シスコサポートコミュニティ (CSC)
Live Expert Webcast
第3版 Cisco LANスイッチ教科書「番外編」

山下 薫 (kaoru@cisco.com)
シスコシステムズ合同会社

2014年9月9日 (Rev. 1)

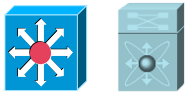
はじめに

- 第3版 Cisco LANスイッチ教科書
  - 3月20日の出版から半年弱が経ちました。
  - 表紙にはCatalystの写真しかありませんが、実はNexus含有率30%くらいの本です。
  - なので書名から "Catalyst" が無くなりました。
- 出版後のアップデートや、時間と体力が足りず書き切れなかった点がいくつかあります。
- 今回のセッション：「番外編」
  - 「質問箱」にいただいたご質問と、プレゼンタが用意した話題の計4つについて解説します。

今日のAgenda

はじめに

難易度
(Nexus含有率)

1. ACLがTCAMに書き込めないとどうなるのか (20分)  1.5

2. STP安定化技術を実際に「発動」させてみる (10分)  2

3. FabricPathの高度なループ防止メカニズムと vPC+ (20分)  3

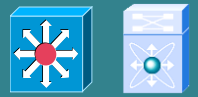
4. HSRP Preempt使用時に、再起動の際の
パケットロスを防ぐには (10分)  2

Q&A (20分) + 受講者プレゼントについてのお知らせ + おわりに

進め方

- 4つのご質問 (話題/テーマ) について、まずは直接ご回答します。
- その後、そのご質問に関して、「LANスイッチ教科書」に書き切れなかった関連情報や、出版後に判明したことをご紹介します。
- 追加のご質問は、このセッションの最後にまとめてお受けします。

[その1]
ACLをTCAMに書き込めない(TCAMの容量が足りない)場合、スイッチはどのように動作するのか



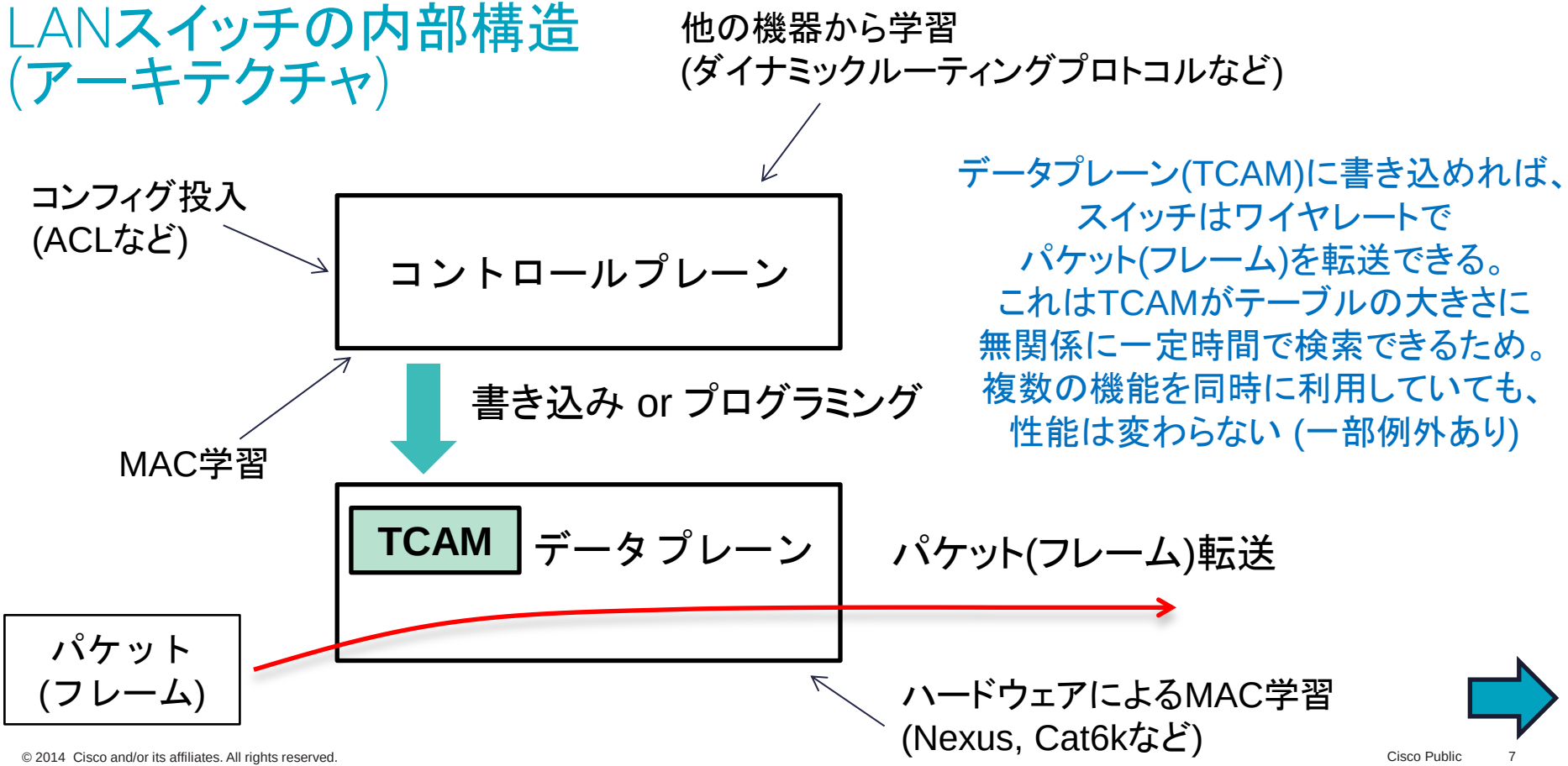
Catalystメイン (一部Nexusにも適用)
難易度 1.5 (やや低)

最初にご回答

- ACL(RACL)をTCAMに書き込めない場合の挙動を、一部の機種で確認しました
  1. 代わりに、該当ACLとパケット転送をソフトウェアで処理する (Catalyst 3750-X)
  2. ACLの設定は入るが、実際にはTCAMに書き込まず、ソフトウェア処理もしない
     → ACLの設定を無視する。OSPF, ICMPなどもともとソフトウェア処理されるパケットには効く
     (Catalyst 3650, IOS-XE 3.6.0E)　※ 3850も同じとのBUの回答あり

- 今後リリースされるIOSやIOS-XEでは、挙動が変わる可能性があります。
- Catalystの他の機種や、Nexusでは挙動が異なる場合があります。
- RACL以外の機能を併用する場合、TCAMを共用している機種では
  制限事項や挙動の変化があり得ます。(参考: P.180 コラム)

P.133 - 135 LANスイッチの内部構造 (アーキテクチャ)



データプレーン(TCAM)に書き込めれば、
スイッチはワイヤレートで
パケット(フレーム)を転送できる。
これはTCAMがテーブルの大きさに
無関係に一定時間で検索できるため。
複数の機能を同時に利用していても、
性能は変わらない (一部例外あり)

Catalyst 3750-X と 3650の挙動の違い (RACLの場合。抜粋)

- 3750-X, 15.2(2)E

```
C3750X-04(config)#int Te1/1/1
C3750X-04(config-if)#ip access-group simple-acl-2k in
%ACLMGR-4-UNLOADING: Unloading ACL input label 1 Te1/1/1,  IPv4/Mac feature
%ACLMGR-4-ACLTCAMFULL: ACL TCAM Full. Software Forwarding packets on Input label 1 on L3 L2
```

RACLは効きますが、ソフトウェア処理になります

- 3650, 3.6.0E

```
C3650-01(config)#int Te1/1/3
C3650-01(config-if)#do show platform tcam utilization asic 1 | inc Security Access
 Security Access Control Entries                    1536              184
C3650-01(config-if)#ip access-group simple-acl-1850 in
%ACL_ERRMSG-4-UNLOADED: 1 fed:  Input IPv4 L3 ACL on interface Te1/1/3 for label 4 on
asic255 could not be programmed in hardware and traffic will be dropped.
C3650-01(config-if)#do show platform tcam utilization asic 1 | inc Security Access
 Security Access Control Entries                    1536              184
```

RACLのコンフィグは入りますが、効きません。

黄色の部分は、「制御パケットに対して」という意味です。(確認中)

P.134　図3.4　--- ACLがTCAMに書き込めている場合

テーマ その1

制御パケット/フレームに対するACLはここで適用

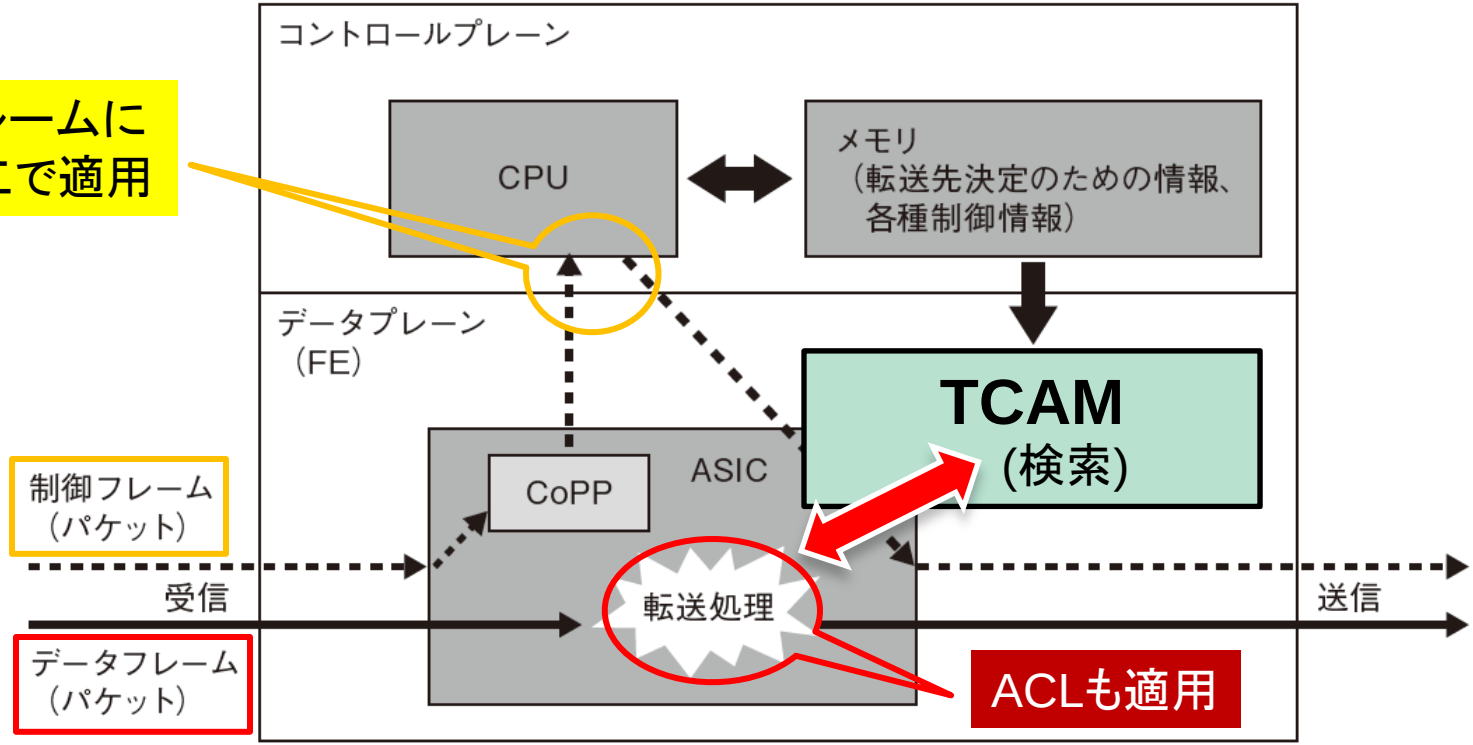
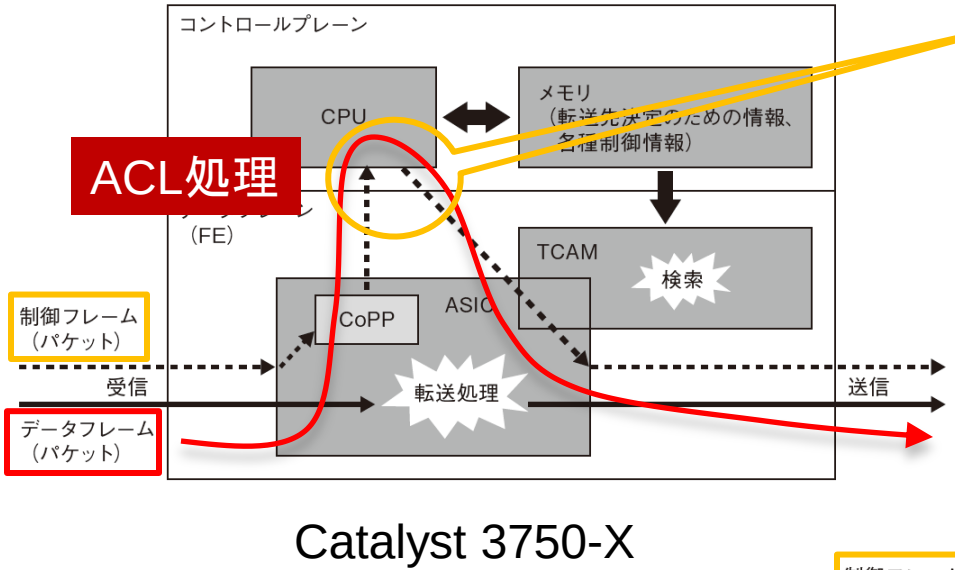


図3.4　コントロールプレーンとデータプレーンの役割分担と実装

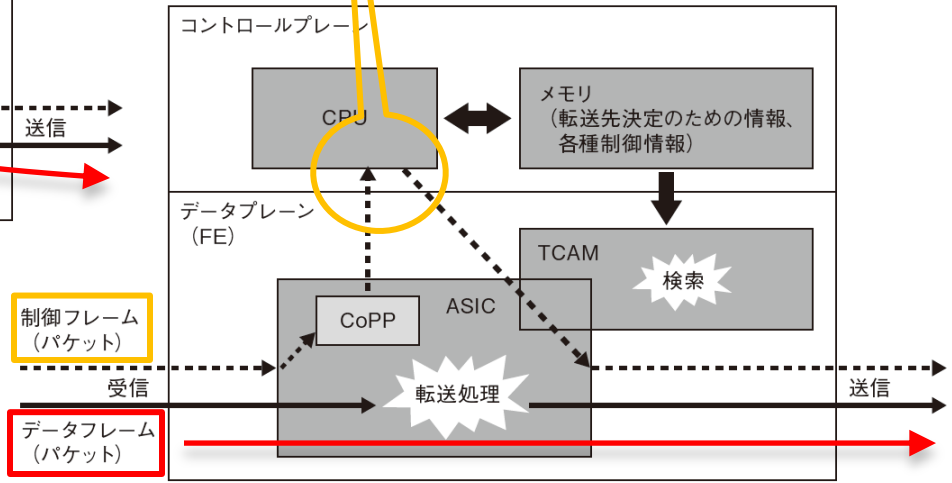
ACLがTCAMに書き込めない場合の 3750-X と 3650 の挙動の違いの詳細

制御パケットにはいずれもACLが効く



Catalyst 3750-X

Catalyst 3650
(3850も同様)



ACL処理なし

TCAMにACLを書き込まずに事前確認する機能があります

- Catalyst 6500/6800 Sup2T --- Dry Run (IPv4 RACLのみ)
- Nexus 7000/7700 --- Session Manager

Cat6k ACL Dry Runの例

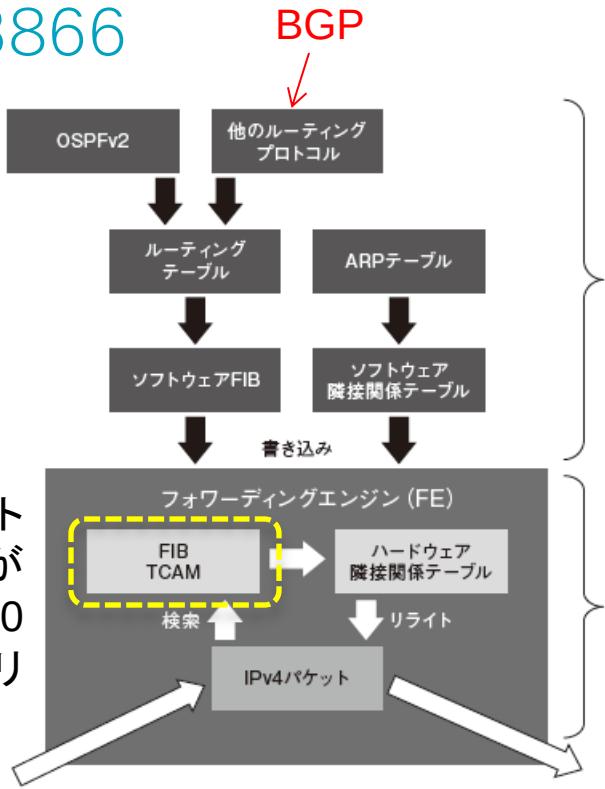
```
Sup2T-VSS#configure session test1
Sup2T-VSS(dry-run-config)#ip access-list extended test-acl1
Sup2T-VSS(dr-config-ext-nacl)#permit ip host 1.1.32.1 host 2.2.33.1
                             (大量のRACL)
Sup2T-VSS(dr-config-ext-nacl)#permit ip host 1.1.48.1 host 2.2.49.1
Sup2T-VSS(dr-config-ext-nacl)#exit
Sup2T-VSS(dry-run-config)#validate
Jan 16 11:39:57.653: %FM-6-SESSION_VALIDATION_RESULT_INFO: Session validation result: Validation Completed..
  Please use 'show config session test1 status' for details.
Sup2T-VSS(dry-run-config)#end
Sup2T-VSS#show config session test1 status

===================================
Status of last config validation:
Timestamp: 2014-01-16@11:39:57
===================================

SLOT = [17]    Result = Configuration will fit in TCAM.
SLOT = [19]    Result = Configuration will fit in TCAM.
SLOT = [33]    Result = Configuration will fit in TCAM.
SLOT = [35]    Result = Configuration will fit in TCAM.

Sup2T-VSS#
```

この構成には、PFCとDFCが
合計4個あるため、RACLは
4つのTCAMにそれぞれ
書き込まれます。

(ACLではなく、ルーティングテーブルとTCAMの話題です)

Cat6k FIB TCAM > 512K : Field Notice 63866

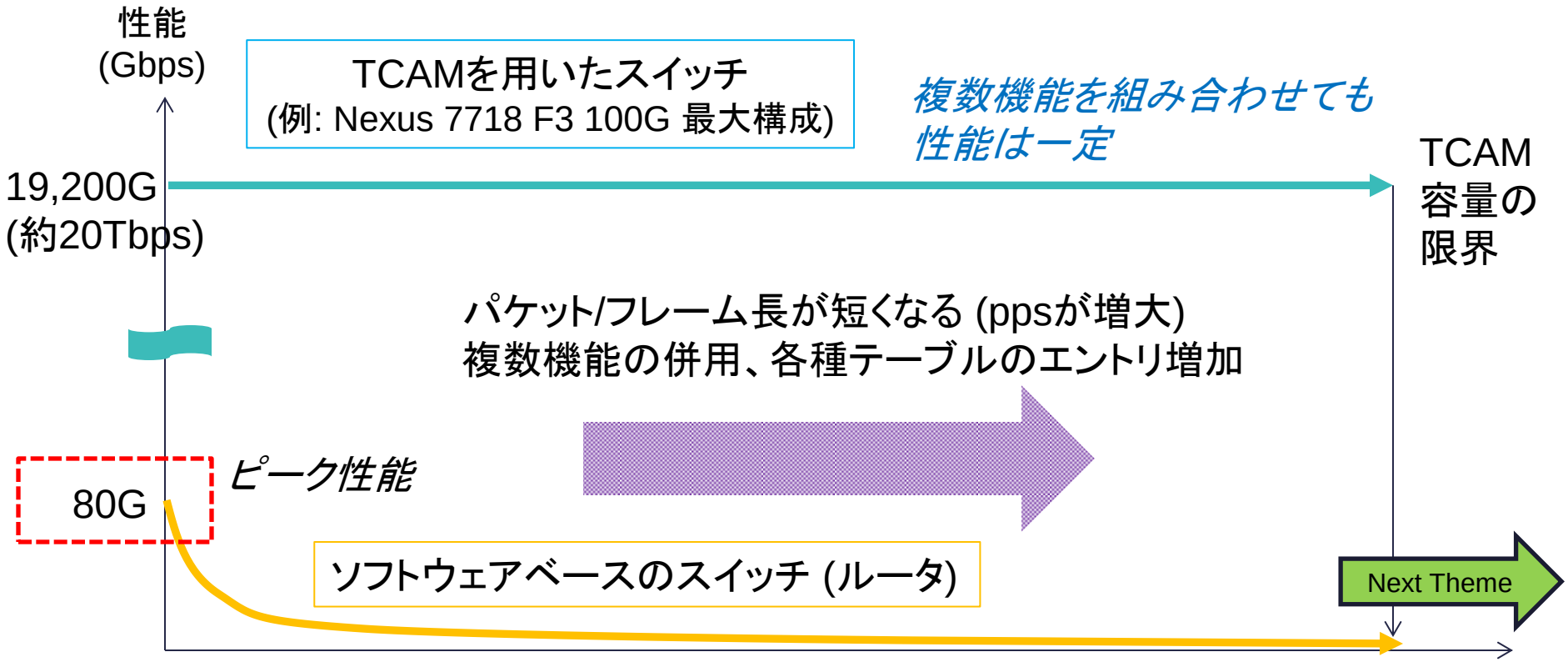
- 8月に、インターネットのIPv4経路数(フルルート)が512,000を超えました。
- このため、SUP720-3BXL等を搭載したCatalyst 6500 や Cisco 7600で、FIB TCAMが溢れ一部のパケット転送がソフトウェア処理になる現象が発生しました。(FIB TCAMの合計容量 = 1M)
- FIB TCAMの割り当ては変更可能なので上記の現象は予防または回避できます。
- Sup2T-XL、Catalyst 6880-X (2M FIB), Nexus 7000 M1-XL, M2-XLモジュールでは挙動がどう異なるかの詳細を、現在調査中です。



デフォルト割り当てが512,000エントリ

P.147 図3.14 FIB方式（CEF）

なぜTCAMを使い続けるのか?



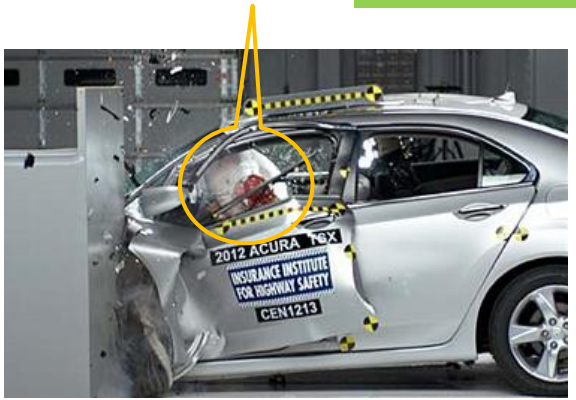
[その2]
STP安定化技術を、実際に「発動」させてみる

Catalyst & Nexus共通
難易度 2 (中)

STP安定化技術 --- 何が難しいのか?

1. 発動させることが難しい場合がある
   → そもそも想定外の状態に対処する機能なので
   正常時は動作しない。エアバッグやブレーカ
   のようなもの
   → BPDUや他の制御フレームを選択的に、
   かつ片方向だけ止められる構成を作り、
   「想定外の状態」を意図的に作り出す必要あり

2. 種類が多すぎて覚えられない
   → 実際に「発動」させられると、
   挙動が肌で分かるので、頭に残ります



(C) IIHS (iihs.org)

Catalystを物理的に壊すわけではないので、ご安心ください

P.280 表7.1 STP安定化技術の一覧

	目的	動作	自動復旧	ポートチャネルの場合
ルートガード（Root Guard）	想定外のスイッチがルートブリッジになることを防ぐ	優先度の高いBPDUを受信している間、ポートを不整合状態にしてブロックする	優先度の高いBPDUを受信しなくなると復旧	ポートチャネル全体に対して動作
BPDUガード（BPDU Guard）	スイッチを接続することを想定していないポートへのスイッチの接続を防ぐ	BPDUを受信すると、そのポートをerrdisable状態にする	なし（手動操作またはerrdisable recoveryを使用）	ポートチャネル全体に対して動作
ループガード（Loopguard）	ブロッキング状態のポートが、意図せずにフォワーディングになることを防ぐ	BPDUの受信が途絶えると、ポートを不整合状態にしてブロックする	BPDUを再び受信すると復旧	ポートチャネル全体に対して動作
UDLD（Normal）	レイヤ1では直接わからないリンクの片通を検出する	相手が自分を認識していない（相手から自分への片通状態）と判断したらerrdisable状態にする	なし（手動操作またはerrdisable recoveryを使用）	メンバーごとに独立して動作
UDLD（Aggressive）	レイヤ1では直接わからないリンクの片通と全断を検出する	UDLD Normalに加えて、相手から一定時間UDLDの制御フレームを受信できなければerrdisable状態にする	なし（手動操作またはerrdisable recoveryを使用）	メンバーごとに独立して動作

この表に加えて、
- Dispute
- BA (Bridge Assurance)

が Cat6kとNexusで
サポートされています。

(P.278 - 279)

Dispute: BPDUが正常に届かず
対向のポートが想定外の状態に
なった時に、ポートをブロックする

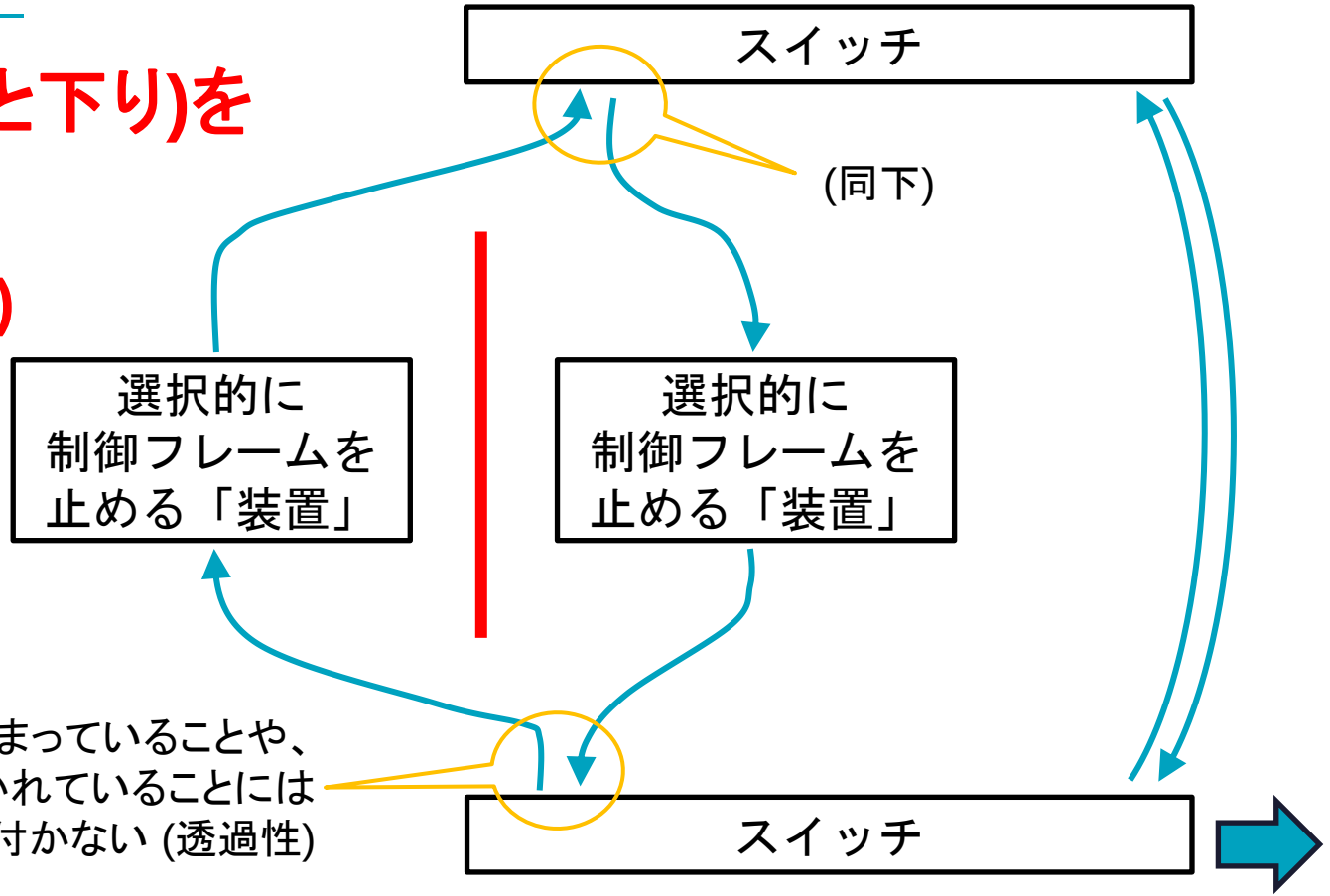
BA: BPDUを双方向に送受信する
ことにより、コントロールプレーン
全体の障害などが原因で正常に
STPが動作していない機器を
認識し、ループを防ぐ

実機で確認せずに、全部暗記するの大変です!!!

試してみよう ---

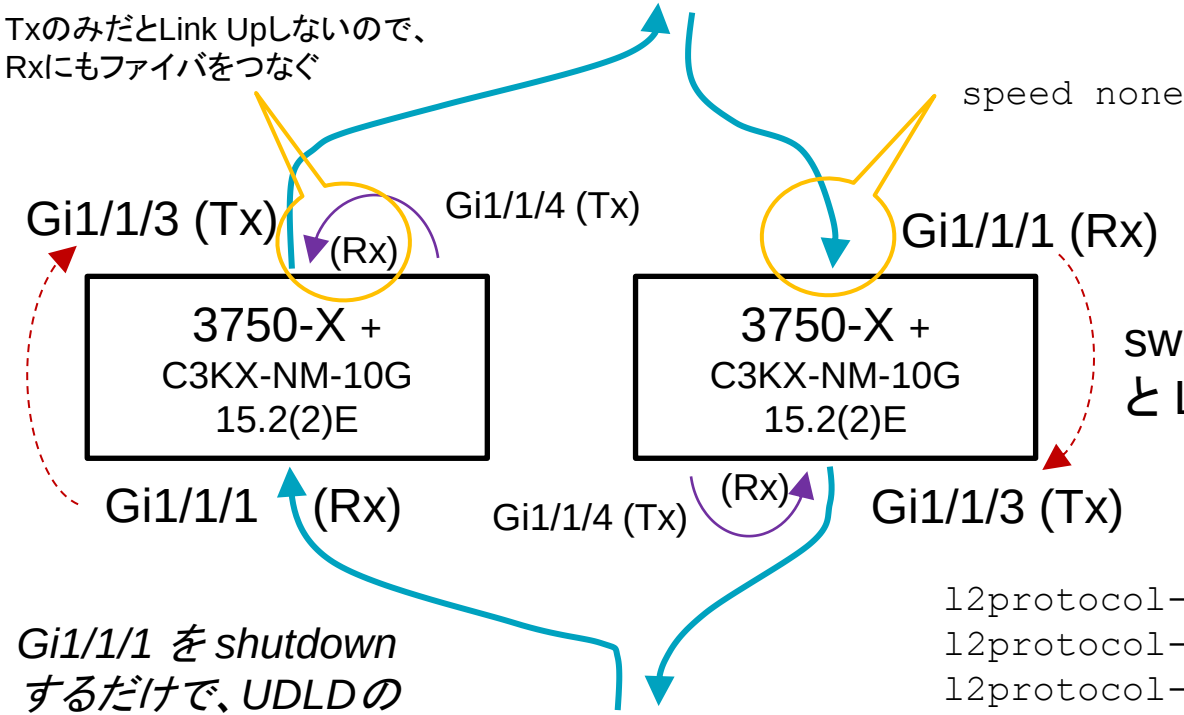
行きと帰り(上りと下り)を分離する

(光ファイバが必要)



途中に何かが挟まっていることや、
行きと帰りが分かれていることには
気付かない (透過性)

選択的に制御フレームを止める構成例

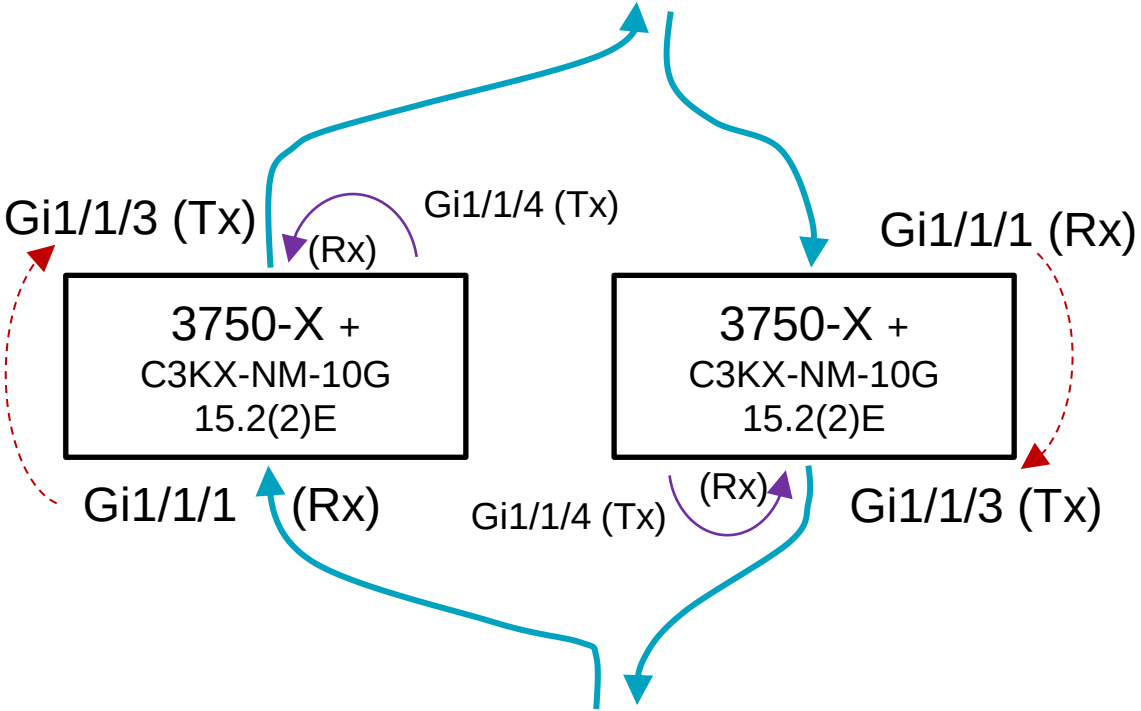


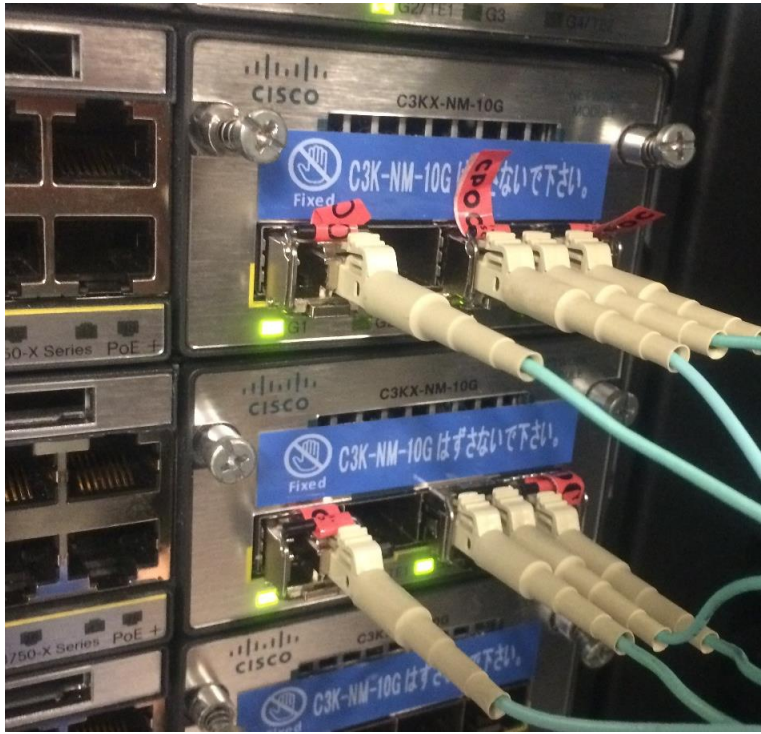
止めたいプロトコルに
対するコマンドを消す

```
l2protocol-tunnel cdp
l2protocol-tunnel stp
l2protocol-tunnel point-to-point pagp
l2protocol-tunnel point-to-point lacp
l2protocol-tunnel point-to-point udld
```


あくまでも試験用です。TACサポートはありません

実際に検証している様子



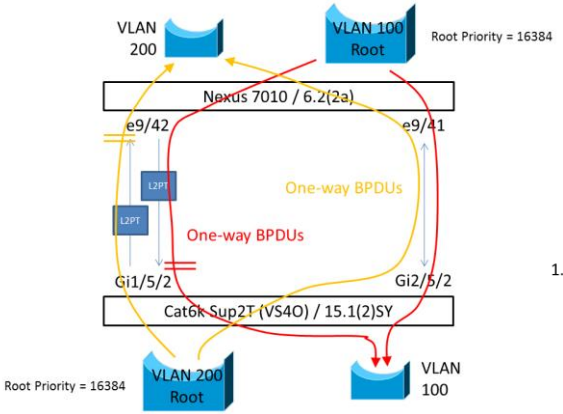


Next Theme

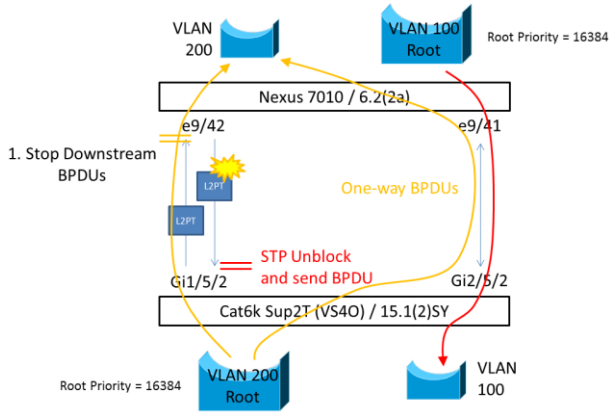
あくまでも試験用です。TACサポートはありません

追加スライド

検証結果の一部 -- Dispute

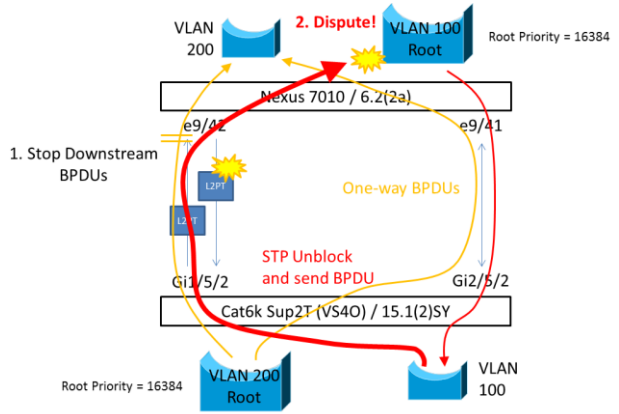


正常状態



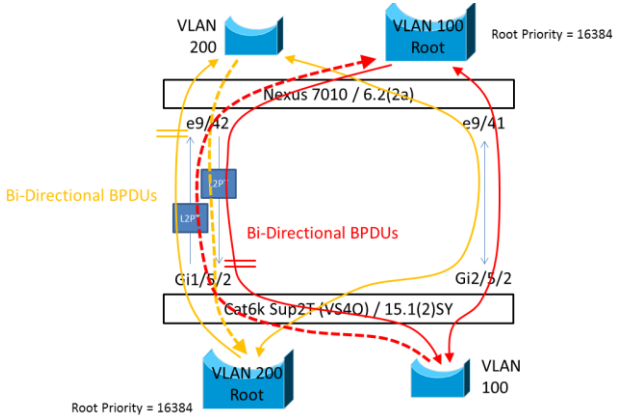
左のリンクの下りだけ
BPDUが通らなくなる

BLKが誤って解除され、
BPDUが上向きに送信
→ Dispute発生

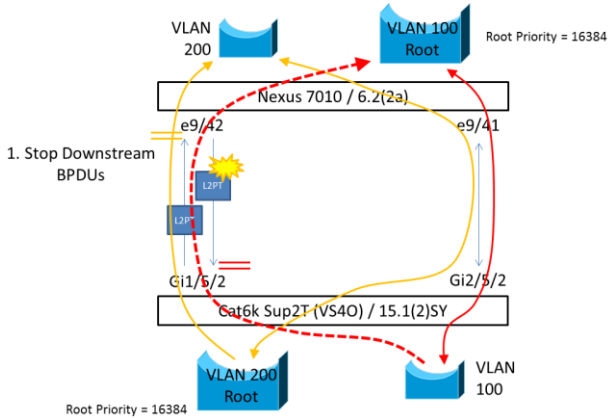


追加スライド

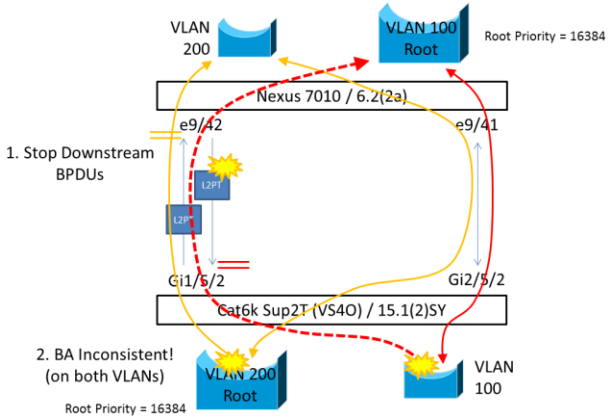
検証結果の一部 -- BA



正常状態。
双方向にBPDUが
飛んでいる



左のリンク経由の
下向きのBPDUが
通らなくなる



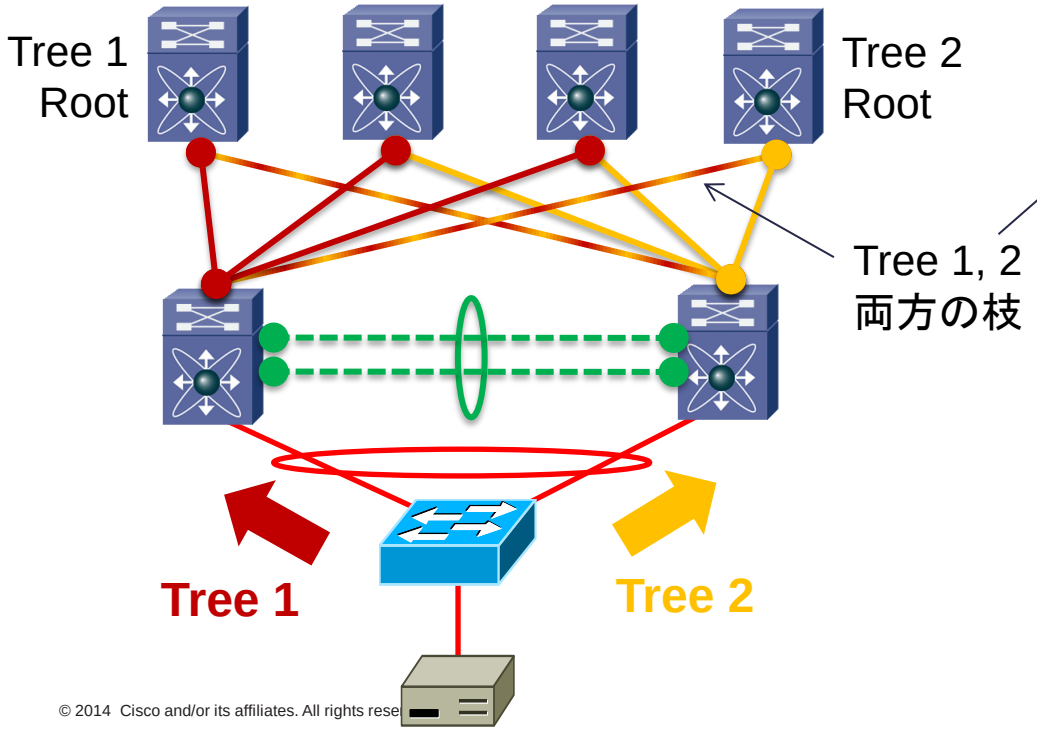
"BA Inconsistent"
発生

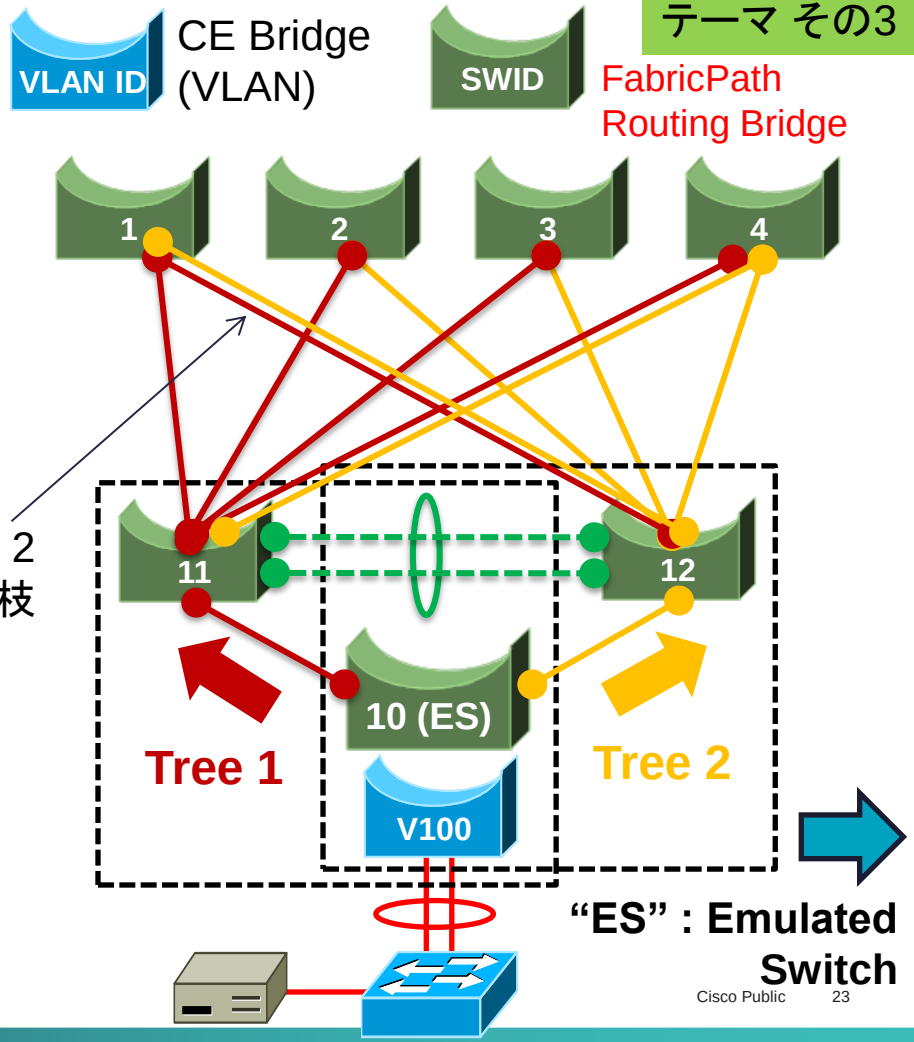
[その3] FabricPathの高度なループ防止メカニズムと vPC+

Nexus限定
難易度 3 (高)

(2013年1月のWebcastより)

FabricPath: vPC+ とツリー (上り)
Broadcast も分散する

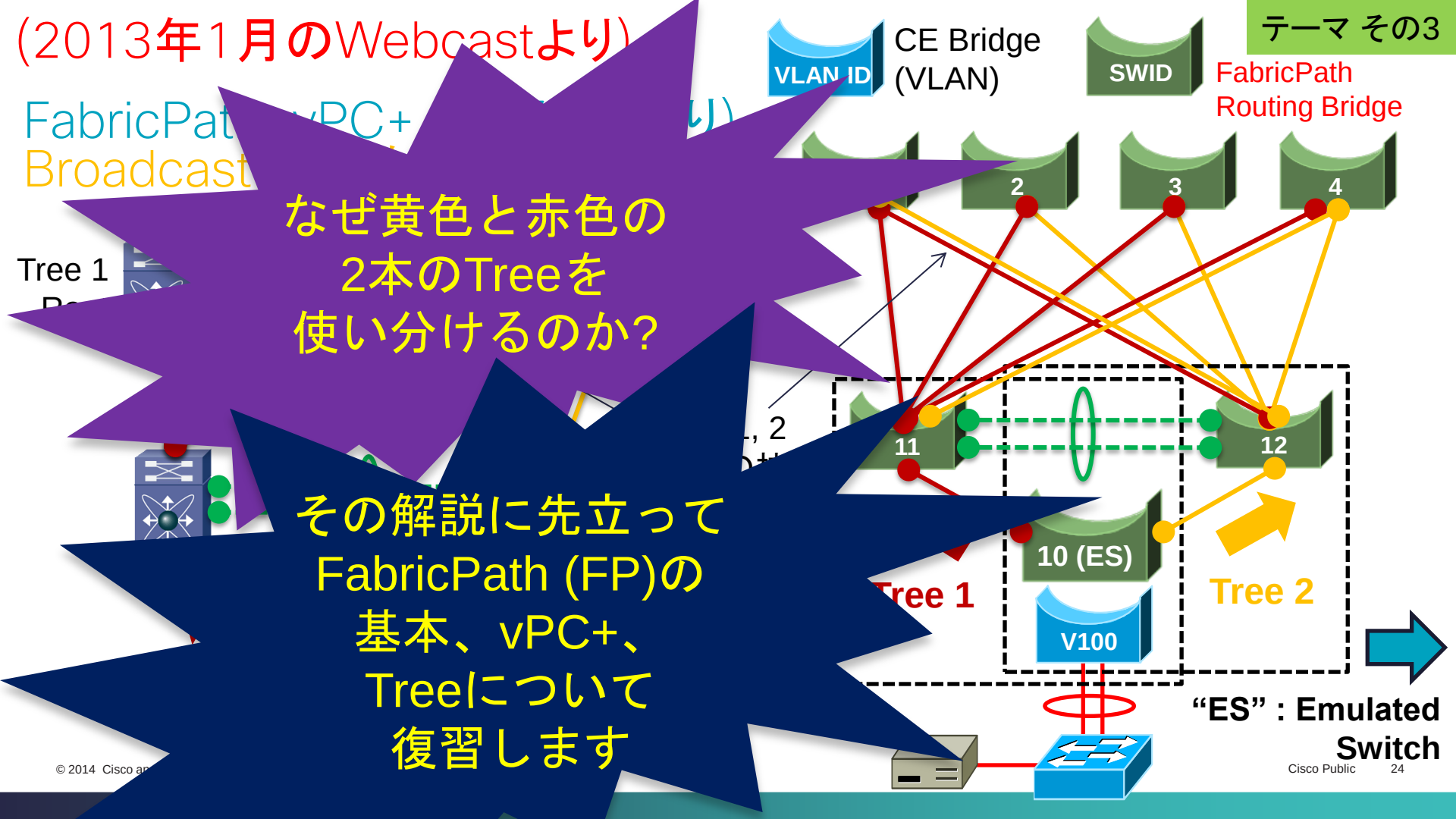




(2013年1月のWebcastより)

FabricPat vPC+
Broadcast

テーマ その3

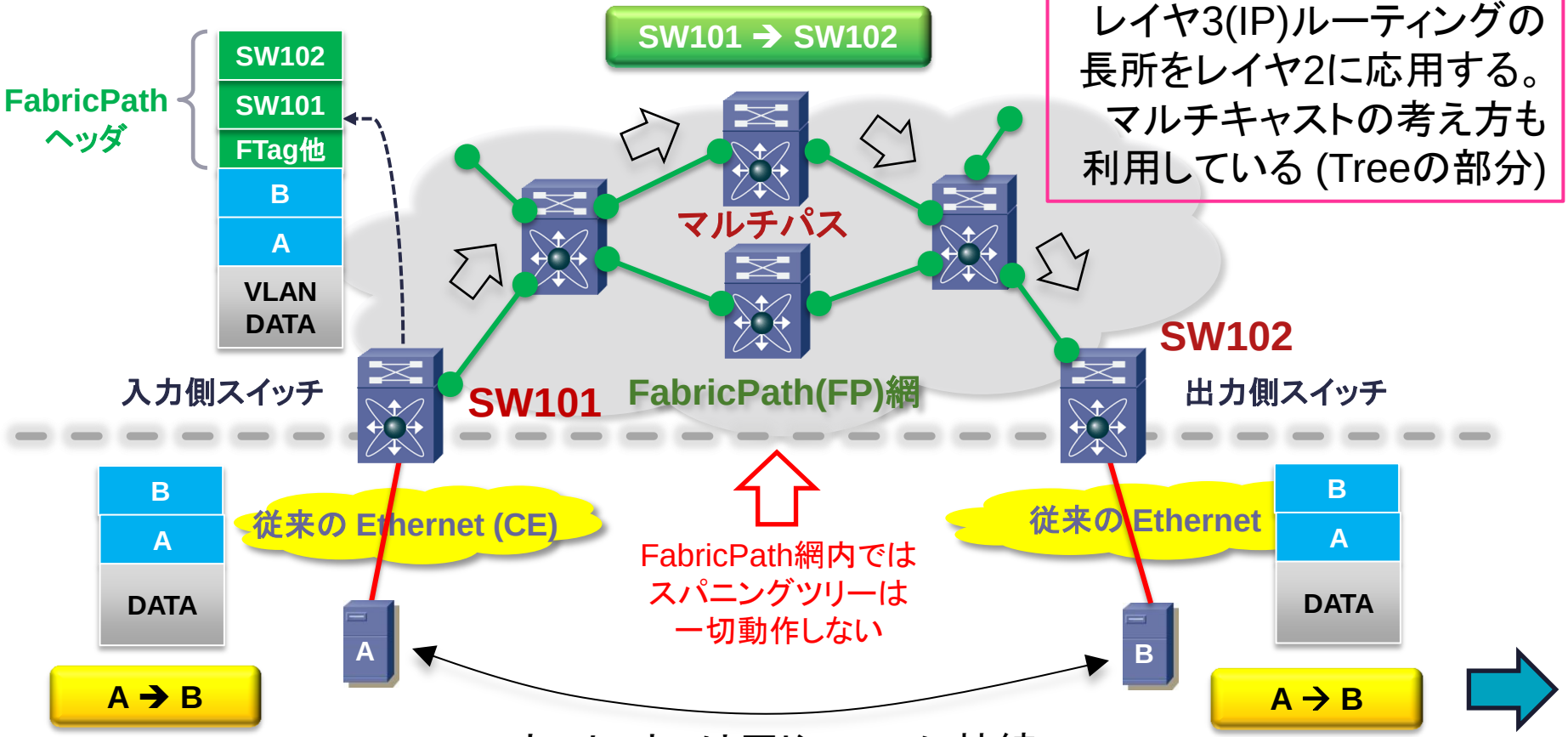


なぜ黄色と赤色の
2本のTreeを
使い分けるのか?

その解説に先立って
FabricPath (FP)の
基本、vPC+、
Treeについて
復習します

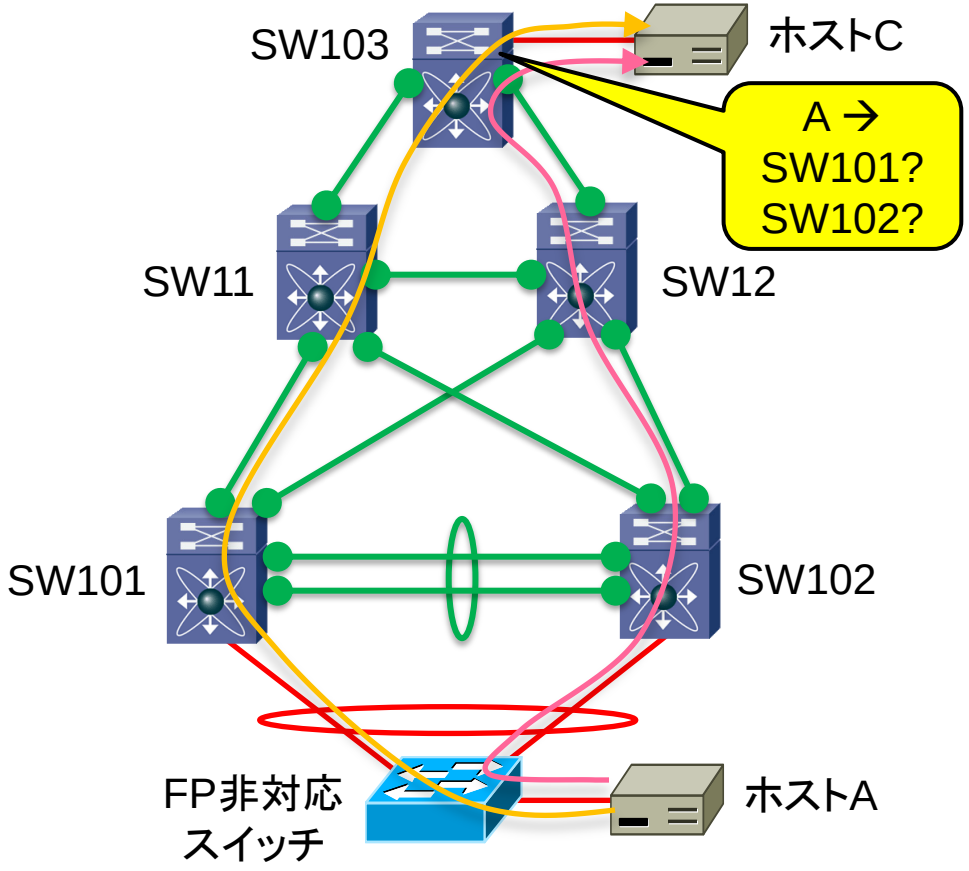
FabricPathの概要 (P.248)

レイヤ3(IP)ルーティングの長所をレイヤ2に応用する。
マルチキャストの考え方も利用している (Treeの部分)



vPC+ とは何か (P.258 図6.16を単純化)

テーマ その3



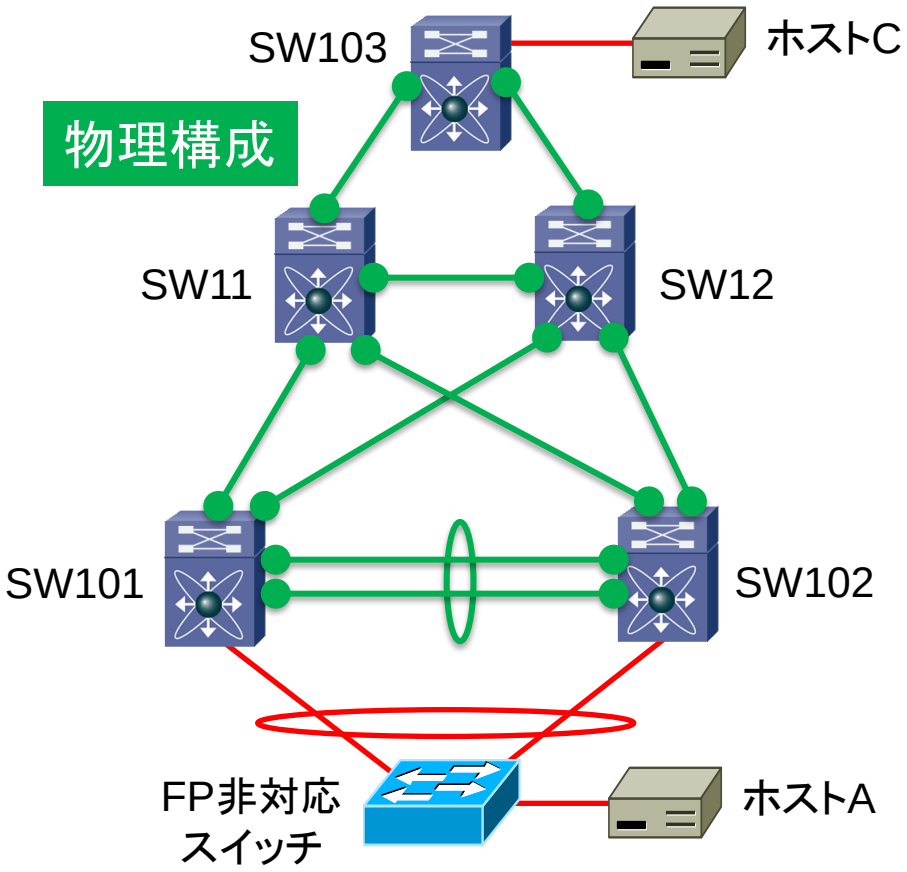
0. vPC+ を使用すると、FP非対応スイッチを従来のイーサチャネル(ポートチャネル)でFP網に接続できる

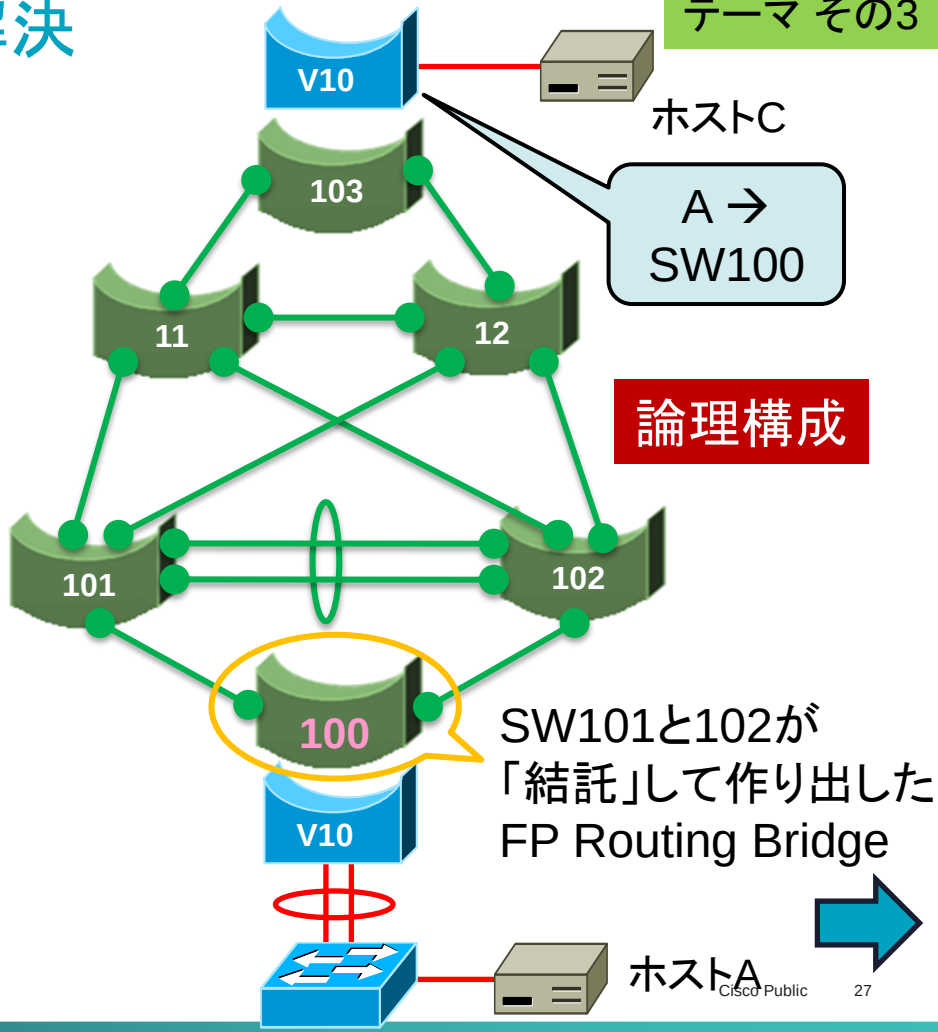
1. FP非対応スイッチ配下のホストAがホストC宛のフレームを送信する。
2. 左側へ振り分けられると、フレームはSW101を経由してFP網に入ってくる
3. ホストCが接続しているSW103は、ホストAをSW101とセットで学習する

4. 右側へ振り分けられると、フレームはSW102経由でホストCに届く
5. SW103では、ホストAがSW101とSW102のあいだでフラップしているように見えてしまう

Emulated Switch (ES) を導入し解決

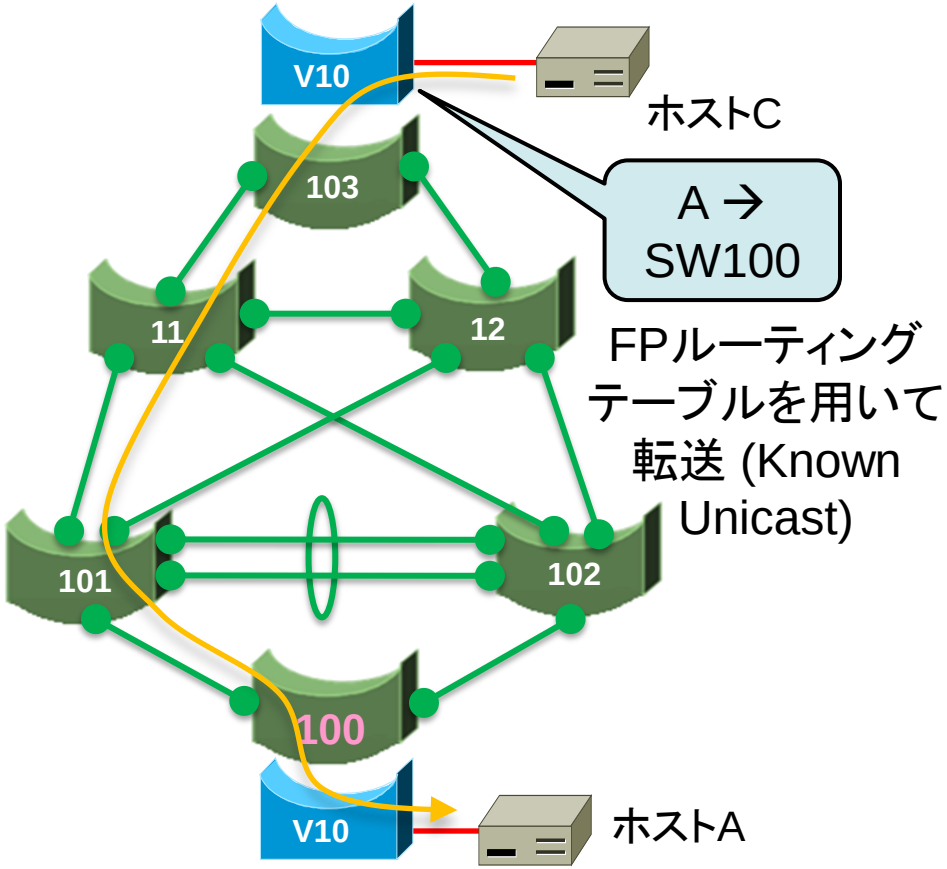
テーマ その3



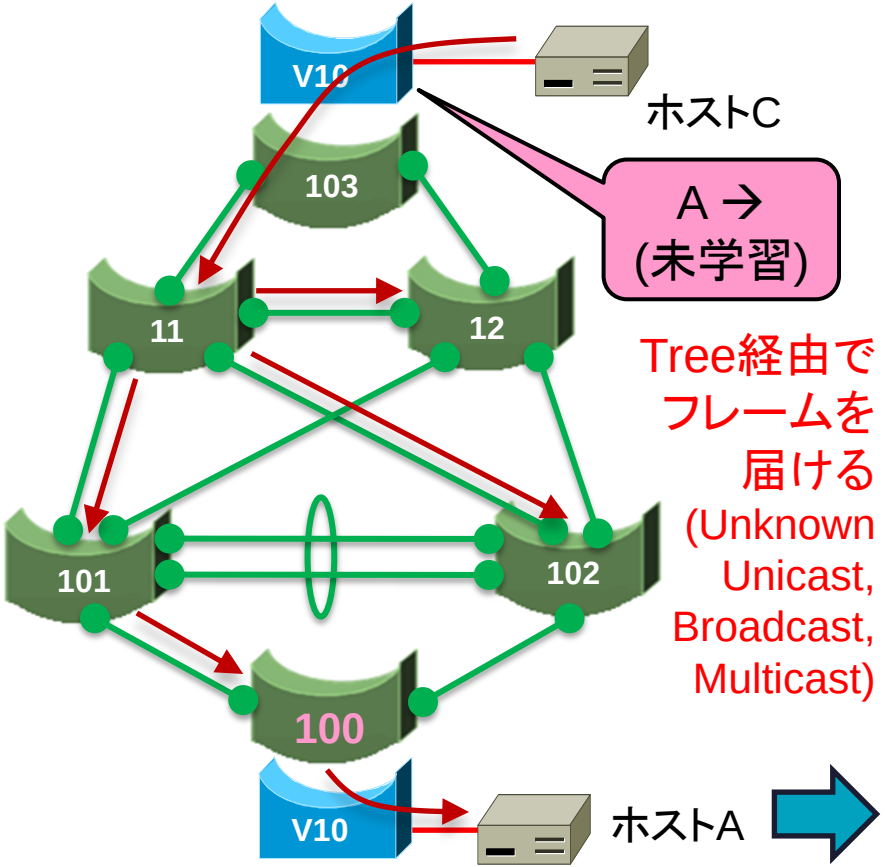


Treeを用いてルーティングできないフレームを転送(P.262)

テーマ その3



FPルーティングテーブルを用いて転送 (Known Unicast)



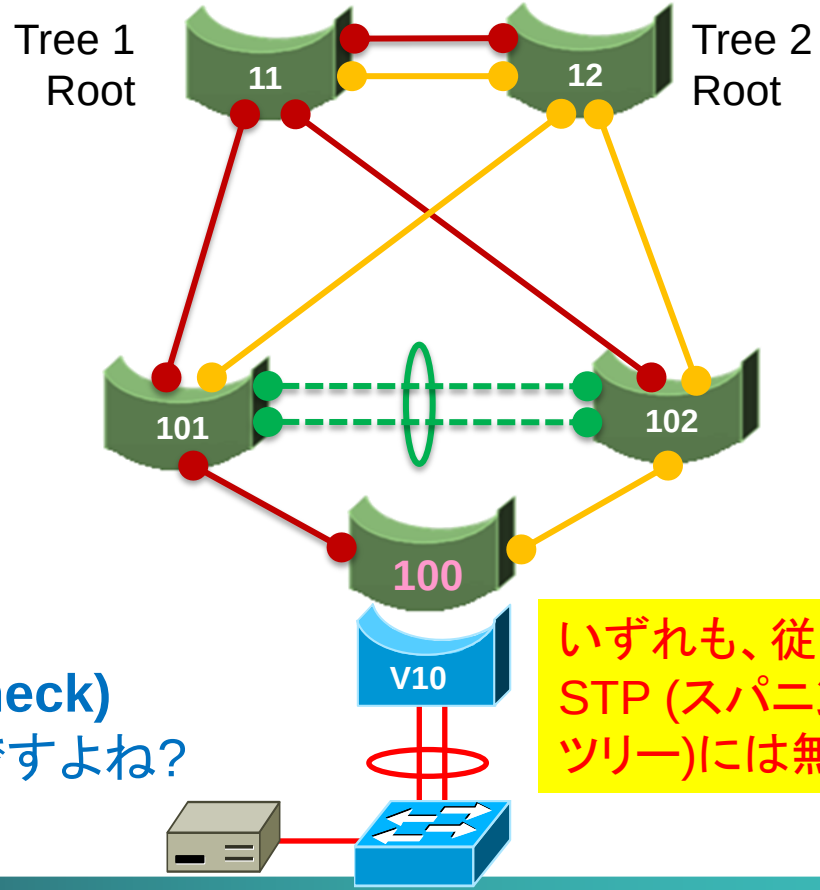
Tree経由でフレームを届ける (Unknown Unicast, Broadcast, Multicast)

FPはTree経由の転送において、3つの方法でループを防止する

1. リンクステート型ルーティングプロトコル(IS-IS)の利用
   → ネットワークの全体像を各スイッチが把握してからそれぞれが Treeを構築

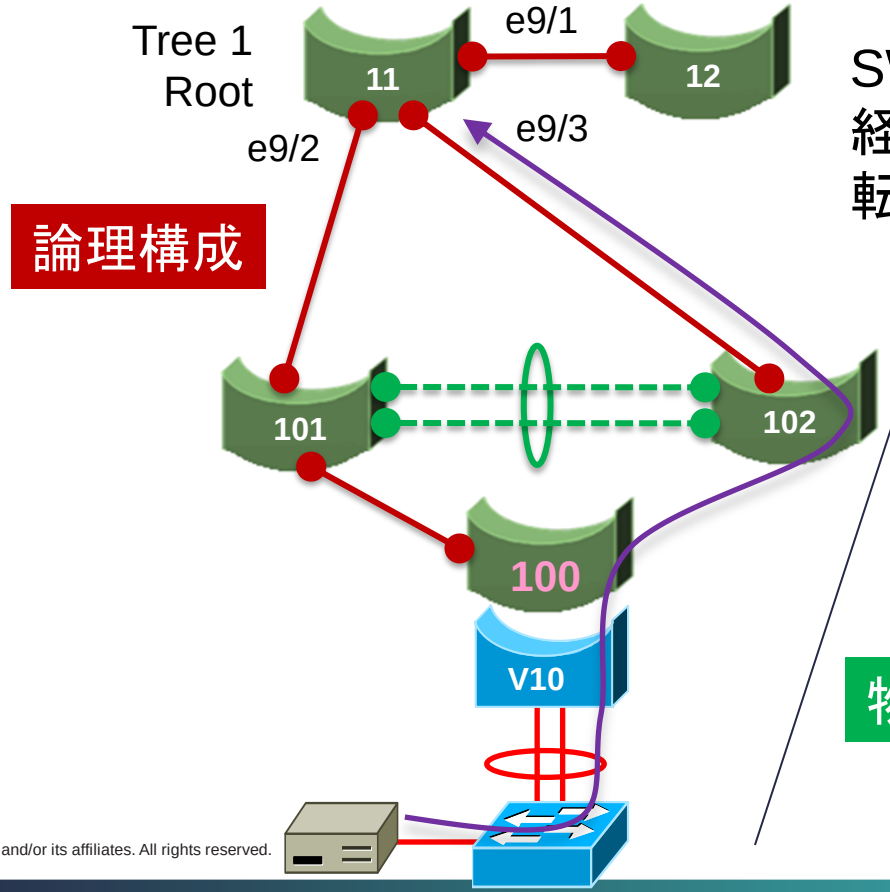
2. TTLの減算 → フレームの破棄

3. RPF Check
(Reverse Path Forwarding Check)
IPマルチキャストでおなじみ --- ですよね?



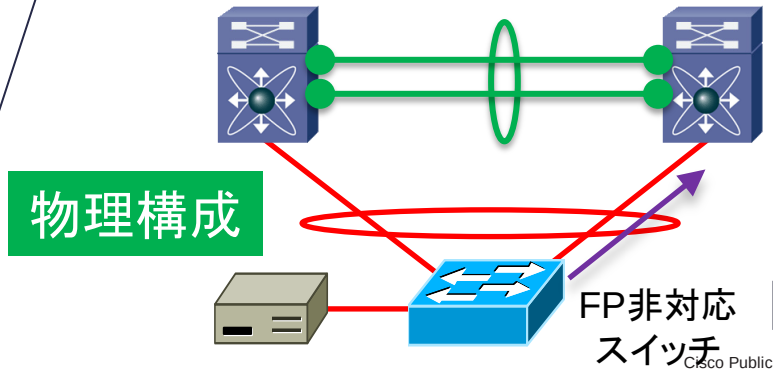
いずれも、従来のSTP (スパニングツリー)には無い

Tree 1のみに着目。vPC+配下からBroadcast等を受信した場合



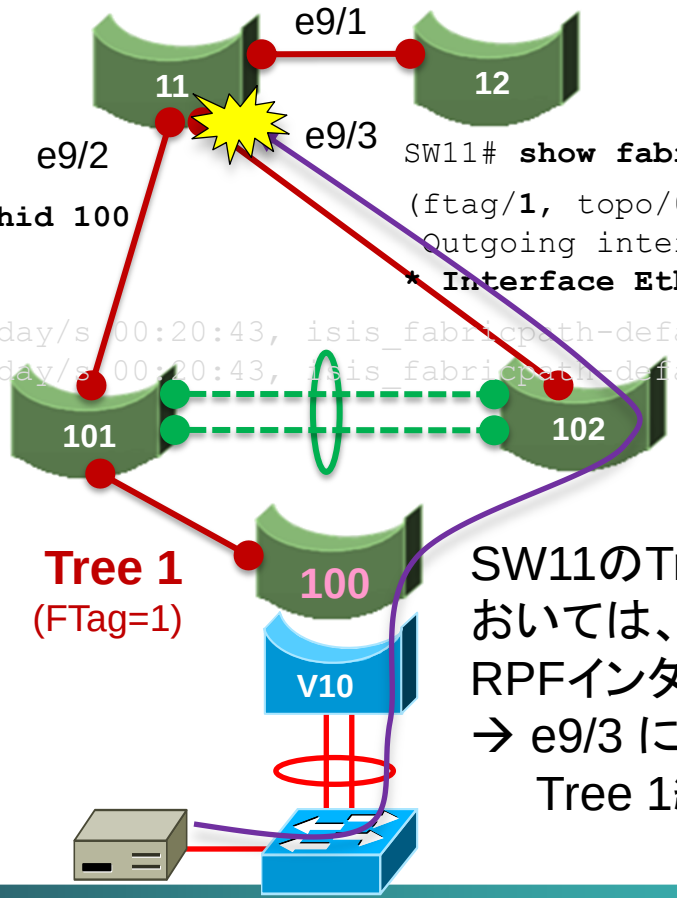
SW100 → SW102 → SW11 の経路でフレーム(Broadcast等)が転送されてきたら?

左右に振り分けているのはFP非対応スイッチなので、FP側からは制御できない



Unicast Route と RPFインターフェイス

```
SW11# show fabricpath route switchid 100
(略)
1/100/0, number of next-hops: 2
        via Eth9/2, [115/400], 0 day/s 00:20:43, isis_fabricpath-default
        via Eth9/3, [115/400], 0 day/s 00:20:43, isis_fabricpath-default
```




```
SW11# show fabricpath multicast trees ftag 1
(ftag/1, topo/0, Switch-id 100), uptime: 00:2
  Outgoing interface list: (count: 1, '*' is t
  * Interface Ethernet9/2, [RPF]  [admin distar
```

Tree 1
(FTag=1)

SW11から見ると、
Unicast Routing Tableでは
SW100への経路は等コストの
エントリ(e9/2, e9/3)が2つある
(SW101 - 100, 102 - 100 のコストは0)

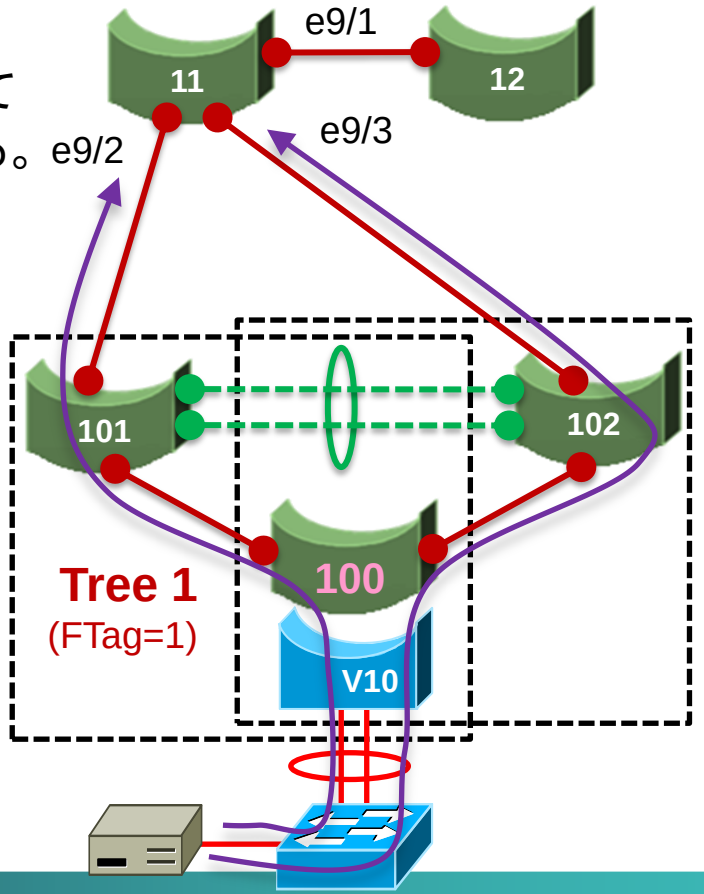
SW11のTree 1 (FTag=1)に
おいては、SW100に対する
RPFインターフェイスは e9/2 のみ
→ e9/3 に誤って着信した
Tree 1経由のフレームは破棄

なぜ vPC+ では2本のTreeを使い分けるのか

SW100からSW101と102の両方を経由して
フレーム(Broadcast等)がSW11に着信する。

→ もし Tree が1本しか無いと、
RPFインターフェイスは1つなので、
片方しかRPF Checkが成功しない
(反対側は常に失敗)

→ 対策:
Treeを2本に分けて、RPFインター
フェイスを1つに限定する



Treeを2本にすることで、RPF Checkも大丈夫

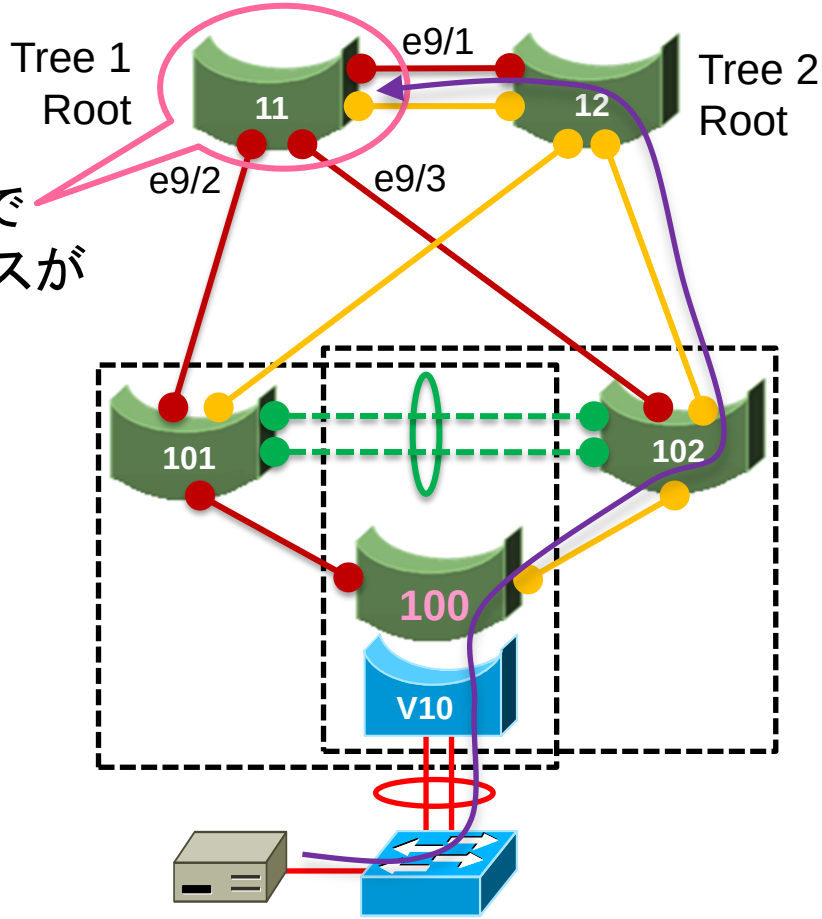
Tree 1とTree 2で
RPFインターフェイスが
分けられる

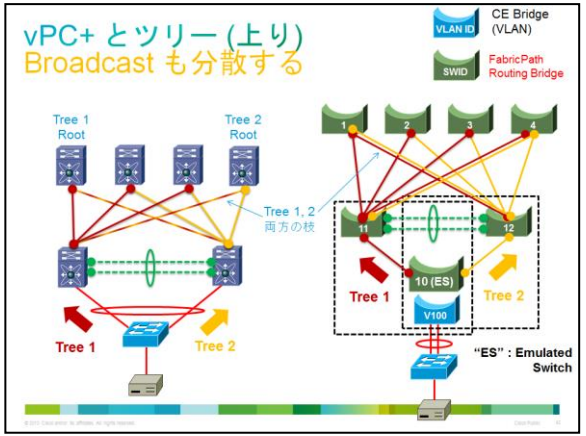
Tree 1 :

SW12	e9/1
SW101と100	e9/2
SW102	e9/3

Tree 2 :

SW12	e9/1
SW101と100	e9/1
SW102	e9/1





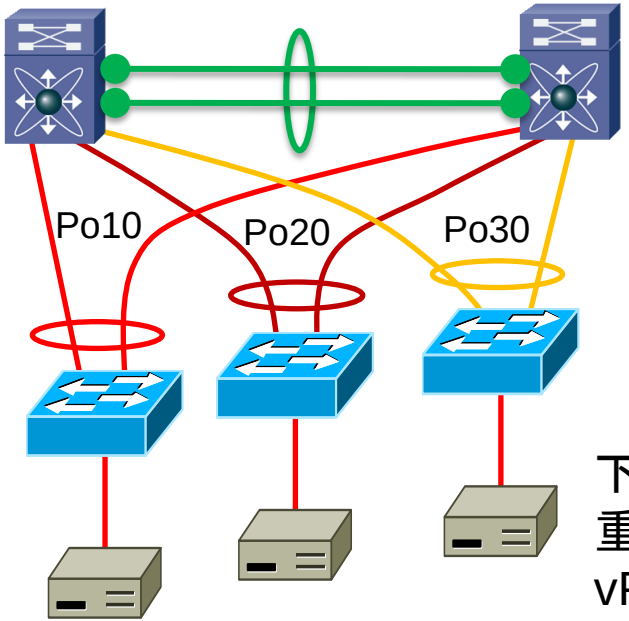
2013年1月のWebcastで
ご説明した赤色と黄色の
2本のTreeには、
このような背景が
ありました。

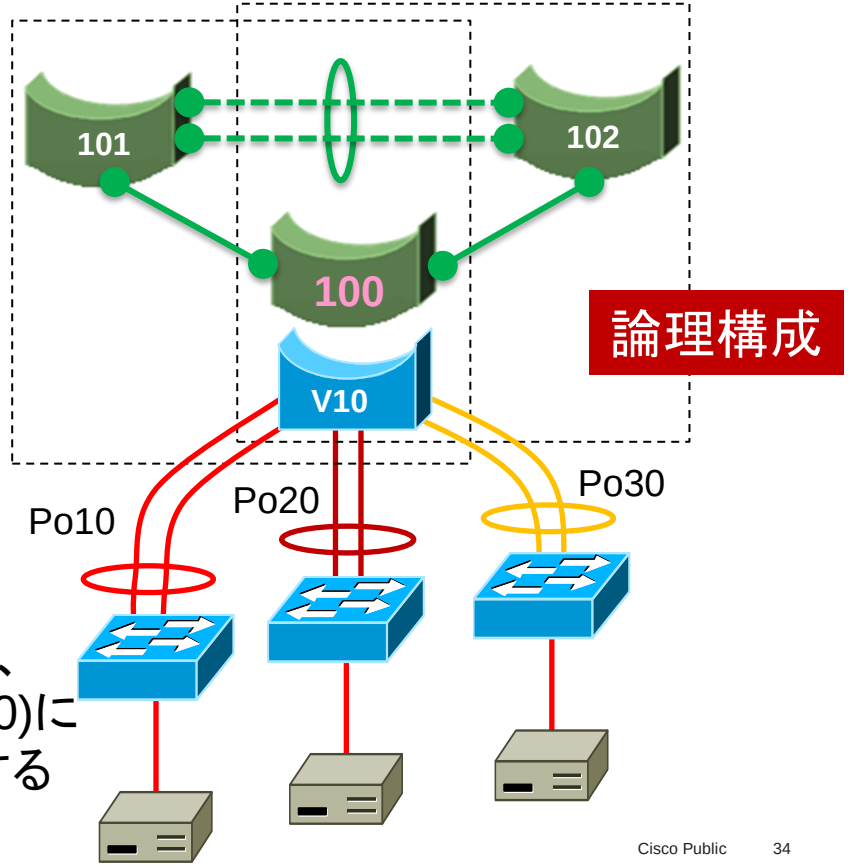
Next Theme

vPC+の下りは Designated Forwarder (DF)で制御

追加スライド

物理構成



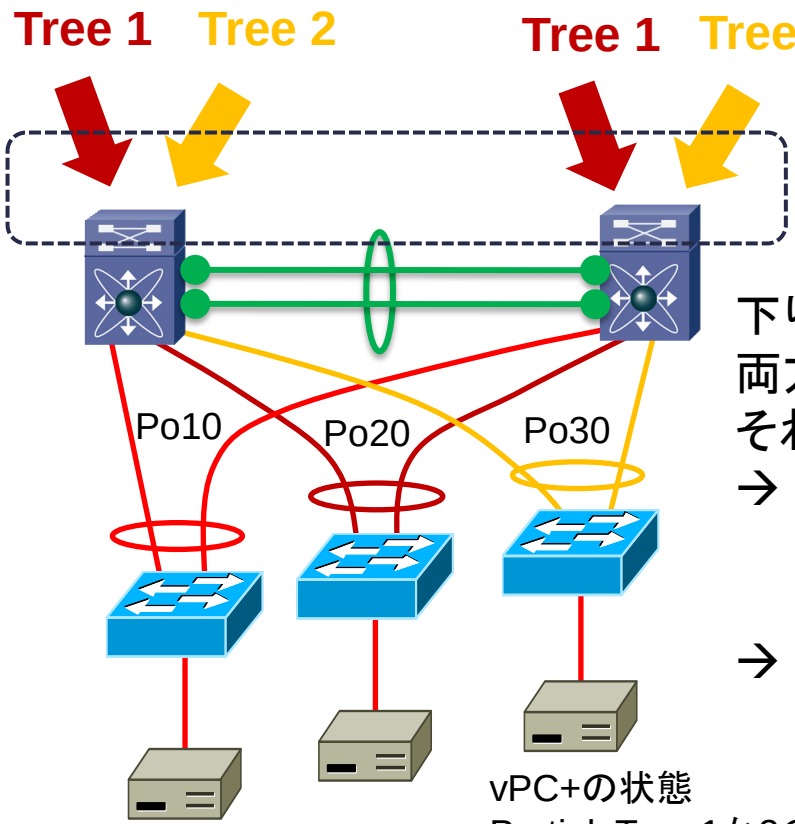


論理構成

下りのフレームが
重複しないために、
vPC+毎 (10,20,30)に
一方だけが転送する
→ DFにより実現

vPC+の下りは Designated Forwarder (DF)で制御

追加スライド

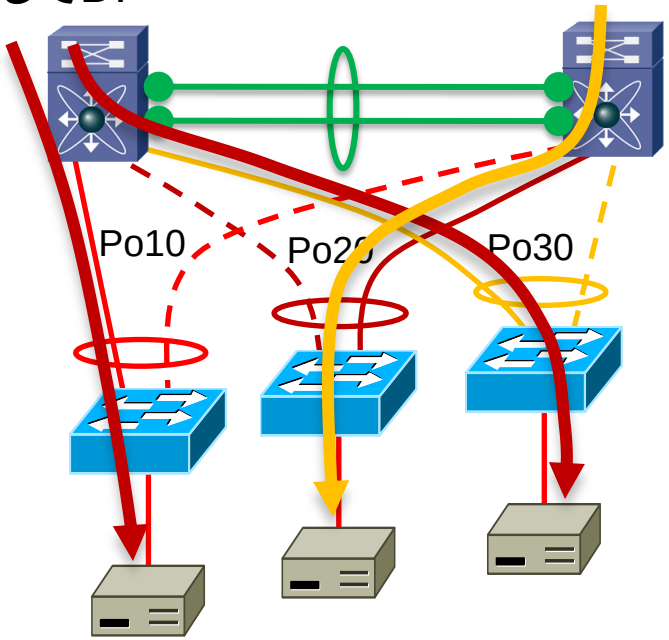


下りはvPC+の
両方のPeerに
それぞれ着信する
→ 両方のPeerが
下へ転送すると
フレームが重複
→ DFだけが転送

vPC+の状態
Partial: Tree 1か2の一方に対してDF
Full: Tree 1と2の両方に対してDF
None: DFではない

Tree 1
Po10, Po30に
対してDF

Tree 2
Po20に対してDF



vPC+のメンバがダウンすると、
Po毎にDFが変わる

[その4]
HSRP Preempt使用時に、再起動の際のパケットロスを防ぐには

Catalyst & Nexus共通 (Nexus分多め)
難易度 2 (中)

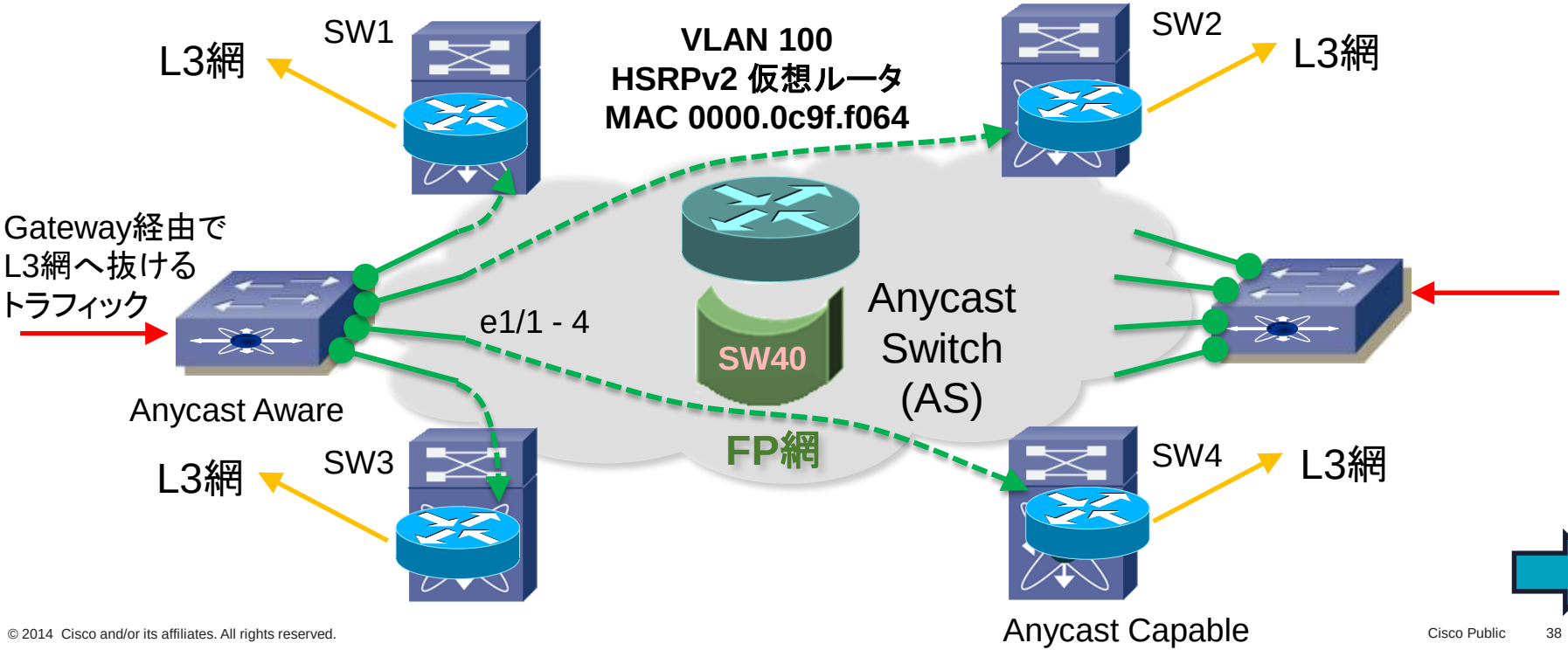
ご回答

- HSRP Preemptは、もし明確な理由がないのであれば、使用しない方が楽です。(プレゼンタ個人の見解です)
- もし「どうしてもPreemptを使わなければならない」のであれば、"preempt delay reload <秒>" コマンドを利用することにより、スイッチ再起動時に指定された時間だけ待ってから、Activeに昇格します。(Catalyst、Nexus共通)
- ただし、SVI にHSRP Preemptを設定する場合には、再起動時にSVIのUpやHSRP Helloの受信開始が遅くなり、タイマーの調整が必要になることがあります。
  - STP Forwarding状態への遷移の遅れ
  - モジュール型機種で、ラインカード/モジュールの初期化に時間が必要な場合
- P.268「FabricPathの拡張機能」による別のソリューションを紹介します。

FP Anycast HSRP を利用すると (P.268, 269)

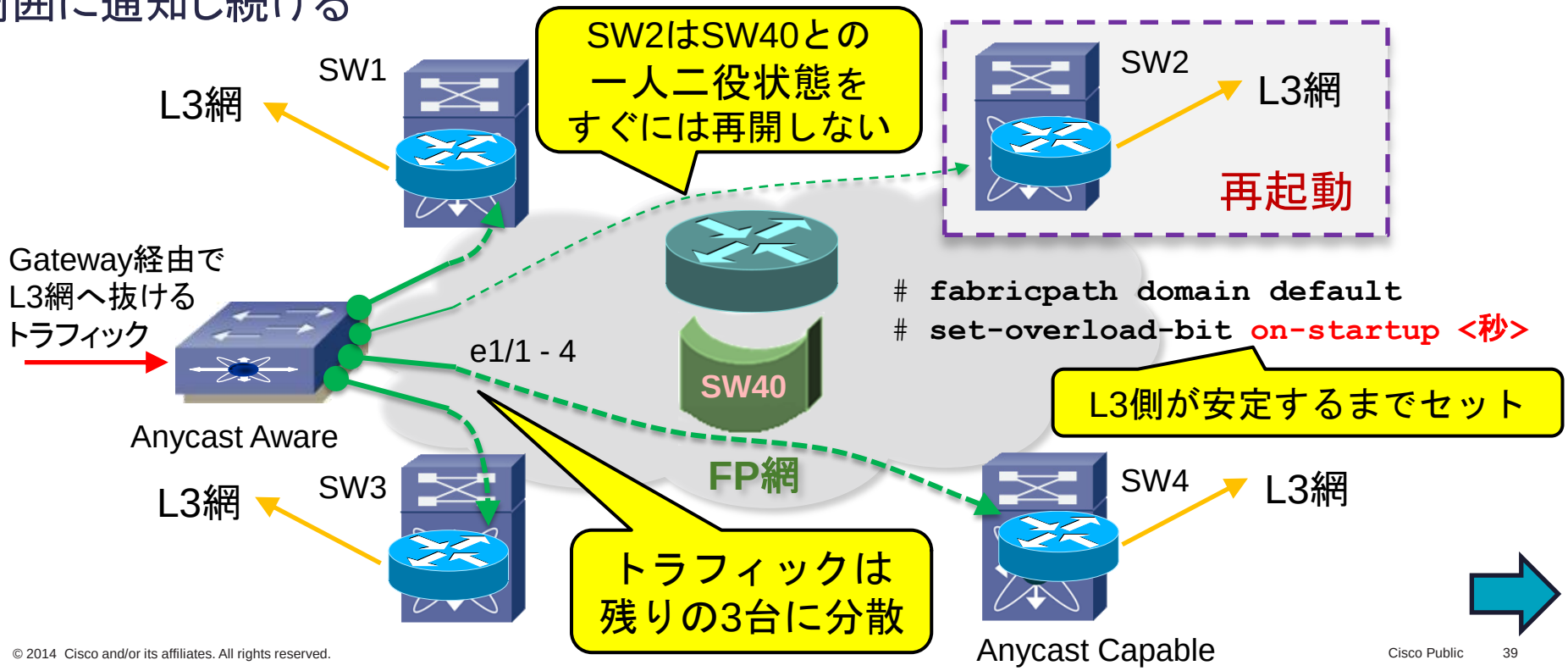
4台が全てActive Gatewayとして動作する

→ 4台いずれも、SW40 (AS)を兼任している



FP Anycast HSRP + Overload Bit (P.270)

Overload Bitがセットされているあいだ、自分がGatewayではないと
周囲に通知し続ける



FP Anycast HSRP の地味な長所

テーマ その4 [終了]

- 2台や3台でも構成可能
- 障害発生時に、Standby が Active に昇格するのではなく、
  Active x 4台が 3台に縮退するので、切り替わりが非常に速い
  - FabricPathルーティングの収束が早いため
  - HSRP Timerをチューニング (e.g. 250msec / 750msec)する必要がない
  - 例えばホストが100台あって、デフォルトゲートウェイを均等に分散して使用していれば、
    25台だけが影響を受ける。再起動後の切り戻りの時も、同じ25台が復旧した
    デフォルトゲートウェイを使うようになり、残りの75台には影響がない。
- Overload Bitは、起動時以外にも必要に応じてOn/Offできるので、
  トラフィックへの影響を最小限にしてメンテナンスや構成変更が可能。

```
# fabricpath domain default
# set-overload-bit always
# no set-overload-bit always
```

End of Theme

お疲れさまでした!

- 以上で、ご用意した4つの話題に関するご説明は終了です。

- それでは、これからご質問をお受けします。
  - 可能な限りこの場で回答します
  - が、即答できない場合は「質問箱」で後日お答えします

- 最後に、受講者プレゼントについてのお知らせがあります。

Thank you.