JANOG50「できるのか？400G光伝送？」

明日やってくる400GbEにどう立ち向かうか

情報処理推進機構 産業サイバーセキュリティセンター サイバー技術研究室
ソフトイーサ株式会社
松本 智(Matsumoto Satoshi)

自己紹介

- **松本 智 (Matsumoto Satoshi)**
  - AS18127 AS59103 AS59120 AS63770
  - IPA産業サイバーセキュリティセンター・サイバー技術研究室
    ソフトイーサ株式会社
    - NW設計構築運用研究開発、サービス開発を行う
    - 最新技術調査・機器選定・検証等々L1～L7まで何でもやる
    - 徹底した自前主義・足りないものは何でも作る
  - 主な領域
    - DF・通信線路・光伝送・長距離光専用線・ xWDM ・MPLS-SR
      ・データセンターNWアーキテクチャ・BGPオペレーション
      ・ ASの構築運用・ルーティングセキュリティ等割と何でもやる
    - インターネットコミュニティーでの活動
      - InternetWeek：2014～、2017-2018副委員長、2019-2022委員長
      - OPENコンソーシアム
      - セキュリティ・キャンプ講師
      - つくば市教育研究所（インディペンデンスサーバーデイ）

400GbEをどうお迎えするか

- **10GbE/100GbEとは違う常識・技術・デバイス**
  - 同じノリで導入できるのか？
    - 光ケーブルを挿せば上がると本当に思っているのか？
  - ちゃんと”使う”にはどういう知識・ノウハウが必要か?
  - 正確な光デバイスに対する運用知識が必要
  - 単純な100GbEからのステップアップではない
  - 400GbEはすでに身近な技術・明日急にやってくるかもしれない
  - 400GbEの動作検証はどうしたらいいのか？
    - NWの高速化が進むほど動作確認は難しくなる
      - トラフィックを発生させること自体も難しい
    - 400Gbpsのトラフィックをどう発生させれば…
      - 無いなら作ればいい ⇨ 作りました

一般的なNW検証

- **NWの構築には必ず正常性確認が求められる**
  - L1～L7まですべてのレイヤーで様々な確認が必要
    - ルーティング、フォワーディングの動作確認
    - ACLやFW、IPS等の遮断動作、漏れがないか
    - Proxyの動作は正しいか、VPNはちゃんと使えるのか
    - そもそもHWがスペック通り動いているのか？
    - 対障害性の確認
      - 冗長系の動作（主にL1～L3）、冗長性を確保するプロトコルの動作確認
      - 電源断の試験、故障模擬
    - ログ監視系
      - ちゃんと監視できているか、ログは取れているのか
    - 運用上の確認
      - 定形オペレーションへの切り出しができているか
      - マニュアルの整備
      - 非定形オペレーションへの対処方法の確認
      - 障害発生時の動きのチェック

一般的なNW検証

- **現代では検証方法も含めて様々な知見が豊富**
  - わかりやすいバグや不具合があることはかなり減った
  - ベンダー間の相互接続性も向上しており定番のProtocolは枯れてきている傾向
    - 作る側使う側共にある程度ちゃんと動くだろうという期待を相互に持っている
  - 一般的な利用範囲ではほぼ想定通りの結果が得られやすい
    - 一方で定番から外れていくと想定通りという的から外れていく
    - 想定から外れる＝想定外 とはなにか
      - 通常想定されないような使い方や環境
      - 世の中に出たばかりの新しいテクノロジー
        - SRv6を代表とするような新たな大規模なProtocol
        - ホワイトボックススイッチやNOS等の新たな考え方の装置
        - 400G/800Gと言ったさらなる高速化の最先端を行くもの
      - このような分野は世の中に十分な知見が蓄積されておらず当然のようにバグや未知の不具合に遭遇する

バグや未知の不具合とはなにか

- **バグ**
  - RFCやDraft等規格どおりに動いていない
    - Statementがおかしくなるパターン
    - IPsecのような複雑なプロトコルにおける相互接続不具合
  - 装置内部での動作における不具合
    - macが学習されない、ARPがおかしい
    - 冗長系のProtocolが正しく動かない
    - show コマンドの結果と実際の動作が違う
- **未知の不具合**
  - 一見するとちゃんと動いているが高負荷をかけるとおかしくなるもの
    - MulticastPacketを100Gbpsくらい投げ込むとおかしくなる
    - 1000万経路学習させるとBGPがちゃんと動かなくなる
    - カタログスペック通りの最大ppsで負荷をかけた時のみ発生する事象
  - 複雑化したNWが起因して発生するもの
    - 仮想回線（VPLS/EVPN等）によるロングホールの存在
    - 冗長化に伴って発生する変な現象
  - 頭のおかしい（褒め言葉）使い方をすると遭遇できる

製品や技術が枯れていく過程

- **枯れるとはなにか**
  - 装置や規格が多くの環境で利用されバグや不具合が取り除かれた状況
    - みんなが使うことで鍛えられ、実用に耐えうる十分な品質が得られる
      - 多くの利用者がサポートに問い合わせることで製品が鍛えられるといいなぁ…
    - 多くの環境という観点から見ると単独での検証には必ず限界がある
      - 例えばShowNetのような環境を単独で維持することは困難
        - 無限のコストがあればできるかも知れないが、様々な最新機器を手に入れるというのは非現実的
        - 故に相互接続性の確認をするために持ち寄って確認する事に価値が出る
      - 広域NWにおける実証実験等を他組織で行うことも同様
  - メーカーの検証＋利用者実環境での不具合の洗い出しの両輪によって枯れていく
    - メーカーや業界にとっての初物は必ずバグや未知の不具合がある
    - 単に買ってきて使う、不具合があれば文句を言うという段階からの脱却が先進的な利用者には求められている

検証はどこまでするべきか

- **厳密な検証**
  - 測定器を使いチップの動作や限界性能を正しく測定する
    - ものすごく高価な測定器が必須
      - 例）高価かつ高精度なパケットジェネレータ
  - 製品の品質や性能を正確に測定することが求められる
- **目の前に見える動作を注視する検証**
  - 多くの人にとってのNW動作検証
    - 利用シーンに近い環境を作って確認を進めていく
    - ケーブルの挿抜や機器の停止起動など手を使って進めていく
    - 実際の業務フローを回してみて不具合がないかのチェックも大事
  - 必要なものさしの正確性はキログラム原器のように正確である必要がない
    - 事務作業等で紙を切るときに0.001mmの精度は必要かという話
      - 例）iperf等ソフトウェアで実現可能な精度のパケットジェネレータ
  - 利用シーンを再現できる検証シナリオを作れるかも重要
    - 測定ではなく利用に重きをおいた検証といえる

検証のためのものさし

- **NWの規模や環境・利用する技術（例えばEthernetの速度）に依存**
  - プロトコル解析やパケットキャプチャー
    - ソフトウェアがとても優秀でOSSも豊富、
    - ラップトップ環境でも実行でき、可搬性もよく使い勝手位が良い
  - 10Gbps程度のトラフィックを流す程度であれば iperfがあれば十分
    - ワイヤーレートを流すためには工夫が必要
    - ラップトップ環境での実現も難しくなる
  - 100Gbps/400Gbpsのトラフィックを流すととても難しい
    - 何らかのHW支援機能を利用しなければならない
    - HWに完全依存する測定器は移動が困難であり可搬性も高くなく
      高価であるので手軽に現場に投入することは難しい
      - 何らかの設備（ラボ）等に整備する必要があり気軽な導入は難しい
  - ソフトウェアで実現可能なものさしとHWでしか実現できないものさしの中間が無い
    - 100Gbps超の世界においてそのような中庸なものさしが存在しない
    - ならば作れば良いと言うのが発端

自作400Gbpsトラフィックジェネレータ

- **Over 100GbE のNWにおいて検証を行うため自作**
  - 高価な測定器とソフトウェアのトラフィックジェネレータの間を埋めたい
    - 我々の利用用途にジャストフィットするような性能と価格を目指す
  - 目指した目標
    - なるべく安く市中で手に入る部材で量産可能であること
      - 自作PCの範疇で実現したい、Xeonの採用は高価になるため回避
    - 単独 200Gbps 送受信、複数台の組み合わせで 400Gbps超を実現できること
      - 400G NIC がまだ市場に存在しないため
    - ワイヤーレートは出なくとも高ppsを実現できること
    - 大量の mac/IP 宛の通信を模倣できること
      - Single IP 100G stream から 1万IP 100G stream くらいまでの柔軟性
    - 可搬できること
      - 回線や広域NWの検証を行うことを想定したため
      - 台車あれば気軽に持っていけることが重要

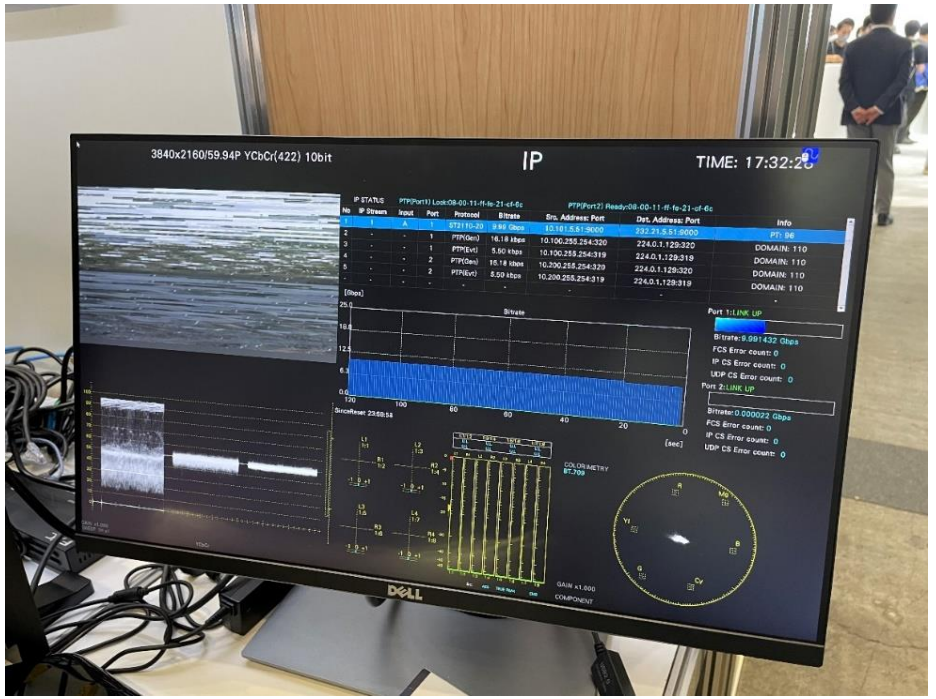
自作400Gbpsトラフィックジェネレータ

- **Spec**
  - Corei9 12900KS / 64GB RAM / M.2 256GB
    - 400GbE NIC対応のためPCIeGen5必須
  - NIC: Mellanox Connect X6 200GbE * 2
    - 将来的には Connect X7 400GbE * 1
  - Ubuntu 22.04 LTS / DPDK 21.11.1
- **性能(PC 1台あたり)**
  - Max 200Gbps / 220M pps(64byte)
    - 100Gbps の単一stream生成
  - 送信機受信機を分離できPCを増やすと 400Gbps/800Gbps のトラフィックを生成可能
- **価格**
  - 1台あたり約40万くらい（NICは中古）
    - 新品でも ~100万以内

実証実験への投入

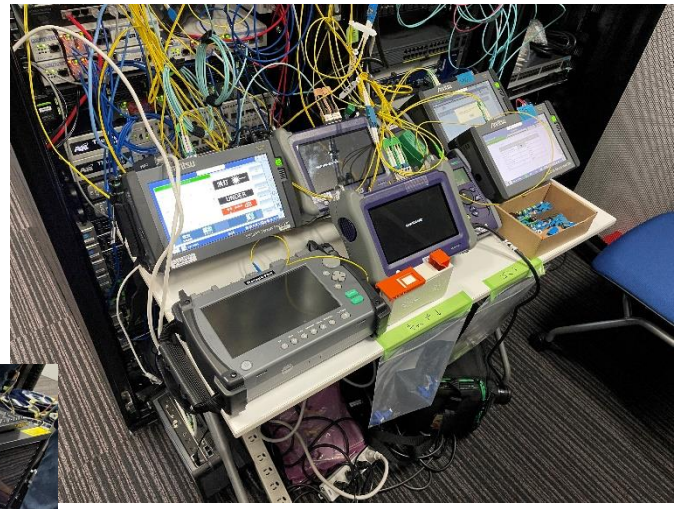
- **Interop Tokyo 2022 400GbE 超広帯域トラフィック伝送実験 2022/6/13-17**
  - 光伝送路に光ホワイトボックス伝送装置を用いた長距離伝送を利用
  - 一部NW構成にLAGを意図的に挿入しLAG IF間に於けるトラフィック分散等を計測
  - 映像伝送と伝送区間併用時における、映像伝送側への攻撃の実現
    - 本装置を使いある特定のトラフィックを流すことで、異なるVLAN上を流れる映像伝送に対して影響を耐えて映像を乱す攻撃デモを実施

DEMO展示として攻撃トラフィック発射ボタンを自作
押すと回線を共有している映像を乱す



実証実験への投入

- **線路長100kmの400G-ZR回線を用いたトラフィック伝送実験 2022/7/6**
  - 400G SW、WDM装置、トランシーバ等を皆で持ち寄って実施
  - DF+ダミーファイバーを用いて100kmの400G-ZR回線を自作
    - トラフィックを印加しながら光のチューニングを実施
    - 400Gbpsのトラフィックを伝送しエラー等計測を行った
      - エラーなし、安定した伝送を実現
    - 400G SWに様々な特殊なトラフィックを印加し動作確認
    - 最終的に無事安定した回線を実現



まとめ

- **何をどこまで考えるかという見極めが重要**
  - NWの動作検証において何をどこまで確認をするのかという見極めが最も重要
    - NWの設計にとどまらずHWに関する知識や市場動向も抑える必要がある
      - 特に昨今の高速Ethernet界隈ではチップのロードマップや半導体の動向からチップ性能限界を把握しておくことが重要化している
      - どのような規格やスペックの製品が何時登場するのかという将来を見据えたNW設計をベースに検証項目を立案していく
        - 上限性能をどこまで想定できるかがNWシステムの将来性を決める
  - 必要なものさしを自分で作るという気概
    - ものさしを作ることは一見車輪の再発明に見えるが手に合うものさしを作ることは様々な学びがありエンジニアの鍛錬にもってこいである