IX相互接続実証実験を通じて見えてきた400G導入で「変わること」「変わらないこと」

IIJ/AS2497編

Internet Initiative Japan

2022年7月19日

IIJ/AS2497 takez

竹﨑 友哉 Tomoya “takez” Takezaki(takez@iij.ad.jp)

- **経歴**
  - 2020年　　IIJ新卒入社
  - 2021年~　　AS2497の構築・運用
    IP Backbone & Peering

- **出身**
  - 福岡県福岡市

- **好きなもの**
  - マンホール
  - 電話局・公衆電話
  - 電信柱・示名条片・架空
  - 海底ケーブル・ケーブルシップ

IIJ目線での400G導入のモチベーション

- **はじまり**
  - ルータ買ったら400Gポートついてきた。
- **モチベーションと期待**
  - ポート単価
    - ポートは100G/400G Combだが…
    - 300Gの損失(ライセンス次第でもあるが)
  - 100G x Nから400Gアグリゲーションによる効率化
    - コスト減
    - ケーブリングの効率化
    - 運用負荷の軽減
      - LACP・ECMP・ロードバランシング
    - ~~節電(雀の涙程度)~~
  - コヒーレントオプティクスへの期待
    - 長距離規格(400G-ZR/400G-ZR+)への期待
    - 既存DWDMシステムとのシームレスな融合・高い親和性?
    - IP over DWDM

100Gと400Gで変わること

400GbEではFECによるエラー訂正が必須

- PAM-4変調の採用により2bit/baudの伝送が可能
- しかし、アイ開口がNRZと比較し1/3となりSNRも悪くなるためエラーフリーの伝送は困難に

FEC(Forward Error Correction)/前方誤り訂正

- 今までOptionalであったFECによるエラー訂正を400Gでは必須とし、伝送品質を担保。
- FEC frame=Codeword
  - 15 Symbol errorsまでは訂正可能(Correctable errors)
  - 16 Symbol errors以上は検知は可能、訂正は不可能(Uncorrectable errors)

➢ **100Gから400Gの大きなChangeとも言えるFEC必須は運用者視点でどれほどのChangeなのか？**

エラーはinput errorsとしてカウントアップ

- Uncorrected Codeword Countとinput errorsは必ずしも連動しない
  - IOS-XRではinput dropsとInput error symbolとしてカウントアップ
    - Uncorrected Codeword Countをクリアできない

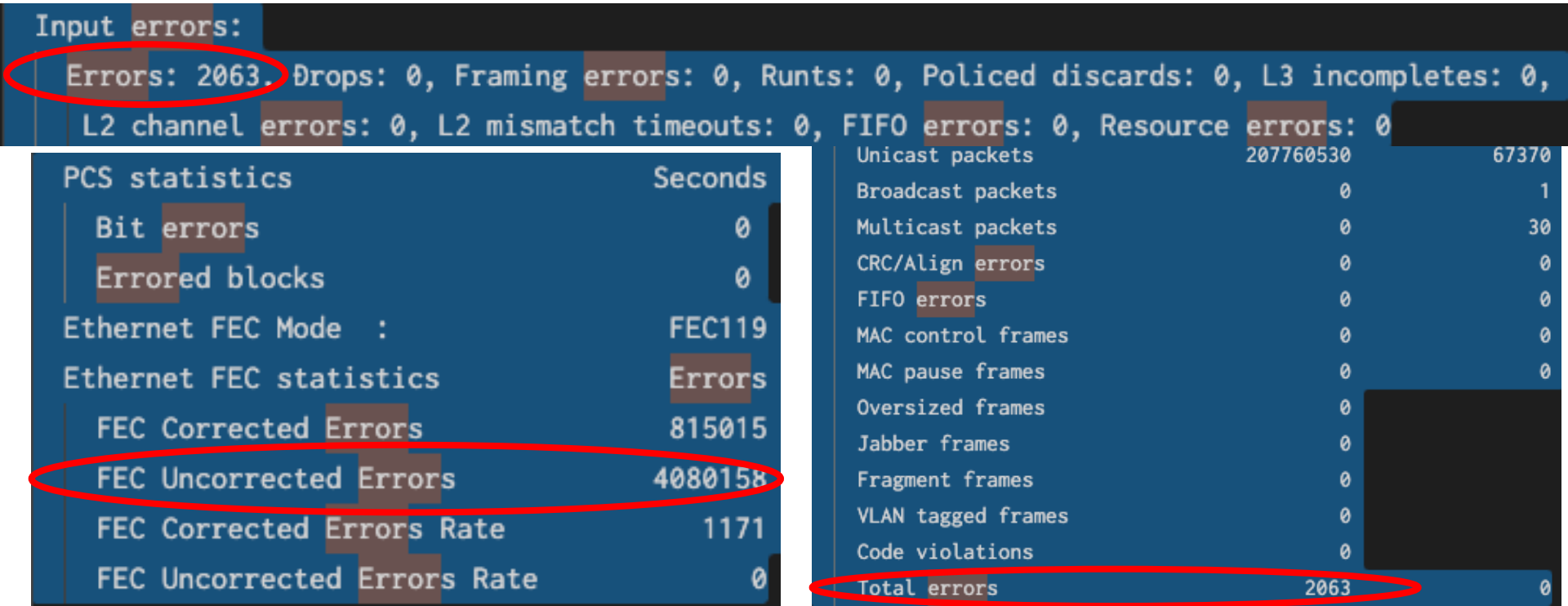
```
Statistics:
    FEC:
        Corrected Codeword Count: 20595101925
        Uncorrected Codeword Count: 1862858
```

```
Input error giant                = 0
Input error runt                 = 0
Input error jabbers              = 0
Input error fragments            = 0
Input error CRC                  = 0
Input error collisions           = 0
Input error symbol               = 484602
Input error other                = 0
```

```
5 minute input rate 166553894000 bits/sec, 13884388 p
5 minute output rate 166709084000 bits/sec, 13897368
  6883602717 packets input, 10325404075500 bytes, 484602 total input drops
  0 drops for unrecognized upper-level protocol
  Received 0 broadcast packets, 0 multicast packets
           0 runts, 0 giants, 0 throttles, 0 parity
  484602 input errors, 0 CRC, 0 frame, 0 overrun, 0 ignored, 0 abort
```

SNMPなどのカウンタはもちろんInput errorsの値を出力

- 単発の16 symbol error(Uncorrectable Codewords)を挿入してもInput errorsが必ずしも上昇するわけではないことを今回の検証で確認
- もちろん継続的な16 symbol errorであれば100Gと変わらずカウントは上昇する
- 下図はJunosのI/Fエラー出力



FEC Uncorrected codewordsで生じるinput errorsの内訳はベンダーにより様々

- ベンダーの仕様やエラー解釈が異なるため
- EOSではrunts, input errors, CRC, symbol, input discardsが上昇した

```
5 minutes input rate 69.2 Gbps (17.5% with framing overhead), 5767198 packets/sec
5 minutes output rate 69.2 Gbps (17.5% with framing overhead), 5767195 packets/sec
    1924452425 packets input, 2886802725284 bytes
    Received 0 broadcasts, 0 multicast
    22 runts, 0 giants
    312174 input errors, 156076 CRC, 0 alignment, 156076 symbol, 9475 input discards
    0 PAUSE input
    1924451353 packets output, 2886679055126 bytes
    Sent 14 broadcasts, 45 multicast
    0 output errors, 0 collisions
    0 late collision, 0 deferred, 0 output discards
    0 PAUSE output
```

```
 FEC corrected codewords      20381271              2833              0:00:03 ago
 FEC uncorrected codewords    20038                   40              0:00:03 ago
```

ベンダーにより実装、カウントは様々

- 解釈・仕様の違いであり良し悪しはない

エラーカウンタの差異

- CodewordはあくまでもFEC frameでありEthernet frameではないと考察
- 今までのエラー監視で検知は可能

NOS	CLIでのFEC Corrected/Uncorrected確認	Uncorrectable codewordsによるinput errors増加	Input errorsの内訳	ifMIBで増加するカウンタ
IOS-XR	○	○ (Uncorrectable codewords > input errors)	input drops, symbol	ifInDiscards
Junos	○	○ (Uncorrectable codewords > input errors)	なし	ifInErrors
EOS	○	○ (Uncorrectable codewords < input errors)	runts, CRC, symbol, discards	ifInDiscards, ifInErrors

思ったほど大きく変わらない…？

FEC関連のshowコマンド

- CLIではCorrected or Uncorrectedしか確認できない…
  - つまり、16 Symbol errors以上か未満か
  - FECエラーマージンの監視は機器だけでは不可

```
RP/0/RP0/CPU0:c8202#show controllers optics 0/0/0/48 | i FEC
 FEC State: FEC ENABLED
RP/0/RP0/CPU0:c8202#show controllers fourHundredGigE0/0/0/48 all | utility egrep -2 FEC
        Receive Power (dBm):      13.000    10.000   -21.549   -23.010
    Statistics:
        FEC:
            Corrected Codeword Count: 3486763984
            Uncorrected Codeword Count: 1209
```

```
takez@ptx10001> show interfaces extensive et-0/1/0 | mat "fec "
  Ethernet FEC Mode  :                 FEC119
  Ethernet FEC statistics              Errors
    FEC Corrected Errors              1747347
    FEC Uncorrected Errors                  0
    FEC Corrected Errors Rate               2
    FEC Uncorrected Errors Rate             0
  Optic FEC Mode  :                    CFEC
```

```
knd-test-ars1#show interfaces ethernet 18/1 phy detail
-snip-
  FEC corrected codewords     20381271          2833          0:00:03 ago
  FEC uncorrected codewords   20038               40          0:00:03 ago
  FEC corrected symbol rate   1.12E-04*
-snip-
```

1. 商用DF環境でのループバックトラフィック印加試験

- 環境
  - 大手町のとあるDF(ハードループ時3.46km,減衰6.6dB)
  - 複数ベンダー、規格のオプティクス(FR4,LR4-10)

- 試験内容
  - 規格範囲内(LR4-10)外(FR4)でのリンクアップ可否
  - トラフィック印加試験でのエラーの有無を検証する

- 試験結果
  - 400GBASE-LR4-10は規格内の試験であるため問題なくリンクアップ
  - 400GBASE-FR4は規格外の試験であったため、用意したベンダーのうち一部はリンクアップせず、一部はリンクアップ

7~8分ほどトラフィックを印加したSymbol error数

- 伝送品質・規格の影響がここまで大きく見える

FR4

Number of Symbols	Corr. A+B CW Errors	Corr. A+B CW Error %
1	863,441,394	98.3449000
2	13,897,227	1.5828800
3	569,332	0.0648462
4	54,482	0.0062054
5	8,220	0.0009362
6	1,620	0.0001845
7	365	0.0000416
8	77	0.0000088
9	27	0.0000031
10	6	0.0000007
11	0	0.0000000
12	0	0.0000000
13	0	0.0000000
14	0	0.0000000
15	0	0.0000000

LR4-10

Number of Symbols	Corr. A+B CW Errors	Corr. A+B CW Error %
1	17,609,634	99.7131000
2	49,868	0.2823740
3	770	0.0043601
4	20	0.0001132
5	1	0.0000057
6	1	0.0000057
7	0	0.0000000
8	0	0.0000000
9	0	0.0000000
10	0	0.0000000
11	0	0.0000000
12	0	0.0000000
13	0	0.0000000
14	0	0.0000000
15	0	0.0000000

上がらなかったFR4では

- 一部Laneの受光値がLow warning
- クロック偏差が-65,003.2(Too Low)
- Degraded SER点灯
- Loss of alignment marker payload sequence
- Loss of alignment
- Uncorrected codewordが上昇

※規格(2km)を超えた試験であることは留意

もちろんDC内やフロア内であれば容易にリンクアップ

- 距離の短い別の大手町のとあるDF(ハードループ時1.78km, 減衰7.2dB)では問題なくリンクアップ
  - 距離は短いが、減衰はこちらの方が大きい

2. 商用トラフィック導通試験

- 環境
  - 大手町のとあるDF(1.73km,減衰3.3dB)
  - IIJ側400GBASE-FR4<>JPNAP側400GBASE-LR4-10

- 試験内容
  - 実利用を想定する環境での商用トラフィック試験

- 試験結果
  - 400GBASE-FR4<>400GBASE-LR4-10混在環境でも問題なく利用可能

➢ IIJではユースケースに応じた使い分けを検討
  ➢ 単価、バージョン対応可否、距離、品質…etc

まとめ&課題

- **通常運用する中で大きく変わることはなさそう**
  - 機器へインストール、接続、初期確認
    - 普段やることと変わらない
  - なにかあればほぼInput Errorsが乗る
- **フロントは少しだけ長く**
  - 皆様お気をつけください
  - QSFP-DD800を見据えるとフロントの余裕は計画的に
- **対応機器も増え転換期？**
  - 10G -> 100Gに比べ100G -> 400Gで増加率は見劣りするが100Gと400Gを備えた製品が多く出ている
  - コモディティ化が進むにつれトランシーバコストも直に下がる(はず)
- **7月に堂島ビル間(2ビル~3ビル)で実網利用開始予定(FR4)**

- **伝送品質管理**
  - FEC Correctable/Uncorrectable監視
    - (ベンダさん)SNMPは非対応です。Telemetryなら
  - FECエラーマージン
    - 運用前に気づければよいが、S-in後は…
  - Degraded-SERどうする(後述)
  - そもそも100G/400G関係なく伝送品質管理って皆様どうされていますか？
- **Breakout**
  - 4X100G-LR or 2X100G-LR4 or 100G-LR4
- **細かいところの400G対応はベンダと協力しながら**
  - 商用利用可能な水準
  - showコマンドの一部結果不備や挙動不審な部分は多かれ少なかれ見られる
    - 今回のように事前検証は大事
    - そしてそれをフィードバックすることも言わずもがな大事

JANOG50後追記

利用したDFのリターンロスについて

- 大手町のとあるDF(ハードループ時3.46km,減衰6.6dB)
  - 1芯目14.85dB、2芯目15.08dB(1310nm)
- 距離の短い別の大手町のとあるDF(ハードループ時1.78km,減衰7.2dB)
  - 1芯目19.87dB、2芯目19.58dB(1310nm)
- 片芯ずつでの測定、FR4のリンクアップに差異があったDFの方が反射減衰量が低かった

FECエラーレートについてトラフィックのレートと時間について

- トラフィックレートはワイヤーレートで7~8分ほど

Symbolエラーの増加が継続的もしくは突発的であったか

- 継続的に増えており、突発的に増えるような減少は見受けられなかった。

日本のインターネットは1992年、IIJとともにはじまりました。以来、IIJグループはネットワーク社会の基盤をつくり、技術力でその発展を支えてきました。インターネットの未来を想い、新たなイノベーションに挑戦し続けていく。それは、つねに先駆者としてインターネットの可能性を切り拓いてきたIIJの、これからも変わることのない姿勢です。IIJの真ん中のIはイニシアティブ――IIJはいつもはじまりであり、未来です。