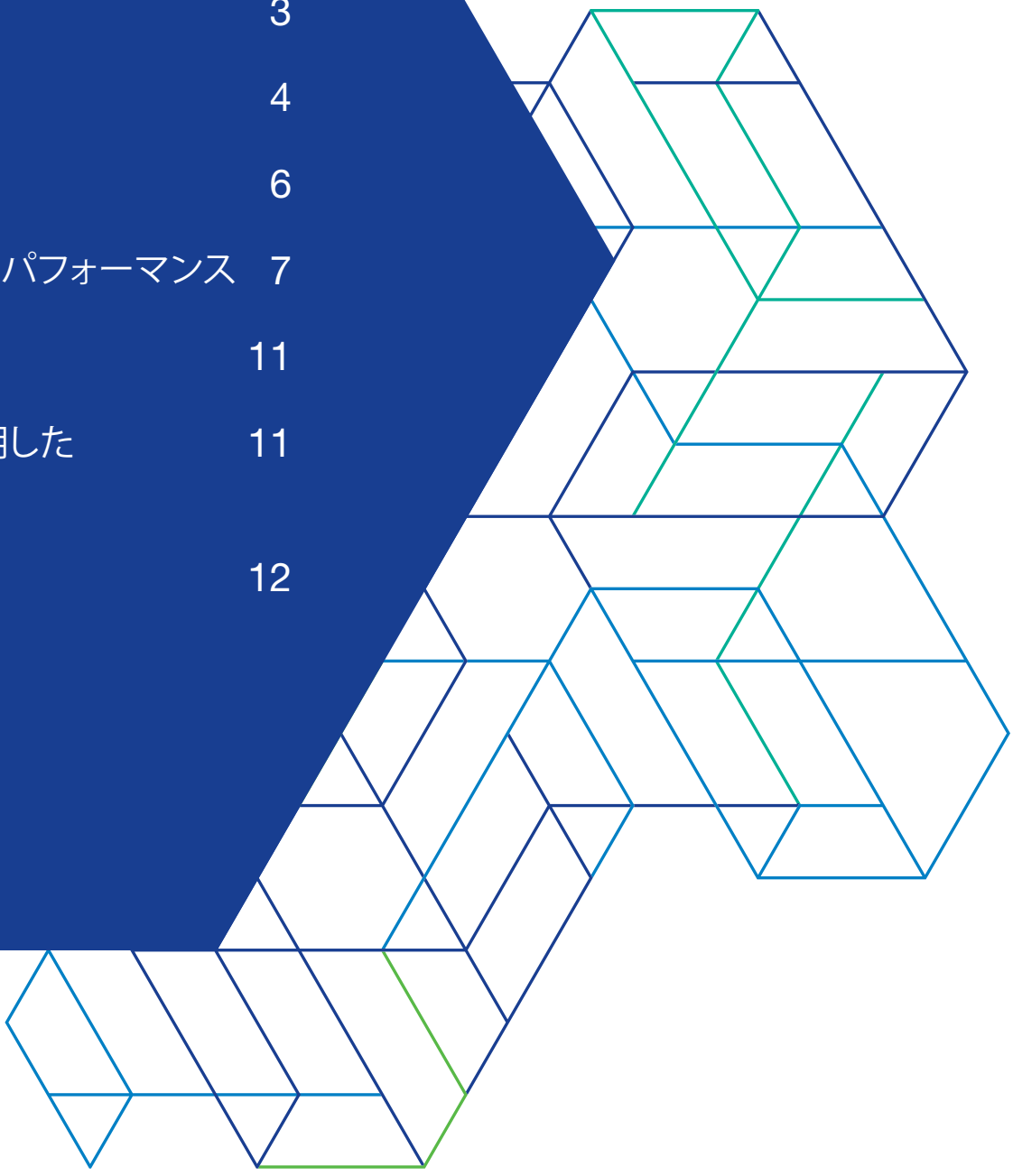
構造化配線が AI ネットワークのパフォーマンスに与える影響

PANDUIT™

ホワイトペーパー

目次



トレーニングや推論を行う最先端の人工知能 (AI) や機械学習 (ML) システムでは、非常に高い帯域幅、低い (テール) レイテンシ、数多くのアクセラレータ (GPU、TPU、その他の種類) を相互に接続できるファブリックトポロジーが求められます。こうしたシステムでは、InfiniBand (IB) リンクなど、バックエンドネットワークと呼ばれる、特化した光ネットワークや、ネットワークトラフィックを最適化する拡張機能を備えたイーサネットが使用されています。イーサネットのアップグレードには、IEEE タスクフォースで既に開発されたものもあれば、Ultra Ethernet Consortium で現在開発中のものもあります。

AI のトレーニングには、逐次計算と、次に進む前に完了しなければならない通信フェーズが含まれるため、そのパフォーマンスは、基盤となるネットワーク物理層のパフォーマンスに大きく依存します。テールレイテンシはこの計算シーケンス内の最も遅いメッセージを送信する時間によって決定されるもので、トレーニング効率に大きく影響します。通信の遅延によって GPU 時間の最大 50% がアイドル状態となる可能性があることが示されています。

通信ネットワークは、電気的リンクと光リンクの双方によって構成されます。電気的リンク (NVIDIA 固有の NVLink など) はノード内の GPU 間を接続し、非常に高い帯域幅と低レイテンシの通信を提供します。ただし、電気的リンクを使用して相互に接続することができる GPU の数は、これらのデータレートで必要となる高い周波数での銅伝導体の信号損失の高さによって制限を受け、有効な距離がわずか数メートルになることがあります。結果として、電気的リンクによる GPU の相互接続の拡張性は、こうした物理的な制限の制約を受けます。

AI の光バックエンドネットワークは、可能性としては数万の GPU への拡張性があり、一般的にスパインスイッチとリーフスイッチを使用し、リーフスイッチとサーバーとの間にはレール最適化されたトポロジーを使用しています。図 1 (a) に示すこのような複雑なトポロジーは、設置が困難になる可能性があります。特に、サーバーノードとリーフスイッチとの間、およびリーフスイッチとスパインスイッチとの間に直接のポイントツーポイント接続を使用する場合はなおさらです。直接接続を使用する AI ネットワークを実装する際の困難さは、将来アップグレードが必要となったときにはさらに困難なものになります。多くのデータセンターで、AI モデルとそれらが提供するサービスの成長と歩調を合わせていくためには、将来的にはインフラストラクチャのアップグレードが必要となります。ネットワークが単一のクラスターから複数のクラスターへと成長するにつれ、直接のポイントツーポイント接続を使用している場合には、ネットワークの拡張と保守がより大きな課題となっていきます。

ここで、構造化配線が重要な役割を果たすことができます。構造化配線には、設置、ドキュメント化、将来のアップグレードとネットワークの拡張性に寄与し、保守と配線管理を容易にするモジュラー性という特徴が備わっています。このホワイトペーパーでは、レイテンシ、信頼性、物理層パフォーマンスに関する懸念についても検証します。

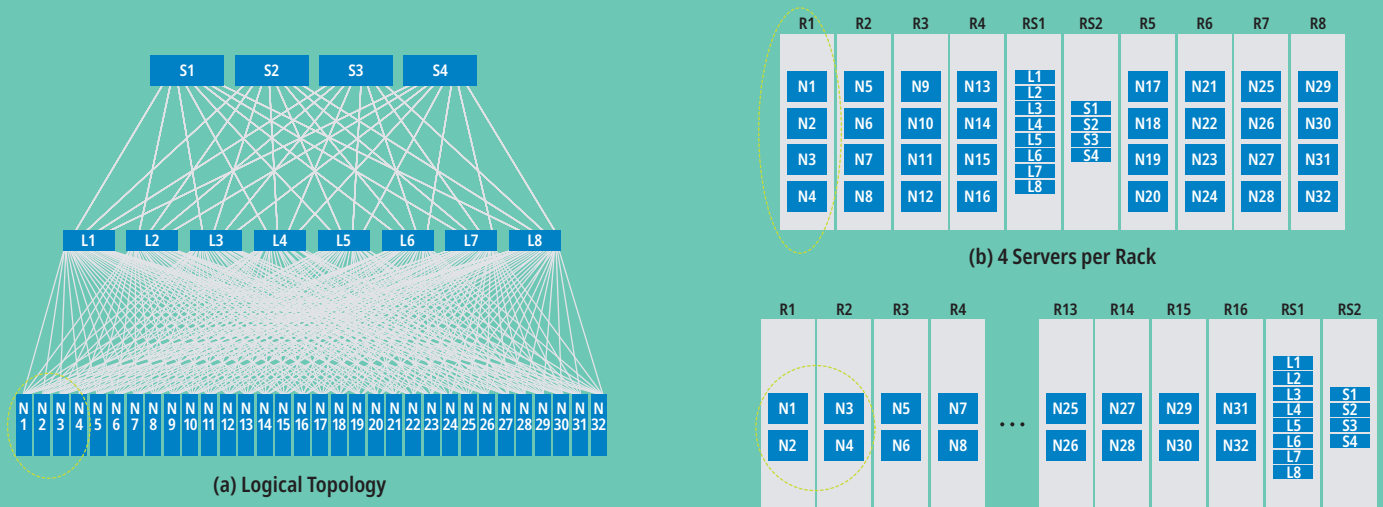


図 1
1 台の NVIDIA の拡張可能なユニットの例
(a) 32 サーバーノード (N) の論理的トポロジー
(b) 32 サーバーノード、8 リーフスイッチ、4 スパイン (ラックあたり 4 サーバーと想定) の物理レイアウト
(c) ラックあたり 2 サーバーを使用する場合の同一の論理トポロジーの物理レイアウト

テストから判明した重要な結果

構造化配線によってレイテンシに悪影響が及ぶことはない

レイテンシ

構造化配線は、ポイントツーポイント接続と比較して、より多くの接続ポイントをもたらします。光接続を追加することで光損失が増加する可能性はありますが、それによってレイテンシが増加することはありません。反対に、構造化配線システムは、ポイントツーポイントの直接接続よりも少ないケーブル余長にもかかわらず同等もしくはより優れた伝播レイテンシを提供するため、配線経路の最適化と管理をより柔軟に行うことができます。

AI ネットワークで使用される距離は比較的短い (SuperPod では 50m 未満、光の伝播遅延で約 250 ナノ秒に相当) ため、トランシーバーやスイッチで発生するレイテンシの原因のほうがより重要となる可能性があります。たとえば、FEC エンコーディングとデコーディングは、それだけで数百ナノ秒かかります。フレームバッファリングやキューイングなどの他のスイッチプロセスは、さらに大きい伝播レイテンシの原因 (数百～数千ナノ秒) となります。そのため、ネットワークを通過するパケットのホップ数を減らすことが、GPU 間で共有されるデータの遅延を低減することになります。

AI ワークロードは相互に接続された複数の GPU 間の通信パフォーマンスに依存し、大規模な分散システムでは特にそれが顕著です。結果として、長い通信遅延のあるネットワークセグメントは AI システムの運用に重大な影響を及ぼします。これらの条件下で、テールレイテンシはレイテンシの絶対値や平均値よりもより重要となります。

Spine-Leaf 構成やレール最適化トポロジーのようなネットワーク設計は、ネットワークをフラット化し、GPU 間通信に必要なホップ数を削減することで、テールレイテンシの低減を図ります。レール最適化ファブリックは、ノード内の高速内部リンク (例: NVLINK) を活用することで、ネットワークのスケールアウト時に必要となるホップ数を削減し、ネットワーク性能を向上させます。レール最適化トポロジーでは、すべてのサーバーの特定の GPU が同じリーフスイッチに接続されている必要があります。たとえば、サーバー A の GPU 0 とサーバー B の GPU 0 はリーフ 0 に接続されます。他の GPU についても同じ順序付けが行われ、図 2 に示すように、GPU 7 はリーフ 7 に接続されます。

図 2 (a-b) では、この構成によって相互接続のレイテンシが低減される仕組みも示しています。パート (a) のサーバー A の GPU 0 とサーバー B の GPU 7 との間の通信について、従来の方法ではリーフスイッチとスパインスイッチを経由する複数回のホップが必要です。ネットワーク上で信号がたどる経路を、経路

1a、2a、3a、4a として黄色の線で強調しています。通信には 3 回のホップが必要です。リーフ 0 を経由して経路 1a から経路 Path 2a に接続するホップ、スパインを経由して経路 2a から経路 3a に接続するホップ、リーフ 7 を経由して経路 3a から経路 4a に接続するもう一つのホップです。それぞれのホップで、電気から光、光から電気の変換、FEC エンコーディング/デコーディング、スイッチキューイングが必要であり、すべてが遅延の増加要因となります。

対照的に、図 2(b) に示しているこのレール最適化構成を使用すると、同じ GPU と通信するのに光ネットワークでの 1 回のホップしか要しません。

これを可能にするために、サーバー A の GPU 0 は内部の高帯域リンク (経路 1b) を使用してデータを同じサーバー内の GPU 7 に直接送信します。続いて、図に示すように、ノード A の GPU 7 とノード B の GPU 7 との間の通信では、リーフ 7 で経路 2b と経路 3b とを接続する 1 回のホップが必要です。

AI ワークロードをスケールアウトするためのフラットなネットワークトポロジーの利点がお分かりいただけたでしょうか。ただし、レール最適化ネットワークのようなフラットなトポロジーを設置するには、正確な接続が必須であり、直接のポイントツーポイント配線を使用する場合、設置が複雑になります。構造化配線によって、こうしたネットワークの設置と管理が容易になります。

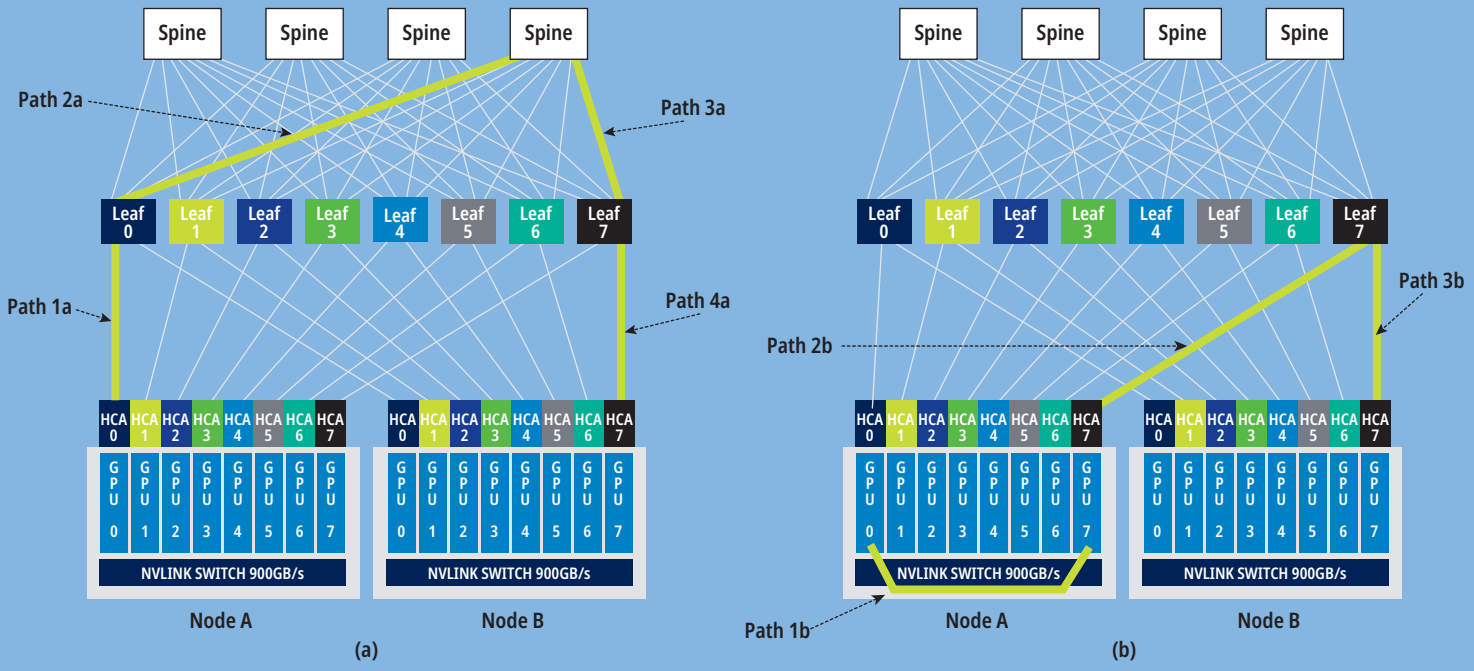


図 2
レール最適化トポロジー構成を示しています。黄色の軌跡は、(a) リーフスイッチとスパインスイッチを使用する場合と、(b) NVLINK とリーフスイッチを使用する場合のレイテンシとを比較するための信号経路を示しています。

信頼性

大規模な AI ネットワークで直接接続を使用すると、多くの場合配線が乱雑で収拾がつかなくなりがちです。相互接続とトランクのためのケーブル余長は、直接接続では一般的に長くなり、推奨値よりも小さい半径で曲げて設置されたケーブルが絡み合った状態になってしまう可能性が高まります。これはストレス、高損失、さらには恒久的なファイバーの損壊を引き起こし、故障率が高まることにつながります。場合によっては、ファイバーにかかるストレスは、時間の経過と共にガラスの亀裂に発展する可能性があります。こうした問題が発生すると、迅速な特定と解決が困難になる場合もあります。対照的に、図 3 に示す、パッチパネルを経由するパッチコードから分離されたトランクを使用する構造化配線では、ケーブルの管理が簡単になり、余長が削減されます。トランクはケーブルトレイを通じて経路が指定され、パッチコードはサーバーやスイッチに接続されます。これによりケーブルの整理、保守の容易さ、さらには将来のネットワークの拡張性に不可欠なアップグレードのしやすさが向上します。

ネットワークのドキュメント化は、大規模なネットワークの信頼性を向上させるのに重要な役割を果たします。数千の光リンクを扱う場合はなおさらです。ネットワークの回路の経路とケーブルの経路が明確にドキュメント化されていれば、ネットワークエンジニアとサービス技術者は、問題が発生したときに迅速に接続を追跡できます。したがって、故障点の特定に費やす時間を回避することで重要な AI ネットワークのダウンタイムが削減される一方、保守とサービスの費用の節約にもなり、AI システムに投じられた莫大な資本の最適化にもつながります。米国電気通信工業会 (TIA) の規格 TIA-606-C (通信インフラストラクチャ向けの管理規格) では、商用ビルでの通信システムとネットワークシステム向けのラベル付けとデータの記録に関するガイドラインが提供されており、これは TIA-568.3-D. などの構造化配線の規格と整合性のあるものとなっています。ケーブルアセンブリの両端を含むネットワークコンポーネントが出荷時点であらかじめラベル付けされている RapidID™ Network Mapping System のようなソリューションは、ネットワークのドキュメント化に大いに寄与します。

構造化配線は、よりよい整理状態とより容易な保守をもたらしますが、固有の課題もあります。一般的に、構造化配線ではより多くの接続インターフェイスが必要となります。ほとんどの設置環境はほこりやごみで汚れており、コネクタの端子面に付着する可能性があります。これによってコネクタの挿入損失 (IL) と反射減衰量 (RL) が増加する可能性があり、ネットワークのパフォーマンスの問題の原因となりかねません。

ただし、こうしたリスクは、規格のガイドラインを完全に実装している高品質の製品を使用すること、およびコネクタの端子面が汚れるのを防ぐ適切な設置手順に従うことで最小化できるものです。結局のところ、優れた拡張性や管理のしやすさといった構造化配線のメリットは、適切な注意を払えばこうした懸念を上回ります。

テストから判明した重要な結果

構造化配線は
高密度の AI 環境で
さらなる重要性を帯びる

図 3
構造化配線を使用したより優れたケーブル管理

ネットワークのパフォーマンス

AI ネットワークは、モデルのトレーニングや推論の最適なパフォーマンスを実現することを目的として、利用可能な物理層の帯域幅を最大限に活用し、テールレイテンシを最小化するよう設計されています。

パケットの再送信が増加すると、テールレイテンシは増大します。したがって、AI ネットワークのパフォーマンスにとって、チャネルのビット誤り率 (BER) を低くすることは、パケット損失、およびその結果としてのパケットの再送信を排除して、テールレイテンシを制御するために、極めて重要です。チャネルの BER は、光ファイバー、コネクタ、トランシーバーの送受信パフォーマンスに起因する信号の障害に依存します。ワーストケースでのチャネルのパフォーマンス要件は、イーサネット、光ファイバーチャネル、InfiniBand の規格団体によって指定されています。

たとえば、2024 年 3 月に公開された最新のイーサネット仕様 (IEEE 802.3df) では、マルチモードチャネル (800GBASE-SR8) およびシングルモードチャネル (800GBASE-DR8) について、8 デュプレックスレーン (16 芯) で、合計 800G のデータレートが盛り込まれています。IEEE 802.3df は、最長許容ファイバー伝送距離を通過した後の信号品質について記述しており、その際にワーストケースのトランシーバーが使用されることを前提としています。これにより、フォワードエラー訂正 (FEC) 前のワーストケースのビット誤り率 (BER) が得られます。IEEE 802.3df では、フォトディテクタ受信回路の帯域幅と感度も、その他のパラメーターと共に規定しています。規格に基づく送受信機の性能テストは、製造現場で広く使用されており、チャネルの BER を迅速かつ比較的正確に推定する手段となっています。これらのチャネル性能仕様によって、ベンダー間の相互運用性が実現しています。

AI ネットワークでの構造化配線に対する懸念の 1 つが、チャネル性能のリスクとなる可能性のあるコネクタ損失の増加です。この議論に対する反論は簡単です。イーサネットチャネル仕様に完全に準拠したトランシーバーでは、MMF チャネル (800GBASE-SR8) では 1.5 dB、SMF チャネル (800GBASE-DR8) では約 2.5 dB の接続損失が割り当てられています。ただし、AI ネットワークで使用される今日のトランシーバーは、固有のソリューションとなっており、こうした懸念に対応するために IEEE に従っているとは必ずしも想定できなくなっています。これを理解するためには、実際のチャネル性能の試験データを IEEE 802.3df のワーストケースのチャネル仕様との比較で示す必要があります。その目的に向けて、パンドウイットではシングルモードとマルチモードの 800Gbps OSFP (オクタルスモールフォームファクタ) トランシーバーの評価を実施しました。これは InfiniBand とイーサネットプロトコルの両方をサポートしており、NVIDIA DGX や HGX サーバーベースの AI クラスターの設置で使用されています (図 4 上部)。図 4 (上部) に示すように、テストは DR8 に準拠したトランシーバーと NVIDIA 専用のトランシーバーの両方について、直接接続アーキテクチャで行われました。

テストから判明した重要な結果

NVIDIA および他の IEEE 準拠のトランシーバーには構造化配線を使用するための十分なヘッドルームがある

NVIDIA はシングルモード 800G 2xDR4 トランシーバーおよびマルチモード 800G SR8 トランシーバーを、IEEE 802.3df で規定された到達距離の 500m (SMF) および 100m (OM4 MMF) ではなく、それぞれ 100m (SMF) および 50m (OM4 MMF) に縮小して提供していることに着目しました。評価を行った NVIDIA の仕様に基づいて、NVIDIA 800G-SR8 トランシーバーは IEEE 800GBASE-SR8 トランスミッターおよびレシーバーの仕様に準拠しており、最大 100m で 1.5 dB のコネクタ損失で動作が可能であるということがわかりました。これを検証するため、いくつかの異なる MMF を使用して、既製の NVIDIA 800G-SR トランシーバーの BER 性能を測定しました。これにはワーストケースを示したモード帯域幅も含まれています。また、異なるコネクタ損失条件をシミュレーションするために、モードに依存しない Keysight 製の減衰器を使用しました。この直接的な BER 計測は、製造者が使用しているオシロスコープベースのテストより時間がかかりますが、チャネル性能をよりよく表しています。

図 5 は、被試験送受信機、被試験ファイバー (FUT)、コネクタ損失をシミュレーションするための可変光減衰器からなる試験セットアップを示しています。マルチモードセットアップでは、FUT は 50m および 100m のワーストケースの規格に準拠した OM4 MMF と、当社の最高のパフォーマンスの、OM4+ Signature Core™ として知られている OM4 ファイバーで構成されています。後者はチャネルの分散を補正するファイバーで、長距離のチャネル向けに長年使用されているものです。

通常、データセンターアプリケーションで許容される BER は、1 兆送信ビットあたり 1 ビットエラー (<1e-12) をはるかに下回ります。最大 25Gbps (レーンあたり) のトランシーバーは、ゼロと 1 の伝送にシンプルな 2 レベル信号を使用し、エラー訂正スキームなしで BER <1e-12 を達成しました。今日の PAM-4 トランシーバーでは、1e-12 より大きい値の BER を達成するために FEC スキームを必要としています。こうした FEC コードでは、100 万送信ビットあたり 240 エラー (2.4e-4) までエラーレートを訂正することができ、1 兆送信ビットあたり 1 エラー (1e-12) より優れた値を達成しています。

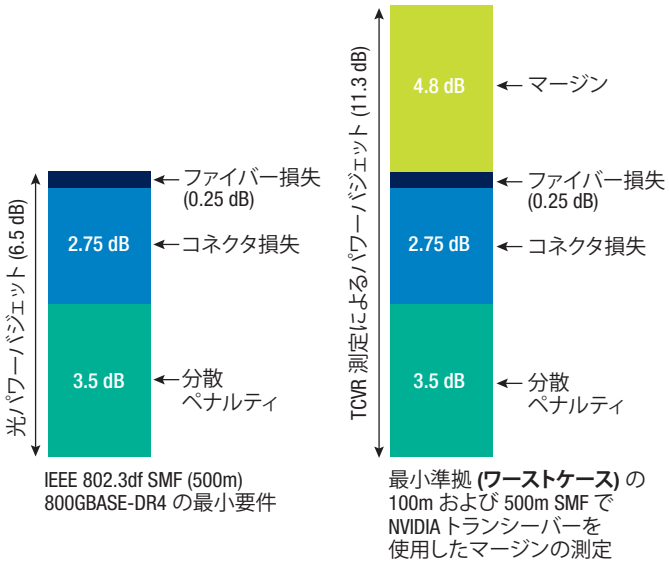
ワーストケースに準拠した OM4 ファイバー 50m を使用した当社の実験では、BER を IEEE 仕様のワーストケースの値である 2.4e-4 まで性能を低下させるのに、5 dB を超えるコネクタ損失を要しました。また、トランシーバーは 100m を超えるワーストケースの OM4 で、最大 3.5 dB のマージンをもって動作することもわかりました。分散補正を備えた 100m の Panduit Signature Core ファイバーを使用すると、そのマージンは 5 dB を超える値となりました。

パンドウイットが 400G の接続を使用する米軍のデータセンターに到達距離とパフォーマンスの面で極めて優れた提供を行った実績についてご覧ください

こうした結果が示しているのは、800G SR8 トランシーバーは指定された 50m の到達距離で 1.5 dB の接続損失に耐えることができ、さらに、経年劣化や温度変動に対応するための大きなマージン (>3.5 dB) があるということです。

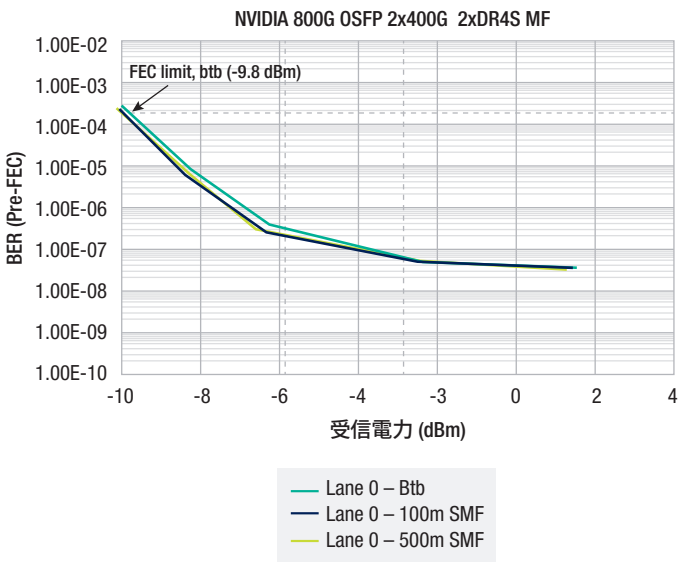
同様に、NVIDIA 800G DR8 トランシーバーを標準 SMF の 100m および 500m で評価したところ、2.5 dB を超えるコネクタ損失にも耐え、その後の経年劣化に耐えるヘッドルームがもたらされることもわかりました。さらに、100m で正常に動作するとされているトランシーバーが、500m でも良好に動作し、その際に生じる不利益は無視できる程度でした。

800G DR4 – 100m および 500m シングルモードファイバー (SMF)



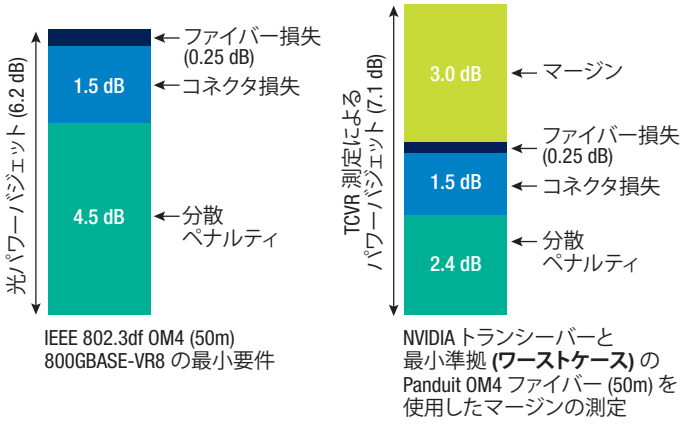
IEEE 802.3df SMF (500m) 800GBASE-DR4 の最小要件

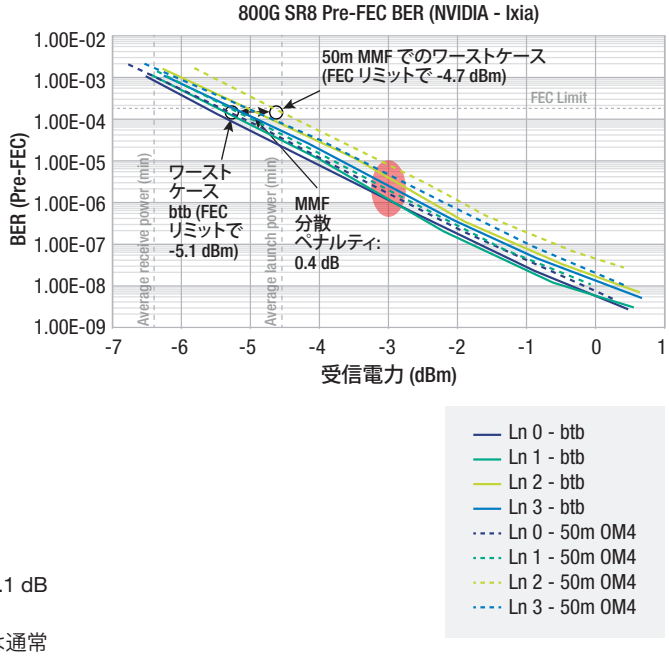
最小準拠 **(ワーストケース)** の 100m および 500m SMF で NVIDIA トランシーバーを使用したマージンの測定



注: NVIDIA のトランシーバーは、IEEE 802.3df 規格よりも 4.8 dB 高い出力を提供します。
トランシーバーの出力は時間とともに劣化するため、安定した性能を維持するには通常 1.5 dB のマージンが必要とされています。
実際の結果は、ワーストケースのケーブルを用いた測定データよりも大幅に改善されることが期待されます。

50m マルチモードファイバー (MMF)



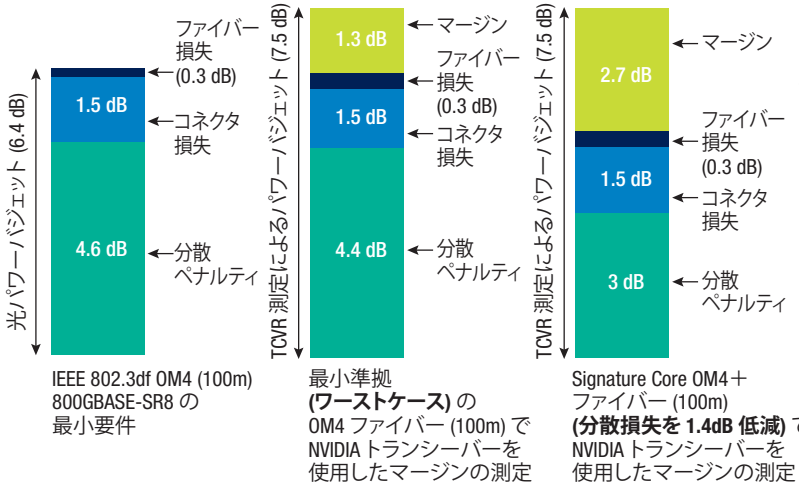


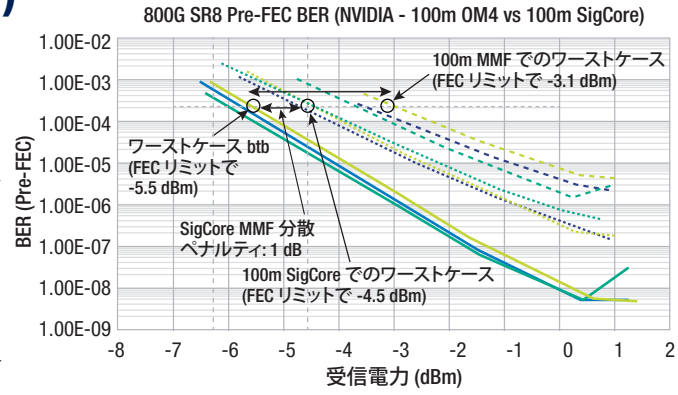
注: NVIDIA のトランシーバーは、IEEE 802.3df 規格よりも 0.9 dB 高い出力と 2.1 dB 低い分散ペナルティを提供します。

トランシーバーの出力は時間とともに劣化するため、安定した性能を維持するには通常 1.5 dB のマージンが必要とされています。

実際の結果は、ワーストケースのケーブルを用いた測定データよりも大幅に改善されることが期待されます。

100m マルチモードファイバー (MMF)





100m ワーストケース MMF: -3.1-(5.5) = **2.4 dB** は 100m MMF の分散による (モード-クロマチック)

100m ワーストケース MMF: -4.5-(5.5) = **1 dB** は 100m MMF の分散による (モード-クロマチック)

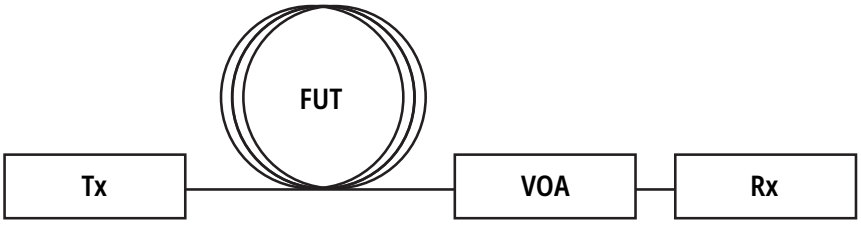
注: NVIDIA のトランシーバーは、IEEE 802.3df 規格よりも 1.1 dB 高い出力と 0.2 dB 低い分散ペナルティを提供します。

トランシーバーの出力は時間とともに劣化するため、安定した性能を維持するには通常 1.5 dB のマージンが必要とされています。

実際の結果は、ワーストケースのケーブルを用いた測定データよりも大幅に改善されることが期待されます。

図 4a

図 4b
(a) NVIDIA トランシーバー (b) チャネル接続スイッチの簡単な例
多くの場合、MPO ケーブルは実装されるトポロジーに従ってさまざまなスイッチに向かう。



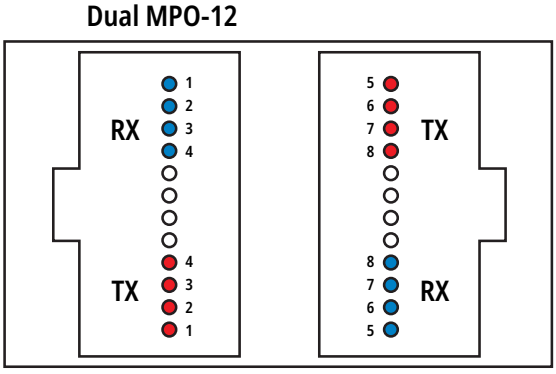
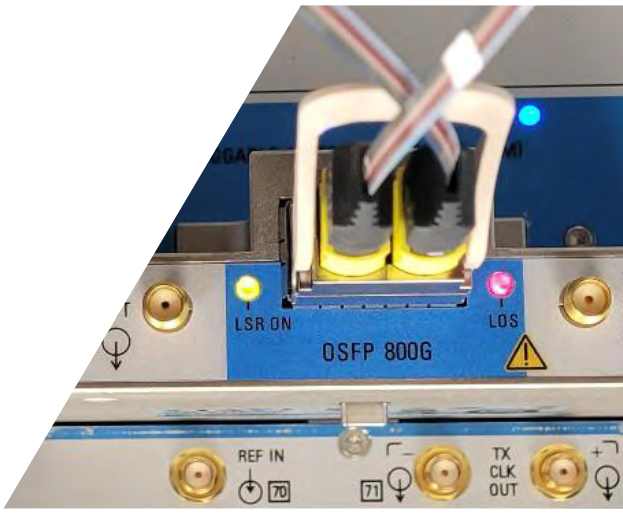


図 5
(上) 800G デュアル MPO 12 モジュールの一般的なテスト セットアップ (デュアル MPO 12 APC が使用されているため、MMF や SMF の場合と類似した セットアップ)。デュプレックス ファイバーあたり 100G の Tx トランスミッター、テスト対象 のファイバー (FUT)、コネクタ 損失をシミュレートする 可変光減衰器 (VOA)、 Rx レシーバーで構成。



トランシーバーの MPO コネクタの写真と
詳細な相互接続マップ

まとめ

大規模な AI ネットワークでは数千本のファイバーケーブルが使用されています。これは従来のデータセンターの 4 倍～ 8 倍の密度です。こうした複雑なネットワークを管理するには、標準規格に基づいた構造化配線を使用することで、より整理された状態にすることができ、ファイバーの保護、余長の収納を行う上で有益です。このホワイトペーパーでは、構造化配線が AI ネットワークにもたらす多くのメリットと、それがレイテンシや BER に悪影響を及ぼすことがないということを詳しく見てきました。

トランシーバーのパフォーマンスに対する光接続の影響を評価するため、MMF チャネルおよび SMF チャネルについて NVIDIA 光トランシーバーのヘッドルームを仕様に基づいて計算し、BER テストを計測する実験も行いました。

MMF チャネルに関するこの分析の結果から、評価対象の NVIDIA のトランシーバーとサードパーティのトランシーバー (このレポートには記載していませんが) は、レーザーの経年劣化や温度依存のペナルティに対して十分なヘッドルームを維持しながら、既定の性能に影響を与えることなく、MMF で 1.5dB の接続損失を許容できることが示されました。同様に、SMF チャネルでも NVIDIA トランシーバーは 2.5 dB の接続損失に耐え、経年劣化に対する十分なヘッドルームがあります。

仕様からの理論的な評価、テスト結果、サードパーティベンダーからの補足的なデータから、当社としては、AI ネットワークの設置、保守、拡張性にとって不可欠な構造化配線は 800G NVIDIA トランシーバーを使用して効果的に実装することができると結論付けます。この実装は、コネクタ損失を指定の範囲内に保ち、コネクティビティのクリーニングガイドラインに従うことが前提です。

テストから判明した重要な結果

- 構造化配線がレイテンシに負の影響を及ぼすことはない
- 構造化配線は、余長の過多、問題のあるケーブルの修理/交換、高密度の管理など、ケーブル配線に関する問題を軽減するのに役立つ
- NVIDIA トランシーバーはワーストケースの挿入損失の閾値よりもはるかに優れている (1+ dB) ため、構造化配線を使用するための十分なヘッドルームがある

このホワイトペーパーは、技術的スキルを持つ作業者が自分の判断と責任においてガイドとして使用することを前提に作成されています。パンドウイットのいかなる製品についても、購入者には、使用前に当該製品が目的の用途に適合することを確認する責任があります。また購入者は、それに伴うあらゆるリスクおよび責任を負うものとします。パンドウイットは、この文書に記載された、または記載されていないいかなる情報から生じるいかなる責任も負いません。

パンドウイットのすべての製品には、当該時点で最新の限定製品保証の利用条件および制限事項が適用されます。詳しくは、www.panduit.com/warranty. をご覧ください。

*本書に記載されたすべての商標、サービスマーク、商号、製品名、およびロゴの所有権は、それぞれの所有者に帰属します。

参考文献
[1] https://docs.NVIDIA.com/networking/display/800gmma4z00ns
[2] IEEE 802.3df: https://standards.ieee.org/ieee/802.3df/11107/

当社には、お客様の
インフラストラクチャ投資の
効果を最大化するための
知識と経験があります。

panduit.co.jp

ぜひご相談ください
jpn-toiawase@panduit.com

PANDUIT™