

1.2.4 研究開発項目①-3「光スイッチネットワークアーキテクチャの最適化の研究開発」

1.2.4.1 研究開発項目①-3-1「光スイッチネットワークアーキテクチャの研究開発」

【成果の概要】

次世代コンピューティングでは各種の AI 関連サービス、クラウド関連サービスを含む広範囲のアプリケーションが想定される。本研究開発項目ではディスタグリゲーション型次世代データセンタへの光スイッチの導入にあたり、様々なアプリケーション、並びに様々な規模のコンピューティング環境に最適化可能な効率的な光スイッチネットワークアーキテクチャを開発する。光スイッチを導入することにより期待される低消費電力・高性能・経済性の最大化、開発光スイッチの導入の促進、並びに広範囲な導入を達成するためには、各種のアプリケーションに応じたアーキテクチャの最適化が必要となる。また大規模スイッチの経済的な導入を図るためには、拡張性（pay-as-you-grow）が重要となる。これを達成するために、波長と空間の次元の組み合わせの最適化を図る。以下に開発した光スイッチネットワークアーキテクチャとその最適化に関する研究開発成果を報告する。

【最終目標】

分散データセンタアーキテクチャへの適用も含め、異なるアプリケーション・規模に応じた、変調形式・ポート数並びに空間スイッチ・波長スイッチ機能分担の最適化を行い、実証実験へ反映させる。将来的に 1000 ポート以上に対応するためのアーキテクチャに関して、試作したデバイスを用いた伝送実験によりその実現性を検証する。

【成果】

<研究開発項目①-3 の位置付け>

ムーアの法則の減速並びに AI/ML のプロセッシングの急増に際し、データセンタ内ネットワークのパフォーマンスと消費電力の課題を抜本的に解決するためには、新しい革新的なアーキテクチャを開発する必要がある。デバイスレベルでは、パッケージ内ではフォトニクス技術は帯域幅密度とエネルギー効率の大幅な改善をもたらすことが期待され、技術開発が加速されている。ネットワークレベルでは、光スイッチング技術の導入がキーとなる[1]。光スイッチング技術は、ルーターや ODU-XC (Optical channel Data Unit-Cross Connect) などの一部の電気スイッチ機能を置き換える ROADM/OXC として、メトロおよびワイドエリアネットワークにすでに広く導入されている。データセンタ内ネットワークでは、光スイッチング技術の役割は ROADM/OXC とは異なり、単純な光カットスルーの実現に加えてアーキテクチャの革新が必要となる。何故なら光によるカットスルーのみで得られる効果は光スイッチング技術の導入を推進するのには十分とは言えない。ネットワークパラダイムを現在のマルチティアからシングルティアに変更することは、電気スイッチング技術では実現不可能であり、ビットレートの透過性および WDM と SDM の効果を最大限活用することにより可能となる光本来の大容量性により初めて可能となる。本委託研究で開発する光スイッチをベースとするデータセンタネットワークは、既存の電気スイッチで構成されるネットワーク機能の一部を置換えたものではなく、新しい機能の導入で有るため、本技術が既存のデータセンタを置換可能でかつ future proof で有ることを示す必要がある。本研究開発項目①-3 は図 1.2.4.1-1 に示すように、開発するアーキテクチャが実現するデータセンタは、本プロジェクトのマイルストーンで実現されるパラメータ領域に加えて、既存のデータセンタを含む幅広い領域で、将来に渡って革新をもたらす技術である事を実証すること

を目的としている。

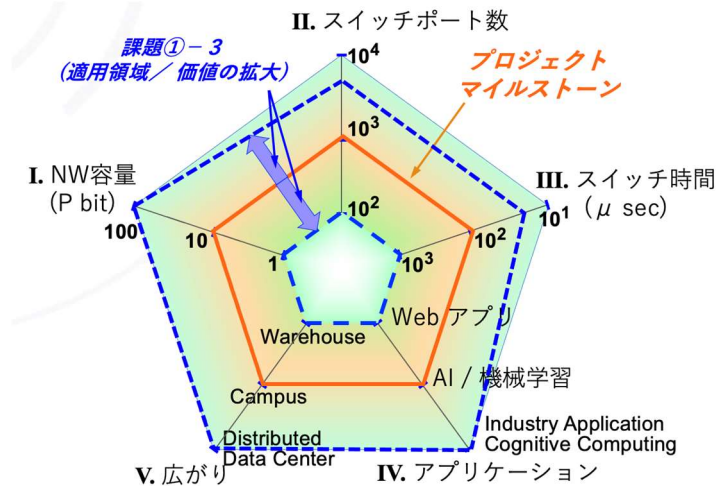


図 1.2.4.1-1 研究開発項目①-3

<光スイッチによるデータセンタ内ネットワークの革新>

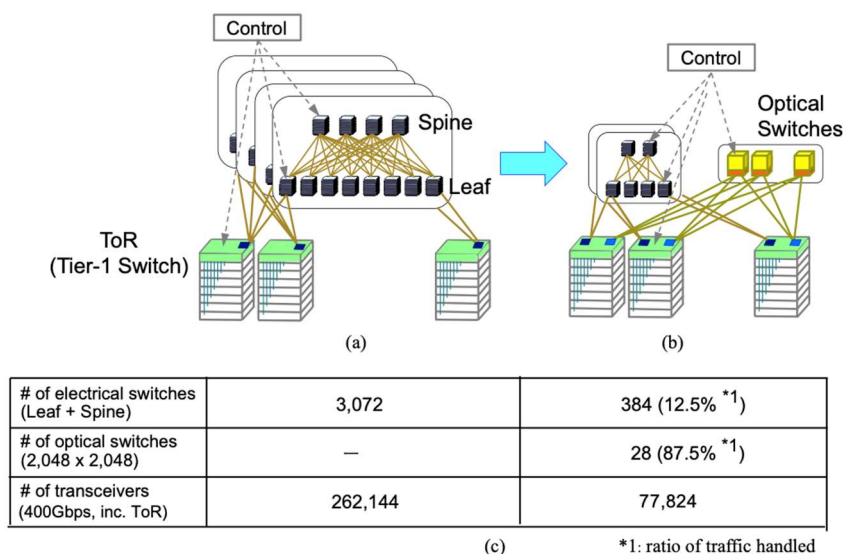


図 1.2.4.1-2 (a) 3 層電気スイッチ ネットワーク、(b) 電気と光のハイブリッドスイッチネットワーク、(c) 各々に必要なハードウェア規模の比較。

光スイッチは、信号形式とビットレートに対する透過性（ビットレートトランスペアレンシーと呼ばれる）を有し、さらに波長分割多重 (WDM) を利用できる。光スイッチは電気スイッチとは異なり、ポート速度とポート数相互の依存性が小さい。実際には、光リンクの速度を上げると、光パワーバジェットが減少し、実現可能な光スイッチポート数が減少するが、電気スイッチチップとは異なりこの関係は反比例とはならない。つまり、光スイッチを使用すると、多ポートと高いリンク速度をより簡単に同時達成できる [1,2]。Folded Clos ネットワークトポロジは、ネットワークコストを大幅に削減するとともに、同一

のスイッチチップを用いて大規模なネットワークを作成できるため、現在のハイパースケールデータセンターネットワークで主に利用されている。このアーキテクチャでは、ネットワーク内のホスト（サーバーやメモリ）、ToR、インタコネクトリンク数、スイッチ数、およびネットワークのバイセクション帯域幅は、スイッチチップの帯域幅、スイッチ radix (k)、およびスイッチング階層 (n) によって決まる。図 1.2.4.1-2(a)と(b) は、それぞれ、典型的な電気スイッチ並びに本研究開発で想定する電気/光ハイブリッドスイッチネットワークを示す。図 1.2.4.1-2(b) では、光スイッチのポート数を増やすと、収容する Tier-1 電気スイッチの数またはネットワークのバイセクション帯域幅が増加する。これは、Tier-1 スwitchの radix, k , に依存しない。このことは、Tier-1 スwitchの最大数 ($k^{n-1}/2^{n-2}$) がスイッチ radix によって決定されるファットツリー (Clos)型電気スイッチネットワークとの根本的な違いである [2]。光スイッチネットワークでは、リンク速度 (C) は $C=Te/2S$ で決まる。ここで、 S は光スイッチの並列数、 Te は電気スイッチチップの帯域である。即ちリンク速度 (C) は Tier-1 電気スイッチ数ではなく、後で説明する光スイッチのブロッキング性能によって決まる。電気スイッチネットワークでは $C=2Te/k$ であるため、スイッチ radix によってリンク速度が決まることに注意する必要がある。ポート数が大きい光スイッチは、2 層 (光スイッチと電気スイッチ) のネットワークを非常に効果的に実現できる。図 1.2.4.1-2(c) は、3 階層の電気スイッチネットワークと、トラフィックの 87.5% を光スイッチが処理すると想定した 2 階層の光および電気ハイブリッドスイッチネットワークを比較している。ここでは、電気スイッチチップの帯域幅(Te)を 25.6 Tbps、収容するサーバー数を 131,072 (262,144)、サーバーの I/O 速度を 200 (100) Gbps、Tier-1 スwitch数を 2,048、電気スイッチの radix とリンク速度はそれぞれ 64 と 400 Gbps とし、実現されるネットワークバイセクション帯域幅は両者とも 52.4 Pbps である。図 1.2.4.1-2 (c) に示すように、ハイブリッドスイッチネットワークでは、電気スイッチチップの総数 (ToR スwitch数 2,048 を含む) は 5,120 から 2,432 (-52%) に減少し、トランシーバー並びにインタコネクファイバー数は 262,144 から 77,824 (-70%) に減少するが、28 個の 2,048 x 2,048 光スイッチが必要となる。その結果、スイッチネットワーク (ToR スwitchを含む) の消費電力は 65% 以上削減される [3]。上で分析したように、光スイッチを使用してスイッチレイヤの数を減らすと、必要なスイッチチップの数、電力消費量、およびインタコネクファイバーとトランシーバー数が大幅に削減できる。ToR 間で通過する電気スイッチチップの数も減少し、これにより、遅延と、頻繁に発生するロングテールの問題が抑制される [2]。トラフィック制御に関しては、可能なルート数が減るため、トラフィック管理がより簡単になり、障害が発生した場合には、その診断がより簡単になる。このスイッチ階層の削減は、適度に高速なリンク速度を使用して一定数の tier-1 スwitchを収容できる多ポートの光スイッチを導入することで効果的に実現される。

<多ポート光スイッチの開発>

これまでにさまざまな材料とメカニズムを使用する種々の光スイッチ アーキテクチャが提案されてきたが [4]、高速 ($\sim 10 \mu s$) スwitchングと大ポート数 ($>1,000$) の要件を同時に満たすアーキテクチャは実現されていなかった。我々の提案では、総ポート数は使用する波長数と空間スイッチポート数の積で与えられる。即ち光技術で利用できる波長と空間軸の組み合わせを併用することにより、多ポートのスイッチが効果的に実現可能となる[1]。Si フォトニクス技術を活用し、必要な高速性を達成する。類似のアーキテクチャは、商用の ROADM (再構成可能な光アド/ドロップマルチプレクサ) または OXC (光クロスコネク) を実現するために広く利用されている。1990 年代初頭に発表された基本的な OXC/ROADM 構成 [5] は、ルート&コンバインアーキテクチャ (入力に波長選択スイッチ (WSS) 機能、出力に光カプラーが配置される) である。データセンタのスイッチに関しては、図

1.2.4.1-3 に示すように、2つの異なるアーキテクチャを提案している。一方は送信機に波長可変レーザーを使用し、他方は受信機に波長可変フィルターを使用して波長ルーティングを実現する。これらのアーキテクチャは、ROADM の Add および Drop 部にそれぞれ類似の、鏡面対象の構成をとる [1]。ここで光スイッチの機能要素は分散して実装されることに注意されたい。チューナブルレーザー、またはチューナブルフィルターとレシーバーは、図 1.2.4.1-3 に示すように Tier-1 (ToR) スイッチに実装されるため、光スイッチ自身は非常に単純な構成部品、小規模の空間スイッチと EDFA、で実現されている。データセンタおよび ROADM 向けに提案された光スイッチは、アーキテクチャにおいては多くの共通点を有しているが、主要なパフォーマンスの違いの一つに、必要なスイッチング速度がある。ROADM のほとんどのアプリケーションでは、秒単位のスイッチング時間が許容されるが、データセンタのアプリケーションでは、後述するように $10\ \mu\text{s}$ 未満が現在の目標となる。従って我々の光スイッチは、現在の ROADM で用いる MEMS (微小電気機械システム) または LCoS (シリコン上の液晶) ベースの WSS ではなく、Si フォトニクス技術をベースとして波長可変フィルター[6][7]と空間スイッチ[8]を実現している。一方、データセンタでは伝送距離が短く ($< 2\ \text{km}$)、必要なホップ数は1であるため、データセンタスイッチに要求される伝送性能は ROADM と比べて大幅に緩和される。

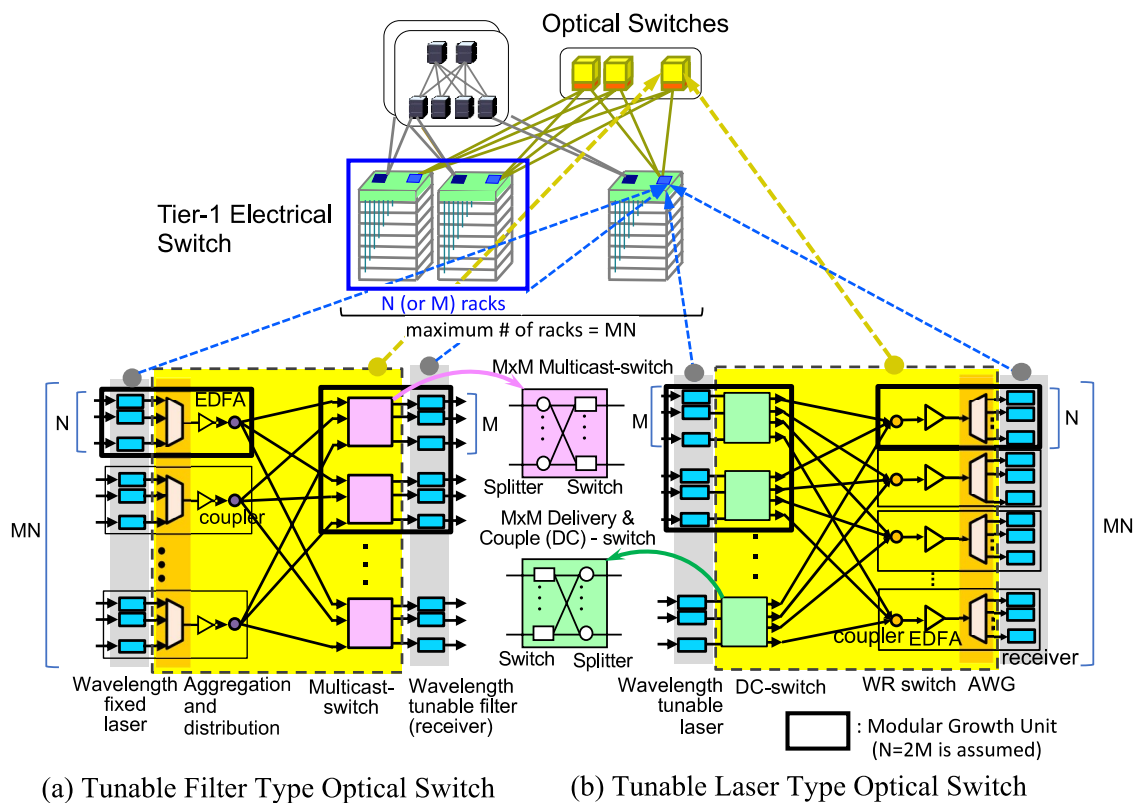


図 1.2.4.1-3 多ポート光スイッチのアーキテクチャと実装：(a) チューナブルフィルター型光スイッチ、および (b) チューナブルレーザー型光スイッチ

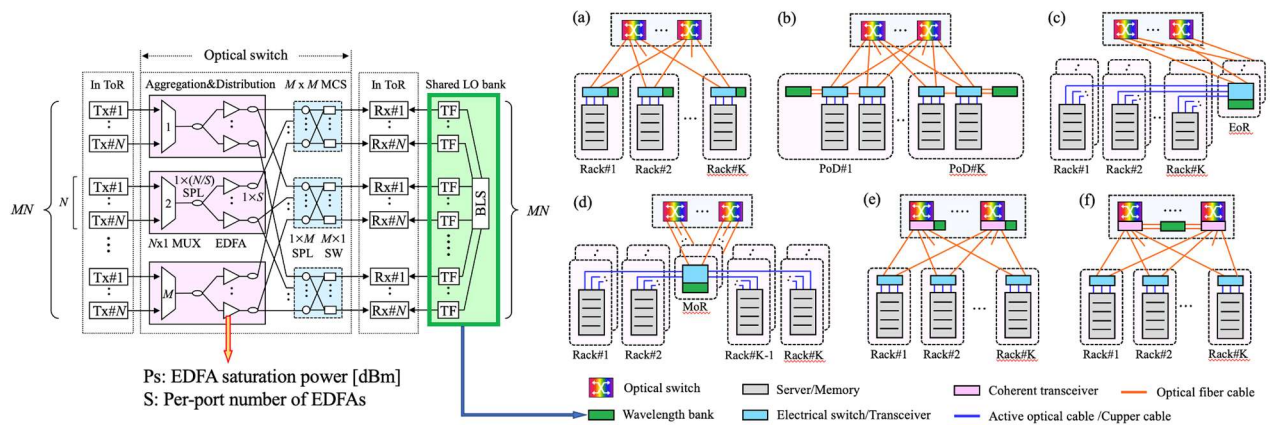


図 1.2.4.1-4 波長可変 LO 光が波長可変フィルターを備えた LD バンクから供給されるカラーレス検波を使用するチューナブルフィルタータイプ光スイッチ。(a)-(f) 異なる LD バンクの実装法。(a) 各 ToR、(b) 複数の ToR、(c) End of Row、(d) Middle of Row、(e) 各光スイッチ、(f) 複数の光スイッチ、に配置される場合。

図 1.2.4.1-4 は、LD バンクから供給される波長可変 LO 光と波長可変なフィルターを使用したカラーレスコヒーレント検波を採用した波長可変フィルターベースのスイッチを示す [9][10]。図 1.2.4.1-4 (a)-(f) に示すように、LD バンクの実装はさまざまな形式が考えられる。TLD（波長可変レーザー）に関しては、現在開発が進められているコパッケージ化された光学部品の進歩に注意する必要がある。そこでは、スイッチを含む電気デバイスは光学部品とコパッケージ化されるが、熱の制約のために、外部レーザー光源を利用することが検討されている。外部光源からの光は偏波保持ファイバーを介して供給される。これらの技術が一般的に利用可能になると、TLD は固定波長 LD とチューナブルフィルターの組み合わせにより効果的に実現される。図 1.2.4.1-5 は、図 1.2.4.1-4 に示される波長可変フィルタータイプのスイッチの実現可能なポート数を示している [9]。異なる変調形式について、Ps（EDFA 飽和電力）、および S（EDFA の複数ポートでのシェア数）をパラメータとしている [9]。費用対効果の観点で、EDFA は共有ポート数を最大化することが望ましい。この様に、波長可変フィルタータイプのスイッチにおいても大きなポート数と帯域幅を実現することができる。たとえば、256 Gbps DP-QPSK 信号の場合、22 dBm の EDFA 飽和電力と EDFA 共有数 2 において、7,424 を超えるポート (1.48 Pbps を超える帯域幅) がサポートできる [9]。最近のデモ実験では、128 Gbaud DP-QPSK 信号用を用いた 7,424 ポートスイッチが、新しく開発された波長可変フィルター、集積された偏波無依存 64×1 シリコンフォトニクススイッチ並びに高速バーストモードコヒーレント受信器を用いて、1.4 μ s のスイッチング時間を実現した [10]。シリコンフォトニクススイッチのファイバー-チップ間の結合損失は 2.6 dB/端面で有ったが、最近の研究 [8] で報告されているように、1.4 dB/端面に低減可能である。光スイッチ ネットワークのバイセクション帯域幅は、光スイッチを並列化することより拡大され、並列数が 64 の場合ネットワークの帯域幅は 48 Pbps となり、現在最大のハイパースケールデータ センタよりもほぼ 1 桁大きくなる。さらに、波長可変レーザータイプのスイッチは、ロスの大い空間スイッチが EDFA の前に配置されるため [11]、波長可変フィルター タイプのスイッチと比べてより広い帯域幅が実現される。

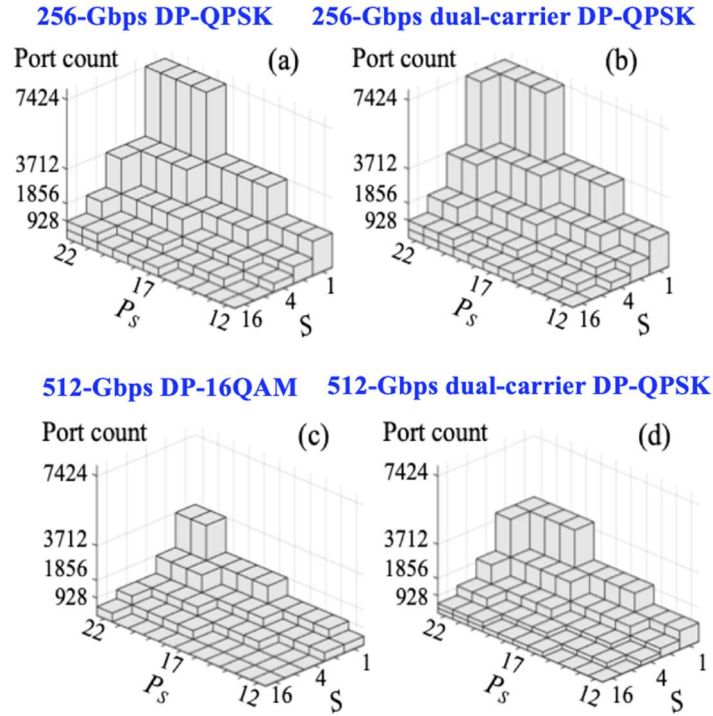


図 1.2.4.1-5 図 1.2.4.1-4 に示す光スイッチの可能なポート数、(a) 256 Gbps DP-QPSK、(b) 256 Gbps デュアルキャリア DP-QPSK、(c) 512 Gbps DP-16QAM、および (d) 512 Gbps デュアルキャリア DP-QPSK 信号を想定

<7,424 × 7,424 光スイッチ伝送実験>

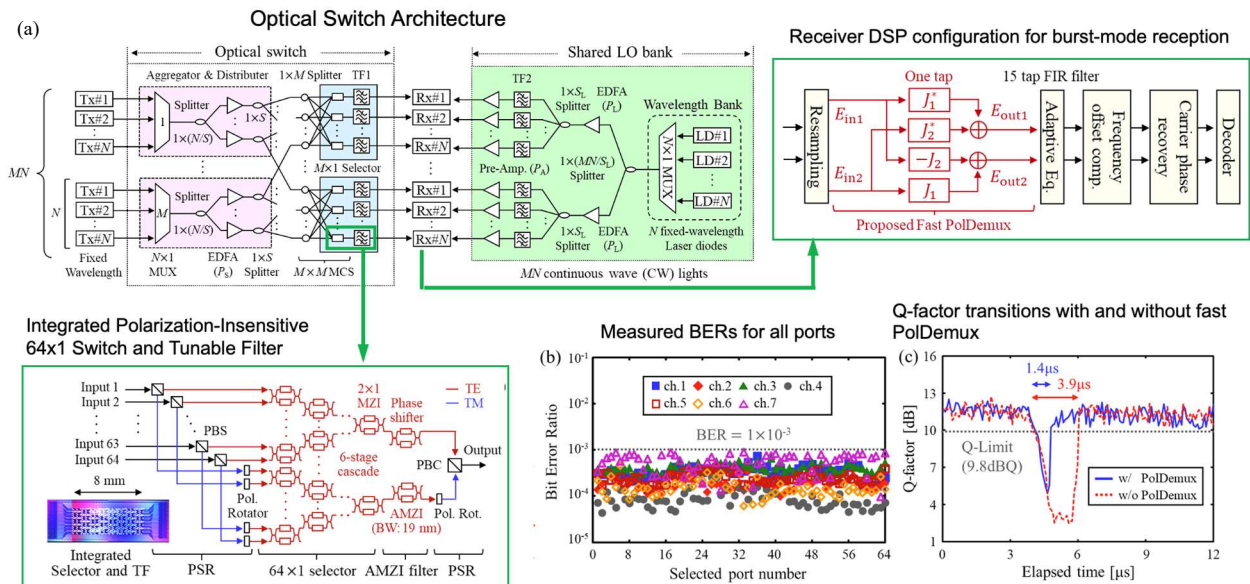


図 1.2.4.1-6 (a) $MN \times MN$ チューナブルフィルタータイプ光スイッチアーキテクチャ並びに受信部バースト受信 DSP 構成、(b) 7 つのチャネルと 64 個のスペーススイッチポートのすべての組み合わせに対する BER、(c) 高速 PolDemux を使用した場合と使用しない場合のスイッチング特性 (Q ファクターの遷移)

Modulation format (SC: Single carrier, DC: Dual carrier)	Channel speed [Gbps/ch.]	Number of wavelength (N)	Channel spacing [GHz]	$M = 64$	
				Switch port count	Throughput [Tbps/switch]
32Gbd SC DP-QPSK	100	116	37.5	7,424	742.4
64Gbd SC DP-QPSK	200	58	75	3,712	742.4
32Gbd DC DP-QPSK				3,712	742.4
43Gbd SC DP-8QAM		88	50	5,632	1126.4
64Gbd DC DP-QPSK	400	32	137.5	2,048	819.2
43Gbd DC DP-8QAM		48	100	3,072	1228.8

図 1.2.4.1-7 各変調方式に対して実現可能な光スイッチの規模

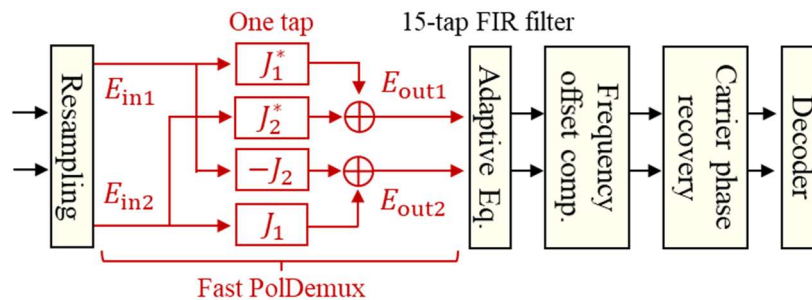


図 1.2.4.1-8 バーストモード受信器 DSP 構成

光スイッチのポート数拡張性を実証するために $7,424 \times 7,424$ ポートの光スイッチの伝送実験を行った[10]。以下その概要を示す。図 1.2.4.1-6 (a)に実験に使用した $MN \times MN$ チューナブルフィルタタイプ光スイッチアーキテクチャ並びに受信部バースト受信 DSP 構成を示す。同図(b) に、実験により得られた 7 つのチャネルに対して測定した 64 個のスペーススイッチポートのすべての組み合わせに対する BER を、(c) に高速 PolDemux を使用した場合と使用しない場合のスイッチング特性（Q ファクターの遷移）を示す。スイッチポート数は MN で与えられるが、空間スイッチ (M) が大きいほど損失が大きくなる。カラーレス検波で利用できる波長数 (N) は、受信機のダイナミックレンジによって制限される。18 dB のダイナミックレンジは 63 チャネル ($10^{18/10}$) に対応する。受信機の直前に TF1 を設定すると、受信機の電力飽和を緩和できる [12]。個別部品の追加は損失を大きく増加させるため、損失を最小限に抑えるために TF とモノリシックに集積した低損失のシリコンフォトニクス 64×1 光スイッチを開発した。ここでは、TE モード設計とオンチップ偏光ダイバーシティ方式を採用した。これにより、 64×64 MCS（マルチキャストスイッチ）が実現され、116 波長を使用することにより、図 1.2.4.1-7 に示すように、32 Gbaud SC DP-QPSK 信号に対する $7,424 \times 7,424$ 光スイッチの開発に成功した。図 1.2.4.1-8 は、今回開発したバースト信号受信器 DSP の構成を示す。この構造は、典型的なデジタルコヒーレントレシーバーで 사용되는構造を利用しているが、CMA (constant-modulus algorithm) ベースの適応イコライゼーションの前に高速偏波分離 (PolDmux) を採用している。DSP の高速応答は、偏波追尾にストークスベクトルを使用すること、トレーニングシーケンスを使用しない適応操作を用いることで実現されている。 64×1 セレクターと TF をチップに集積し、偏波無依存多ポート空間スイッチを構成した。図 1.2.4.1-9 (a) に示すように、126 個の対称マッハツェンダ干渉計 (MZI) エレメントスイッチと 2 個の非対称 MZI (AMZI)を用いて、偏波無依存を実現するために TE モード用に設計された 2 つの同一の 64×1 セレクターと TF を用い、偏波スプリッター (PSR) を

利用した偏波ダイバーシティ回路を構成している。空間スイッチでは、PSR が 2 つの直交偏光を分割し、TM 光は TE 光にチップ上で変換されるため、任意の入力偏光に対して動作する。各 64×1 セレクターは、ツリー構造の 2×1 MZI エlementスイッチ 6 段のカスケードにより構成される。AMZI フィルターの 3 dB 帯域幅は約 19 nm で、受信機の電力飽和を防ぐのに十分な特性を示す。すべての MZI は、位相シフターによって熱光学的に制御される。図 1.2.4.1-9 (b) にオンチップ偏光ダイバーシティを備えたチップの顕微鏡写真を示す。実現回路面積は、 $3 \times 8 \text{ mm}^2$ であった。図 1.2.4.1-9 (c) は、1547 nm の信号光に対してすべての入力ポートにおいて測定されたファイバー間挿入損失 (F2F-IL) と偏波依存損失 (PDL) を示す。64 入力ポートの平均 IL は 9.4 dB で、平均 PDL は 0.27 dB と良好な値が得られた。クロストークは、C バンド (35 nm) で -20 dB 未満であった。

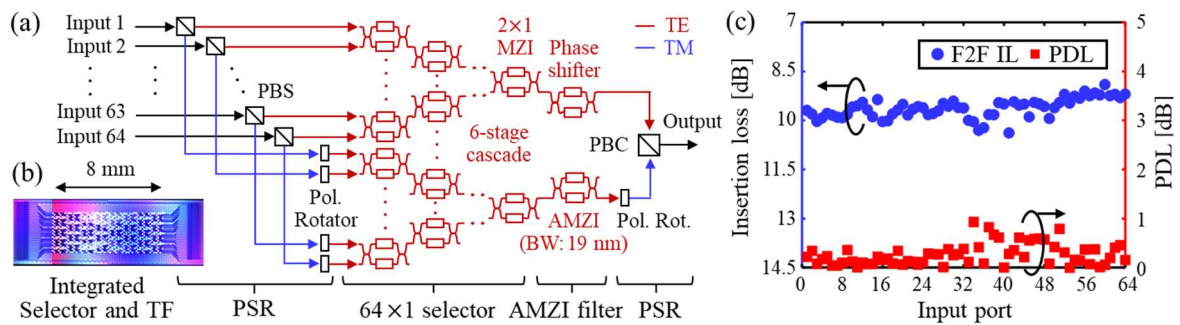


図 1.2.4.1-9 集積化偏光無依存 64×1 セレクターとチューナブルフィルター：(a) 構造、(b) 作製したシリコンフォトニックチップ、(c) 挿入損失と PDL のポート依存性

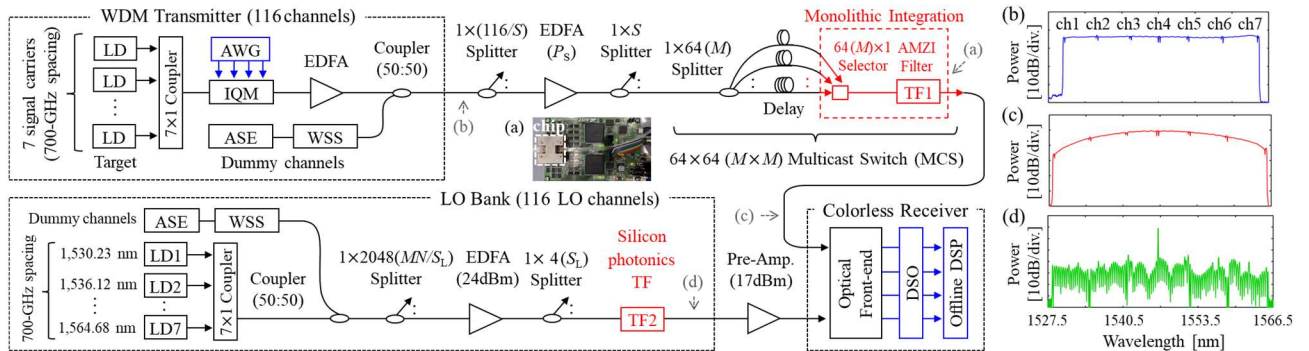


図 1.2.4.1-10 32Gbaud DP-QPSK 信号を使用した $7,424 \times 7,424$ 光スイッチ実験 ($M = 64$ 、 $N = 116$ 、 $SL = 4$)：(a) 制御電子回路を備えたプリント回路基板に取り付けられたシリコンフォトニック 64×1 空間スイッチ、(b) WDM 信号の生成、(c) ポート 35 を介した空間スイッチング、(d) 1547.116 nm (ch.4) でのシリコンフォトニック TF による波長チューニング後の光スペクトル

図 1.2.4.1-10 に作製した 64×1 空間スイッチを使用した $7,424 \times 7,424$ 光スイッチの実験セットアップを示す。IQ 変調器 (IQM) と任意波形発生器 (AWG) を使用して、1,530 nm から 1,565 nm まで 700 GHz 間隔で配置された 7 つのチャネルを一括変調した。生成された信号は、ロールオフ係数が 0.05 のルートレイズドコサインフィルターによって整形された 32 Gbaud DP-QPSK である。ダミーチャネルを組み合わせることで、全体で 14.85Tbps ($116\text{ch} \times 128\text{Gbps}$ DP-QPSK) の WDM 信号を構成した [図 1.2.4.1-10(b)]。ダミーチャネルには波長選択スイッチ (WSS) を使用してスペクトル整形された増幅器自然放出 (SS-ASE) 光を用いた。送信された WDM 信号は $1 \times (116/S)$ スプリッターで分配され、 P_s

の飽和電力を有する EDFA で増幅し、さらに 1×S スプリッターで分配される。64×64 MCS では、入力信号は 1×64 スプリッターを介して 64 パスにブロードキャストされ、1 つのパスが 64×1 セクターと TF で構成される試作モジュールを使用して選択される [図 1.2.4.1-10 (a)]。クロストークをエミュレートするために、64 パスの信号は、ファイバー遅延線を使用して少なくとも 100 シンボル分無相関化された。受信機の前に挿入される TF1 の機能により受信機はダイナミックレンジ内で適切に動作する [図 1.2.4.1-10 (c)]。LO バンクでは、37.5 GHz 間隔の 116 チャネルの LO を、7 個の LD 光と SS-ASE 光とを多重することによってエミュレートした。多波長光は、スプリッターの損失を補償するために、2 段の分配器 (1×2,048 および 1×4 スプリッター) と EDFA を利用してブロードキャストされ、作製したシリコンフォトニック TF [13] により目的の波長チャネルが抽出される [図 1.2.4.1-10 (d)]。バーストモードプリアンプを介して 17dBm に増幅されたターゲットチャネルは、コヒーレント受信機の LO として使用された。信号は、光フロントエンドでコヒーレント受信され、デジタルストレージ オシロスコープ (DSO) を使用してデジタルデータとして記録した。検出された信号は、図 1.2.4.1-8. の DSP にてオフラインで処理された。

図 1.2.4.1-10 (a) において EDFA 飽和電力 (P_s) が、22.9 および 24.8 dBm、 $S = 1$ および 2 に対してそれぞれ 1×10^{-3} の BER が得られることがわかった。図 1.2.4.1-6 (b) に、チャネル周波数と各空間スイッチポートに対して測定した信号品質を示す。448 チャネル (7 WDM×64 ポート) のスイッチングが 1×10^{-3} 以下の BER で実現されることを確認した。光パスをポート 1 からポート 64 に変更するとともに、TF の通過帯域を ch.3 から ch.4 に同時に変更して、スイッチング性能を測定した。1.2.4.1-6 (c) に、高速 PolDemux のオンとオフを切り替えたときの時間分解 Q ファクター値を示す。高速 PolDemux 無しの場合は 3.9 μ s であったが、バーストモード DSP を用いることにより 1.4 μ s の高速スイッチング時間が実現できた。

<フィードフォワード制御を使用した高速シリコンフォトニクスダブルリングチューナブルフィルターの開発>

上記に示した様に、これまでシリコン (Si) マッハツェンダ干渉計 (MZI) フィルターを主に利用してきた。シリコンマイクロリング共振器 (MRR: micro-ring resonator) は、フットプリントが非常に小さく、広帯域の波長チューニングが可能であるため、狭帯域 TF として有望なデバイスである [10,9]。Si-TF の高速チューニングは、光スイッチのみならずチューナブルレーザーやラボオンチップなどの多くのアプリケーションでも重要である。ただし、熱光学 (TO) 狭帯域 MRR のチューニング速度は、帯域幅が狭いため非常に遅く (数十または数百マイクロ秒)、TO ベースの MZI とは大きく異なる。TO ベースの MZI では、高エネルギーの短時間の矩形電気パルス (ターボパルス) とそれに続く一定の加熱電流を使用して、高速チューニングを実証した [14]。ここでは、指数関数的に減衰するフィードフォワードアナログパルスを採用することにより、MRR のチューニング速度を大幅に向上できることを理論的および実験的に示した [7]。時間分解分光法を用いて、最適な制御パルスを得る方法を開発した。提案方法を使用することにより、フル C バンド (1530~1565 nm) の波長範囲で約 20 GHz の帯域幅を持つダブルリングタイプの光フィルターの高速 (<9.4 μ s) チューニングを実現した。その結果、単純なステップ関数型制御と比較して、チューニング速度が 25 倍以上向上した。以下にその概要を示す。

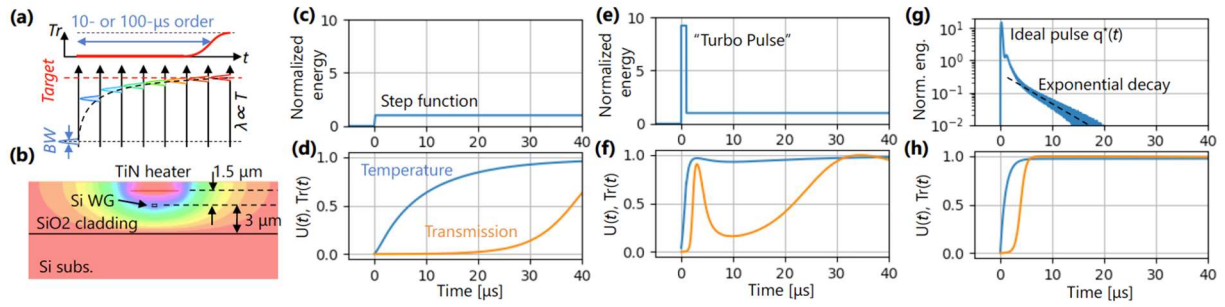


図 1.2.4.1-11 試作リングフィルタに関する解析 (a) 光学応答の概略、(b) 定常温度分布、(c) ステップ入力、(d) (c)に対する温度と光出力応答、(e) ターボパルス入力、(f) (e)に対する温度と光出力応答、(g) 理想的な入力パルス、(h) (g)に対する温度と光出力応答

図 1.2.4.1-11 (a)に狭帯域 MRRの応答が遅い理由を示す。図中の下部の垂直矢印は、ターゲット温度まで加熱する際の各時間ステップの狭帯域幅MRRのスペクトルを示す。温度は指数曲線 (黒い破線) でターゲットに近づくものの、光スペクトルが狭いため、温度が十分に収束した後に光透過率 (Tr) が増加し、応答が 10 または 100 μs オーダーと遅くなる。これを検証するために、有限要素法 (FEM: Finite Element Method) を使用してさまざまな時間関数を持つ電力をヒーターに注入するとき、図 1.2.4.1-11 (b)に示すフィルタ構造に関して非定常解析を実行した。これらの分析には Freefem++ を使用した。同図は定常状態の温度分布を示す。最終温度 T_∞ で正規化された導波路コアでの温度の「ステップ応答」 $U(t)=T(t)/T_\infty$ は、同図 (d)の青線で示されている。単一 MRR の正規化された光透過率 $Tr(t)$ はローレンツ関数 $f(\phi) = \phi^2/(\phi^2 + [\pi/F]^2)$ と仮定される。ここで、 ϕ は MRR の round-trip の位相、 F はフィネスで、同図 (d) にオレンジ色の線でプロットされている。図に示すように、透過率は数十マイクロ秒後に増加する。MZIの場合、Turbo Pulseの適用[同図(e)] は、迅速な応答を得る上で効果的な方法で有ったが、同図 (f) のオレンジ色の線で示されているように、MRR の場合温度の緩やかな変化が数十マイクロ秒にわたって残る。これにより、フィネスが大きい場合 (実験において約 25)、光学応答が大幅に低下する可能性がある。したがって、単純な Turbo Pulse を狭帯域 MRR に使用することは困難となる。ここでは、以下に示す逆設計されたパルス形状の利用を提案する。ステップ応答 $U(t)$ を用いて、 $G(s)=L[U(t)]s$ によりシステムの応答関数を計算できる。ここで、 L はラプラス変換を表す。逆設計された理想的な入力関数 $q^*(t)$ は、 $q^*(t)=L^{-1}[U^*(s)/G(s)]$ で計算できる。ここで、 $U^*(s)=L[U^*(t)]$ は目標関数、 L^{-1} は逆ラプラス変換である。ターゲット関数を $U^*(t)=1-e^{-t/\tau_0}$ と見なした場合に計算した $q^*(t)-1$ 関数を同図 (g) に示す。ここで目標時定数 τ_0 を 1.0 μs に設定した。同図 (h)に示されているように、短いパルスの後に数マイクロ秒の時定数を持つ指数関数的減衰が見られた。この指数関数的減衰は、デバイスの遅い応答をキャンセルするために追加されたことになる。これは、Si ワイアグリッド周囲の熱容量によって引き起こされたものである。これらの数値解析から得られた重要な結果は、ステップ応答を実験的に評価することにより[同図 (d)]、任意の応答関数 $U^*(t)$ を取得するための理想的なパルス形状 $q^*(t)$ を自動的に計算することが出来るということである。この理想的なパルスをデジタル アナログコンバーター (DAC) を介してヒーターに印加することにより、Si-MRR ベースの光フィルタの時間応答を高速化できることになる。

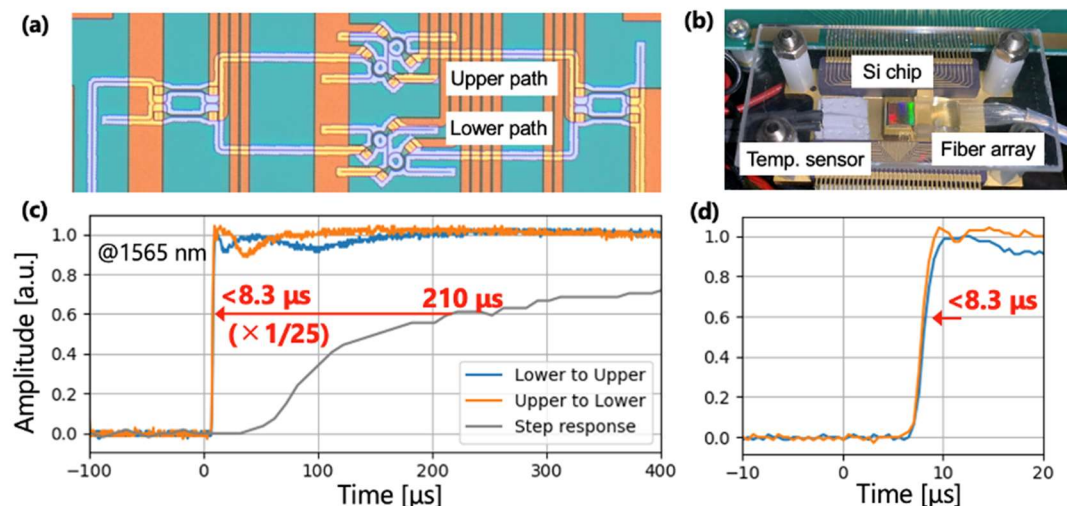


図 1.2.4.1-12 (a) 試作デバイスの顕微鏡写真、(b) パッケージ化されたデバイスの写真、(c)、(d) 25 倍高速な応答を示す、MZIを搭載した二重リング型光フィルターの高速チューニング測定結果

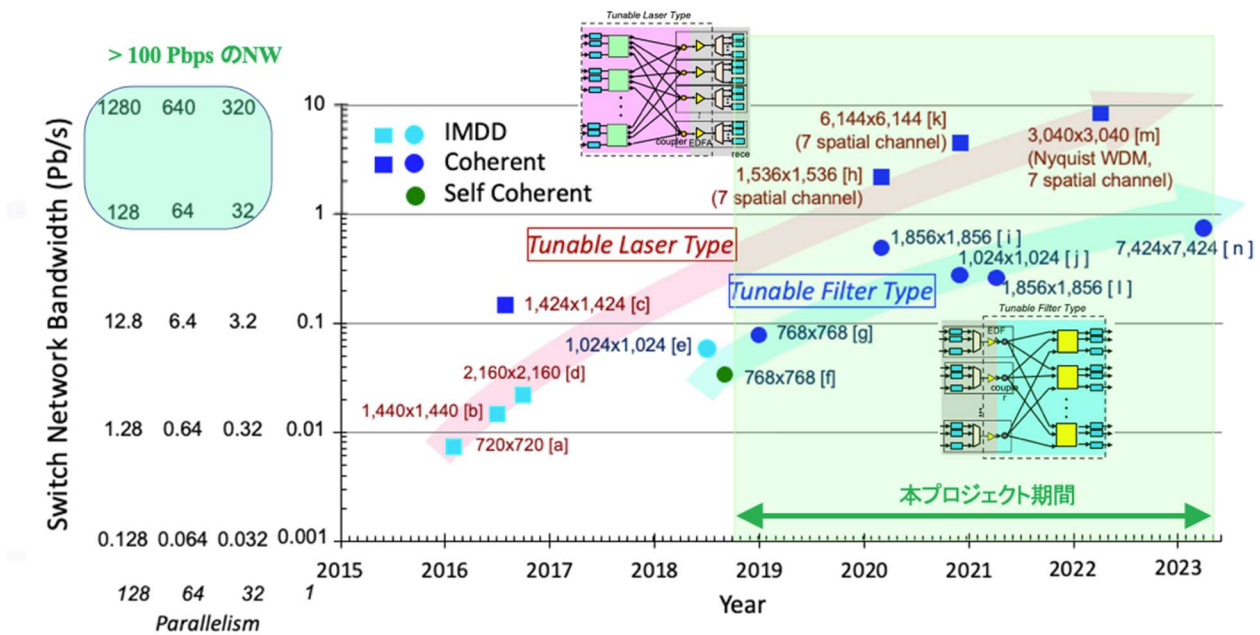
図 1.2.4.1-12 (a)と(b)に今回試作したデバイスの顕微鏡写真とパッケージ全体の写真を示す。温度安定化にはTECを使用した。同図 (c) は、光学応答特性を示しており、ステップ応答は灰色の線でプロットされている。ターゲット波長は 1565 nm であり、これは最も応答が遅かった波長である。同図 (c) と (d) より、60% の立ち上がり時間は 8.3 μs 未満で、ステップ応答に比べて 25 倍高速化が達成されている。

以上、指数関数的に減衰する逆設計のフィードフォワードアナログパルスを採用することにより MRR のチューニング速度を大幅に低下できることが示された。開発した時間分解分光法を用いて最適な制御パルスを得ることにより、フルCバンド (1530～1565 nm) の波長範囲で約20 GHzの帯域幅を持つダブルリングタイプの光フィルターの高速 (<9.4 μs) チューニングが可能なることを実験的に示し、単純なステップ関数型制御と比較して、チューニング速度の25倍以上の高速化を達成した。

<本プロジェクトによる光スイッチスループットの拡大>

図 1.2.4.1-13 に本委託研究で達成された、イントラデータセンタネットワーク用光スイッチのスループットの実験結果をまとめる。これまで IMDD、Self-coherent、および Coherent を含むさまざまな変調形式を調査した。コヒーレント技術は、高ビットレート (マルチレベル変調) と高い受信感度を可能にするため、ポート数とスイッチのスループットを拡張する上で重要である。光スイッチネットワークによりリンク速度を高速化できるが、リンク速度と光スイッチの並列性は、以下の項で説明するように、ネットワークの所望のブロッキング率並びに目標とするリンク使用率の両者の観点から導出される。したがって、大規模なネットワークを形成するために高いスイッチ radix が必要な市販のシリコン電気スイッチベースのネットワークよりも、費用対効果の点でコヒーレント光リンクが有望となる。提案アーキテクチャに関して、LO (Local Oscillator) レーザー ダイオードバンク [9]、空間スーパーチャネル [15]、および共有 LO バンクから供給される伝送信号と LO チャネルの協調フィルタリング方式[16]を使用するものを含む、さまざまなバリエーションがテストされてきた。光スイッチネットワークの設計と性能の詳細については、文献 [1, 9, 11, 10] を参照されたい。同図に示されるように、数千ポートのスイッチの実現可能性が検証された。これらのアーキテクチャと光スイッチの並

列配置を使用して、約 100 ペタ ビット／秒のバイセクション帯域幅を実現できる。この値は現在最大規模のデータ センタ ネットワークの 10 倍以上である。



[a] K. Ueda, et al., SPIE Photonics West OPTO, Proc. SPIE 9775. [b] K. Ueda, et al., IEICE Transactions of Communication, Vol.E100-B, No.10, 2017. [c] E. Honda, et al., OECC/PSC, WA1-3, 2019. [d] K. Ueda, et al., 2016 Photonics Society Conference (IPC), MB4.2. [e] M. -E Ganbold, IEEE Photonics Tech. Lett., Volume: 30, Issue: 16, 2018. [f] Y. Mori, et al., ECOC 2018, Mo3H.4. [g] Y. Mori, et al., IEEE/OSA JLT, Volume 37, Issue 2, 2019. [h] E. Honda, et al., OFC 2020, W1F.2. [i] R. Matsumoto, et al., OFC 2020, W2A.36. [j] R. Matsumoto, et al., ECOC 2020, Th1J-2. [k] E. Honda, et al., ECOC 2020, Th1J-1. [l] R. Matsumoto, et al., OFC 2021, W1A-6. [m] T. Kuno et al., OFC 2022, Th3A.2. [n] R. Matsumoto, et al., OFC 2023, M4J.5

図 1.2.4.1-13 波長可変レーザータイプと波長可変フィルタータイプの光スイッチのスループットの向上

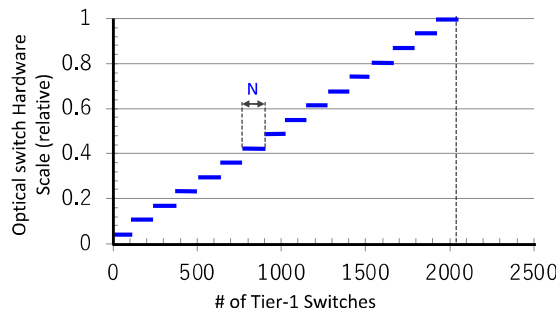


図 1.2.4.1-14 2,048 x 2,048 スイッチのモジュラー拡張の例 (N=128、M=16)

このような大規模なポート数の光スイッチを費用対効果の高い方法で導入するには、pay-as-you-grow のネットワーク拡張の実現が望ましく、即ち光スイッチのモジュラー拡張機能が重要である。最大ポート数の大きな光スイッチは、Tier-1 スイッチラック (またはクラスタースイッチ)の増設に対応できる。この利点は、各光スイッチ自体がモジュラー拡張機能を提供できる場合に最大化される。提案光スイッチは、図 1.2.4.1-3 に示したように、波長ルーティングと空間スイッチングを組み合わせたもので有り、それぞれの独立した次元はモジュラー単位を提供する。図 1.2.4.1-3 では、通常波長数 N は空間スイッチスケール M よりも大きく、モジュラーユニットは太字の長方形で示されている (簡単のために $N = 2M$ を想定)。図 1.2.4.1-14 に $N=128$ および $M=16$ の 2,048 x 2,048 光スイッチのモジュラー拡張の例を示す。

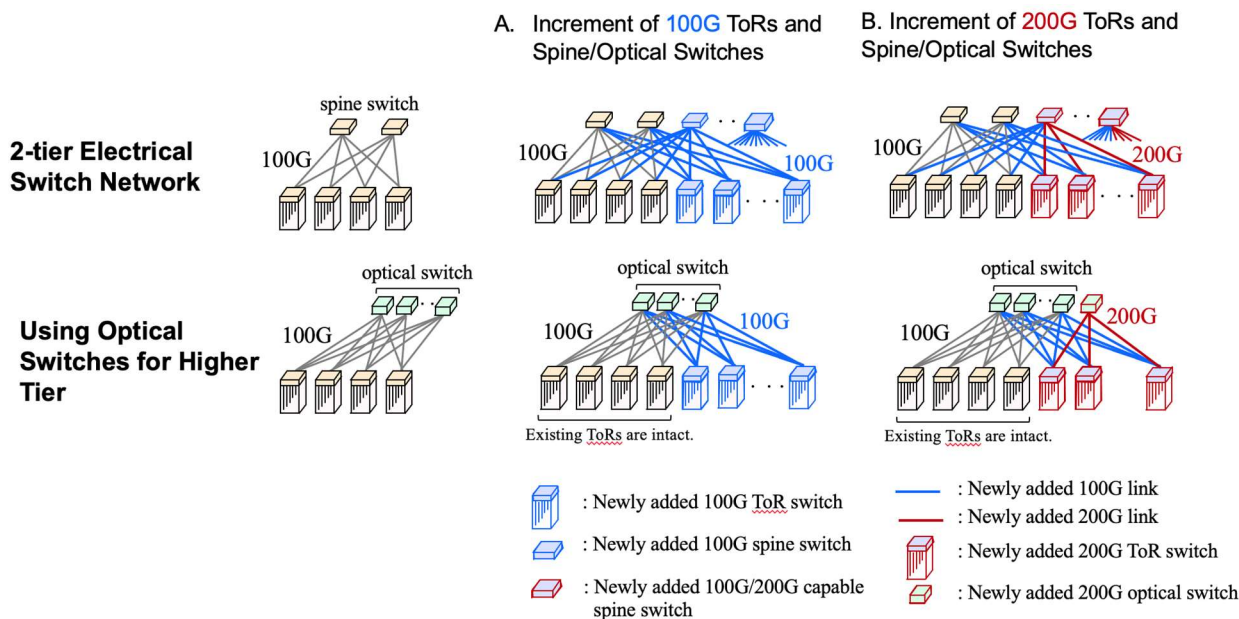


図 1.2.4.1-15 Comparison of Switch Network Increment

上記では光スイッチ自体の拡張性を論じたが、次に、スイッチネットワークにおける新しい技術の導入（例えば電気スイッチのレーン速度の高速化）に関する拡張性を、電気スイッチネットワークと光スイッチネットワークで比較する。図 1.2.4.1-15 A, B で、単純化された 2 層スイッチネットワークにおける ToR の増設を比較する。一方のネットワークは 2 層の電気スイッチで構成され、他方のネットワークは光スイッチを上位層のスイッチとして使用している。光スイッチの場合、ToR の増設は容易に実現できる。各光スイッチにおいては必要なモジュラー拡張ボードをキャビネットに挿入し、ToR を追加して相互に接続する（光スイッチラック内のボード間のファイバー接続は予め実装されている必要がある）。一方、電気スイッチについては、必要な新規 Tier-2 スイッチを追加し、既存および新規追加の ToR と接続し、また新たに追加した ToR は既存並びに新規追加の Tier-2 スイッチに接続する必要がある。このプロセスは、光スイッチのプロセスよりもはるかに複雑となる。また、データセンタでは異種テクノロジーの追加も重要である [17]。図 1.2.4.1-15 B は、100 Gbps ベースの既存ネットワークに 200 Gbps リンクとそのシステムが追加される場合を説明している。光ネットワークでは、200Gbps 対応の光スイッチを必要に応じて追加し、新たに追加した 200Gbps 対応の ToR に接続し、追加した 200Gbps 対応の ToR を（100Gbps のリンクにより）既存の 100Gbps 対応の光スイッチに接続する。この時光スイッチはモジュラー的に効率的に増設される。新たに追加した ToR の 200Gbps ポート同士の接続は 200Gbps 対応の光スイッチを介して行われる。この様に光スイッチネットワークでは、新しいリンク速度と旧来のリンク速度は異なるスイッチで収容されるために、シンプルな接続で実現できる。本例では電気スイッチネットワークを 2 階層としているが、上位層スイッチの帯域幅は、光スイッチと比べ電気スイッチははるかに小さいため、大きなバイセクションバンド幅のネットワークを電気スイッチで作成するには、スイッチング層の数を多くする必要があり、実際のネットワークでは電気スイッチとして 3 階層以上が多く用いられるために、増設はずっと複雑化する。これらに比べて、ToR より上位を光スイッチ 1 階層で構成する場合は、異なる速度の ToR の増設に対して

も、電気スイッチに比べて増設が容易になる。

＜光スイッチコントロールネットワークの開発＞

データセンタ内ネットワークで光回線スイッチを実現するには、効率的なコントロールプレーンネットワークの開発が重要である。光回線交換および光タイムスロット/セル交換 [18-21] については、さまざまな研究が行われてきた。図 1.2.4.1-16 は、光スイッチネットワーク制御メカニズムを分類する。本テーマでは光パケットスイッチングは用いない。光スイッチの制御遅延はデータセンタへの光スイッチの適用範囲を決定する上で重要な指標となる。初期の研究のほとんどは集中制御によるトラフィックスケジューリングに焦点を当てていたが、トラフィックマトリックスの作成と動的更新の手順、光回線割当て計算 (計算負荷が大きいと大きな遅延が発生する)、およびネットワーク構成ハードウェアの設定制御が必要となるため、トータルの制御遅延が大きくなる。分散制御を採用するとそれらの制御の複雑さが緩和され、より短いレイテンシが実現される。これに関連して、光タイムスロット (またはセル) スwitching が研究された [19-21]。[20] では、各光スイッチはすべてのエンドポイント間で均一な帯域幅を提供する固定された静的な割り当てを循環させる。この方式は一様 (または固定) なトラフィック分布に効果的だが、それからずれば場合の特性は劣化する。固定されたスケジューリングではネットワーク帯域幅を浪費し、時として必要な複数ホップにより遅延が増加する。[19] は、専用リンクとアダプティブなリンクを使用して、それぞれ小さいフローと大きいフローを伝送する。

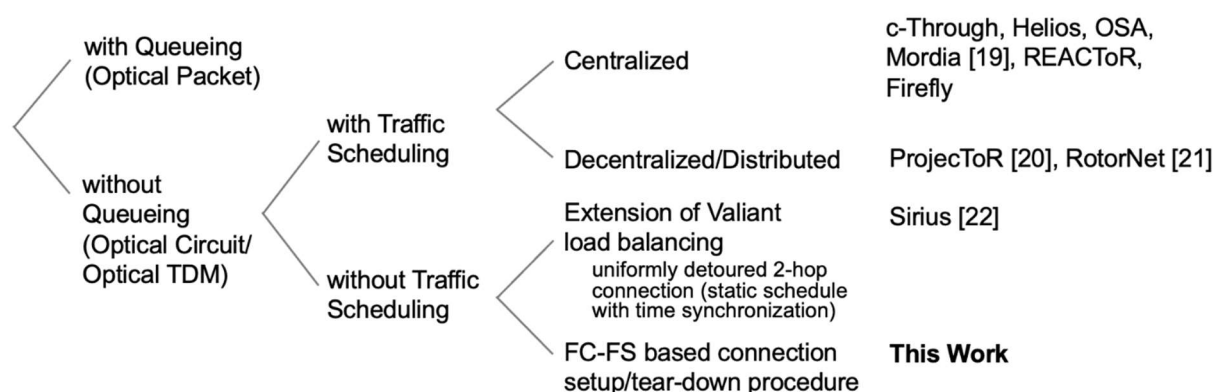


図 1.2.4.1-16 光スイッチングネットワーク制御アプローチの分類

大規模なフローに対する簡素化されたフロースケジューリング (オフラインのソリューションベース) は、帯域幅のロスを犠牲にして待ち時間を短縮できる。トラフィックのスケジューリングがなければ、制御の遅延を最小限に抑えることができる。[21] は、ロードバランシングルーティングを使用したラウンドロビンベースの静的制御を使用して、各ノードからのトラフィックが宛先に向かう時に他のノードに一樣に迂回をする。これには常に 2 ホップが必要となるため、必要な帯域幅はシングルホップネットワークの 2 倍になる。さらに、このセル (固定パケット) スwitching には、データセンタ全体でのスケジューリングが必要となる。すべてのノードは約 100 ピコ秒の精度で同期する必要があるため、ノードペア間の伝搬遅延を正確に推定する必要がある。

我々が提案する制御スキームはユニークな方式である。スケジューリングが不要で完全に分散された制御であるため、latency が短くなる。回線の切り替えには、回線のセットアップ/ティアダウン手

順が必要となる。これらの手順を簡素化し、高いパフォーマンス (低ブロッキング確率) を達成するための鍵は、単一層 (ポート数の多い) 光スイッチネットワークと光スイッチの並列処理である。これらがどのようにしてシンプルで高性能な光接続制御を実現できるかを図 1.2.4.1-17 で説明する。単一層の光ネットワークにより送信元と宛先の Tier-1 (ToR) スイッチ を直結することにより、ネットワーク全体の使用状況に基づくルーティング制御が不要になる。以下に示すように、光スイッチの並列処理により、光スイッチのブロック率が大幅に低減する。接続セットアップ制御は、ソース Tier-1 スイッチによりトリガーされ、完全分散で非同期的に行われる(ネットワーク全体の同期は不要)。接続設定要求は、アプリケーション駆動型とトラフィック駆動型の 2 つのメカニズムに基づく。後者については、フロールーターメカニズム [22,23] の利用を想定している。詳細は、本報告では省略する。トラフィック制御は、光回線スイッチ接続セットアップに関して、フロー間の遅延を最小限に抑えた単純な先着順 (FC-FS) メカニズムによって自動的に行われる。より多く発せられた接続設定要求は、より多くの接続を自動的に確立する (需要駆動型)。もちろん、インキャスト輻輳 (多くの送信元が共通の宛先に同時に送信する) は、SDN (Software Defined Networking) を使用する現在のデータセンタコントロールネットワークと同様の手法で制御する必要がある。制御システムの実装並びに光スイッチおよび可変デバイスの制御には、後述するように FPGA を使用することにより処理のレイテンシを最小限に抑えることができる。

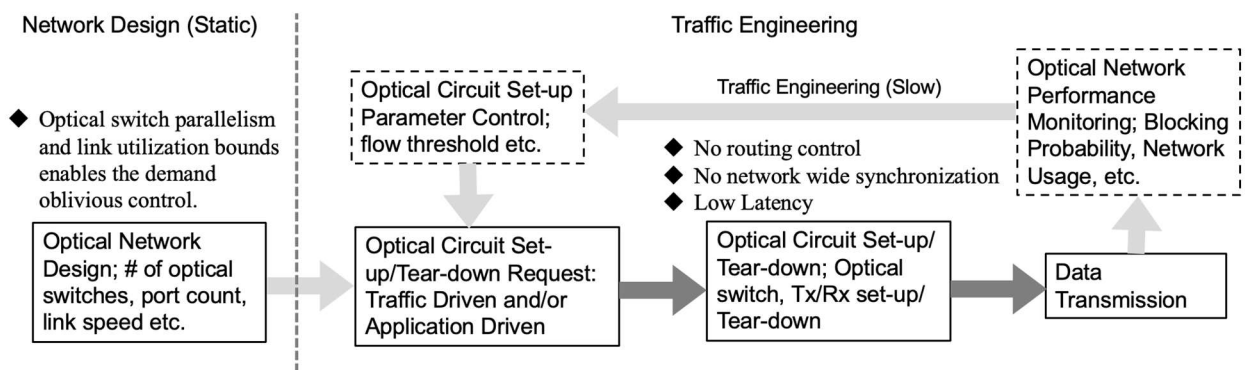
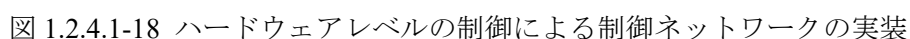


図 1.2.4.1-17 データセンタにおける光回線設定制御の戦略

FPGA を使用すると QSFP28 トランシーバーからの制御パケットの受信と、接続テーブルの検索と設定の処理を含めて、500 ns 以内に接続設定の決定を完了することができる[2]。この単純な接続セットアップ手順により、接続遅延が最小化され、一定レベルの光スイッチの並列動作により優れたブロッキングパフォーマンスが得られる。Tier-1 スイッチアップリンクの数、または光スイッチの並列数が 64/(128) の場合、平均リンク使用率 0.57/(0.68) で 10^{-6} のブロッキング確率が達成されることが示されている [3]。ここで、接続設定要求はパケットではなくフローに関連し、要求は ToR の任意のペア間でランダムに生成された。各フローでのパケット生成は、ランダムプロセスよりバースト性が高く構わないがそれはブロッキング率とは関連しない。さらに、光回線設定がブロックされた場合、ブロックされたフローは、ハイブリッドスイッチングネットワークを構成する電気パケットスイッチングネットワークによって処理でされる。ブロッキングレートを非常に小さく保つことができるため、電気スイッチネットワークの平均使用量はほとんど増加しないため、電気スイッチネットワークのパフォーマンスにほぼ影響を与えない。この様に、光スイッチの並列性を最大限に活用することで、FC-FS 接続の簡単なセットアッププロセスにより、低遅延でほぼ損失のないネットワークを構築でき

＜制御ネットワークの設計と性能検証＞



- 59 -

μs 未満のスイッチ切り替え時間とチューナブルフィルターチューニング時間を実現した。これは、ターボパルスがない場合の $28.5 \mu\text{s}$ から大きく短縮されている。

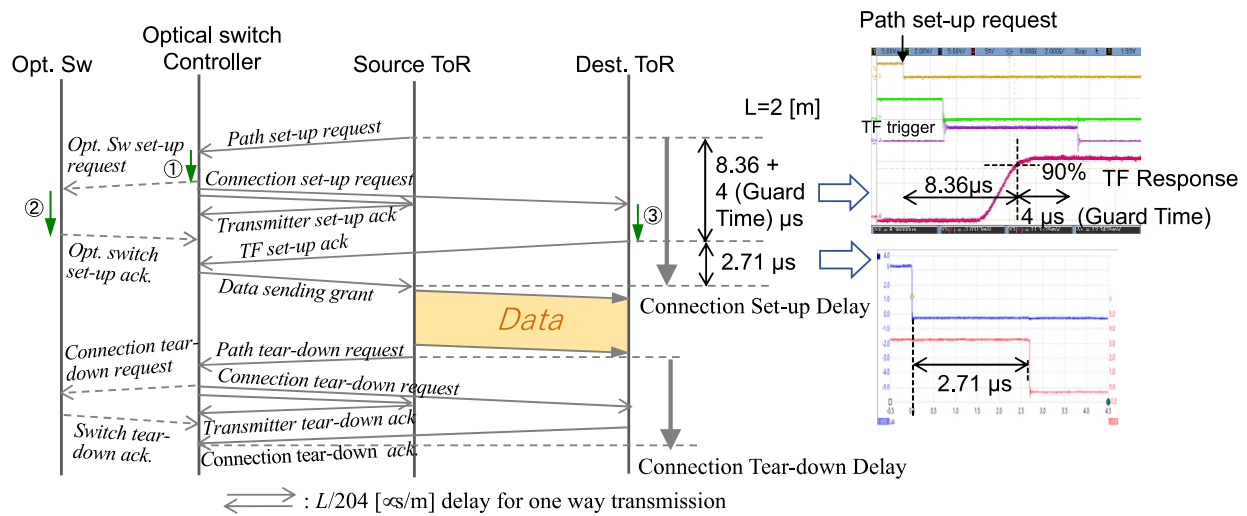


図 1.2.4.1-19 基本的な光回線設定/切断手順 (イレギュラー処理は含まれない) と測定された遅延の例

図 1.2.4.1-19 に、基本的な光回線の設定/切断手順を示す (イレギュラー処理は含まれない)。TCP 接続で使用する堅牢なスリーウェイハンドシェイクタイプの接続セットアップ手順が採用されている。実験では、光スイッチコントローラは 128 個の $2,048 \times 2,048$ 光スイッチを制御できるように設計した。コントローラには、 $2,048 \times 128$ のスイッチ入力ポートまたは出力ポート接続設定テーブルがあり、エントリは 1 (占有) または 0 (フリー) となる。必要なメモリ量は 0.5 Mbit ($128 \times 2,048 \times 2$ ビット) 未満で、使用可能な FPGA メモリの約 0.2% であった。同テーブルを利用して、3 FPGA クロック ($20 \text{ ns} \times 3 = 60 \text{ ns}$) 内で接続判断を行うことができる。図 1.2.4.1-20 に、光コネクションのセットアップ時間の測定例を示す。ソース ToR が接続設定要求を送信した後、宛先 ToR 受信器が TF 波長の設定を完了するまでに $8.36 \mu\text{s}$ かかり、追加の $4 \mu\text{s}$ (ガードタイム) 待機した後で TF 設定 ACK が送信される。送信元の ToR が光スイッチ コントローラからデータ送信許可を受信するまでに $2.71 \mu\text{s}$ が必要となる。実験では、ToR と光スイッチ コントローラ間のファイバー長 L は 2 m で有った。図 1.2.4.1-19 では、(1) 測定された接続セットアップレイテンシと、(2) 解析的に評価されたレイテンシの内訳を示す。後者には、光ファイバー伝送遅延、 $L [\text{m}] / 204.4 \mu\text{s}$ が含まれる。実験では、制御ネットワーク用のイーサネットスイッチの使用率は非常に低い。最悪の遅延を評価するには、最も混雑した状況でのスイッチのキューイング遅延を求める必要があり、これは 100 Gbps のリンク速度を想定して解析的に評価した。なお制御パケット長は 64 バイトとした。制御パケットの最悪キューイング遅延は、同図 (2) に示すように、 10^{-9} の確率となる値で評価した。接続確立に必要な全体の遅延は、同図 (3) $16.9 + L [\text{m}] / 51.1 \mu\text{s}$ となる。ここで、 $16.9 \mu\text{s}$ の内訳は、全体のコントローラ処理遅延 $6.9 \mu\text{s}$ (コントロールのプロセッシング遅延とネットワークが最も混雑した状態での制御パケットのキューイング遅延) 並びに $10 \mu\text{s}$ の TF 設定時間 (TF チューニング時間と約 $4 \mu\text{s}$ のガードタイム) で有る。 $L [\text{m}] / 51.1 \mu\text{s}$ は、ToR と光コントローラ間の 2 往復分の制御パケット伝送遅延である。これにより、 $L=200 \text{ m}$ (フロア内) で $20.8 \mu\text{s}$ 、 $L=2,000 \text{ m}$ (キャンパス内) で $56.0 \mu\text{s}$ のトータルのレイテンシが生じる。簡略化された

一方向 (ハンドシェイクなし) 接続設定手順を採用し、シリコンフォトニクスの特光学効果の代わりに電気光学効果を使用して、1 μs (0.5 μs の光スイッチング時間と 0.5 μs のガードタイム) の光スイッチまたは光チューナブルフィルタの設定遅延が実現された場合、全体の遅延はそれぞれ 200 m (2 km) のリンク長で約 7 (25) μs に短縮される。上記の評価では、1 マイクロ秒未満のバーストモードコヒーレント検出器の遅延 [10]を含んでいない。データセンタ内ネットワークに用いる近距離伝送用簡易 DSP の最適化は今後の研究課題である。

(1) Measured connection set-up latency component

Connection setup process	Latency
100G Packet Transmission (FPGA1) and Receiving (FPGA2) Processing Time (excl. Transmission Delay)	0.26~0.32 μs
Ethernet Switch Latency (Baseline)/sw	0.40~0.42 μs
Optical Switch Connection Table Search	0.06 μs
Si Photonics Tunable Filter Setting Time (0-90%, transition)	5.75~6.1 μs (Typical)

(2) Analytically evaluated connection set-up latency component

Connection setup process	Latency
Optical Fiber Transmission Delay	Fiber Length [m] / 204.4 μs
Ethernet Switch Worst Queueing Delay: the delay where the probability that exceeds the value is 10^{-9} .	$D_9 = 0.251 \mu\text{s}$ (M/M/1 queueing model, link speed 100Gbps, packet length = 72B)

(3) Total connection latency for L [m]

$$= 16.9 + L [\text{m}] / 51.1 \mu\text{s} \text{ (three way hand shaking)}$$

$$= 5.4 + L [\text{m}] / 102.2 \mu\text{s} \text{ (one way simple procedure)}$$

図 1.2.4.1-20 接続セットアップ時間の内訳: (1) 実験的および (2) 解析的に評価された接続セットアップ時間

レイテンシは、データセンタのサービスにおける重要なパフォーマンスメトリックスの 1 つであり、レイテンシの最悪値は、全体的なパフォーマンスを決定する上で重要である。1 つのサービスが、全体のプロセスが並列化されたサブプロセスで構成されている場合、サービスパフォーマンスは、多くの異なるノード/サーバーにわたるサブプロセスの最後の応答によって決まる。多層の大規模な分散システムによって提供されるサービスの場合、ロングテールのレイテンシ特性が見られる。つまり、99 パーセントまたは 99.9 パーセントのレイテンシは、レイテンシの中央値よりも 1 桁以上大きくなる可能性がある。そして、これらの比較的まれなイベントが全体的なパフォーマンスを決定することになる。実際、ローカルデータセンタの場合、レイテンシの 50% と 99% 値は大きく異なり、たとえば、10 μs と数ミリ秒 (キャッシュされていない 1 バイトのアクセスのレイテンシ) [24] となる。Google での測定 (完了時間) 例では、単一のランダムリクエストが完了するまでの 99 パーセンタイル

レイテンシが 10 ミリ秒であっても、すべてのリクエストが完了するまでの 99 パーセンタイルレイテンシが 140 ミリ秒となることがあることを示している [25]。応答時間は、サービスクラスとキューイング優先処理の設定、ヘッドオブザラインブロッキングの低減、バックグラウンドアクティビティを管理することによって短縮できる[25]。パケット ネットワークのレイテンシに関しては、1MB のパケットバッファは 50 Gbps の光レーン速度で 160 μ s のレイテンシを導入し、3 層のイントラデータセンタネットワークの通信には最大 5 ホップの転送が必要となる。フローの遅延は、フローに含まれるパケット遅延の最大値によって決まる。回線交換ネットワークでは、パスにキューイングやパケット競合がないため、一旦確立された回線上を流れるパケットは同じ一定の遅延 (伝送遅延のみ、ジッターなし) となる。その結果、回線交換ネットワークのフローレイテンシは、ロングテール効果を持たず、接続セットアップ時間と伝送遅延によって決まる。したがって、上記の実験で示されているように、接続セットアップの遅延が 20 ~ 50 μ s と適度に小さく保たれている場合、最後のパケット到着によって決定されるパケットネットワークと比較して、フローの遅延を大幅に減らすことができる。このように、制御遅延が小さい光回線交換は、データセンタ内ネットワークの遅延性能を大幅に改善することができる。

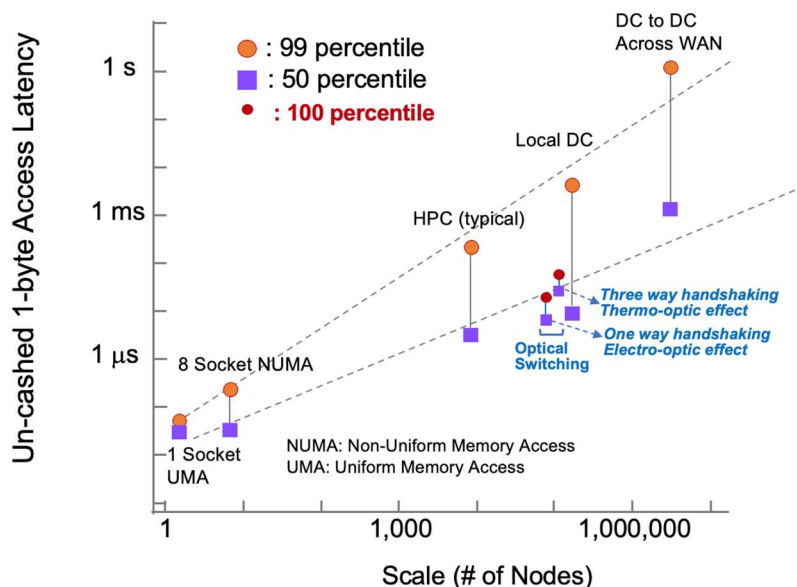


図 1.2.4.1-21 ネットワークスケール (ノード数) に対するレイテンシの依存性

レイテンシのロングテール効果は、ネットワークの規模が大きくなるにつれて拡大し、サービス品質を決定する重要な指標である。図 1.2.4.1-21 は、ネットワークスケール (構成ノード数) に対するレイテンシの例を示す [24]。ネットワークスケールは、物理ネットワークの規模も反映している。パーセント値が増加するにつれて、テールの長さが拡大される。レイテンシの 99 パーセント値は、HPC およびローカルなデータセンタの 50 パーセント値のレイテンシよりも 2 ~ 3 桁大きくなる傾向にある。光回線交換では遅延はコネクションセットアップのレイテンシで決まるほぼ一定値 (コントロールパケットのキューイング遅延は 0.25 μ s 程度と極めて小さい (図 1.2.4.1-20 参照)) である。つまり、コントロールネットワークの最大の通信距離により 100 パーセント値の遅延が決定される。光回線のセットアップレイテンシをスリーウェイハンドシェイク手順を用いた場合に、ファイバー長が

2 km (キャンパスネットワーク) の場合に発生する 100 パーセント値の遅延と、ファイバー長が 165 m (ハイパースケールデータセンタにおける平均ファイバー長) の場合に発生する 50 パーセント値の遅延について図 1.2.4.1-21 にプロットする。現在のローカルデータセンタの例と比較すると、光回線スイッチにおける遅延は、現在の電気スイッチネットワークの 99 パーセント値の遅延と比較しても大幅に (2 桁) 小さいことが分かる。一回のハンドシェイクで、さらに電気光学効果によるシリコンフォトリクスデバイスの制御を想定すると、同図に示すように、レイテンシパフォーマンスはさらに向上する。

<まとめ>

本課題で開発した光スイッチアーキテクチャは、広範囲に導入されている ROADM と多くの共通点があるが、データセンタのレイテンシ要求を十分カバーするためには、シリコンフォトリクスを利用した新しいデバイスが必要となる。パラダイムシフトを実現する上で、もう 1 つの重要な技術はネットワーク制御技術である。必要な特性を実現する上では、光スイッチの並列処理を最大限に活用することが重要であるが、これは Tier-1 (ToR) に使用する電気スイッチチップの最近の進歩とマッチしている。

本課題では、将来のハイブリッドスイッチングに基づくデータセンタ内ネットワークを構築する上で必要な一連のテクノロジーを開発した。本課題で開発したアーキテクチャは、100 Pbps を超える帯域への拡張性を有し、エネルギー効率の高い将来のデータセンタネットワークを実現する道を開くもので有る。AI や ML のワークロードの爆発的な増加と、そう遠くない将来に予想されるムーアの法則の終焉に起因する問題を解決するものである。本検討の結果、分散データセンタアーキテクチャへの適用も含め、異なるアプリケーション・規模 に応じた、変調形式・ポート数・空間スイッチ・波長スイッチ機能分担等の最適化が可能なアーキテクチャが実現された。また、試作したデバイスを用いた光スイッチ伝送実験により、将来的に 7,000 ポート以上に拡張可能なアーキテクチャ並びに現状のデータセンタの遅延特性を大幅に改善するコントロールネットワークの実現性が検証された。