社会課題を解決する革新的コンピューティング

Innovative Computing for Solving Social Issues

● 井上　淳樹　● 三吉　貴史　● 石原　輝雄　● 本田　育史

あらまし

1940年代後半に実用的なプログラム蓄積方式の計算機が開発されて以来，電子計算機は70年の間に約10^{12}倍という驚異的な性能向上を実現してきた。しかし，半導体の微細化技術が限界を迎え，ムーアの法則の終焉が近づいていると認識されている。このような技術的な背景にもかかわらず，IoT時代に生み出されるデータ量の爆発的な増加は今後も続くと予想され，そのデータを使った新たな価値創造やサービスへの期待が高い。この期待に応えるには，ムーアの法則に頼らない性能向上が必要になっている。本稿では，新しいコンピューティングパラダイムとしてドメイン指向コンピューティングを提案する。ドメイン指向コンピューティングでは，知識処理のような厳密な数値処理結果を得ることを目的としない分野において，必要な処理に特化したアーキテクチャーを採用することによってムーアの法則の限界を突破することを目指す。例として，ディープラーニング学習エンジン，高速画像検索エンジン，組み合わせ最適化問題専用マシンへの取り組みについて述べ，従来比50 ～ 12,000倍という高い性能となることを実証した。

本稿では，新しいコンピューティングパラダイムとしてのドメイン指向コンピューティングの方向性と，具体的な取り組み事例について述べる。

Abstract

Since the development of practical stored-program computers in the late 1940s, performance has risen amazingly by about 10^{12} times over a period of 70 years. However, it is generally recognized that semiconductor transistor scaling is reaching its limits and that Moore's law is coming to an end. Regardless of these technical issues, the explosive increase in the amount of data generated in today's IoT era is expected to continue, and it is highly anticipated that this data will be used to create new value and novel services. Meeting these expectations will therefore require improvements in performance independent of Moore's law. This paper proposes domain-oriented computing as a new computing paradigm. The aim of domain-oriented computing is to break through Moore's law by adopting architecture specific to the type of processing needed in fields such as knowledge processing whose objective is not to obtain rigorous numerical results. For example, in application to deep learning engines, high-speed image search engines, and machines dedicated to combinatorial optimization problems, it has been shown that domain-oriented computing can improve performance by 50–12,000 times that of existing approaches. In this paper, we describe the direction of domain-oriented computing as a new computing paradigm and present specific application examples.

ま え が き

実用的なプログラム蓄積方式の電子計算機EDSAC（Electronic Delay Storage Automatic Calculator）が1949年に開発されてから約70年の歳月が流れ，今や老若男女を問わずスマートフォンのような電子計算機を内蔵した電子機器を手軽に手にすることができるようになった。EDSACは3,000本の真空管と水銀遅延線をメモリとして使い，消費電力が12 kWという巨大な計算機であった。その後，基本電子素子の真空管は，ほぼ同時期にW. B. Shockleyらによって発明された固体電子素子であるトランジスタで置き換えられ，1950年代から1960年代に多くの第2世代の商用電子計算機を生み出した。J. Kilbyらの発明によるモノリシックな集積回路技術は，電子計算機の急速なコストダウンをもたらし，より多くのトランジスタを有効に使って計算機の高性能化に向かう方向の開発が加速した。

この頃，インテル社の創業者の一人であるG. Mooreによって，経験則に基づく将来予測として提唱されたのがムーアの法則である。Mooreは「部品（トランジスタ）あたりのコストが最小になるような複雑さは，毎年およそ2倍の割合で増大してきた。少なくとも今後10年間ほぼ一定の率を保てないと信ずべき理由はない。」[(1)]と述べ，後にIBMのR. H. DennardのScaling則によって理論的な裏付けが与えられた。その後，その限界が何度となくささやかれたが，2000年代初頭まで約30年にわたって半導体微細加工技術の開発を後押ししてきた。半導体の加工寸法を縮小することにより，トランジスタの性能と電力効率の向上，集積密度の向上，コストダウンを同時に実現し，一石三鳥のメリットがもたらされてきたからである。

この結果，EDSACの時代から2010年までの計算機の性能をプロットすると，平均して約1.5年で2倍という指数関数的な向上を維持してきており，[(2)] 70年間の性能向上は10^{12}倍という驚異的数字となった。計算機の電力効率もほぼ同様に約1.5年で2倍の速度で向上してきており，このことが同程度の大きさと価格を保ったままで計算機の性能を向上させてきた。すなわち，今日に至るまでの計算機性能の驚異的な向上は，半導体微細加工技術の進歩によるものが第一義的な要因であったのである。

理想的なScaling則が適用されていた時代には，電力密度一定のもとで集積密度の向上ができていた。すなわち，電力効率と性能の2乗の積は一定になると考えられる（**図-1**）。この積の値はその時代に実現できる半導体微細加工技術によって決まるため，図では直線で表すことができ，積が一定の線を越えて高い性能と高い電力効率を同時に実現できないことを示している。この意味で，この線のことをMooreの限界線と呼ぶことができる。しかし，この積の値が半導体微細加工技術の進歩とともに大きくなるため，図のように電力効率の向上と性能を時代とともに同時に向上できたわけである。

しかし，2000年代に入ると電源電圧を理想的に低下させることが難しくなり，トランジスタの集積度を上げることで消費電力が急激に増大してしまうようになった。このため，消費電力の制約によって性能を向上させることが困難になってきた。更に，加工寸法が原子の大きさの100倍のオーダーに近付いてくるとトランジスタの性能を決めるゲート長を短くすることが困難になり，45 nmノードの時代からはトランジスタのゲート長はほとんど短くすることができなくなっている。すなわち，ムーアの限界線が壁になって，半導体微細加工技術の進歩だけに頼っていては，これ以上の性能向上が難しくなると考えられるようになってきた。

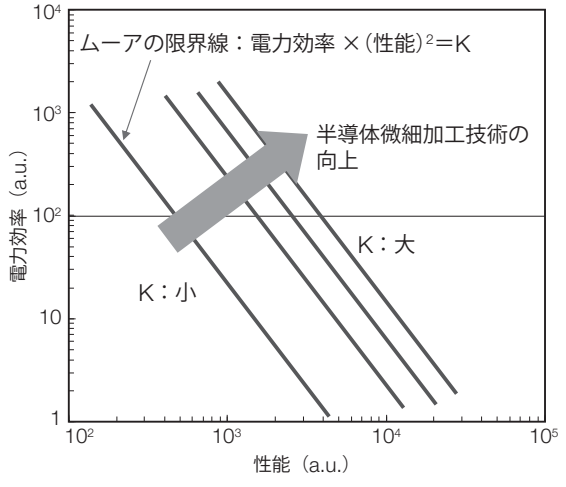


図-1　ムーアの限界線と微細加工技術による向上

これらのことにより，ムーアの法則の終焉と，ムーアの限界線を越えた超ムーアコンピューティングの可能性について，その研究開発の方向性について多くの議論[3]が行われるようになってきた。例えば，ドメイン指向コンピューティングを目指した命令セットRISC-Vの開発や，D-Wave社の超伝導回路技術を用いた量子アニーラの開発，Google社によるAI（人工知能）専用のLSIであるTPU（Tensor Processing Unit）の開発などが挙げられる。

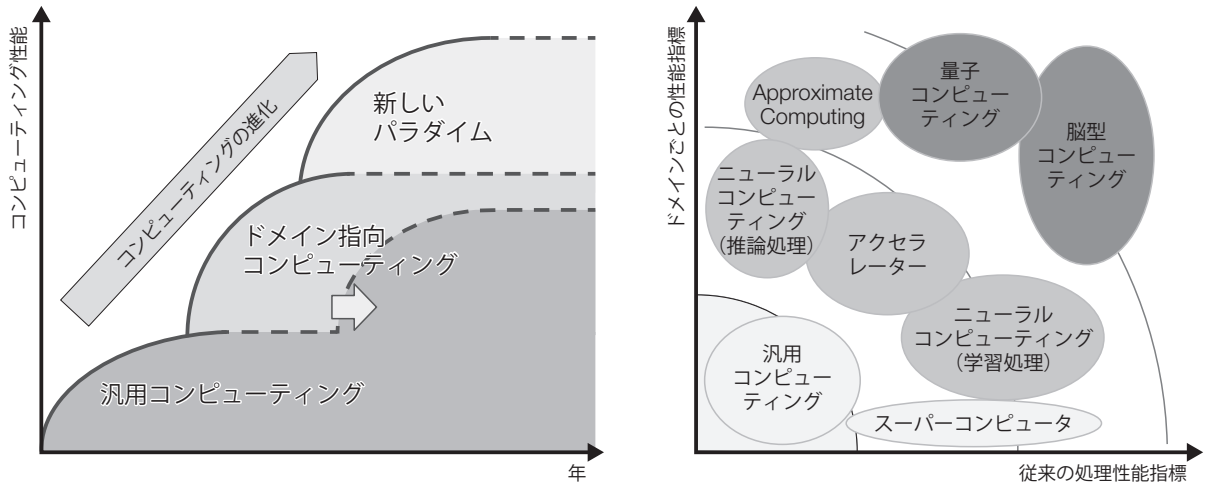
本稿では，まずドメイン指向コンピューティングの考え方について述べた後，具体的なドメインに適用した事例として，AI，メディア，組み合わせ最適化の三つの領域に適用した場合の効果について述べる。

ドメイン指向コンピューティング

ムーアの法則の終焉が目前になり，テクノロジー面では半導体プロセス，ネットワーク帯域，消費電力，計算性能など多くのチャレンジが存在している。一方，処理しなければならないデータは爆発的増加を続けている。IoTにより生成されるデータ量は，2020年には40ゼタバイト以上，2030年には1ヨタバイトに達すると予想されている。[4] データ量が爆発的に増加していくと，従来型のICTのデータ処理能力を上回るのは明らかであり，大量データの中から貴重な情報を見つけるための新しい情報処理の創造が課題となる。例えば，IoTデバイスによって生成されるデータが持つ情報の密度は薄い。しかし，クラウド上にデータを集約し，AIを用いて大量データからそのエッセンスや意味を抽出することで，データは知識や知能に昇華される。その知識・知能を応用することで，高度なICTソリューションの実現が可能となる。

この新しい情報処理に対応するには，コンピューティングも変化が求められる。従来のアーキテクチャーは数値処理にとても優れていたが，知識や知能を効率良く作り出すことに適したアーキテクチャーに進化しなければならない。大量データの処理手法の進化と，アーキテクチャーの進化を相互に補完し合うことによって，知識と知能を活用した新しいアプリケーションやサービスを生み出すことができる。

図-2（a）は，コンピューティングアーキテクチャーの進化の方向性を示している。特定の分野で新たなアーキテクチャーが生まれると，コンピューティングの処理能力が飛躍的に向上する。つまり，新しいアーキテクチャーの革新によってパラダイムシフトが発生する。このパラダイムシフトは徐々に汎用コンピューティングでも取り込まれ，新しいアーキテクチャーによって汎用コンピューティング自体が強化されていく。新しいアプリケーションやサービスを生み出すために，アーキテクチャーの継続的な革新とパラダイムシフト



（a）ドメイン指向と汎用コンピューティングの関係　　（b）ドメイン指向と性能指標

図-2　ドメイン指向コンピューティング

の創出が必要である。そのアプローチとして，富士通研究所はドメイン指向コンピューティングを提案する。

従来，コンピューティングの性能は汎用のアプリケーションを対象に，主に整数演算性能，浮動小数点演算性能，メモリ帯域などの指標で評価されてきた。これらの指標は，半導体の性能に密接に関わるため，ムーアの法則終焉後においては従来のような性能向上を継続することは困難である。しかし，対象とする処理の領域，つまりドメインを絞り，その領域で頻繁に利用される処理に着目してコンピューティングアーキテクチャーを定めることによって，ムーアの法則を超えて桁違いの高性能化をもたらすことが可能である。これをドメイン指向コンピューティングと呼ぶ。

図-2（b）は，コンピューティングにおけるアーキテクチャーごとの特性を表している。横軸は従来のコンピューティングの処理性能を表し，縦軸はコンピューティングの目的に応じてドメインごとに定義される新たな指標を示す。新たな指標とは，例えばメディアドメインにおいては，シーンに応じた要求画像品質の達成度合いなどを指し，従来の周波数，数値演算性能といったCPU性能指標では単純に測ることができない。

半導体の微細化による性能向上の限界を超えるために，富士通研究所は横軸方向ではスーパーコンピュータによる処理能力向上を進め，縦軸方向では特定ドメインへの専門化を追求している。コンピューティングは，従来の数値処理中心からドメインに特化した情報・知識・知能の処理へと変化する。このため，ドメイン指向コンピューティングの第一段階として，アクセラレーターやニューラルコンピューティングに取り組んでいる。更に，新世代のアーキテクチャーとして，量子コンピュータ，脳型コンピューティングなどが注目を集め始めている。

ドメイン指向コンピューティングの考え方を基にアーキテクチャーを見直すと，従来とは異なる方向性が見出せる。従来のアーキテクチャーでは，様々なワークロードを想定し，そのワークロードに共通して性能を発揮するよう設計された汎用CPUを用いる。そこでは，逐次処理と並列処理を組み合わせ，高精度（あるいは均一的な精度）な解を探索している。

一方，ドメイン指向コンピューティングのアーキテクチャーでは，ドメインに必要とされるコア処理に着目する。その特性に特化して無駄を排除したシンプルな専用コアを大量に並べることによって，膨大な演算を高並列かつ高電力効率で処理することを可能にする。更に，対象とするドメイン処理に応じた精度を追求することで，最適な状態で処理を効率化できる。これらの方針により，従来のコンピューティングに比べて，桁違いに高性能かつ高電力効率の実現を目指す。

ドメイン指向コンピューティングを実現する上で，汎用CPUに代わるコンピューティングデバイスは重要な役割を担う。GPGPU（General-Purpose computing on Graphics Processing Units）による高並列プログラムやFPGA（Field-Programmable Gate Array），ASIC（Application Specific Integrated Circuit）を用いた専用ハードウェアはその一例である。更に，対象ドメインのコア処理となるアルゴリズムは，従来の汎用CPU向けのアルゴリズムからハードウェアの構造を意識したアルゴリズムに変わっていく。アルゴリズムとハードウェアを密接に連携する要素として扱い，最適な処理手法の提供が桁違いの高速化を実現するためのポイントとなる。

以降の章では，富士通研究所がドメイン指向コンピューティングで開発した技術としてDL（Deep Learning）専用エンジン，メディアサーバ，デジタルアニーラについて説明する。

ドメイン指向コンピューティングの適用例

コンピューティングの基本動作は，データの入出力を除くと，制御，演算，記憶（メモリ）である（**図-3**）。ドメイン指向コンピューティングでは，これらの一つ，あるいは複数をその特性に合わせてアーキテクチャーを特化し，従来では実現し得なかった高い性能を獲得する。

以下に，ドメイン指向コンピューティングの適用例を示す。

- **DL専用エンジン[5]　DLU（Deep Learning Unit）[6]**

ディープラーニングの学習処理は条件判断によって処理フローを逐次変えるのではなく，決まっ

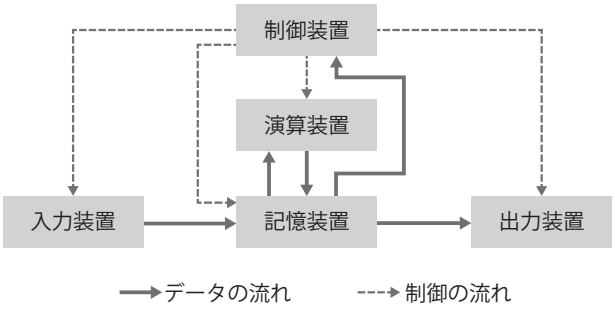


図-3　コンピューティングの基本動作

た処理フローに学習データを入力して処理を繰り返すことで，少しずつ学習していくという特徴がある。そして，ディープラーニングに代表されるAIを使ったアプリケーションでは，ディープラーニングの推論の特性からある一定以上の確率で正解が得られることを前提に考えられており，必ず正解が得られることを基本にしたものではない。

推論精度が高いことは高い価値につながるが，トレーニングは膨大な処理量となる。したがって，アプリケーションを満たす推論精度を，少ない処理量の学習で実現することによって様々な種類のトレーニングを実行したり，少ない消費電力で実現することで利用できる適用領域を拡大したりすることで，価値を高めることが可能なコンピューティングである。

一般的に，ディープラーニングで行われているコンピューティングは，CPUやGPUを用いて32ビット浮動小数点フォーマットで演算が行われてきた。しかし，推論精度向上に向けたディープニューラルネットワークの規模が拡大するのに伴い，演算処理量，メモリ量，メモリバンド幅，消費電力が増大してきている。このことから，これらの効率を改善することで電力性能の向上を可能にするアーキテクチャーが必要になっている。

DLUはディープラーニング処理に特化したプロセッサであり，高い処理性能を少ない消費電力で実現するため，独自アーキテクチャーを採用している。DLUの制御と記憶では，膨大な学習データに対してコンスタントに性能を維持するために，演算と独立してDPU（Deep learning Processing Unit）内のレジスタファイルへのデータ共有や保存を，ソフトウェアによる制御で実行できる仕組み（**図-4**）と，大規模なレジスタファイルをベースに演算を並列して行う（**図-5**）ことができるアーキテクチャーとしている。演算としては，GPGPUによるディープラーニング処理では，浮動小数点32ビット（以下，FP32）が標準となっている。しかし，処理性能の向上に特化した低ビット化が進んでおり，16ビットの浮動小数点演算をベースにしたNVIDIA社のTensorCore[7]や，16ビットの整数演算をベースにしたIntel社のFlexpoint[8]などが発表されている。いずれも，FP32演算と同等の学習性能を維持しながら，より低ビットで学習することを工夫している。

DLUでは，DL-INT（Deep Learning Integer）という演算中の統計情報を利用することで演算精度を確保し，8ビットまたは16ビットによる学習を可能にする仕組みを搭載している。これにより，同一のメモリ量，メモリバンド幅で最大4並列の演算を可能にする。それに加えて，整数演算をベースにして消費電力を半減することで，FP32演算に対してアーキテクチャーによる約8倍の電力性能の向上を図っている。更にこれは，半導体チップを複数用いるマルチチップによる学習処理において，ボトルネックとなるチップ間のデータ通信量を減少させる。このため，規模が増大するディープニューラルネットワークに対して，マルチチップによる高効率な学習にも寄与する。

● メディアサーバ[9]

メディアサーバは，画像や音声などのメディアデータを高速に検索することで大量データの再利用を可能にし，業務効率化の実現を狙いとしている。

このメディア検索を高速化するアーキテクチャーの特徴は，メディアデータ検索処理に関わるアルゴリズムを並列化してハードウェアエンジンに実装し，演算をCPU処理からオフロードすることにある。エンジンは，デバイスとしてFPGAを利用し，記憶データを高速なメモリにタイミング良く転送するバランスの良いパイプラインスケジューリングを行うことで，その処理効率を高めている。

具体的には，32並列の特徴量計算と6並列のマッチング処理をハードウェアリソースに割り当てている。このようなアーキテクチャーにより，アル

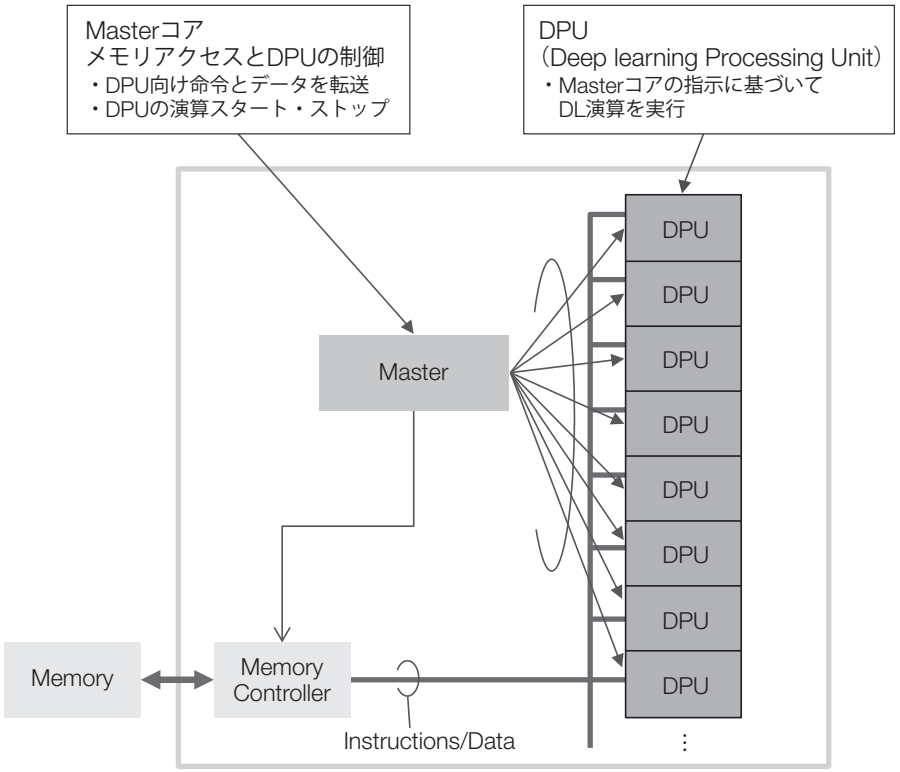


図-4　制御系とマルチDPU構成

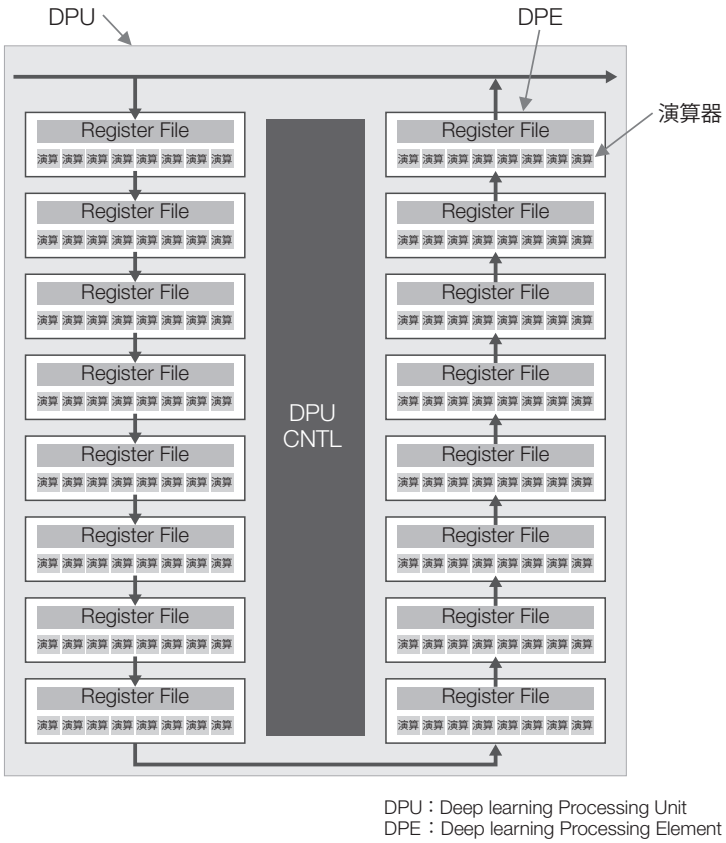


図-5　DPUアーキテクチャー

ゴリズムのハードウェア並列処理を実現し，50倍に高速化した検索に成功している。

● **デジタルアニーラ**[10]

組み合わせ数が少ない場合には，従来のコンピューティングで容易に解ける問題であっても，組み合わせ数が多くなると途端に解を得るまでに膨大な時間がかかってしまう組み合わせ最適化（巡回セールスマン）問題などがある。

従来の汎用コンピューティングでは処理対象から外れていたものを，全結合イジングモデルをハードウェアアーキテクチャー化し，更に速度を飛躍的に向上させることで，金融ポートフォリオ最適化問題（500銘柄）などの実問題を解くことを可能にした。

この組み合わせ最適化問題を高速に解決するアーキテクチャーの特徴は，演算のコア部分を専用ハードウェア化することで単一試行（基本処理）の高速化と，その演算処理を1,024並列で実行可能にしているところにある。更に，アルゴリズムとしてダイナミックオフセット手法を併用することで，全体として約 12,000倍の高速化を実現している。

む　す　び

本稿では，半導体の微細化による性能向上の限界を突破する，富士通が考える新しいコンピューティングアーキテクチャーであるドメイン指向コンピューティングの方向性と，その実現例を紹介した。

コンピューティングに要求される性能向上は，今後も続くことが予想される。しかし，対象とする処理の領域を知識処理などの最適化が必ずしも必要でない領域に絞り，その領域で頻繁に利用される処理に着目すれば，デバイスの性能向上に頼らず桁違いの高性能化が可能であることが示された。

今後は，アプリケーション自体を大きく変更することなく，性能を引き出すことのできるソフトウェアおよびライブラリの開発を行う。それとともに，このような高性能な処理をお客様に低コストで提供できる仕組みの実現が鍵になると考えている。

参考文献

（1）G. Moore：Cramming More Components onto Integrated Circuits. Electronics, Vol.38, No.8, p.114（1965）.

（2）J. G. Koomey et al.：Implications of Historical Trends in the Electrical Efficiency of Computing. IEEE Annals of the History of Computing, p.46（2011）.

（3）T. H. Theis et al.：The End of Moore's Law：A New Beginning for Information Technology. Computing in Science & Engineering, Vol.19, p.41-50（2017）.

（4）二瓶真理子ほか：情報の質・価値をめぐる概説的試論．信学技報，Vol.116，No.290，p.9-12（2016）.

（5）池 敦ほか：Deep Learningのための高効率化技術. FUJITSU，Vol.68，No.5，p.15-21（2017）.
http://www.fujitsu.com/jp/documents/about/resources/publications/magazine/backnumber/vol68-5/paper03.pdf

（6）T. Maruyama：Fujitsu HPC and AI Processors. ISC 2017.
http://www.fujitsu.com/global/Images/fujitsu-hpc-and-ai-processors.pdf

（7）P. Micikevicius et al.：Mixed Precision Training. arXiv preprint arXiv:1710.03740，2017.

（8）U. Köster et al.：Flexpoint: An Adaptive Numerical Format for Efficient Training of Deep Neural Networks. In: Advances in Neural Information Processing Systems. 2017. p.1740-1750.

（9）渡部康弘ほか：FPGAアクセラレーターによるドメイン指向コンピューティング．FUJITSU，Vol.68，No.5，p.22-28（2017）.
http://www.fujitsu.com/jp/documents/about/resources/publications/magazine/backnumber/vol68-5/paper04.pdf

（10）塚本三六ほか：組み合わせ最適化問題向けハードウェアの高速化アーキテクチャー．FUJITSU，Vol.68，No.5，p.8-14（2017）.
http://www.fujitsu.com/jp/documents/about/resources/publications/magazine/backnumber/vol68-5/paper02.pdf

著者紹介

井上　淳樹（いのうえ　あつき）
（株）富士通研究所
コンピュータシステム研究所
新コンピュータアーキテクチャーの研究開発に従事。

三吉　貴史（みよし　たかし）
（株）富士通研究所
デジタルアニーラプロジェクト
ドメイン指向コンピューティングの研究開発に従事。

石原　輝雄（いしはら　てるお）
（株）富士通研究所
コンピュータシステム研究所
知能コンピューティングの研究に従事。

本田　育史（ほんだ　やすふみ）
富士通（株）
AI基盤事業本部
DLU開発に従事。