Ethernet Switch 技術～冗長性とループフリーの実現～

Internet Week 2002

2002／12／19
株式会社NTTデータ
吉野　誠吾

Ethernet Switch 技術～冗長性とループフリーの実現～

1. はじめに
2. 冗長性とループフリーの実現
3. L2SW 技術のおさらい
4. さいごに

1．はじめに

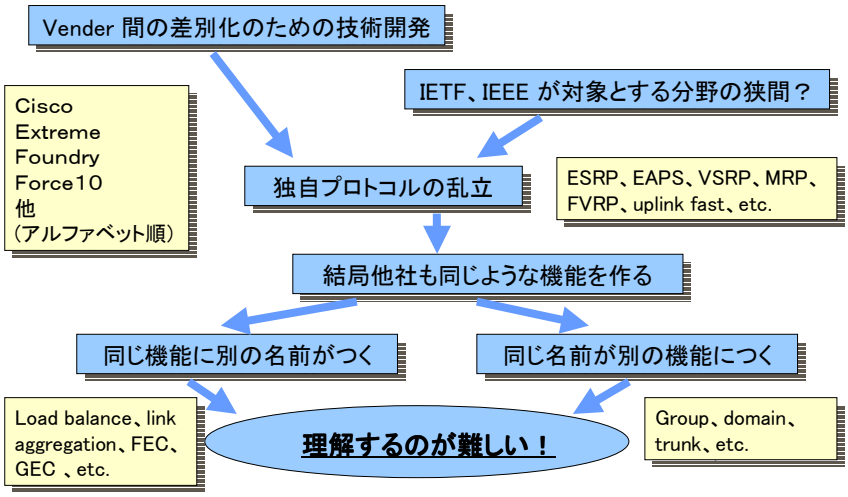
1．1　本チュートリアルの概要
1．2　動向
1．3　自己紹介

1．1　本チュートリアルの概要

Ethernet Switch（以下L2SW）はキャリア（第一種通信事業者）にも広域Ethernet サービス等で使われるようになっています。
これを支えているのは冗長性を確保する技術と vMAN に代表されるようなタグ付与技術です。
冗長性の確保はL2SWの弱点であるループという問題を排除することによって得られます。

本チュートリアルではループを排除して冗長性を確保する技術、それに加えて最近の複雑化するL2SW技術を解説することで、キャリアのみならずエンタープライズ（企業等）ネットワーク構築の参考としていただくことを目的としています。

1．2　動向



1．3　自己紹介

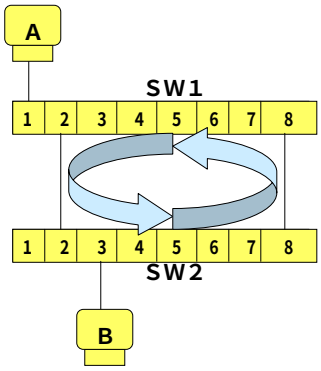
1991　NTTデータ通信株式会社（現：株式会社NTTデータ）入社
1992　現シスコシステムズ社との技術窓口担当（-1998）
1993　Cisco 初の ATM ルータ導入プロジェクトに参加
1994　CCIE取得（CCIE No.1234）
1994　ATMメガリンク最初の全国ユーザ？の立ち上げ支援
1998　Gigabit Ethernet 標準化前の沖縄での実証実験参加
1998　VoIP 等、社内評価を担当
2000　ISP サービスのネットワーク設計を担当（現職）
2002　JANOG10 での発表がきっかけでこの場に至る

2. 冗長性とループフリーの実現

2．1　ループは危険
2．2　STP
2．3　STP の拡張
2．4　メッシュトポロジー
2．5　リングトポロジー
2．6　リンクマネージメント
2．7　RSTP、MSTP

2．1　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



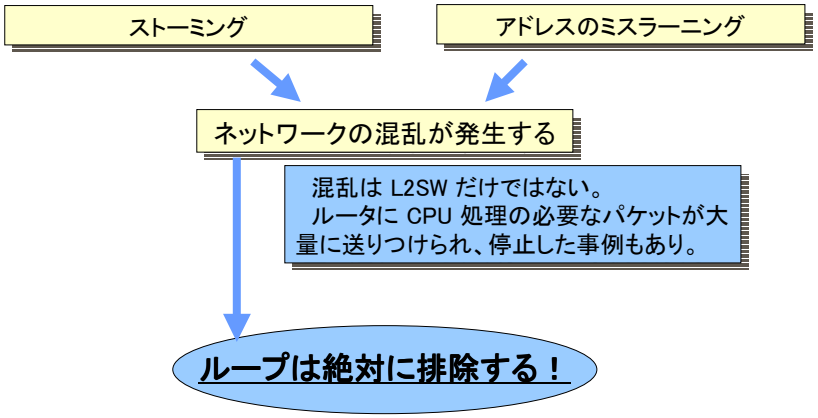
ストーム：未学習のパケットを全てのポートに送信するので、無限に回りつづける。

ミスラーニング：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。

再学習はCPU負荷も上げる

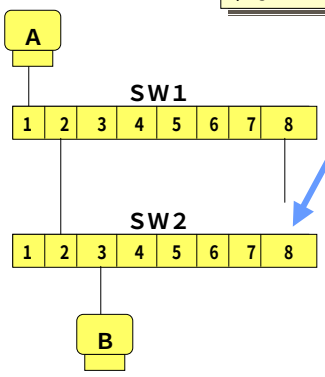
2．1　ループは危険

ループは絶対作ってはいけない。ループがあると・・。



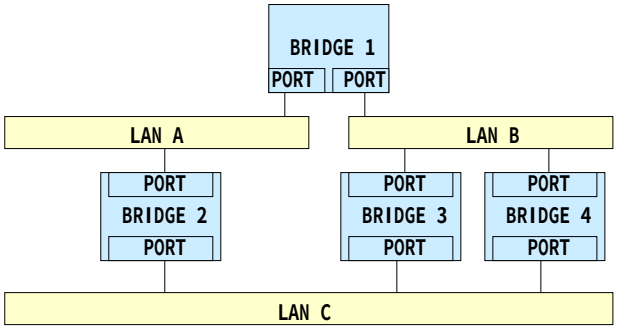
2．2　STP　Spanning Tree algorithm and Protocol

STPは、ループにならないように、ポートをパケットの送受信を行わない BLOCKING ステータスとするのが大きな仕事の1つ。



DEC で開発され IEEE で標準化された。
IEEE802.1D（1990 Edition、1993 Edition とあり、現在は 1998 Edition）の中で定められ、随時拡張されている。
Root Bridge と呼ばれる 1 つの装置をルートとした木(Tree)構造のネットワーク構成を作るプロトコル。

2．2　STP　説明用ネットワーク図



この例における LAN セグメントは L1 レベルのセグメント。ルータと L2 セグメントの関係ではないので注意。
また、Bridge=L2SW として Bridge という表現で説明。

2．2　STP　BPDU フォーマット

Configuration BPDU

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

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

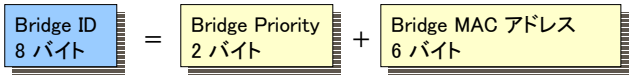
BPDU は Bridge Protocol Data Unit の略。Configuration BPDU は Hello パケットとも呼ばれる。

Flag は bit で表現し、Topology Change flag と、Topology Change Acknowledgement flag の 2 種類が定義されている。

2．2　STP　パラメータ

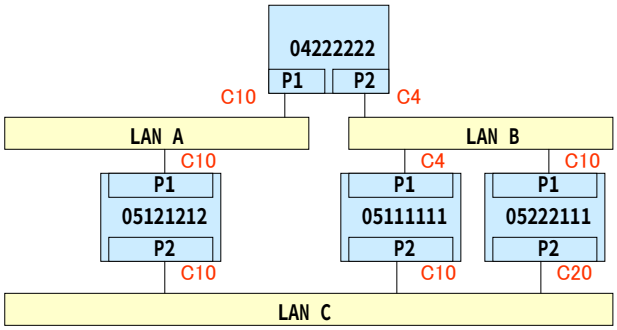
Bridge には以下の重要なパラメータがある。

Bridge ID	下記参照
Port Path Cost	高速なインタフェースほど小さい値となる
Port ID	ポート番号



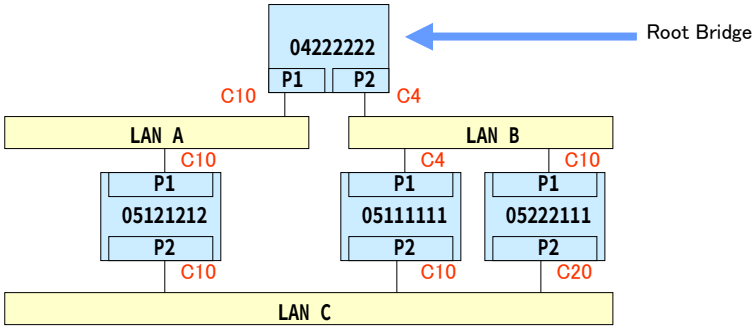
Bridge Priority と Port Path Cost は設定で変更できる。
Bridge ID は比較に用いられ、値が小さいほどプライオリティが高くなる。
Port ID も比較に用いる場合があり、値が小さいほどプライオリティが高くなる。

2．2　STP　例に値を入れる



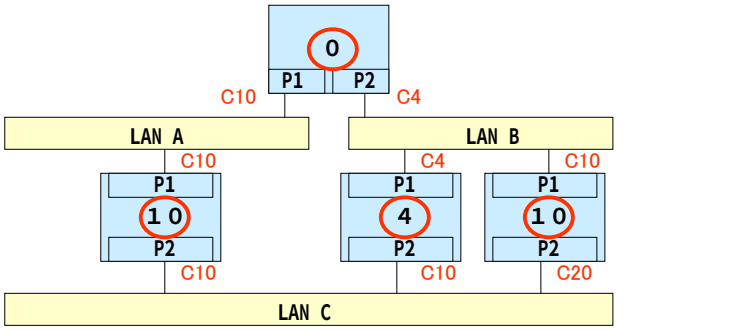
Bridge ID は簡略化した表記とした。

2. 2 STP Root Bridge を選ぶ



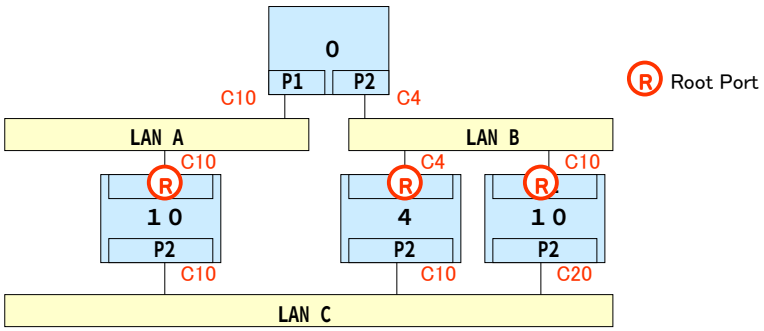
Bridge 間で Configuration BPDU をやり取りして、Bridge ID のプライオリティが高い（数値が小さい）BRIDGE 1 が Root Bridge になる。
自分が Root と信じている情報より劣った情報（Root のプライオリティが低い）を受信したときは、自分の情報を送り返してすぐに劣った情報を打ち消す。

2. 2 STP Root Path Cost を計算する



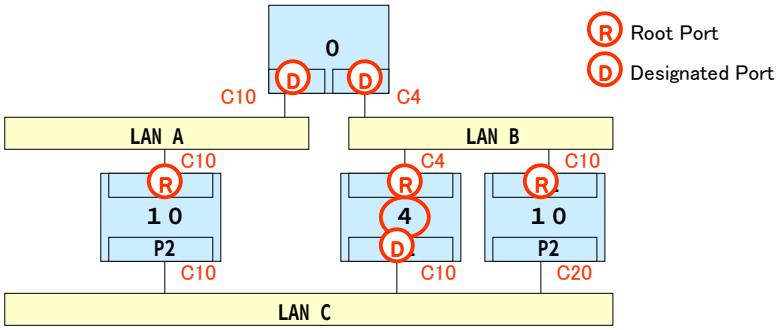
Root Bridge を 0 とした Root Path Cost を Configuration BPDU で送信。受信した各 Bridge で Port Path Cost を足し、受信した Bridge の Root Path Cost とする。

2. 2 STP Root Port を決める



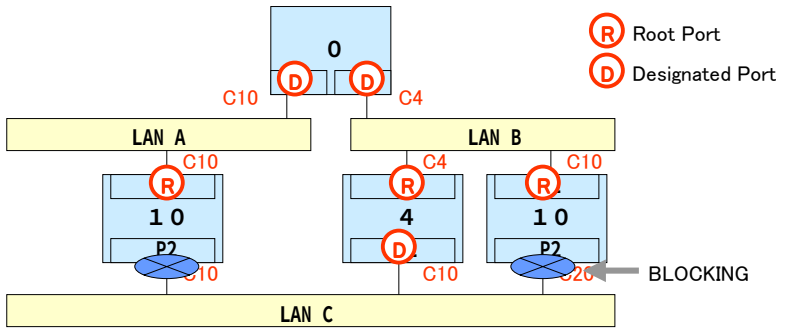
各 Bridge の Root に近いポートが 1 つだけ Root Port となる。

2. 2 STP Designated Port を選ぶ



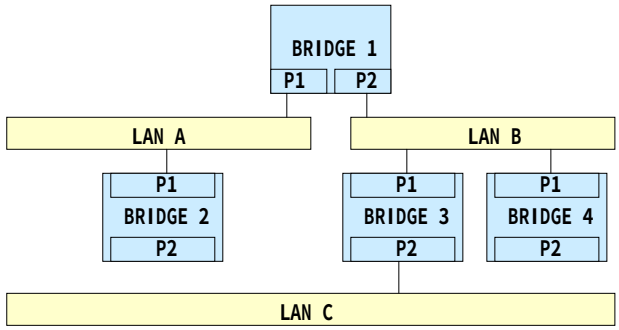
各 LAN で一番 Root Path Cost が小さいポートがその LAN の Designated Port となる。
Root Bridge のポートはすべて Designated Port となる(Root Path Cost が 0 だから)。

2．2　STP　Port ステータスを決める



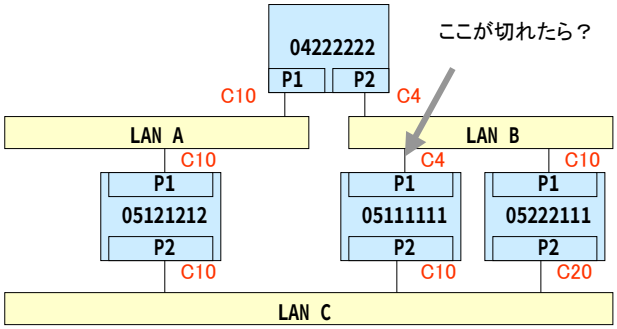
R や D ポートは FORWARDING、それ以外は BLOCKING というステータスになる。

2．2　STP　ループのない構成に安定する



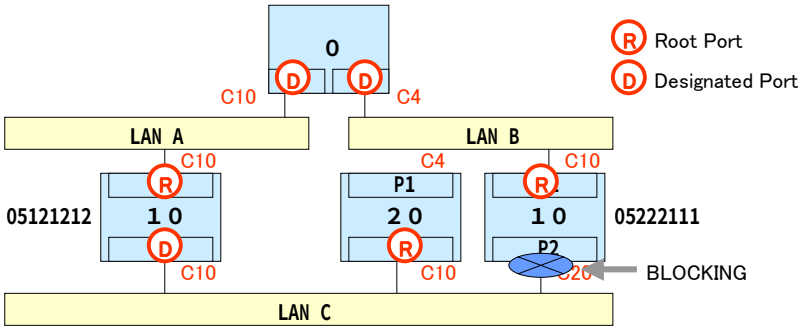
BLOCKING のポートはパケットの送受信を行わない。
これでループはなくなり、論理的には上図のようになる。

2．2　STP　ネットワーク故障発生



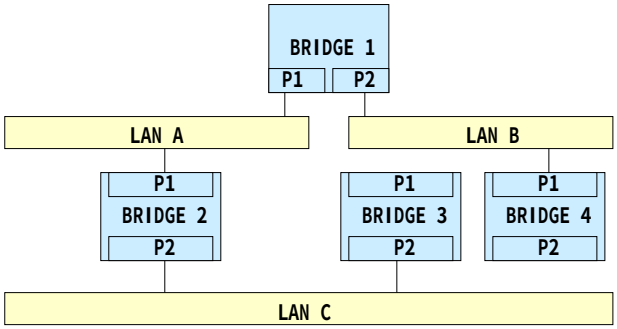
状態が変化したことを知るのは BRIDGE 3 だけ。

2．2　STP　再構成を行う



再構成を行う。BRIDGE 3 は Root Path Cost が 20 に変化。
LAN C において、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。
BRIDGE 3 は P2 が Root Port となる。

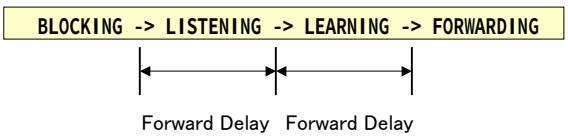
2．2　STP　安定する



結果はこうなる。
この時状態が変化したのは BRIDGE 2 と 3 だけであることに注意。

2．2　STP　ステータスの変換 (1)

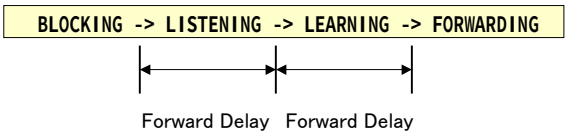
BLOCKING から FORWARDING に変化するとき（前述の BRIDGE 2 の P2）は、ループが完全になくなったことを確認するために時間をかけて変化する。このため、LISTENING、LEARNING という二つのステータスを経由する。



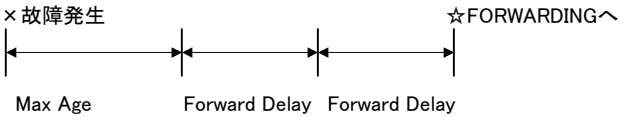
LISTENING も LEARNING もパケットの送受信は行わないのは BLOCKING と同じ。ただし LEARNING は MAC アドレスの学習は行い、FORWARDING に変わった際に無用に Unicast の flooding を起こさないようにしている。
Forward Delay は default 15 秒。

2．2　STP　ステータスの変換 (2)

Bridge の起動時は BLOCKING からスタートするので、最低 30 秒は通信ができない。



また、上位の Bridge から Max Age（default 20 秒）の時間の間 BPDU の更新がなかった場合、故障と判断する。このため、故障を検知してから FORWARDING になるには、最大 50 秒かかる。



2．2　STP　動作説明 (1)

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]
802.1D（1998 Edition）では 100Mbps は 19、1Gbps は 4、10Gbps は 2。
802.1T では 10Gbps が 2000、1Tbps が 20、10Tbps が 2 などと拡張されており、将来 802.1D に統合される可能性あり。

2．2　STP　動作説明 (2)

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 の入りロインタフェースの 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 Age 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）

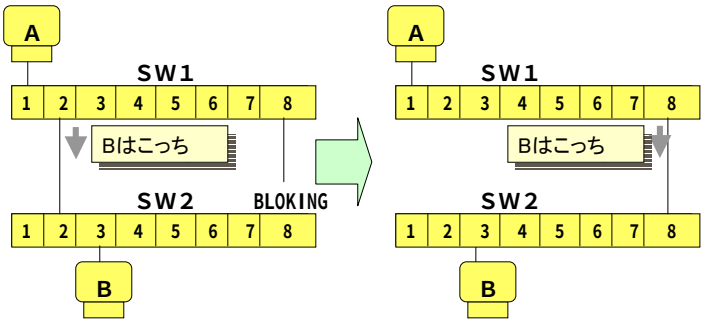
2．2　STP　動作説明 (3)

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 Forward 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 Forward 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。

2．2　STP　学習テーブルの更新（1）

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

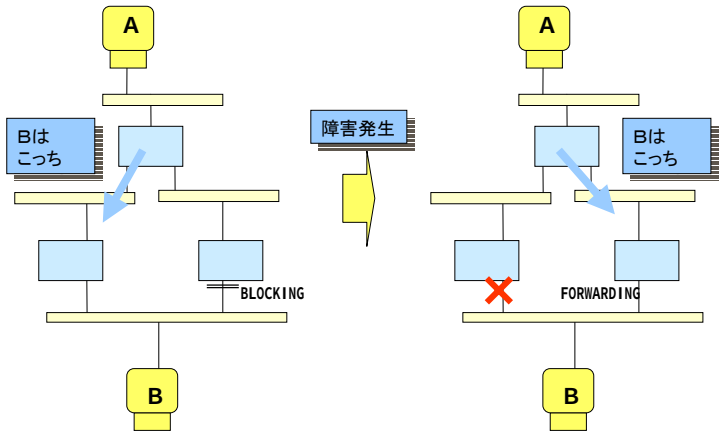


この問題はポートのステータス変化だけでは解決できない。

Topology Change Notification BPDU を使う！

2．2　STP　学習テーブルの更新（2）

SW 自身が検出できない障害でも再学習が必要な場合がある。



2．2　STP　学習テーブルの更新 (3)

ポートのステータスが変わったら Root Port から Topology Change Notification 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
Forward Delay（デフォルト 15 秒）
の時間に変更して早めに忘れる。

2．2　STP　まとめと補足

まとめと補足

・BLOCKING でループを抑える
・FORWARDING はループがないことを十分に確認する時間を経過してから移行する
・変化があった場合は MAC アドレスの学習テーブルを早く忘れる
という点がポイント。

障害時に最大50秒かかるのは長すぎる（STPの欠点）。設定変更で、Hello を 1 秒、Max Age 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・・。

新しいプロトコルを使う

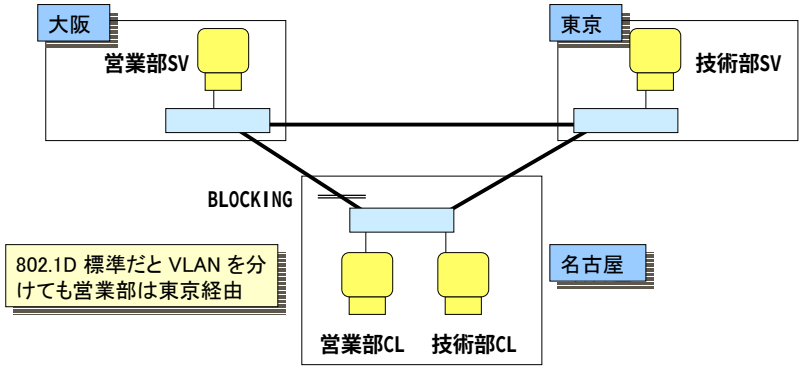
現在の L2SW では PC やサーバが直接つながる場合も多い。
STP が動いているとリンクがあがっても FORWARDING になるまでに 30 秒はかかる。よって、DHCP などでアドレス取得に失敗する場合がある。

STP を拡張する

ネットワーク全体で同期を取るため、タイマー値は Root Bridge の値を全ての Bridge が使う（設定値に関わらず）。Configuration BPDU で Root の値が送られる。

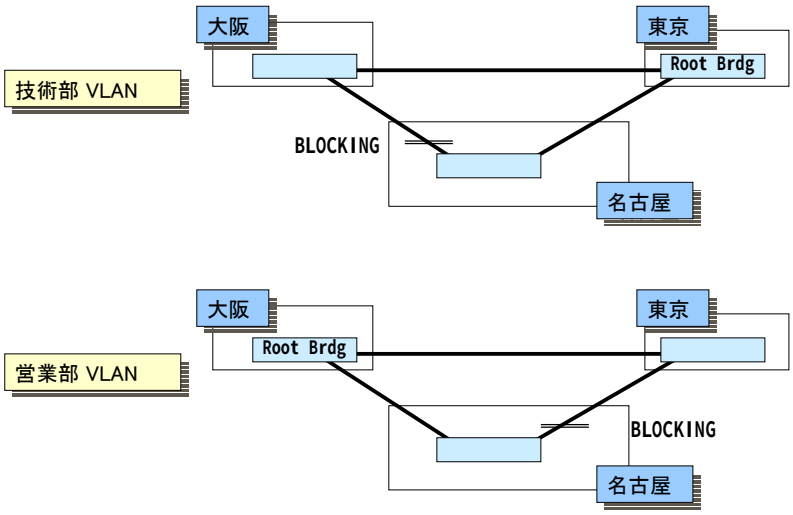
2．3　STP の拡張　VLAN と STP (1)

802.1D では VLAN に関係なく 1 つのトポロジーとなる。技術部 VLAN は東京中心(Root)、営業部 VLAN は大阪中心(Root)、というように VLAN ごとに違うトポロジーにできない。



vendor の拡張機能(PVST+)や IEEE802.1T、もしくは IEEE802.1S を使う。

2．3　STP の拡張　VLAN と STP (2)



2．3　STP の拡張　VLAN と STP (3)

PVST（Per Vlan Spanning Tree）はVLAN ごとに STP を動かすためのベンダー独自の拡張で、ベンダーによって実装が違うが、Cisco の実装にも対応している場合が多い。
Cisco の PVST は ISL トランク用に開発され、802.1Q の仕様に合わせるため PVST+ に拡張したもの（VLAN 1 の取り扱いが PVST と違う）。
802.1T は VLAN ごとに STP を動かすための実装方法等を標準化している。

PVST+ も 802.1T も VLAN ごとに STP のトポロジーを計算し BPDU も飛ぶので、多くの VLAN を設定すると重たい。
このため、いくつかの VLAN を 1 つのトポロジーにまとめ上げる技術もある（Cisco MISTP、IEEE MSTP）。

2．3　STP の拡張　VLAN と STP (4)

VLAN ごとに STP を動かす。この時 Bridge ID は VLAN ごとにユニークである必要がある。

↓

PVST+ では予め MAC アドレスを複数予約してある。これでは、
・数に制限があり、STP を動かせる VLAN 数が制限される
・MAC アドレスが無駄
という問題がある。

↓

802.1T は Bridge ID のプライオリティ 2 バイト（16ビット）を

プライオリティ 16ビット -> プライオリティ 4ビット ＋ VLAN ID 12ビット

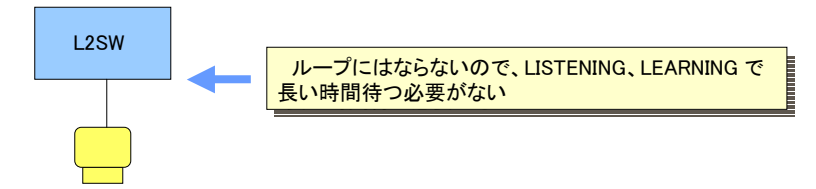
に変更し VLAN ごとに Bridge ID はユニークとなる。
数の制限はなくなるが、負荷の問題は残る。

2．3　STP の拡張　VLAN と STP (5)

STP は全ての VLAN で 1 つ	VLAN ごとに STP を動かす	いくつかの VLAN をグループ化して STP を動かす
802.1D や 802.1Q 標準	PVST+ などベンダー独自だが概ね互換性がある	MISTP（ベンダー独自のもの）と MSTP（802.1S draft)がある
IEEE 標準	現状では一般的。 設定が簡単。	多くの VLAN を少ない STP インスタンスで処理できる（スケーラビリティ）。
現状では現実的でない。機器のデフォルトが PVST なので。	多くの VLAN を処理できない。	標準化が終わっていない。 設定量が増える。

MISTP = Multiple Instance STP

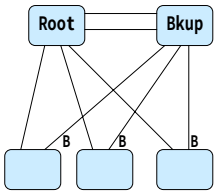
2．3　STP の拡張　port fast



実装1　ポートに設定することで、LISTENING、LEARNING を飛ばしてすぐ FORWARDING に移行する。このポートで BPDU を受信すると異常として処理する。

実装2　自動で検出。MAC アドレスは 1 つ、BPDU を受け取らない、link aggregation していない、等の時は Forward Delay を 2 秒にして動作する。

2．3　STP の拡張　uplink fast



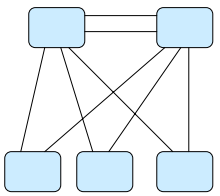
上位の SW は通常の STP を動かし Root Bridge とする。
左図の FORWARDING のポートで障害を検知すると、BLOCKING のポートをすぐ FORWARDING に変更する。

このようなエッジの SW で使用する

実装1	設定すると自動的に Bridge priority が低くなり、Root Bridge にならなくなる。 切り替え時、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。
実装2	切り替え時の Forward Delay を 2 秒に変更する。 Root Port に関連していた MAC アドレスはすぐに忘れる。

2．4　メッシュトポロジー　ベンダー共通事項 (1)

数百 msec ～数秒で切り替わることを目標としたプロトコル。
ベンダー固有なので、相互接続性はないが、高速に切り替えられる。



上位の SW に設定する機能

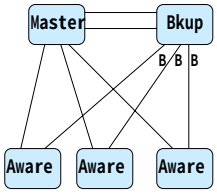
メッシュと言ってもこのようにデュアルホームの接続に限定されている。

ESRP：Extreme
VSRP：Foundry
FVRP：Force10

Cisco は RSTP 中心

下位の SW はプロトコルを理解できるもの（Aware という表現が使われる）の場合、より早く切り替わるが実装は必須ではない

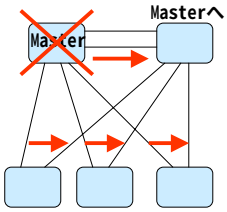
2．4　メッシュトポロジー　ベンダー共通事項 (2)



Master、Backup

Master と Backup（複数可。Standby という場合もあり）の役割がある。
Master がパケットの送受信を受け持ち、Backup は全て BLOCK するのでループにならない。
Master、Backup は VLAN ごとに存在する。

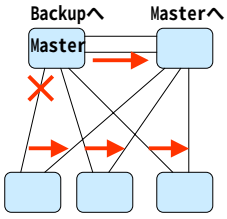
- Active ポート数
- プライオリティ（設定可）
- MAC アドレスの大小



障害時 (1)

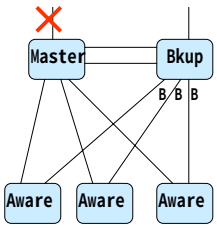
Master は keepalive を一定間隔で送信し、Backup は一定回数受け取れないと Master に昇格し、パケット転送の役割を引き継ぐ。
下位（Aware）の SW は Master との障害を検知すると即座に MAC アドレスの学習テーブルを削除もしくは Backup 側に変更する。

2．4　メッシュトポロジー　ベンダー共通事項 (3)



障害時 (2)

Active なポート数が多い方が Master になるので、Access SW とのリンクが切れても Master は移る。

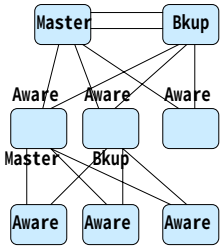


Tracking

指定した情報の変化（up->down 等）により priority 値を増減し、Master を選び直す機能。

Master のアップリンクが切れたら、右側の SW が Master になった方が効率がよいので、切れたらプライオリティが下がるように設定しておく。

2．4　メッシュトポロジー　ベンダー共通事項 (4)



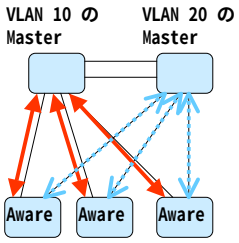
階層化

ネットワークを大規模化するため階層化し、上位のレイヤで Aware な装置が下位のレイヤでは各プロトコルを実行する装置となる。

VLAN のグループ化

MISTP に相当する機能で、複数の VLAN をまとめて 1 つのインスタンスで管理する。ベンダーによって、Group と言ったり、Domain と言ったりするので注意。

2．4　メッシュトポロジー　ベンダー共通事項 (5)



負荷分散

Master は VLAN（or VLAN グループ）ごとに選ぶことができるので、別々の装置を Master とすることで通常時において全てのリンクにトラフィックを流すことができる。

2．4　メッシュトポロジー　ESRP (Extreme)

L3 も	L2 だけでなく、L3 のバックアップ機能（VRRP のような）を提供する。
Tracking	選択肢が豊富（アクティブポート数、Tracking、プライオリティ、装置の MAC アドレス。Tracking は、リンク状態、IP ルーティングテーブル、ping、ハードウェアの動作状況、他）。
Host Attach	AFT 機能（後述）の NIC を持つ host を接続するための設定がある。
Domain	複数の VLAN をグループ化し、Master を共有する機能。
Group	ESRP を階層化するための機能。
Port Restart	Master が切り替わる際、下位の装置が Aware でない装置だと MAC アドレスの flash がすぐには起きない。これを促すために、短い時間意図的に link を落とす機能。下位の装置は link が落ちることで MAC アドレステーブルを flash する。

2．4　メッシュトポロジー　VSRP (Foundry)

L3 も	L2 だけでなく、L3 のバックアップ機能（VRRPE と同等）を提供する。
Tracking	リンク状態による Tracking 機能。
Traffic Group	複数の VLAN をグループ化し、Master を共有する機能。
Domain	VSRP を階層化するための機能。
タイマー値	100msec 単位での Hello など、ASIC 処理により高速切り替え
フラッシュしない	Aware な装置は、Master 側から Backup 側に切り替える際に、MAC アドレスの学習テーブルをフラッシュせずに、Backup 側に書き換える。

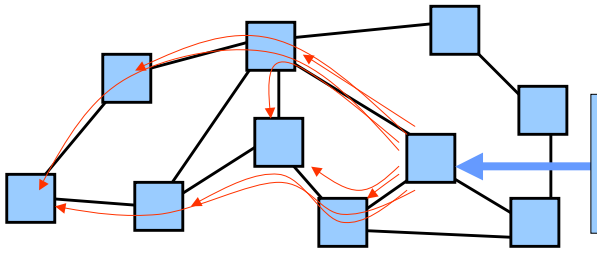
OS は 7.6.01（2002/11 リリース）以降を使う。
VRRPE を拡張した機能

2．4　メッシュトポロジー　FVRP (Force10)

Tracking	リンク状態による Tracking 機能。
VLAN Grouping	複数の VLAN をグループ化し、Master を共有する機能。
Hierarchical FVRP Domain	FVRP を階層化するための機能。
Link Flapping	Master が切り替わる際、下位の装置が Aware でない装置だと MAC アドレスの flash がすぐには起きない。これを促すために、短い時間意図的に link を落とす機能。下位の装置は link が落ちることで MAC アドレステーブルを flash する。
Core Loop Avoidance	Master は Backup との通信ができなくなったと認識した場合、Backup に対して Access SW 経由で keepalive を送ることができる（通常は直接のリンク間のみ）。このため何らかの不具合で Master は生きているのに keepalive だけが届かなくなり Master-Master の状態でループになるのを防ぐ

2．5　リングトポロジー　ニーズ

メトロネットワークを作る場合、拠点間のファイバは有限であり、中心拠点を中心としたスター状（メッシュ）のネットワークは作りづらい。隣接する拠点を接続し、中継する形態となる。
冗長化の考えを加えるとリングトポロジーとなる。

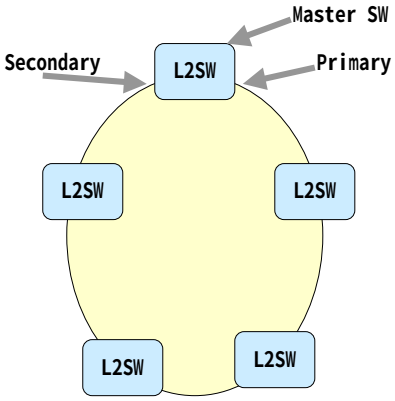


ここを中心拠点とした場合、リングトポロジーでないとこの拠点周辺のファイバは多く必要となる。

FDDI ネットワークからのリプレース。日本かつこれから、という意味ではそれほどない・・。

２．５　リングトポロジー　EAPS (Extreme)

リングトポロジーで高速に切り替える。MAN サービス向け。



リング内で master SW を選ぶ。
master SW の一方を Primary、もう一方を Secondary とする。Secondary を BLOCKING する。
Primary からリングに対して 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 master 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。

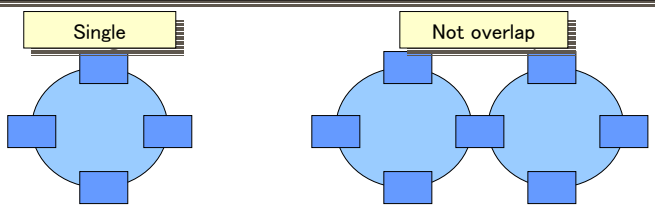
２．５　リングトポロジー　MRP (Foundry)

EAPS と動作はほぼ同じ(互換性はない)。
keepalive の間隔が ASIC 利用により 100msec 単位と短く、切り替わり時間は早い。

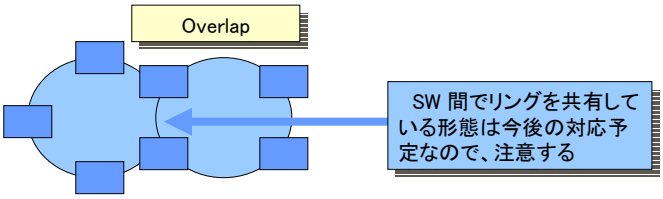
OS は 7.6.01(2002/11 リリース)以降を使う。
VLAN をグループ化する機能は MSTP(802.1S)ベース。

2．5　リングトポロジー　EAPS、MRP 共通

現状対応しているトポロジー



現状未対応のトポロジー



2．5　リングトポロジー　RPR 802.17 (1)

RPR（Resilient Packet Ring）はリングトポロジーにおける通信規格。
リング上の障害を検出して高速（50msec以内）に切り替える冗長性を提供する。
L2SW の世界ではまだ実装されている製品はない（or 少ない）。

10G Ethernet の物理層などを使い、低価格化する方向性もあるようなので、近い将来使える技術になっている可能性はある。SONET インフラにも対応。

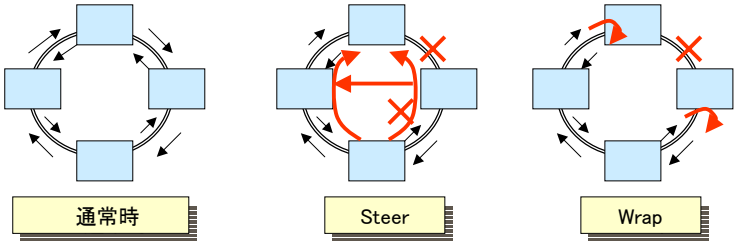
RPR Alliance（www.rpralliance.org）では 2003/3 の標準化を目指している模様。

障害検出、切り替えは RPR のレイヤで行うので、EAPS や MRP のようなプロトコルを SW が実装する必要はなくなる

価格次第か・・。

2．５　リングトポロジー　RPR 802.17 (2)

障害対策



Steer は障害部分を通らない方向に向きを変更する。Wrap は送る向きは変えずに、障害点を折り返して通信を継続する。

Steer はこの切替えの際に多少のパケットロスが出るが、遅延の変動は小さく、最適な経路を通るので効率がよい。Wrap はこの逆でパケットロスは少ないが、場合によっては通信経路が長くなり遅延が増える。

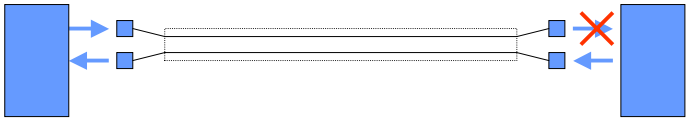
2．５　リングトポロジー　RPR 802.17 (3)

Class of Service	3 つのクラスが定義されており、帯域を確保する機能も定義される。
Spatial Reuse	宛先の SW に届いたら、パケットはそこでリングから取り除かれる。リングの帯域を有効活用できる。
Weighted Fairness	リング内で輻輳が起きた場合、トラフィック量が公平になるように各 SW の送信量を調整する。この時 CoS の設定によって、優先順位の低いものから落とされる。

その他、リングの最短経路を選択して送信する機能がある。どのようにしてパケットごとに最短の経路を計算するのかは？？？

2．6　リンクマネージメント

ケーブルは送信と受信が別の線になっている。これを挿し間違えたり、片側だけつながっていないと片方向通信の状態ができる場合がある。



本来右側の SW が BLOCK しなければならない状況でも、BPDU を受け取れないと検出できない場合がある。この結果ループが発生する。

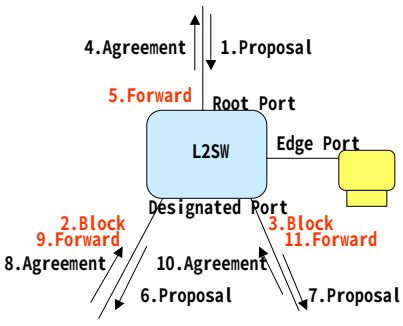
片方向通信は絶対にあってはならない。

auto negotiation（Gigabit Ethernet）を使えば、刺さっていない場合は反対側もリンクが上がらないが、別装置への挿し間違えは検出できない。

UDLD(UniDirectional Link Detection：Cisco)、Link Keepalive(Foundry)、FEFD（Far End Failure Detection：Force10）などを使う。

2．7　RSTP、MSTP

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。



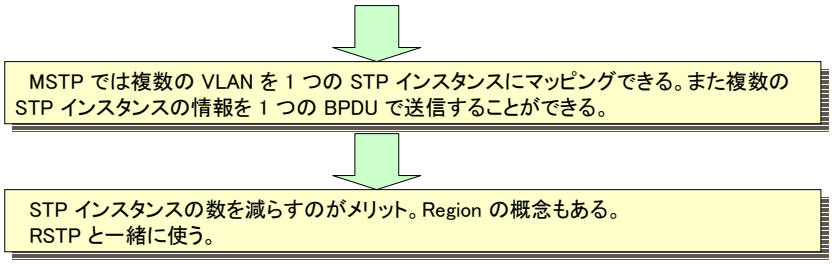
上位の SW から Proposal を受信すると Edge Port 以外は全て Block し、Agreement を返す。Block しているため loop はなくなるので、上位の SW はすぐ Forwarding に移行する。
下位の装置に Proposal を送信し・・、と続いていく。

ベンダーによっては RSTP の一部の機能（Root fast failover）だけしかサポートしていない場合があるので、注意。

2．7　RSTP、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。。

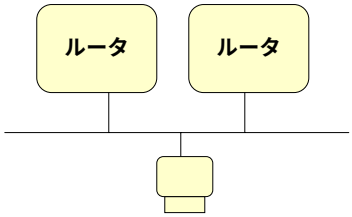


3．L2SW 技術のおさらい

3．1　default gateway redundancy
3．2　VLAN 802.1Q
3．3　QoS 802.1p
3．4　VLAN トンネリング技術
3．5　GVRP、VTP
3．6　link aggregation 802.1ad
3．7　flow control 802.3x
3．8　policing、shaping
3．9　packet filtering
3．10　port based authentication 802.1X
3．11　broadcast storm control
3．12　traffic mirroring、switched port analyzer
3．13　RMON
3．14　AFT
3．15　CWDM GBIC と ADM
3．16　10G Ethernet 802.3ae

3．1　default gateway redundancy (1)

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。

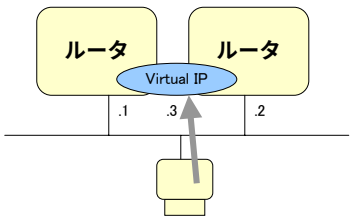
ルータが 2 台あるので、二重化したい・・！

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)
を使う

3．1　default gateway redundancy (2)

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し、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。

VRRP は、3 つ目のアドレスを必要としない。完全な Act-Stdby

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。

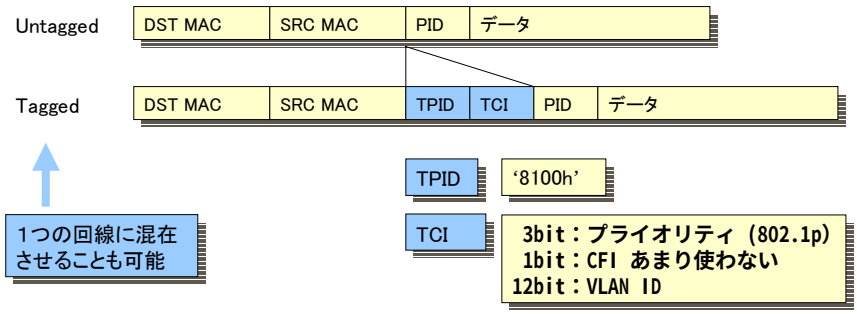
Standby が Active になった場合は？ → MAC アドレスが変わる。ARP REPLY を送りつける

ルーティングプロトコルの Src IP は？ → ルータの実アドレス

認証機能:
Man in the middle 対策

3．2　VLAN 802.1Q

ポートベースの VLAN を実現するための規格。
通常の Ethernet パケットに 4 バイトのヘッダがつく。
Tag つきのパケットが流れるリンクを VLAN トランクと言う。



VLAN ID は 0 は未使用、1 はデフォルト VLAN、4095 も予約されている。
ベンダーによってはそれ以外も予約されている場合あり。

3．3　QoS 802.1p

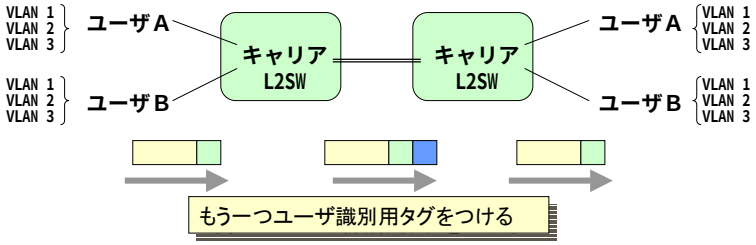
802.1Q ヘッダ内のプライオリティ bit の値（0-7）を元に、優先制御を行うことができる。 → ・パケット送信の優先制御
・Traffic policing
等

VoIP などのリアルタイム系トラフィックを優先したり、ということが可能。

L2SW によっては、Queue が 8 つまでなく 4 つという実装もある。この時値がどのように丸められるかは注意が必要。

3．4　VLAN トンネリング技術（1）

キャリアの Ethernet 接続サービスで VLAN が使えるのはこの機能のおかげ。



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。
Extreme の vMAN が最初だが、各社同様の機能を持つ。

3．4　VLAN トンネリング技術（2）

SW は MAC アドレスとユーザ識別用のタグ情報だけを見てスイッチングしている。

ユーザが付与した VLAN タグはスイッチング時には見ていない。

ユーザが複数の VLAN で VRRP を使用していたとする。

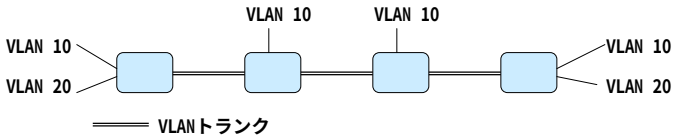
VRRP は Hello パケットのあて先 MAC アドレスとして、VRRP で決まった特定の値を使う。

ユーザ内では MAC + VLAN ID でユニークになるが、通信事業者内では MAC + ユーザ識別子となり、重なってしまい通信できない。

3．5　GVRP、VTP

IEEE は様々な情報を交換を汎用的に行う手順として GARP（Generic Attribute Registration Protocol）を規定した。

GVRP（Generic Vlan Registration Protocol）は GARP のアプリケーションの 1 つで、ネットワーク内にどのような VLAN が存在するかを動的に交換するためのプロトコル。

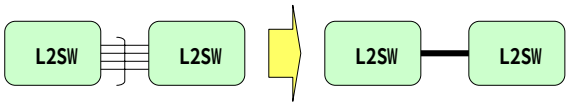


間の SW は VLAN 20 の存在を知らない。このため、VLAN 20 同士の通信ができない。
VLAN を作る度に全ての SW の設定を変更するのは特に台数が多い場合面倒なので、GVRP で情報を動的に伝える。
不必要なところまで伝えないよう（broadcast がネットワーク全体に流れないよう）に不要なところは pruning（枝刈り）する機能もある。

VTP（Cisco）は同じ目的の物。VLAN につけた名前の情報も交換できる。

3．6　link aggregation 802.1ad

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。ベンダーによって、機能名が違う。



ルータでも対応しているベンダーあり

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右表（ベンダー、機種にも依存）の通り。このためトラフィックが均等にはならない場合がある。
両端で違うアルゴリズムでも構わない。

- SRC MAC
- DST MAC
- SRC&DST MAC
- IP アドレス
- TCP/UDP ポート

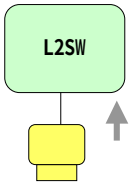
LACP　論理チャネルを組み上げるための動的プロトコルで、リンク状態のチェックや挿し間違えの防止にも利用できる。実装は必須ではない。

ベンダー間の相互接続性や、(R)STP との組み合わせでうまく動かない事例があるので、利用にあたっては注意する。

3．7　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



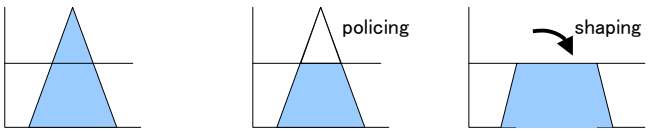
「何msec の間送信を止めてください」とお願いする。

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場合もある。
QOS 機能を設定しているポートでは使わない方がいい。

3．8　policing、shaping

トラフィックの状況を監視し、設定値を超えたら、

policing	パケットを廃棄する。
shaping	パケットをバッファに入れ、設定値の間隔でバッファから引き抜いて送信する。バッファがいっぱいの場合は廃棄される。



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。
policing は一般的だが、shaping できる機種は少ない。

広域 Ethernet で契約帯域が設定されている場合、網で policing されるのか shaping されるのかは確認が必要。Policing の場合は、ユーザが網に送る時点で shaping が必要。

3．9　packet filtering

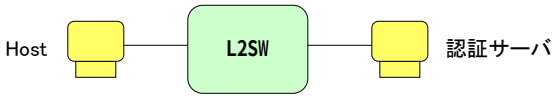
MAC アドレスやプロトコルフィールドの値を指定してフィルタリングする。

未知のマルチキャストは自動的でフィルタリングする‥という機能もある。
（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‥。

3．10　port based authentication 802.1X

L2SW は誰でもつなげられるものだったが、認証する機能を定義した。
無線 LAN などでも使われる場合がある。



Host は 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要求をあげる。
WindowsXP には実装されている。
L2SW は認証サーバに問い合わせ、認証されなかった場合は接続させない。

3．11　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？

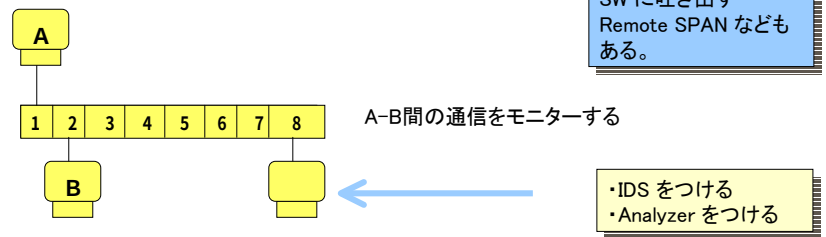
ハードが壊れた場合などに、ブロードキャストパケットが送信される場合がある。

ブロードキャストはポートの全体帯域の何%、という設定ができる。

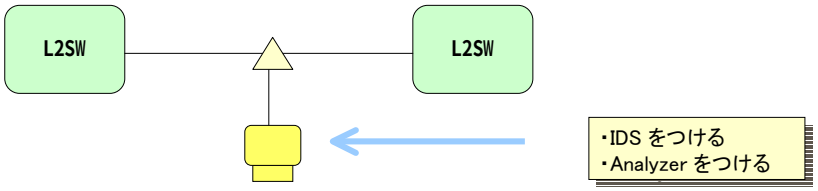
3．12　traffic mirroring、switched port analyzer

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。

VLAN を利用して別の SW に吐き出す Remote SPAN などもある。



TAP（分光器）でも可能。



3．13　RMON

RMON（Remote MONitering）は SNMP で L2 情報を取得するための規約。RFC2819。

L2SW は RMON エージェントとして動作する。RMON マネージャ（SNMP のマネージャにオプションになっている場合や、専用のソフトの場合もある）で以下の情報を取得可能。

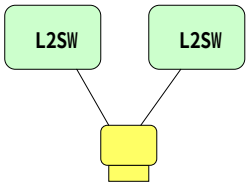
RMON Group

- ethernet statistics
- history control
- ethernet history
- alarm
- host
- hostTopN
- matrix
- filter
- packet capture
- event

SW は全ての RMON Group には対応していないのが普通。

閾値を設定して、超えたら SW から TRAP を上げさせることもできる。

3．14　AFT



AFT（Adapter Fault Tolerance）は、コンピュータに NIC を 2 枚挿して、論理的に 1 つのインタフェースとして冗長化する機能。

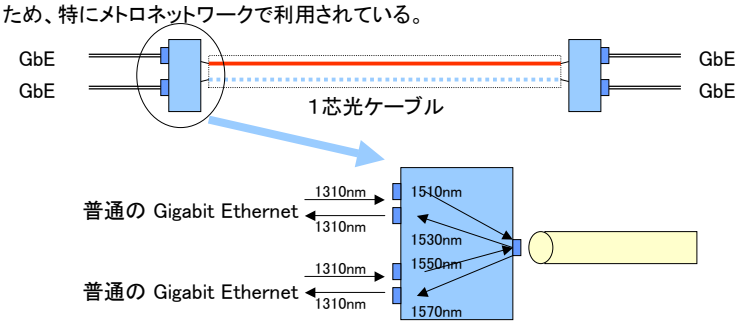
コンピュータ側ではパケット転送は行わないので、ループにはならない。

ESRP、VSRP、FVRP 等を設定していると、Master 以外の SW は BLOCKING になる。コンピュータ側が必ずしもこれらのアルゴリズムとは同期を取っていないので、Backup 側に送信する可能性もあり BLOCKING は不都合。

SW では以下のどちらかの対応が必要
- AFT 用の専用設定をポートに行う
- 該当ポートをアルゴリズムからはずす

3．15　CWDM GBIC と ADM (1)

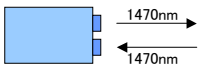
CWDM（Coarse Wavelength Division Multiplexer）は、通信チャネルごとに波長を変えて 1 つのファイバに多重して通す技術で DWDM に比べて波長の間隔が広く、安価に装置が作れるため、特にメトロネットワークで利用されている。



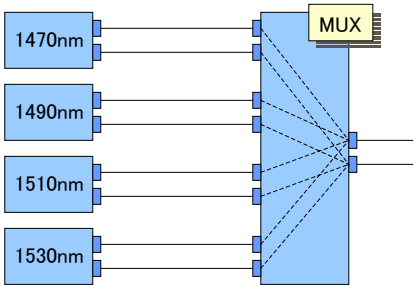
1 芯で Gigabit Ethernet×2 本を多重する CWDM の例

3．15　CWDM GBIC と ADM (2)

CWDM GBIC は通常の 1310nm ではなく、別の波長で送受信するように作られた GBIC。



波長が違う GBIC を並べて、1 つのファイバに多重する MUX を使うと、CWDM の装置と同等になる。

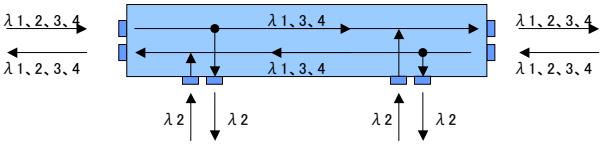


MUX は光学フィルタだけで実現できるので、電源不要で壊れにくい？！

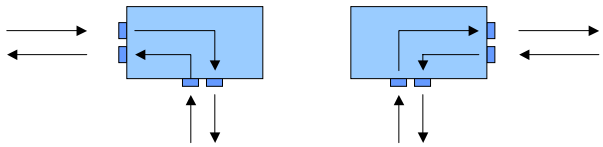
1470-1510nm の 4 波が多重されている。

3．15　CWDM GBIC と ADM (3)

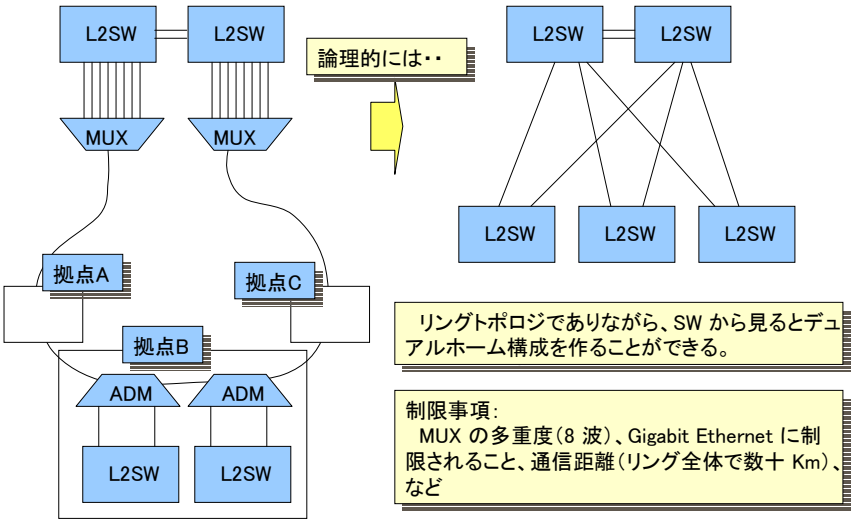
ADM は指定の波長だけを抜き出すためのフィルタ装置。



λ2 だけを見ると以下のイメージ



3．15　CWDM GBIC と ADM (4)



リングトポロジでありながら、SW から見るとデュアルホーム構成を作ることができる。

制限事項:
MUX の多重度(8 波)、Gigabit Ethernet に制限されること、通信距離(リング全体で数十 Km)、など

3．16　10G Ethernet 802.3ae (1)

LAN PHY	Short Reach	MM SM	数十～300m	波長は 850nm
	Long Reach	SM	10Km	波長は 1310nm
	Extended Reach	SM	40Km	波長は 1550nm
	WDM	MM SM	10Km(SM)	波長は 1310nm で複数の波長を多重して利用する
WAN PHY	Short Reach	MM SM	数十～300m	波長は 850nm
	Long Reach	SM	10Km	波長は 1310nm
	Extended Reach	SM	40Km	波長は 1550nm

LAN PHY（ファイ）は Ethernet フレームを使う。
WAN PHY は SONET フレームを使うので、SONET 設備があればそのまま長距離通信が可能。
Full Duplex のみ。

3．16　10G Ethernet 802.3ae (2)

GBIC のようなトランシーバが規格化されている。Hot swappable。今後の主流となるか？
・XENPAK アライアンス（www.xenpak.org）
・XPAK アライアンス（www.xpak.org）

当面は光ファイバのみ。メタルによる 10G は今後 IEEE で進められる予定だが時間がかかりそう？！。Copper はトランシーバが市販されている（25m まで）。
10G の次は 40Gbps か 100Gbps か未定。

XAUI、XGMII（10Gigabit Media Independent Interface）といった汎用インタフェースが定義されている。XAUI は 1000BASE-X 用のインタフェースの速度を 2.5 倍したものを 4 本並行にしたようなもの。

PC 用の NIC もある。

MAC レベルでは 10Gbps で、XAUI などを通して物理インタフェースと接続する。
LAN PHY は、10.3125Gbps を 64b/66bコードでエンコードして 10Gbps になる。
WAN PHY は、9.95328Gbps なので、Ethernet フレーム間の Inter Frame Gap の長さを調整する。

4. さいごに

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

■STPの参考文献
- •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

■ご質問は
yoshinos@nttdata.co.jp へ。