最近はやりのVLAN技術」
VLAN技術を使ったL2-VPNサービスの構築例
課題と今後（Web掲載用）

JANOG 10

2002/7/26

POWEREDCOM,Inc.

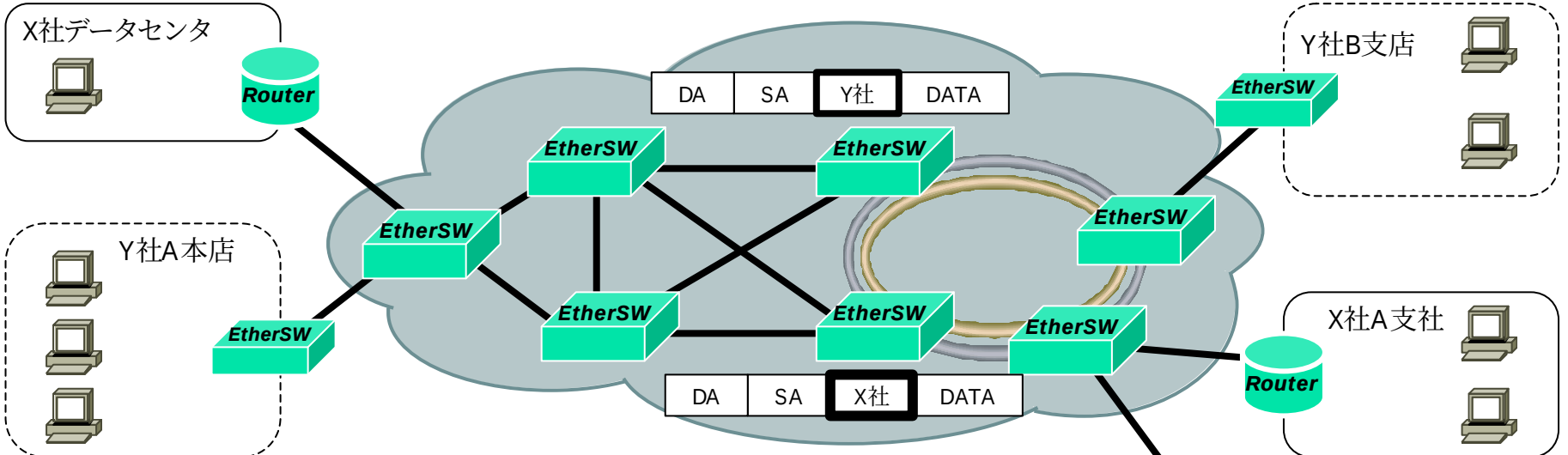
Masato Ando

1.4

VLAN技術を使った広域Ethernet 概要）

広域EthernetとはEthernetをWANにまで拡張したマルチアクセス型のVPN

- 現在提供されている広域Ethernetの多くが802.1Q TagVLAN技術を使って構築されている。
  (POWEREDCOMではEthernetSwitchとVLAN-tag(802.1Q) 技術を使って網を構築している。)



VLAN-tag技術により加入者ごとに論理的にネットワークを分ける

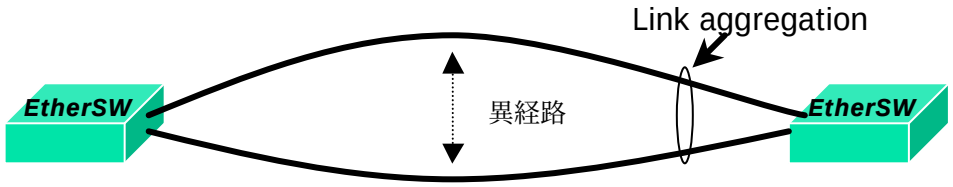
- その他の広域Ethernetの実現の方法
  - RPR(Resilient Packet Ring:IEEE802.17)
  - EoMPLS(Ethernet over MPLS:IETF ppvpn wg)
  - Etc…

※単一の技術で構築される場合もあるし、組み合わせて構築される場合もある。）

如何にして広域Ethernetに冗長性を持たせるか(1)

VLAN型VPNでどうやって信頼性の高いネットワークを構築するか？

- **伝送路の多重化の実現**
  - Link aggregation(802.3ad)によりリンク断対策と伝送容量の確保を行う。**(*1)**



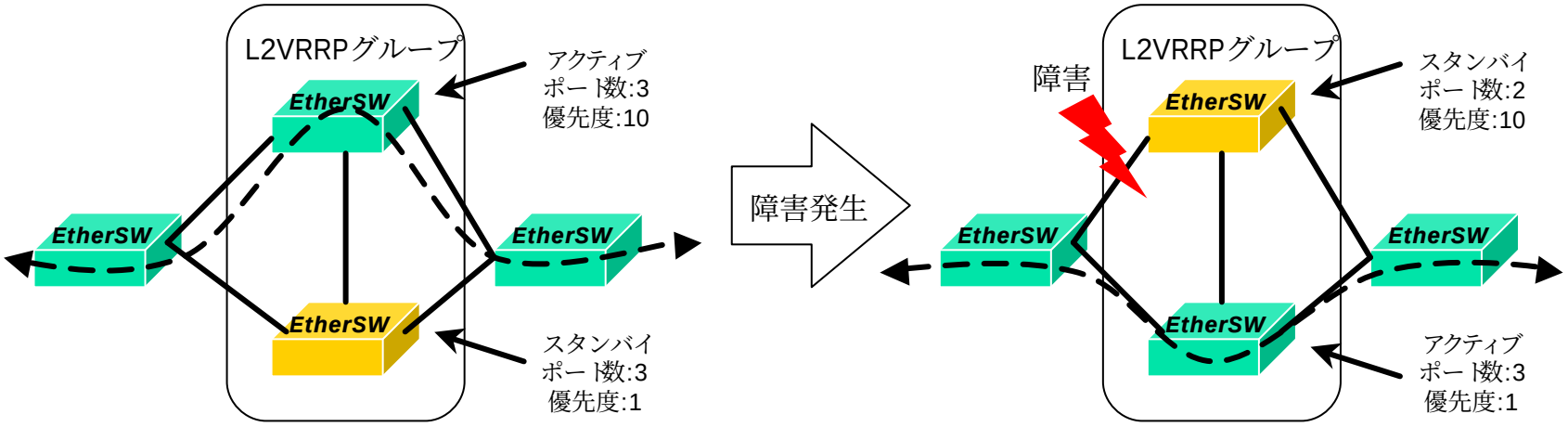
- **スイッチの二重化の実現**（バックアップを持ったままloopfreeなネットワークを作りたい）
  - STP(Spannig Tree Protocol)はどうだろう?
    - 標準だけれども、オペレーションしにくい
    - 当時802.1sのように複数のVLANをグループ化して、グループ単位でSTPの制御を行う方法がなかった。（加入者VLANに個々にSTPを走らせるのはCPUの負荷的にも論外）
  - スター構成でネットワークを作りたい。
    - L2で動作するVRRPに似た冗長化プロトコルを使ってみる…（便宜上ここではL2VRRPと呼ぶ）**(*2)**

(*1)最近は冗長化を持ったEthernet伝送装置もありそれらを利用することも出来る。

(*2)L2で動作するVRRPプロトコルは標準化されておらず、残念ながら今のところベンダ依存の実装を利用するしか方法がない。ESRP(extreme),VSRP(Foundry),FVRP(Force10),etc…その他、RPR、EAPSやMRPなどのリング型冗長プロトコルも最近は増えたのでそれらを使って、リングをスター状に展開する様な選択肢もある。

如何にして広域Ethernetに冗長性を持たせるか(2)

L2VRRPの基本動作の例（復習）

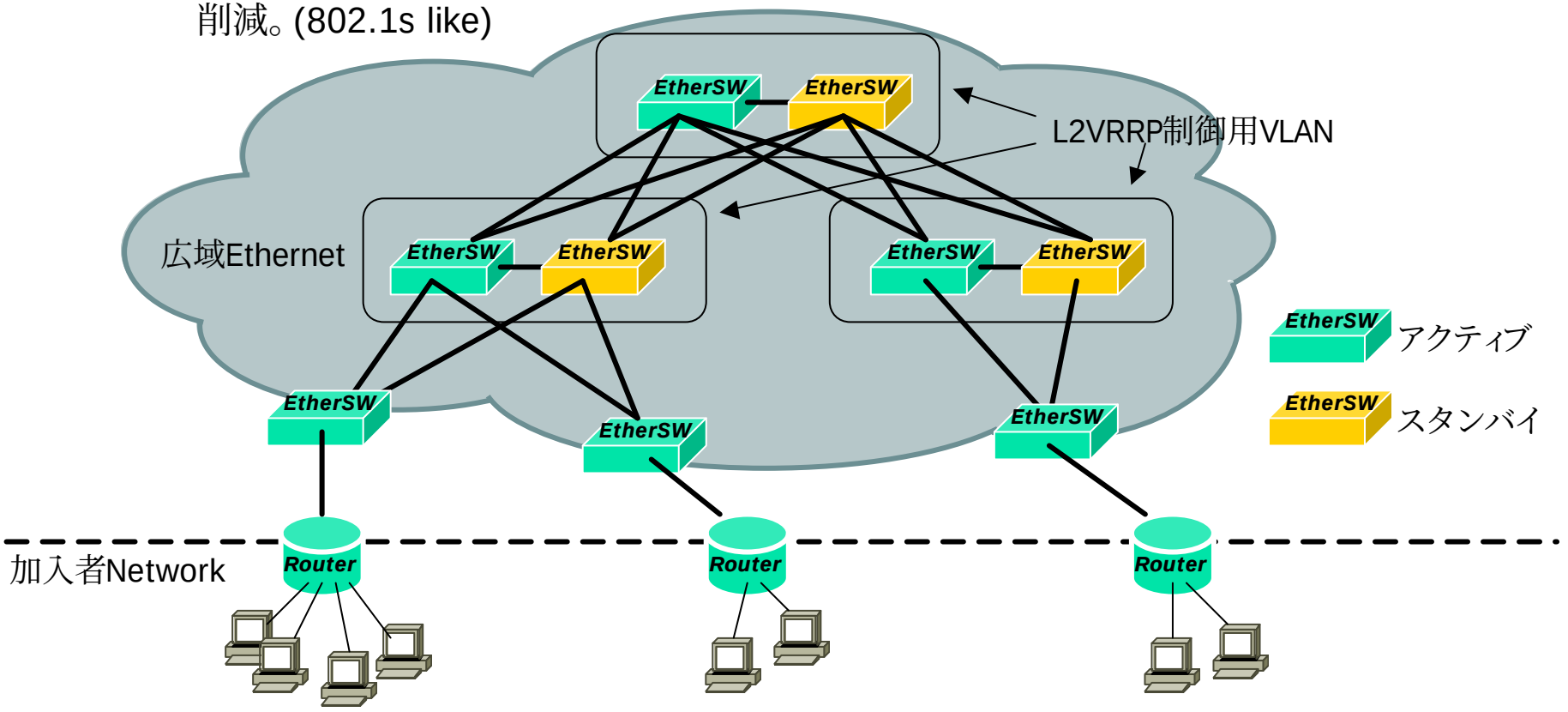


- 同じL2VRRPグループのスイッチの中で生きているポート数が一番多いスイッチ1個がアクティブとなり、他のスイッチはスタンバイとなる。（スタンバイのスイッチはフレームの転送を行わない）
- 生きているポート数が同じ場合は事前に各スイッチに設定した優先度を比較し、優先度が最も高いものがアクティブとなり、他のスイッチがスタンバイとなる。
- アクティブのスイッチが機能しなくなるとスタンバイのスイッチがアクティブになる。
- L2VRRPの切り替えが発生した場合には配下のスイッチでBridge Tableがリセットされる。

如何にして広域Ethernetに冗長性を持たせるか(3)

L2VRRPで巨大なEthernetSwitch網を構築する

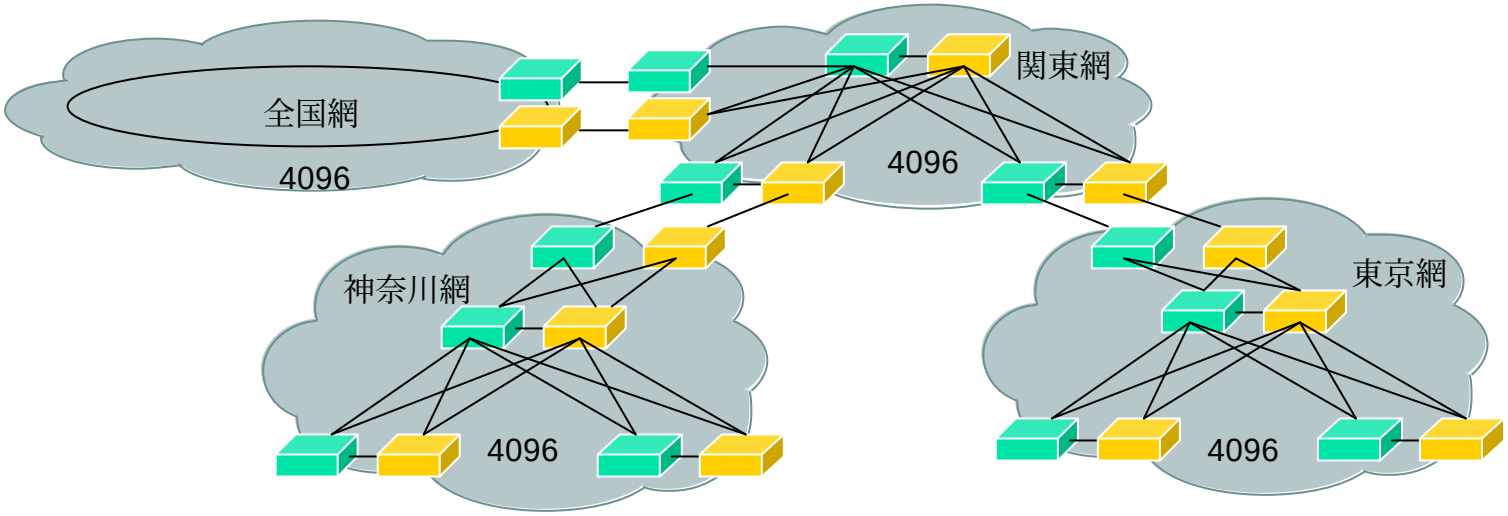
- L2VRRPのアクティブスタンバイの組を沢山繋ぎ合わせていけばよい。
- L2VRRP制御用VLANを作りL2VRRPの制御をまかせ、事業者設定VLANは主体的にL2VRRPの動作を行わず、代表のL2VRRP制御VLANの挙動にならう事により制御のオーバーヘッドを削減。(802.1s like)



広域Ethernetはどれくらいの規模までスケールするか？(1)

どのようにして4096の壁をやぶるか？

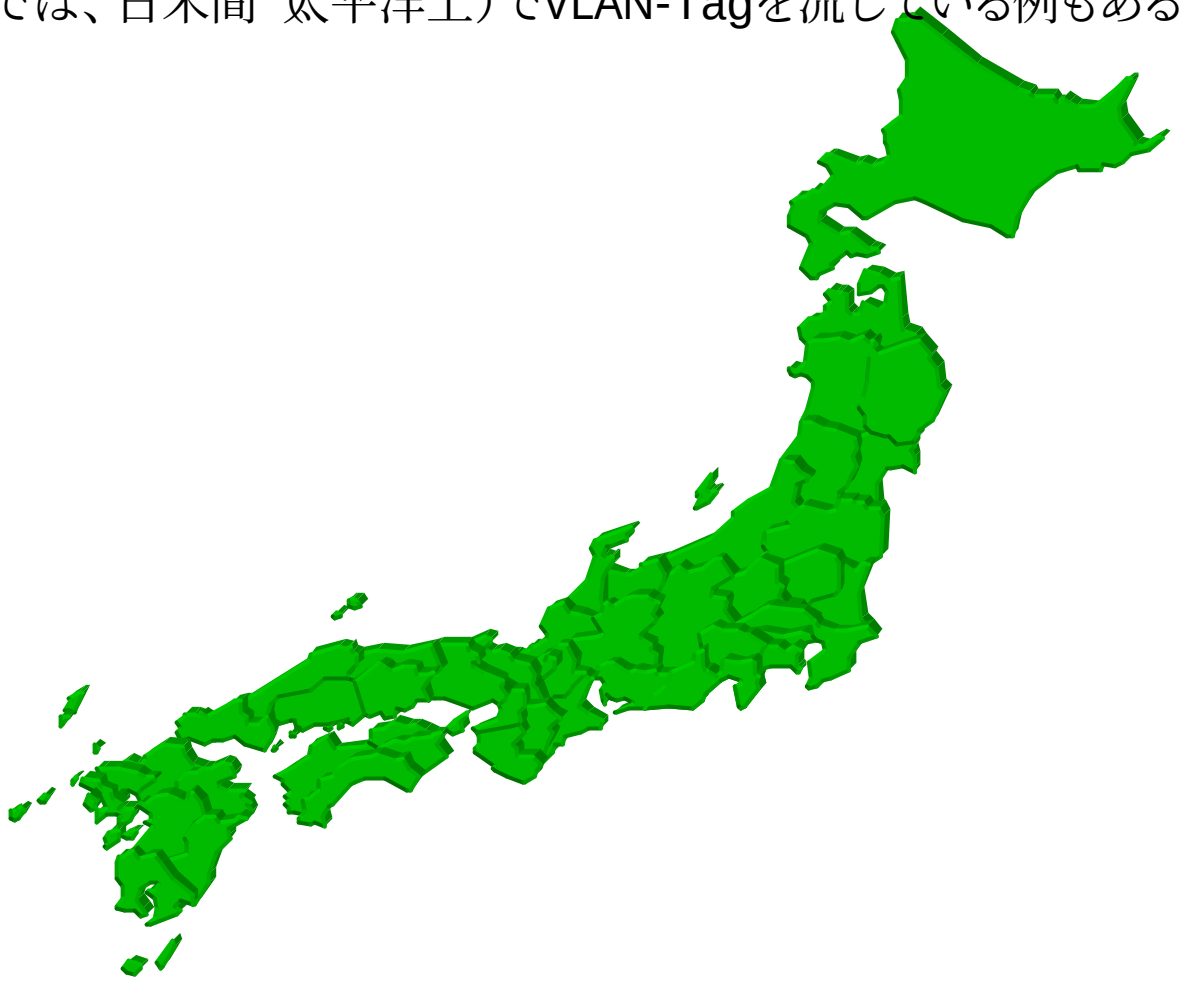
- 802.1QのVLAN技術を使っている限り、プロトコルの仕様上4096個のVLANしか作成出来ない。
- 網を地域ごとに分けて連結する事により、網全体として、4096以上のVLANを収容する。
- VLANが足りなくなったら新しい網をオーバーレイ。



※広域Ethernet VPN網はこんなに簡単に作れてしまう。(Simple is Best!!)

広域Ethernetはどれくらいの規模までスケールするか？(2)

- 日本全国の規模で実際に運用中。
- 数百台のスイッチで構成。
- 実験では、日米間（太平洋上）でVLAN-Tagを流している例もある。



現在のVLAN型VPNの課題（1）

VLAN型VPNは簡単で安価だが、いくつかの課題も抱えている

- **VLAN数の上限**
  - 802.1Q VLANをのタグの機構を利用している為仕様上(12bits)各ネットワークで4096個しかVLANを作れない。
  - ネットワークを分ける事によって、VLAN数を増やしていくにしても、ネットワークやスイッチの伝送容量やポート数に十分な空きがあるにもかかわらず、新しいネットワーク作らなければならない場合もあり、不経済。

- **Bridge Tableで学習出来るMACアドレス数の限界**
  - 広域Ethernet上のスイッチ、特にコアにあるスイッチのBridgeTableでは、VLAN数が多くなるそれに従って学習しなくてはならないMACアドレス数も多くなる。
  - BridgeTableがあふれ始めると、Unknown Unicast通信の割合が増加し、網としての性能の劣化につながる。（一般的にEthernet SwitchでBridgeTableがあふれると古いエントリーが消されて新しいエントリーによって上書きされる）
  - 大抵の広域Ethernet網提供事業者は、アクセス単位や加入者ネットワーク単位で使用出来るMACアドレスの個数を制限している。（本当はあまり制限しない方が加入者の自由度は上がる）

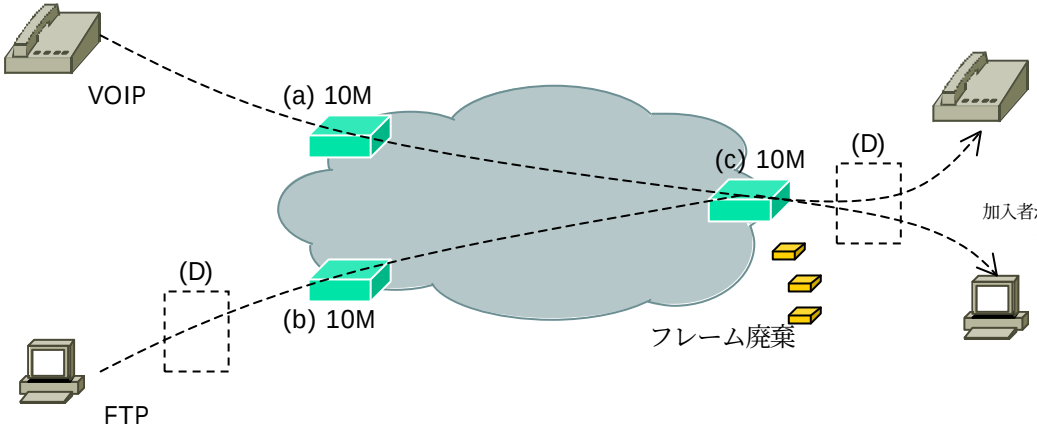
- **Loop発生時の影響**
  - Loopが発生する事による問題はループトラフィックの増殖**(*1)**によって通信帯域を浪費される事にだけあると思われがちだが、実際にはBridgeTableの内容が破壊されてしまう事によるインパクトも大きい。つまり、トラフィックが増殖しない程度の短時間でもLoopを発生させてはいけない。（この場合、不正な内容を持ったBridgeTableがAgeOutするまで通信が不可能になってしまう場合がある）

(*1)増殖したトラフィックはしばしばルータに対するDoS攻撃となる。

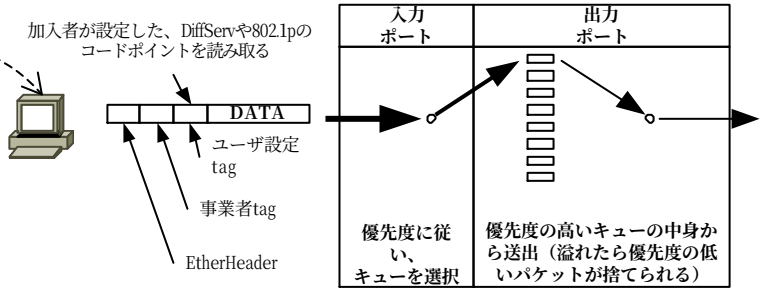
現在のVLAN型VPNの課題（2）

アクセスでのQoSの欠如

- 広域Ethernetのバックボーンは帯域をWDMとLinkAggregatinにより十分に確保（帯域確保型）し、フレームのロスを発生させないようにする事が出来るが、アクセス部分では複数の拠点から特定の拠点にトラフィックが集中した場合にフレームロスが発生する可能性がある。
- 下図のように、(a)及び(b)からそれぞれ10Mづつの入力があった場合に(c)の出力の性能が10Mしかなければ、(c)のエッジスイッチでフレームの廃棄が発生する。
- 加入者がVOIPのトラフィックを守りたい場合は、加入者宅の(D)の位置にパケットシェーパを設置しTCPの流量を制限する事により、UDP(RTP)の流量を守ると言う間接的な手段によって、VOIPのトラフィックを守らなくてはならない。（TCPはウインドウサイズの制御により速度コントロールが可能）



※本来は（C）のスイッチが(a)や(b)から流れてくるフレーム上で加入者が設定した802.1P(802.1D-1998)やDiffservの優先度を調べて選択的にフレーム廃棄を行う事により、特定のトラフィックを守る方法が考えられるが、今の広域Ethernetではサービスされていない。

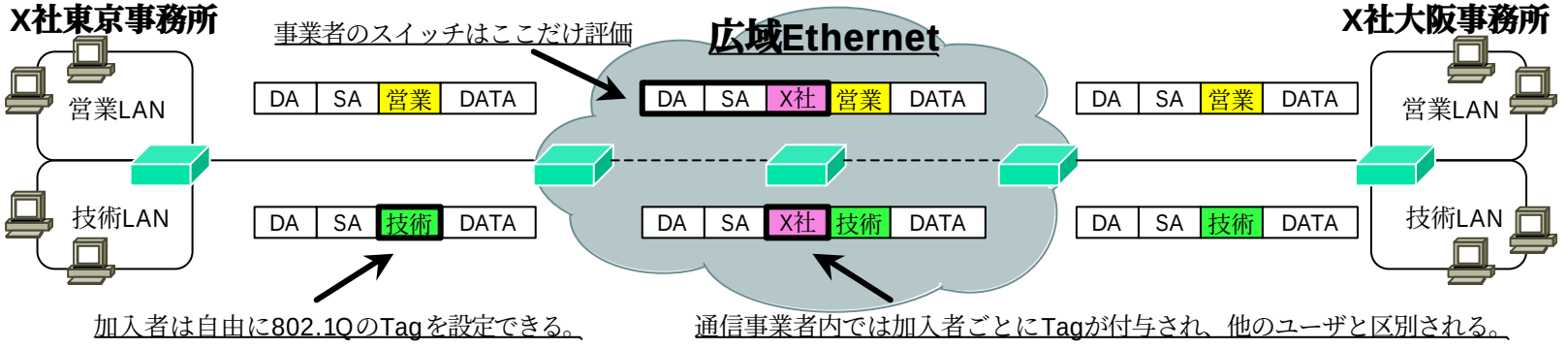


※足回りに専用線やATMを使っている場合にはそのメディア変換時、速度差によって、フレームは破棄が発生する（そういったメディア変換機にもQoS機能が求められる）。

現在のVLAN型VPNの課題（3）

加入者が設定VLAN間でのMACアドレスの重複不可

- 多くの広域Ethernetでは、多重VLAN技術**(*1)(*2)**を使い加入者が設定した802.1Qタグ付きフレームを透過的に流す事を可能としている。この機能によって加入者の設定する遠隔地の複数LANを論理的に分割したまま転送出来るが、広域Ethernetのスイッチはユーザの802.1Qタグを透過するだけで、認識しない為、**加入者の設定する802.1QVLAN間でのMACアドレスの重複は許されない**。（VRRPなど仮想MACを使う場合に注意）



(*1)事業者が使用する802.1QTagと加入者が使用する802.1QTagの２つのタグをスタックして使えるようにする技術は標準化されておらず、様々なベンダが独自の実装を行っている。
(vMAN(Extreme),QinQ(Cisco),Super AggregatedVLAN(Foundry)などがある、呼び名が異なっても相互接続出来る場合もある）

(*2)加入者設定VLANを透過させる為、広域Ethernet網内ではMTUが拡張されている。（加入者は1522bytesまで利用可能）

現在のVLAN型VPNの課題を解決する技術(1)

MP to MP EoMPLS(Multipoint to Multipoint Ethernet over MPLS)

EthernetフレームをMultipoint to MultipointのMPLSで転送する技術

- Tunnel LabelとVC Labelの使用
  - 理論上多量のVPNの設定が可能（LSPが何本張れるかは装置による）
- 送り元PEから送り先PEに対してPtoPのLSPでの通信+Split Horizon(標準)
  - 非常にLoopが発生しにくい。**(*1)**
- PEでのブリッジング（標準）
  - コアではMACを学習しない為、MACアドレスの学習量は少なくてすむ
- qualified ラーニング（オプション）
  - 加入者設定VLANごとにVPLSインスタンスを作る、つまり加入者設定VLANも広域Ethernet内で一つのVPNとして扱う事により、加入者設定VLAN間でMACアドレスの重複を可能とする。（スケールするかどうかは知らない）
- トラフィックエンジニアリング（オプション）
  - 特定のトラフィックの保護

(*1)EoMPLSでもHub and Spokeなどの階層化手法の導入の仕方によってはLoopの危険性が高まる。

現在のVLAN型VPNの課題を解決する技術(2)

VLAN型VPNソリューションの改良

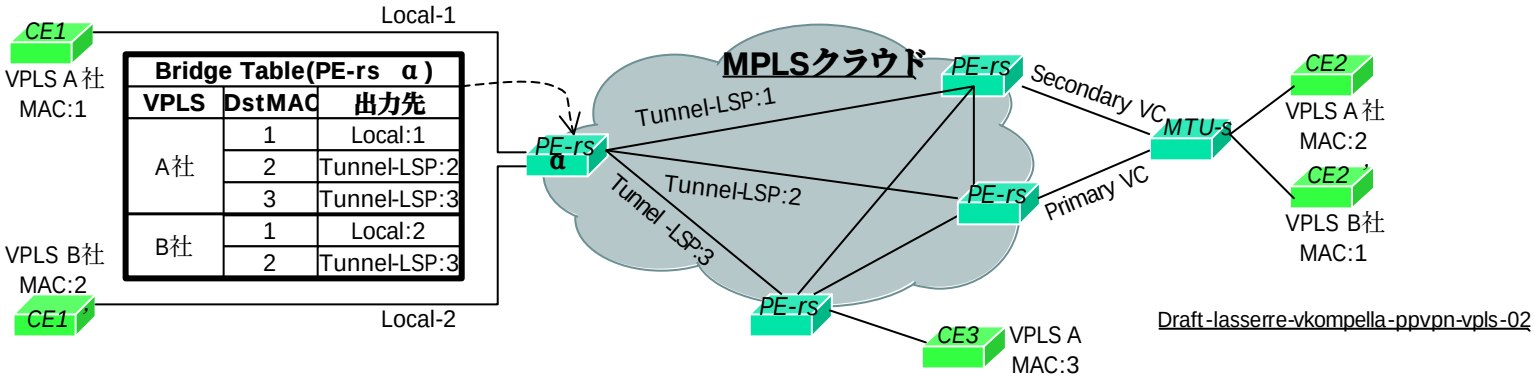
- Tag Swapping(VLAN cross connect）:(VLAN-Tag利用の効率化）
  - スイッチ間でTag番号を変換する事により、Tag番号の効率的利用を図る。(一本の伝送路にのせる事の出来るVLAN数は4096個のままなので、VLAN数を増やせるかどうかは網の設計次第。）
  - この技術は各ネットワーク間をTagのまま接続する用途に用いられる場合もある。

- Tagのbit幅拡張:(設定VLAN数の増加）
  - 802.1QのVLANタグの領域を拡張してしまう。あるいは、MPLSのラベル部分をVLANタグの代わりに使用する、…… あまり現実的ではなさそうだが）
- Tagのスタック:(設定VLAN数の増加）
  - 単純にスタックするだけでは、MACアドレスが加入者間で重複できなくなってしまうので、Tagラーニングと組み合わせるか、スタックしたTag全部をまとめて1つのTagとして扱えるスイッチが必要となる。
- 階層化ブリッジング(Ethernet over Ethernet):(学習MAC数の低減）
  - コアスイッチでのMACアドレス学習量の低減
- Tagラーニング：加入者設定VLANでのMACアドレス重複の許可）
  - MACアドレスだけでなく加入者が設定するTag部もラーニングする事により、加入者が設定する802.1Q VLANでのMACアドレス重複を可能とする。
- PEでの802.1pやDiffServのコードポイントに基づいたキューイング :(QoS)
  - VOIPなど特定のトラフィックの保護

MP to MP EoMPLS(Multipoint to Multipoint Ethernet Over MPLS)

EthernetフレームをMultipoint to MultipointのMPLS上で転送する技術

- IETFで標準化が進行中（ドラフトはいくつかある）
- PE間でフルメッシュLSPを張り、ブリッジングはPEで行い、MPLSのコアのではラベルスイッチングのみを行う。
- VPLSインスタンスごとに別々のBridgeTableを持つ。



- 特徴
  - Tunnel LabelをFull Mesh LSPに使い、VC LabelをVPN識別に用いる事で多量のVPNを作る事が出来る。
  - MACを学習するのがPEなのでMACアドレスの学習量の問題はほとんどない。
  - 既存のMPLSのインフラを利用出来る。
  - トラフィックエンジニアリングが可能。
  - MPLSベースの強力な冗長化機能が使える。（ファーストリルートなどを使い高速な経路切り替えが可能）
  - MPLS網側から受け取ったフレームをMPLS網に戻さない、**Split Horizon** の機能により、Loopが発生する要素が非常に少ない。

MPtoMP EoMPLSでのn-square問題

N-square問題

マルチポイントのEoMPLSを利用する場合、現在のIETF-Draftベースでの実装では階層化したとしても、PE-rs間でLSPをFull Meshではらなくてはならない(n-square問題)。これはMPtoMP EoMPLSの強みでもあるが、いくつかの弱点をともなっている。

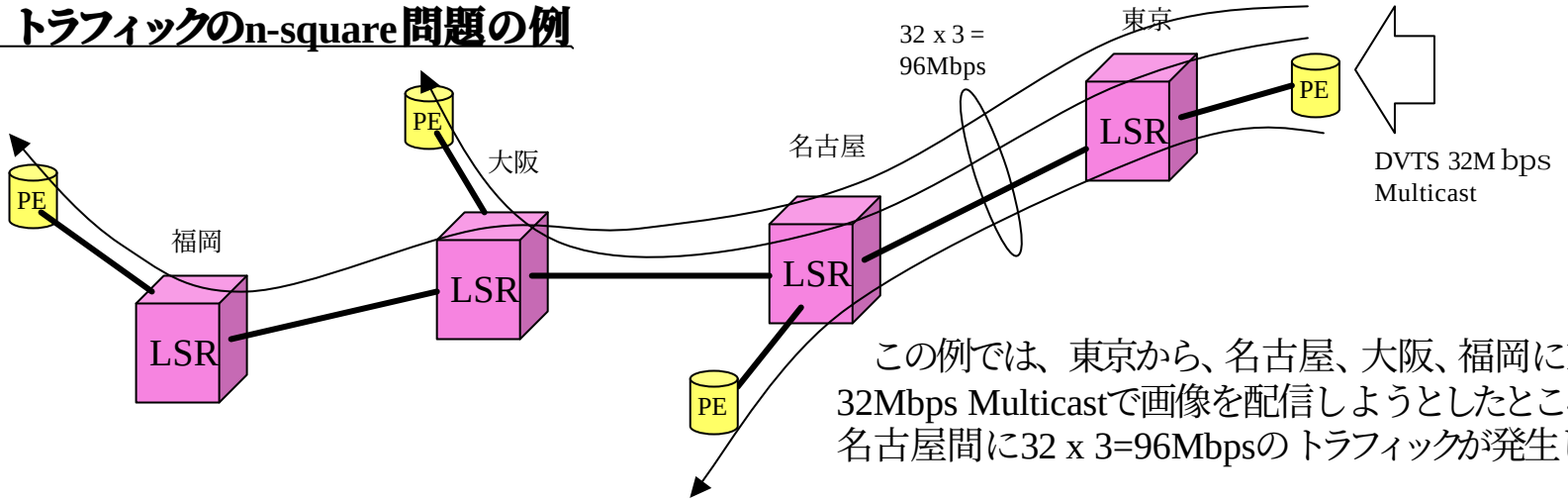
(1)LSPのn-square問題

PE-rsを追加する事により、n-1箇所での設定、シグナリングの追加が必要。 網全体ではn*(n-1)のLSPが必要）

(2)トラフィックのn-square問題

MulticastやBroadcas tUnknown Unicastをあて先にEthernetフレームを転送すると、網内では送り先のPE単位ごとのLSPにパケットがコピー転送される為、通信帯域を浪費してしまう。また、規模によりPEのメモリが余計に必要になったり、ジッタの増加を招く場合がある。

トラフィックのn-square問題の例



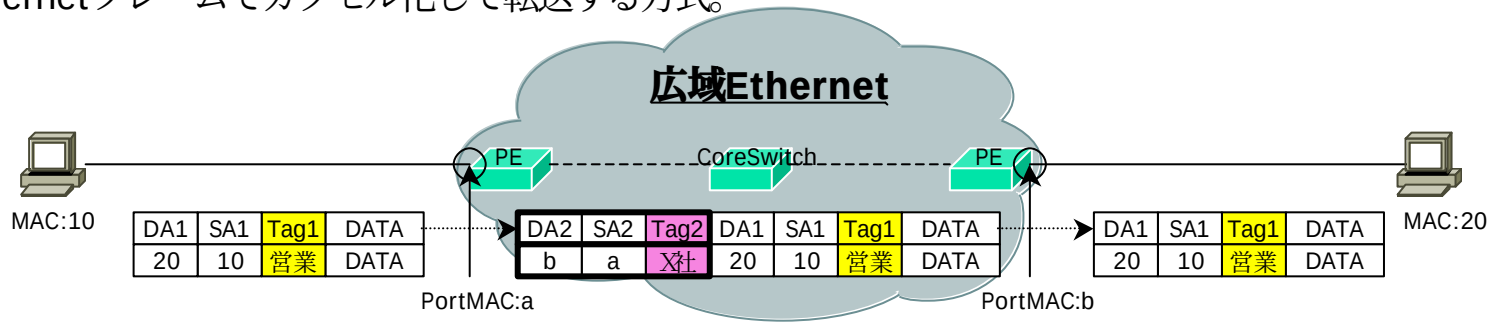
この例では、東京から、名古屋、大阪、福岡にDVTS 32Mbps Multicastで画像を配信しようとしたところ、東京名古屋間に32 x 3=96Mbpsのトラフィックが発生している。

※n-Square問題の軽減措置として、PEの機能を2つに分けて、パケットのコピー数やLSPを減らすDistributed Model(**Decoupled Model**)などが提案され、実装もあるが、今のところ根本的な解決策ではない。 (Multicast LSP や Hierarchical LSPがほしい！）

階層化ブリッジング(Multipoint Ethernet over Ethernet)

802.1Q Tag VLANを使ったVLAN VPNの改良方式

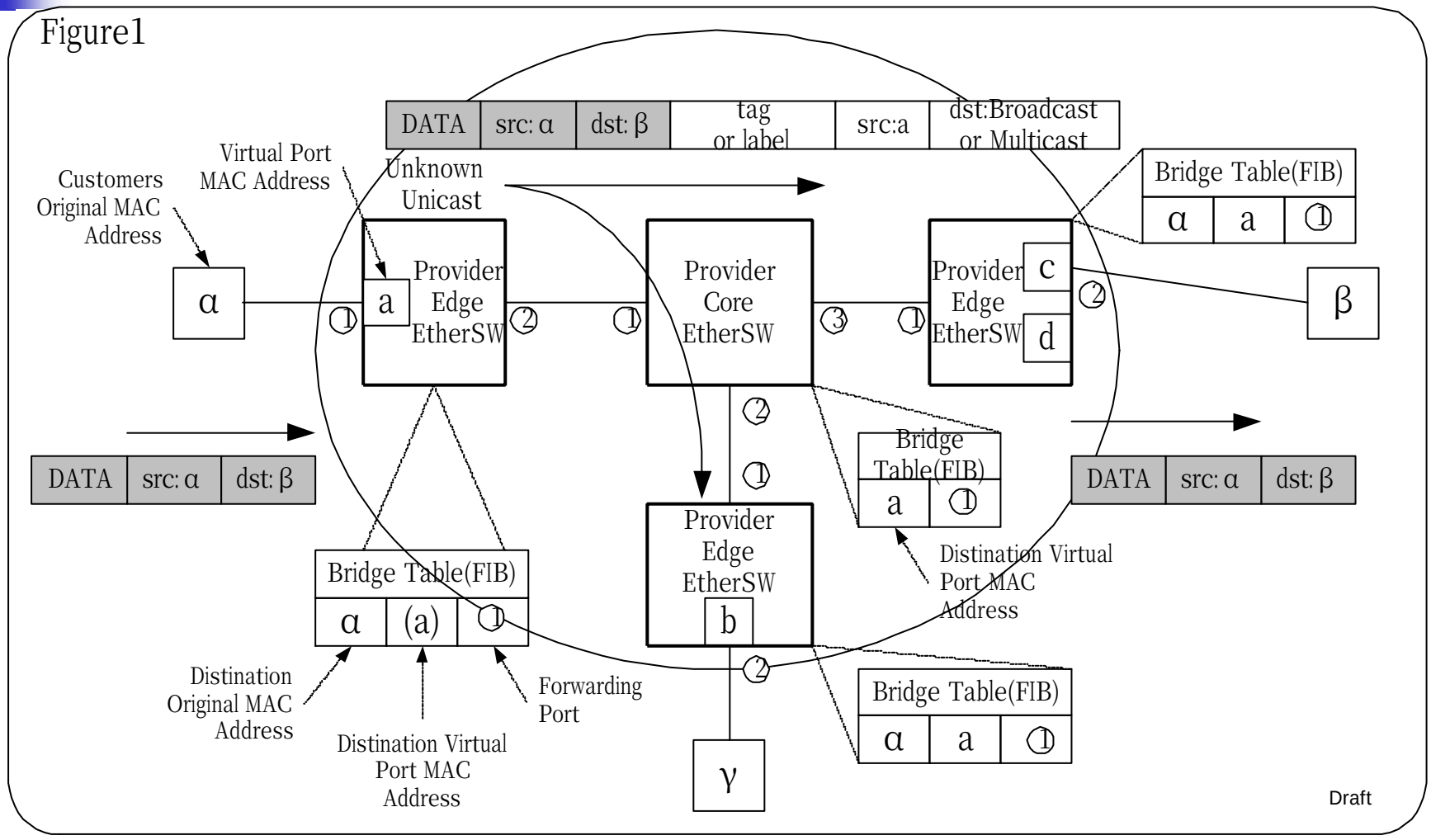
- PEの加入者向けポートそれぞれにユニークなMACアドレスを定義し、加入者から受け取ったEthernetフレームをその入力ポートに定義されたMACアドレスをソースとし送り先のPEのポートのMACアドレスをディスティネーションとするEthernetフレームでカプセル化して転送する方式。



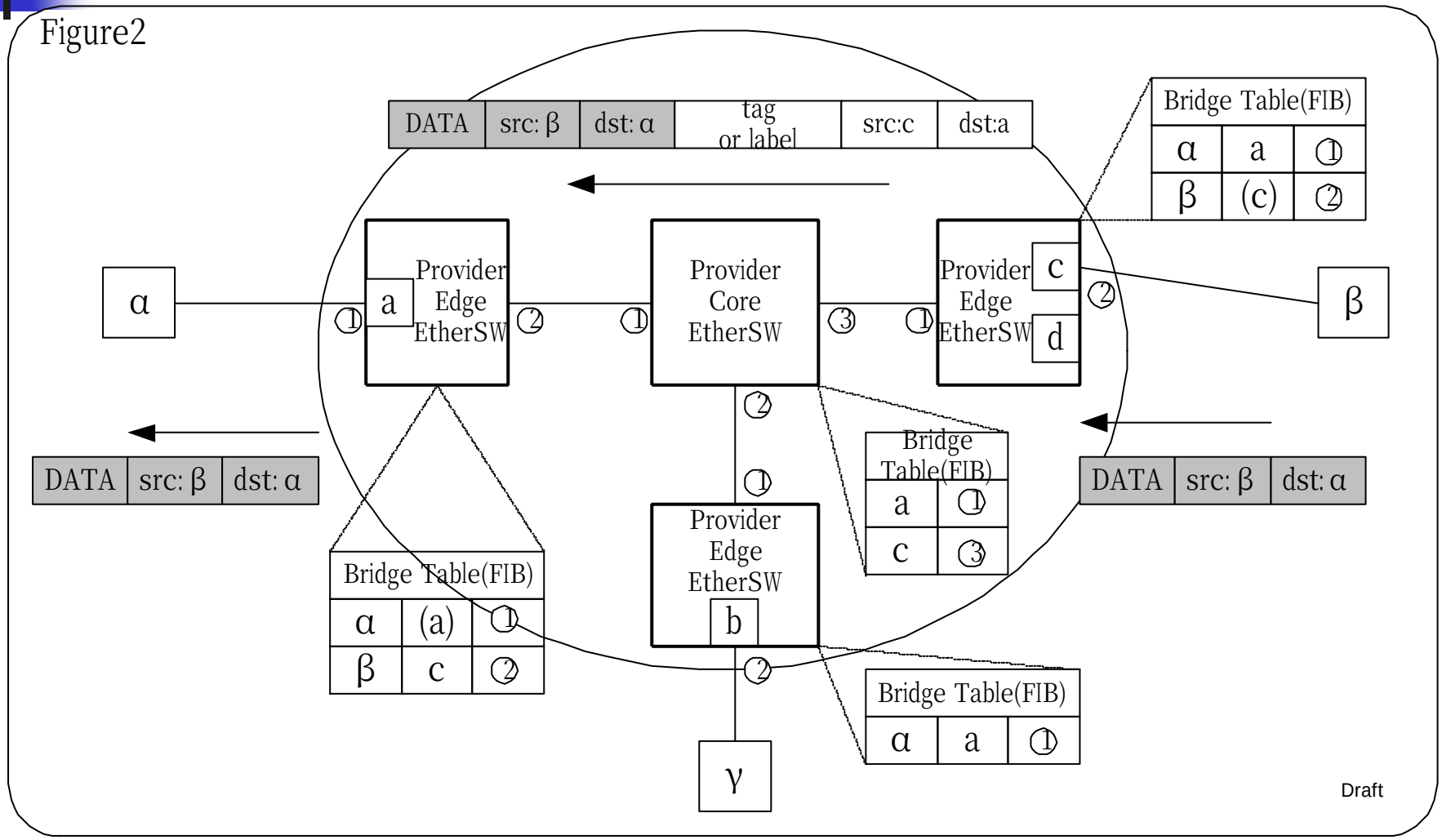
- 特徴
  - コアスイッチで学習しなくてはならないMACアドレスを劇的に減らす事が出来る。（最大でもPEが持つ加入者ポートの数だけのMACを学習すればすむ。）
  - 障害切り分け時にBridgeTableの内容を人間が見て確認する事が容易。
  - 特殊な処理を意味するあて先MACアドレスを持つパケットを安全に転送する。（STPで使用する制御用パケットであるBPDUやベンダが制御用に使っているフレーム、その他広域Ethernet上のスイッチや伝送装置がCPUに転送したり、ブロックさせたりする恐れのあるフレームを広域Ethernet上で安全に透過させる。）(*1)
  - コアスイッチは単にジャンボフレームを転送出来る普通のスイッチでかまわない。

(*)海外のIXで過去に発生した、ISP->IX向きのBPDUによるDoSにもこの技術を使って対応出来る。

参考:Multipoint Ethernet over Ethernet動作(1)



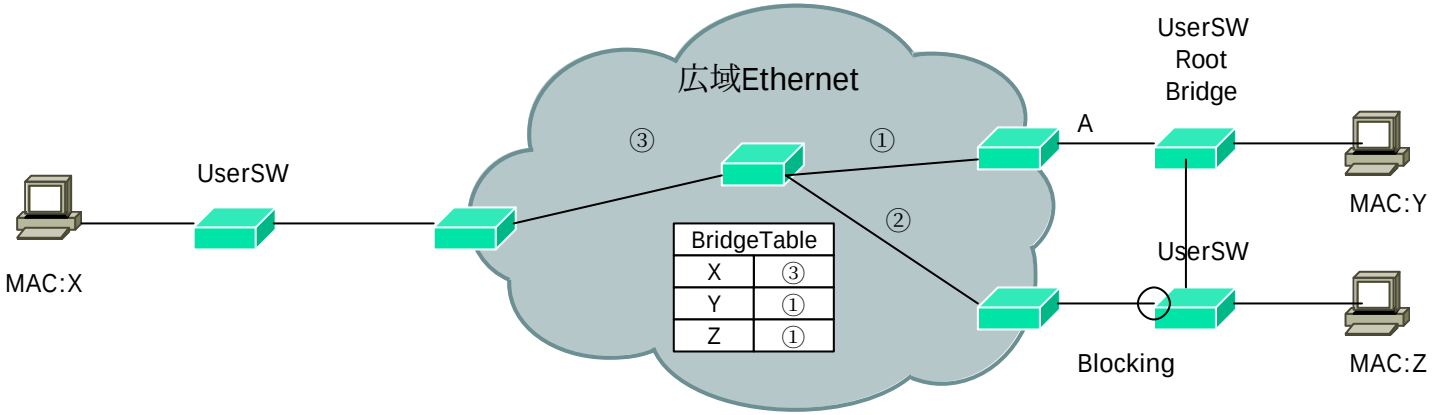
参考 :Multipoint Ethernet over Ethernet動作(2)



参考 :広域Ethernet上でSTPを使った場合の留意点

広域Ethernet上でSTPを使用する場合はBridgeTableの動きに注意する必要がある。

- 広域Ethernet上でBPDUは透過する**(*1)**が、広域Ethernet上のスイッチはBPDUを転送するだけでBPDUを解釈しない。その為、STPのBridgeTableの状態には注意する必要がある。
  - STPのトポロジー変更はBPDUで加入者スイッチ間を伝達されるが、広域Ethernet上のスイッチはBPDUを解釈しない為、トポロジー変更に伴うBridgeTableのフラッシュ（リセット）は行われない。
  - 加入者のSTPトポロジー変更時に広域Ethernet上のBridgeTableのフラッシュが行われないと、条件によって、BridgeTableのAgeOutまで通信が停止する。**(*2)**
- 組み方によっては、BridgeTableのAgeOutを待たなくても切り替えを出来る構成もありえるが、事業者としては本当はあまりSTP使ってほしくない。（加入者の設定ミスで発生したループを制御出来ない為）



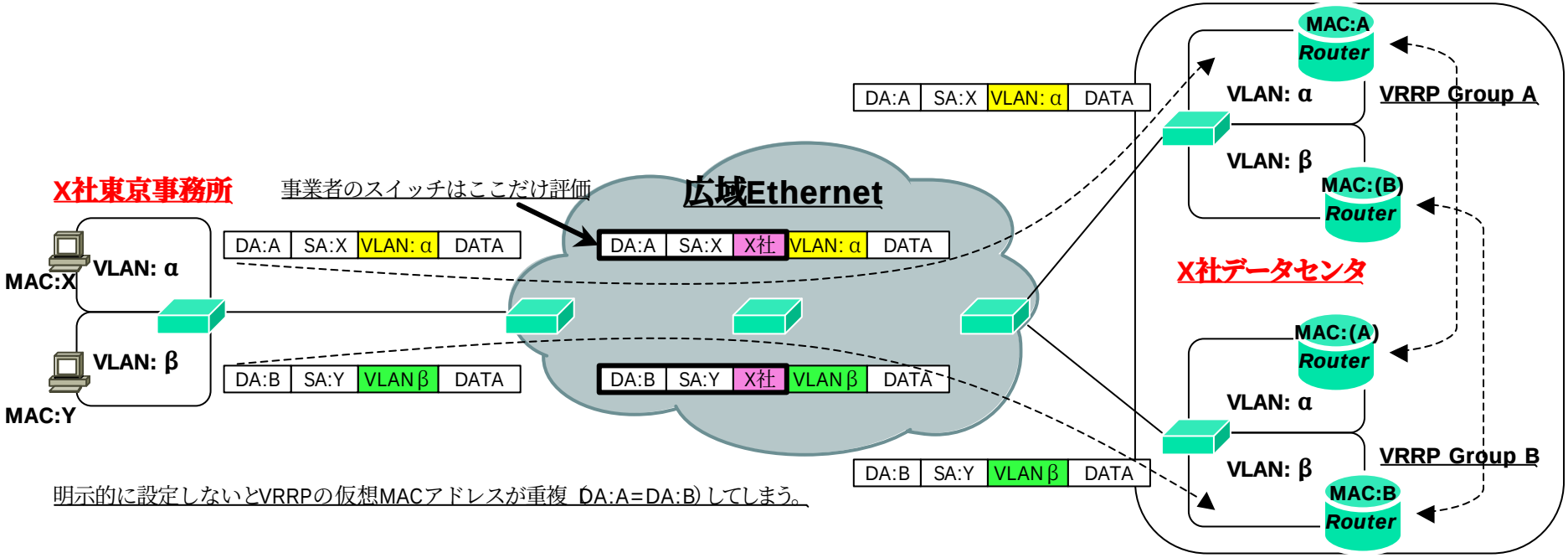
このような構成でAの部分が切断されると、UserSWのトポロジー変更が上手くいったとしても、広域Ethernet網内のBridgeTableが古い状態のまま保持されてしまう為、AgeOutするまで通信が出来なくなってしまう事がある。

(*1)POWEREDCOMの場合、足回りでATMを使っている場合にBPDUが通らない場合がある。
(*2)BridgeTableのAgeOut時間はPOWEREDCOMでは5分に設定している。（将来変わる可能性あり）。

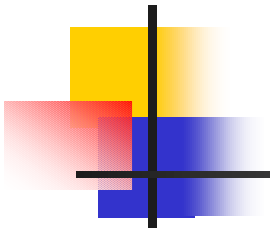
参考 :広域Ethernet上でVRRPを使った場合の留意点

加入者設定VLAN間でMACアドレスの重複が発生する例（VRRP）

- 加入者設定VLANを複数作り、その上でVRRPのようにシステムが仮想MACアドレスを自動的に設定してしまうようなプロトコルを使用する場合には注意が必要となる。



- 異なる加入者設定VLANで別々のVRRPを動作させると、異なる加入者設定VLANで同じ仮想MACアドレスが使われる。
- 異なる加入者設定VLANで重複するMACアドレスがあった場合、広域Ethenetr内では加入者設定VLANを認識しない為、上手く処理されない。（認識するのは、事業者設定VLANのみ）
- このような場合は重複が発生しないようにVRRPの仮想MACアドレスを明示的に設定するなどの、対策を講じなくてはならない。



Questions?

masaty@tec.poweredcom.net