ナノフォトニック・ニューラルアクセラレーション構想

川上 哲志[1,a)]　磯部 聖[2]　浅井 里奈[2]　小野 貴継[1]　本田 宏明[3]　井上 弘士[1]　納富 雅也[4,5]

概要：本稿では，ナノフォトニクスを用いたニューラルネットワーク向けアクセラレータの構想を提示する．ナノフォトニクス技術の発展に伴い，ナノフォトニック・デバイスは低遅延，小型，低消費エネルギー化が進められている．これまで，光デバイスは通信技術への応用が殆どであり，演算への応用事例は少ない．我々は，ナノフォトニック・デバイスの特徴に着目し，演算に適用する．これにより，CMOS を凌駕する演算性能を狙う．具体的には，ニューラルネットワークを対象とし，光学干渉機器を用いた行列積演算手法を応用することで，高性能かつ低消費電力なアクセラレータの実現を目指す．また，簡易的な性能モデルに基づいた性能評価によって，既存のニューラルアクセラレータに対して面積・電力効率で優位性があることを明らかにするとともに，今後の光コンピューティングの方向性を示す．

1. はじめに

情報処理通信技術の発展に伴い，組込みシステムからデータセンタ，スーパーコンピュータなど様々な階層で大量のデータが処理されるようになり，あらゆる社会システムの高性能化・効率化へ寄与している．特に，実世界から大量の情報を獲得し，高度な計算（分析・解析）に基づいて，再び実世界に情報をフィードバックするサイバーフィジカルシステム（CPS）は，医療・交通・金融といった様々な応用分野が期待され，今後もその応用範囲は拡大すると予想される．CPS の基盤となる計算機システムにおいては，大量のデータを低レイテンシで処理することが重要な課題の 1 つとなる．

近年の CMOS 技術に基づくプロセッサの性能向上は，主に並列化によるスループットの向上によって実現されており，CMOS 技術による計算処理のレイテンシの改善は困難である．さらに，今後の微細化による配線抵抗の顕在化 [4] や微細化そのものが限界を迎えることにより，劇的な低レイテンシ化を期待できない．

[1] 九州大学 大学院システム情報科学研究院
Faculty of Information Science and Electrical Engineering, Kyushu University
[2] 九州大学 大学院システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
[3] 九州大学 情報基盤研究開発センター
Research Institute for Information Technology, Kyushu University
[4] NTT ナノフォトニクスセンタ
NTT Nanophotonics Center
[5] NTT 物性科学基礎研究所
NTT Basic Research Laboratories
[a)] satoshi.kawakami@cpc.ait.kyushu-u.ac.jp

そこで，極低レイテンシを実現する手段として光技術が注目されている．光子は，電場や磁場の影響を受けにくいという性質を持つため，電子配線で生じる寄生抵抗や浮遊容量に依存することなく，光の速度での信号伝達が可能となる．光技術を応用した光コンピューティングの研究は，空間伝搬光を用いた超並列演算や CMOS の論理回路機能を模した光型トランジスタなど，80〜90 年代に活発に行われたが，集積度や消費電力で CMOS に対して大きく劣っており，90 年代後半には衰退してしまった．

しかしながら，近年のナノフォトニクス技術の発展により，フォトニック結晶に代表される光の伝搬を制御可能な物質の大規模集積化，小型化，低消費エネルギー化が可能になりつつある [11]．ナノフォトニック・デバイスは，かつては困難であった光の波長程度の極微小領域への光の閉じ込めを可能にすることが報告されており [17]，今後のさらなる発展が期待されている．これにより，ナノフォトニクス技術は，すでに実用化しているインターネット通信のみならず，CPU-メモリ間やコア間といった比較的短距離な通信にもその適応範囲を広げ，CMOS 技術に対する優位性を示している [2]．一方で，ナノフォトニクス技術の演算処理応用はこれまで殆どされておらず，現状では未知なる領域である．

そこで，ナノフォトニクス技術に基づいた光演算処理を含む光コンピューティング手法を提案し，CMOS 技術を凌駕する高性能・低消費電力な計算機システムの実現を目指す．我々は，ナノフォトニック・デバイスを用いた演算処理応用の 1 つとして，光パスゲート理論による低遅延回路を提案している [22]．これは，従来の電気回路をベース

に，演算のレイテンシに影響するクリティカルパスにナノフォトニクス技術を適用することで，光による極低レイテンシ性と電気による高スループットの両立を狙うものである．これに対し本研究では，ナノフォトニック・デバイスの性質に基づくアナログ行列演算方式に着目し，光演算処理の実現を目指す．光波の情報密度を最大限に活かすことのできるアナログ処理方式は，依然として CMOS 回路に対し集積度の劣るナノフォトニック・デバイスの欠点を補い，高い演算処理性能を達成できる有効な処理方式であると考える．本稿では，ナノフォトニック・デバイスを用いたアナログ行列演算の応用例としてニューラルネットワーク向けアクセラレータの構想を提示する．さらに，その性能モデルを作成することによって，CMOS 技術による既存のニューラルアクセラレータに対し，同程度の面積効率と劇的な電力効率が達成可能であることを明らかにする．また，これらの結果に基づいて光コンピューティングの展望を示す．

2. 過去の光コンピューティングとその問題

2.1 研究の変遷

現在，光通信はインターネット社会を支える極めて重要な要素技術として広く普及している．また，広域通信のみならず，計算ノード間接続やチップ内コア間接続（いわゆる Network-on-Chip）といったコンピュータ・システム内部の比較的狭域な通信路においても光の適用が進みつつある [1]，[18]，[19]．その一方，光の特性をデータ処理に利用する「光コンピューティング」の研究開発も盛んに行われた次期があったが，90 年代後半以降は衰退の一途を辿っているのが実状である．そこで本節では，これまでの光コンピューティング研究の変遷を鑑み，問題点を整理する[*1]．

光学に関する歴史は古いが，特に 1960 年代以降，レーザーの発明等もあり光による情報処理の研究開発が活発化した．1970 年代頃までには，レンズやフィルター，フォログラムなどの光学系機器を用いた光アナログ処理に関する研究が多く行われた．便宜上，本稿ではこれを**空間系光アナログコンピューティング**と呼ぶ．光波の伝搬・干渉・回折などの自然法則に基づく処理モデルであり，多次元フーリエ変換や多次元相関演算といった高機能性を有する．また，自由空間での光データ通信により入出力経路の配線制約を完全排除でき（光の 3 次元性），演算用光学系機器の搭載数を増加することで，空間方向にスケール可能な並列処理の実現が可能となる．このような高性能化アプローチを本稿では**スケールアウト型光並列処理**と呼ぶ．

空間系アナログコンピューティングでは高機能かつ高並列な情報処理を期待できる一方，演算精度保証の難しさ，反復処理の難しさ，操作性（プログラム可能性）の低さが問題であった．これらを解決すべく，また，電子式コンピュータにおけるデジタル化の流れも影響し，1980〜1990 年代にかけて**空間系光デジタルコンピューティング**に関する研究が盛んに行われた．光の強度値，偏光状態，波長，空間周波数などを利用して離散化されたビット情報を表現し，非線形光学素子や空間光変調素子を用いた光論理ゲート，投影光学系を用いた並列光論理演算，ビームスプリッタ/ホログラフィックフィルター等を用いた記号置換論理（ビットパタンの変換）などを用いて光演算処理を実現する．空間系光アナログコンピューティングと同様，2 次元空間に並列配置した光論理ゲートに対し光入力データを直接照射し演算することでスケールアウト型光並列処理が可能となる．応用としては，空間並列処理に相性の良いデジタル画像処理が主流であった．

また，1980 年代後半から 1990 年代にかけて，空間系光アナログコンピューティングの発展系として**空間系光ニューロコンピューティング**の研究開発も活発化した．自由空間を用いた高密度な光配線，高機能な光アナログ処理，さらには，スケールアウト型並列処理といった空間系光アナログコンピューティングの特徴を活用し，高スループットなニューラルネットワーク処理の実現を目指すものである．たとえば，空間光変調素子によりニューラルネットワークのシナプスを模倣し，シナプス荷重行列を電気信号として与える光電ハイブリッド方式による実現が報告されている．空間系光アナログ/デジタルコンピューティングと同様，スケールアウト型光並列処理を実現できるといった特徴がある．

2.2 問題点

これまでに光コンピューティングに関する多くの研究開発が行われてきたが，CMOS による電子式コンピュータを凌駕する性能を達成するには至っていない．その最大の原因は，現代の電子式コンピュータと同じ高性能化手法を指向するアーキテクチャ的アプローチにあると考える．空間系光アナログ/デジタルコンピューティングでは，処理機構を物理的に空間並列配置した SIMD 処理を高性能化の拠り所としている．一般に，SIMD 処理における実効性能は，「演算器の数（演算並列度）」とそれに見合った「入出力バンド幅」に依存する．スケールアウト型光並列処理では，自由空間データ通信により十分な入出力バンド幅を提供できる一方，演算素子のスケーリングには光学装置の大規模化や増加が必要となる．これは，小型化が求められる近年の情報処理システムにおいては，非常に厳しい要求となる．これに対し，電子式コンピュータにおいては，ムーアの法則に従って搭載する演算器数を順調に増加してきた．ここで，チップサイズ一定で搭載する演算器数を増大するアプローチを**スケールアップ型電子並列処理**と呼ぶ．また，メモリチャネル数の増加やメモリバス動作周波数の

[*1] 光コンピューティング研究開発の変遷については主に文献 [21] を参考にし整理した．

向上，階層メモリ構造の採用など様々な工夫により高い入出力バンド幅を実現している．その結果，スケールアウト型光並列処理とスケールアップ型電子並列処理の性能差は「演算並列度」に大きく依存する形となり，半導体集積度の向上を拠り所とする後者が圧倒的優位性を獲得したものである．これに加え，光アナログ処理から光デジタル処理への転換も負の要因である．集積度の観点では半導体に対し圧倒的な差が存在するにも関わらず，機能レベルを電子式コンピュータと同程度にまで低下させたため，光本来が有する利点を失う結果を招いた．その一方で，光の特性上，DRAM や SRAM といった半導体メモリのような高い制御性を持つ記憶素子の実現や，複雑な順序回路の実装は依然として難しく，光デジタル処理方式導入の狙いの 1 つであった制御容易性の向上には十分貢献できていない．以上を整理すると，演算精度や可制御性の向上を目的として光デジタルコンピューティングを導入した結果，電子式コンピュータと比較して演算粒度は同レベルとなり，かつ，実現可能な並列度に圧倒的な差が生じたことが，光コンピュータの可能性を制限している最大の問題である．

2.3 今後の方向性

光の利点を最大限に活用すると同時に欠点を隠蔽し，電子式コンピュータを凌駕する高い性能や電力効率を達成するには，デバイス/アーキテクチャレベルでのコデザインが必要不可欠である．特に以下に関する検討が重要となる．

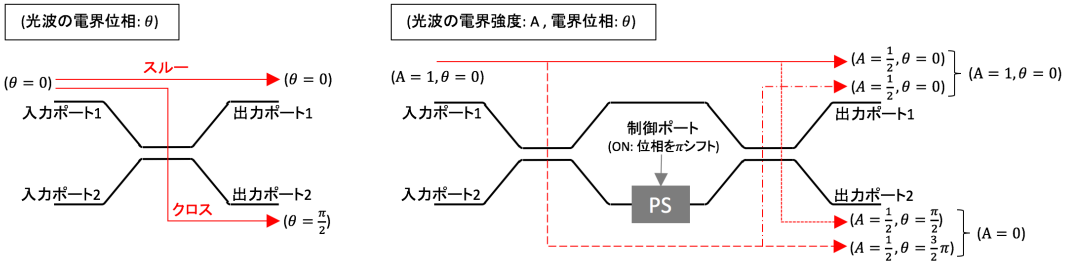
- **低レイテンシ集積系光デバイスの実現：**より高い光並列演算を実現するには空間系アプローチではなく，ナノフォトニクス技術を用いた低レイテンシな集積系デバイスが必要となる．電子式コンピュータは，2000 年以降，デナードスケーリングの破綻により消費電力問題が顕在化し，動作周波数が頭打ちとなっている．そのため，半導体の微細化に基づく高スループット化により並列処理性能は向上し続けている反面，逐次処理性能は停滞した状況にある．その一方，近年では CPS に代表される高度なリアルタイム処理の要望は年々高まっている．この要求に応える一つの解は，光速で情報処理可能なコンピューティング・プラットフォームを実現することである．
- **光アナログ処理の導入：**ナノフォトニクスの導入により集積化光コンピューティングが可能となるが，依然として電子式トランジスタの微細化に対しては 3〜4 桁の開きがあり，これはそのままスループットの差として顕在化する．このギャップを埋めるためには次元の異なる最適化が必要であり，その有効な手段として光アナログ処理による高機能演算の実現が挙げられる．
- **不完全計算モデルの導入：**第 2.1 節で述べたように，光アナログ処理では演算精度の低下が最大の問題となる．そこで，演算精度の低下を許容するコンピューティング・モデルの導入が必要不可欠となる．その代表例として第 2.1 節で述べた光ニューラルネットワークの活用が挙げられる．ただし，スケールアウト型光並列処理ではなく，ナノフォトニクスを用いた集積型の（つまり，スケールアップ型電子並列処理と同様の）光ニューラルネットワークの実現が必要不可欠となる．これに加え，演算精度の低下を許容する概略計算（Approximate Computing）用アクセラレーションと組み合わせることで，アプリケーションレベルでの演算精度保証とプログラム容易性（可制御性）の向上を狙う．Approximate Computing は，演算精度を犠牲にして演算処理量を削減することで性能向上や消費電力削減を達成する処理形態である．FFT，JPEG エンコーディング，k-means クラスタリングや sobel フィルタによるエッジ検出といった演算を Approximate Computing に適応する例が報告されている [10]．そこで，Approximate Computing の演算精度の劣化を許容可能な性質を利用し，光アナログ処理の欠点を隠蔽する．
- **極めて高い光入出力バンド幅の活用：**過去の光コンピュータと同様，光による高い入出力バンド幅（低レイテンシ化も可能）を有効活用すべきである．特に，光通信路に直結した光速処理機構を搭載することで，光電変換のオーバヘッドを隠蔽しつつ，極めて低いレイテンシでの情報処理が可能となる．すなわち，光通信中の情報処理（In-Optical-Network Computing）の実現である．
- **光演算における多重化技術の適用：**光通信技術で用いられる波長多重技術 (DWDM: Dense Wavelength Division Multiplexing) をナノフォトニック・デバイスによる演算処理へ適用することで，同時刻に同一デバイスで MIMD もしくは SIMD と同等の機能が実現可能となる．これは，異なる波長の光信号はお互いに干渉しないという性質を有効に活用しているため，従来の電気回路では考えられなかったことである．DWDM を含めた In-Optical-Network Computing によって，高いスループットの実現を目指す．

3. ナノフォトニック・デバイス

3.1 基本素子

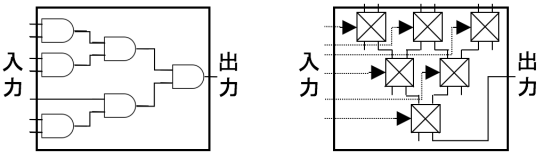
方向性結合器

方向性結合器（DC: Directional Coupler）とは，1 つの光の伝達経路（以下，導波路）からの光信号を 2 つの導波路に分岐したり，あるいは 2 つの導波路からの光信号を 1 つの導波路に結合する機能を有するデバイスである．**図 1** (a) のように 2 つの導波路を十分に近い距離で平行に並べると，光波は 2 つの導波路間を移動する．入力ポート 1 に入力された光波が，他方の出力側導波路である出力ポート 2 に移動することをクロス，



(a) 方向性結合器　　(b) マッハツェンダ干渉器

図 1: ナノフォトニック・デバイス.



(a) CMOS論理の例　　(b) パスゲート論理の例

図 2: CMOS 論理とパスゲート論理.

同一の出力側導波路である出力ポート 1 を通過することをスルーと呼ぶ．入力光のエネルギーに対するクロスした光波のエネルギー比率を透過率，スルーした光波のエネルギー比率を反射率と呼び，これらの比率は 2 本の並行する導波路の長さと結合係数に依存して決定される．また，導波路型の DC においては，クロスへの透過で位相が $\pi/2$ シフトするという特徴を持つ．

マッハツェンダ干渉器

マッハツェンダ干渉器 （MZI: Mach-Zehnder Interferometer）の概略を図 1 (b) に示す．導波路型 MZI の回路は，図 1 (a) の DC と位相シフタ（PS: Phase Shifter）から構成される．PS では，制御ポートからの信号によって任意の位相変化が可能であり，これによって生じた位相差に従って出力光の強度を変化させるという特徴を有する．たとえば，DC の反射率ならびに透過率が 0.5 であり，PS は制御信号により位相を π シフトする（ON 時）もしくは位相シフトしない（OFF 時）という 2 値制御が可能な構成を考える．この際，入力ポート 1 から入力された光波は，PS が ON 時に出力ポート 1 から出力され，逆に OFF の時には出力ポート 2 から出力される経路切り替えスイッチとしての機能を果たす．これは，図 1 (b) に示すように，2 つの DC によって分割された光波が出力ポートにて強め合う（もしくは，弱め合う）ことで実現されている．

3.2 光パスゲート理論とその応用

ナノフォトニック・デバイスを用いた演算処理応用の 1 つとして，光パスゲート理論による低遅延回路 [22] の概要について述べる．**図 2** (a) に示す CMOS 論理では，各論理ゲートの切り替え時間が入力信号が出力されるまでのゲート段数に応じて増加する．光パスゲート理論とは，図 2 (b) に示すように，MZI に代表される光信号経路切り替えスイッチを直列接続することにより，任意のブール演算を実現するものである．したがって，入力信号は経路切り替え制御信号として入力されるため，全ゲートの切り替えが同時に行われる．パスゲート理論は一部の CMOS 回路でも用いられていたが，伝搬遅延が段数の二乗に比例する性質から汎用的な回路に至らなかった．しかしながら，光回路における伝搬遅延は段数 × 素子長に比例するため，光回路とパスゲート理論は親和性が非常に高い．

この光パスゲート理論による低遅延回路として，加算器を提案する [22]．加算器においては，ある桁の計算のためには，下位の全ての桁の入力から桁上げ信号を計算する必要があるため，桁上げ回路の遅延が支配的となる．そこで，加算器のクリティカルパスとなる桁上げ回路をナノフォトニック・デバイスのみで実現する．これにより，同機能の CMOS 回路と比較して約 10 倍高速に動作することが示されている．

4. ナノフォトニック・ニューラルアクセラレーションの実現

4.1 ニューラル・アクセラレーションの現状と問題点

ニューラルネットワークは，学習による情報の獲得により任意の連続関数を任意の精度で近似できる情報処理技術 [9] で，ノイマン型コンピュータでの処理が困難であった分野に対して有効な手段として注目されている．ニューラルネットワークは，処理手順（アルゴリズム）そのものが明確でない対象にも適用可能であるが，このような領域においては，（1）十分な量のデータと（2）十分な計算機性能の双方を活用できることが重要である．これら 2 つの条件を満たす計算機基盤と適用範囲に限り，ニューラルネットワークは人間並みの精度を持つ帰納推論が可能となる．たとえば，画像認識や音声認識に代表されるマルチメディア処理では，劇的な認識精度向上を達成している [13], [14].

ニューラルネットワーク処理を高速化する計算機アーキ

テクチャ（ニューラル・アクセラレーション）の研究は，その歴史も古く多角的に行われている．これらの先行研究において，基盤となる計算機システムに対する要件は，高メモリバンド幅，高スループット，低消費電力と共通している．すなわち，前述の（1），（2）の要件に対応すべく，現在のニュラル・アクセラレーションの研究が行われている．具体的には，コア間のデータのやり取りに要する時間と電力の消費が大きいことに着目し，各コアに隣接したローカルメモリを介してコア間通信を可能にした，ディープラーニング向けのアクセラレータが提案されている [6]．データ通信のコストを最重要に考え，ニューラルネットワークにおける各データの再利用性を上げた際の性能向上について評価を行っている．DaDianNao [5] は，ディープラーニング向けの ASIC であり，劇的な電力効率の改善を達成している．しかしながら，実行可能なニューラルネットワークは ASIC に搭載された eDRAM の 36MB に限られる．ISAAC [15] や PRIME [7] では，メモリインテンシブな特徴を持つニューラルネットワーク処理を高速化するため，メモリスタや ReRAM を利用しデータ読み出しと同時にアナログ演算を実行するアーキテクチャを提案している．

このように，既存のニューラルアクセラレーション研究は，データの圧縮や再利用によってメモリアクセスを如何に削減または効率化するか，専用回路やアナログ演算によって如何に性能向上を達成するかに着目している．より高度な識別・認識を達成すべくニューラルネットワークの層数・ノード数は増加傾向にあり，計算機システムに対するメモリバンド幅や性能の要件はより厳しくなることが予想される．たとえば，現在のニューラルネットアプリケーションは数百万のノード（ニューロン）を有するものもあるが，単純に人間の脳全体の数百億個のニューロンを達成するには最低でも 100 万倍のデータと演算が必要になる．したがって，ニューラルネットワークアクセれレーションにおいては，メモリバンド幅とスループットの改善，ならびに，低消費電力化が重要な課題となる．

4.2 ナノフォトニック・ニューラルアクセラレーション

ナノフォトニクス技術を用いた計算機基盤は，ニューラルネットワーク処理の高性能化のための有力な候補の 1 つである．なぜなら，第 4.1 節で述べたニューラル・アクセラレーションの課題であるメモリボトルネックの改善と高性能化を同時に達成できる可能性を有するためである．ニューラルネットワークに発展に伴い，扱うユニット数が増大すれば，現在のようなオンチップメモリに収まる範囲での処理形態は実行不可能であり，オフチップメモリアクセスが必須となる．その際，大容量性と高速性を兼ね備える光通信技術が活用されると同時に，光信号を電気へ変換することなくそのまま演算を実行可能なナノフォトニック・ニューラルアクセラレーションは大きな優位性を持つ

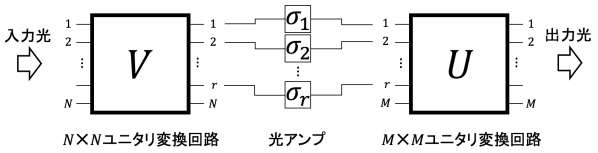


図 3: 特異値分解による行列演算光回路の構成図．

ことが予想される．ニュラルネットワーク処理の根幹となる行列積は，MZI を組み合わせる事によりアナログ的に演算可能となる．詳細は第 4.3 節で述べるが，基本は MZI で構成されるため，回路遅延は通信路と変わらず素子長で決定される．したがって，光の速さ（正確には，真空中における光速/屈折率）での行列演算が可能となる．また，アナログ演算は光波の情報密度を最大限に活かすことができるので，素子あたりの演算機能性が高く集積度が比較的低いナノフォトニック・デバイスにおいて有効な処理形態である．一方で，アナログ演算では計算結果の精度が重要な課題となるが，ニューラルネットワークに代表される Approximate Computing においてはある程度の精度劣化が許容できるので，親和性の高い応用例だと考えられる．

ニューラルネットワークの出力は，入力ベクトルに対して行列積を行い，その結果に活性化関数を適用することで得られる．すなわち，ニューラルネットワークの処理はベクトル行列積に関する処理と活性化関数に関する処理とに大別することができる．ナノフォトニック・デバイスを用いた回路において任意のベクトル行列積は，第 4.3 節で述べるユニタリ変換回路と特異値分解を活用することによって実現できる．特異値分解により，任意の $M \times N$ 行列 A を次式の通り分解できる．

$$A = U\Sigma V \tag{1}$$

ここで，U は $M \times M$ ユニタリ行列，V は $N \times N$ ユニタリ行列を表す．また，Σ は $M \times N$ 行列であり非対角要素は 0，かつ，対角要素は非負で降順の特異値（$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$）を持つ行列である．$r$ は A のランクに等しい．したがって，光アンプとユニタリ変換回路を**図 3** の通りに構成することで，任意のベクトル行列演算が可能となる．ニューラルネットワークにおける活性化関数は，単調増加する非線形関数が一般的に用いられている．そこで，このような特性を有する可飽和吸収体や双安定素子を用いる事によって活性化関数が実現可能であると考える．可飽和吸収体は，強度が小さい入力光に対しては吸収体として働き，強度が大きい入射光に対しては透明体として働く物体である．また，双安定素子は入力光の強度に応じて出力光の強度が異なった 2 つの安定値になるヒステリシス特性をもつ素子である．以上により，ニューラルネットワークにおける基本的な処理は，ナノフォトニック・デバイスによって実現することが可能である．なお，本稿では MZI を用いた光アナログ演算器に着目するが，可飽和吸収体を用いた活性化

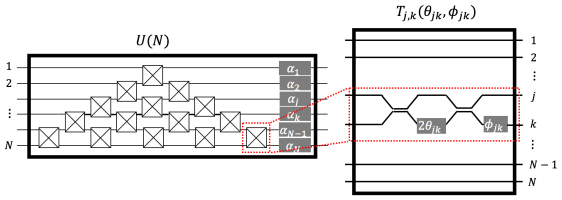


図 4: MZI によるユニタリ行列 U と $T_{j,k}(\theta_{jk},\phi_{jk})$ の構成図.

関数の実現に関する検討については文献 [20] を参照されたい.

4.3 MZI によるユニタリ行列演算

本節では，MZI 回路による任意の $N\times N$ ユニタリ変換を実現するための MZI 構成法の概略についてまとめる [12]. 第 4.1 節で説明した MZI では，PS の位相操作により入力光波のエネルギーを任意の比率で出力ポートに分配することが可能であり，透過した光波のユニタリ変換を行うアナログ回路とも見なされている．この MZI における N 個からなる入出力ポート内の光波をそれぞれ N 次元複素ベクトルと考え，入力光に MZI を作用させることを入力複素ベクトルに対し MZI で定義されたユニタリ行列との行列積を計算するをみなすことで，このユニタリ変換と MZI による計算操作との関係付けが可能である.

ユニタリ変換回路の基本ゲートとして，図 1 (b) で示した回路の片側出力ポートに PS を追加した素子を考える（**図 4** の赤点線枠）．なお，2 つの DC の反射率と透過率はともに 0.5 である．この基本ゲートを図 4 のように三角形状に接続することで任意の $N\times N$ ユニタリ行列 $U(N)$ を表現することが可能となる．したがって，$N\times N$ のユニタリ変換回路を構成するためには，$N(N-1)/2$ 個の基本ゲートが必要となる.

ユニタリ変換回路の原理説明のため，図 4 右に示すように基本ゲートを N ポートに拡張したユニタリ行列 $T_{j,k}(\theta_{jk},\phi_{jk})$ を導入する．$T_{j,k}(\theta_{jk},\phi_{jk})$ は，入出力ともに N ポートを有するスイッチであるが，j，k 番目のポートが基本ゲートのポートに対応し，それ以外の入力ポートはそのまま出力ポートへと接続される．$T_{j,k}(\theta_{jk},\phi_{jk})$ の伝達行列は次式で示される.

$$T_{j,k}(\theta_{jk},\phi_{jk}) =$$

$$\begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & 0 & \\ & & e^{i\phi_{jk}}\sin\theta_{jk} & \cdots & e^{i\phi_{jk}}\cos\theta_{jk} & & \\ & & \vdots & \ddots & \vdots & & \\ & & \cos\theta_{jk} & \cdots & -\sin\theta_{jk} & & \\ & 0 & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ j\text{ 行} \\ \\ k\text{ 行} \\ \\ \end{matrix}$$

j 列　　k 列

ここで，θ_{jk} と ϕ_{jk} はそれぞれ PS での位相差を示し，θ_{jk} は基本ゲートにおける反射率 R ($\sqrt{R}=\sin\theta_{jk}$) と透過率 T ($\sqrt{T}=\cos\theta_{jk}$) を決定するパラメータとなる．$T_{j,k}(\theta_{jk},\phi_{jk})$ は，恒等行列の要素 $I_{jj},I_{jk},I_{kj},I_{kk}$ を 2×2 の MZI の伝達行列の要素で各々置き換えたものに等しく，ユニタリ行列 $U(N)$ と次の関係式が成り立つ.

$$U(N)\cdot\prod_{k=N-1}^{1}T_{N,k}(\theta_{Nk},\phi_{Nk}) = \begin{pmatrix} \boxed{U(N-1)} & 0 \\ 0 & e^{i\alpha_N} \end{pmatrix} \tag{2}$$

式 (2) に随伴行列 $(\prod_{k=N-1}^{1}T_{N,k}(\theta_{Nk},\phi_{Nk}))^{\dagger}$ を右からかけることにより，$U(N)$ の N 行目の行ベクトル $\langle N|$ は，次式で示される.

$$\langle N| = \begin{pmatrix} e^{-i(\alpha_N+\phi_{N1})}\cos\theta_{N1} \\ -e^{-i(\alpha_N+\phi_{N2})}\cos\theta_{N2}\sin\theta_{N1} \\ \vdots \\ (-1)^N e^{-i(\alpha_N+\phi_{NN-1})}\cos\theta_{NN-1}\sin\theta_{NN-2}\ldots\sin\theta_{N1} \\ (-1)^{N+1}\sin\theta_{NN-1}\ldots\sin\theta_{N1} \end{pmatrix}^{T} \tag{3}$$

式 (3) により，任意の $U(N)$ に対して $\langle N|$ の各パラメータ (θ,ϕ,α) を決定することができる．さらに，式 (3) で求めたパラメータと式 (2) により，再帰的に $\langle N-1|$ のパラメータを決定することが可能となる．すなわち，$U(N)$ は次式で示される.

$$U(N) = (\prod_{j=N}^{2}\prod_{k=j-1}^{1}T_{j,k}(\theta_{jk},\phi_{jk})\cdot D(N))^{\dagger} \tag{4}$$

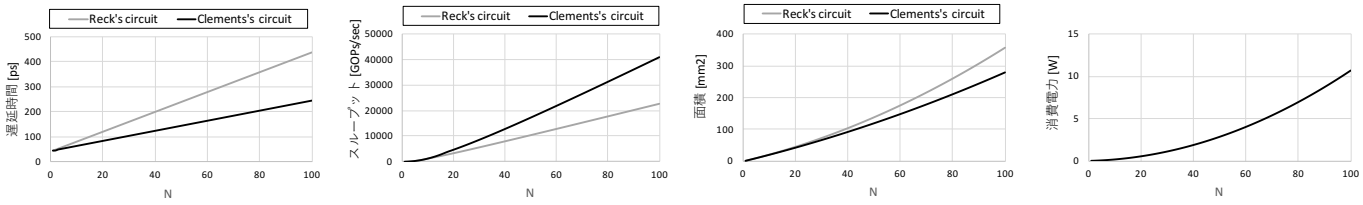
ただし，$D(N)=diag(e^{i\alpha_1},e^{i\alpha_2},\cdots,e^{i\alpha_N})$ である．$T_{j,k}(\theta_{jk},\phi_{jk})$ と $D(N)$ はユニタリ行列であるため，その随伴行列は各デバイスの出力側から入力光波を入力した際の伝達行列と等しくなる．すなわち，式 (4) にしたがい，$T_{j,k}(\theta_{jk},\phi_{jk})$ と $D(N)$ に対応するデバイスを逆向きで直列接続するとその系の伝達行列は $U(N)$ となる.

図 4 で示すユニタリ変換回路は，6 ポートの回路が実装されており，ランダムなユニタリ行列に対して $99.9\pm0.1\%$ の忠実度を達成している [3]．さらに，図 4 の回路と比較し，実装面積が小さく，入力光波の通過する最長経路が短く，精度が高い四角形状のユニタリ変換回路実装方式も提案されている [8].

5. モデルに基づくナノフォトニック・ニューラルアクセラレータの性能推定

5.1 性能モデリング

本節では，ナノフォトニック・ニューラルアクセラレータの遅延時間 L[s]，スループット T[OPs/sec]，面積 S[m^2]，消費電力 P[W] の近似モデルを作成し，既存のニューラルアクセラレータに対する優位性を明らかにする．ナノフォトニック・ニューラルアクセラレータにおける遅延時間は，



(a) 遅延時間　(b) スループット　(c) 面積　(d) 消費電力

図 5: $N \times N$ 次元行列-ベクトル積を前提としたナノフォトニック・ニューラルアクセラレータの性能評価

表 1: 性能モデルのパラメータ一覧

L_{MZI} [ps]	2×2 MZI のゲートパス時間	1
L_{AMP} [ps]	$20dB$ の光アンプのパス遅延時間	20
L_{SA} [ps]	SA のパス遅延時間	0.1
L_{PD} [ps]	受光器の応答時間	25
f_{PS} [GHz]	MZI 内の PS のスイッチング周波数	12.5
f_{PD} [GHz]	受光器のスイッチング周波数	40
S_{LS} [μm^2]	光源の面積	1000
S_{AMP} [mm^2]	$20dB$ の光アンプの面積	2
S_{SA} [μm^2]	SA の面積	100
S_{PD} [μm^2]	受光器の面積	1000
W_{MZI} [μm]	MZI のゲート幅（長）	100
D_{MZI} [μm]	MZI のゲート奥行	40
P_{PS} [mW]	MZI 内の PS の消費電力	0.5
P_{SA} [mW]	全入力の透過率が 50%の SA の消費電力	0.02
P_{AMP} [mW]	$20dB$・変換効率 50%の光アンプの消費電力	8

入力光が入射されてから図 3 に示す行列変換回路と可飽和吸収体 (SA: Saturable Absorber) に代表される非線形応答素子を通過し，受光器で検出されるまでの時間とする．光波は，伝達経路の屈折率と経路長によって遅延時間が決定する．本稿では，導波路中の平均屈折率を 3（すなわち，導波路中の光信号の伝達速度は $1 \times 10^8 [m/s]$）とする．また，各光素子に関しては**表 1** に示すパラメータ値を用いる．このとき，遅延時間は次式で示される．

$$L = L_{MZI} \times (LP_{MZI}(N) + LP_{MZI}(M)) + L_{AMP} + L_{SA} + L_{PD} \quad (5)$$

$$LP_{MZI}(N) = \begin{cases} 2N-3 & (Reck's\ circuit) \\ N & (Clements's\ circuit) \end{cases}$$

ただし，$LP_{MZI}(N)$ は $N \times N$ ユニタリ変換回路の最長経路上に存在する MZI 数を示す関数である．また，Reck ら [12] の回路と Clements ら [8] の回路によって最長経路が異なるため各々の $LP_{MZI}(N)$ を示す．

スループットは一秒間あたりの積和演算 (MAC: multiply-accumulate) 数とし，$M \times N$ 次元行列-ベクトル積回路を有するナノフォトニック・ニューラルアクセラレータにおいては式 (6) で表される．

$$T = M \times N \times min(f_{PS}, f_{PD}, 1/L) \quad (6)$$

回路の周波数は，MZI 内の位相シフタのスイッチング時間，受光器のスイッチング時間，式 (5) で示す回路遅延の何れかに律速される．

回路面積は配線を除く各素子の合計とし，ユニタリ変換回路面積は Reck ら [12] と Clements ら [8] の両方を示す．また，光アンプ数は，演算を行う $N \times M$ 行列のランクに依存するが，最大ランク数（N もしくは M の最小値）を想定し式 (7) で示す．

$$S = S_{MZI}(N) + S_{MZI}(M) + S_{LS} \times N + S_{AMP} \times min(N, M) + S_{SA} \times M + S_{PD} \times M \quad (7)$$

$$S_{MZI}(N) = \begin{cases} W_{MZI} \times (2N-3) \times D_{MZI} \times (N-1) & (Reck's\ circuit) \\ W_{MZI} \times N \times D_{MZI} \times (N-1) & (Clements's\ circuit) \end{cases}$$

消費電力は，位相シフタと SA での電力損失ならびに光アンプでの利得により決定される．実際の回路においては，各素子に入射される光波のエネルギーや行列計算値によって消費電力は変動するが，本稿では簡単のため各素子への入射光の電力は 0.04[mW]，アンプは常に最大利得での増幅を行うものとする．

$$P = 2 \times P_{PS} \times (\frac{N \times (N-1)}{2} + \frac{M \times (M-1)}{2}) + P_{SA} \times M + P_{amp} \times min(N, M) \quad (8)$$

5.2 性能評価

本節では，第 5.1 節で述べた性能モデルと表 1 のパラメータを用いて，ニューラルアクセラレータの面積あたりの積和演算処理数（面積性能）ならびに消費電力あたりの積和演算処理数（電力効率）の評価を行う．なお，式 (5) から式 (8) は $N \times M$ 次元行列-ベクトル積可能な回路の性能モデルであるが，本節では $N \times N$ 次元行列-ベクトル積を前提とする．

図 5 に N（横軸）に対する遅延時間，スループット，面積，ならびに，消費電力（縦軸）の変化を示す．遅延時間は，光信号が通過する経路長によって決定されるため，N に対して線形に増加している（図 5 (a)）．また，Clements らの回路構成は，光信号の最長伝達経路を短くすることが可能なため Reck らの回路に対して遅延時間が小さくなる．

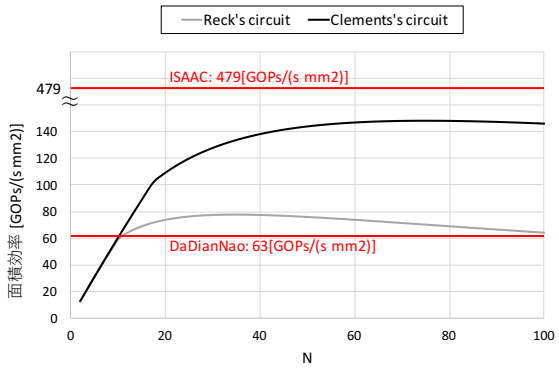


図 6: 性能モデルに基づく面積効率.

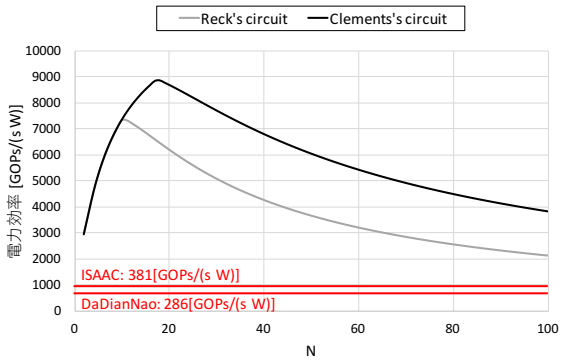


図 7: 性能モデルに基づく電力効率.

スループットは，N が小さい領域においては，N^2 に比例して増加する一方，N が一定数を超えると回路の遅延時間が増加し周波数が律速されるため，スループットは線形増加となる．具体的には，Reck らの回路においては $N = 11$，Clements らの回路においては $N = 18$ 以上の領域において，スループットは線形増加となる（図 5 (b)）．面積は，N の増加に応じて入出力ポート数と MZI の段数がともに増えるため N^2 で拡大する（図 5 (c)）．消費電力に関しても，面積と同様に回路を構成するナノフォトニック・デバイスの増加に応じて N^2 で拡大する（図 5 (d)）．また，行列演算回路を構成するためにに必要な MZI 数は，Reck らの回路と Clements らの回路で等しいため，消費電力は等しくなる．

図 6 と**図 7** に N(横軸) に対する面積効率と電力効率 (縦軸) を示す．Clements らのユニタリ変換回路は面積効率と電力効率の両方において Reck らの回路よりも優位であることがわかる．前述のように，N が増加すると処理可能な Operation 数は N^2 で増加するが回路遅延に応じて周波数が低下するため，スループットは N のオーダーで増加する一方で，面積と消費電力は N^2 のオーダーで増加する．したがって，N が十分に大きな領域においては，面積効率・電力効率ともに低下傾向となる．しかしながら，N の小さな領域 (面積効率は $N = 35[Reck'scircuit]$，$N = 75[Clements'scircuit]$，電力効率は $N = 11[Reck'scircuit]$，$N = 18[Clements'scircuit]$) では，回路の周波数は位相シフタのスイッチング周波数で決定され，スループットの改善率が大きいため．面積性能・電力効率ともに増加していることがわかる．また，図 6 と図 7 に CMOS のニューラルアクセラレータである DaDianNao [5] と ISAAC [15] の性能を各々示す．ナノフォトニックニューラルアクセラレータは，CMOS 回路に対する集積度の低さゆえ面積効率で ISAAC に劣るものの DaDianNao よりも高い面積効率を達成できる．さらに，ナノフォトニックデバイスの電力損失の少なさと極低レイテンシ性によって電力効率では両者に対して大きな優位性がある．

5.3 光コンピューティングの今後

第 4.2 節で述べた回路と同構成のナノフォトニック・ニューラルアクセラレータは文献 [16] でも概要が記されている．ナノフォトニック・ニューラルアクセラレータは，第 2.2 節で述べた「**低レイテンシ集積系光デバイス**」「**光アナログ処理**」「**不完全計算モデル**」を満たすものであり，本稿では性能モデルを用いることで電力効率に関して圧倒的な優位性があることが明らかとなった．また，ナノフォトニック・デバイスは単一素子サイズで比較すると CMOS トランジスタの微細化に対して 3 桁以上の開きがあるにもかかわらず，面積効率も同程度であることが判明した．さらに，第 2.2 節で述べた波長多重技術を用いることで，理想的には面積ならびに消費電力は一定のまま，波長多重度に比例してスループットの向上が可能となるため，さらなる面積・電力効率の向上が期待できる．

現在の計算機の演算処理は，CMOS による電子式コンピュータを基本とし構築されている．したがって，光コンピュータがこれに取って代わる，もしくは部分的に光コンピューティングを導入するには，情報媒体としての電気信号を光信号へ変化させる必要もあるため，圧倒的な性能改善を見込めることが前提となる．通信技術に着目して計算機基盤を俯瞰すると，ナノフォトニクス技術は，比較的短距離な通信にもその応用範囲を広げている（**図 8**）．たとえば，文献 [1] では NoC によるコア間と CPU-DRAM 間を波長多重の光通信で実現することで，高バンド幅・低消費電力を達成している．また，文献 [19] でもコア-L2 キャッシュ間と CPU-3 次元積層メモリの通信を光化する提案がなされている．このような，あらゆる通信が光によってなされ，演算が従来の電気回路で行われる計算機環境においては，ナノフォトニック・ニューラルアクセラレータは光電変換のオーバヘッドを隠蔽する事が可能になる．さらに，光パスゲート理論に基づいた演算回路は，前述のように配線長によって遅延が決定される．これは，従来の通信経路をナノフォトニック・アクセラレータで置き換えた場合で

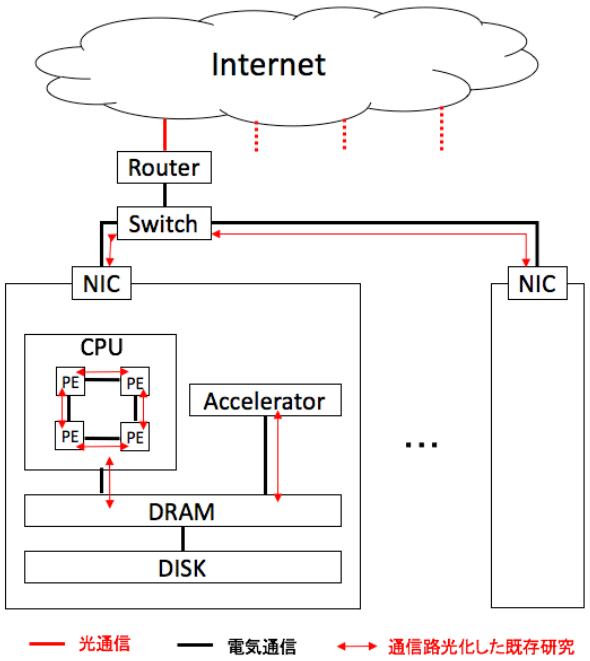


図 8: 計算機システムへのナノフォトニクス技術の導入.

も，理想的にはレイテンシが変わらないことを意味する．加えて，図 4 で構成される回路はスイッチとしての機能も果たすため，演算が必要でないデータに対してはただの通信路として振る舞うことができる．したがって，データの移動と同時に情報処理を実行すること（In-Optical-Network Computing）が可能である．In-Optical-Network Computing において，ナノフォトニック・アクセラレータを用いることを想定すると，図 6，図 7 で示すアクセラレータそのものの優位性を利用できるだけに留まらず，波長多重化や光電変換不要といった更なる性能向上の機会を獲得できる．よって，In-Optical-Network Computing の応用を拡充することがナノフォトニック・デバイスの特性を活かした光コンピューティングの有効な活用法である．

6. おわりに

本稿では，ナノフォトニクス・ニューラルアクセラレータの構想を提示し，その近似モデルによって CMOS 技術を用いた既存のニューラルアクセラレータに対する面積・電力性能の優劣を明らかにした．具体的には，面積効率は CMOS 回路におけるアナログ演算を活用したアクセラレータに対して劣るものの，電力効率では圧倒的な優位性があることが判明した．さらに，古くからの光コンピューティングの歴史，ならびに，現在のナノフォトニック・デバイスの研究を整理し，今後の光コンピューティングの将来展望である In-Optical-Network 構想を示した．

今後，ナノフォトニック・ニューラルアクセラレータの精度と詳細な消費電力を実機で計測し，計算精度も考慮した実現可能性について検討する．また，光 NN 向けメモリシステムや CPU インタフェースを含むシステムアーキテクチャの考案，プログラミングモデルの検討，ならびに，試作チップやシミュレータを用いた詳細評価を行い，提案する光ニューラル・アクセラレーションの有効性を実証する。さらに，In-Optical-Network Computing の応用拡充のため，ナノフォトニック・デバイスによる様々な演算手法についても検討する予定である．

謝辞 本研究の一部は，科学技術振興機構の戦略的創造研究推進事業「新たな光機能や光物性の発見・利活用を基軸とする次世代フォトニクスの基盤技術」の助成により行われた．

参考文献

[1] C. Batten, A. Joshi et al.: Building Many-Core Processor-to-DRAM Networks with Monolithic CMOS Silicon Photonics, *IEEE Micro*, Vol. 29, No. 4, pp. 8–21, 8 (2009).
[2] S. Beamer, C. Sun et al.: Re-architecting DRAM Memory Systems with Monolithically Integrated Silicon Photonics, *Proceedings of the 37th Annual International Symposium on Computer Architecture (ISCA)*, pp. 129–140, 6 (2010).
[3] J. Carolan, C. Harrold et al.: Universal linear optics, *Science*, Vol. 349, No. 6249, pp. 711–716, 8 (2015).
[4] A. Ceyhan, and A. Naeemi: Cu interconnect limitations and opportunities for SWNT interconnects at the end of the roadmap, *IEEE Transactions on Electron Devices*, Vol. 60, No. 1, pp. 374–382, 8 (2013).
[5] Y. Chen, T. Luo et al.: Dadiannao: A machine-learning supercomputer, *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 609–622, 12 (2014).
[6] Y. H. Chen, T. Krishna et al.: Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks, *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 262–263, 2 (2016).
[7] P. Chi, S. Li et al.: PRIME: A Novel Processing-In-Memory Architecture for Neural Network Computation in ReRAM-based Main Memory, *Proceedings of the 43th Annual International Symposium on Computer Architecture (ISCA)*, pp. 27–39, 6 (2016).
[8] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley: An Optimal Design for Universal Multiport Interferometers, *arXiv preprint arXiv:1603.08788* (2016).
[9] G. Cybenko: Approximation by superpositions of a sigmoidal function, *Mathematics of control, signals and systems*, Vol. 2, No. 4, pp. 303–314, 2 (1989).
[10] H. Esmaeilzadeh, A. Sampson, L. Ceze, and D. Burger: Neural Acceleration for General-Purpose Approximate Programs, *Proceedings of the 45th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 449–460, 12 (2012).
[11] M. Notomi, K. Nozaki, A. Shinya, S. Matsuo, and E. Kuramochi: Toward fJ/bit optical communication in a chip, *Optics Communications*, Vol. 315, pp. 3–17, 3 (2014).
[12] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani: Experimental realization of any discrete unitary operator, *Physical Review Letters*, Vol. 73, pp. 58–61, 7 (1994).
[13] G. Saon, H. K. J. Kuo, S. Rennie, and M. Picheny: The ibm 2015 english conversational telephone speech recognition system, *arXiv preprint arXiv:1505.05899* (2015).

[14] F. Schroff, D. Kalenichenko, and J. Philbin: Facenet: A unified embedding for face recognition and clustering, *arXiv preprint arXiv:1503.03832*, (2015).
[15] A. Shafiee, A. Nag et al.: ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars, *Proceedings of the 43th Annual International Symposium on Computer Architecture (ISCA)*, pp. 14–26, 6 (2016).
[16] Y. Shen, S. Skirlo, M. Soljacic, D. R. Englund, and N. Harris: On-Chip Optical Neuromorphic Computing, *CLEO: Science and Innovations*, pp. SM3E–2, (2016).
[17] Y. Takahashi, Y. Inui, M. Chihara, T. Asano, R. Terawaki, and S. Noda: A micrometre-scale Raman silicon laser with a microwatt threshold, *Nature*, Vol. 498, No. 7455, pp. 470–474, 6 (2013).
[18] M. A. Taubenblatt: Optical interconnects for high-performance computing, IEEE/OSA Journal of Lightwave Technology, vol. 30, no. 4, pp. 448–457, 2 (2012).
[19] D. Vantrease, R. Schreiber et al.: Corona: System implications of emerging nanophotonic technology, *Proceedings of the 35th Annual International Symposium on Computer Architecture (ISCA)*, pp. 153–164, 6 (2008).
[20] 磯部聖，川上哲志，小野貴継，井上弘士，納富雅也：可飽和吸収体の利用を前提としたナノフォトニック・ニューラルアクセラレータ向け活性化関数の評価，情報処理学会研究報告（デザインガイア），Vol. 2016-ARC, 11 (2016).
[21] 稲場文男，一岡芳樹：光コンピューティングの事典「普及版」，朝倉書店 (2006).
[22] 石原亨，新家昭彦，井上弘士，野崎謙悟，納富雅也：光パスゲート論理に基づく並列加算回路の提案と光電混載回路シミュレータによる動作検証，Vol. 116, No. 94, pp. 109–114, 6 (2016).