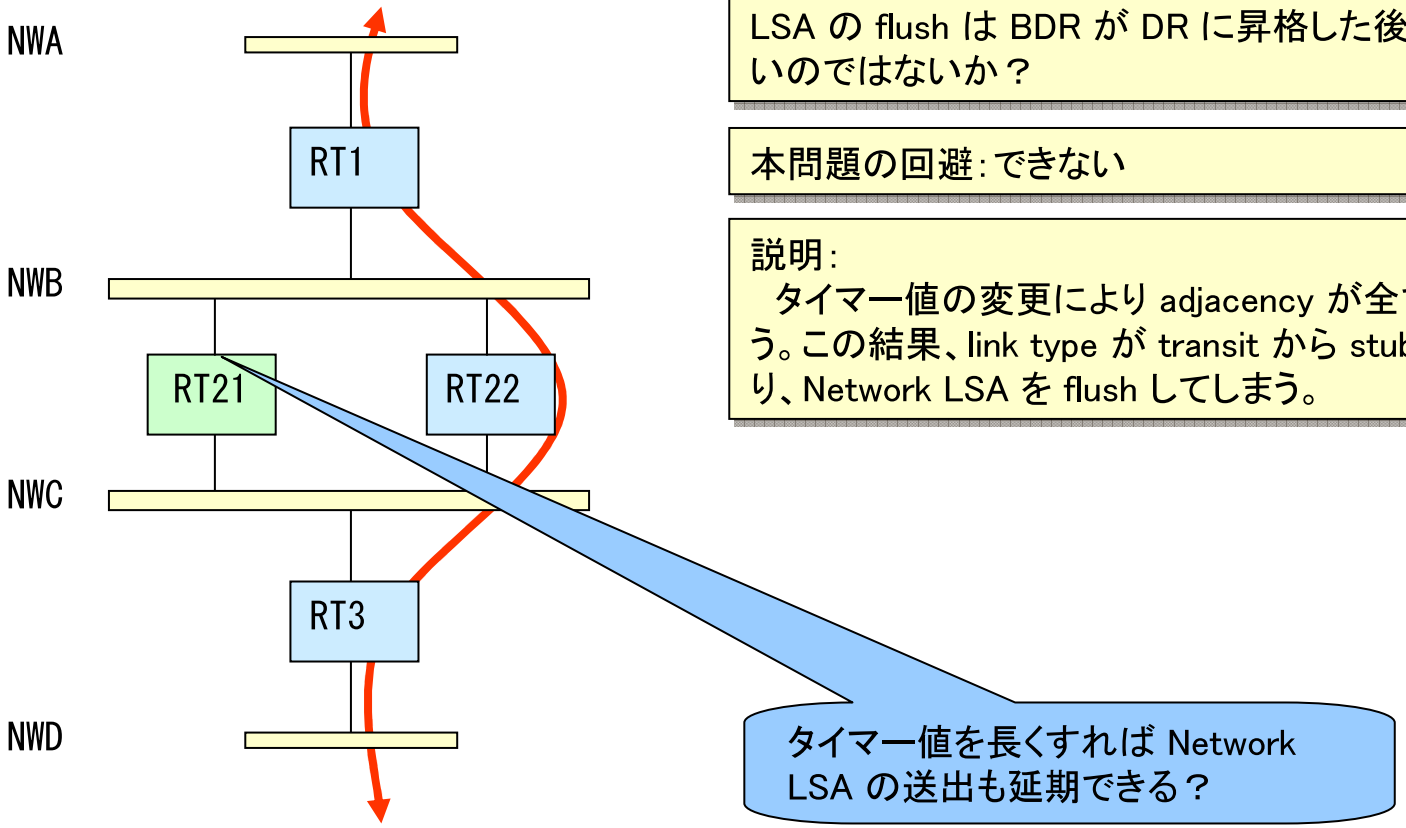


考え方:DR になっているインタフェースの Router Priority を 0 に変更し、DR を他のルータに譲ってから shutdown するという考え。

本問題の回避:可能な実装とできない実装がある。

説明:
Router Priority を 0 に変更すると DR は DRother となり、BDR が DR に昇格するが、DR が DR でなくなる際に Network LSA を flush する実装と、flush しない実装があり、flush してしまうと本問題の解決にはならない。

4. 回避策案4：DeadInterval を長く変更

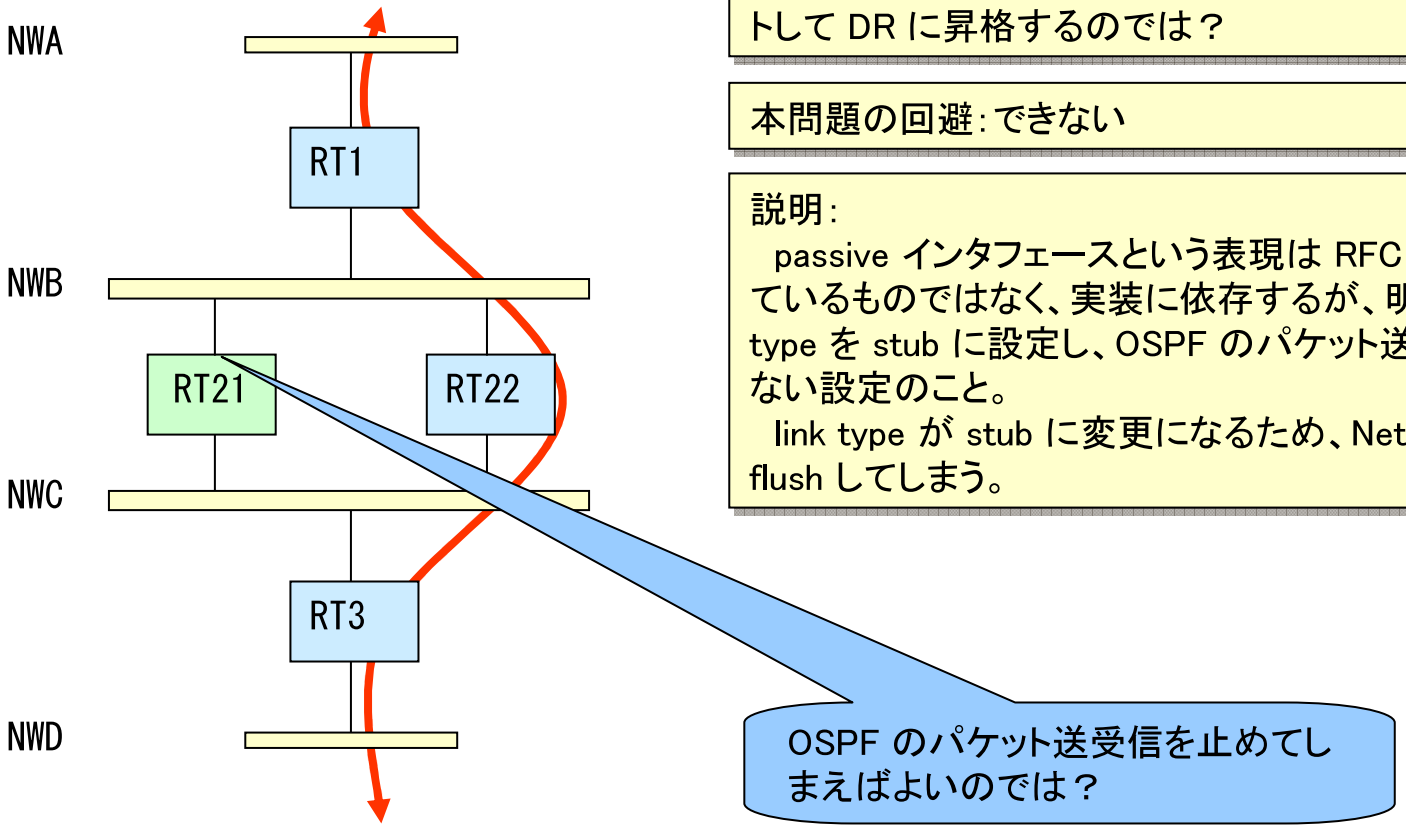


考え方：タイマー値が一致しないと neighbor(adjacency) が切れるが、DeadInterval を長くしておけば Network LSA の flush は BDR が DR に昇格した後で影響はないのではないか？

本問題の回避：できない

説明：
タイマー値の変更により adjacency が全て切れてしまう。この結果、link type が transit から stub に変更になり、Network LSA を flush してしまう。

4. 回避策案5:passive インタフェースに変更

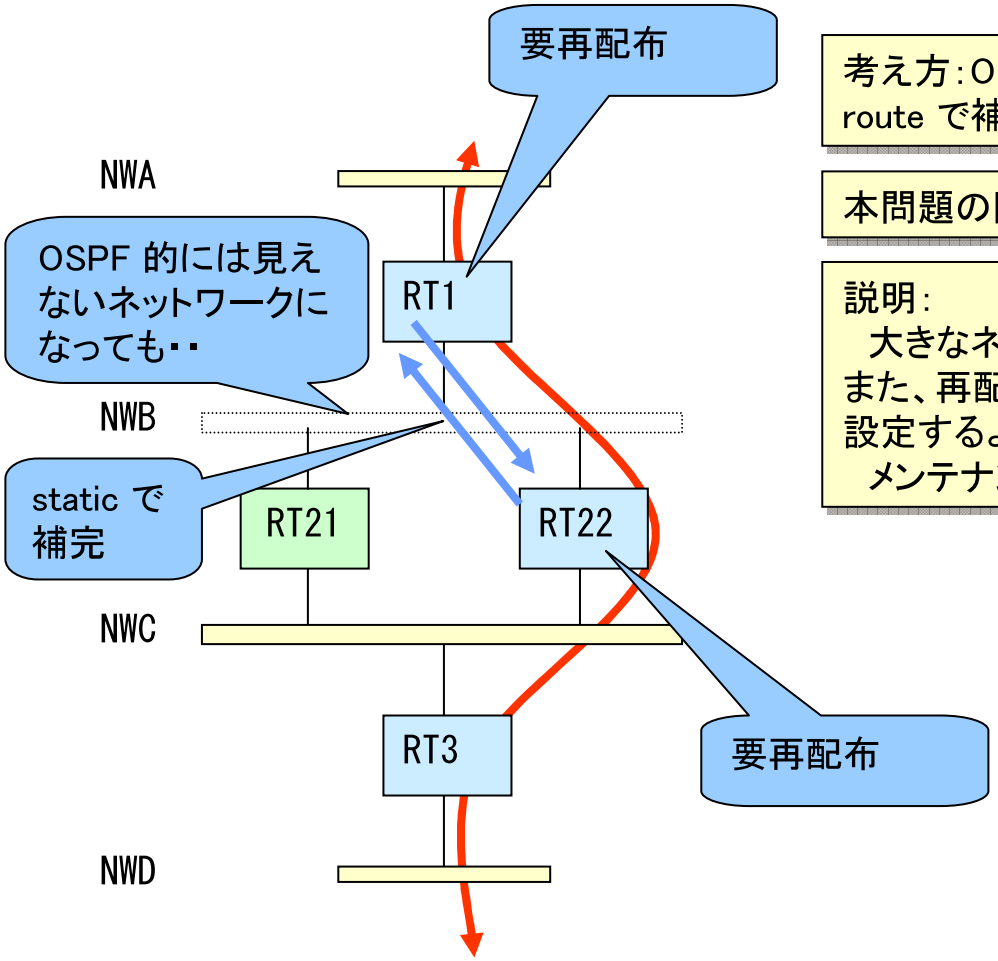


考え方:passive インタフェース設定に変更すれば、OSPF のパケット送出が止まるので、BDR がタイムアウトして DR に昇格するのでは？

本問題の回避:できない

説明:
passive インタフェースという表現は RFC に規定されているものではなく、実装に依存するが、明示的に link type を stub に設定し、OSPF のパケット送受信を行わない設定のこと。
link type が stub に変更になるため、Network LSA を flush してしまう。

4. 回避策案6:(floating) static で補完

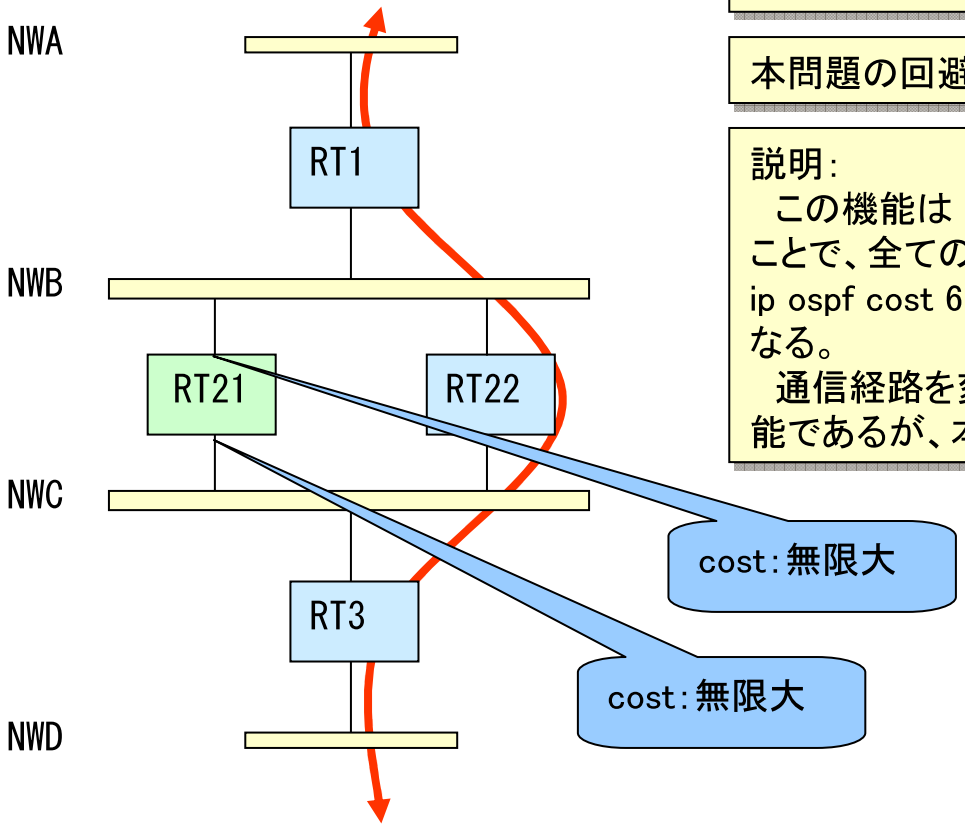


考え方:OSPF で消えてしまう経路情報を予め static route で補完する

本問題の回避:可能

説明:
大きなネットワークでは設定する経路情報が多くなる。また、再配布(redistribute)設定など矛盾のないように設定するよう注意が必要である。
メンテナンス対象以外の設定変更も必要である。

4. 回避策案7：Cisco graceful shutdown

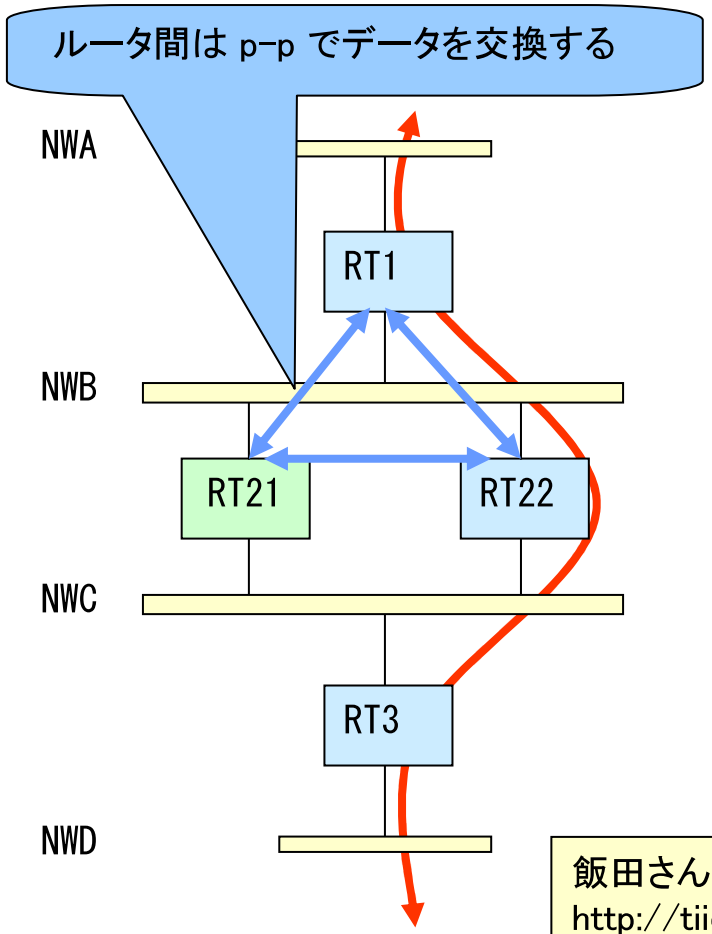


考え方：機能名称からして周囲のルータに通知するなど、通信停止を回避して shutdown できるのではないか？

本問題の回避：できない

説明：
　この機能は wait for BGP 機能の派生の使用方法のことで、全てのインタフェースに、ip ospf cost 65535 と同時に設定するのと同等の動作となる。
　通信経路を変更するという作業においては有用な機能であるが、本問題の回避にはつながらなかった。

4. 回避策案8：point-to-multipoint mode



考え方：ルータ間を point to point で扱えればよいのではないか？ATM や FR などの broadcast 機能のないメディア向けの動作 mode だが、LAN でも使える

本問題の回避：可能

説明：
point-to-multipoint（p-mp）では、ルータ間の link を p-p の集合として扱うため、DR や Network LSA の概念が存在しない。このため、RT21 のインタフェースを shutdown しても RT1-RT22 間の通信には影響がない。
ただし、実装によって LAN インタフェースで設定した場合の動作が異なるため、neighbor の設定、タイマー値の調整、IP MTU の調整、等の設定が必要となる。
また、p-mp では該当ネットワーク（左図の NWB）の IP について、host route（stub）でしか扱われないため、NWB 上に別の IP host が存在する場合通信できない。

飯田さんの解説：
http://tiida.cocolog-nifty.com/netblog/2004/10/7_p2mp.html

OSPFのDRインタフェース断が他の通信に与える影響

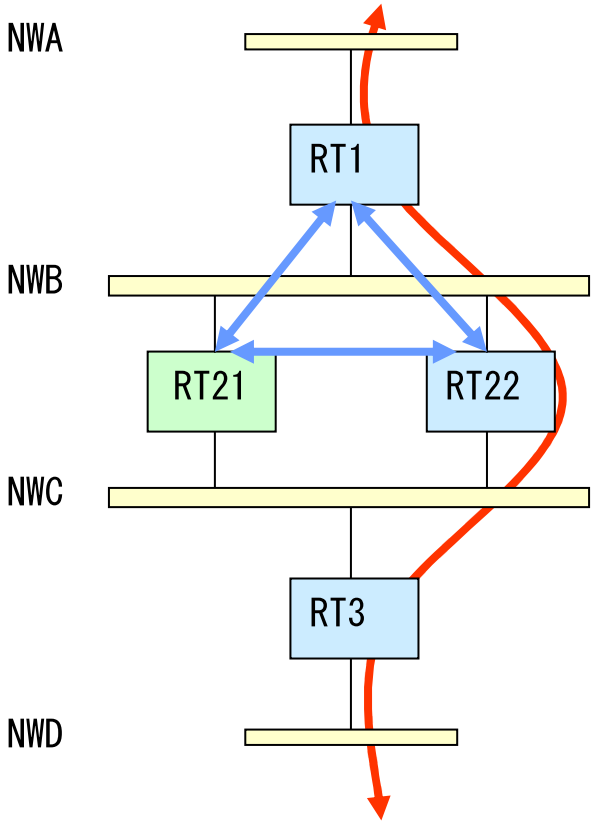
1. 問題の概要
2. OSPF
3. 問題発生のメカニズム
4. 解決策、回避策
5. ルータの実装改善提案

5. ルータの実装改善提案

本問題はそもそも OSPF の仕様の問題によるものである。
回避策の議論から見えてきた、以下の点について提案する。

1. point-to-multipoint mode の broadcast networks への適用
2. DR を graceful に譲る実装

５－１　point-to-multipoint mode の broadcast networks への適用

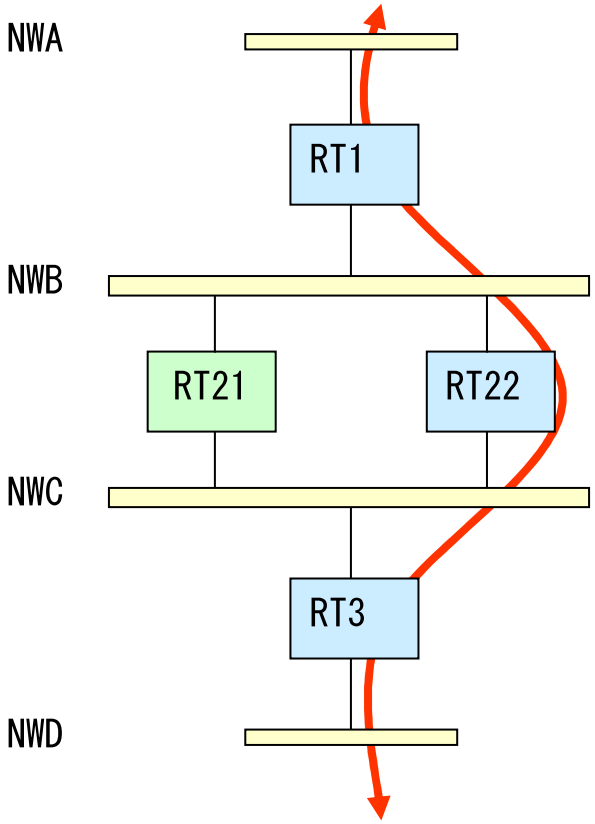


回避策案8の p-mp mode を LAN で動かすための共通仕様を定めることを提案する。

1. Hello は LAN の場合と同じく、自動で neighbor を検出し 2Way まで確立する（通常の LAN の場合と同じ）。
2. 2Way の neighbor とは全て p-p の mode と同様に adjacency を確立する（データベースを交換する）
3. タイマー値は LAN の場合と合わせる
4. IP MTU も LAN の場合と合わせる
5. 該当 LAN の情報を stub link として IP subnet mask の情報を含めて生成する

備考：
機器故障時においても、メンテナンス時においても本問題を回避可能となる。
また、ルータベンダの実装変更に要する工数もそれほど多くないと考える。

５－２　DR を graceful に譲る実装



回避策案3で DR を譲る際に、本問題を回避できる実装とできない実装があることを説明したが、以下の実装とすれば回避できる。

0. 管理者が Router Priority を 0 に変更
1. Hello Packet の Router Priority を 0 として送信
（DR の選出アルゴリズムが起動され、BDR が DR に昇格する）
2. Router LSA の該当 link について DR が現在の BDR が昇格するものとして Link ID を更新
3. Network LSA については flush しない。

備考：
Network LSA を flush しないことにより、ネットワーク内に古い情報が残ってしまうが、残ったとしても最大 MaxAge の時間が経過すれば消えるので、インパクトは小さいと考える。
また、OSPF の標準を変えずに実装の変更のみで対応可能な範囲と考える。