LAN　Switch技術
～冗長化手法と最新技術～

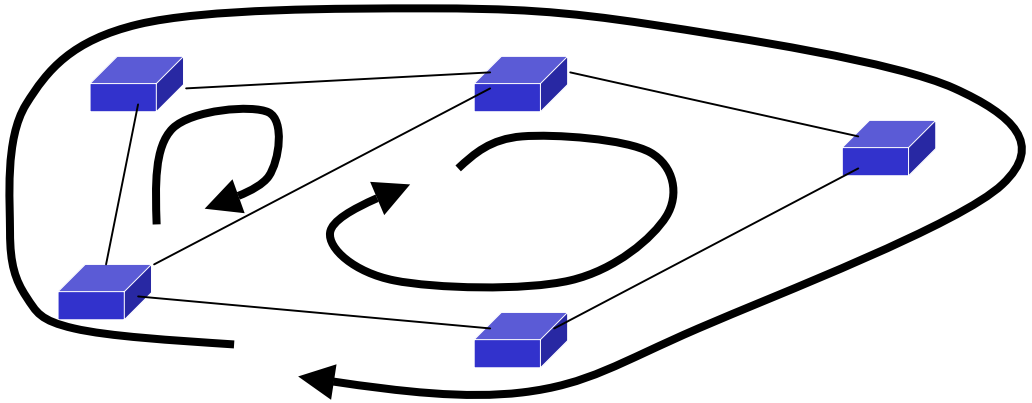
Version 2.0c

POWEREDCOM, Inc.

安藤　雅人

Ethernetにおけるループ発生の発生と弊害

- リンクやノードの故障の影響を防ぐ為に、イーサネットスイッチを冗長を持たせて接続すると、ループ部分ができてしまう。



- 何故ループが駄目か？

（1）FDB（Forwarding Data Base=MAC学習テーブル）が狂う。

ユニキャスト通信が出来なくなる。

（2）フレームが増殖する。

帯域が圧迫される

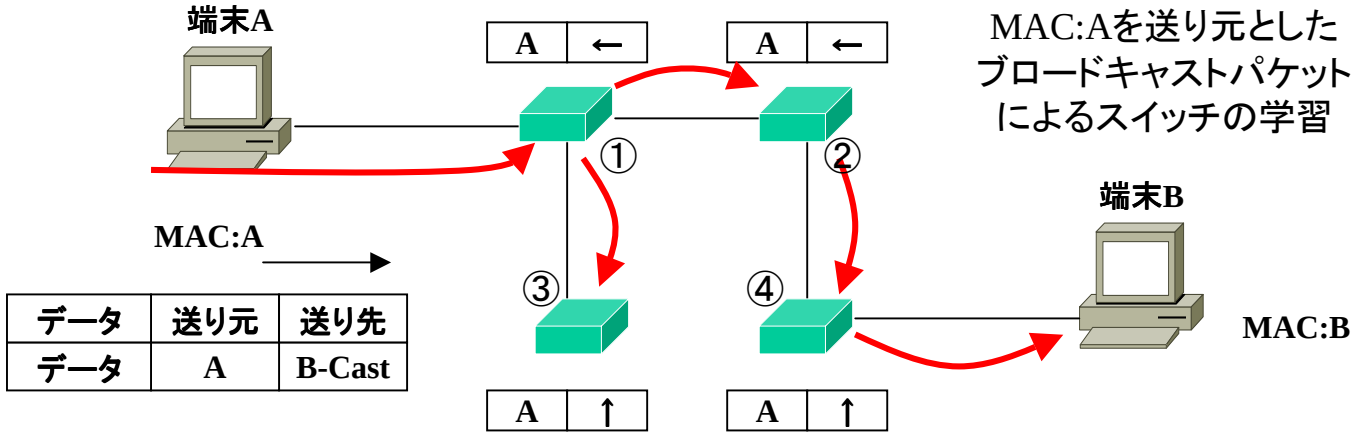
アプリケーションへの悪影響（システムダウンする場合もある）

→絶対にループは発生させてはならない！（一瞬でも）

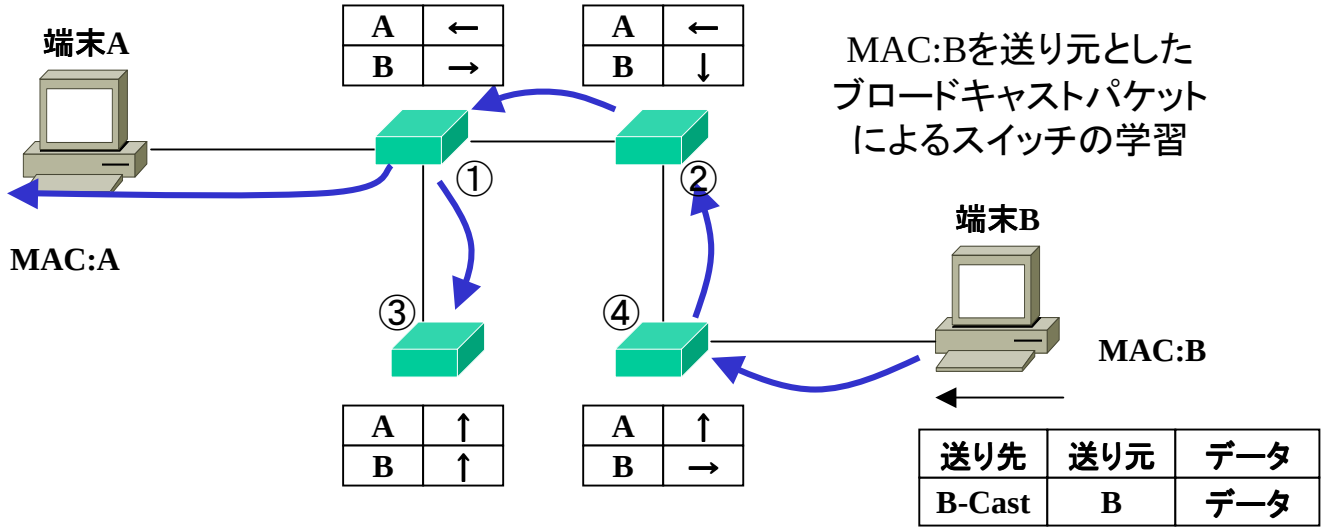
冗長を組みながら、ループを防止する方法を最新技術を交え紹介

正常時のスイッチ網のFDB学習状況

通常時の学習



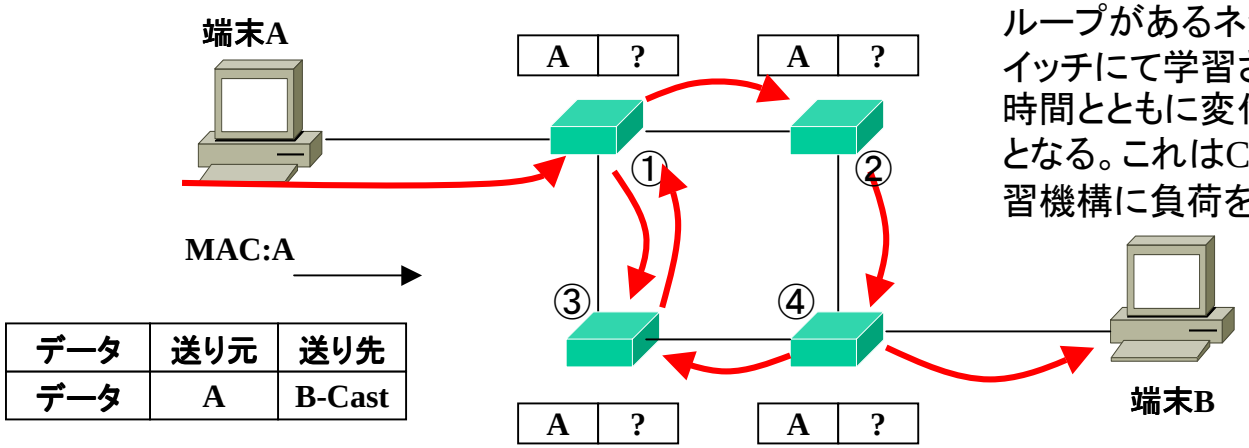
データ	送り元	送り先
データ	A	B-Cast



送り先	送り元	データ
B-Cast	B	データ

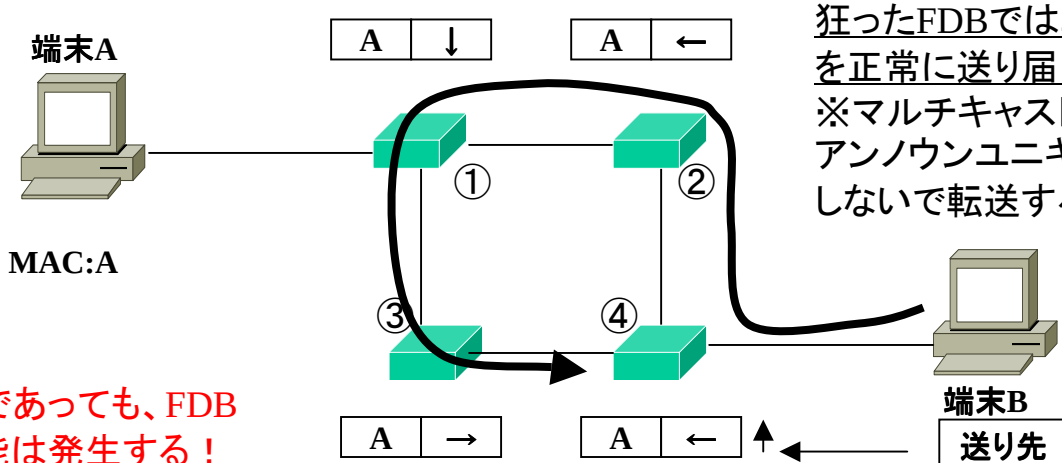
ループ発生時のFDB

ループ発生時の学習



データ	送り元	送り先
データ	A	B-Cast

ループがあるネットワークでは、スイッチにて学習されている内容は時間とともに変化する、狂ったFDBとなる。これはCスイッチのFDB学習機構に負荷を与える事にもなる。



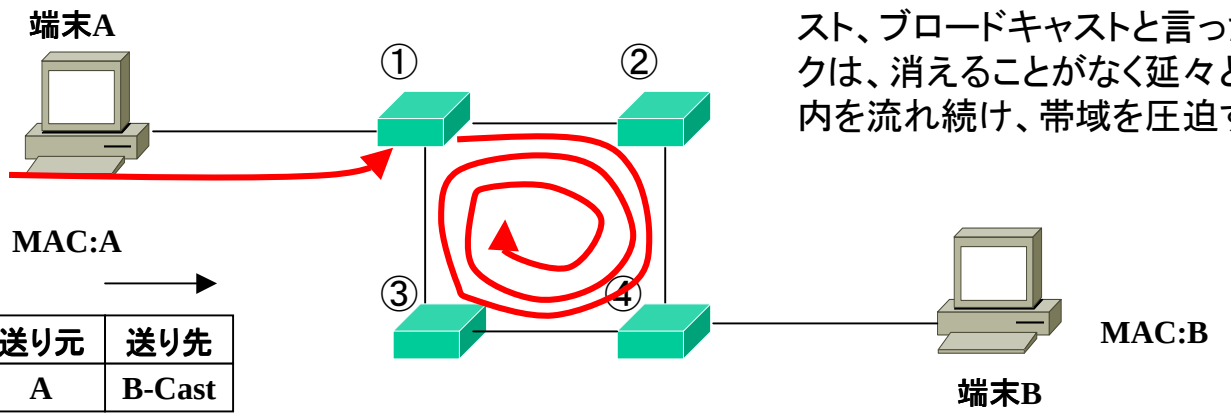
送り先	送り元	データ
A	B	データ

狂ったFDBではユニキャストフレームを正常に送り届ける事は出来ない。
※マルチキャスト、ブロードキャスト、アンノウンユニキャストはFDBを参照しないで転送する為届くことは届く。

瞬間的なループであっても、FDBが狂い、通信不能は発生する！

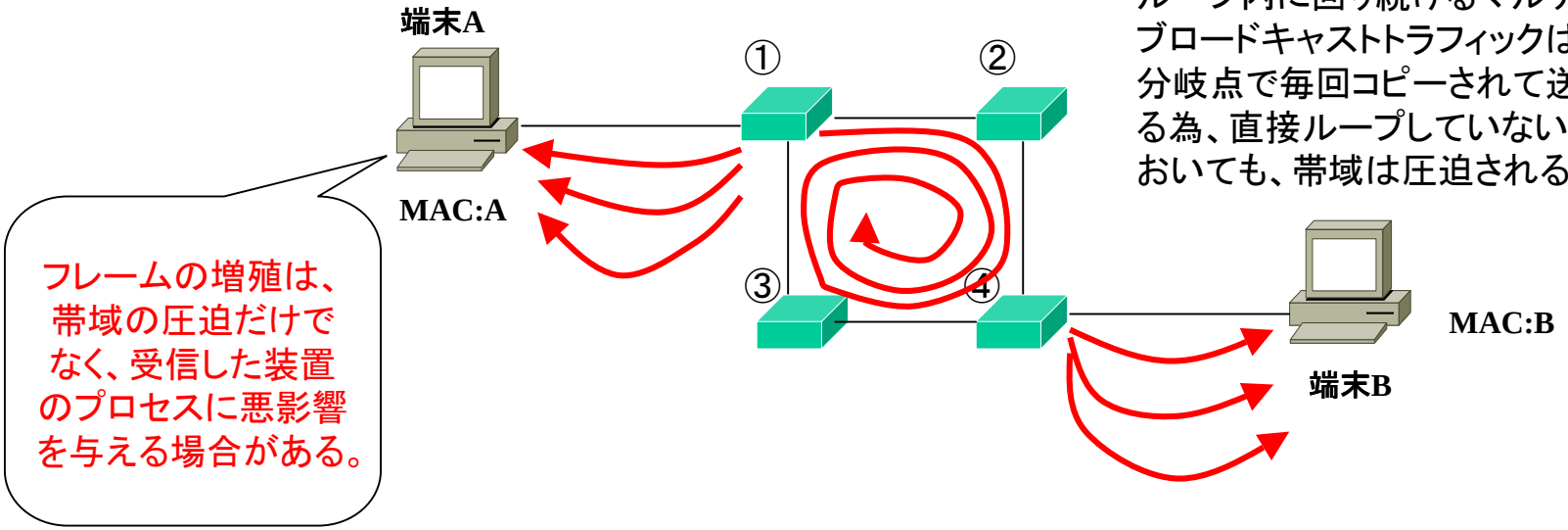
ループによる帯域圧迫

ループによる増殖と帯域圧迫



データ	送り元	送り先
データ	A	B-Cast

ループ発生時において、マルチキャスト、ブロードキャストと言ったトラフィックは、消えることがなく延々とループ内を流れ続け、帯域を圧迫する。

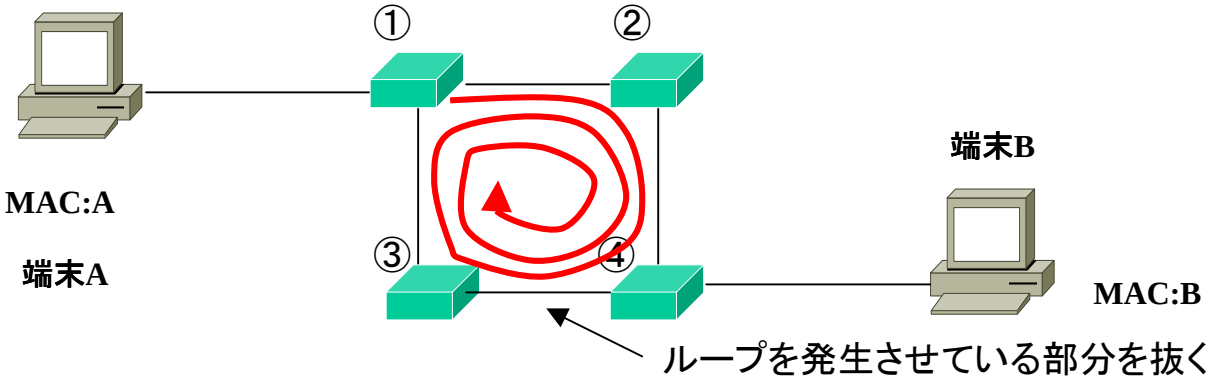


ループ内に回り続けるマルチキャスト、ブロードキャストトラフィックはノードの分岐点で毎回コピーされて送出される為、直接ループしていない箇所においても、帯域は圧迫される。

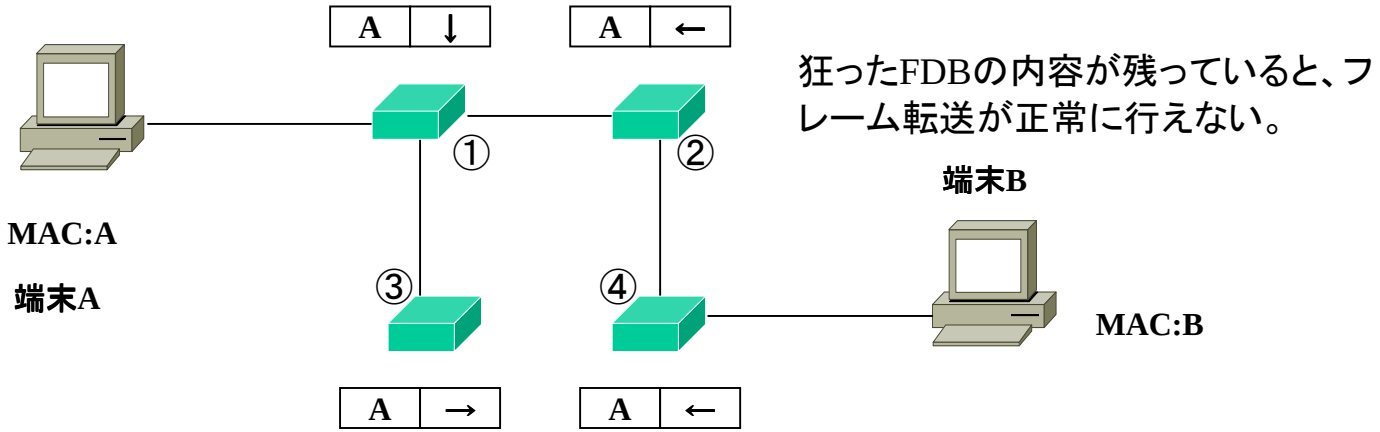
フレームの増殖は、帯域の圧迫だけでなく、受信した装置のプロセスに悪影響を与える場合がある。

ループが発生したら何をすべきか？（手動の場合）

(1)ループを構成している部分のリンクを切断したり、スイッチの転送を止める



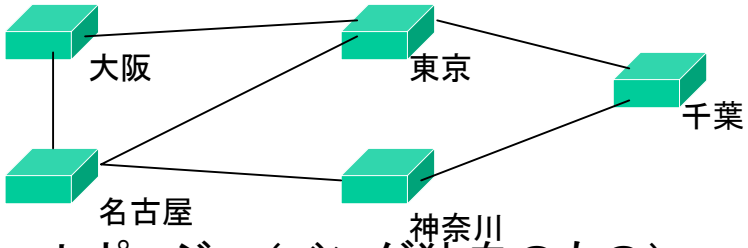
(2)FDBの内容を一度フラッシュする。（フラッシュしなければ、FDBのエージアウトを待つか、内容が上書きされるのを待たなければ正常な通信が行えない）



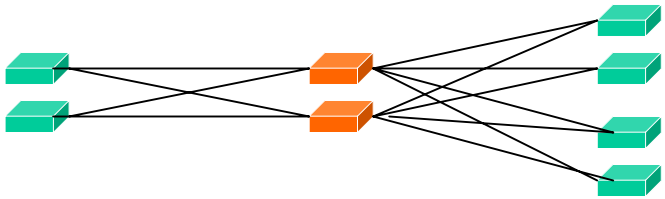
※何もしなければFDBのエージアウトは5分で AgeOutするのが一般的。(設定によって変更可能な物が多い)

イーサネット網における冗長方式の分類

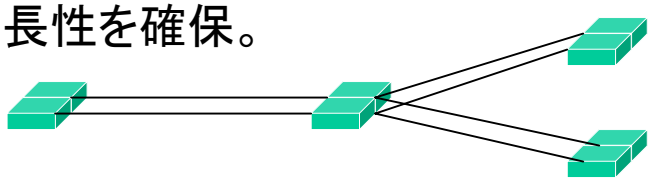
- ネットワーク冗長（STPファミリ、RPR、VPLS、リング型冗長など）
  ネットワークとして、冗長性を確保する。
  爆撃やテロなどに対しても比較的強い。



- ノード冗長化=メッシュトポロジー（ベンダ独自のもの）
  コアのスイッチの冗長化（二重化）
  装置の信頼性、伝送路の信頼性を補う為に主に用いられる。



- ノード内完全冗長+リンクアグリゲーション
  装置内部を完全に冗長化し、冗長単位で交換可能とする。伝送路はリンクアグリゲーションなどで冗長性を確保。



冗長なネットワークでのループを防止する三つの機構

1. ループフリーな論理トポロジーを維持する機構
   STPをはじめとする、様々な論理ネットワーク維持機構

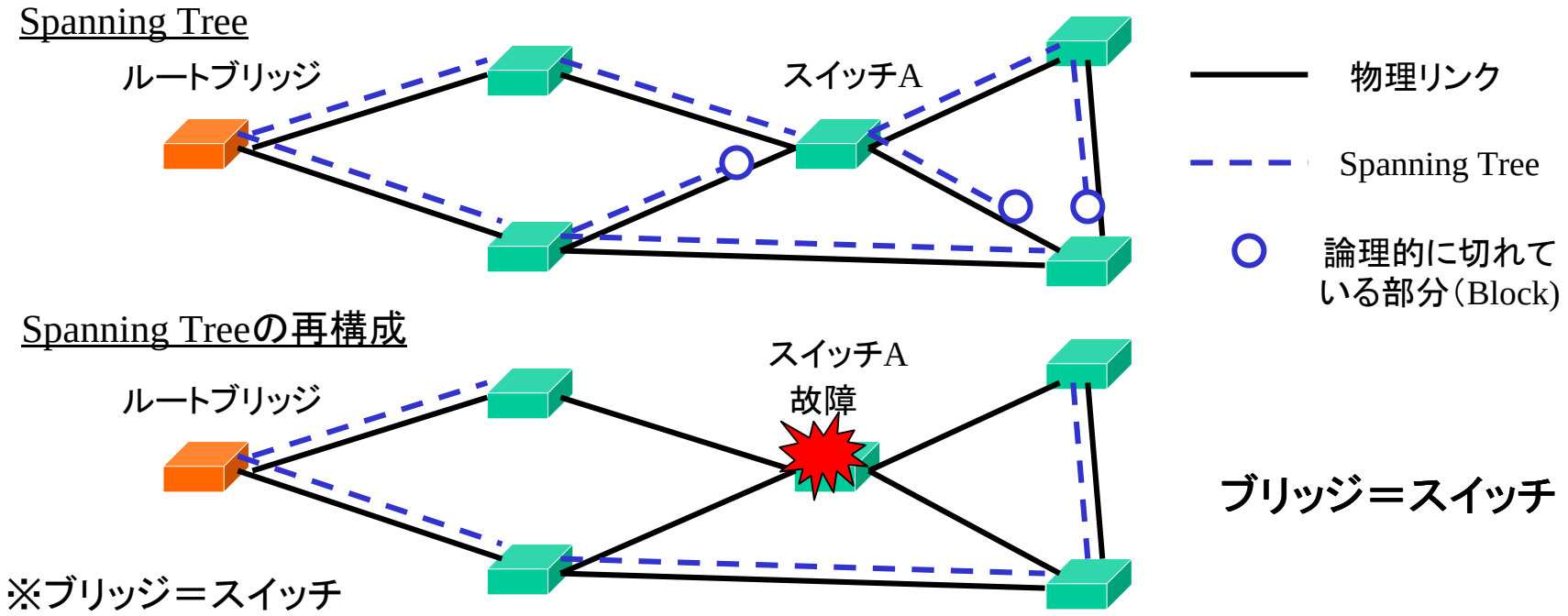
2. ループが発生した場合ループを検出し、論理トポロジーに働きかける機構
   ループの検出を行い、網のトポロジーに働きかける機構

3. ループしたフレームを検出し、フレームを破棄する機構
   TTLを利用したフレーム破棄、フィルタリング

Spanning Tree Protocol
ループフリーな論理トポロジーを維持する機構

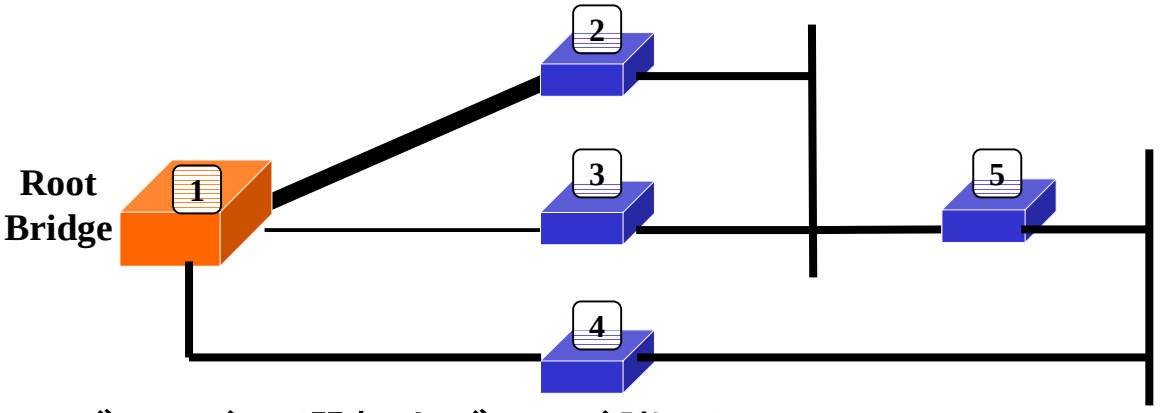
STP（Spanning Tree Protocol）IEEE 802.1D

- ループフリーな論理トポロジーを維持する機構の一つ。
- IEEE 802.1D標準
- 中心となるスイッチはルートブリッジと呼ばれそこから木のように枝分かれしていくので、Spanning Tree (広がる木）と呼ばれる。
- 木は、枝分かれはするが、一度分かれた枝が先で再度くっつく事は（普通）ないので、ループが発生するようなトポロジーとならない。
- 利用中のリンクが断したり、ノードが停止したら、論理トポロジーを自動的に再構成



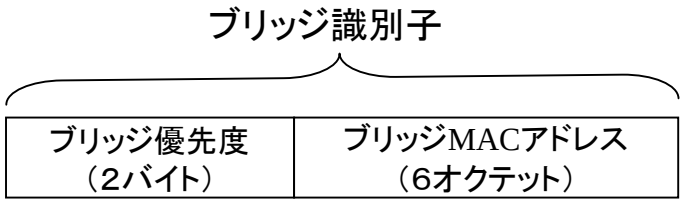
STPでのトポロジーの構築

- ルートブリッジの選択
  - トポロジーの中に1台だけ存在出来る
  - ルートブリッジを中心に木構造を作るのがSpanning Tree Protocol



BID(ブリッジ識別子)の数字は簡略化した表現とした。

- ルートブリッジの選択とブリッジ識別子
  - ブリッジはそれぞれ固有のブリッジ識別子を持ちその値のもっとも小さいものがルートブリッジになる。
  - ブリッジ識別子は、2オクテットのブリッジ優先度とブリッジMACアドレス（6オクテット）をつなげたものになる。



※ブリッジ優先度はデフォルトでは、32768(0x8000)となっています。

STPでのトポロジーの構築（リンクコスト）

- リンクコスト
  - スイッチには様々な速度(10M 100M 1G 10G等)のポートが存在する。
  - ポートの速度に応じてリンクのコストを付ける。
  - Bridge IDとともにSpanning Treeのトポロジー決定の重要な要素の一つ。
  - 網内でポリシーをそろえる事が重要（ポリシーが揃っていないと障害時の切り分け作業が大変）
  - かつては、リンクコスト=1000/速度(Mps)とされていたが、メディアの高速化に伴い、IEEE802.1[IEEE98a]では推奨コストを変更している（16bit ショート法）
  - さらに、IEEE802.1tでは、さらなる高速化に対応した推奨コストを提示している（32bitロング法）
  - 802.1tでは、Link Aggregationを組む場合には、速度によるリンクコスト/Link Aggregationのコストとなる。
    - GbE(1Gbps)のコストを20.000とすると、2本のGbEで組まれたLink Aggregationのリンクコストは、10.000となる。

STPでのトポロジーの構築（リンクコスト表）

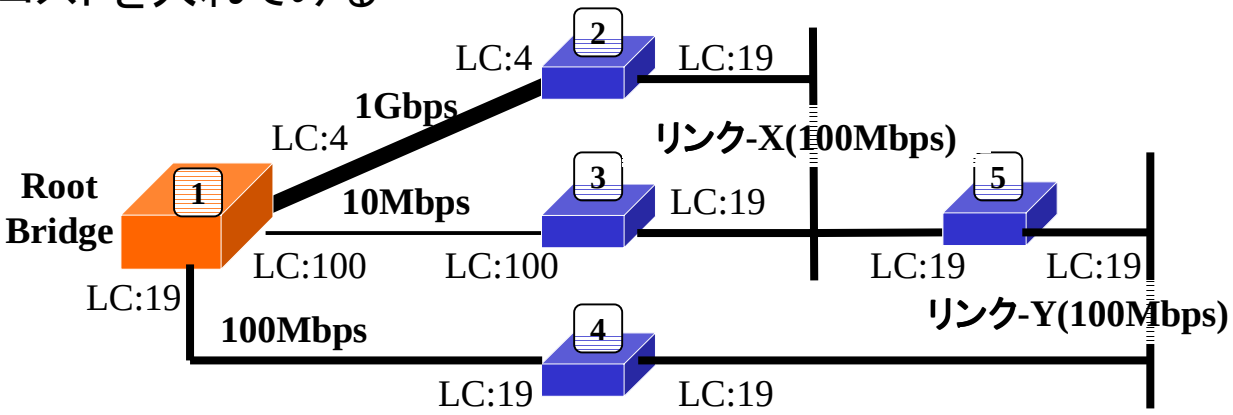
リンクコストの推奨値

データレート	IEEE 802.1D 98a（ショート法） 推奨リンクコスト範囲（推奨値）	IEEE802.1t（ロング法） 推奨リンクコスト範囲（推奨値）
4Mbps	100～1000(250)	
10Mbps	50～600(100)	200.000-20.000.000(2.000.000)
16Mbps	40～400(62)	
100Mbps	10～60(19)	20.000-2.000.000(200.000)
1Gbps	3～10(4)	2.000-200.000(20.000)
10Gbps	1～5(2)	200-20.000(2.000)
100Gbps		20-2.000(200)
1Tbps		2-200(20)
10Tbps		1-20(2)

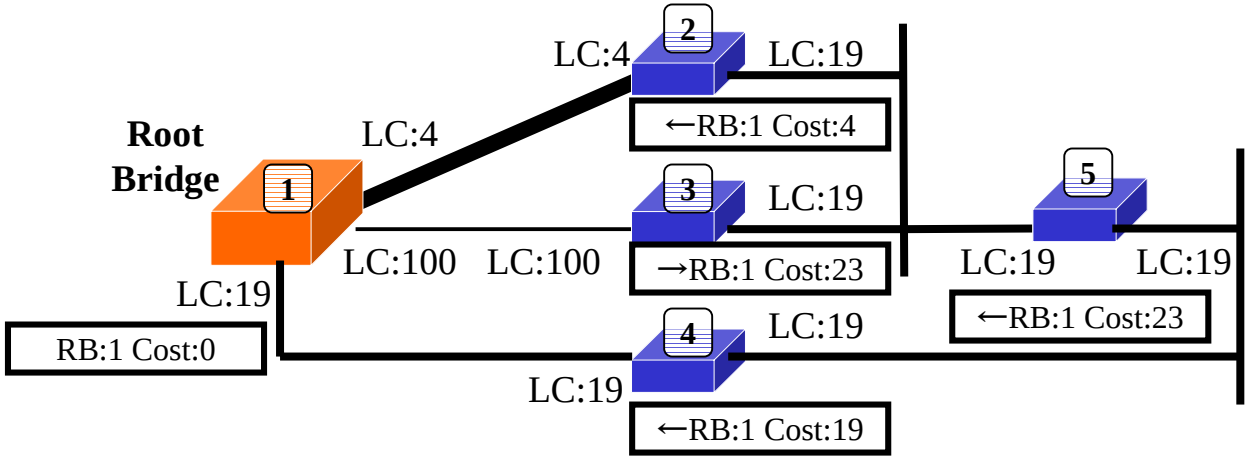
STPでのトポロジーの構築

コスト計算

- リンクコストを入れてみる



- スイッチにてリンクパスコスト（ルートブリッジへの最短距離）と方向を計算



STP　BPDU

- BPDU（Bridge Protocol Data Unit）
  - Spanning Treeの作成維持管理を行う。
  - スイッチ（ブリッジ）間で通常2秒間隔で送信、交換される。
- 二種類のBPDU
  - Configuration BPDU（Hello パケットとも呼ばれる）
  - Topology Change Notification BPDU

宛先MAC 01:80:C2:00:00:00
LLC　TYPE1
DSAP,SSAP 0x42

Configuration BPDU（トポロジー構築に使用）

Protocol ID=0000h	2
Protocol Version ID=00h	1
BPDU Type=00000000b	1
Flags	1
Root ID	8
Root Path Cost	4
Bridge ID	8
Port ID	2
Message Age	2
Max Age	2
Hello Time	2
Forward Delay	2

Topology Change Notification BPDU
(トポロジー変更時に使用）

Protocol ID=0000h	2
Protocol Version ID=00h	1
BPDU Type=10000000b	1

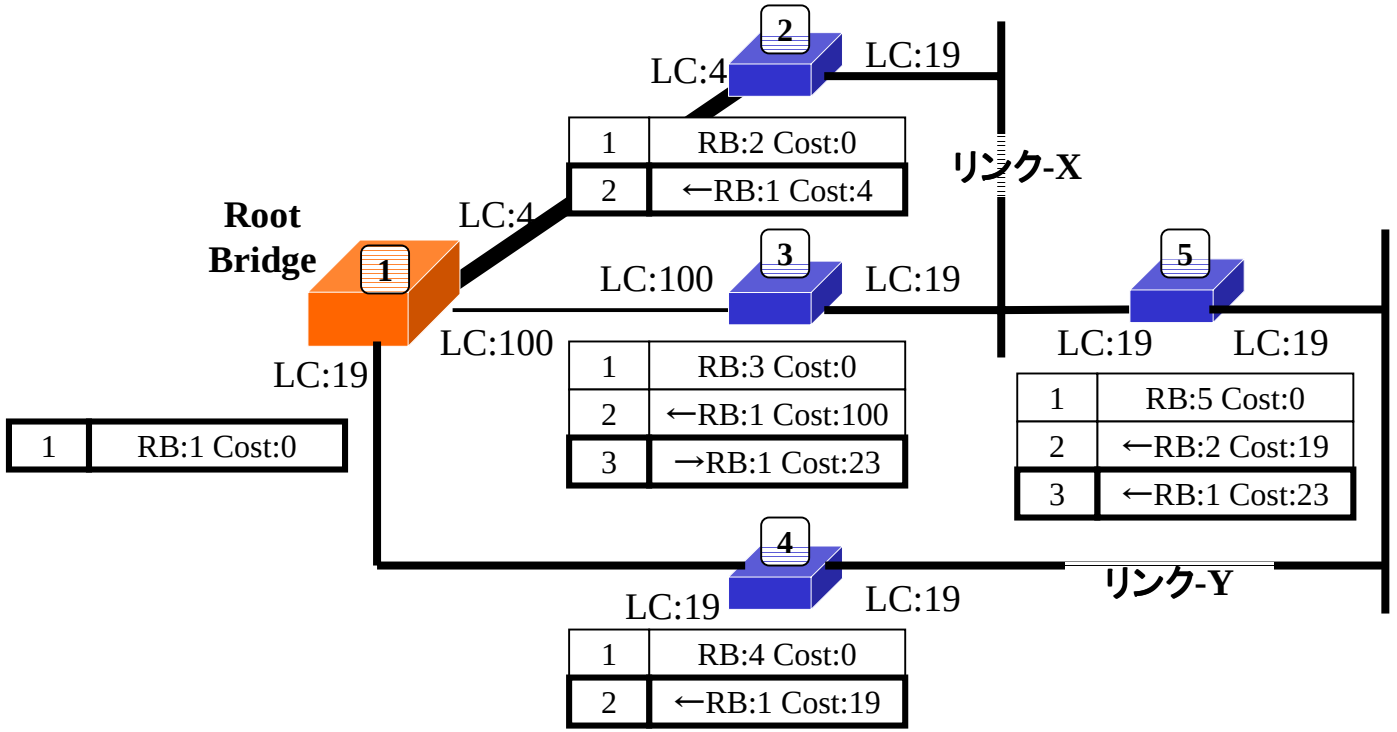
発信するスイッチが知りえる最もブリッジIDが小さいスイッチのID（Root Bridgeと思われるブリッジID)

Root BridgeまでのPathコスト

Flagには、Topology Change FlagとTopology Change Acknowledgement Flagの二種類が定義されている。

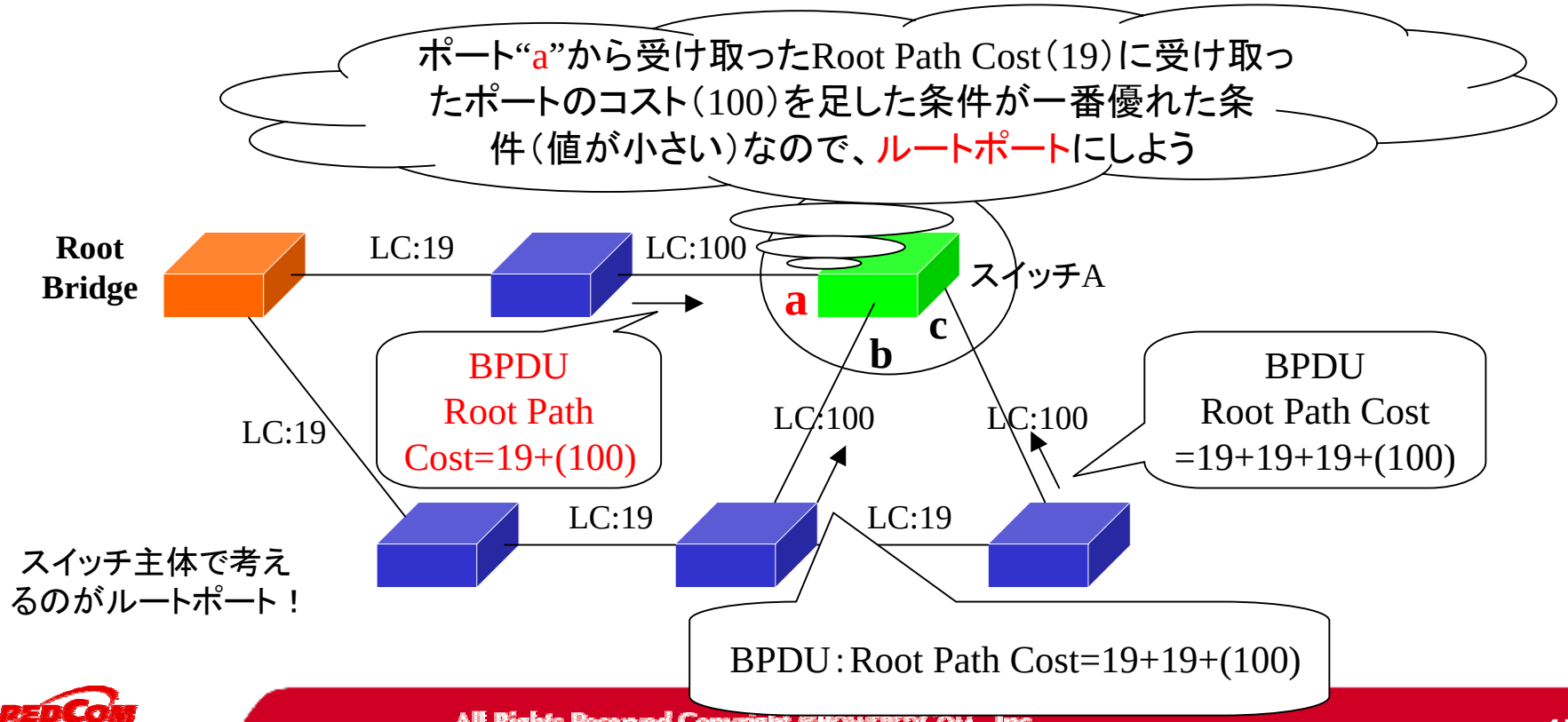
STPでのBPDUの交換

- Spanning Treeの構築はConfiguration BPDUの交換によって行われる。
- 各スイッチは、自身が知っている最も条件のいいルートへのコストをBPDUを使って広報する。
- 自身が知っているよりもより条件の良いコストを示すBPDUを受け取ったら受け取ったポートのコストを足して自身が採用するとともに、他のスイッチへその情報を広報する。



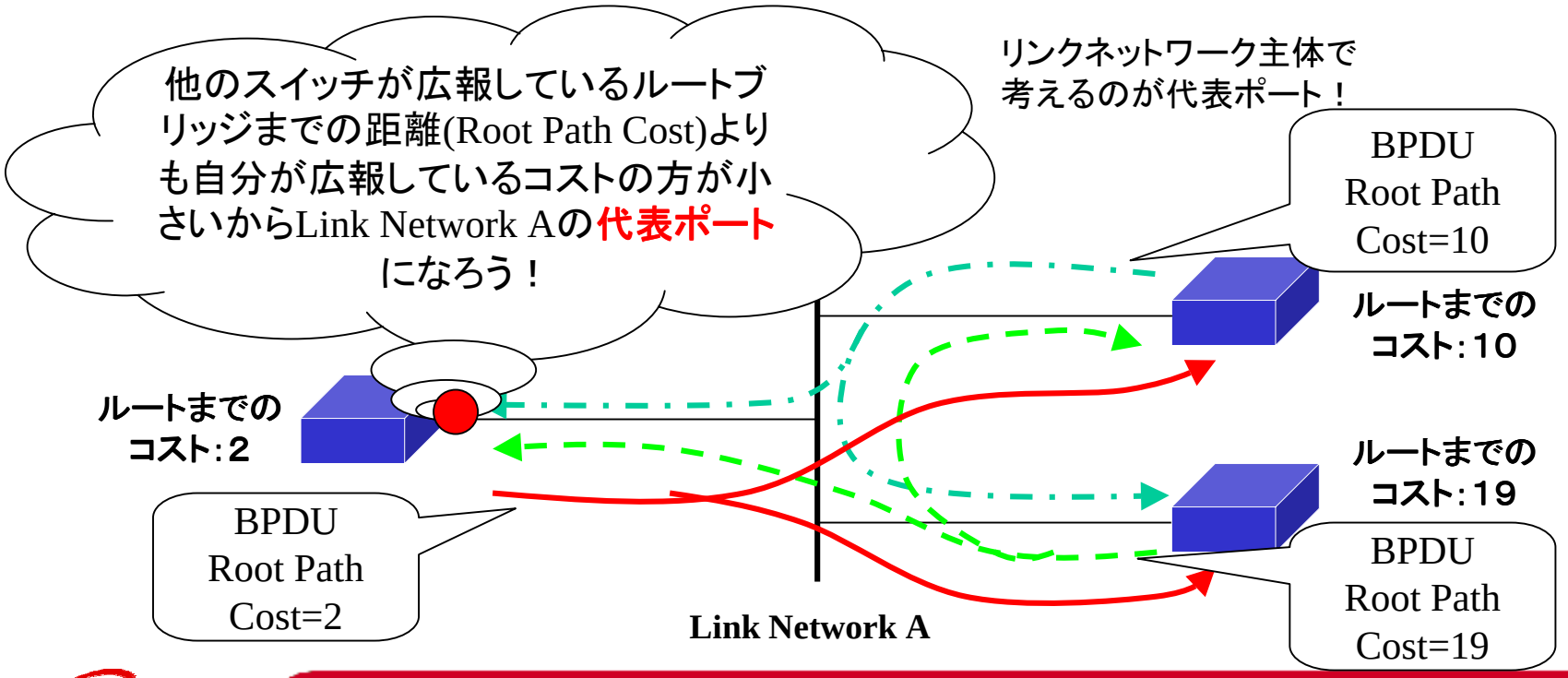
STPでのトポロジーの構築（ポートの役割の決定）

- ルートポート（Root Port）
  - スイッチからルートブリッジに到達するのに利用出来るポートを1つだけにする事によりループを防ぐと言う方法をSTPは取っているが、このポートの事をルートポートと言う。
  - スイッチの中で最もルートブリッジに少ないパスコスト（リンクコストの合計）で行けるポートがルートポートとなる。
  - スイッチが受信したBPDUにその受け取ったポートのコストを足した結果スイッチの中で一番コストが小さくなったポートが一つルートポートとして選択される。



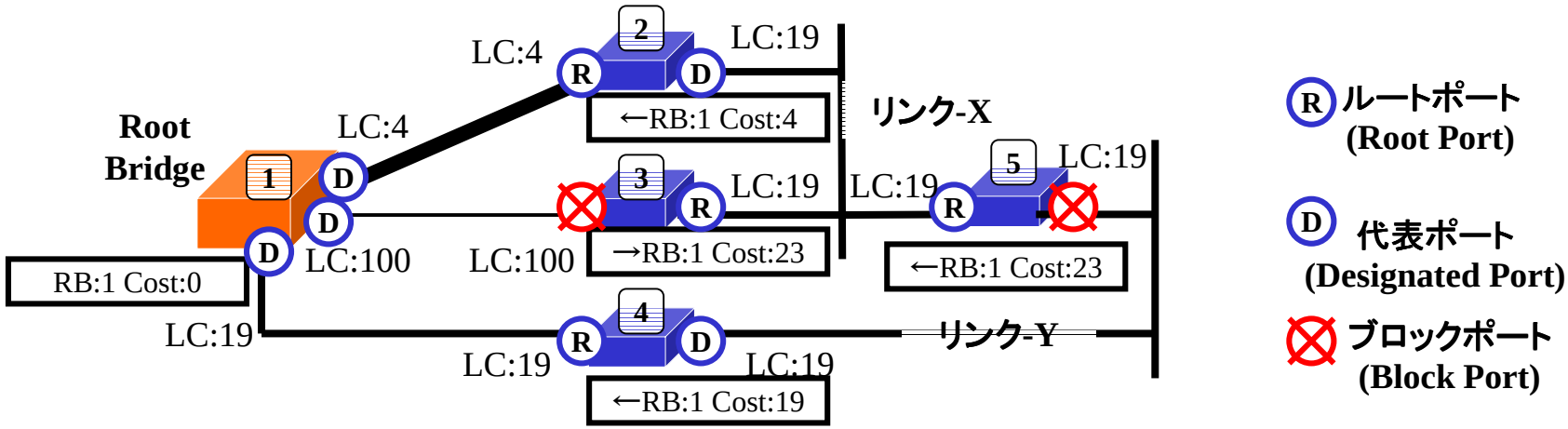
STPでのトポロジーの構築（ポートの役割の決定）

- 代表ポート(Designated Port)
  - リンクからルートブリッジに到達するのに利用出来るポートを1つだけにする事によりループを防ぐと言う方法をSTPは取っているが、このポートの事を代表ポートと言う。
  - ルートブリッジのポートは全て代表ポートとなる。（ルートブリッジ以外の代表ポートを持つブリッジを代表ブリッジ=Designated Bridgeと呼ぶ）
  - リンクに流れているBPDUをそのポートにて観察しそのブリッジが持っているコストより優れたものが流れていない場合に、そのリンクからルートブリッジに到達するのにそのポートが一番有利な条件と考え、そのポートを代表ポートとする。

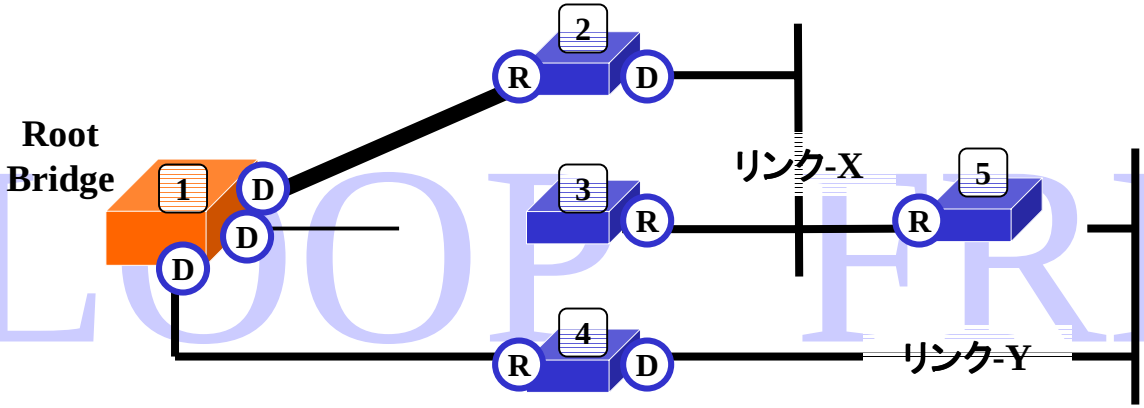


STPでのトポロジーの構築（ポートの役割の決定）

- ルートポートでも代表ポートでもないポートはブロッキング状態（非転送）になる。

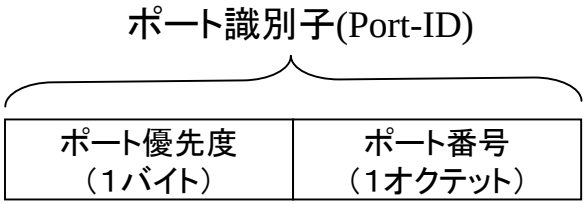


- ルートポートと代表ポートだけがフォワーディング状態（転送）になる。
- フォワーディング状態にあるポートとリンクだけを繋ぐとループのない木構造になっている。

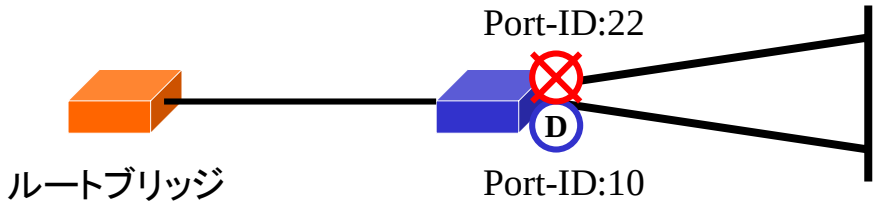


STPでのトポロジー構築

- ポート識別子
  - ブリッジ内のポートはそれぞれシャーシ内で固有のポート識別子を持つ。
  - ポート識別子は、1オクテットのポート優先度と1オクテットポート番号をつなげたものになる。
  - 代表ポートの選択がパスコストの比較だけではつかなかった時にこの数値が低い方のポートが代表ポートになる。



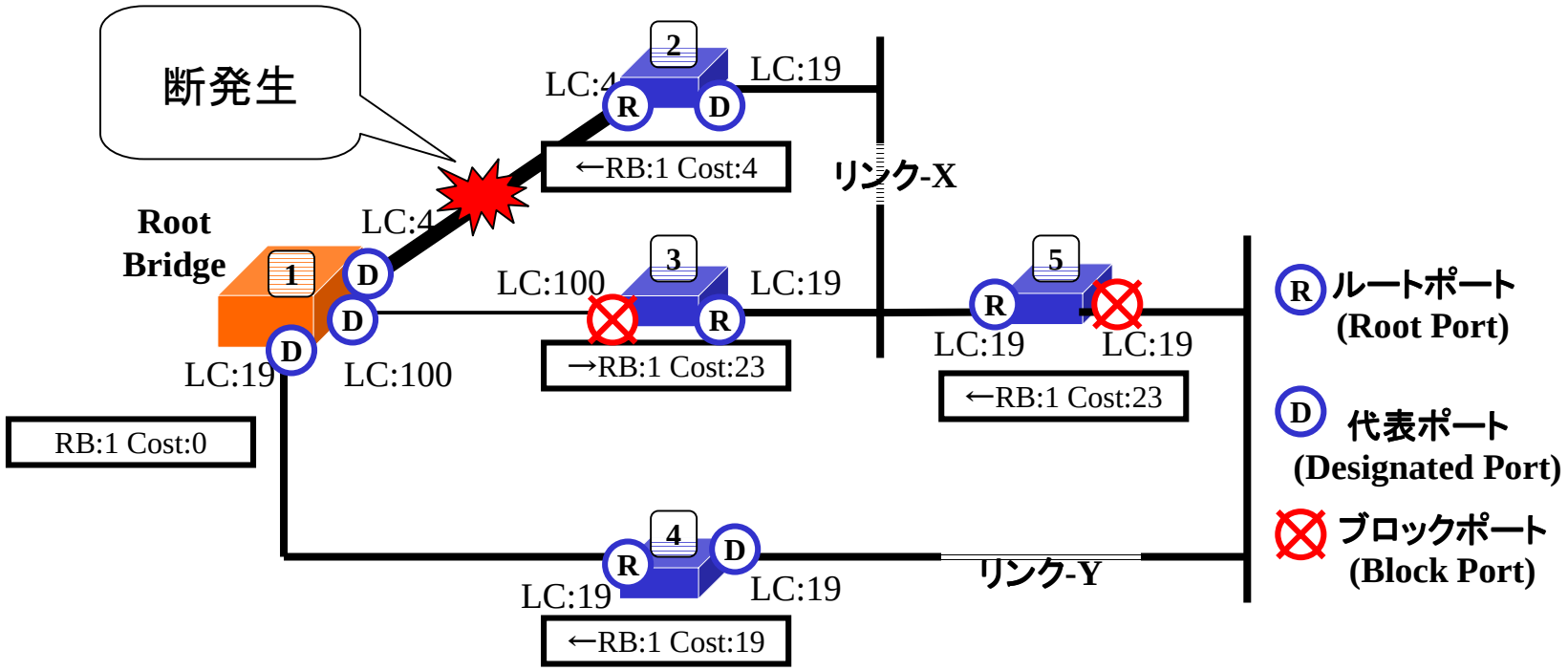
※ポート優先度はデフォルトでは、
128(0x80)となっている。



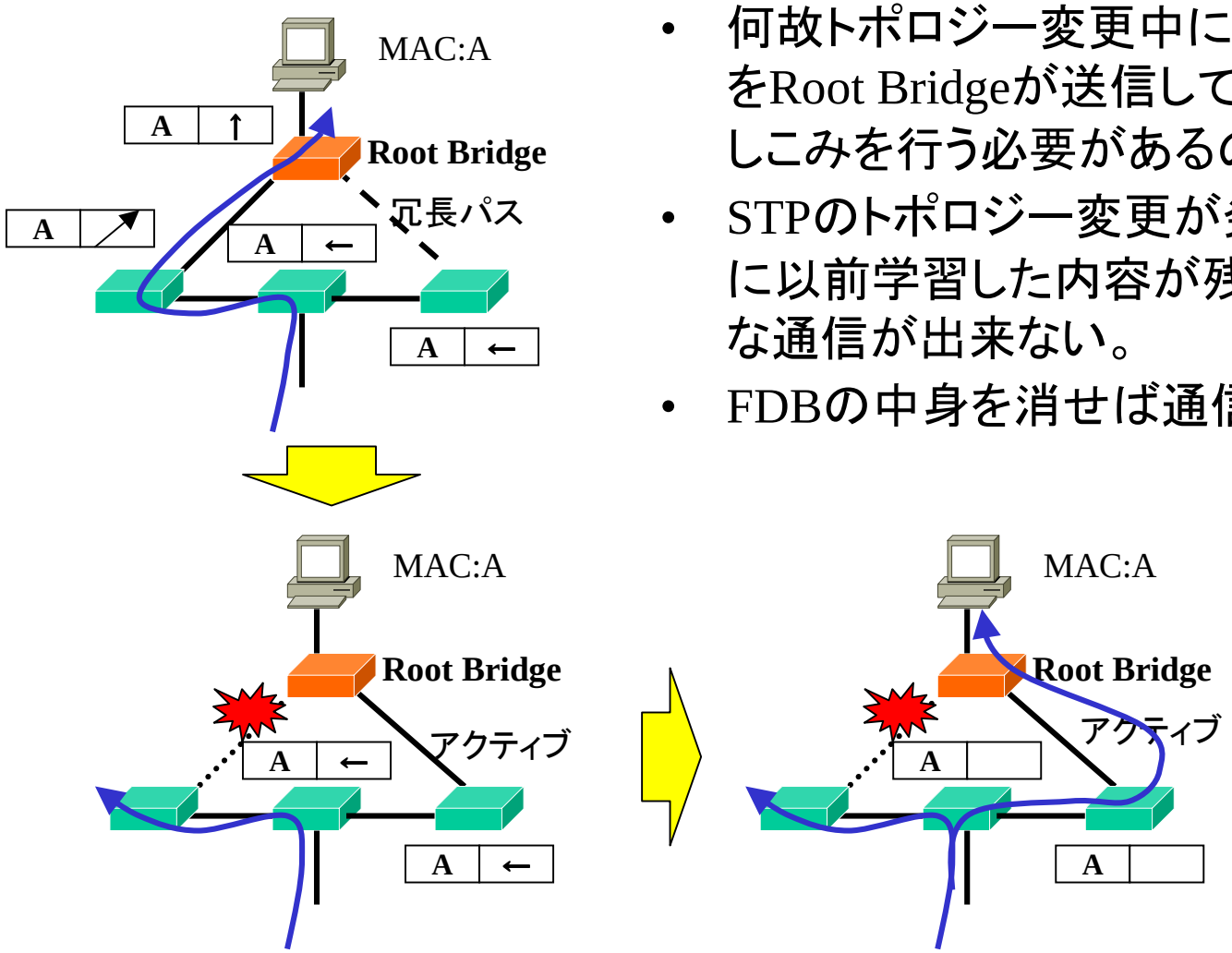
※代表ポートの選択でパスコストが同じ場合（同シャーシの場合など）Port-IDが小さいほうが、代表ポートとなる。

STPでのトポロジー再構築

- 障害発生の場合
  - スイッチ2はルートポートの断を検出すると、Spanning Treeの再計算を開始する
  - ここでの計算の方法は通常のSTP構築と同じ



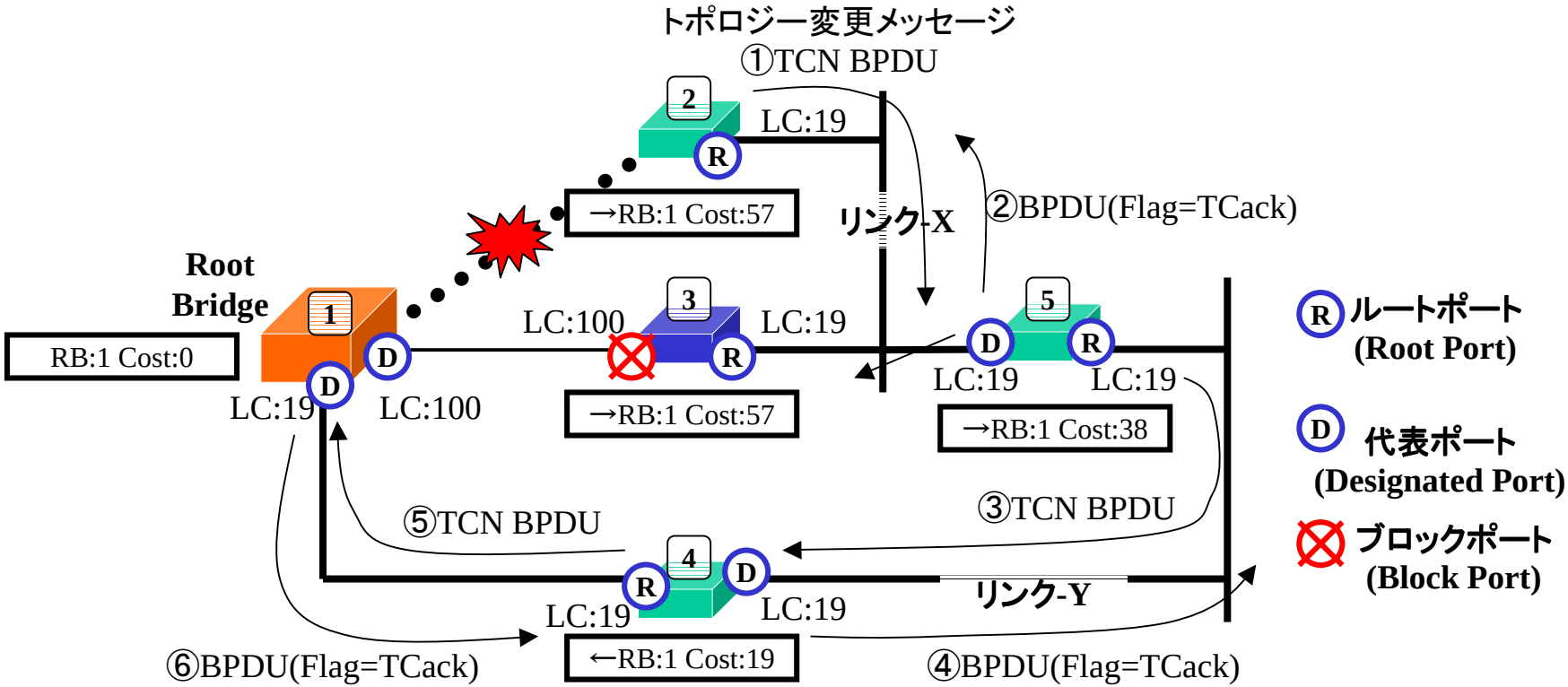
STPでのFDBフラッシュ（消しこみ）の必要性



- 何故トポロジー変更中にFlag=TCNのBPDUをRoot Bridgeが送信して、FDBの内容の消しこみを行う必要があるのか？
- STPのトポロジー変更が発生した後で、FDBに以前学習した内容が残っていると、正常な通信が出来ない。
- FDBの中身を消せば通信出来るようになる。

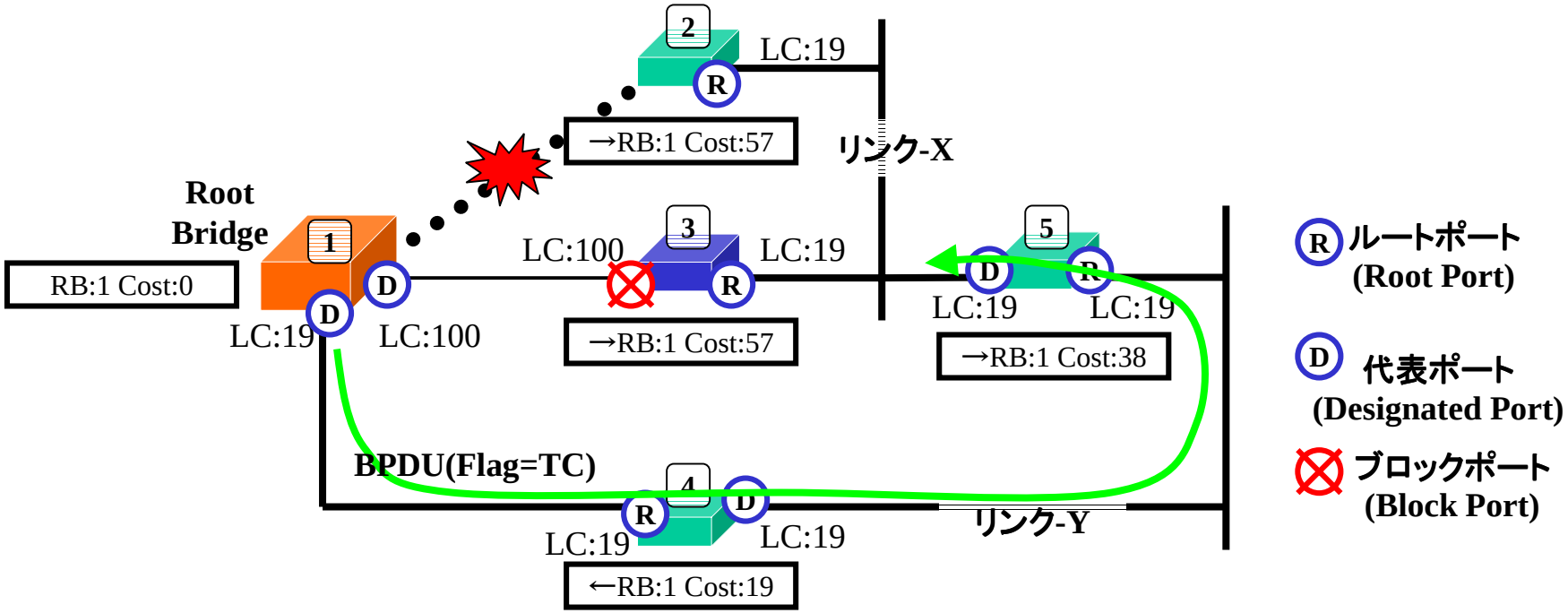
STPでのFDBフラッシュ

- STP構築以外に網全体のFDBをフラッシュする為の機構が必要となる。
- トポロジーの変更を検出したスイッチは、ルートポートより、Topology Change Notification BPDU(TCN BPDU)を送出する。
- 代表ポートはTCN-BPDUを受け取ると、BPDUのFlag=TCack（Topology Change Acknowledgement)を設定し送り返しTCN-BPDUを受信した事を通知し、さらに上位のスイッチにTCN-BPDUを送信する。



STPでのFDBフラッシュ

- TCN BPDUを受信した、ルートブリッジは、一定期の間、送信するConfiguration BPDUをFlag=TC(Topology Change)として送信し、全てのスイッチにトポロジー変更が発生した事を通知する。
- Flag=TCのBPDUを受信したスイッチはフラグが設定されている期間中、FDBの中身をより短い時間でAge Outするようにする。
  - Aging Time（一般には5分）を、Forward Delay（Default=15秒）に変更する事により、速やかにFDBを忘れさせる。



STPでのポート状態

- STPでブロッキングポートが、ルートポートや、代表ポートに変更されたとしてもすぐに転送状態（Forwarding）とはならない。
  - ループ防止
  - 無駄なフラッディングを防ぐ

ルートポートであるとか代表ポートであるとかいったポートの役割とは別に、ポートにはいくつかの状態がある。

- DISABLED状態
  - シャットダウンされているか、電源の入ってない状態。（このポートは使えない）
- BLOCKING状態
  - データフレームの転送を行わない
  - ルートポートや代表ポートになっていないポートはこの状態に落ち着く
  - BPDUの送信は行わないが、BPDUの受信は行っており、その処理も行われる;
  - 電源投入時は全てのポートがこの状態
- LISTENING状態
  - データフレームの転送は行わない
  - BPDUの受信を行う状態、必要であればBPDUの送信も行う
  - Spanning Treeを構築中のスイッチはこの状態にある。

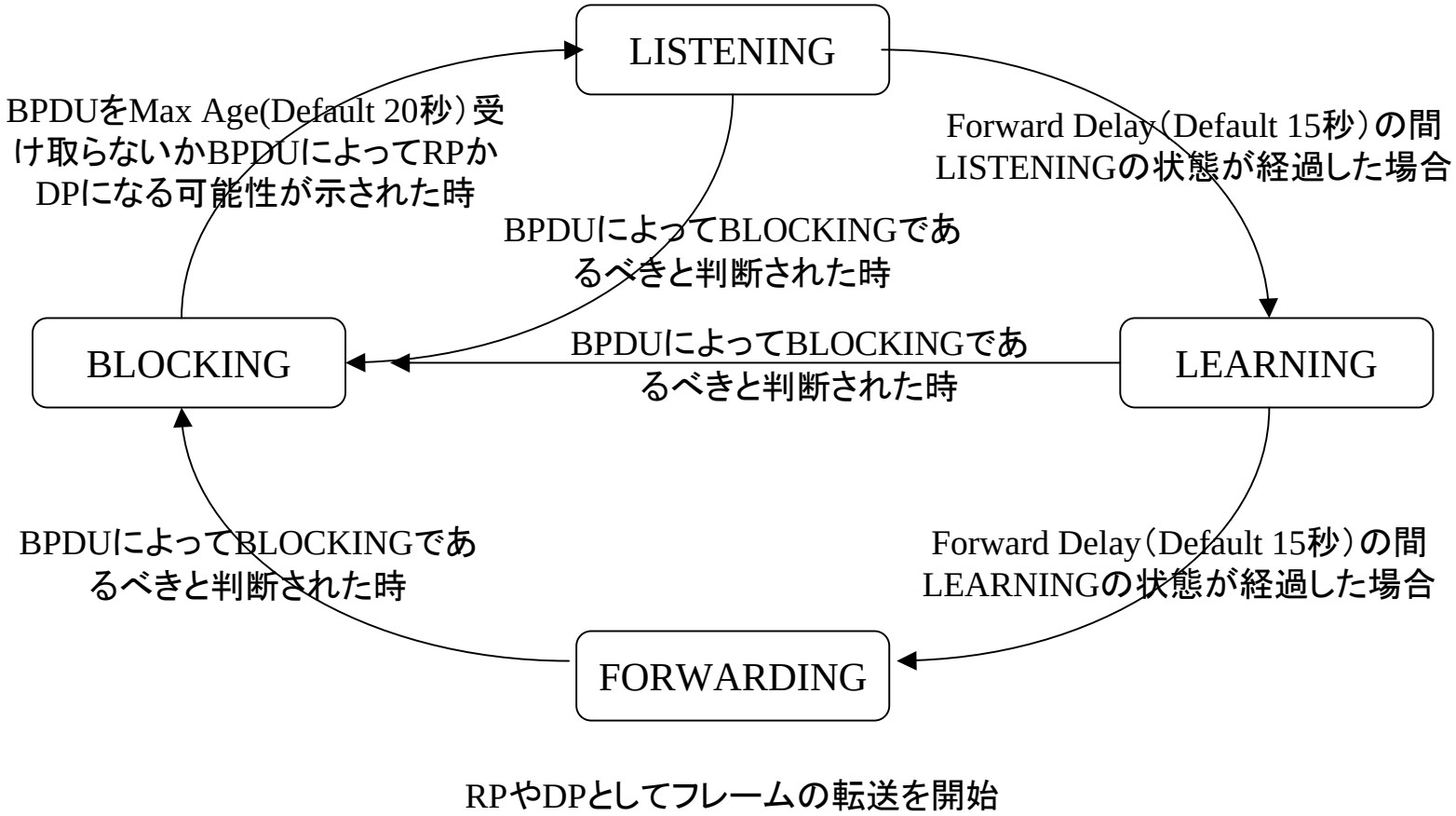
STPでのポート状態

- LEARNING状態
  - 転送を始める前はFDBの内容が空である為そのまま転送をはじめるとフラッディングが多発する。これをおさえる為、転送を開始する前に流れているフレームからFDBの内容の学習を行う。
- FORWARDING状態
  - 通常の転送状態。

	利用可能？	BPDU処理	MAC学習	データ転送
DISABLED状態	×	×	×	×
BLOCKING状態	○	△（受信のみ）	×	×
LISTENING状態	○	○	×	×
LEARNING状態	○	○	○	×
FORWARDING状態	○	○	○	○

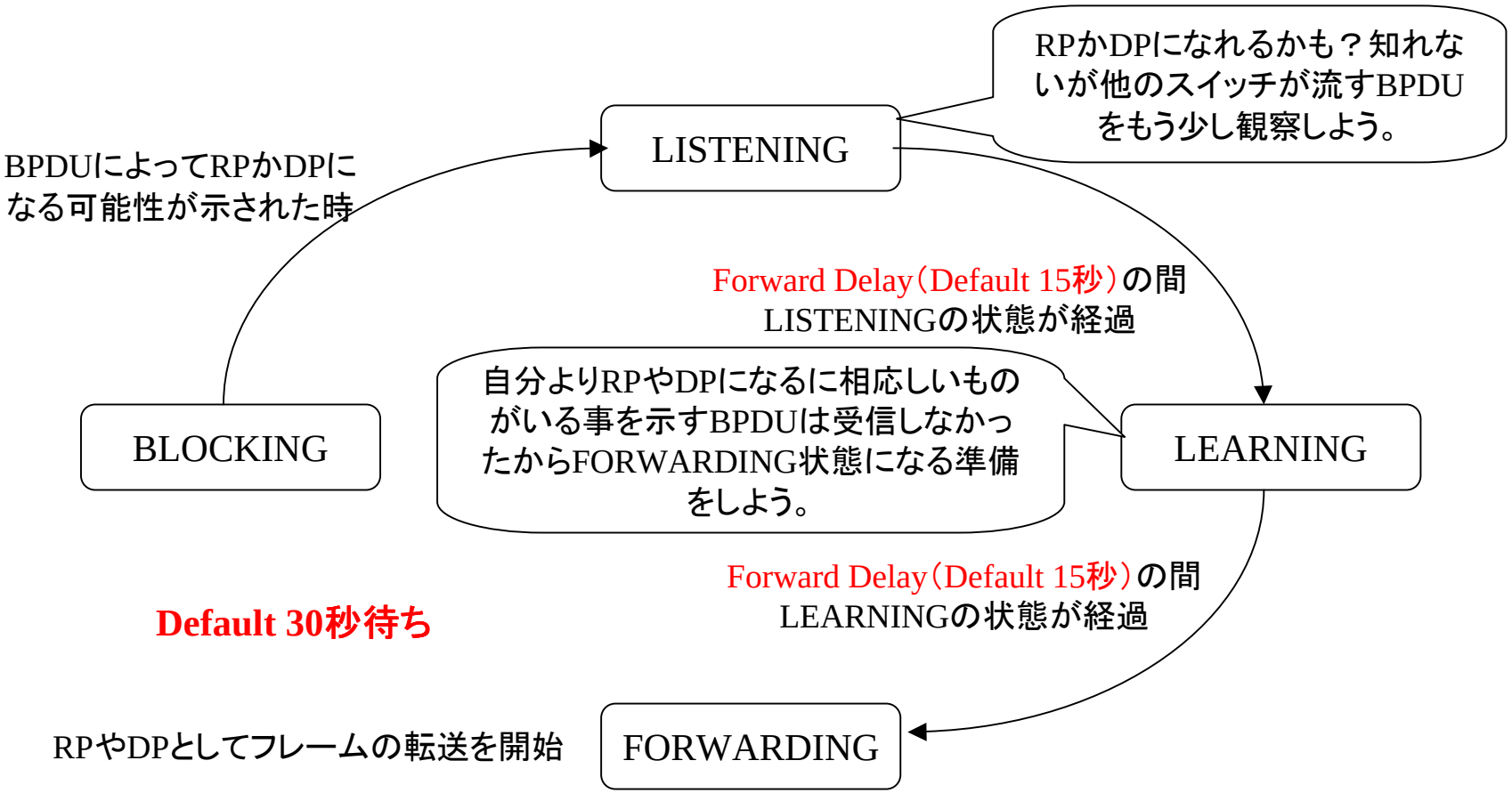
STPでのポート状態

- STPポート状態遷移



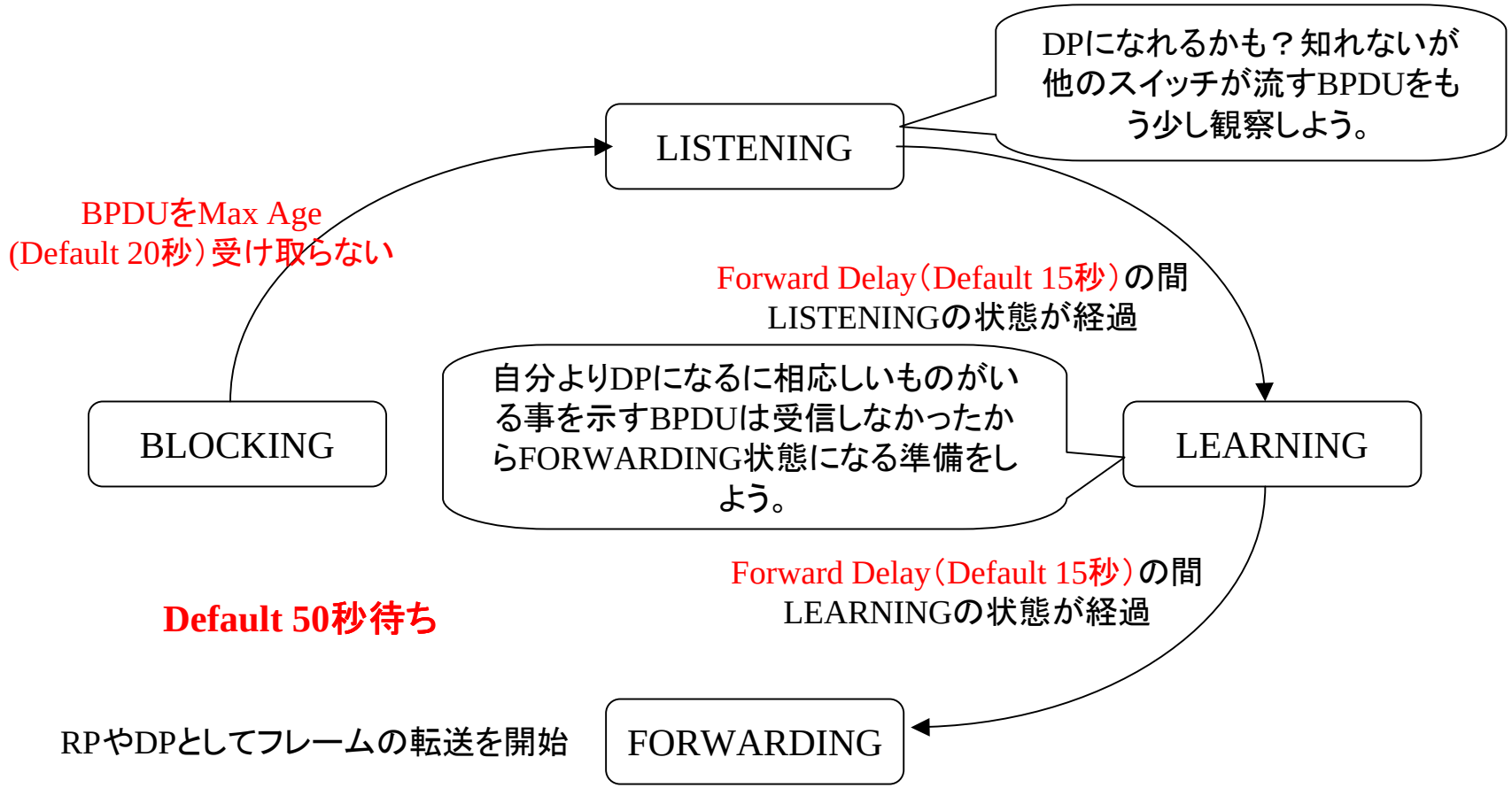
STPでのポート状態

- トポロジー変更により、BLOCKING状態であったポートがルートポート(RP)や代表ポート(DP)に変更され、転送状態になるにはForward Delay x 2待たなくてはならない。
- スイッチの起動時やポートをリンクアップさせてすぐの状態も同様。（スイッチにPCに付けてすぐに通信出来ないのはこれが原因の事もある。）



STPでのポート状態

- 上位のブリッジから、BPDUをMAX Age時間受け取らなかった場合は結果的にMAX Age + (Forward Delay x 2)待たなくては転送を開始出来ない。



STPパラメータの確認

パラメータ	説明	Default値 （設定可能範囲）
Hello time	BPDUの送信間隔	2秒（1-10）
Forward Delay	LISTENNINGやLEARNINGに使う時間	15秒（4-30）
MAX Age	BPDUがタイムアウトする時間	20秒（6-40）
Bridge Priority	スイッチの優先度、小さいほど、ルートブリッジになりやすい	32768（0-65535）
Port Cost	そのポートのコストを示します。小さいほど選択されやすくなる。	速度に応じて設定
Port Priority	ポート間でパスコストが同じだった場合に比較される値、小さいほど選択されやすくなる。	16(0-255)

STPのまとめ

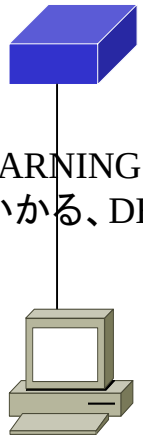
- 通常STPが止まるのは、LISTENING→LEARNING→FORWARDINGにかかる、Forward Delay x 2=Default 30秒の時間、ただし、最悪の場合、MAX Age + Forward Delay x 2=Default 50秒の時間止まる。Forward DelayとMAX Ageを設定変更する事も可能。それでも14秒(6+4x2)を切る事は出来ない。
- パラメータはルートブリッジに設定されたものが採用される。
  - 他のスイッチはルートブリッジが流すBPDUに記述されたパラメータを採用する。
- STPは特に何も設定しなくても動作するが最低限ルートブリッジとバックアップでルートブリッジになる装置がネットワークの適切な位置で適切な性能のスイッチになるように設計されなくてはならない。
- STPのパラメータはむやみに変更しない（障害時の解析が大変になる）。変更する時はきちっと設計を行い、全てのスイッチが同じポリシーで動作するようにする事。

- STPのメリット
  - 標準的なプロトコルであり、異ベンダ機器の相互接続が可能になっている。
  - 物理トポロジーを選ばない。

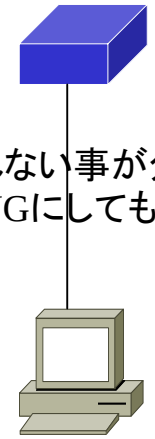
- STPのデメリット
  - 標準のSTPは切り替えに時間がかかる
  - トポロジー全体の事を考えて、オペレーションを行わなくてはならない。
  - PCなどをつなげる場合もForward Delay x 2の時間待たされる。

STPの拡張(Port fast)

- スイッチにPCなどを接続した後、LISTENING→LEARNING→FORWARDINGと状態変化する時間（通常30秒）待たなくてはならないのは使いにくい。
- スイッチが接続される可能性のないポートに関しては、事前に設定しておくことにより、いきなり、FORWARDINGになるようにしておく。
- BPDUをポートで受信した場合はポートをブロッキング状態にする。
- Ciscoで実装しているが多くのスイッチで同様の効果を得る設定は出来る。



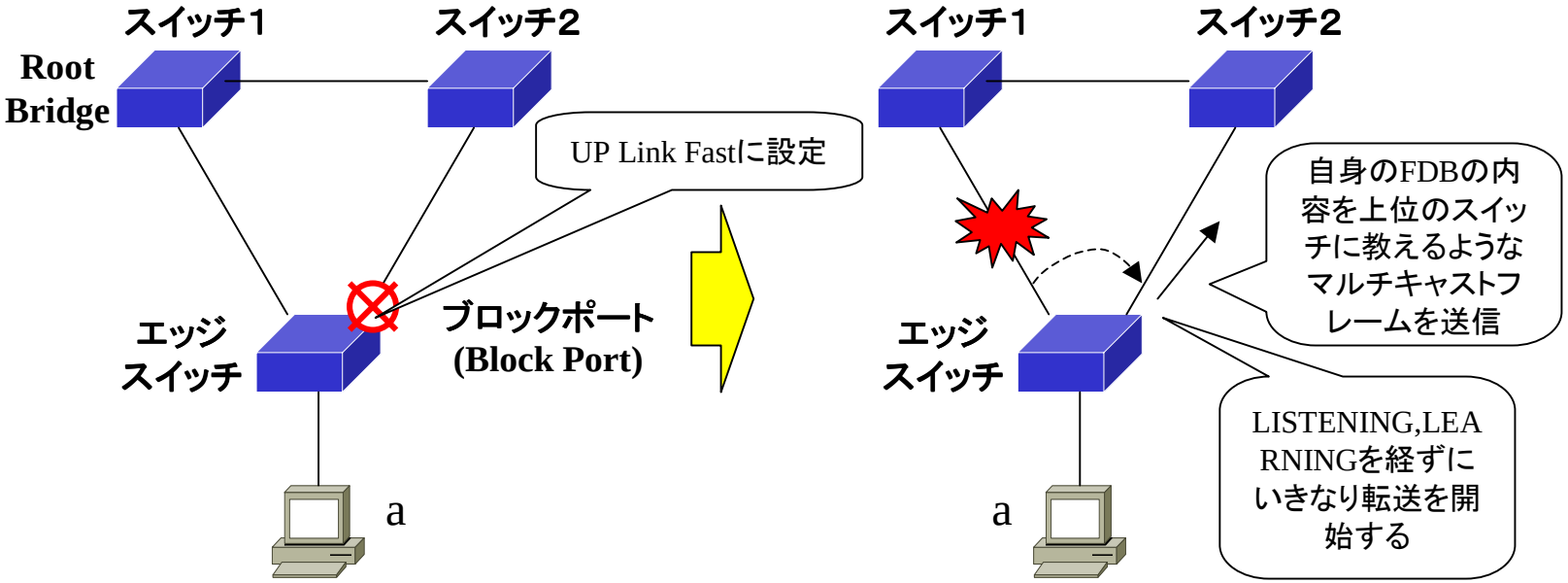
LISTENING→LEARNING→FORWARDING
と待つとDefault 30秒かかる、DHCPに失敗したりする。



スイッチが接続されない事が分かっているのなら、
いきなりFORWARDINGにしても問題ない。（すぐ使える）

STPの拡張(Uplink fast)

- CiscoによるSTPの拡張
- アクティブなリンクが断となった場合に、バックアップのリンクにすぐ切り替わる機能。（ LISTENING→LEARNING→FORWARDINGの状態変化にかかる時間をとばす）
- エッジに設置したスイッチが上位のスイッチに2本のリンクで接続されている場合にエッジに設定出来る
- 上位のスイッチのFDB構築を支援する為、Uplink fastでエッジスイッチが切り替わりを発生させた場合に、エッジスイッチは自身がFDB内に学習済みのアドレスを送り元アドレスとする、マルチキャストフレームを新しくアクティブにしたリンクに流す。

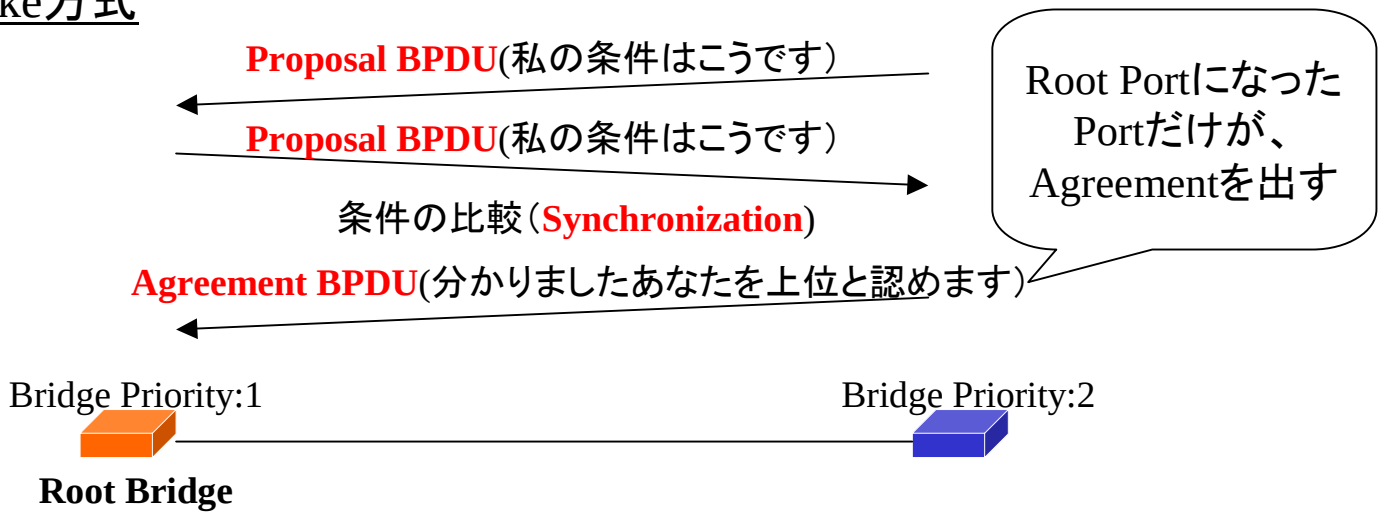


Rapid Spanning Tree Protocol
ループフリーな論理トポロジーを維持する機構
（STP高速化）

RSTP(Rapid Spanning Tree Protocol)802.1w

- IEEE 802.1w標準
- STPの切り替わり動作を高速化する為に作られた。(30秒や50秒かかるのは遅い)
- 構築される論理木構造はSTPと同じ(同じパラメータを使う)
- Max Age、Forward DelayのパラメータはBPDUを受信しないポートを代表ポートとする場合か、あるいはSTPと混在して使う時のみ有効
- 802.1wは802.1Dの上位互換性がある
- Point-to-Pointの接続が基本
- BPDUの交換にHandshake方式の導入
- ポートの役割の種類の追加

Handshake方式



RSTPにおけるポートの役割（Port Role）

ポートの役割	説明	定常状態
ルートポート (Root Port)	Rootブリッジへ最も少ないコストで到達出来る経路を提供するポート、STPと同じ	Forwarding
代表ポート (Designated Port)	リンクからRootブリッジへ最も少ないコストで到達出来る経路を提供するポート、STPと同じ	Forwarding
アルタネートポート (Alternate Port)	ルートポートに変わる二番目に少ないコストでルートブリッジに到達出来るルートブリッジへの経路を提供するポート。 Next Root Port	Blocking
バックアップポート (Backup Port)	指定ポートが提供するリンクへの経路に変わるリンクへの経路を提供するポート Next Designated Port	Blocking
ディスエーブルドポート (Disabled Port)	故障しているか、シャットダウンされているポート、STPと同じ	Disabled

RSTPで使用されるBPDU

- RSTP-BPDU
  - BPDU ver2として定義
  - Hello 間隔ごとにスイッチ間で交換（3Hello timeで過去の情報は無効になる）
  - Flagの部分を拡張
  - Topology Change Notification BPDUは使わない。(STPとのインターワークでのみ利用）

RSTP BPDU

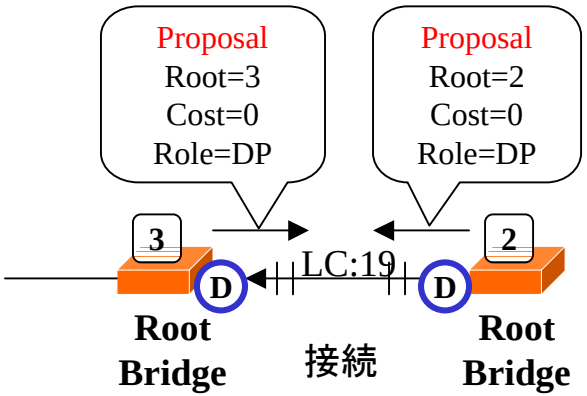
Protocol ID=0000h	2
Protocol Version ID=02h (2)	1
BPDU Type=0000 0010b (2)	1
Flags	1
Root ID	8
Root Path Cost	4
Bridge ID	8
Port ID	2
Message Age	2
Max Age	2
Hello Time	2
Forward Delay	2

FLAGを大きく拡張

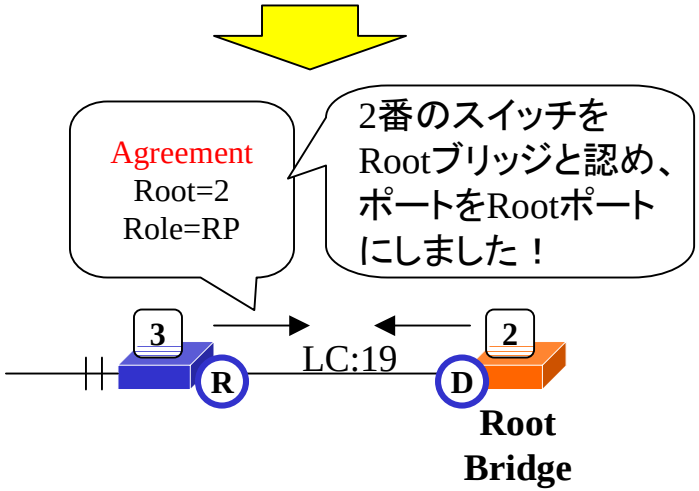
Bit位置	Flagの意味
0	Topology Change
1	Proposal
2-3	Port Role
00	Unknown Port
01	Alternate or Backup Port
10	Root Port
11	Designated Port
4	Learning
5	Forwarding
6	Agreement
7	Topology Change Ack

RSTPでのトポロジー構築

RSTPにおけるHandshake



条件比較（Synchronization）

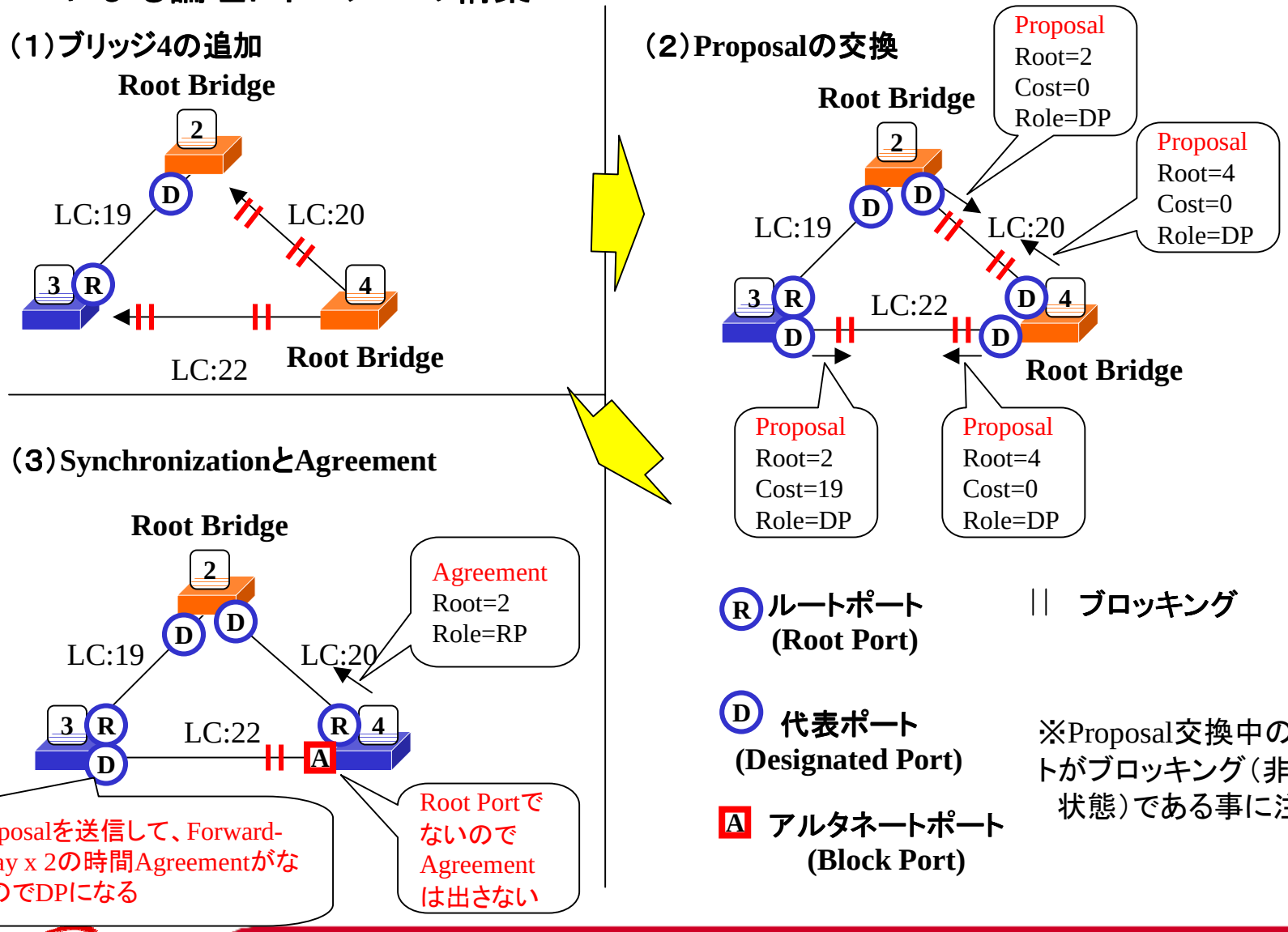


- 二つのスイッチが接続されると、ポートを代表ポートとし状態を非転送（ブロッキング）にした上で双方のスイッチが自身がRoot Bridgeであり、当該ポートが代表ポート(DP)であるとするProposal BPDUを送信する。

- Proposal BPDUを受信すると、自身の持つRoot情報及びコストと比較し。
  - 相手が勝る場合
    - 相手を代表ポート(DP)と認める
    - “Proposal BPDU”を受信したポートをRootポート（RP）として転送状態にし、その他のポートをブロック状態にする。
    - 相手に“Agreement BPDU”を送信する。
  - 自身が勝る場合
    - 相手から“Agreement BPDU”を受け取ると代表ポートとしてすぐに転送を開始する。

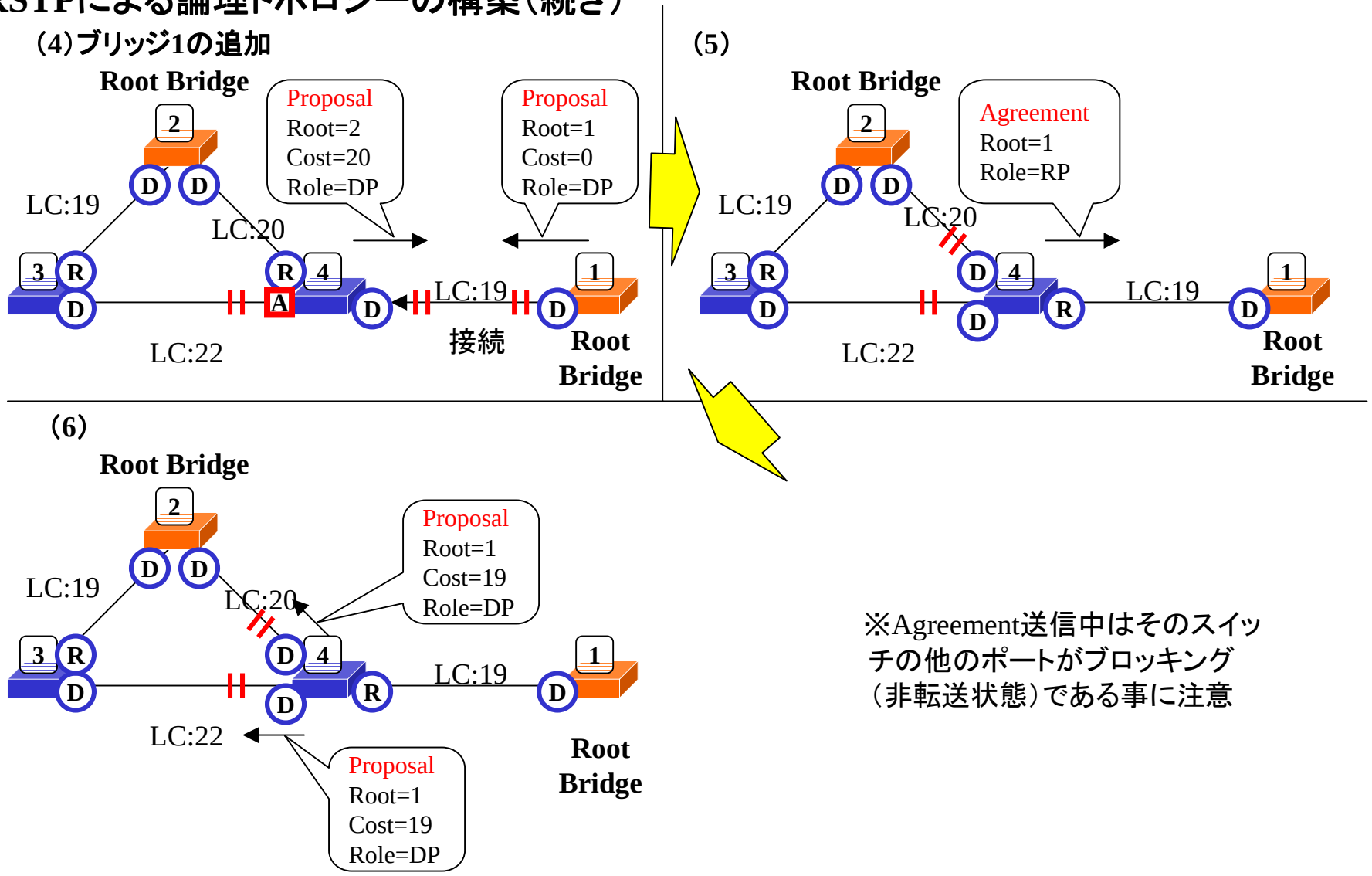
RSTPでのトポロジー構築

RSTPによる論理トポロジーの構築



RSTPでのトポロジー構築

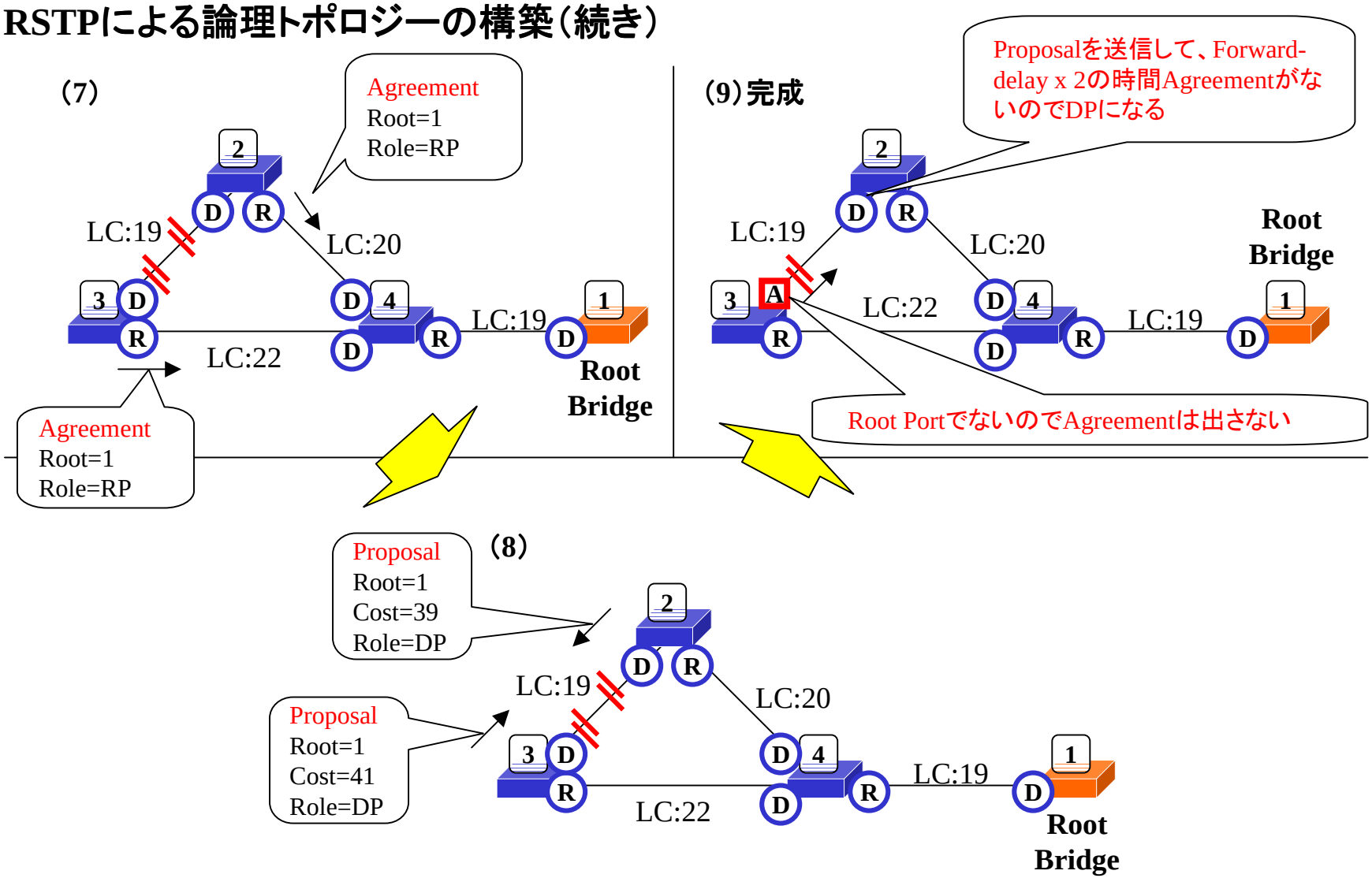
RSTPによる論理トポロジーの構築（続き）



※Agreement送信中はそのスイッチの他のポートがブロッキング（非転送状態）である事に注意

RSTPでのトポロジー構築

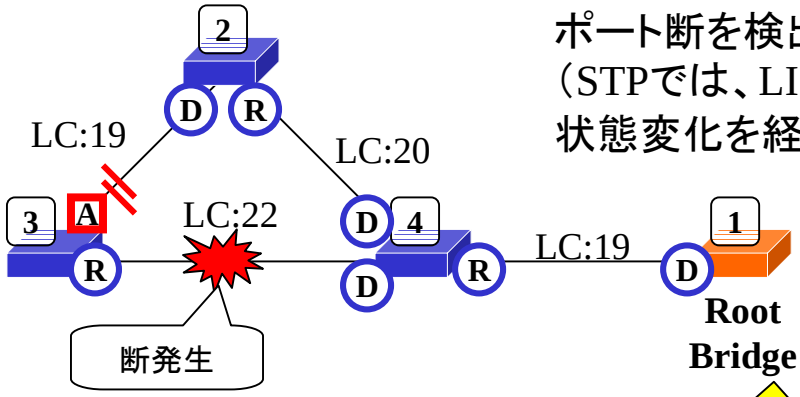
RSTPによる論理トポロジーの構築（続き）



RSTPでの障害回復（アルタネートポート）

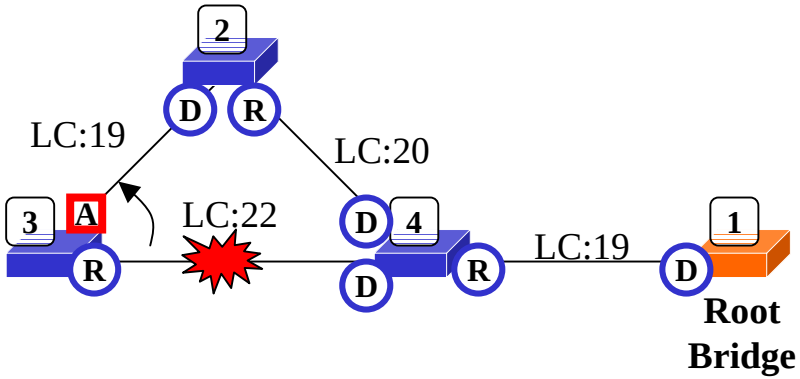
障害発生の場合（アルタネートポートがあるとき）

（1）障害発生



アルタネートポートは二番目にRootブリッジに近いポート。ルートポート断を検出するとすぐに転送状態に切り替える。
（STPでは、LISTENNING->LEARNING->FORWARDINGと言う状態変化を経なくては切り替えられなかった。）

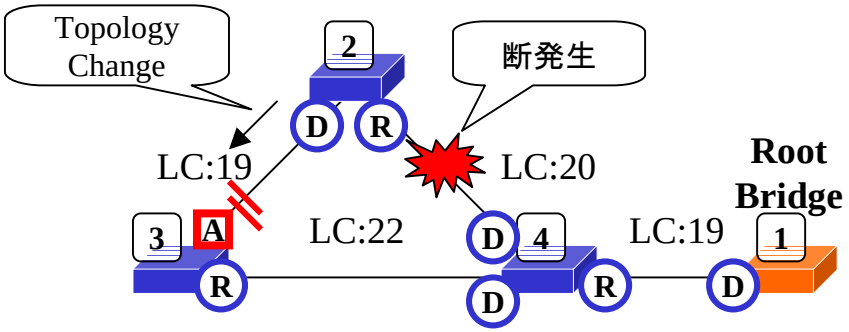
（2）すぐにアルタネートポートを転送状態にする



RSTPでの障害回復

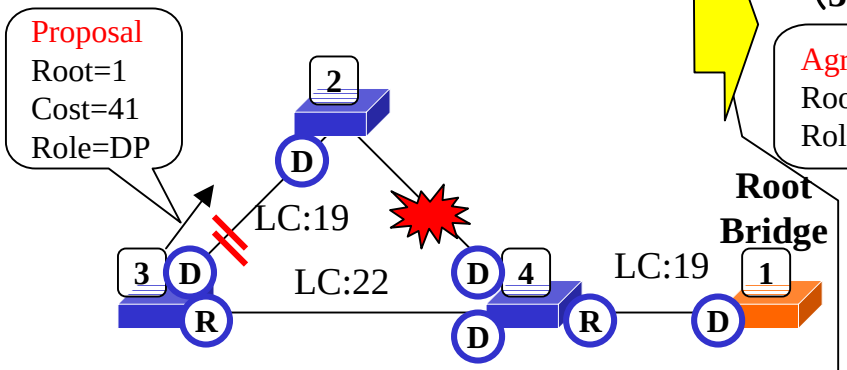
障害発生の場合2

（1）障害発生



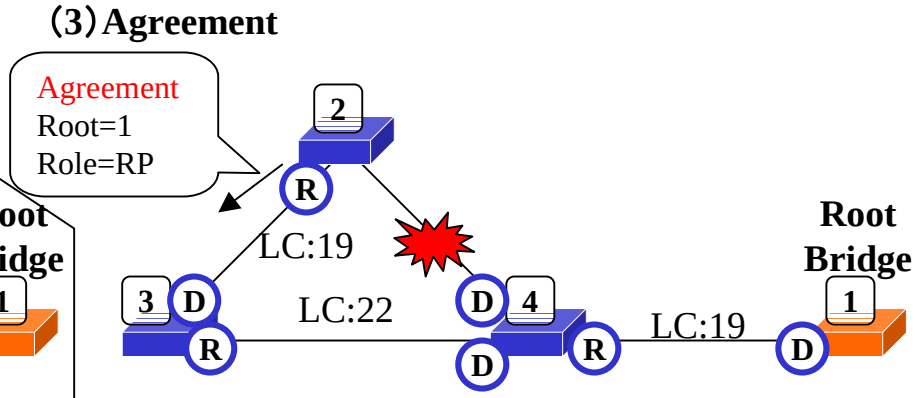
•スイッチ2はルートポートの断を検出すると、トポロジー変更があった事を示すBPDUを送信する。

（2）Proposal



•スイッチ3はTCNを受信後、ただちにHandshakeを開始

（3）Agreement



•Handshake完了で、新トポロジー完成

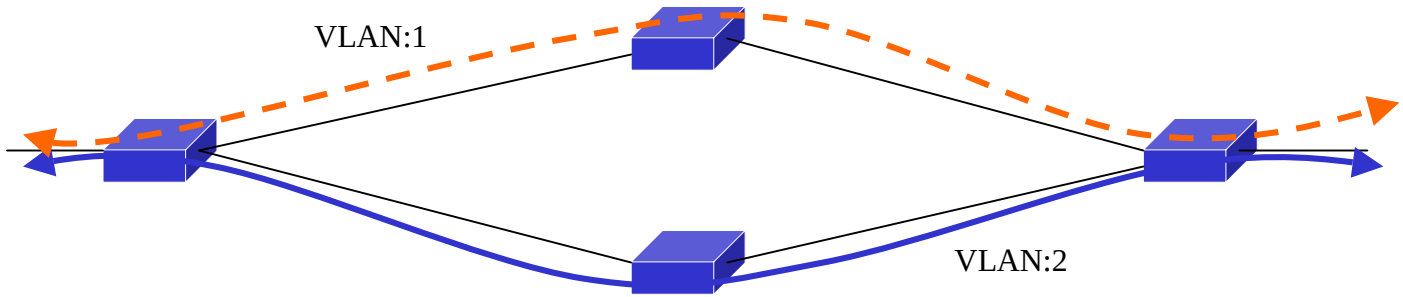
RSTPのまとめ

- トポロジー変更
  - LISTENNING->LEARNING->FORWARDINGのステートをHandshakeを導入したことによりはぶけるのでSTPよりも高速。

- トポロジー変更に伴うFDBのフラッシュ
  - RSTPでは、Topology Change Notification BPDUは使わない。そのかわり、Topology Change Flagを立てたBPDUを使って、他のスイッチにトポロジー変更が発生した事を教える。
  - Topology Change Flagを立てたBPDUを受信したスイッチは他のポートにTopology Change Flagを立てたBPDUを送信するともに、FDBのフラッシュを行う。
- 802.1Dとの接続
  - Proposalを投げて、Agreementを返してこなければ(Forward Delay x 2の時間)、802.1Dの動作をする。

MSTP(Multiple Spanning Tree Protocol)802.1s

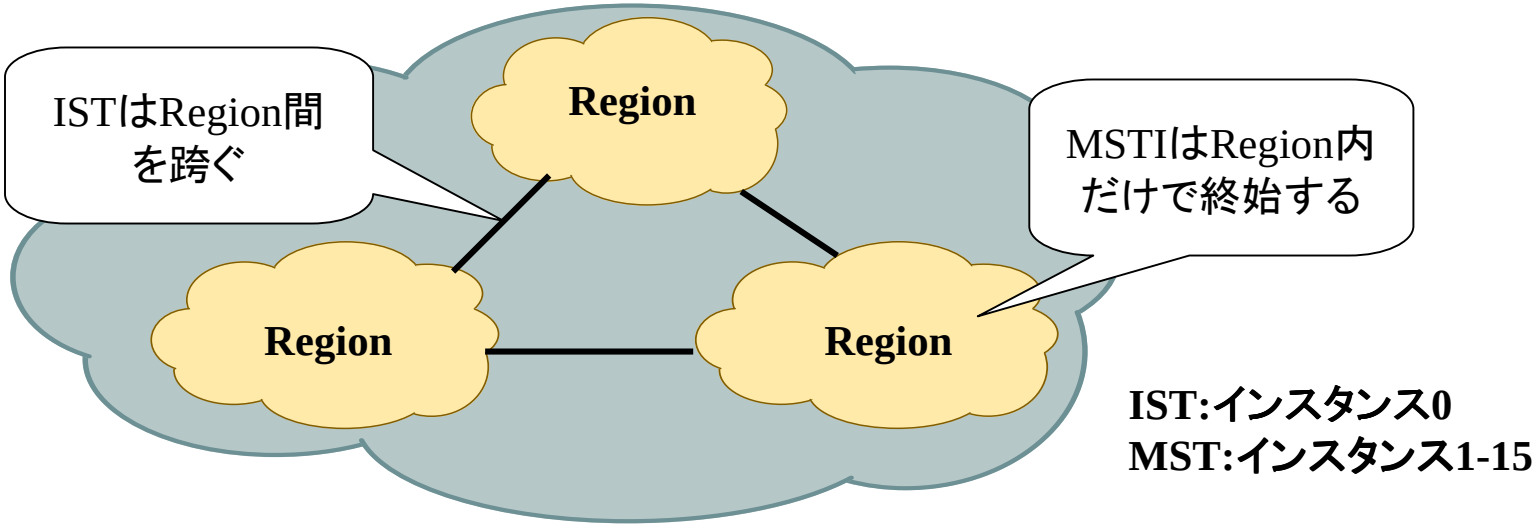
- STPで複数のトポロジー（インスタンス）を扱いたいと言う要求に答える為に登場
- IEEE 802.1s標準
- 802.1sは802.1Dの上位互換性がある。
- RSTP 802.1wと連携して使われる。
- VLANごとに別々のSTPのインスタンスを動作させる方法（PVSTなど）もあるがVLAN数が増えるとそれなりに負荷が大きくなるので、MSTPでは複数のインスタンスを一つのBPDUで扱えるようにしている。



負荷分散などの為にVLANごとに経路を変えたい事がある

MSTP

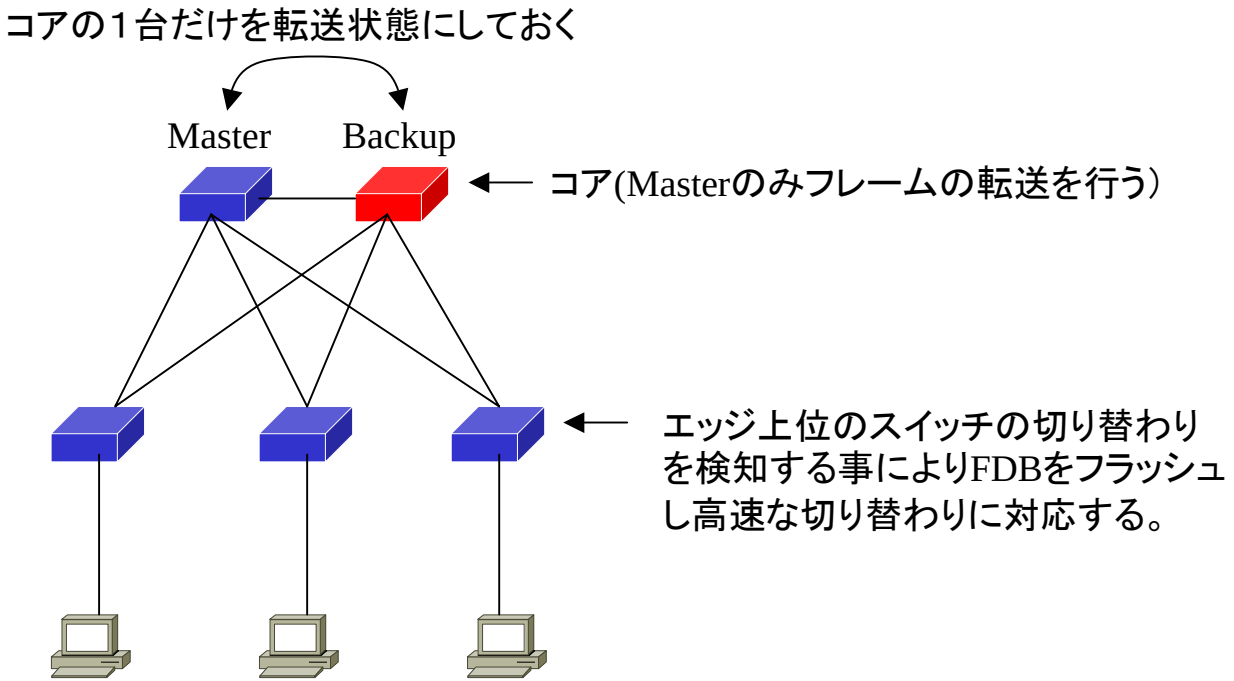
- MSTPには802.1D互換の親玉になる1個のInternal Spanning Tree(IST)と多数のMultiple Spanning Tree Instance(MSTI)がある。
- 個々のVLANはIST（すべてのVLANがマッピングされている）と任意のMSTIインスタンス1つにマッピングされ、それらのインスタンスの挙動に同期した挙動を行う。
- MSTIごとに、BPDU(MSTPのBPDUはversion3)にM-recordと呼ばれるレコードが追加される。
- 1個のBPDUに多数のM-recordが搭載される為、インスタンスが増えてもBPDUは増えない。
- Regionと呼ばれる概念がありMSTIはリージョン内に閉じ込められるが、ISTはRegionをまたいで存在する。



STPファミリ以外の冗長化プロトコル
ノード冗長化（メッシュトポロジー）
　ESRP、VSRP、FVRP、GVRP

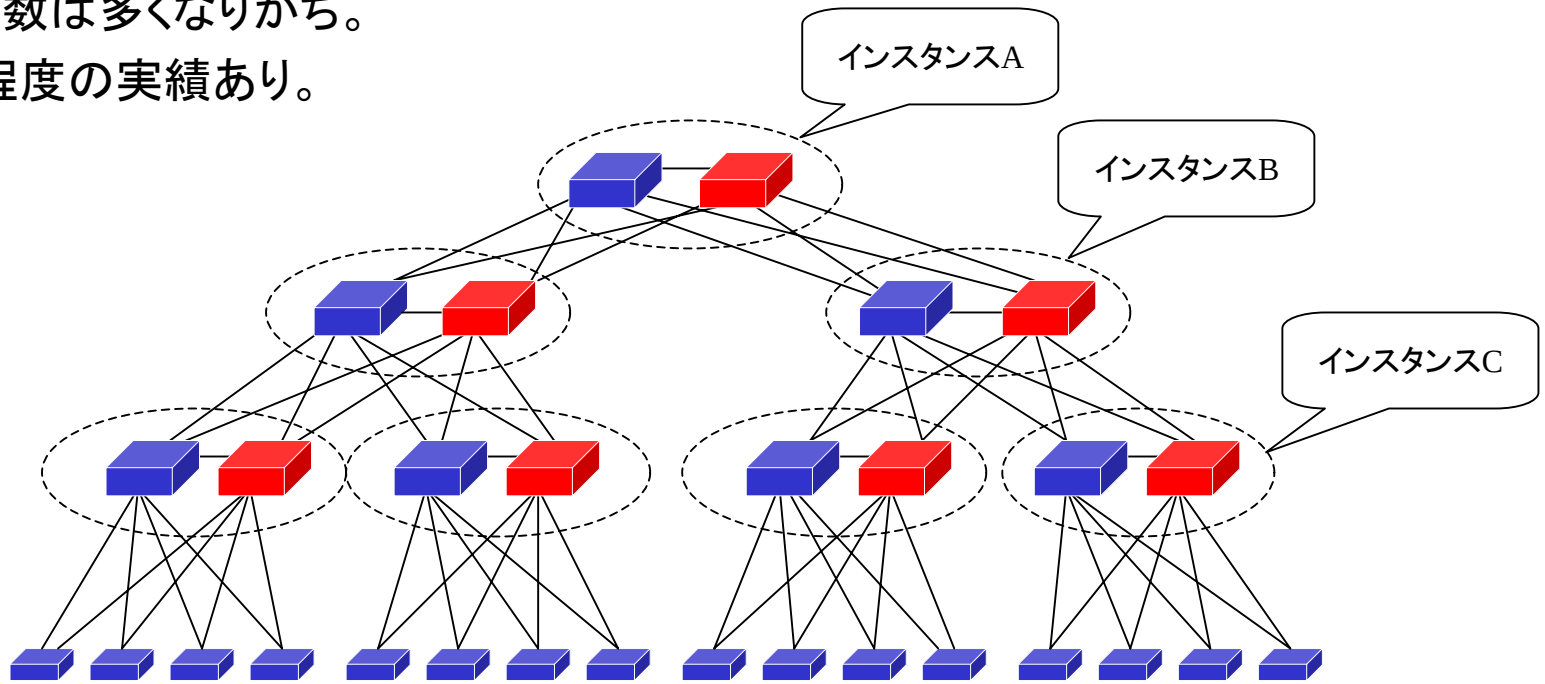
ノード冗長化プロトコル（メッシュトポロジー）

- ネットワークコアにあるスイッチを二重化し、コアに接続されるエッジ側のスイッチはコアに二重帰属する構成が基本
- Edge-Coreトポロジーとも呼ばれる。
- 標準的なプロトコルはない、ベンダ独自の実装
  - ESRP : Extreme
  - VSRP : Foundry
  - FVRP : Force10
  - GVRP（仮称）：日立製作所



ノード冗長化プロトコルの大規模化

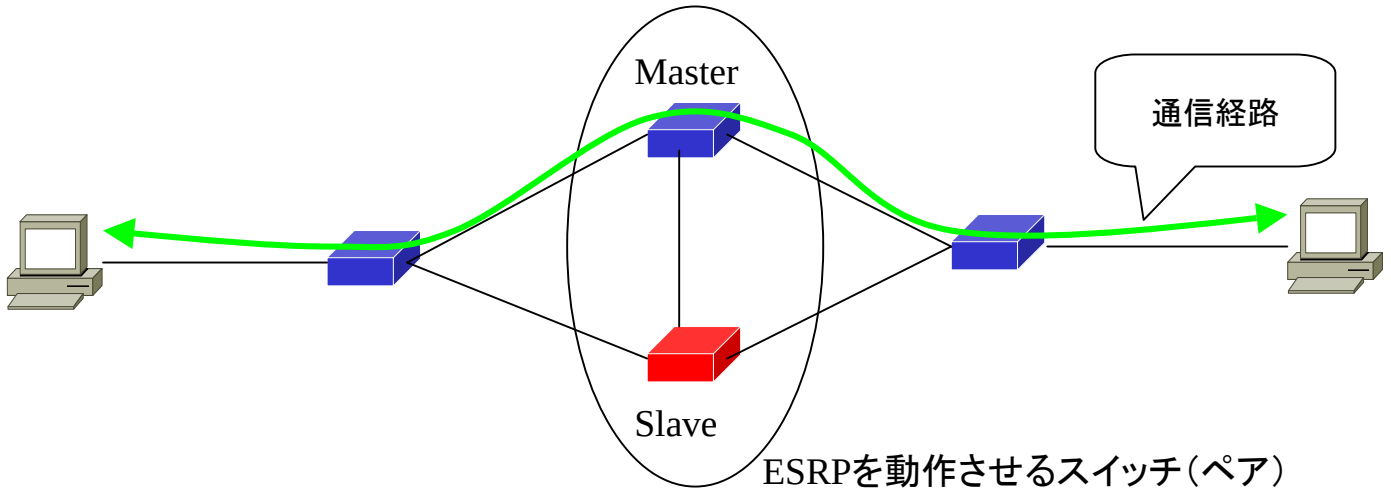
- 大規模な冗長を組む場合は、スイッチのペアごとに独立したのノード冗長化のインスタンスを作成し、そのペアを通過するVLANはそのインスタンスの挙動に同期して動作をするようにする。
- STPと比較して、スイッチやリンクの障害がネットワーク全体のトポロジーの再構成に引き起こさない（トポロジー変更が局所に閉じる）と言うメリットがある。
- リンク数は多くなりがち。
- ある程度の実績あり。



ノード冗長化プロトコル（ESRP）

- ESRP（Extreme Standby Router Protocol）
  - Extreme社が開発した、ノード冗長化プロトコル、レイヤー2とレイヤー3の冗長機能の両方の機能を提供している。

  - 冗長機能を必要とするスイッチにESRP機能を持たせ、冗長を持たせている。
    - Master スイッチ：データの送受信を行っているSW
    - Slave スイッチ：　データの送受信を行わず、予備状態となっているSW(Standbyとも言う)
    - マスタvlan：　ESRPを管理するvlan、マスタvlanのみESRPのアルゴリズムを計算し、他のvlanはマスタvlanの動作に同期してMaster Slaveの選択を行う事が出来る。

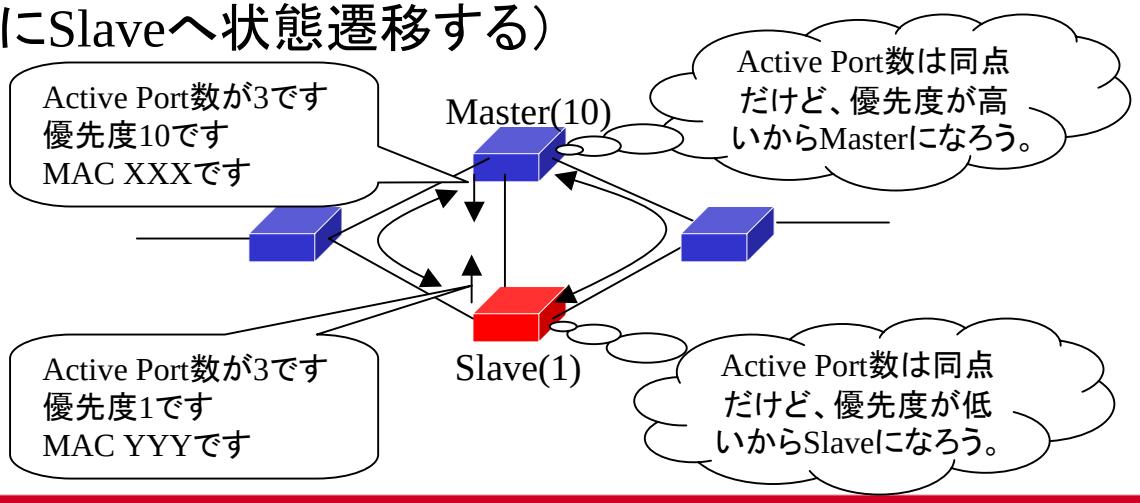


ESRPを動作させるスイッチ（ペア）

ノード冗長化プロトコル(ESRP)

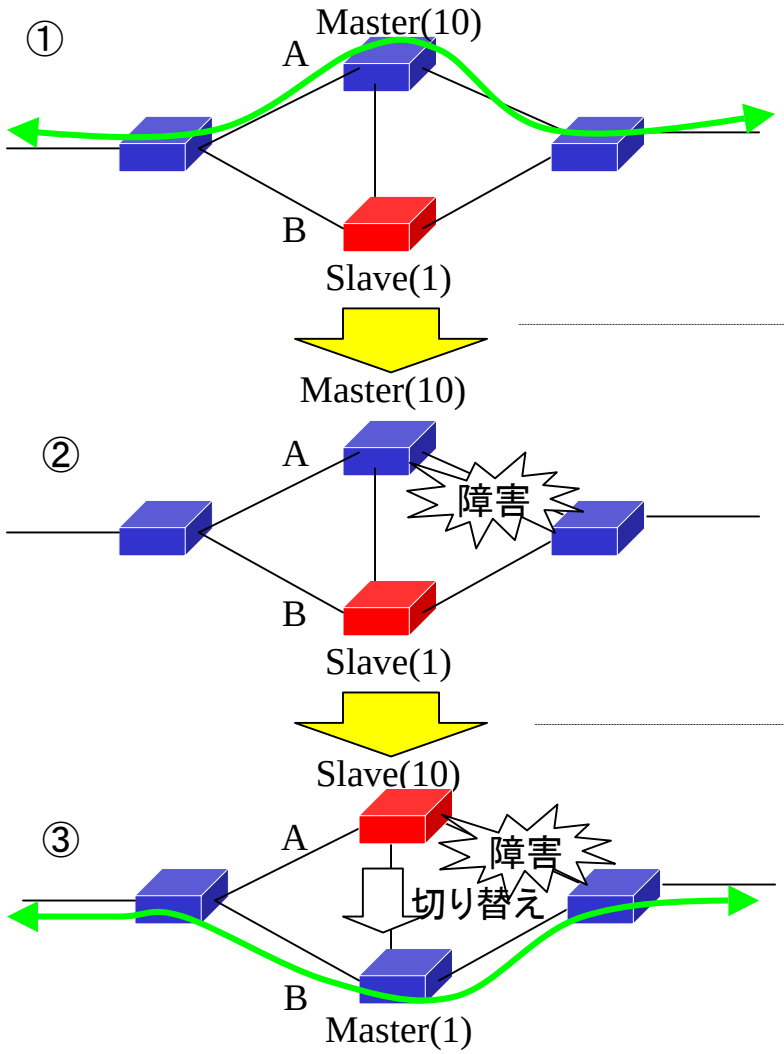
- ESRPマスターの選択
  - ESRPを動作させるスイッチでは、定期的に制御フレーム(ESRP hello packet)を交換し、どのスイッチが最もMasterにふさわしいかを判断している。
- ESRPマスターを決定する要素
  - Active Port数
  - スイッチの優先度(Priority)
  - トラッキング情報(pingなど)
  - システムMACアドレス（大きい番号のものが優先）

  これらの要素をタイブレークルールで比較していく。（比較順は変更可能）
- 相手スイッチがMasterに遷移したと言う通知を受けた時、自身がMasterであったら、即座にSlaveへ状態遷移する）



ノード冗長化プロトコル(ESRP)

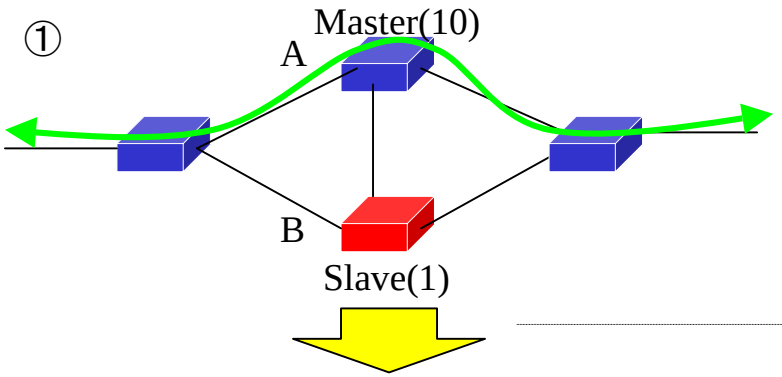
リンク障害による切り替え



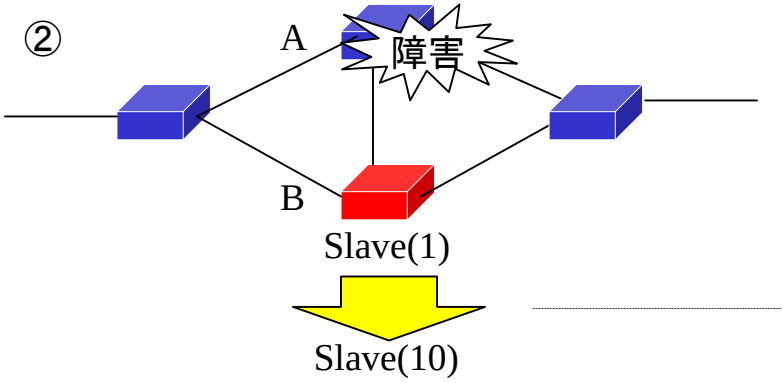
- フレームはMasterとなっているスイッチAを経由して流れている。
  - Active port数 3:3　→ 同点
  - Priority 数 10:1 →10の勝ち(スイッチAがマスター)
- スイッチAを経由する通信経路で障害が発生。
  - 勝負前
- ESRPの機構により、Master/Slaveの切り替えが発生し、フレームの流れる経路も変わる。
  - Active port数 2:3 →3の勝ち(スイッチBがマスター)

ノード冗長化プロトコル(ESRP)

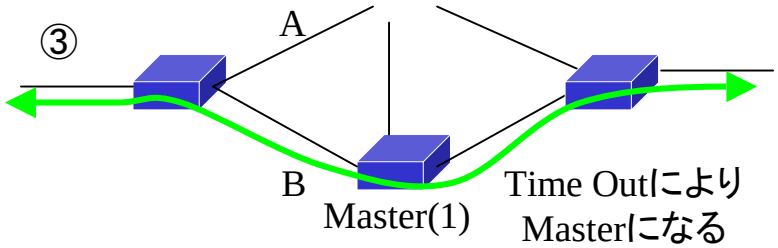
ノード停止による切り替え



- フレームはMasterとなっているスイッチAを経由して流れている。
  - Active port数 3:3　→ 同点
  - Priority 値 10:1 →10の勝ち(スイッチAがマスター)



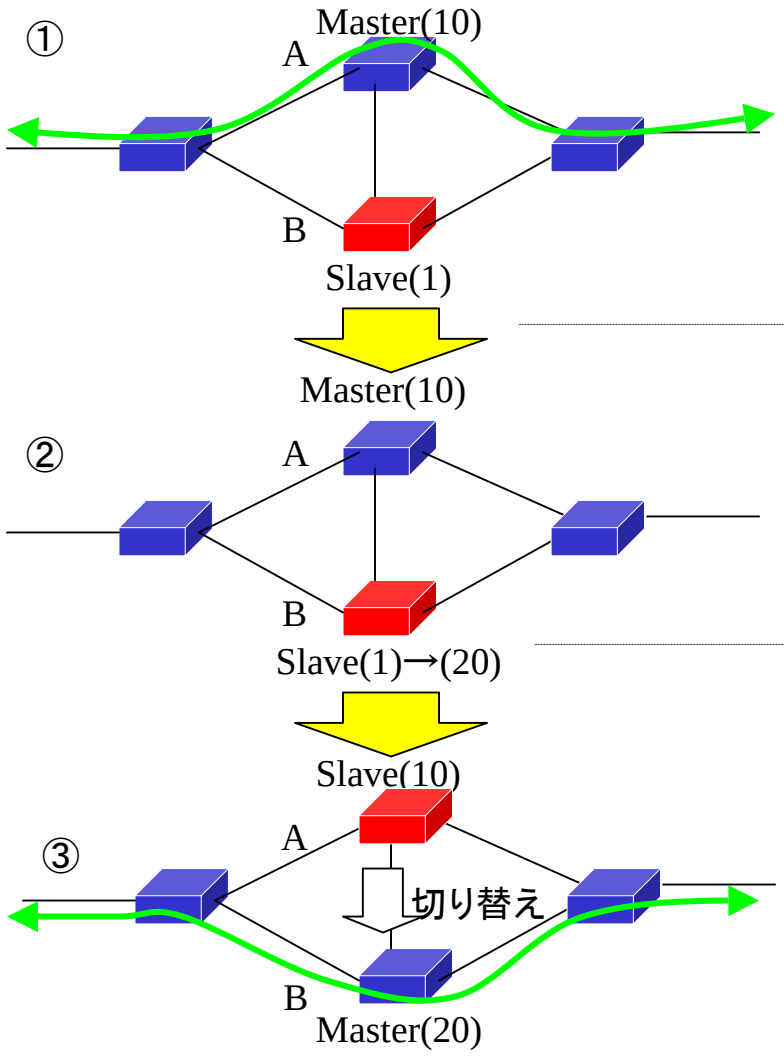
- スイッチAが停止



- スイッチBでhello packetが未受信となってあらかじめ設定した時間を越えると、SlaveスイッチはMasterスイッチに障害が発生したと認識してMasterとなる。

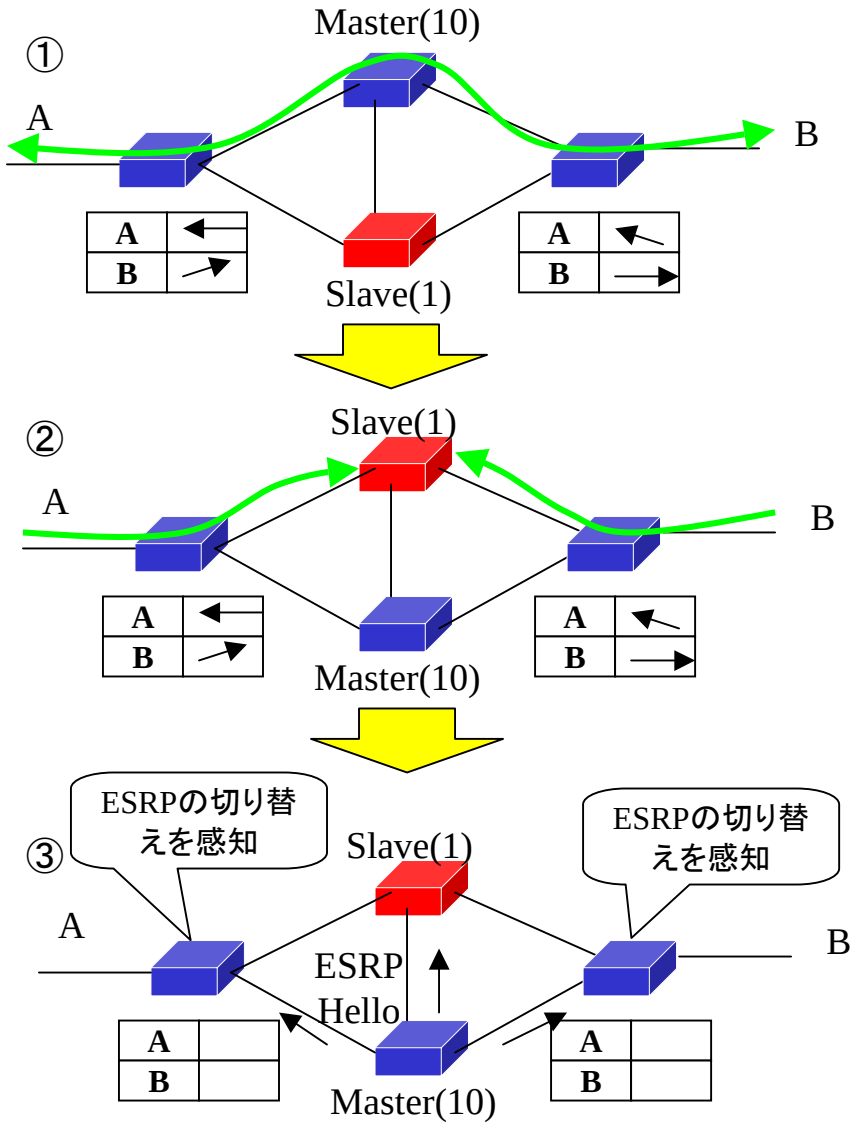
ノード冗長化プロトコル(ESRP)

Priority変更による切り替え



- フレームはMasterとなっているスイッチAを経由して流れている。
  - Active port数 3:3　→ 同点
  - Priority 値 10:1 →10の勝ち(スイッチAがマスター)
- スイッチBに系を切り替える為にPriority値を 1から20に変更する。
  - 勝負前
- ESRPの機構により、Master/Slaveの切り替えが発生し、フレームの流れる経路も変わる。
  - Active port数 3:3 →同点
  - Priority値 10:20→20の勝ち(スイッチBがマスター)

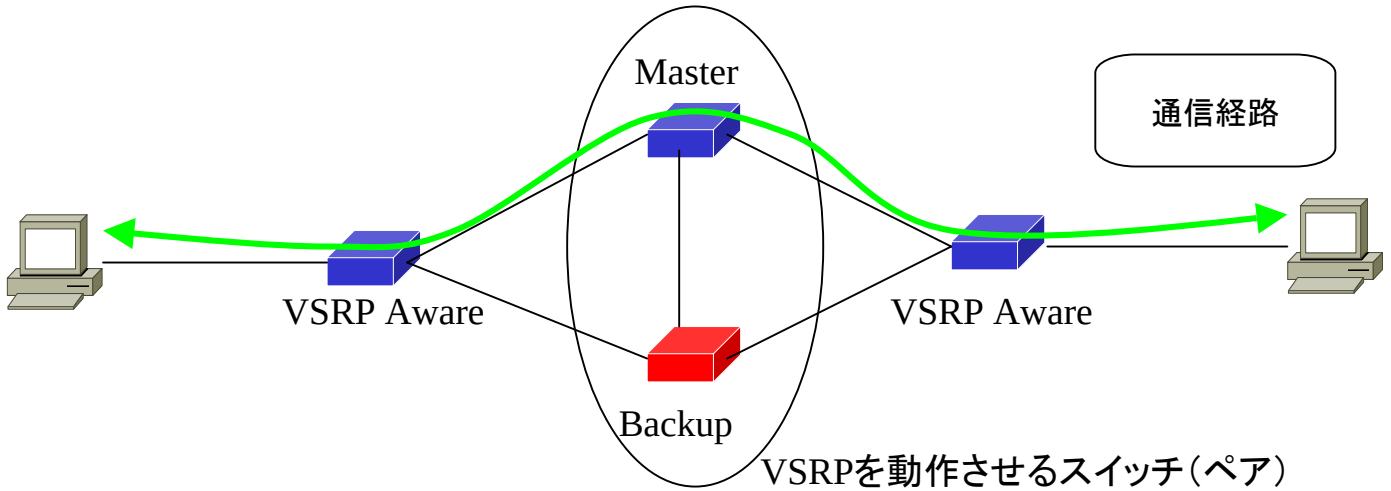
ノード冗長化プロトコル（ESRP AWARE）



- ESRPのMasterを経由してトラフィックが流れると当然その経路に従ったFDBの学習がなされる。
- ESRPのPriorityなどを変更して、Master/Slaveの関係が入れ替わった場合に、そのままでは、古いFDBの内容に従ってフレームの転送が行われる為、フレームが結果的に転送不能となる。
- FDBの矛盾状態を防ぐ為、両端のスイッチは、上位のスイッチがMaster/Slaveの関係を変更した事をESRP Helloの内容変更を見る事により、FDBの内容をFlushする（消す）。これにより再度学習が行われ、正常な通信が行えるようになる。

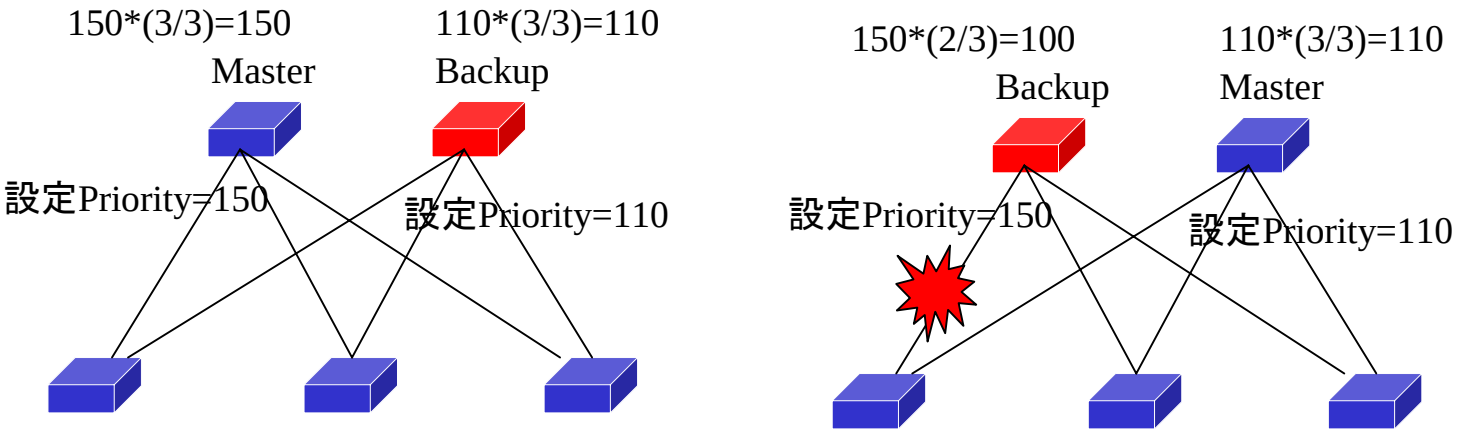
ノード冗長化プロトコル（VSRP）

- VSRP（ Virtual Switch Redundancy Protocol ）
  - Foundry Networks社が開発した、ノード冗長化プロトコル、レイヤー2とレイヤー3の冗長機能の両方の機能を提供している。
  - 冗長機能を必要とするスイッチにVSRP機能を持たせ、冗長を持たせている。
    - Master スイッチ： データの送受信を行っているSW
    - Backup スイッチ： データの送受信を行わず、予備状態となっているSW
    - Topology Group：複数のVLANをグループ化し、Masterを共有する機能。
  - VSRP Awareな装置は、系が切り替わりMasterとなったスイッチが送信するTC packetsを受信すると、FDBをフラッシュするのではなく、Backup側に書き換える。



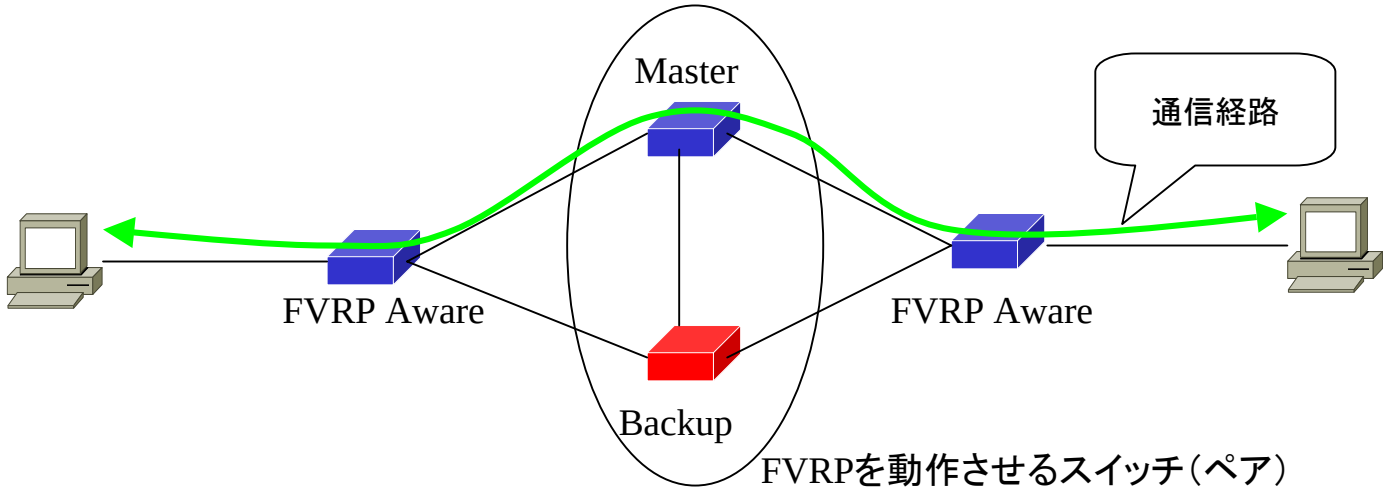
ノード冗長化プロトコル（VSRP）

- VSRPの priority(3-255 Default 100)の高い方がMasterとなる。
- ポートがダウンするとpriorityが減る。Priority x（利用可能Link数/設定Link数)
- Tracking portにより、特定のリンクのDownによりpriority値を制御可能
- VSRP Helloを使って　priority情報を交換(Default 1秒間隔)
- Active 決定後はHelloはMasterからのみ送信
- Backup側のスイッチはMasterから、Dead interval時間Helloを受信しないと、Hello packetを送信しはじめ、さらに、Hold-down interval時間自分よりpriorityの高いHelloを受けとらなければ、Masterとなる。



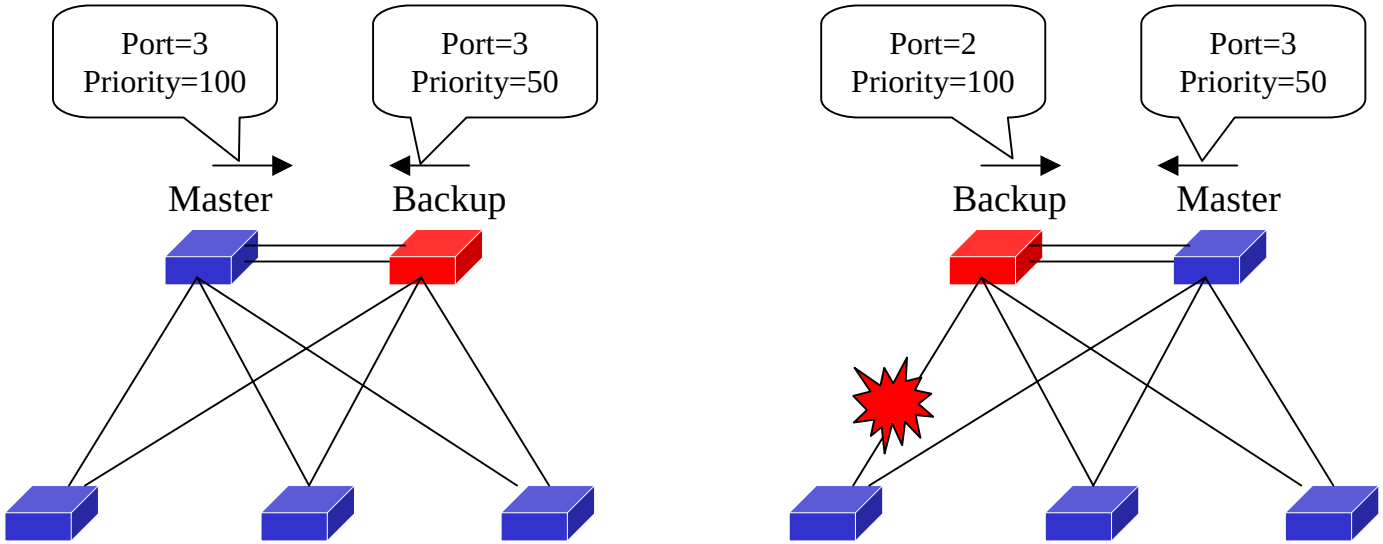
ノード冗長化プロトコル（FVRP）

- FVRP（ Force10 VLAN Redundancy Protocol ）
  - Force10社が開発した、レイヤー2ノード冗長化プロトコル。
  - 冗長機能を必要とするスイッチにFVRP機能を持たせ、冗長を持たせている。
    - Master スイッチ： データの送受信を行っているSW
    - Standby スイッチ： データの送受信を行わず、予備状態となっているSW
    - FVRP Domain： 複数のVLANをグループ化し、Masterを共有する機能。
  - FVRP Awareな装置は、コアスイッチより、flush address messageを受信するとFDBをフラッシュする。



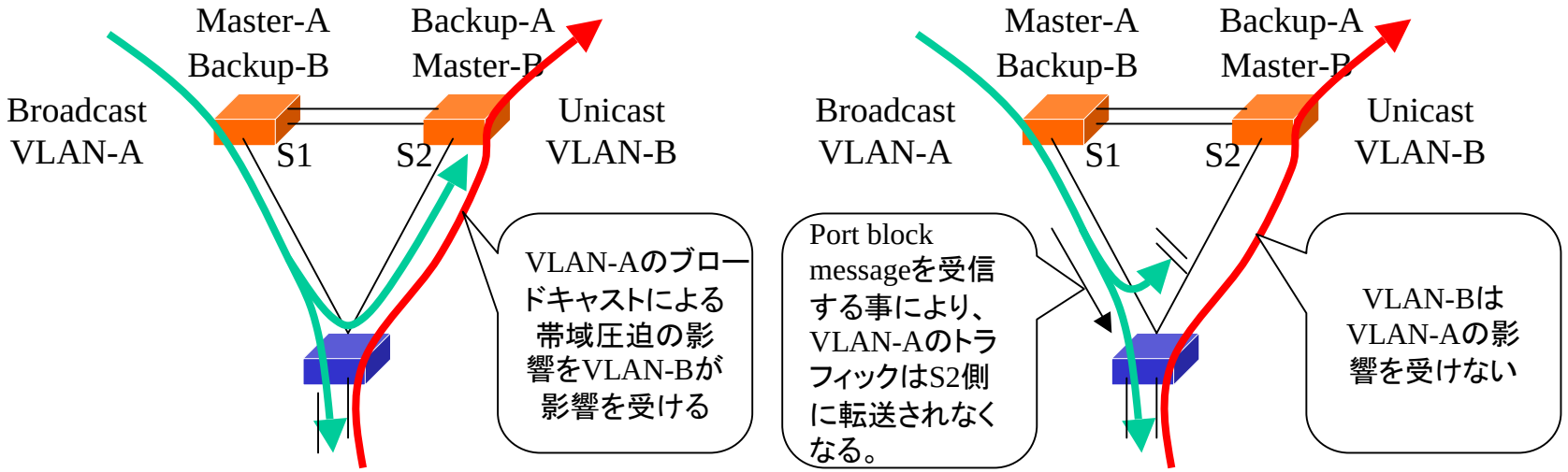
ノード冗長化プロトコル（FVRP）

- ポート数、priority、制御ポートのMACアドレス（低い方が有利）の順でタイブレークルールで比較し、勝った方がMasterとなる。
- priority(1-255 ただし255は強制Slave)は高い方がMasterとなりやすい。
- FVRP Helloを使って　priority情報を交換(Default 1秒間隔)
  - 通常は、Master-Standby間に張られたCore Linkを使ってHelloのやり取りを行う。
  - Core Linkが断になった場合はアクセスリンク上のコントロールVLANを使ってHelloのやり取りを行う。(Dual Masterを防ぐ為）
- Standby側のスイッチはMasterから、Message Age Timer時間Helloを受信しないと、遷移プロセスに移る。

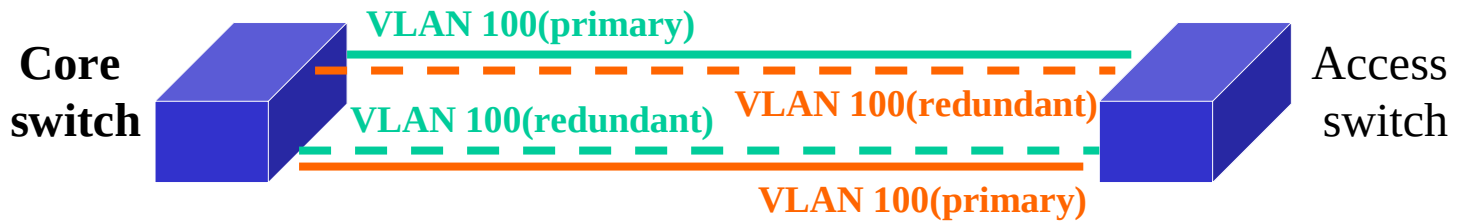


ノード冗長化プロトコル（FVRP）

- ブロードキャストが帯域を無駄使いする事を防止
  - Masterから port block messageを受信すると該当するVLANのStandby側のポートをブロックする。（これによって余計なブロードキャストが回りこまずに、帯域を有効活用出来る）

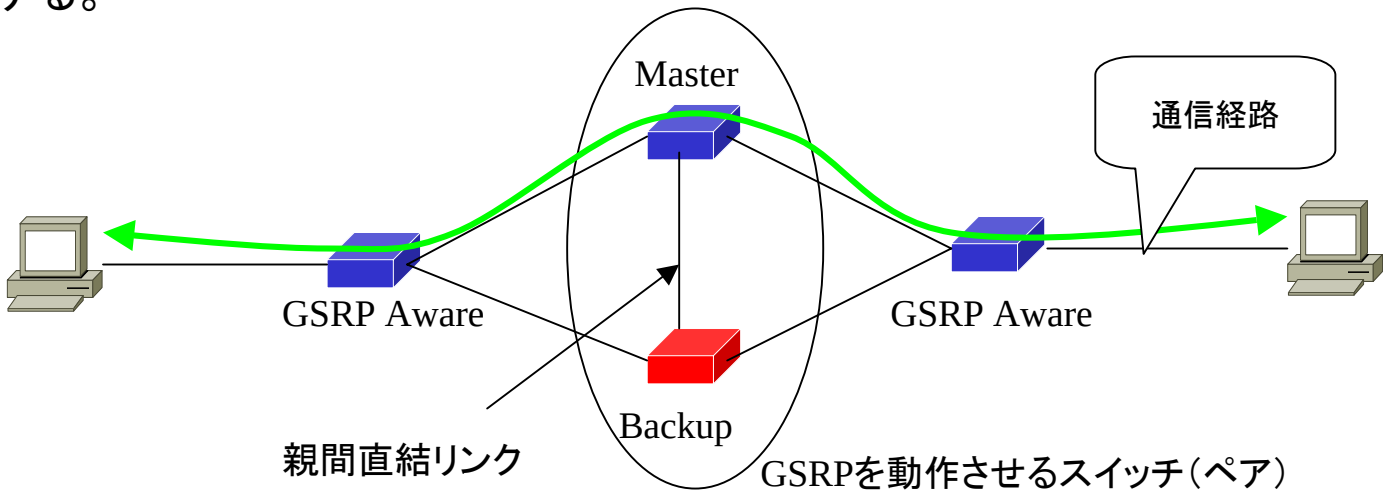


- FVRPはスイッチの対向でも使用出来る！



ノード冗長化プロトコル（GSRP）

- GSRP（ GS4000 Switch Redundant Protocol =仮称 ）
  - （株）日立製作所が開発した、ノード冗長化プロトコル。
  - 冗長機能を必要とするスイッチにGSRP機能を持たせ、冗長を持たせている。
    - Master スイッチ： データの送受信を行っているSW
    - Backup スイッチ： データの送受信を行わず、予備状態となっているSW
    - VLAN Group： 複数のVLANをグループ化し、制御する機能。
  - マスターの切り替えが発生すると、新マスタは全隣接スイッチに、GSRP Flush Requestを送信する、GSRP Awareな装置はそれを受信するとFDBをフラッシュする。



ノード冗長化プロトコル（GSRP）

- GSRPマスターの選択
  - GSRPを動作させているスイッチでは、定期的に制御フレーム(GSRP Advertise)を交換し、どのスイッチが最もMasterにふさわしいかを判断している。
- GSRPマスターを決定する要素
  - Active Port 数
  - スイッチの優先度（Priority）
  - 装置MAC（大きい番号のものが優先）

  これらの要素をタイブレークルールで比較していく。（比較順は変更可能）
- デュアルマスター防止策
  - 切り替え時のデュアルマスターの可能性（デュアルマスターはループになる）を排除する為の機構を持っている。（瞬間ループもFDBが狂うので絶対に駄目！）

  （1）マスターになろうとするスイッチはまず、マスター待ち状態になる。（ブロック状態のまま）

  （2）バックアップになろうとするスイッチはすぐにバックアップになり、バックアップになった事をGSRP Advertiseを使って広報

  （3）マスター待ちのスイッチは相手側のスイッチがバックアップになった事を示すGSRP Advertiseを受信すると、マスターとして動作しはじめる。
- GSRP Advertiseを規定回数受信しないと（1-255 Default=3）で相手不定状態になる。
- オプション指定時は、相手不定状態と親間直結リンク断条件組み合わせでバックアップスイッチはマスターとして動作しはじめる。