安定した800G DCNWを求めて
~トランシーバのBER計測~

ソフトバンク株式会社
内田 泰広

自己紹介

内田 泰広 (Yasuhiro Uchida)

2024/10~ ソフトバンク株式会社入社
ネットワークエンジニア

担当業務:

- AI計算基盤のNW設計・構築
- GPUサーバの構築・検証
- スイッチの最新技術調査・検証
- ラックスケールGPUサーバ

「NVIDIA GB200 NVL72」を搭載したAI計算基盤が稼働開始
https://www.softbank.jp/corp/news/press/sbkk/2025/20251225_01/

リンク障害で困っている事

BER: Bit Error Rate(ビット誤り率)
数値が低いほうが品質が良い

リンクフラップ問題

- **AI計算基盤で数万本以上トランシーバを利用**
  - 毎週トランシーバの交換・リシート対応を実施
  - 光レベルだけでは切り分け出来ないトラブル
  - 同時多発的にトラブルが発生

⇒ **稼働工数** が大きい、インフラの品質を上げたい

性能出ない問題

- **なぜか特定のリンクのみで性能が出ない**
  - 別のリンクで再計測すると性能が変わる
  - BERが高くエラーパケットが多発していたことが判明しトランシーバ交換で解消
  - 構築時にBERの高いトランシーバは交換している

⇛ リンクアップしたままのトラブルだと根本原因に気付きにくい

安定したトランシーバ選定が必要、自分たちでトランシーバの性能を確認したい

多種多様な800Gトランシーバ

光トランシーバ + 光ケーブル

DR8	SMF MPO16 * 1	DR8
SR8(VR8)	MMF MPO16 * 1	SR8(VR8)
2xFR4	SMF LC-LC * 2	2xFR4
2xFR4 Base	SMF LC-LC * 2	2xFR4 Base
2xFR4 Lite	SMF LC-LC * 2	2xFR4 Lite
2xDR4	SMF MPO12 * 2	2xDR4
2xDR4+	SMF MPO12 * 2	2xDR4+
2xDR4++	SMF MPO12 * 2	2xDR4++
2xSR4(VR4)	MMF MPO12 * 2	2xSR4(VR4)

DR8 TRO/LRO	SMF MPO16 * 1	DR8 TRO/LRO
FR8 TRO/LRO	SMF LC-LC * 2	FR8 TRO/LRO

MCF	MCF * 1	MCF

2xDR4 LPO	SMF MPO16 * 1	2xDR4 LPO
2xFR4 LPO	SMF LC-LC * 2	2xFR4 LPO

2xDR4 OSP	SMF MPO12 * 2	2xDR4 OSP

Packaged Solution

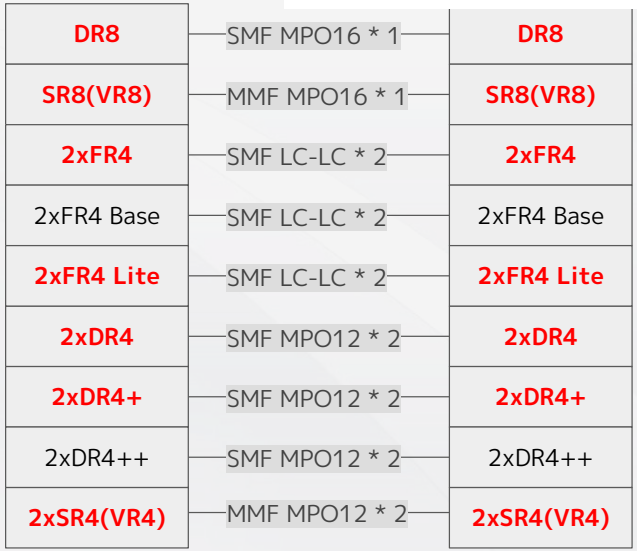
NPO	NPO
CPO	CPO
CPC	CPC

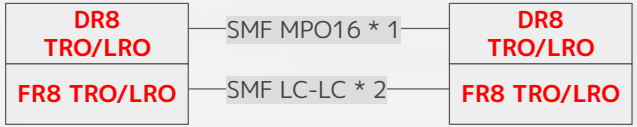
ケーブルレス一体型

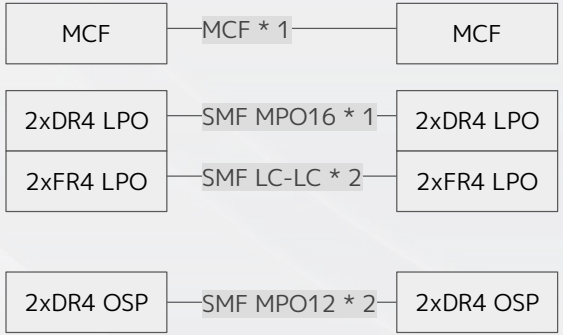
Active DAC	Active DAC
AOC	AOC
AEC	AEC
LED	LED
RF	RF

計測対象のトランシーバ

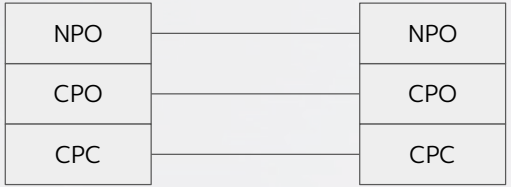
複数の800Gトラシーバを確保することは大変
手元に集めれた全トランシーバでBER性能を計測



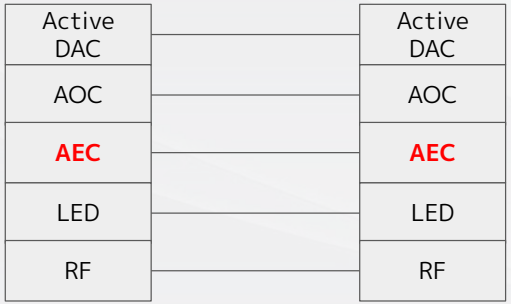




Packaged Solution



ケーブルー体型



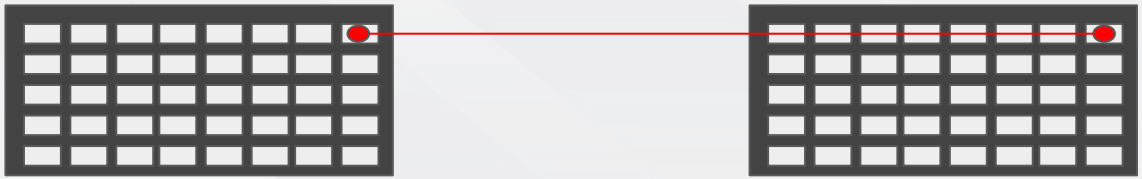
BERの計測方法

1. 2台のスイッチの同じポートにトランシーバを取り付け
   a. スイッチのポート毎に多少のBERの品質が変わる問題を回避

2. 一定時間経過後、SWのコマンドでBERを計測
   a. 実際のトラフィックを処理するのはスイッチの為、スイッチで計測

3. 計測が終わったらトランシーバを取り外し、1から繰り返す

出来るだけApple to appleとなるような計測



BERの測定結果

E-XXのXX部分が大きい数字だと性能が良い

	ベンダ	Optics	ファイバ	ケーブル長[m]	SW1 BER	SW2 BER
①	A社	AEC	-	3	**6.00E-12**	1.00E-11
	B社	2SR4	MMF 2xMPO12	5	5.00E-10	6.00E-10
	B社	SR8	MMF MPO16	5	2.00E-09	1.00E-09
	B社	2DR4	SMF 2xMPO12	15	5.00E-10	1.00E-10
	B社	DR8	SMF MPO16	5	2.00E-08	2.00E-09
	B社	DR8+	SMF MPO16	5	5.00E-09	2.00E-09
②	B社	DR8 TRO/LRO	SMF MPO16	5	1.00E-08	1.00E-08
	B社	2FR4	SMF 2xLCLC	10	5.00E-10	2.00E-09
②	B社	2FR4 TRO/LRO※1	SMF 2xLCLC	10	9.00E-09	**1.00E-06**
	B社	2FR Lite	SMF 2xMPO12	10	**計測予定**	**計測予定**
③	C社	2SR4	MMF 2xMPO12	5	2.00E-08	8.00E-09
	C社	2SR4	MMF 2xMPO12	40	3.00E-09	3.00E-09

① AECケーブルがBERの比率が良かった
② TRO/LROはトラシーバによって BERの品質低下が見られた
③ 光ケーブルはケーブル長によって BERによって大きさは無かった

※1 片方のリンクのみUP = 400G
BERの計測はリンクアップした方のFのみで計測
他社スイッチでは問題なくリンクアップ

安定したDCNWを作るために

- **トランシーバ毎に BERに性能差がある**
  - TRO/LROは製品によって性能劣化があるため慎重な選定が必要
- **トランシーバの選定において BERも重要な指標の一つ**
  - 消費電力、ケーブル（取り回しも重要）、価格、サポート、**BER**
- **BERを取得できないスイッチ**
  - より深い性能切り分けとトランシーバ切り分けが出来ない
  - **スイッチの選定条件にも関わる**
- **最新トランシーバで注目している技術をどのように見ていますか？**
- **安定したデータセンタネットワークの為に取り組んでいる事はありますか？**
  - **製品選定、オペレーション・監視、運用ナレッジ**

SoftBank
for Biz

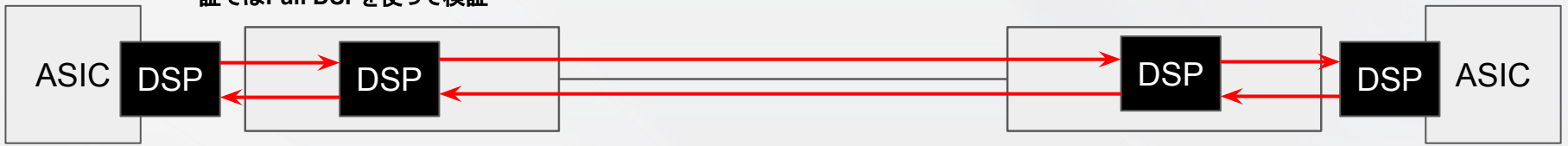
Appendix

リンク障害の物理原因箇所

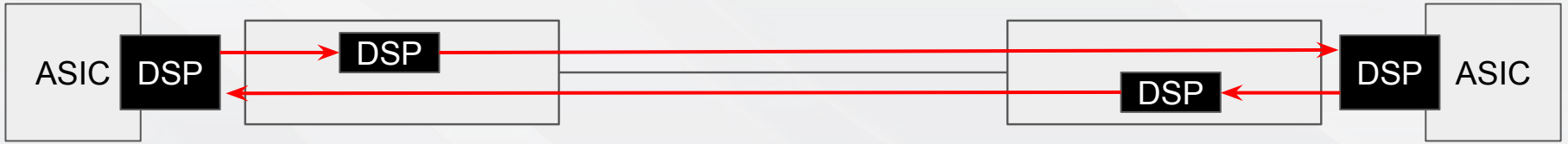
- SERDES(PCB+NIC)の問題
  - Lane故障
- トランシーバの問題
  - TOSA/ROSA、LD/PDなど
  - DSP/MCUの不具合・故障
- ケーブルの問題
  - コネクタの汚れ
  - ファイバの折れ
  - パッチパネル
  - リンクバジェット
  - コネクタが設置していない
  - 極性

TRO/LRO(Full DSP/Half DSP)、LPO

TRO Full DSP: 受信側だけDSPで行っている処理を無効にする。本検証ではFull DSPを使って検証



TRO Half DSP: 送信側だけDSPを経由する、受信側はDSPは通らない



LPO: トランシーバにDSPは存在しない、ASIC側で信号の補正を実施