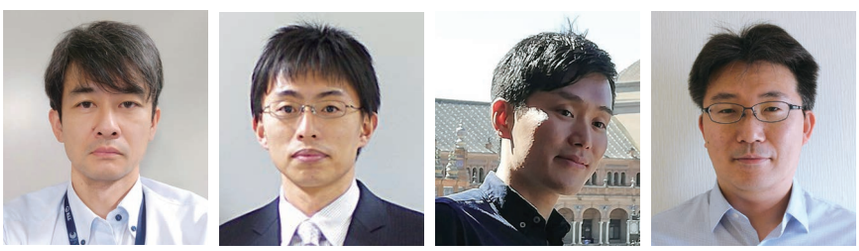
2 ポストムーア時代 / 光インターコネクト / 分散処理コンピューティング

高速光インターコネクトがこれからのコンピュータを変える

光電子融合研究部（融合部）では、ポストムーア時代のコンピュータ技術について分散処理コンピューティングを目指して、高速光インターコネクトの研究開発を行っている。ビッグデータの時代に、AI 活用は必須であるが、NTT としてはこれまで培ってきた超高速光ネットワーク通信技術を活用することで、ポストムーア時代のコンピューティング技術に対応する。



NTT 先端集積デバイス研究所　光電子融合研究部
（左から）部長　坂本　健氏
研究主任　有川 勇輝氏
研究員　田仲 顕至氏
研究主任　伊藤　猛氏

ポストムーア時代のコンピューティング技術の動向

――高速光インターコネクトによる分散処理コンピューティング

数十億、数兆個の IoT（Internet of things）デバイスがネットワークに接続される時代になれば、これまでとは桁違いな量のリアルワールドのビッグデータが得られ、従来では考えもつかなかった驚きや感動を与えるアプリケーションが実現できる可能性がある。これを実現するには、この桁違いな量のデータを短い時間で処理する必要がある。つまり、従来の常識を大きく超える演算能力を実現しなくてはならない。

電力密度の制限などによりムーアの法則が限界に近づき、LSI の大規模化・高クロック化では処理能力向上の限界を迎える時代（ポストムーア時代）が到来しつつある。2010 年ころから、ムーアの法則は飽和傾向にあり、ポストムーア時代に向け、複数の演算リソースで処理を分担する分散処理が情報処理の性能向上をもたらすアーキテクチャとして期待されている。この分散処理においては、演算リソース間のデータ共有が高速に行えるインターコネクトの性能が重要となる（図1）。

IOWN 時代には、端末やエッジサーバからデータセンタまで、数多くの演算リソースが広帯域の光通信で接続されるようになる。この光により高速につながれた多数の演算リソースを協調させることで、演算を早くすることがポストムーア技術として期待されている。

このポストムーア技術の先行的なトライアルとして、NTT がこれまで光通信用として研究開発してきた高速プロトコル技術・通信処理回路技術を活用し情報処理システムの性能向上を目指す、光インターコネクト技術が注目されている。

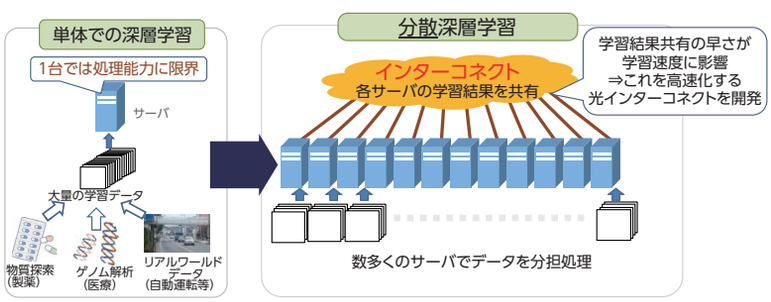


図1　AI 用光インターコネクトの構成

――大量のデータを高速に分散処理する技術

今回は、アプリケーションとして AI に注目し、自動運転やゲノム解析などといった、リアルワールドの大量なデータを高速に処理するために

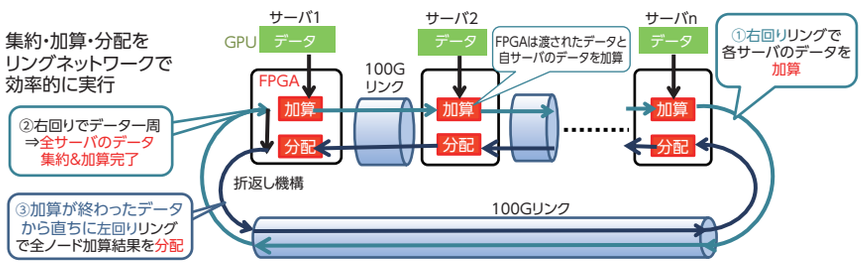


図２　高速リングネットワークで接続

数多くのサーバで分散処理を行う「分散深層学習」に適した光インターコネクトをソフトウェアイノベーションセンター（SIC）と組み新たに開発、AI 学習の高速化を目指した。

ポストムーア時代ではプロセッサの性能向上が限界に達しているため、特定領域に特化した処理回路やアーキテクチャを用いる、いわゆる「ドメインスペシフィックコンピューティング」が重要になる。CPU（Central processing unit）がこのまま高速化し続けるなら CPU に投資するべきであるが、それが期待できなくなった現在、深層学習の領域では深層学習の演算に適した GPU (Graphics processing unit) をネットワークで接続する、並列コンピューティングを取り入れる動きが加速している。これまでは情報処理の負荷をプロセッサに集中していたが、これからはプロセッサとネットワークで分担するというものである。

今回の光インターコネクトの研究では、光ネットワークをうまく活用し、並列コンピューティングを深化させるアーキテクチャを提案することを研究指針としている。

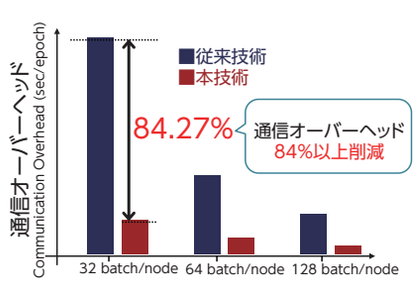


図３　通信オーバーヘッド量比較

――情報処理時間を短縮するための３つのポイント

本成果の光インターコネクトは、分散深層学習の高速化を、以下の３つの技術的ポイントで実現している。

第1の工夫のポイントは、学習処理を行う GPU と光インターコネクトをより密接に結びつけるダイレクト通信にある。既存のコンピュータでは、サーバ間通信を実行するにあたり、サーバ内で演算を行うデバイスとサーバ間の通信を担う光インターコネクト用デバイス間でのデータの授受に CPU やメインメモリが介在するため、大きな遅延が発生していた。そのため、学習処理を行う GPU と光インターコネクト用デバイス間を密接に結びつけ、CPU やメインメモリを介することなく直接データ授受ができる構成とすることで、伝送遅延を削減した。

第2に、複数サーバでデータ共有を行うのに適したプロトコルと通信アーキテクチャを構成した。分散深層学習では、各サーバが分担した学習結果の集約・加算・分散が必要で処理の多くを占める。具体的な解決方法としては、各サーバをリングネットワークで接続し、右回りで加算処理を行いながら、左回りでデータの分配を行った（図2）。全サーバのデータを集めてから加算するよりも加算処理時間を短縮することが可能である。さらに、逆回りルートで分配するため、加算と分配を並行して実施できる。

さらに第３のポイントとして、データ共有処理を専用のハードウェアでデータの集約・加算・分配を高速に実行できるアクセラレータ回路を開発した。アクセラレータでは、光通信速度 100 Gbps のスループットで動作可能な高速加算器を開発し低遅延化に貢献した。

――情報処理システムの基盤技術として

分散深層学習においては、各サーバの学習結果で光インターコネクトを介し共有する通信を行う。この通信の早さが AI の学習速度に大きく影響する。現在用いられている市販品で最速の構成との性能比較を行った結果、４台のサーバ（1 台当たり 1 GPU）を接続した場合においては、通信のために生じる演算待ち時間（通信オーバーヘッド）が 84%以上削減されることを確認できた（図3）。この結果、学習速度が 7%向上することを確認した。この測定結果をもとに、GPU 台数を増やした場合の見積もりを行うと、32 GPU 利用時に 40％以上学習速度が向上する見積もりが得られた。

――将来に向けた技術の展望

今後は、この技術で AI 学習を行うデータセンタに導入することで、自動運転・遺伝解析・気象予測など、大量のデータを扱う AI 学習処理の高性能・低消費電力化が期待される。さらには今後、ポストムーア時代に向け、ネットワークにつながれた多数のリソースを協調させて性能向上を図る。