汎用プロセッサから専用プロセッサへ

米グーグルが、ディープラーニングのための専用プロセッサ「TPU」を自社開発していることが明らかになった。金融機関も今後、特定用途に特化した専用プロセッサを開発する可能性がある。

ディープラーニング専用プロセッサの衝撃

米グーグルは2016年5月に開催した同社の開発者向け会議「Google I/O 2016」で、第三次AIブームのコア技術「ディープラーニング」のための専用プロセッサ「TPU（テンソル処理ユニット）」を自社開発し、1年以上前から同社のデータセンターで使用していることを明らかにした。

TPUはASIC（Application Specific Integrated Circuit、特定用途向けIC）の一種である。これまで、ディープラーニングの処理に使用されてきたCPU（Central Processing Unit＝中央処理装置）やGPU（Graphic Processing Unit＝グラフィック処理装置）、FPGA[1)](Field-Programmable Gate Array)に比べ、消費電力あたりの性能が桁違いに優れているという。現在、TPUは、囲碁の世界チャンピオンである韓国のプロ棋士イ・セドル氏を打ち負かした囲碁AIの「AlphaGo（アルファ碁）」やグーグル・ストリートビュー、音声検索など、グーグルの100以上の開発チームで活用されている。

ASICは用途ごとに回路を設計してLSIを発注するため、非常に高性能であるものの、開発費用、開発期間を要する点がこれまでの課題であった。それを、インテルのようなプロセッサメーカーではないグーグルが独自に開発していたことは驚きに値する。

ムーアの法則の終焉

グーグルが「ディープラーニング」専用プロセッサの開発に乗り出した背景には、「ムーアの法則」の終焉がある。「ムーアの法則」とは、「半導体集積回路の集積密度は2年ごとに2倍になる」というもので、「集積密度」を「性能」に置き換えて、「コンピュータの性能は18～24ヶ月で2倍になる」と表現されることもある。

インテルの創業者の一人であるゴードン・ムーア氏が1965年に発表した論文で初めて提唱したもので、これまで50年以上にわたって通用してきた法則である。

しかし、今後も引き続き、ムーアの法則を維持していくことは難しいと見られている。集積回路の回路幅は2016年に14nm（ナノメートル）に到達し、今後、10nm、7nmとさらなる微細化が求められるものの、微細化が進むにつれて、消費電力の低減が困難になり、さらに製造コストも膨れ上がるためだ。

つまり、半導体の集積密度が指数関数的に向上していく時代が終焉を迎え、ソフトウェアの進化にハードウェアの進化が追いつかなくなってきたのである。そのため、今後は用途に特化したハードウェアの実装が重要になってくる。今回、グーグルが発表したTPUは、同社が2015年11月にオープンソースソフトウェアとして公開した機械学習ソフトウェアの「TensorFlow」に特化したものである。

汎用プロセッサから専用プロセッサへ

グーグルに限らず、ムーアの法則の終焉が囁かれ始めたこの数年、汎用品であるCPUの代わりに、用途に特化したプロセッサを使用しようとする動きが盛んになってきている。

たとえば、マイクロソフトは2010年頃から前述

NOTE

1）製造後に購入者や設計者が構成を設定できる集積回路。
2）コンピュータの単位時間当たりの処理量。
3）詳細は、https://www.youtube.com/watch?v=9NqX1ETADn0を参照。

したFPGAを搭載したサーバーの研究を進めており、2014年には、同社の検索エンジン「Bing」のページランク処理の高速化に向けてパイロットプロジェクトを実施し、ページランク処理のスループット[2)]が2倍に向上したことを報告している。2015年には、これをBingの本番システムに投入したほか、グーグル同様、ディープラーニングにも採用している。マイクロソフトは、GPUを採用した場合とほぼ同等の性能を実現しつつ、電力効率はGPUの2倍前後と発表している。また、中国最大のオンライン検索サービス・プロバイダのバイドゥ（百度）も、オンライン検索を高速化するために、ディープラーニング向けに回路を設計したFPGAを使用することを2014年に発表している。

これまで、ディープラーニングで広く使用されてきたGPUは大量の浮動小数点演算のスループットに優れるが、ASICやFPGAのように用途に応じた自在な回路設計ができないため、適用分野によってパフォーマンスに大きく違いがでる（たとえば、グラフィクスの並列処理には強いが、低遅延の実現には向いていない）。また、デバイス1個あたりの消費電力が大きいため、電力消費量の増大が課題となっているデータセンターで大量に採用することは難しい。

これに対しFPGAは、性能はASICに及ばないものの、GPUに比べた場合、デバイス1個あたりの消費電力が小さく、アプリケーションの用件に応じて、回路設計を日々更新できる柔軟性を備える。ただし、HDL（Hardware Description Language）という低レベルのハードウェア記述言語による開発が必要となり、GPUに比べ、技術者の確保が難しい。

このように既存のプロセッサに一長一短がある中で、今回のグーグルによる専用プロセッサの開発は、性能向上を目指し、熾烈を極めるディープラーニングの開発競争に打ち勝つために、コストにはある程度目を瞑り、性能と電力効率を最重視した結果ともいえる。

金融業界における可能性

金融業界では2010年頃から、大手投資銀行等を中心にミリ秒単位で高頻度の売買注文を繰り返すHFT（High Frequency Trading, 高頻度取引）でデータのレイテンシ（遅延）を一定以下に抑えるために、FPGAの導入が進められてきた。

また、ミリ秒単位の高速処理だけでなく、金融業務で多く使用されているバッチ処理やHPC（High Performance Computing）でもFPGAが使用されてきた。たとえば、JPモルガン・チェースでは、デリバティブのリスク計算専用のサーバクラスタをFPGAで構築していることを明らかにしている[3)]。同社では、それまで全社規模のポートフォリオ評価のリスク計算に汎用CPUによる大量のサーバーを使用していたが、計算が終わるまでに8時間もの時間を要していた。それがFPGAを実装したことによって、わずか238秒で計算が完了したそうだ。今後、金融の世界でもFPGAに飽き足らない企業が、グーグルと同じく、用途に特化した専用プロセッサの開発に乗り出しても不思議ではない。

Writer's Profile
城田 真琴 Makoto Shirota
デジタルビジネス開発部
グループマネージャー
専門はFinTech動向の調査
focus@nri.co.jp