「最近はやりのVLAN技術」

― ループフリー実現手法とL2SW に関する技術の紹介 ―

2002／07／26 JANOG10
株式会社NTTデータ
吉野 誠吾

ループフリー実現手法とL2SW に関する技術の紹介

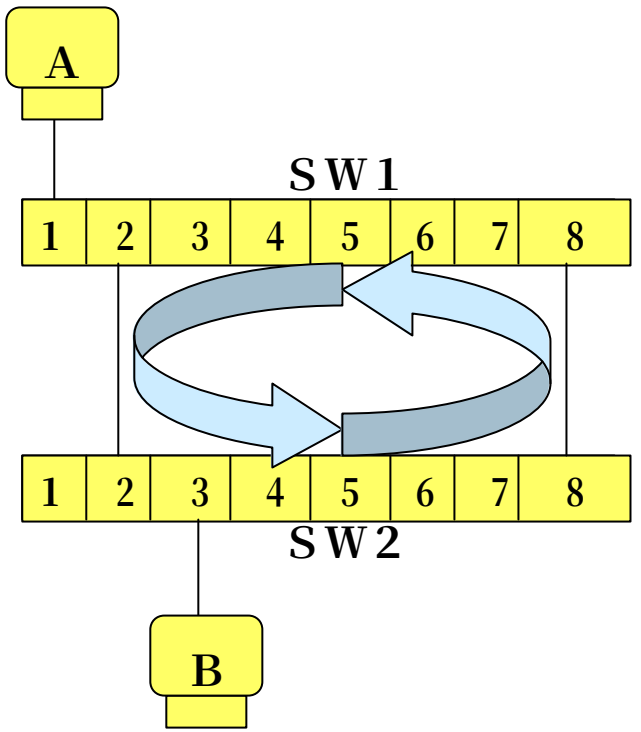
1. ループフリー技術の紹介
2. L2SW に関する技術
3. その他

1. ループフリー技術の紹介

1. 1 ループは危険
1. 2 STP
1. 3 EAPS（Extreme）
1. 4 UplinkFast（Cisco）
1. 5 その他（Cisco）
1. 6 RSTP
1. 7 MSTP

1. 1 ループは危険

学習していないアドレス宛てのパケットはブロードキャストと同じ



未学習のパケットを全てのポートに送信するので、無限に回りつづける（ストーム）。

A が送信したパケットが SW2 から送り返され、間違ったポートに A のアドレスを学習する（ミスラーニング）。

1．１ ループは危険

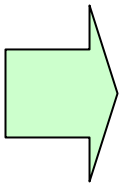
ループは絶対作ってはいけない。ループがあると・・。

ストーミング

と

アドレスのミスラーニング

による



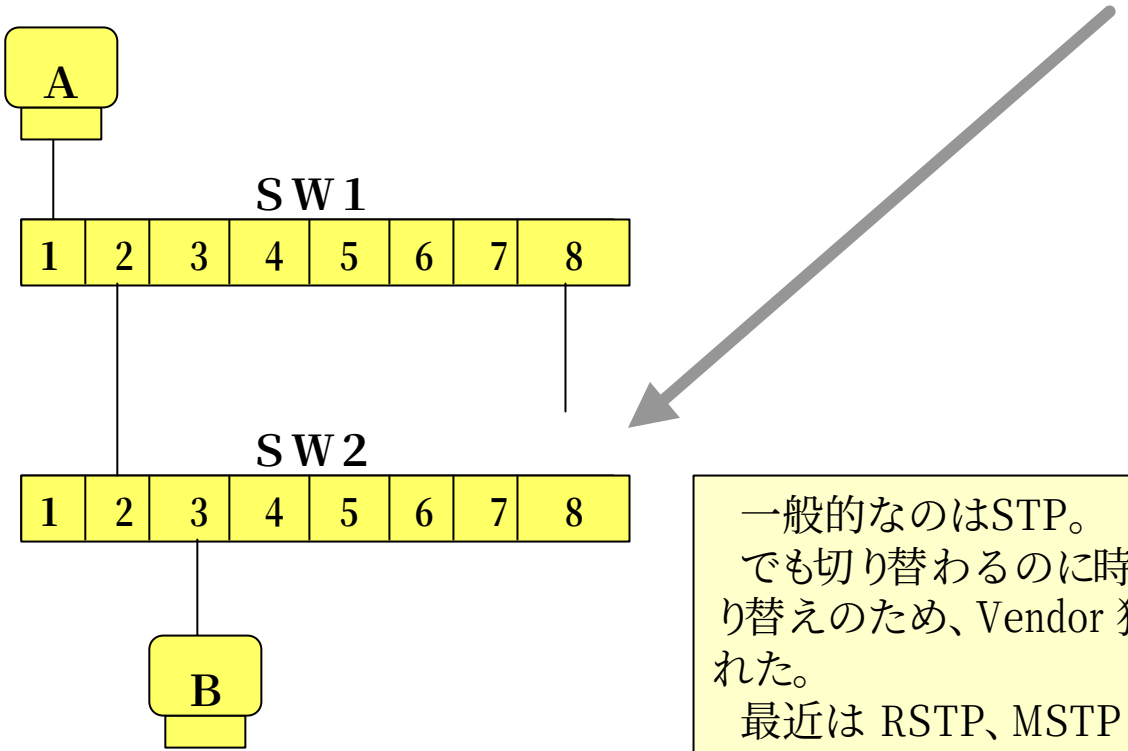
ネットワークの混乱が発生する

例えば、HSRP など CPU 処理の必要なパケットが大量に送りつけられ、ルータが停止した事例もあり。

ループは絶対に排除する！

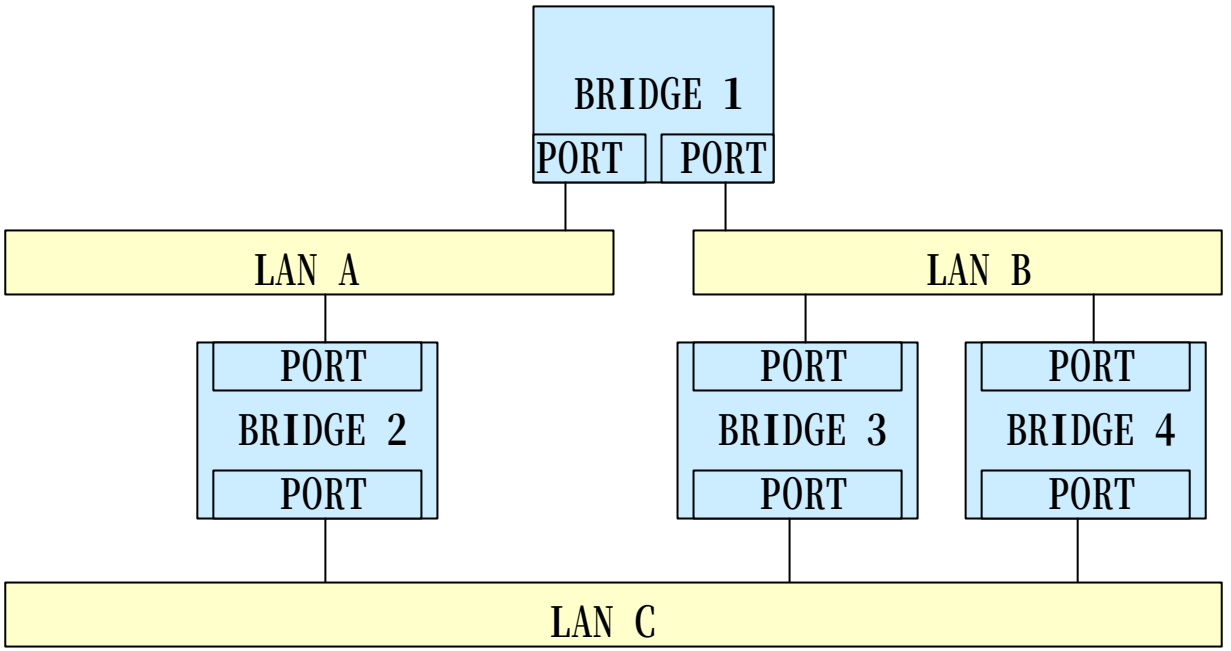
1. 1 ループは危険

ループフリーを実現するプロトコルではループを検知し一部を切り離す。

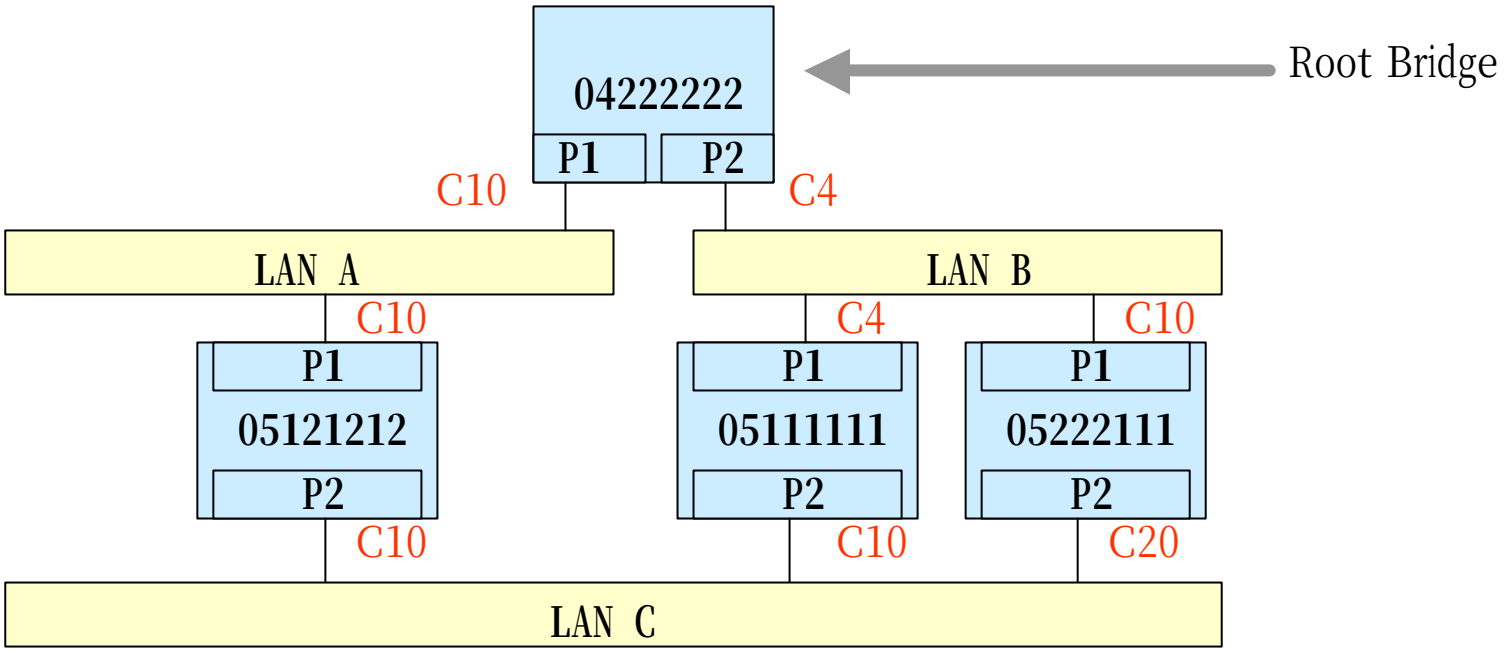


一般的なのはSTP。
でも切り替わるのに時間がかかるので、高速切り替えのため、Vendor 独自のプロトコルが実装された。
最近は RSTP、MSTP というIEEE によるプロトコルもあり。

1. 2 Spanning Tree algorithm and Protocol

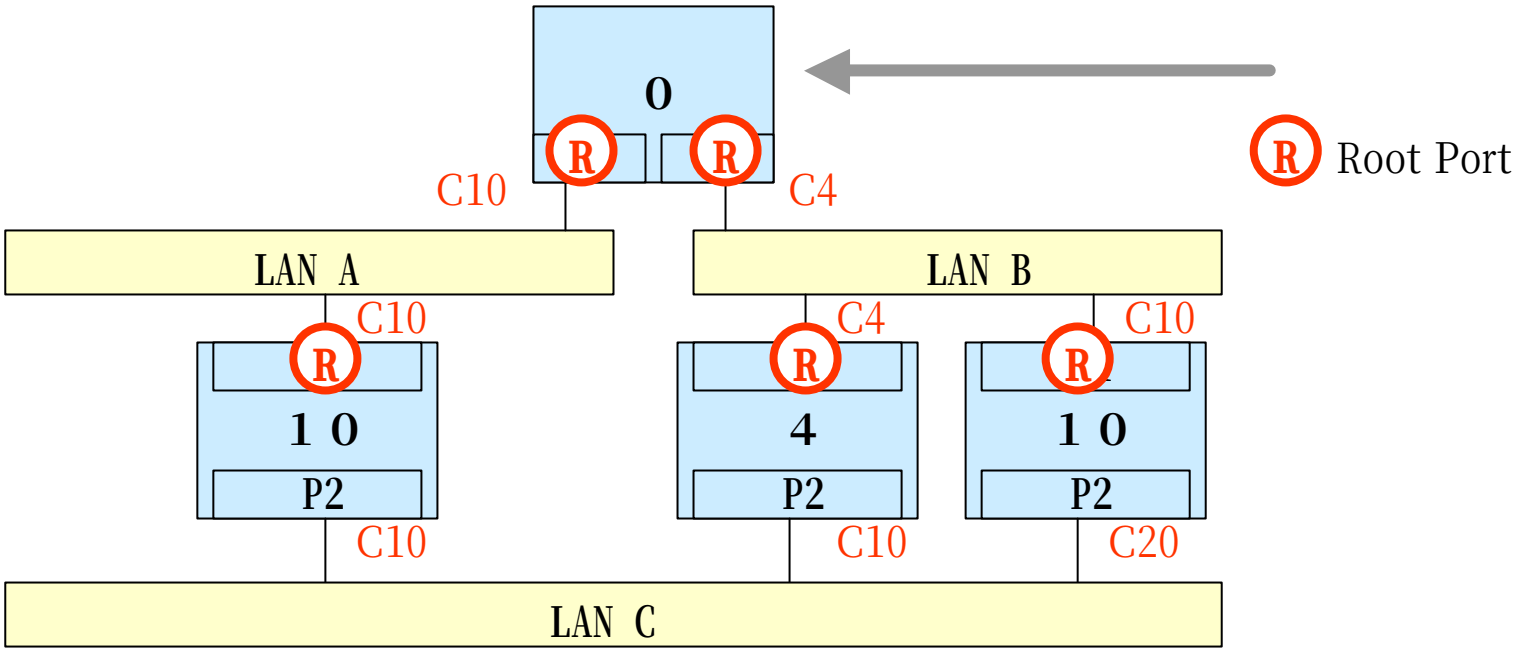


1. 2 Spanning Tree algorithm and Protocol



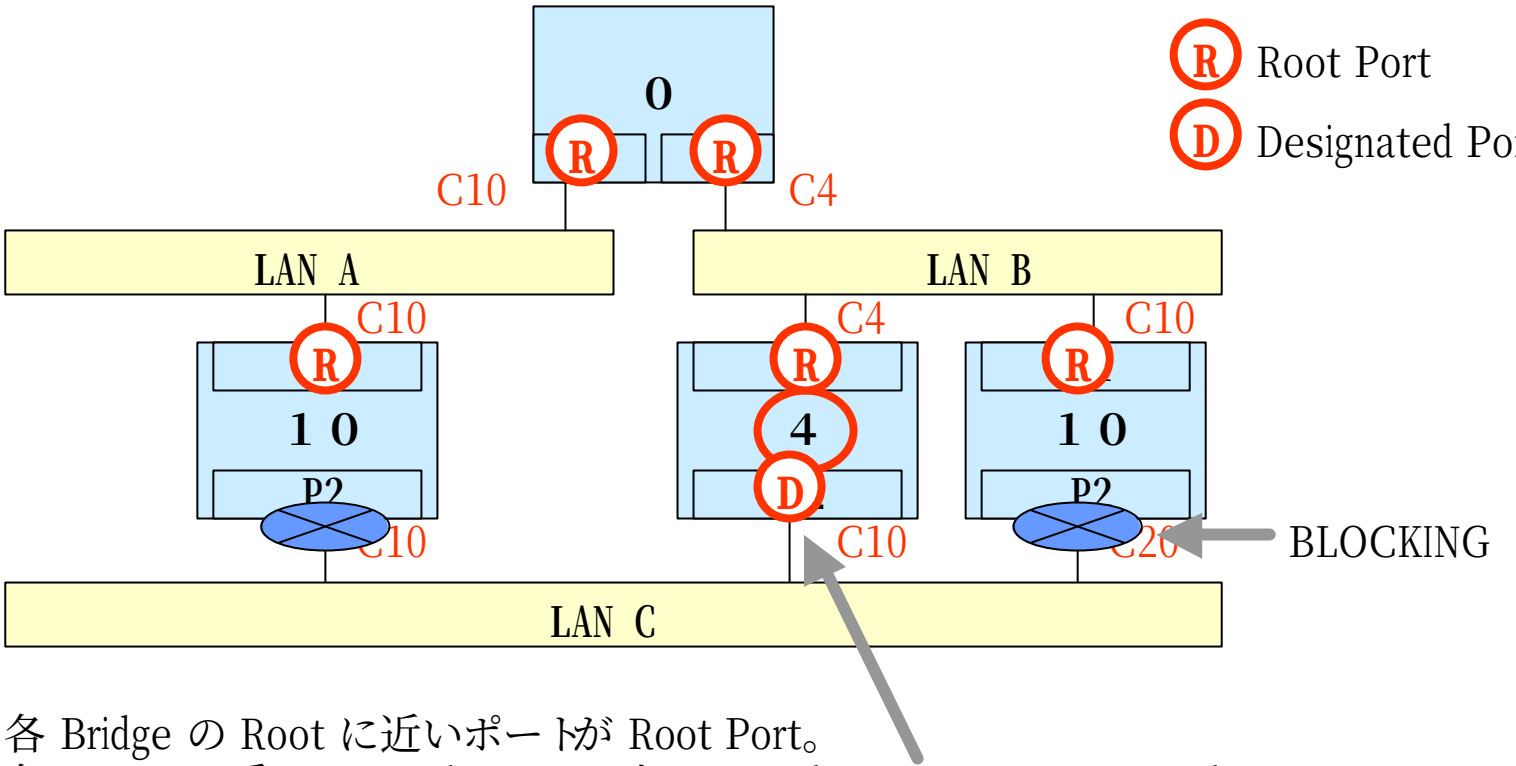
Bridge には、
・Bridge ID （2バイトのプライオリティ＋6バイトの MAC アドレス）図では簡略化
・Port ID
・Port Path Cost （高速なポートほど小さい数値）
の値がある。Bridge ID のプライオリティが高い（数値が小さい）BRIDGE1 が Root Bridge になる。

1. 2 Spanning Tree algorithm and Protocol



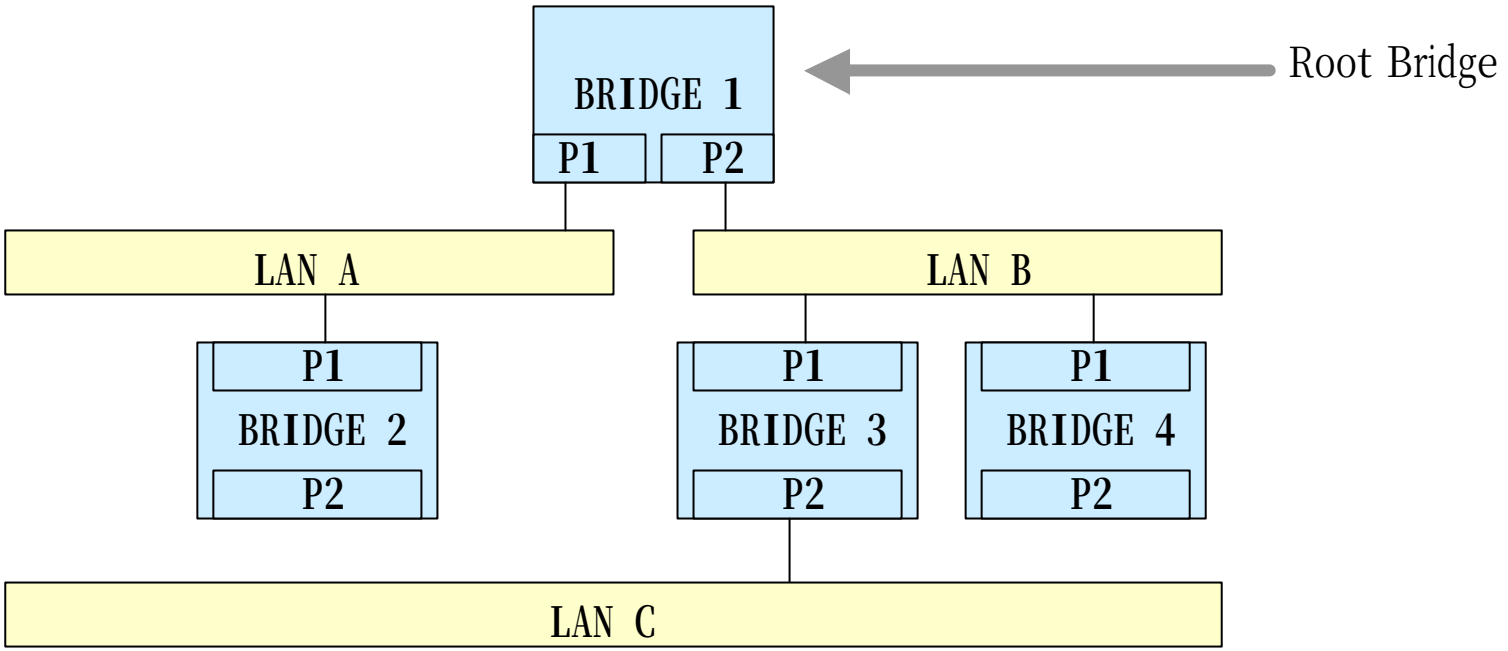
Root Bridge を 0 として各 Bridge で Port Path Cost を足し、Root Path Cost を求める。

1. 2 Spanning Tree algorithm and Protocol



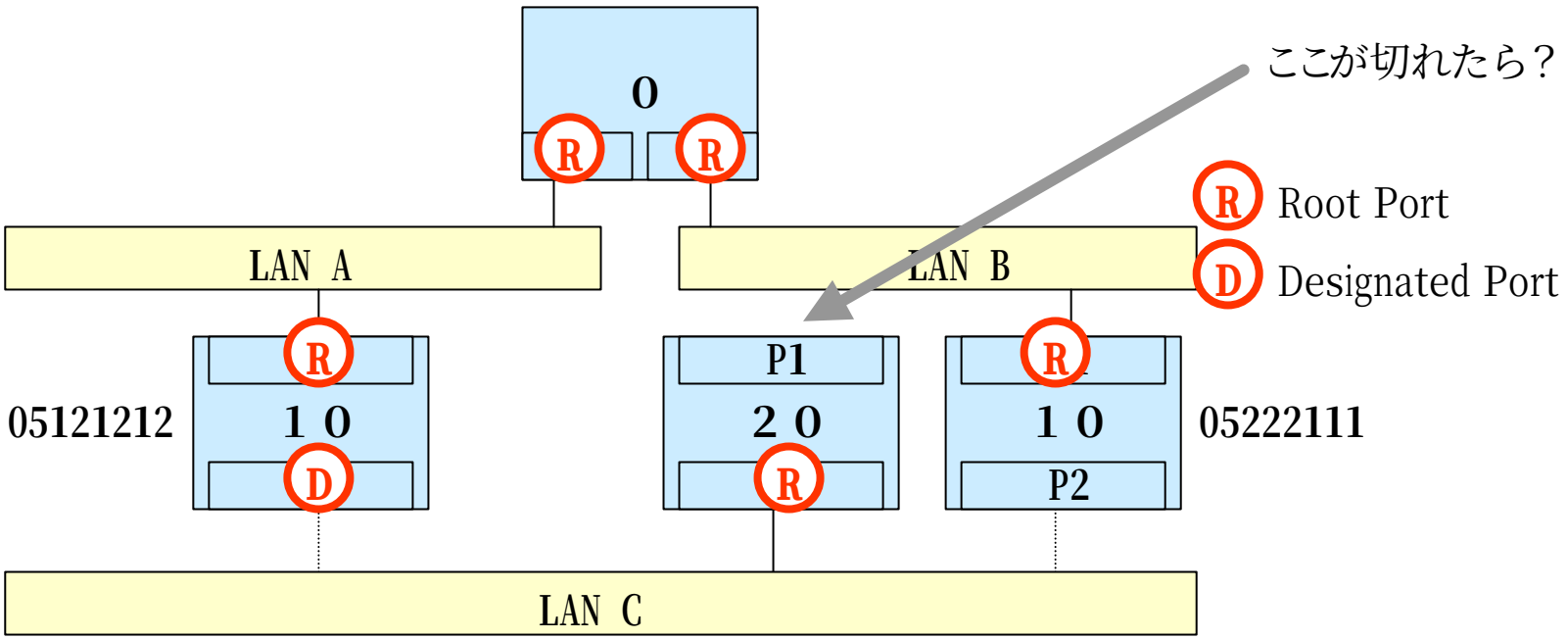
各 Bridge の Root に近いポートが Root Port。
各 LAN で一番 Root Path Cost が小さい Bridge Port が Designated Port。
これ以外が BLOCKING になる。

1. 2 Spanning Tree algorithm and Protocol



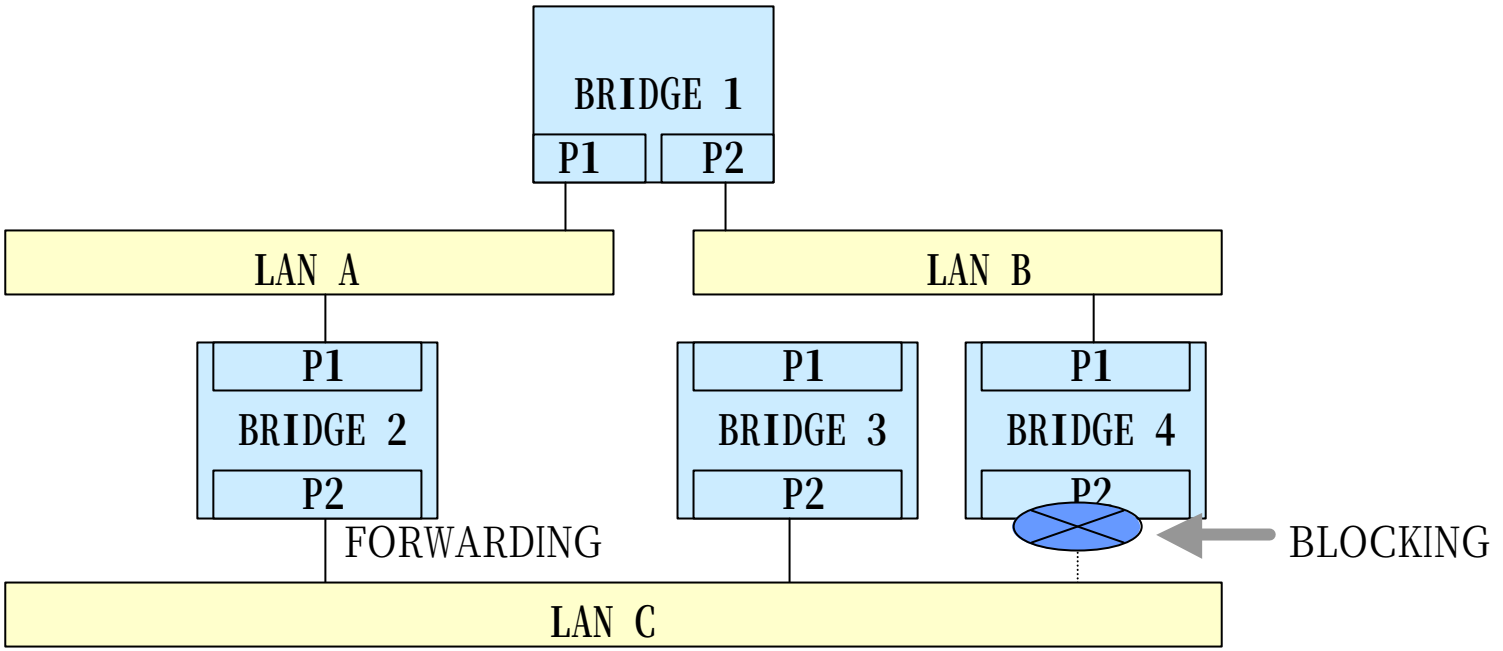
これでループはなくなる。

1. 2 Spanning Tree algorithm and Protocol



再計算。
BRIDGE 2、BRIDGE 4 ともに Root Path Cost は 10。
この時は Bridge ID のプライオリティが高い BRIDGE 2 が Designated Port となる。

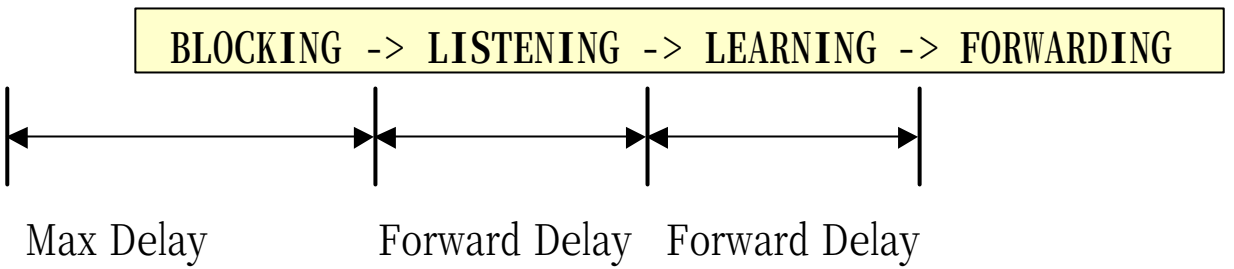
1. 2 Spanning Tree algorithm and Protocol



結果はこうなる。

1. 2 Spanning Tree algorithm and Protocol

BLOCKING からFORWARDING は周囲を確認しながら・・



LISTENING、LEARNING で、周りの Bridge とトポロジー情報を交換する。

リンクが切れなくても、Max Delay の間 Hello が届かないと障害と認識する。
Max Delay は default 20 秒
Forward Delay は default 15 秒。

1. 2 Spanning Tree algorithm and Protocol

IEEE802.1D で規定されている。Bridge と呼ばれていた時代からあるプロトコル。
Configuration BPDU を Bridge 間で交換し、Root Bridge を選定する。Root Bridge を Root とした Tree 上のトポロジーとなるよう、ループ上の 1つのポートを BLOCKING というパケットの送受信を行わない状態に変更しループを回避する。

Bridge 間で BPDU（Bridge Protocol Data Unit）と呼ばれるパケットをやり取りする。これは Hello パケットとも呼ばれる。
Root Bridge は Bridge ID のプライオリティが高いもの（数値の小さいもの）がなる。Bridge ID は 2 バイトのプライオリティ（設定可能）と 6 バイトの MAC アドレスを組み合わせたもの。
ポートには Port ID が割り当てられている。また、Path Cost という値が設定される。この Path Cost のデフォルト値は帯域が大きいほど小さい数値となる（実装によって違う場合がある）。Path Cost が小さいほど望ましいパスとなる。

続く

以前の Path Cost 計算式
Path Cost = 1000/LAN速度[Mbps]
最近はメディアの高速化に合わせて装置によって違う。今の 802.1D では 100Mbps は 19、1Gbps は 4、10Gbps は 2。802.1T では 10Gbps が 2000、1Tbps が 20、10Tbps が 2 などと拡張されている。

1. 2 Spanning Tree algorithm and Protocol

各 Bridge の Root Bridge までの Path Cost を Root Path Cost と言い、Root Bridge を 0 として Bridge の入り口インタフェースの Cost を加えたものが Root Path Cost となる。

各 Bridge には必ず 1 つは一番 Root Bridge に近いポートという意味で Root Port がある。Root Port は必ず FORWARDING というパケットの送受信ができるステータスとなる。

ある LAN セグメントに複数の Bridge が接続している場合、Root Path Cost が小さいものがその LAN における Designated Bridge と呼ばれ、このポートは Designated Port と呼ばれて FORWARDING ステータスとなる。

Root Path Cost が等しい場合は、Bridge ID の優劣（値が小さいほうがえらい）、Port ID の大小でタイブレークする。

Root Port でも Designated Port でもないポートは BLOCKING ステータスとなりパケットの送受信を行わない。

Hello Packet（通常 2 秒間隔）が Max Delay の時間（通常 20 秒）届かないと障害と認識する。Root Bridge が停止する場合やリンクが切れる場合もあるが、新しい状態で再度 Root Path Cost などを評価し、Root Port や Designated Port を再度選択する。

続く）

1. 2 Spanning Tree algorithm and Protocol

この時 BLOCKING から FORWARDING に変化するポート（新しくRoot Port もしくは Designated Port になった）はいきなり FORWARDING にはならず、絶対ループがあってはならないので、LISTENING という周りの言うことをしばらく確認するステータスを経由する。

LISTENING ステータスは Forwarding Delay （通常 15 秒）の時間を経過すると LEARNING ステータスというステータスに移行する。LISTENING も LEARNING もパケットの送受信ができない状態が続くが LEARNING 時は受信したパケットの MAC アドレスの学習プロセスは動作する。

こうして FORWARDING になった時にはある程度は学習が終了しており、無用なパケット転送は避けることができる。LEARNING ステータスも Forwarding Delay の時間が経過後 FORWARDING に移行する。

一般に STP の再構成で通信が途絶える、というのはこの

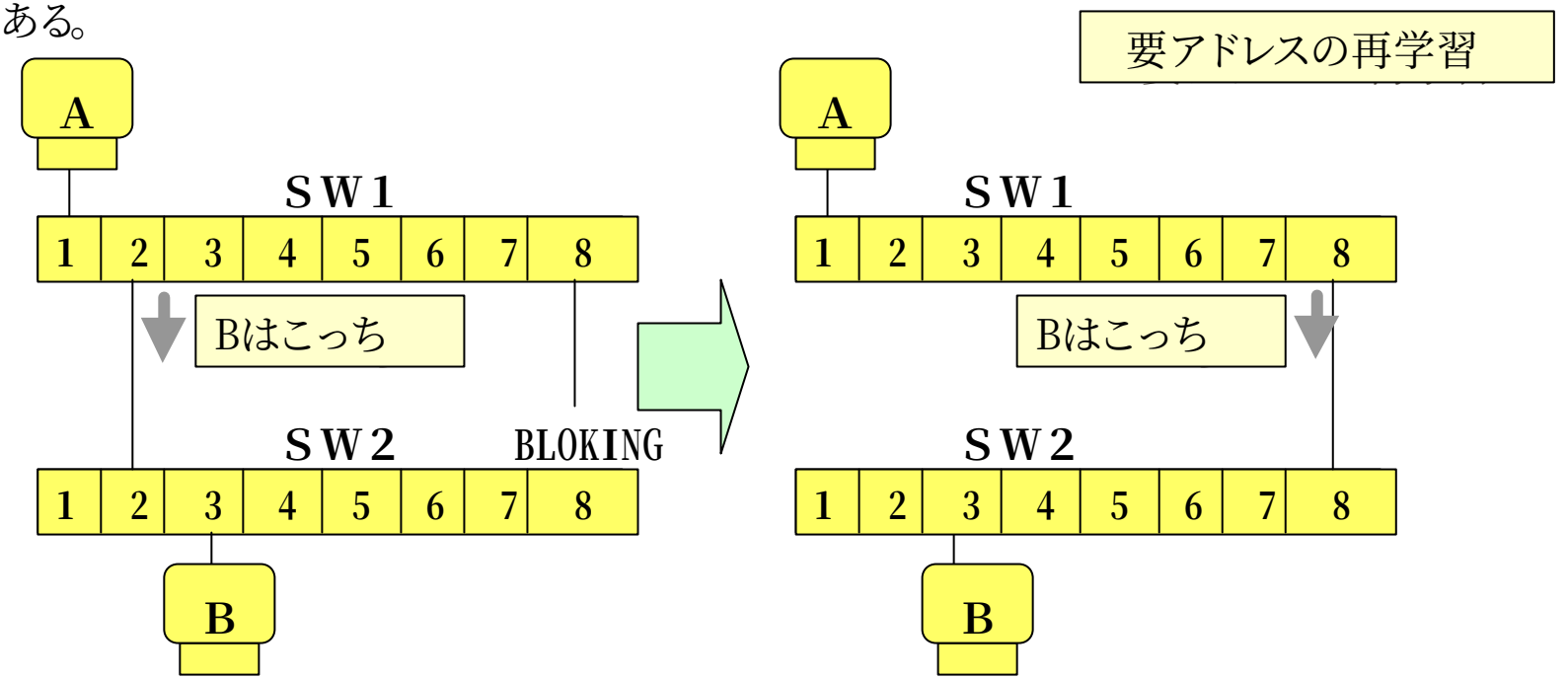
BLOCKING->LISTENING->LEARNING->FORWARDING

に要する変化を意味しており、通信が途絶えるのはこのポートを通る必要があるトラフィックだけである。FORWARDING ステータスのままのポートは再構成の前後でも通信が途絶えることはない。

続く）

1. 2 Spanning Tree algorithm and Protocol

トポロジーが変化すると、今までポート2にいると学習していた装置がポート8に現れることもある。



この問題はポートのステータス変化だけでは解決できない。

Topology Change Nortification BPDU を使う！

1. 2 Spanning Tree algorithm and Protocol

ポートのステータスが変わったらRoot Port からTopology Change Nortification BPDU を Root Bridge 方向へ送信する。上位の Bridge はこれを受け取ったらConfiguration BPDU (Hello パケット)の ACK ビットを立てて受け取ったことを伝える。
これを Root Bridge まで繰り返して Root Bridge に変化があったことを伝える。
Root Bridge は Configuration BPDU の Topology Change flag を一定時間立てて全てのBridge にトポロジーの変化を教える。

Topology Change flag が立っている間、Bridge は MAC アドレス学習テーブルの
aging time (通常 300 秒)
を
Forwarding Delay (デフォルト15 秒)
の時間に変更して早めに忘れる。

1. 2 Spanning Tree algorithm and Protocol

トポロジーの再構成に最大どれだけの時間がかかるか・・。

Max Delay + Forward Delay + Forward Delay = 20 + 15 + 15 = 50 秒

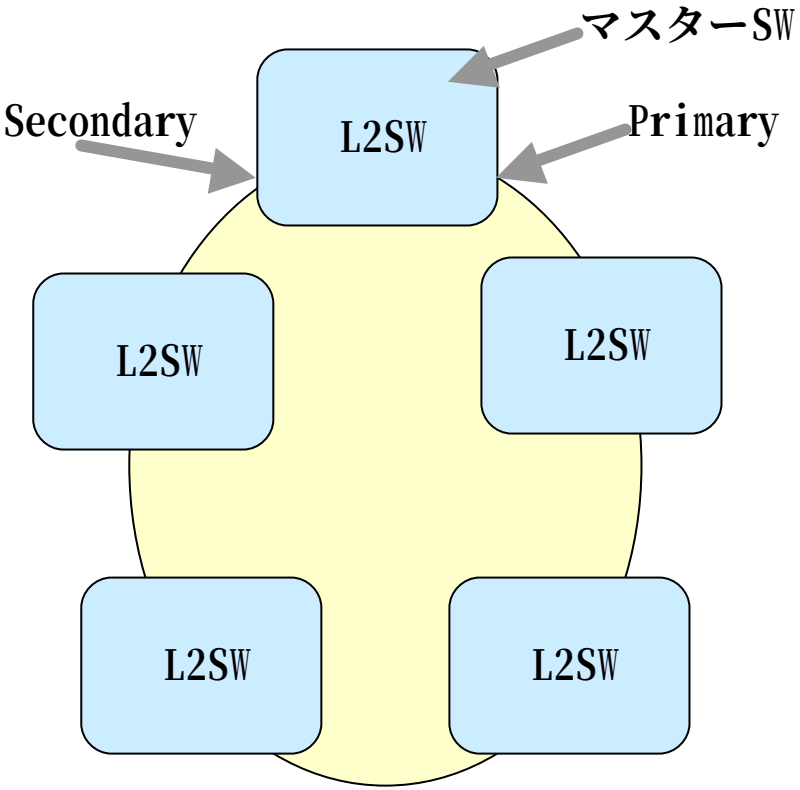
トポロジーがシンプルな場合、Hello を 1 秒、Max Delay を 6 秒、Forward Delay を 4 秒までは短縮できるが、それでも 14 秒はかかる・・。

時間がかかるのが STPの欠点！

PC やサーバがつながっているポートは？
STP が動いているとリンクがあがっても FORWARDING になるまでに 30 秒はかかる。
よって、DHCP などでアドレス取得に失敗する場合がある。
Vender によっては、対処策あり・・。

1. 3 EAPS (Extreme)

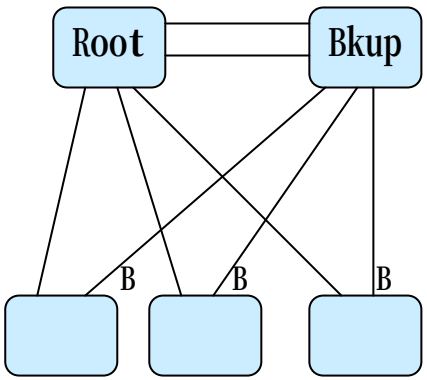
RING トポロジーで高速に切り替える。MAN サービス向け。



RING 内でマスターSW を選ぶ。
マスターSW の一方を Primary、もう一方を Secondary とする。Secondary を BLOCKING する。
Primary から RING に対して Hello パケットを投げ、一定時間内に Secondary に戻ってこなければ障害を検知する。
また途中の SW は障害を検出すると TRAP をマスターSW にあげることができ障害をより早く(1秒未満)に検出することもできる。
障害を検出したらSecondary をすぐ FORWADING にする。
トポロジーが変化したらMAC アドレスの学習テーブルは一旦 flash する。

1. 4 UplinkFast (Cisco)

下位 SW で上位と2 本で接続する構成。1 本が FORWADING、もう1本は BLOCKING。



Core SW

Access SW

この辺で使う

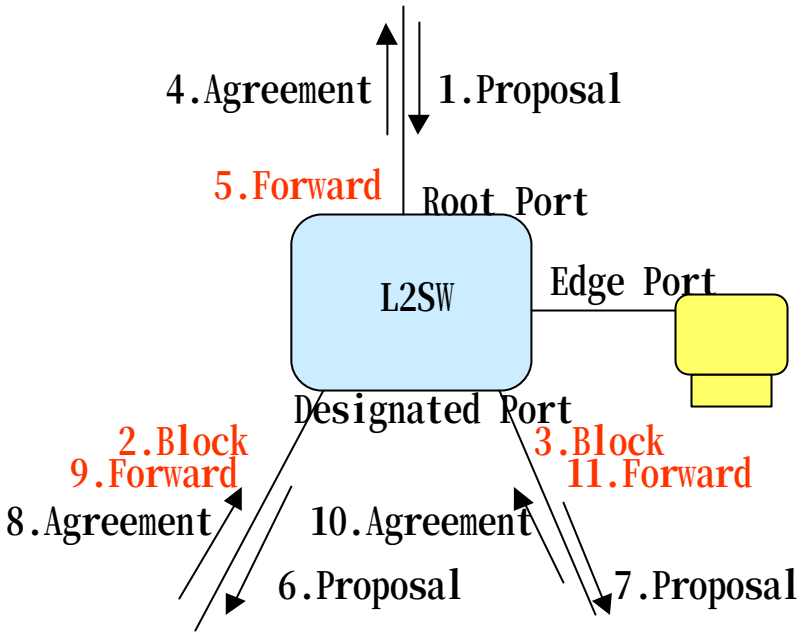
設定すると自動的に Bridge priority が低くなり、絶対に Root Bridge にならないようになる。
FORWARDING のポートで障害を検知すると、BLOCKING のポートをすぐFORWARDING に変更する。
他の Bridge の MAC アドレス学習テーブルの更新を助けるため自分が持っているアドレスを Source アドレスとしてマルチキャストを送信する。他の Bridge はこれを受信して再学習する。

1. 5 その他（Cisco）

Port Fast	端末（サーバ）しかつながっていないポートを指定。LISTENING、LEARNING ステータスを飛ばします。
Backbone Fast	ちょっとだけ切り替わるのが早くなります。
BPDU Guard	Port Fast の設定のつもりが BPDU が飛んでくると、おかしいということで自動的に切り離します。誤接続によるトラブルを回避。
BPDU Filtering	BPDU を filter します。STP が機能しないのと同然なので注意。
Root Guard	サービスプロバイダが L2 のサービスを提供するときなど、ユーザの L2SW が Root にならないように防ぎます。
Loop Guard	よくわかりません。すいません(^^;

1. 6 RSTP

IEEE802.1w で規定されている。Rapid STP という名のとおり、高速に再構成することを目的としている。1 秒以下で切り替わることも可能。



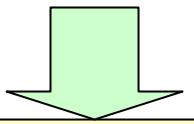
上位の SW から Proposal を受信すると Edge Port 以外は全て Block し、Agreement を返す。Block しているため loop はなくなるので、上位の SW はすぐ Forwarding に移行する。

下位の装置に Proposal を送信し・・、と続いていく。

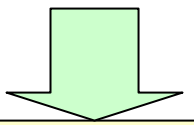
1. 7 MSTP

IEEE802.1s で規定されている。
VLAN を使用している場合、802.1D では STP のインスタンス（プロセス）は 1つでよかった。というか 1つしか定義されていない。このため、VLAN ごとにトポロジーを変えたい場合、Vender 独自の拡張に頼っていた。

VLAN ごとに別々の STP インスタンスを動作させる・・、と多くの VLAN を使った場合に STP の処理が重くなる。。



MSTP では複数の VLAN を 1 つの STP インスタンスにマッピングできる。また複数の STP インスタンスの情報を 1 つの BPDU で送信することができる。



STP インスタンスの数を減らすのがメリット。Region の概念もある。
RSTP と一緒に使う。

2. L2SWに関する技術

2. 1 loop free（1章で説明）
2. 2 VLAN（802.1Q）
2. 3 default gateway redundancy（HSRP、VRRP、ESRP）
2. 4 QOS（802.1p）
2. 5 link aggregation
2. 6 traffic mirroring、switched port analyzer
2. 7 traffic rate limit
2. 8 filtering
2. 9 flow control 802.3x
2. 10 port based authentication（802.1X）
2. 11 broadcast storm control
2. 12 802.1Q tunneling

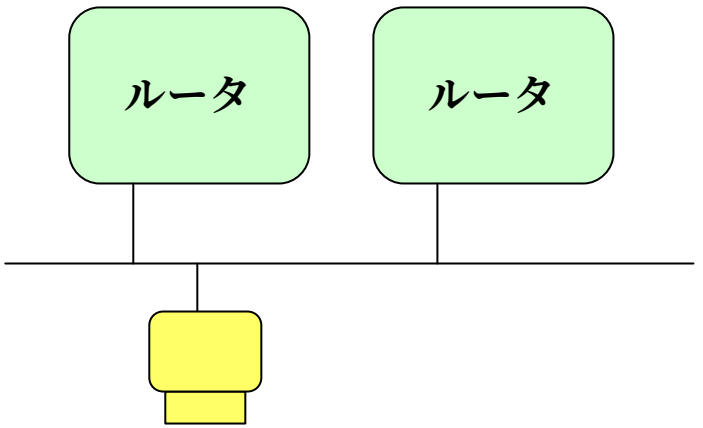
2. 2 VLAN (802.1Q)

さすがにこれは説明不要だと思うので省略します(^^;

パケットに VLAN タグを付加し、VLAN ID で論理セグメントを識別する技術。

2. 3 default gateway redundancy (VRRP、HSRP、ESRP、FSRP)

厳密には L2 ではないが・・。



PC などルーティングプロトコルを動かさないものは、default gateway 設定で外部とつながる。

ルータが 2 台あるので、二重化したい・・!

方法は、
- default gateway を複数書く
- proxy arp を使う
- ICMP でルータを探す

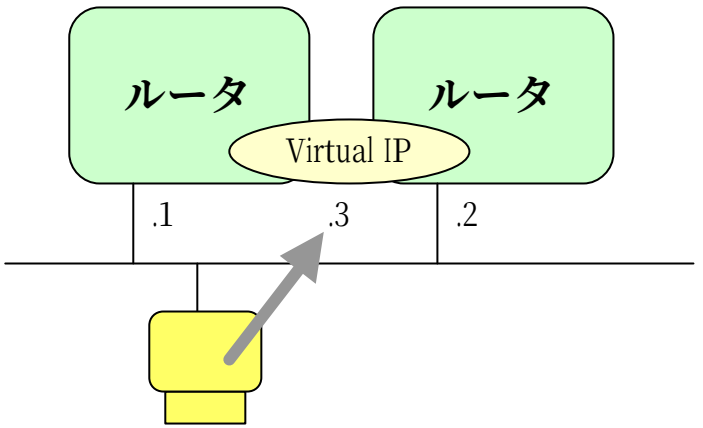
どれもイマイチ・・。なので・・

VRRP (RFC2338)
HSRP (Cisco)
ESRP (Extreme)
FSRP (Foundry)

を使おう!

2. 3 default gateway redundancy（VRRP、HSRP、ESRP、FSRP）

HSRPを例に取ると、



ルータにはまず普通に IP アドレスを付与する。
Virtual IP アドレスを重ならないように付与する。
ルータは、Hello パケットでお互いの状態を確認する。また、Priority を交換してどちらが Active になるのかを決定する。
Active なルータが Virtual IP アドレスの動作を受け持つ。
PCは、default gateway に Virtual IP を設定する。

Active ルータの MAC アドレスは HSRP 用の計算で作られたものが設定される。つまりハードウェアに割り当てられたものは使わない。

Standby が Active に変わったときは？
ルーティングプロトコルの Source IP は？
ルータは動的ルーティングプロトコルが必須

VRRP には、3 つ目のアドレスを必要としないモードもある。完全な Act-Stdby

2. 4 QOS (802.1p)

802.1Q のタグフィールド内に 3bit のプライオリティを示すビットがある。

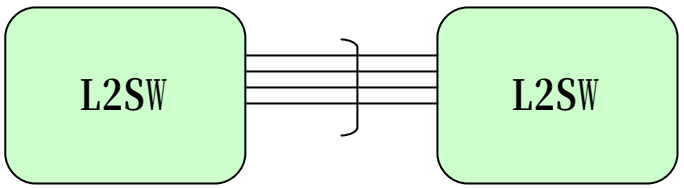
L2SW はこの値を元に処理を変えることができる。

- traffic rate の調整
- 複数 Queue に分けて優先制御

L2SW によっては、Queue が 8 つまでなく4 つという実装もある。

2. 5 link aggregation

100Mbps もしくは 1Gbps が 1本では足りない場合、複数本を束ねて論理的に 1つのインタフェースとして扱う技術。

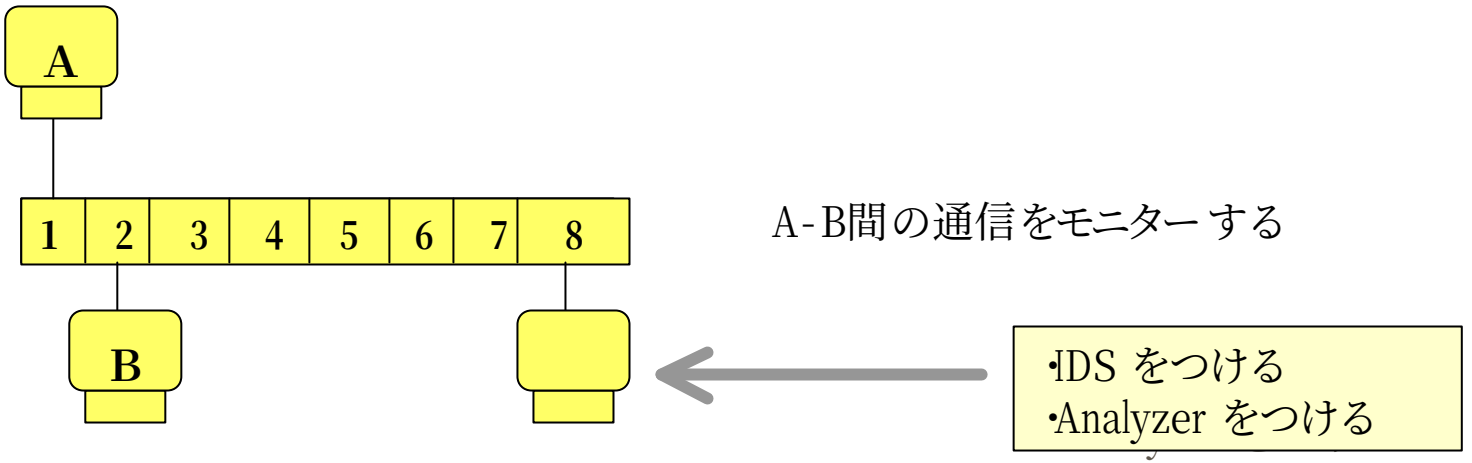


複数本に bit 分割するわけではなく、1 つのパケットはどれか 1 本のラインを通る。
どのラインを通すかのアルゴリズムは、source や destination の MAC アドレスで見る。このためトラフィックが均等にはならない。均等に近くなるかは、アドレスの数次第・・。

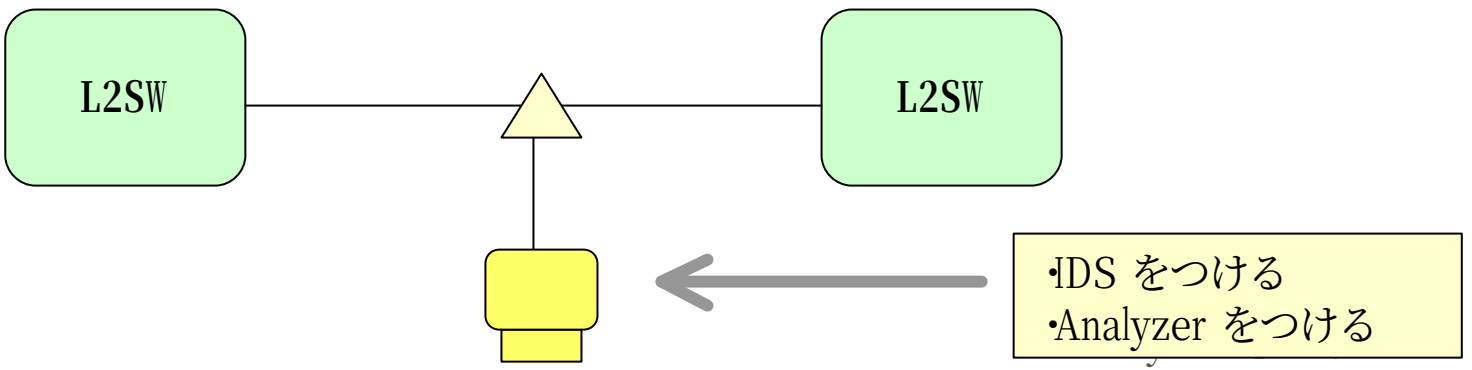
ハードウェアの構成に制約を受ける場合がある。(ポート1から4まで連続でないとだめ、など)。
異ベンダー間の接続は・・(auto negotiation 辺り)難しい場合もある。

2. 6 traffic mirroring、switched port analyzer

SW 内を流れるトラフィックを指定したポートに吐き出す機能。



TAP（分光器）もありますね・・。



2. 7 traffic rate limit

access-list でパケットを classify（類別）して、
指定した速度を超えていたら、
捨てる（policing）とか、ヘッダのプライオリティビットを書き換えるとか。

Input だけとか、Output もできるとか、機器によって仕様に違いがある。

2. 8 filtering

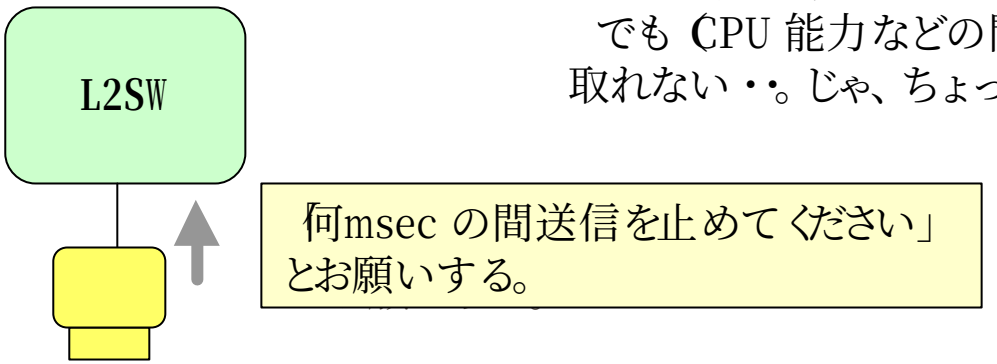
MAC アドレスやプロトコルフィールドの値を指定してフィルタリングする。

未知のマルチキャストは自動的でフィルタリングする・・という機能もありますね。
勝手にフィルタリングされると困る時もあるが・・)

IP のマルチキャストの枝刈をする機能も・・。

2. 9 flow control 802.3x

XON、XOFF みたいなもの・・？ ！

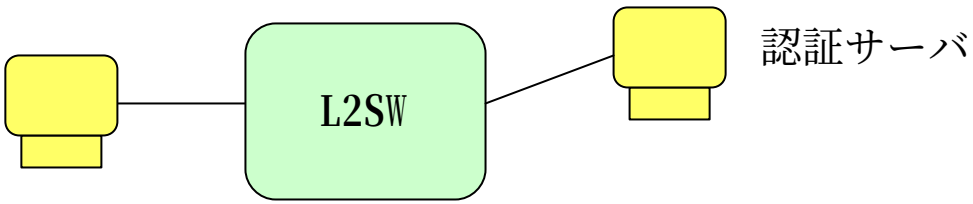


L2SW はワイヤーレートでパケットを送れる。
でも（CPU 能力などの問題で）PC or サーバは全てを受取れない・・。じゃ、ちょっと待ってもらおう

L2SW のバッファも無限ではないので、L2SW がパケットを取りこぼすことになるかも・・。
でも PC or サーバの CPU 使用率を上げて悲鳴を上げさせるくらいならネットワークで捨てたほうが効率がいいのかな？
QOS 機能を設定しているポートでは使わない方がいい。

2. 10 port based authentication (802.1X)

L2SW って誰でもつなげられるからセキュリティが心配・・。



L2SW に接続した場合、EAPOL (Extensible Authentication Protocol Over LAN) プロトコルを使って認証サーバに認証を受ける。
WindowsXP には実装されている。

2. 11 broadcast storm control

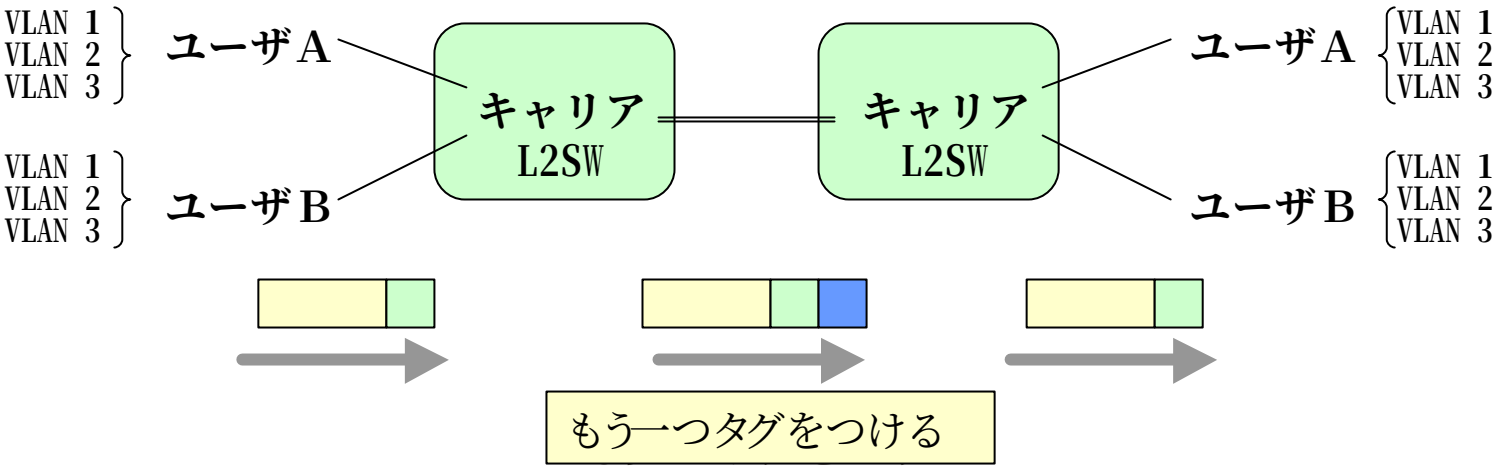
ループを排除すればブロードキャストは防げるか？

ハードが壊れた場合などに、ブロードキャストパケットが送信される場合がある。

ブロードキャストはポートのキャパシティーの何%にしよう、という設定ができる。

2. 12 802.1Q tunneling

キャリアの Ethernet 接続サービスで VLAN が使えるのはこれ？



複数ユーザを収容してもユーザ同士が任意の VLAN ID を付与できる。

L2SW 間の MTU が大きく取れないといけないので、二種事業者が同じ事をする場合は回線の仕様を調べる必要あり。

3. その他

■IEEE802の ドキュメント入手方法は？
http://standards.ieee.org/getieee802/portfolio.html

■ STPの参考文献は？
- IEEE802.1D の ドキュメント
- マニュアル （大体要は足りる）
- 「Interconnections: Bridges, Routers, Switches, and Internetworking Protocols」 by Radia Perlman