広域イーサネットとキャリアサービス

Janog11

2003年1月23日

河野 美也 Miya Kohno (mkohno@cisco.com)

はじめに

- **祝！**

  Ethernet 生誕 30周年

  http://www.digitalcentury.com/encyclo/update/metcalfe.html

  http://www.ethermanage.com/ethernet/

  Internet 生誕 20周年

  http://www.columbia.edu/~rh120/other/tcpdigest_paper.txt

- **という訳で、**

  – この歴史的な年の年頭に当たって、最近ホットな「広域イーサネットサービス」の今後について考察してみたい。

  – 前回(Janog10)での、Ethernet関連議論(↓)が非常に興味深かったので、今回の「論」の場で、もう少し議論を続けてみたい。

  http://www.janog.gr.jp/meeting/janog10/janog10-programs.html

  …というのが、今回の発表の主旨です。

そもそもイーサネットって…(1)

Ether – エーテル、気体、空中波

その部屋（場）にいる誰もが、誰とでも自由に話しができる。その部屋に入るために特に許可は要らず、その部屋に入りさえすれば話ができる。

→ とてもリベラルで画期的。☺

- **CSMA/CD**

  「えー」とか言って話始める合図をする (プリアンブル)。

  他の人が話していたら自分はやめて、少し様子を見てから (Exponential Backoff)、また話し始める。

  → cf. sdlc polling-selecting（司会制）

  Token ring, FDDI（マイクを順に廻す）

  DPT/RPR（合議制）

そもそもイーサネットって…(2)

- **L2 Switchの登場**

  他人が話していようがいまいが、とりあえず話せるようになった。

- **Remote Bridge**

  部屋同士を中継。Video Conference !

- **vlanの登場**

  その部屋(場)の中で、班･グループを作る。

- **広域イーサネットサービス**

  巨大な一部屋

一方、ルータは…

- 玄関。（もともとGatewayと言われていた。）
  - 部屋と部屋の行き来を可能とする。しかし、関係ない人は通さない。
  - ある人と話したい時は、どの部屋を通って行けばよいかを、各玄関が明確に知っている。
  - 玄関や通路、部屋を識別するための住所は、階層性を持つ。
  - 迷子(ループ)抑止機能 (トポロジー計算) 、検出機能 (TTL)あり。
- LANを広域化するときは、“歴史的には” ルータが優位。
  - Bridging over Leased Line ： SRB->DLSw, Appletalk -> AURP, IPX->NSLP
  - Bridging over ATM： 一部の展開に留まる。
  - LAN Emulation over ATM (LANE)： 流行ったこともあったが、廃れた。

広域イーサネットサービスに、ユーザは何を求めるか？

- IP以外(NetbeuiやIPX, Appletalk)をそのまま通したい?
- 「これからはルータレスネットワークだ!」…?
- 複雑な設定をせずに、サイトの追加変更等を行ないたい？
- プロヴァイダと、ルーティングテーブルの交換をしたくない?
- OSPFのデザインを変えたくない?
- Multipoint Connectivity ?
- マルチキャストしたい?
- 音声を通したいので、QoS制御も欲しい?
- 専用線代わりに使いたいから、帯域保証も欲しい?
- 基幹業務を載せるので、ダウンは許されない?
- 早くて安ければ何でも良い???!!

広域イーサネットサービスへの、プロヴァイダの思いは?

- そろそろ運用管理がしんどくなってきた。
  - Loopは何とか防がないと。STPは限界あるなぁ。(RIPより辛そう(from L3屋)。)
  - 障害解析手法が非力。
- 設備コストはとにかく安くないと。
- しかし、SLA(*)的なものも提供しないと、価格競争に陥るだけかも。
  - (*) High Availability, Performance(Delay/Jitter), QoS, 帯域保証に関するService Level Agreement
- Traffic Engineeringしたいし、マルチキャストにも、効率よく対応したい。
- ユーザ要求は、殆ど「無いものねだり」になりつつあるけれども、結局は「安ければよい」、というところもある。
- まぁ、とにかく売れればいいか。(from 営業)
- でも、このままだと破綻するかも。(from 運用者)

根源的な問題 (1)

- UNIという概念が無い
  - Control Signalの扱いをどうするか?

    廃棄? 透過? 連携?
    - Control Signalの種類や、提供するサービス形態よって異なる。
    - BPDUは、廃棄または透過。連携はたぶんあり得ない。
    - Link Aggregationは、連携もありか。
    - .1xのようなauthentication mechanismは、UNIでの認証に使えるかも。(=連携)
  - vlanの扱いは?

    Translate ? Stack ?
  - MACアドレスの扱いは?

    そのまま? Translate ? Stack ?

根源的な問題 (2)

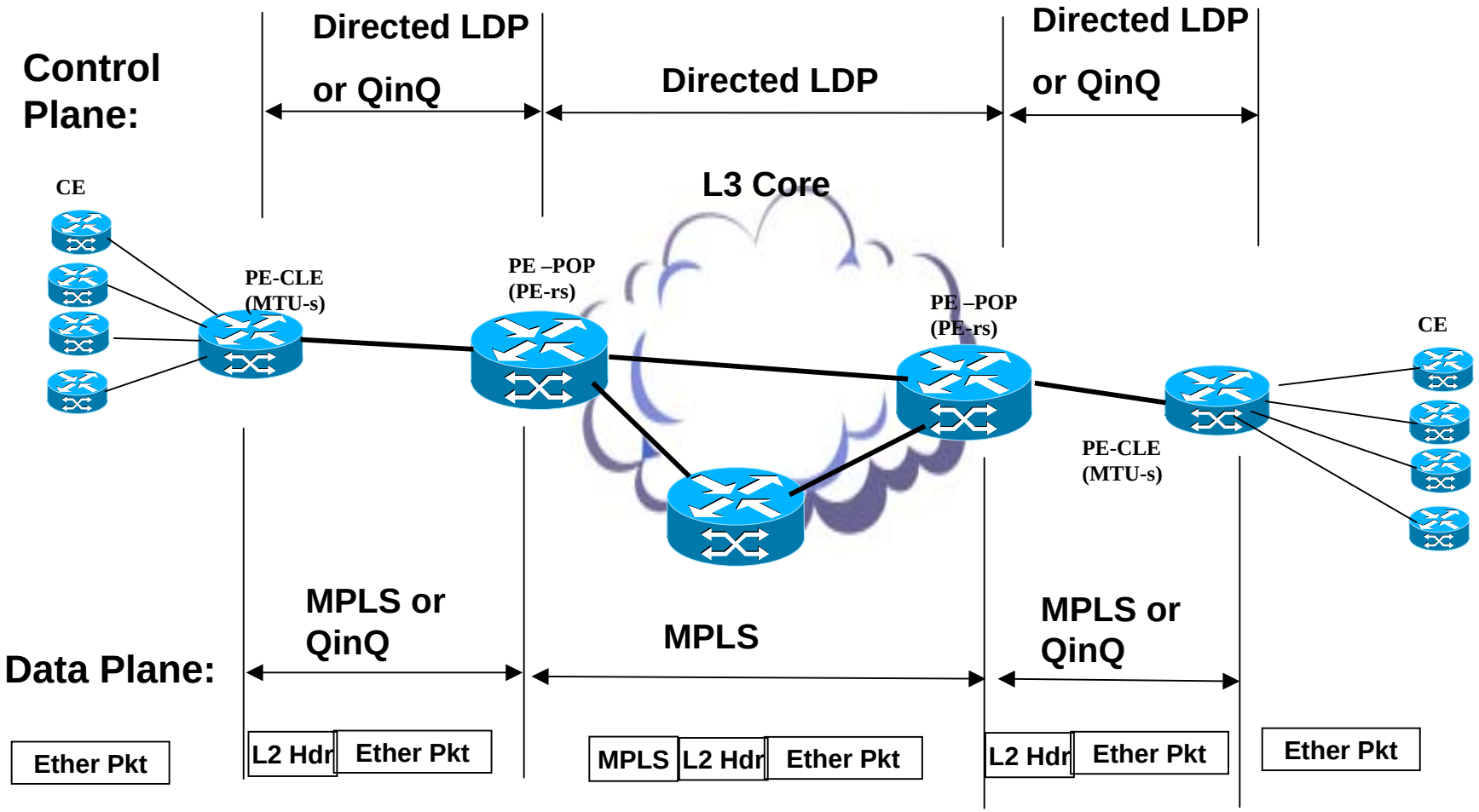
- どこまでスケールする?
  - そもそも、「巨大な一部屋」ってアリなのか。
  - VLAN id(4k)の問題は?
    - → とりあえずtranslateすれば凌げる? でも手作業で?
  - MAC table entryは?
    - → 実際の状況としてはどんなものでしょう?
- 冗長性確保とループ回避の問題は、一筋縄ではいかない。
- そもそも現実的に、既存IEEE802標準に対して、どこまでの変更が許されるものだろうか。
- 障害解析ツールは欲しい(ping,tracerouteやLMI)。

いくつかの提案

- H-VPLS (IETF PPVPN)
- EoE (Janog10 安藤さん!)

- IPLS (IETF PPVPN)
- Pseudo Wire (IETF PWE3)

H-VPLS



draft-lasserre-vkompella-ppvpn-vpls

H-VPLS – 特徴

- L3 Coreによる、stabilityの向上、Multiservice性
- PE-PE間にfull meshのPseudo Wire (MPLS or L2TPv3) VC

  Full mesh決めうちにすることにより、ATM-LANEのような複雑な概念(LES, LECS, BUS)を排除。

  Split Horizonによるループ抑止
- Emulated vlan over IP/MPLS
- コアにおけるMAC learning数の削減
- 階層化(distributed-PE)による、Full Mesh VC数の削減
- VCの両端で、vlan IDは異なっても良い(vlan id rewrite)
- 「“L2 island”同士を“L3 core”で相互接続する」という、IEEE 802.1の動きと協調

H-VPLS – 問題点

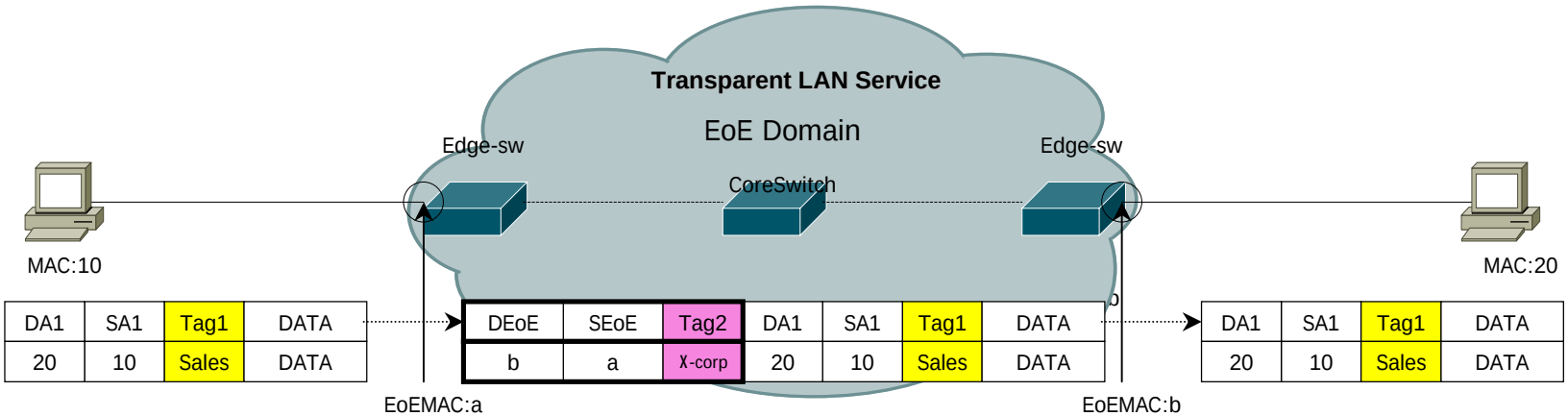
- L3 Core構築のコスト
- Multicast/Broadcast/Unknown unicastに対する非効率さ (*)

  (*) PWにmGREを使うという提案もある (-> draft-sajassi-mvpls)
- 階層化のための、distributed PEモデルの際、PE-rs/MTU-s間回線の冗長性確保をどう考えるか。

  MPLS(LDP)を使えば解決し易いが、MTU-sがMPLS対応する必要がある。

  .1qであれば小型LAN switchで対応可能であるが、その場合、冗長性の問題は別途解決しなければならない。
- Full Bridge Emulationであるので、Bridgingそのもの(巨大な一部屋)に起因する問題は、解決されない。

EoE (Ethernet Over Ethernet)



出典: http://www.janog.gr.jp/meeting/janog10/pdf/janog10-l2-ando.pdf

EoE – 特徴

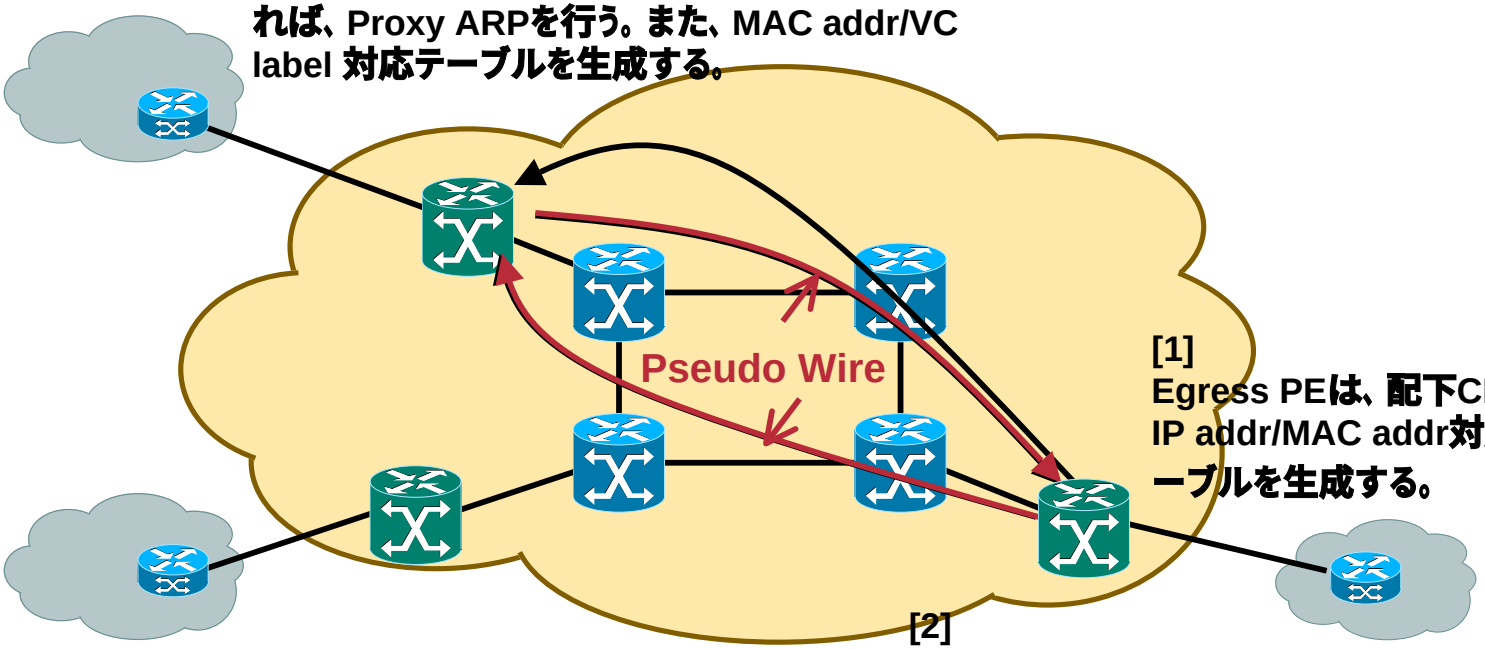
- コアスイッチで学習すべきMACエントリー数を削減する。
- 物理的MACアドレスと論理的MACアドレスの分離
  - 特殊な処理を意味する宛先MACアドレスを持つフレームを安全に転送する。
  - 階層化やbitmaskによる操作が可能
- TTLやVPN-id fieldも定義。
- コアスイッチは、EoE対応である必要は無い。基本的には、普通のスイッチでよい。
- IETF PPVPNにdraftとして提出された、VHLS(draft-sodder-ppvpn-vhls)と、共通点あり。(MAC-in-MAC, VPN-id)

EoE – 問題点

- EoE実現のための開発コストと、得られる結果のバランス
  - 本当にMACアドレス数が問題なのか?
    - 実際のところどうなのでしょうか。
    - 少なくとも、CEの7割~8割はルータと違う?
  - EoEで解決できることは、他の方式でも解決できる。
    - コアでのMACアドレスエントリーの削減
      - ← VPLS/IETF
    - 物理MACアドレスと論理MACアドレスの分離
      - ← MAC address Translation/IEEE
  - または、EoEでも解決できない
    - 冗長とループ回避
  - 逆説的な悩ましさ
    - 比較的High end Specが許され易いコアでのMAC数を削減するが、経済性(ポート当たり単価)が激しく追及されるエッジでは、MAC数を削減できるどころか、却って負荷が高まる。
- VPN-id数の制限
  - 網で4k vpn。網を分ければ良いが、そのための運用負荷が発生する。
- Forwarding Planeでの、VPN-id解析。

IPLS – control plane

[3]
Ingress PEは、配下CEからARP要求を受けた際、テーブルにあればその値を返答し、テーブルに無ければ、Proxy ARPを行う。また、MAC addr/VC label 対応テーブルを生成する。

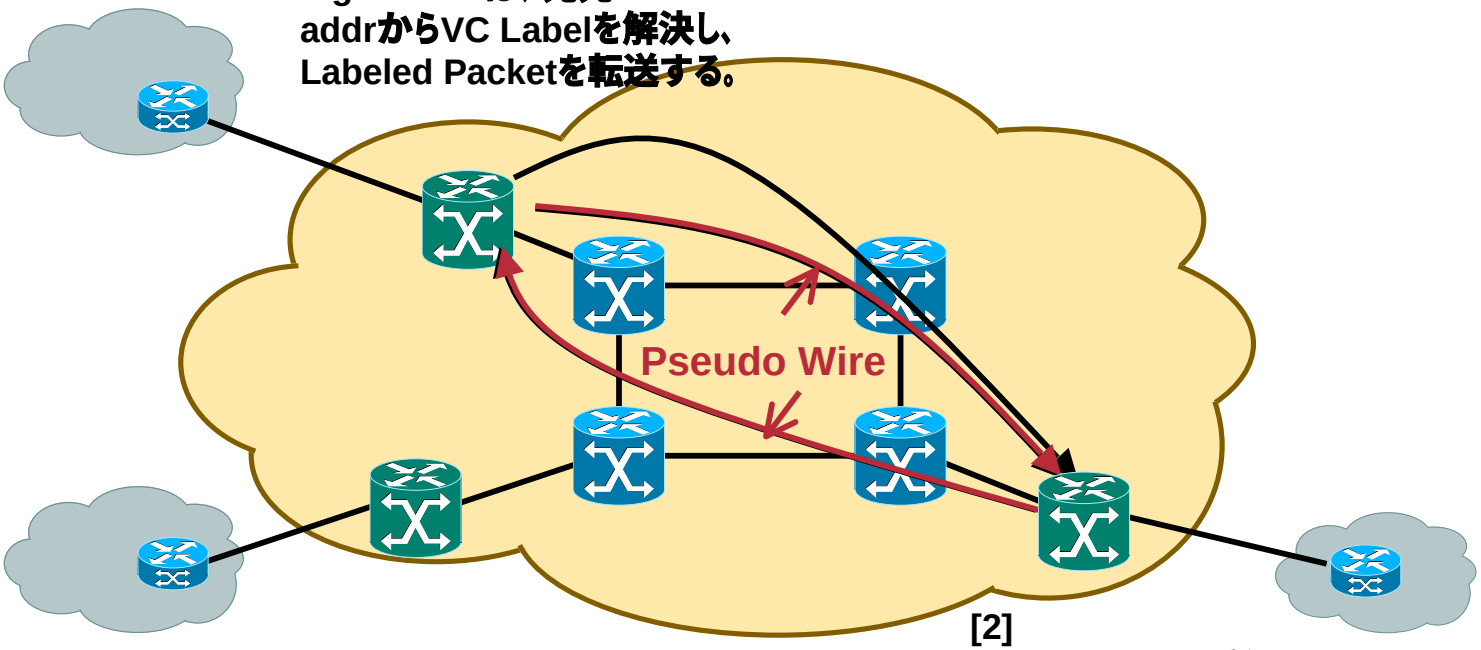


[1]
Egress PEは、配下CEのIP addr/MAC addr対応テーブルを生成する。

[2]
Egress PEは、各CEに対応するVC LabelをBinding。Ingress PEに対し、そのVCを介して、IP addr/MAC addr対応テーブルをadvertiseする。また、VC Label/MAC addr対応テーブルを生成する。

IPLS – forwarding plane

[1]
Ingress PEは、宛先MAC addrからVC Labelを解決し、Labeled Packetを転送する。



[2]
Egress PEは、受信LabelによりMAC addrを解決し、Ethernet Packetを転送する。

IPLS – 特徴

- 本当にFull Bridge Emulationが必要なのか、という根源的問い。

  少なくとも7~8割のCEはL3 deviceでは?

  顧客が本当に必要としているものを抽出して…

  - 複雑な設定をせずに、サイトの追加変更等を行ないたい
  - プロヴァイダと、ルーティングテーブルの交換をしたくない
  - OSPFのデザインを変えたくない
  - Multipoint Connectivity

  もしも上記が主要なポイントなのであれば、Scalability, 運用性, 安定性の面で、もっと良いアーキテクチャも考えられる。

IPLS – 特徴

- DataplaneでのMAC学習の必要なし。
- 必要MAC学習数は少ない。

  もっともこれは、IPLS自体の特徴ではない。Full Bridge Emulationであっても、接続CEがルータであれば、実質的に同様。但し、“予測可能”であるところは大きい。

- Egress PEでの処理がシンプル。

  recursive lookup(Label resolution -> MAC resolution)や、MAC学習のためのsource MAC lookupの必要が無い。

- Distributed PEモデル(scalability向上のための階層化)の必要が無い。

  これは、H-VPLSにおけるPE-rs/MTU-s間回線冗長化問題を排除する。

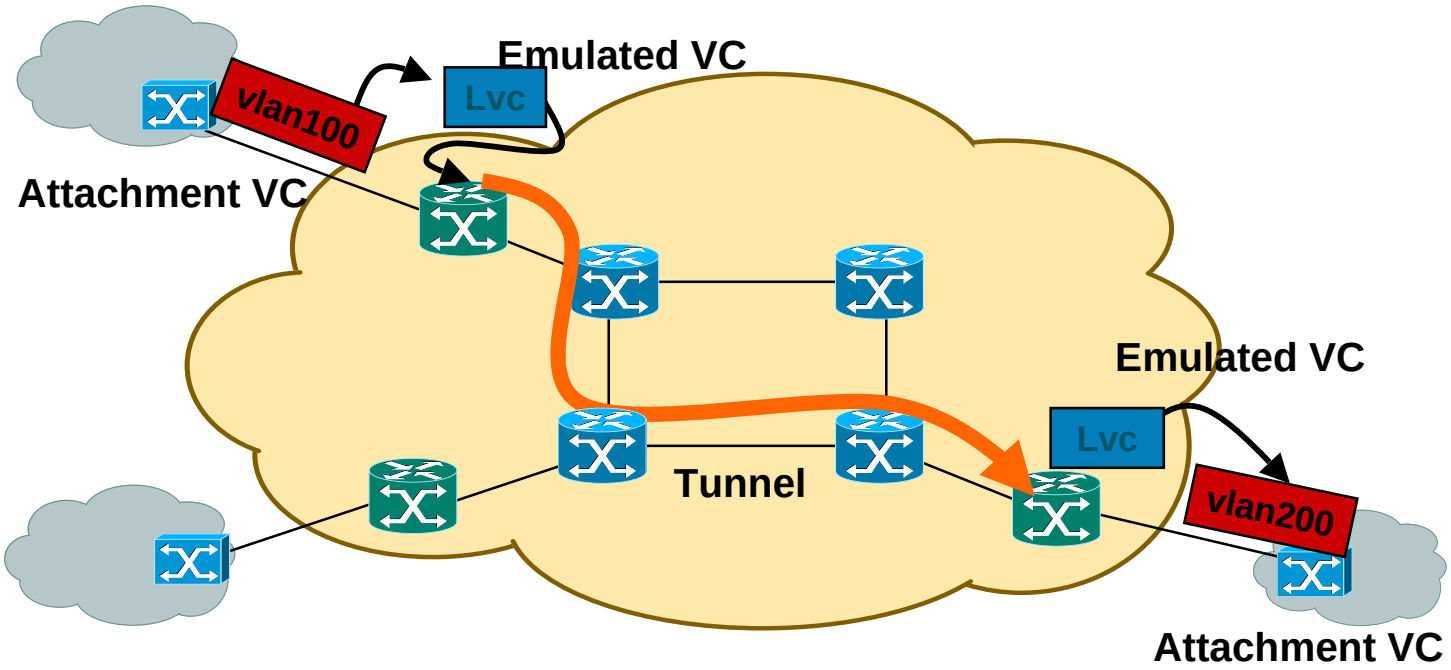
IPLS – 問題点

- Just idea レヴェル。
- 実際、「CEはL3」という仮定が「真」とすると、VPLSでも、ほぼ同程度スケールすることになる。
- IPLSが解決する、下記の命題のうち、
  - 複雑な設定をせずに、サイトの追加変更等を行ないたい -- 1)
  - プロヴァイダと、ルーティングテーブルの交換をしたくない -- 2)
  - OSPFのデザインを変えたくない -- 3)
  - Multipoint Connectivity -- 4)

  1)と4)は、L3 VPNで解決できてしまう。

- どなたか欲しい方いらっしゃいませんか ?! ☺

Pseudo Wire



- Ethernet vlan -> VC mapping
- Ethernet port -> VC mapping

Pseudo Wire – 特徴

- **QoS対応可能**
  - .1p -> mpls exp copying/mapping
- **帯域保証、Traffic Engineeringし易い**
  - 各VCの、TE LSPへのマッピング可能
- **高速迂回**
  - IGP Fast Convergence
  - MPLS TE FRR
- **管理機能充実**
  - 統計・課金データ取得可
  - End-to-End Link Management
    - 例えば、片端のS-LOS検知したら、もう片端のTxを落とす、等。

Pseudo Wire – 問題点

- **Point to Point only !!!**

So, what solves what ?!

Full Bridge Emulation(*)

(*)
- L2 only network
- VPLS

IP以外(NetbeuiやIPX, Appletalk)にも対応

「これからはルータレスネットワークだ!」??

複雑な設定をせずに、サイトの追加変更等を行なう

プロヴァイダとルーティングテーブルの交換を行なわない

OSPFのデザインを替えない

Multipoint Connectivity

音声を通したいので、QoS制御が欲しい

専用線代わりに使いたいから、帯域保証が欲しい

基幹業務を載せるので、ダウンは許されない

So, what solves what ?!

IPLS

- IP以外(NetbeuiやIPX, Appletalk)にも対応
- 「これからはルータレスネットワークだ!」??
- 複雑な設定をせずに、サイトの追加変更等を行なう
- プロヴァイダとルーティングテーブルの交換を行なわない
- OSPFのデザインを替えない
- Multipoint Connectivity
- 音声を通したいので、QoS制御が欲しい
- 専用線代わりに使いたいから、帯域保証が欲しい
- 基幹業務を載せるので、ダウンは許されない

So, what solves what ?!

Pseudo WIre

- IP以外(NetbeuiやIPX, Appletalk)にも対応
- 「これからはルータレスネットワークだ!」??
- 複雑な設定をせずに、サイトの追加変更等を行なう
- プロヴァイダとルーティングテーブルの交換を行なわない
- OSPFのデザインを替えない
- Multipoint Connectivity
- 音声を通したいので、QoS制御が欲しい
- 専用線代わりに使いたいから、帯域保証が欲しい
- 基幹業務を載せるので、ダウンは許されない

So, what solves what ?!

IP VPN
(.1q access)

- IP以外(NetbeuiやIPX, Appletalk)にも対応
- 「これからはルータレスネットワークだ!」??
- 複雑な設定をせずに、サイトの追加変更等を行なう
- プロヴァイダとルーティングテーブルの交換を行なわない
- OSPFのデザインを替えない
- Multipoint Connectivity
- 音声を通したいので、QoS制御が欲しい
- 専用線代わりに使いたいから、帯域保証が欲しい
- 基幹業務を載せるので、ダウンは許されない

「論」- 広域イーサネットサービスはどこに向かうのか (1)

迷った時は、基本精神に立ち返って考える。

- **Carrierの精神 ?!**
  - **網/ユーザ分解点の明確化**
  - **Traffic Engineering/Optimization**
  - **付加価値サービスによる差別化**
- **The Internetの精神 ?!**
  - **The End-to-End principal (by David D. Clark, et.al)**

    http://citeseer.nj.nec.com/saltzer84endtoend.html
    - **CEの機種や実際の使われ方をもっと探り、それに合ったアーキテクチャを考える。**
    - **IP以外のトラフィックってどのくらい?**
    - **CEはL3 device ? L2 device ?**
    - **実際のアプリケーションは何? 何が一番重要な要求?**
  - **Do not reinvent.**

    The role of the architect is to study all the existing pieces and to make sure that the adjunctions will fit. (by Christian Huitema)
    - **トポロジー計算、ループ抑止はL3に任せる。**
    - **Access AggregationはL2、CoreはL3で…。**

「論」 - 広域イーサネットサービスはどこに向かうのか (2)

- **注意**

  「Carrierの精神」と「The Internetの精神」は、全くヴェクトルが違います。

- **広域イーサネットサービスの基本精神は?**

  Carrierに近い?

  The Internetに近い?

  それとも、全く別なもの?

- **ご意見を!!!**