SONiC + ホワイトボックススイッチで実現する EVPN/VXLANロスレスイーサ

APRESIA Systems株式会社

大渕 光希(koki.obuchi.tc@apresiasystems.co.jp)

2023.11.10

◆氏名

◇ 大渕　光希（おおぶち　こうき）

◆経歴

◇ 2015年　日立金属株式会社※に入社

◇ 2015年-2021年 APRESIAシリーズの製品ソフトウェア開発を担当

◇ 2021年-現在　オープンネットワーキング(SONiCなど)の業務を担当

※2016年にAPRESIA Systems株式会社として通信機器事業をカーブアウト

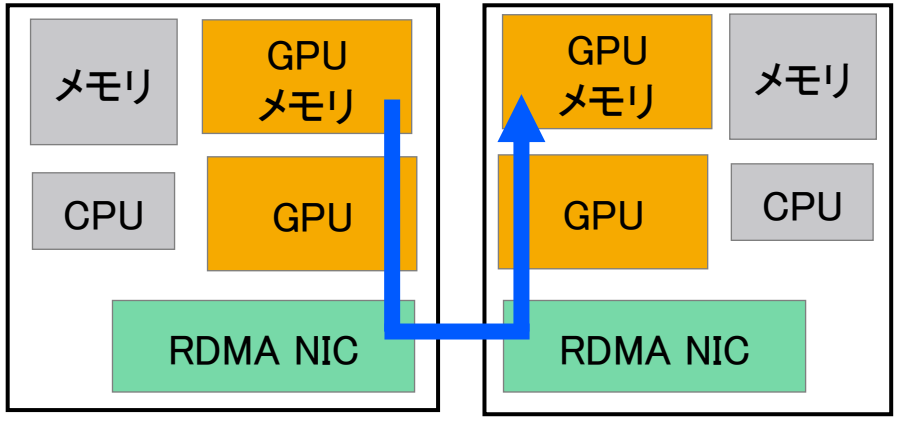
AI/ML基盤の事例・発表

- 大規模AI/ML基盤の構築
  - 2023/6/16　さくらインターネット、生成AI向けクラウドサービス開始へ

    ～NVIDIA H100 GPUを搭載した2EFの大規模クラウドインフラを石狩データセンターに整備～
    - https://www.sakura.ad.jp/corporate/information/newsreleases/2023/06/16/1968211860/
  - 2023/7/7 経済産業省による「クラウドプログラム」の
    供給確保計画の認定について（ソフトバンク株式会社）
    - https://www.softbank.jp/corp/news/press/sbkk/2023/20230707_01/
- AI/MLのネットワークの関連プレゼン
  - 2023/7/7 JANOG52 AI/ML基盤の400G DCネットワークを構築した話（株式会社サイバーエージェント）
    - https://www.janog.gr.jp/meeting/janog52/aiml400/
  - 2023/10/26 MPLS Japan 2023 LLM と GPU とネットワーク（ソフトバンク株式会社）
    - https://mpls.jp/2023/presentations/mpls2023-yuyarin.pdf

- AI/ML基盤のネットワーク構築の事例が出てきている

- RDMA(Remote Direct Memory Access)
  - リモートデバイスのメモリに対してCPUを介さずデータを書き込む技術
    - 低レイテンシ、高スループット、CPU負荷の削減
  - RDMAのデータ転送技術：Infiniband, RoCEv2



- Infiniband
  - クレジットベース(ロスレス保証)のフロー制御が特徴
  - Infiniband対応のスイッチベンダーは少ない
- RoCEv2(RDMA over Converged Ethernet version2) ※本講演のテーマ
  - RDMAの通信をL3プロトコルで中継(Ethernet/IP/UDP)
  - Ethernet自体はロスレスの保証がないため
  - 別途PFC/ECN/ETSなどによりロスレスイーサネットの実現が必要
  - Ethernetスイッチベンダーを採用することが可能

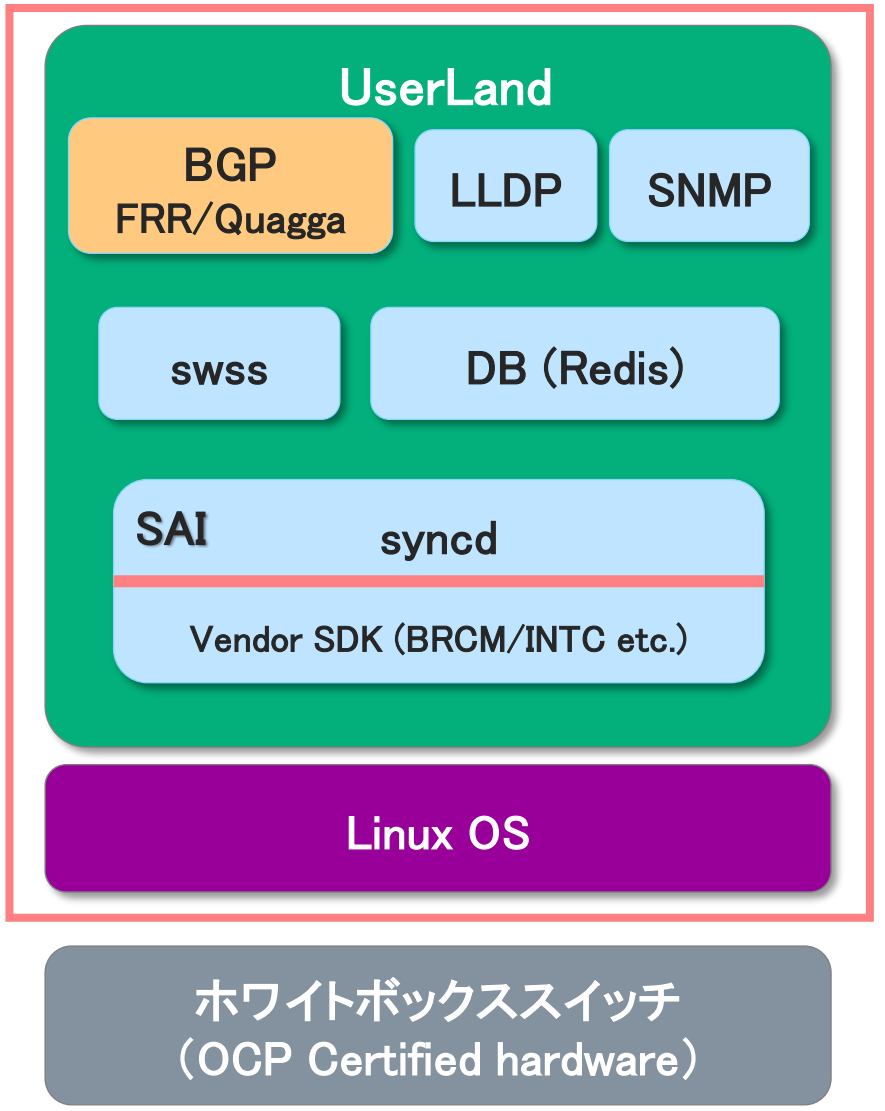
◆AI/ML基盤のネットワークに必要な技術と検証結果の紹介

課題	解決策
ロスレスイーサネットの実現	PFC/ECN/ETS
マルチテナント可能な ファブリックネットワークの構築	IP CLOS Fabric + EVPN/VXLANを併用
オープン技術による実現	SONiCとホワイトボックススイッチの 組み合わせ

- ホワイトボックススイッチ対応OSSベースNOS
- Microsoft公開のソースコードが母体
  - OCP内のプロジェクトから
    Linux Foundationに移行（2022/4）
- マルチスイッチベンダ対応を実現
- BGPベースのIP CLOS Fabricを構築可能

- 本検証では、以下を使用
  - Enterprise SONiC Distribution by Edgecore
    - （以降、Edgecore SONiCと記載）

SONiC（イメージ図）



Edgecore SONiC対応ホワイトボックススイッチ

		400G			100G		
		AS9736-64D	AS9726-32DB	AS9716-32D	AS7816-64X	AS7726-32X	AS7712-32X
製品写真							
ポート構成		64x QSFP-DD	32x QSFP-DD, 6x10G SFP+	32 x QSFP56-DD	64 x QSFP28	32 x QSFP28	32 x QSFP28
SoC		Tomahawk Ⅳ	Trident Ⅳ	Tomahawk Ⅲ	Tomahawk Ⅱ	Trident Ⅲ	Tomahawk
CPU		Intel® Xeon® Processor D-1548 8 cores 2.0 GHz	Intel® Pentium® Processor D1519 4-cores 1.5 GHz	Intel Xeon D-1518 quad-core 2.2 GHz	Intel Xeon D-1518 quad-core 2.4 GHz	Intel Xeon D-1518 quad-core 2.2 GHz	Intel Atom C2538 quad-core 2.4 GHz
メモリ		32GB	16GB	16GB	16GB	16GB	16GB
スイッチ容量（全二重）		51.2Tbps	25.6Tbps	25.6Tbps	12.8 Tbps	6.4 Tbps	6.4 Tbps
パケットバッファ		113.66MB	132MB	64MB	42MB	32MB	16MB
電源	AC	○	○	○	○	○	○
	DC	—	—	—	—	○	○

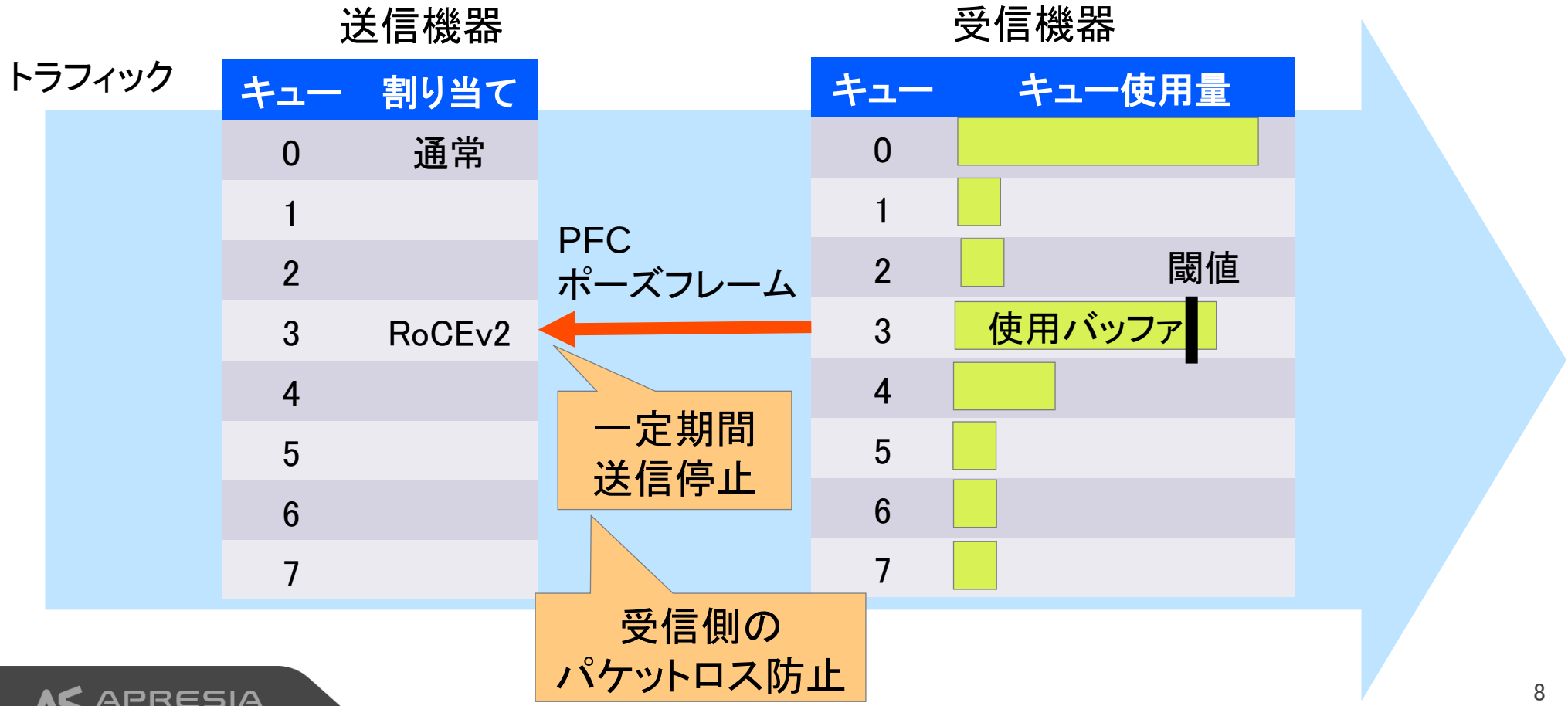
現在検証中

本講演対象

https://www.apresia.jp/products/whitebox/edgecore/

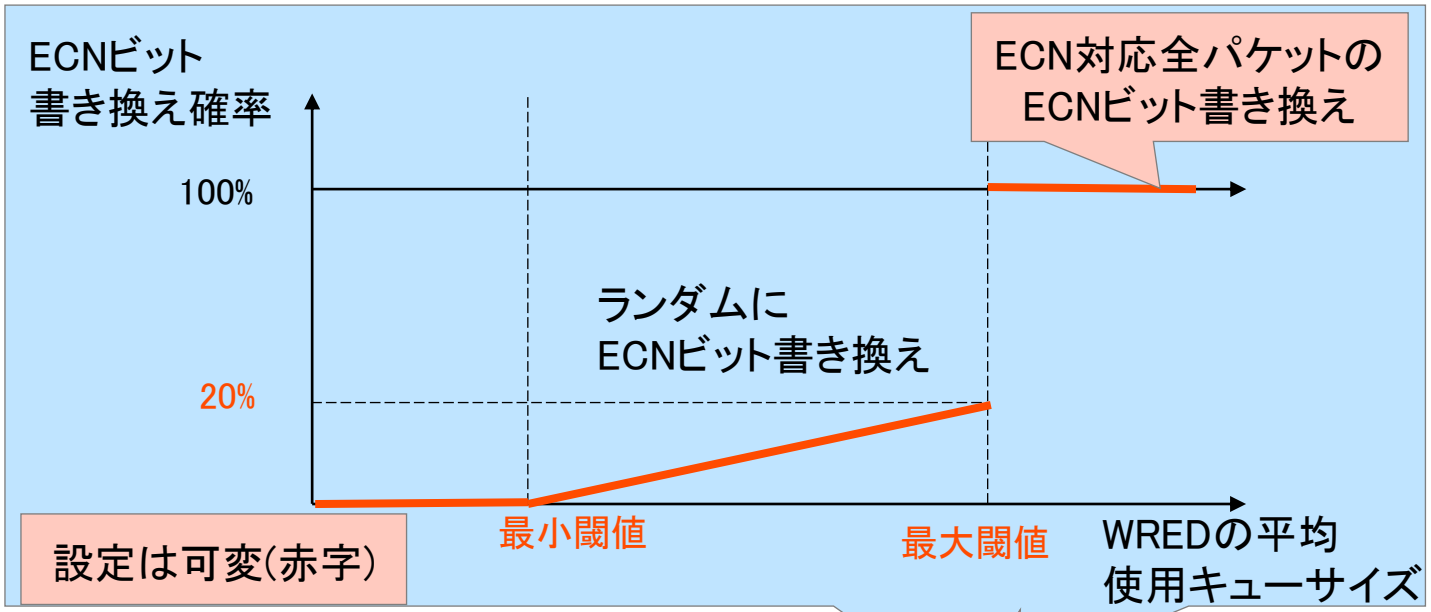
◆ポーズフレームを使ったリンクレベルでのキュー単位の輻輳制御機能

◆ロスレス対象のトラフィックをCoSまたはDSCPで分類し
対象のキューでPFCが動作することでパケットロスを防止



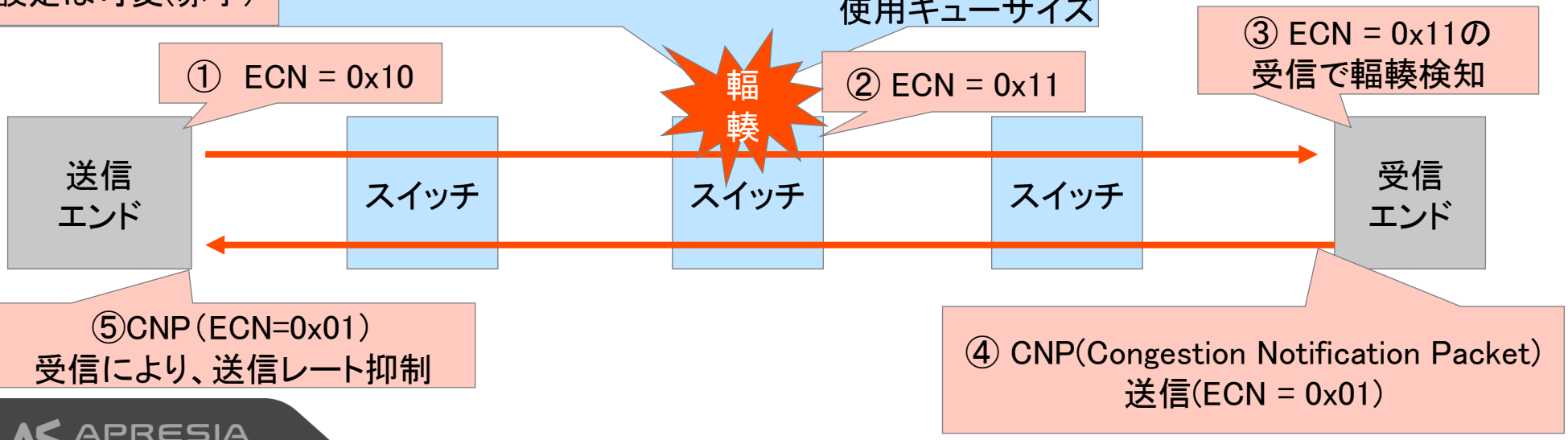
ECN（Explicit Congestion Notification）

◆送信側に輻輳を通知し、送信レートを抑制することでパケットロスを防止



ECNビット	動作
0x00	非ECN対応
0x10	ECN対応(0)
0x01	ECN対応(1)
0x11	輻輳発生

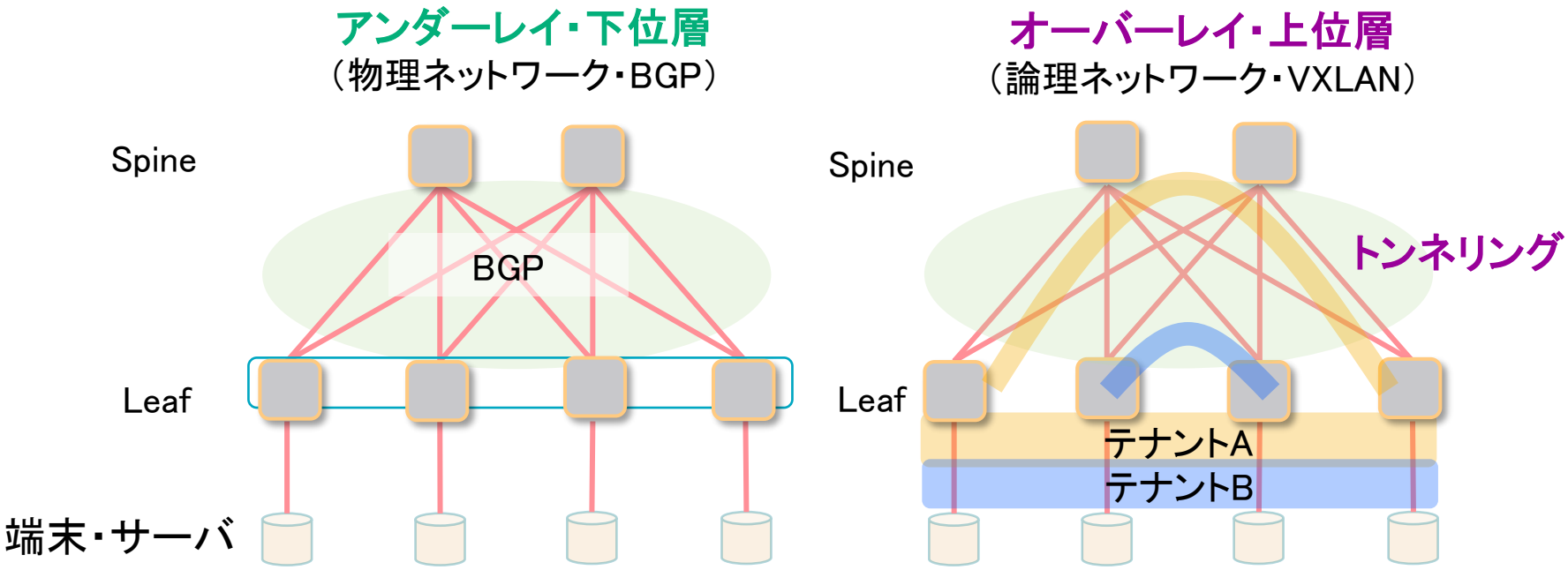
WRED
(Weighted Random Early Detection)



- プライオリティグループ毎にキューの優先順位を定義
- WRR(Weighted Round Robin)やSTRICT(絶対優先)でキューの優先制御

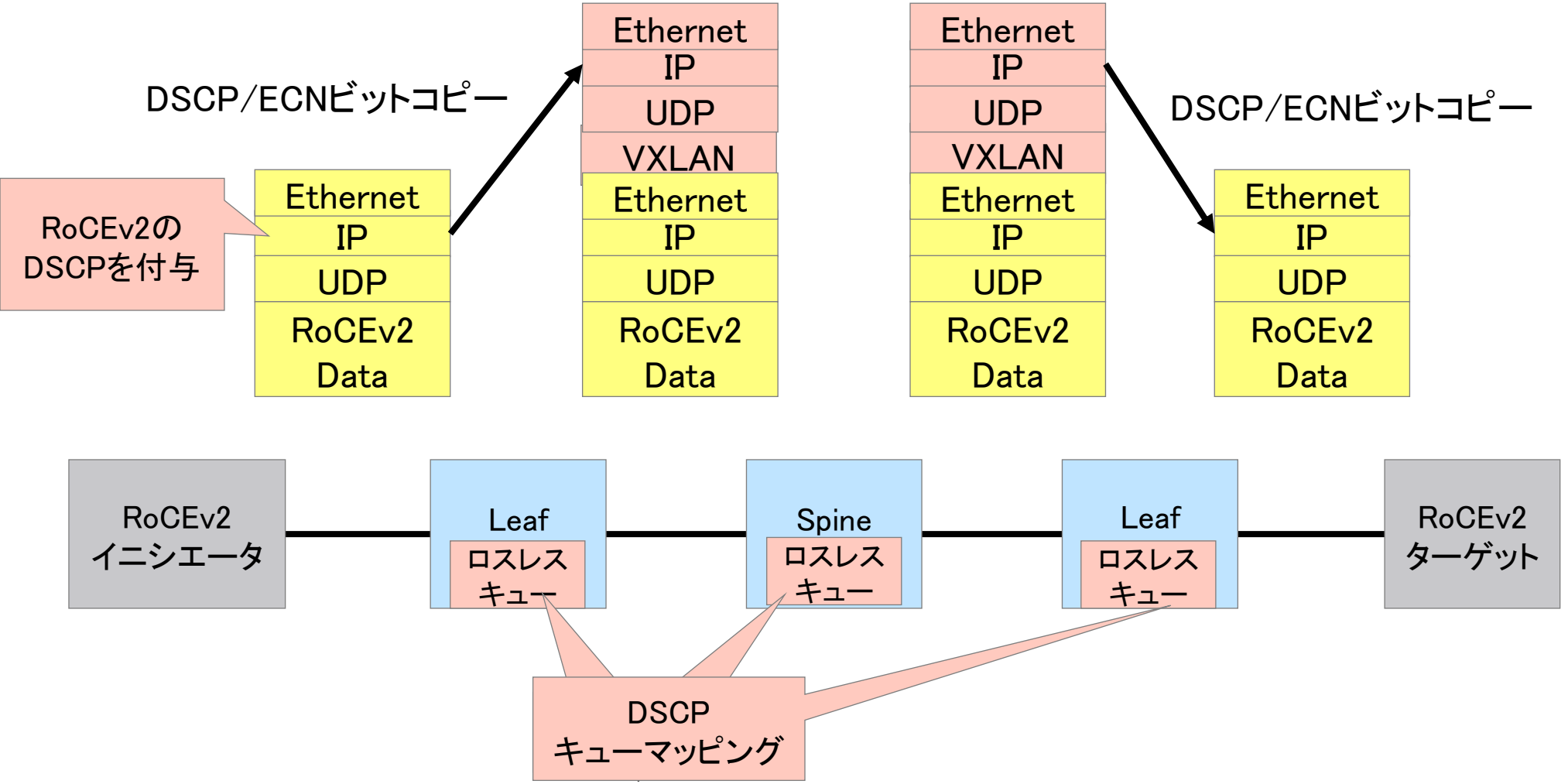
Traffic Class	キュー	優先制御
0	0	WRR:30%
1	1	
2	2	
3	3	WRR:70%
4	4	STRICT
5	5	
6	6	
7	7	

- L3ネットワークの上にオーバーレイネットワーク(L2VPN,L3VPN)を構築する技術
- マルチテナントに対応



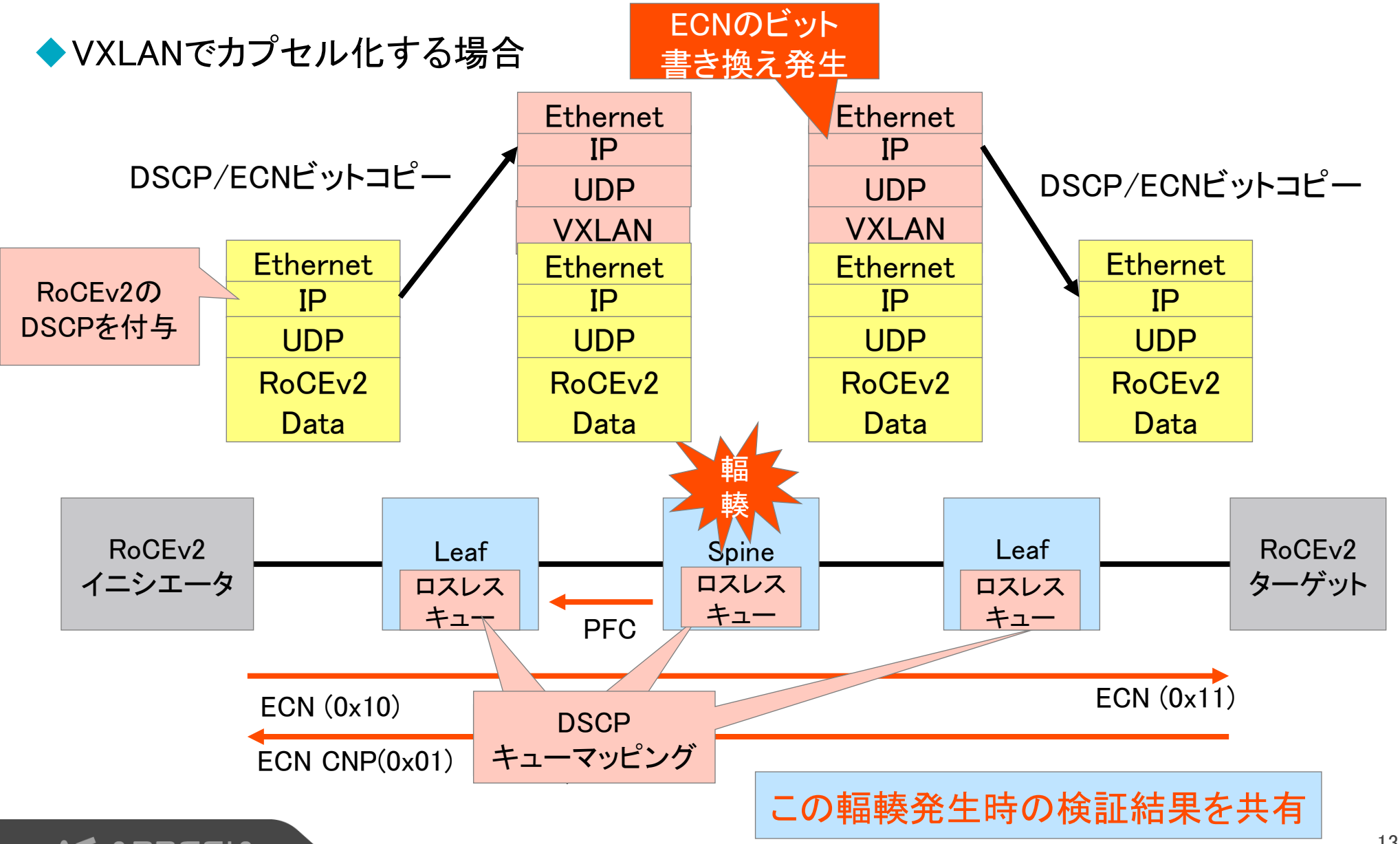
今回はEVPN/VXLANとPFC/ECN/ETSで
ロスレスイーサネットの動作を確認

◆VXLANでカプセル化する場合



EVPN/VXLAN併用時のロスレスイーサネットの動き

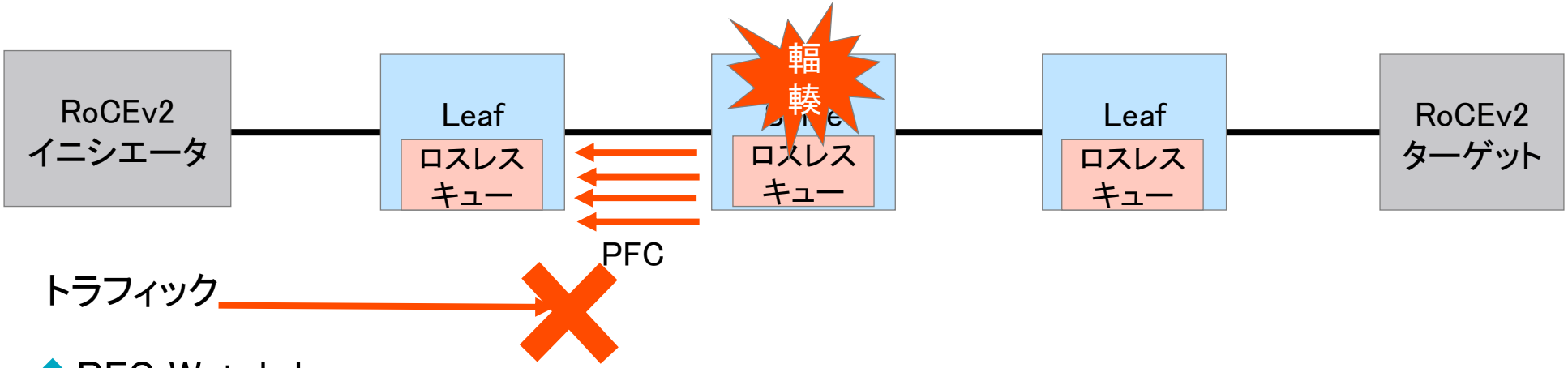
◆VXLANでカプセル化する場合



この輻輳発生時の検証結果を共有

◆PFC Storm

◇ 大量のPFCポーズフレームが送出される(PFC Storm)と
トラフィックの一部もしくは全てが流れなくなる



◆PFC Watchdog

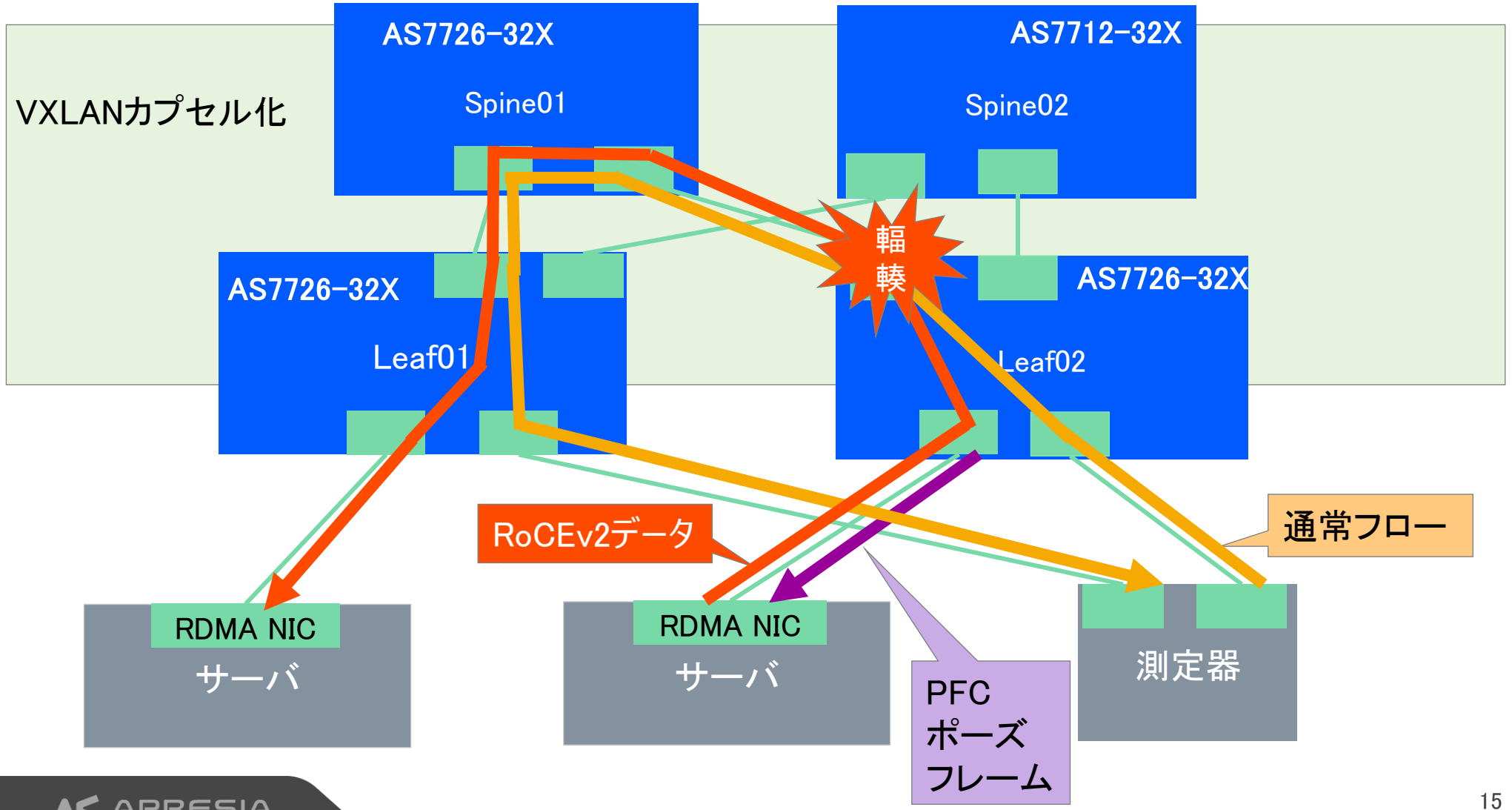
◇ PFC Stormを検知して緩和する機能

◇ Edgecore SONiCはBroadcom ASICの場合
緩和機能はforward(PFCポーズフレームを無視)を選択可能

https://github.com/sonic-net/SONiC/wiki/PFC-Watchdog

https://github.com/sonic-net/sonic-mgmt/tree/master/docs/testplan/pfcwd

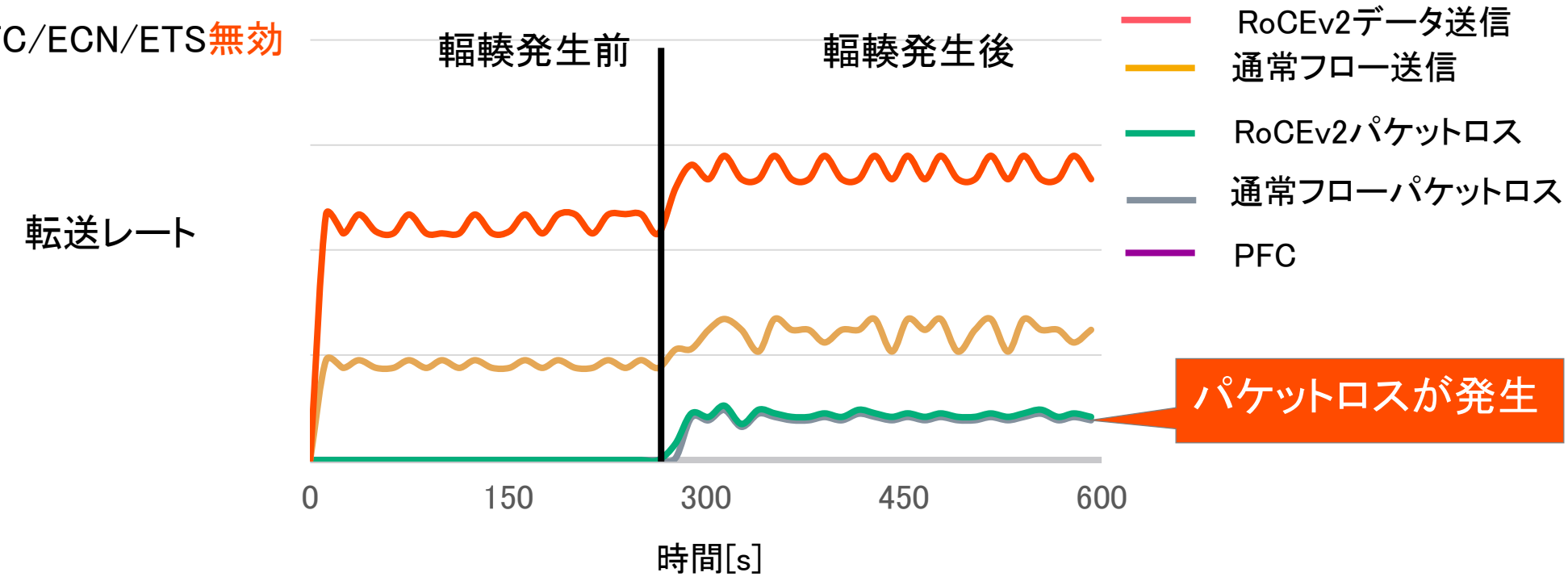
◆測定器から通常フロー、サーバからRoCEv2データを流した際の動作確認



トラフィック	Loss-less	DSCP	Priority Group	キュー	優先制御
通常フロー	×	0,63	0	0	WRR:30%
RoCEv2（データ）	○	26	3	3	WRR:70%
RoCEv2（CNP）	○	48	4	4	STRICT

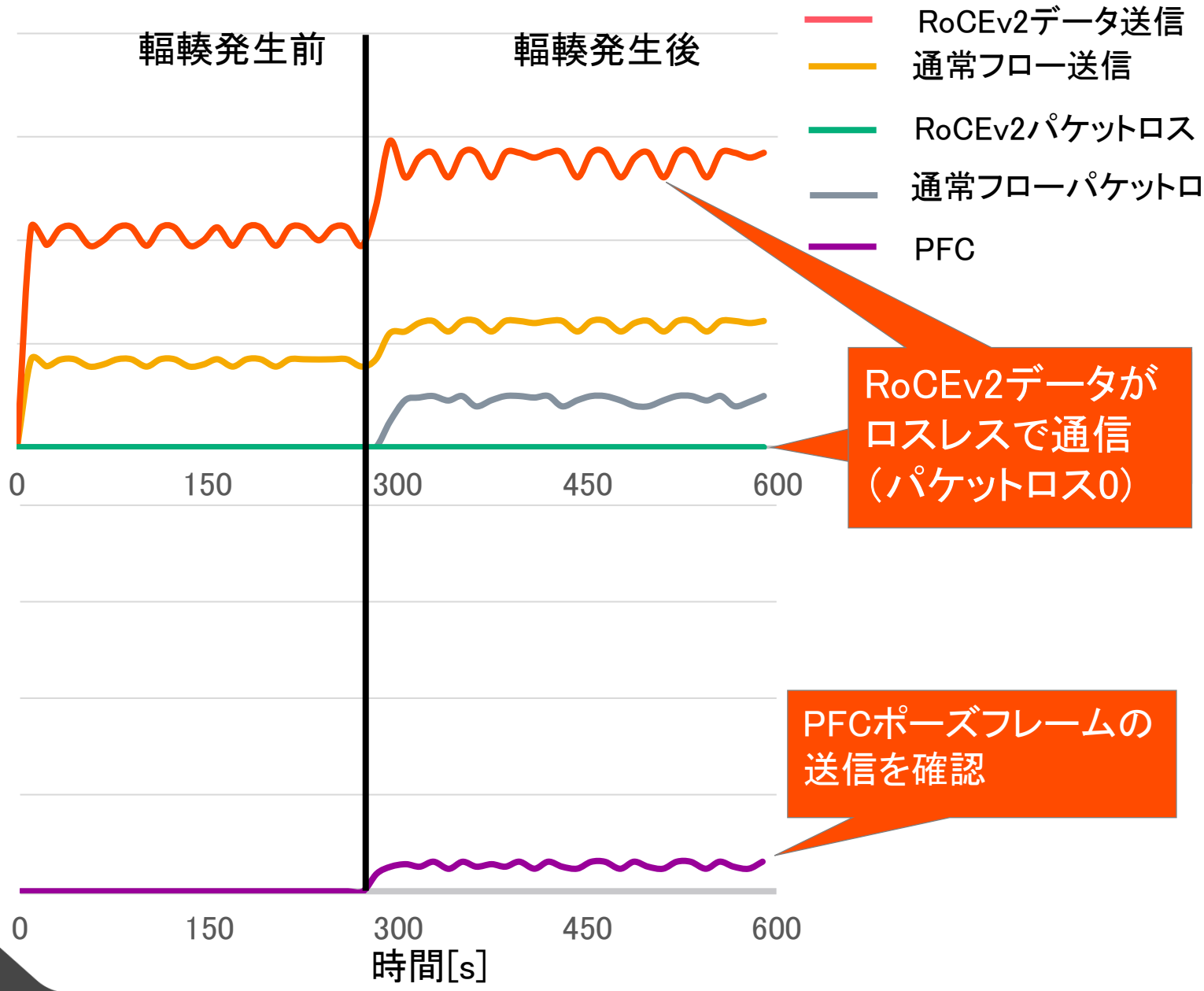
EVPN/VXLAN RoCEv2検証結果（ロスレスイーサネット無効）

PFC/ECN/ETS無効



EVPN/VXLAN RoCEv2検証結果（ロスレスイーサネット有効）

PFC/ECN/ETS有効



- ロスレスイーサネットの技術であるPFC/ECN/ETSの紹介
- PFC Watchdogの紹介
- SONiC+ホワイトボックススイッチで
  PFC/ECN/ETSとEVPN/VXLANとの併用で検証

- 今後の予定
  - EVPN/VXLAN Multihomingの検証
  - 400Gスイッチ(AS9726-32DB)の検証
  - PFC Watchdogの検証