2.5.2 プロセッサーアーキテクチャー

（1）研究開発領域の定義

コンピューティングにおいてプロセッサーは中心的な役割を果たし、長らくフォンノイマン型アーキテクチャーが大勢を占めていた。アーキテクチャー（Architecture）という言葉は、元来は建築学の分野において建築様式を意味する言葉であるが、情報処理分野では、計算機ハードウェアの基本様式、基本構造、設計思想などを指す言葉として使われている。ソフトウェアは「アーキテクチャーをターゲットとしてコンパイルされる」ものであり、ハードウェアは「アーキテクチャーをもとにしてデザインされる」ものであると理解することができ、ソフトウェアとハードウェアとを結びつける抽象モデルがアーキテクチャーであると言える。その位置付けは極めて重要であり、プロセッサーを特徴付ける概念である。

（2）キーワード

フォンノイマン型アーキテクチャー、ドメイン・スペシフィック・アーキテクチャー、深層ニューラルネットアクセラレーター、リコンフィギュラブル・コンピューティング、インメモリー・コンピューティング、エッジコンピューティング

（3）研究開発領域の概要

［本領域の意義］

プロセッサーにおけるアーキテクチャー研究は、1980年代から1990年代にかけて、CISC（Complex Instruction Set Computer）対RISC（Reduced Instruction Set Computer）アーキテクチャー論争、RISCアーキテクチャーをベースにしたプロセッサーの高実行効率化技法、命令レベル並列化、スレッドレベル並列化等の並列実行手法などの研究で大いに盛り上がった。その後、ムーアの法則（トランジスタ数は1.5年で2倍になる）に従ったプロセッサー単体性能の着実な向上の勢いに隠れ、アーキテクチャー研究は次第にその輝きを失っていった。しかし、2010年頃より、主に1）ムーアの法則に陰りが出たこと、2）アーキテクチャーの工夫を必要とする新しいタイプの情報処理課題がメインストリームになったこと（後述）、の二つの事象が並行して進行し、従来のアーキテクチャーから新しいアーキテクチャー（DNN、DSA、脳型など）への期待、展開が広がり、新たな波になりつつある。特にここ数年は「アーキテクチャー研究の黄金時代」とも呼ばれる活況を呈している。 Society 5.0というキーワードで近未来の超スマート社会のビジョンが産官学で議論されているが、その議論は、情報処理能力のこれまで通りの指数関数的な発展を前提にしている。その前提を支えてきたムーアの法則の今後が心もとない現状においては、アーキテクチャー技術の果たすべき役割は大きい。

［研究開発の動向］

コンピューティングアーキテクチャーは、いわゆるフォンノイマン型を王道として発展してきた。これは、メモリー内に蓄えられた命令列（処理プログラム）を順次解釈・実行していくことを基本的特徴とする手続き処理型のアーキテクチャーであり、チューリングマシンを源流とする極めて強力な問題記述能力・汎用性を誇る。1980年代に、多様なプログラムをより少ない命令数で実行することを目的として命令数が膨れ上がってしまったIBM等の汎用コンピューター（CISC）に対するアンチテーゼとして、アーキテクチャーを単純化・規則化して命令数を減らしたマイクロコンピューター（RISC）が提案された。 RISCは実行命令数が増えるものの、命令当たり実行時間の短縮によりプログラム処理性能が向上することが定量的に示され、脚光を浴びた。ムーアの法則の力を大いに借りて、クロック周波数向上により性能向上を図るRISCドリブンなアプローチが、その後のアーキテクチャーを席巻することとなった。

フォンノイマン型アーキテクチャーでは、処理プログラムと処理データの双方をメインメモリーに蓄えるため、

頻繁なメインメモリーアクセスが性能律速要因になる（フォンノイマンボトルネック）ことが知られている。RISCアーキテクチャー登場以来のアーキテクチャー研究の主要な分野の一つは、このボトルネックを緩和するためのメモリーシステム階層に関するものであり、キャッシュメモリーの工夫やその他さまざまな命令・データのバッファリング手法が提案されてきている。その他の主要分野としては、手続き型処理の根幹となる分岐命令の先読み・予測、制御ハザードの回避、複数命令並列実行や複数スレッド（スレッドとは一塊の手続きのこと）並列実行、プロセッサーを多数並べた大規模並列システムなどが挙げられる[1)]。

フォンノイマン型に代わる「非ノイマン型」のアーキテクチャー思想を打ち立てる研究も古くから続いている。その一つは、1980年代から1990年代にかけて盛んに研究されたデータフローアーキテクチャーである。その基本思想は入力データがそろって実行可能になった命令から実行する点にあり、あらかじめ定められた手続き順に命令を実行するフォンノイマン型に比べると、自然に並列化が可能というメリットがあった。一方、デメリットとしては、プログラム実行の際のさまざまな局所性（特にメモリーアクセス）が担保されないという点が挙げられる。ムーア則に基づく単体プロセッサー高速化の波と、当時の処理対象ワークロードが手続き処理型向きだったことで、情報処理アーキテクチャーの主流とはならなかったが、さまざまな並列化技術としてプロセッサー内に埋め込まれる形で、現在でも広く影響を与えている。その当時、電総研（現産業技術総合研究所）をはじめとして日本でも有力な研究がいくつも進められ（SIGMA–1[2)]、EM4[3)] など）、日本のデータフロー分野の技術・アイデア・知見の蓄積は厚い。

FPGA（Field Programmable Gate Array）は、ユーザーが手元で自在に回路を実装できる集積回路として、ハードウェア設計のプロトタイプ目的で1985年に登場し発展してきた。視点を変えて、所望の情報処理をソフトウェアプログラムにではなくハードウェア構造に設計することを考えるならば、FPGAは新たなコンピューティングデバイスと考えることもできると登場当初から提唱されていた。トランジスタ微細化の進展につれFPGAに搭載できる回路規模が爆発的に増大し、特に2010年頃からはこの「FPGAコンピューティング」の考え方が実用的な意味を持ち始めている。

FPGAコンピューティングは、あくまでプロトタイプ目的だったFPGAの別用途利用であり、コンピューティングに使うならばそもそもプログラマブルハードウェアのアーキテクチャーから再定義すべきと考える研究が1990年代から2000年代にかけて盛んに進められた。一つの方向性は、1–2バイトのデータ処理に適した粗粒度の演算器アレイ（Coarse Grained Reconfigurable Array: CGRA）アーキテクチャーであり、もう一つは動的再構成アーキテクチャーである。後者は、ソフトウェアにおける仮想メモリーの考え方に倣い、ハードウェアを仮想化することで汎用性を高めようとするアプローチであり、特に日本で研究が盛んに進められた（慶応大学WASMII[4)]、NEC（現ルネサスエレクトロニクス）の動的再構成プロセッサーDRP[5)] など）。これらの分野はリコンフィギュラブル・コンピューティングと呼ばれている（FPGAコンピューティングもその中に含んで使われる場合が多い）。

コンピューターの出力を人間が視覚的に理解するために必要となるのがグラフィックス処理であり、GPU（Graphics Processing Unit）はその専用アーキテクチャーとして発展してきた。グラフィックス処理には、画素、線分、頂点、視線方向などさまざまなレベルでデータ並列性が存在するため、主にSIMD（Single Instruction Multiple Data）型のデータ並列アーキテクチャーとして進展してきている。既にコンピューターに組み込まれたGPUをグラフィックス目的以外でも使用して他のデータ並列性の高い応用を処理できるように、2010年頃からいわゆるGPGPU（General Purpose GPU）アーキテクチャーやその利用環境が広まってきた（GPU コンピューティング）。

信号処理分野でも独自のアーキテクチャーが発展してきた。信号処理には積和演算が演算の大半を占めることや、あらかじめ静的に処理時間を計算できるリアルタイム性が要求されるという特徴があり、これに応え汎用プロセッサーから分化して誕生したのが信号処理プロセッサー（DSP）アーキテクチャーである。主にベースバンド信号、音声、画像などが処理対象であり、特に静止/動画像を加工したり圧縮・伸長したりする処理では、2次元画像を効率よく処理するためのさまざまなアルゴリズム–アーキテクチャー連動の工夫が創案

されてきており、イメージ処理プロセッサー、ビデオ処理プロセッサーなどと呼ばれている。

これらコンピューターから出発したアーキテクチャー以外に、回路の集積度がある程度進んできた段階で、それをアーキテクチャーに対する境界条件の大きな変化と捉え、集積回路の効率実装の観点からアーキテクチャーを作り直そうと考える研究分野が1980年代に勃興した（VLSIアーキテクチャー）。特に、フォンノイマンボトルネックの解消を狙ってメモリーとロジックが一体となったインメモリー型の処理を標榜する場合が多く、知能メモリーアーキテクチャーと呼ばれている[6)]。メモリーアレイ内で演算を行うことから、データの移動が少なく並列性を高めやすいという特徴があるが、応用が限られるという難点があり、大きなブレークスルーを起こすには至っていない（部分的にシステムLSIに取り込まれることはある）。

また、同じく集積回路ドリブンで、正統的なアーキテクチャー研究からははみ出た研究としてニューロモーフィックアーキテクチャーが挙げられる。1980年代の第2次ニューラルネットブームの際に、網膜神経回路のアナログ集積回路化で注目を集め、シナプスの動作を精密に模倣する回路の試作などが報告されているが、実用化とは距離のある研究であったため、単発的な研究にとどまっていた。しかし、近年になり人工知能分野において注目を浴びるようになり、IBMのTrueNorth、マンチェスター大学のSpiNNaker（Spiking Neural Network Architecture）、IntelのLoihiなど機械学習のアクセラレーターとして開発が進められている。

従来のコンピューターとは異なる計算原理で動作し、問題によっては圧倒的な高速化を実現すると期待されているのが量子コンピューターである。1980年代に理論的可能性が示され、その後、因数分解への適用可能性、量子誤り訂正符号の提案があり、2000年代には量子コンピューターの研究が活発化したが、スケーラビリティーを確保する技術的な見通しの悪さから研究開発は停滞していた。近年になり、ハードウェア技術や量子誤り訂正符号などの進展により、その可能性が再認識されている（詳細は、2.5.3を参照）。

（4）注目動向

［新展開・技術トピックス］

現在のアーキテクチャー研究の活況は、情報処理性能向上に対する社会的要求に応えるためには今後アーキテクチャーで差分を生み出すしかない、という状況を反映したものである。一方、時代の変化により情報処理の対象ワークロードが変化することで、アーキテクチャーの工夫で性能向上ができる余地が生まれたからでもある。

その一つは、ビッグデータを活用しクラウドサービスを支える大規模分散並列コンピューティング技術であり、それを先鋭化したデータセンタースケール・コンピューティングアーキテクチャーの考え方である[7)]。すなわち、今やネットワークでつながった巨大な数のコンピューティングノードの集合体そのものがコンピューターシステムであり、その構成のみならず、空調を含めた電力モデル、故障に対する冗長性、機器のライフタイムマネジメント等、システム全体のオペレーション最適化を考えることがアーキテクチャーの一つの大きな分野となっている。また、光通信の持つ特性を利用してCPUやGPUなどの演算リソースを接続し、通信オーバーヘッドを減らしつつ、柔軟なリソース制御を行えるというディスアグリゲーテッド・コンピューティングというアーキテクチャーも提案されている。

もう一つの分野がドメイン・スペシフィック・アーキテクチャーの新展開である[8)]。その背景には、機械学習、深層ニューラルネット（Deep Neural Network: DNN）、大規模グラフ処理、アニーリングマシンなどに代表されるビッグデータ時代の新たな情報処理ワークロードが、従来の手続き型処理から離れ、構造型処理に適した特徴を有するようになっている点が大きい。例えば、DNNは、大量・多層に並べられたニューロン間の複雑な結合網という「構造」の中に入力データストリームを流し込んで学習や推論を行うことを特徴とする。処理の中に、分岐を含む手続きはほとんど存在せずDNNという構造そのものを並列ハードウェア構造の上に適切にマッピングすることで大幅な処理能力向上を見込むことができる。このような処理対象領域の特徴をアーキテクチャーに反映させることで、大幅に処理効率を向上させることがドメイン・スペシフィック・アーキ

テクチャーの狙いであり、前述のグラフィックス処理、信号処理、イメージ/ビデオ処理などはそのはしりとも言える。なお、プログラミング言語の世界でも、アーキテクチャーの世界の動きに呼応して、ドメイン・スペシフィック言語（Domain Specific Language: DSL）が発展していることも注目される。このような流れは、2020年代を迎えてさらに加速しており、正統的なアーキテクチャー研究（メモリー階層、分岐予測、並列処理）を抑えて、ドメイン・スペシフィック・アーキテクチャーに関する研究成果が主要国際会議での注目分野・注目発表として位置付けられている。

深層ニューラルネット（DNN）アクセラレーター

深層ニューラルネット（DNN）が画像分類精度で従来手法を大きく超えることが2011年に報告され、本技術は一躍脚光を浴びることとなった。その成功の鍵となった学習手法は1980年代に提唱されたバックプロパゲーション（BP）技術であるが、大規模学習データ、高性能計算機、さまざまなBP改善手法（いわゆるディープラーニング/深層学習技術）等が相まって急速に技術発展し、今や多様な応用分野（画像・音声認識、自動翻訳、自動運転など）でDNN活用が広がっている。また、DNNの学習・推論処理の加速、低電力化を目指して多くのDNN処理エンジンが提案され（Google社TPU[9]、MIT Eyeriss[10] など）、新しい情報処理アーキテクチャー技術として大きな注目を集めてきており、ドメイン・スペシフィック・アーキテクチャーの代表的な存在であると言える。2020年には、学習の高速化をターゲットにしたGoogle社のTPUv2、v3の技術内容が公開された[11]。技術的な新しさはさほどないが、学習環境やDNNのモデル開発を中心的にドライブしている立場を利用し、トータルな解を提供している点で大きな強みを見せている。膨大な並列性を有するという点でGPUコンピューティングがまずその中心的アーキテクチャーとなり（特に学習処理）、エッジ側での推論処理を対象として、組み込み機器（特に画像処理）の積和演算アクセラレーターとして発展してきたDSPベースのアプローチも提案されている。また、構造型の情報処理であるという特徴に注目して、データフローマシンをベースとしたものや[12]、FPGAコンピューティング[13]、リコンフィギュラブル・コンピューティング[14] など、さまざまなアプローチがしのぎを削っている状況である。国内では、東京工業大学がFPGAコンピューティング[15]、北海道大学（発表当時：2019年より東京工業大学）がリコンフィギュラブル・コンピューティング[16] ベースの研究を活発に進めている。また、産業界ではルネサスエレクトロニクス社が動的再構成プロセッサーをDNN処理の差別化エンジンとする技術やマイコンの製品ラインを発表し[17, 18]、注目を浴びている。また、プリファードネットワークス社が、国内で開発されてきた並列処理マシンのアーキテクチャーの系統を継ぐDNNの深層学習（ディープラーニング）アクセラレーターチップを発表し、これを搭載したスーパーコンピューターMN-3がGreen500で1位となるなど大きな注目を集めている[19]。

このように深層ニューラルネット（DNN）技術の爆発的進展が続いている中、より大きなDNNモデルの方がより高い汎化性能、すなわち未学習のデータに対して正しく予測できる能力を持つことが分かり、過剰なパラメーターは忌避すべしという従来の機械学習の基本的理解（オッカムの剃刀）に反する新発見として大きな話題になっている。この「DNNのスケーリング則」の発見を理解する鍵とされているのが宝くじ理論、すなわちDNN学習を母体DNNに無数に存在する部分ネットワーク群の中から「良い部分ネットワーク（NW）=宝くじ」を削り出すプロセスであると位置付ける理論である。この理論では、NW接続とその重みパラメーターが増えるほど部分NWの数が組み合わせ爆発的に増えていくため、その中に存在する宝くじの数は増えることが示唆される。この宝くじ理論に基づく推論チップが2022年に東京工業大学から発表された[20]。これは、宝くじ理論では乱数初期化された重みパラメーターをそのまま使えることに着目し、重みの乱数生成により推論実行時のメモリーアクセスを大幅に削減して電力効率を向上するものであった。このように、DNN理論の爆発的な進展は続いており、その進展をうまくコンピューティング手法の革新に転換したアーキテクチャーの研究は今後も活性化すると予想される。

ハイパーディメンジョナル・コンピューティング

一方、DNNの学習処理の重たさや推論時に入力摂動に弱い、すなわちだまされやすいという課題を解決しうる別の機械学習アプローチとして、1980年代に提案された超高次元（ハイパーディメンジョナル）コンピューティング（HDC）も注目されている。大脳の中で各種情報が超高次元のベクトルで分散表現・想起されているとの仮説から模擬して、学習や推論の対象データを超高次元のハイパーベクトル（HV）にランダム写像し、そのベクトル間の演算により分類・推論等を行う仕組みである。典型的にはHVの各要素は{1, 0}のバイナリ変数で表される。高次元になればなるほどランダムなHV同士はほぼ必ず直交することを利用して、要素毎多数決でビット融合、すなわち判定を行う簡便な学習とHV間のハミング距離の近傍探索に基づくロバストな推論とを実現している。注目ポイントは、その軽量性・超並列性を活かし、HV記憶機構の中（もしくは近傍）でHVを並列処理するイン（ニア）メモリー・コンピューティングによる実現である。 HDCについては、Stanford大学が15年に米国Rebooting Computingムーブメントの中で発表したN3XT構想[21)]の中でカーボンナノチューブベース3次元集積システムの計算モデルとして再発見された印象である。ただ、機械学習モデルとしては精度に改善の余地があり、インメモリー計算ハードウェアを志向する上で「使える計算モデル」としてのみ利用されてきた感が強い。今後の別視点での発展が期待される。

ニューロモーフィック・ハードウェア

DNNの興隆の影響を受けて、集積回路の上で生体神経回路網の動作をできる限り精密に模擬しようとするニューロモーフィック・ハードウェア分野も活性化している（DNNアクセラレーターと混同される場合が多いが、区別して理解する必要がある）。生体模倣の目的については慎重に考える必要がある（例えば鳥を忠実に模倣しても飛行機は実現できない）が、脳がDNNより桁違いに（一説に10^4倍）エネルギー効率が良い理由を探求し、その本質を新しい時代のアーキテクチャーとして昇華していく方向の研究ならば工学的な意義も持ち得る。この分野では、IBM社のTrueNorth[22)]や、清華大学のTianji[23)]、Intel社のLoihi[24)]などが知られている。これらはデジタル回路を採用しているが、不揮発性メモリーを用いたアナログ回路アプローチも、特に新規デバイスの出口戦略的な位置付けで、活発に研究されている。

アニーリングマシン（量子、非量子）

DNNの勃興と並行して、種々の組合せ最適化問題を二値スピン格子のイジングモデルにおけるエネルギー最小化問題に置き換え、その近似解を求めるアニーリング計算機分野も広く注目を集めている。格子状に並べられ互いに相互作用するスピンが安定状態（すなわちエネルギー最低状態）に自然に収束するという物理現象を情報処理と見なし、短時間で質の高い近似解を得ようとする方法と理解できる。基本的には物理現象を利用して情報処理を実現するナチュラルコンピューティングの流れをくむものであり、これまでに量子現象を使うアプローチ[25)]と、集積回路で疑似的に再現するアプローチ[26), 27)]が提案されている。「量子アニーリング」の原理は1998年に東京工業大学・西森教授が発表したことで知られ、量子力学的なトンネル効果によりエネルギー極小値にはまらずに与えられた最適化の解（最低状態）を探索できるという特徴がある（ただし、実際のD-waveマシンはこの「量子アニーリング」の原理とは異なる動作をしていると考えられている）。一方、量子効果に頼らなくとも、成熟した集積回路技術とアーキテクチャーの知見を活かし、スケーラビリティー、結合数、結合の階調などの観点で量子アプローチよりも実用的な計算機として開発を進めているのが日立や富士通などである。2020年には、東京工業大学[15)]と同じチームや北海道大学等が並列にスピンを更新できる新しいスピン更新モデル・確率的セルラーオートマトンとこのモデルに基づく全結合・全スピン並列型アニーリング集積回路を発表し注目を集めた[28)]。

後者のような、凝縮系（Condensed Matter）の物理現象に内在する協働現象の理論を並列計算システムに持ち込む考え方を、ナチュラルコンピューティング（Natural Computing）やPhysics-Inspired Computingと呼ぶ。この分野は、大量の計算資源をチップ内に集積化できるようになった時代にふさわしい

新しい並列計算原理アプローチとして、今後注目を集めていく可能性がある。また、DNNを中心とする機械学習分野とアニーリング分野とは、共に「目的関数の最適化」という共通基盤技術を背景として持っており、今後、その計算モデルがどのように融合・協創していくかは、今後のアーキテクチャー研究トレンドの大きな注目ポイントである。

これらは、大きく捉えるならば、「計算するとは何か?」を新しい視点で捉え直すイニシアチブでもあり、長期的には、前述のPhysics Inspired Computingのようなアプローチが、人工知能を支えるアーキテクチャー基盤技術として大きな発展を遂げる可能性があるとも言える。

量子コンピューター

従来のコンピューターの論理素子（bit）では「0か1か」の2状態の情報を計算に用いるのに対し、量子コンピューターでは「0でありかつ1でもある」状態を任意の割合で組み合わせた量子ビットを用いる。量子コンピューターの計算原理としては、量子ビットに位相の回転、量子もつれ、量子干渉などの量子ゲート操作をすることにより情報処理を行う量子回路型量子計算（量子チューリングマシン）が代表的な計算原理であり実装も進んでいるが、量子断熱計算や測定型量子計算など等価な計算モデルも多く知られている。理論通りに動作すれば、現在のコンピューターよりも本質的に高速な計算が可能になると証明されているが、現在のところ、量子性に基づく量子コンピューターの高速性を実験実証するまでには至っていない。 Shorの素因数分解やGroverの検索などの典型的な量子アルゴリズムが要求する量子ビット数やエラー率と、現状の技術との間には大きな隔たりがある（詳細は2.5.3を参照）。今後、量子コンピューターアーキテクチャーの研究開発の充実が期待される。

以上のような主要なドメイン・スペシフィック・アプローチの中で、アーキテクチャー的な手段として特に注目されているのがリコンフィギュラブル・コンピューティングである。これは、HWの構造に問題を落とし込んで解くというこのアプローチの基本的特質が、構造型の情報処理ワークロードと非常に相性が良いからである。日本では、1990年代から2000年代にかけてさまざまなリコンフィギュラブル・アーキテクチャーが活発に研究されてきたという歴史的な経緯があり、相対的に技術・人材の蓄積が厚い分野であるということは注目に値する。

また、演算自体は単純かつ並列化しやすく、入出力データに強くバインドされたものであるという特徴から、ニアメモリー・コンピューティング、ないしはインメモリー・コンピューティングというアプローチも重要になっている。これらのアーキテクチャー概念は知能メモリー・アーキテクチャーとして古くから存在するが、現実的な応用の中で重要性が増してきたため改めて注目されている状況である。

さらに、クラウド一極集中に対するアンチテーゼとして、エッジコンピューティングの概念も注目されている。利点は、ネットワークバンド幅の削減、データの秘匿性や安全性の向上、リアルタイム性の向上などであるが、エッジ側の情報処理能力をどの程度持たせるべきかなどの定量的な答えが見えていないなど、課題はまだ多い[29)]。日本の産学が有する強みを今後のスマート社会に活かしていくという意味では、エッジコンピューティングのシナリオが技術競争的に非常に重要なことはほぼ間違いがない。これを成立させるユースケースや応用スタディー、対応する社会的プラットフォーム等の研究開発に期待が集まるところである。スマートエッジやエッジコンピューティングの動向については、「2.5.5 IoTアーキテクチャー」で整理した。

［注目すべき国内外のプロジェクト］

米国では、IEEEが中心となって、Rebooting Computingキャンペーンを2013年から始めた。これに呼応する形でDARPAが各種のコンピューティングプラットフォームに関するプロジェクトを年々増やし始め、2018年にはこれをまとめる形で電子技術の復権を目指すERI（Electrics Resurgence Initiative）を立ち上げ[30)]、これから数年間で1600億円もの研究費をつぎ込むと報道されている。対象は、前述のドメイン・ス

ペシフィック・アーキテクチャーの各分野やそのLSI設計手法、テストシステム実証を中心として、それ以外にも新規半導体デバイスや3次元実装技術のシステム応用を含んでいる。

また、UC Berkeley発のイノベーションとして、ピュアなプロセッサーアーキテクチャーの世界で、オープンなプロセッサープラットフォームを標榜するRISC–Vアーキテクチャーが急速に求心力を高めていることも注目に値する。アーキテクチャーに新規性があるのではなく、寡占の度を強めるデファクト・プロプライエタリIPであるARMアーキテクチャーに対するアンチテーゼとして、オープンソースIPであることが最大の特徴であり、UC Berkeley発であるという正統性を強みにMakersムーブメントの上げ潮に乗ることに成功したように見える[8)]。 AI系ハードウェアのアクセラレーターを構成する際にもシステム全体の管理を行うプロセッサーは必須部品であり、ここにARMではなくRISC–Vを選択するプロジェクトが急速に増えている。2022年段階では、RISC–Vエコシステムの成熟化に伴い、IoT向けからデータセンター向けチップまで、RISC–V搭載をうたうチップの数が急激に増えてきている。同プロセッサーコアをライセンスする米国SiFive社によれば、2020年から年率73.6%で伸び、2025年には600億個を増えるという[31)]。例えばデータセンター向けAIアクセラレーターチップの1000並列プロセッサーコアとしてRISC–Vを採用する事例[32)] なども増えてきており、今後AI処理向けCPUコアとしてARMと並んでRISC–Vが大きな位置を占める可能性も出てきた。

中国では、中国政府や有力都市の行政府が、大規模な人工知能ハードウェアプロジェクトを始めている。その予算は年間1000億円以上といわれ、米国以上の規模を誇る。清華大学にはAI/ニューロモーフィック分野のハードウェア研究センターが設立され、北京市内の狭いエリアに隣接する清華大学、北京大学、中国科学院の関係者がキャンパス周辺にスタートアップ企業を次々に立ち上げる生態系が形成されている（一説には中国では現在30を超えるAIハードウェア系スタートアップがあるとのこと）。北京では国策によってこれらの動きが推進されている状況であるが、上海・深圳でもHuawei社やBAIDO社が中心となって、北京と同規模の民間ムーブメントが起きている模様である。このようなAI分野における中国の活況が、2020年になってからの米中の政治的軋轢の一つの原因だという指摘もあるが、AI分野をリードせんとする中国の勢いにブレーキがかかることになるのかどうか、現時点では先行き不透明の状況である。

欧州は、Human Brainプロジェクト等、ニューロモーフィック系のプロジェクトが歴史的に盛んに進められてきており、相対的には現在のアーキテクチャー革新の動静からは少し距離を置いたポジションに見える。英国に位置するARM社は米中からの遅れに危機感を感じてこの分野のR&Dに力を入れ始めている。イスラエルでは、画像処理・信号処理技術の強みを活かして、小規模ながらスタートアップ企業が蓄積し始めている。

日本では、2018年度から文科省の戦略目標「Society5.0を支える革新的コンピューティング技術の創出」をもとにJSTにおいてCREST、さきがけ研究領域が立ち上がり、これと並行して経産省–NEDOでも「高効率・高速処理を可能とするAIチップ・次世代コンピューティングの技術開発」事業が立ち上がった。投資規模で米中にはるかに劣ることは明白であり、日本が蓄積してきた技術的強みや産業界でのポジショニングを明確に意識した、勝てるシナリオ作りとそれに沿った研究開発戦略が求められるところである。

（5）科学技術的課題

爆発的なスマート化の進展に呼応した情報処理対象のドラスチックな変化により、情報処理アーキテクチャーは今大きな変革期を迎えている。いわゆる人工知能（AI）ブームの下、米国・中国を中心に活発な開発競争が数年来続いているが、今見えているAI技術は単なる氷山の一角であり、将来にはより豊潤な知能コンピューティングの世界が広がっていると推定して研究を強力に進めなければならない。なぜならば、一つには日進月歩で過去の常識を覆し続ける深層学習技術の爆発的発展がそれを示唆しているからであり、また一つには、それが真ならば、今後数十年にわたって発展し続けるであろう新しい情報処理アーキテクチャーのイニシアチブを取ることが世界的な社会のスマート化の大競争の中で決定的に重要だからである。

DNN技術の勃興以来、10年強がたとうとしており、これを主ターゲットとするいわゆるAIチップの開発は、もう飽和しているのではないかという観測が2018年頃から浮上していた。しかしながら、2022年においても、

DNN分野は日々新しいネットワークモデルや学習技術が浮上し、進化を続けている（例えば、ネットワークの枝刈り技術により深層学習を刷新する提案[28]、宝くじ仮説に基づく軽量DNN推論の提案など）。このような新規技術は今後も登場し続けることが予想され、現時点で最適なAIチップのアーキテクチャーが数年後も最適だという保証は全くない。

科学技術上の課題としては、短期的な「AIチップ開発競争」に勝つことが目標ではなく、この新しい時代にふさわしい本質的で持続可能な情報処理アーキテクチャーの変革を生み出すことを目標としなければならない。アーキテクチャーの世界は、Winner–Take–Allの世界であり、アーキテクチャー的に正しいからデファクトスタンダードになるというのではなく、近未来の応用分野に適した尖ったアーキテクチャーでニーズに応え、そのニッチな応用の爆発的な拡大とともにデファクトスタンダードに成長する。この観点から、エッジにおける知的なデータ処理を、そのデータ処理に適したドメイン・スペシフィックなアーキテクチャーで、リコンフィギュラブル、インメモリーなどの尖ったアーキテクチャー的な特徴を武器としながら、応用課題やソフトウェアのエコシステムと連携して発展させていくことが重要である。

（6）その他の課題

アーキテクチャー革新の好機との認識に立つ海外の著名計算機アーキテクチャー研究者は、データと研究資金を持つGoogle、NVIDIA、FacebookなどのAIプラットフォーマー企業を足場に次世代アーキテクチャーの研究を進めている。また、米国・中国は、AIハードウェアに1000億円規模の国家予算の投資を始めており、そのような企業群に比べて国家投資余力を持たない日本の立ち遅れは大きい。コンピューター産業競争力低下の影響を受けてアーキテクチャー研究分野から人材が流出してきた国内ではアーキテクチャー人材が払底しており、学生にも不人気な時期が長く続いていた。長らくアーキテクチャー研究を主導してきた米国では、アーキテクチャーを含むコンピューター科学の分野は継続的に人気先行であり、例えばMITのアーキテクチャー講義では、大型教室に学生が入りきらない状況になっている。一方、半導体とコンピューター技術振興を国策に掲げる中国では最優秀の学生層がこの分野に集中している。人材や次世代教育の点でも彼我の差は非常に大きい。

このような状況で日本が取るべきアプローチは、まずキャッチアップしなければならない状況を正しく反省することを出発点に、短中期的には過去の技術蓄積を再度掘り起こしながら尖ったアイデアの創出をプロモーションする戦略が重要である。また、長期戦略的にはコンピューティングアーキテクチャー分野の若手世代の育成を目立った形で始めることだと考える。後者に関して、中国がいかに若手世代を育成し技術をキャッチアップしてきたかを正しく理解することは、重要な一手ではないだろうか。

（7）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	△	→	・アーキテクチャー分野の国際学会ではプレゼンスはないに等しく、集積回路分野では過去の影響力を辛うじて保ってはいるが、新しいムーブメントを起こすような指導的立場は持てておらず、チャレンジャーポジションである。どちらの技術階層でも中高年層に優秀な人材は存在するが他の技術階層へのシフトが顕著であり、特に若年層については人材の層が薄い。
	応用研究・開発	△	→	・DNNのエッジコンピューティング分野ではルネサスエレクトロニクス社がEmbedded AIとその差別化IPとしてDRPコアを戦略的に開発・事業展開している。プリファードネットワークス社のMNコアも世界的レベルの競争力を有している。アニーリングマシン分野においては、集積回路型で日立・富士通のプレゼンスが見える以外は、総じて世界的プレーヤーとして活躍できていない。

米国	基礎研究	◎	→	・Stanford、MIT、UCB、Harvardなどの主要大学が積極的に先進技術を発信し続けている。産学連携、官学連携等も活発で研究資金や人材育成・供給の面で研究開発のエコシステムが活発に機能している。 ・DARPAの研究資金投入が突出して目立っている。
	応用研究・開発	◎	→	・Google、Facebookなどのプラットフォーマーがアーキテクチャー分野に投資し続け、NVIDIAもGPGPUへの先行投資が実ってDNN分野でデファクト企業の位置を積極的に狙って影響力を高めている。また、多数のスタートアップ企業が誕生し、世界の技術開発をリードしている。
欧州	基礎研究	○	→	・ニューロモーフィック分野での活動が目立つ。 DNN分野ではKU Lueven、EPFL等の少数の大学が世界的に見てレベルの高い研究を進めている。
	応用研究・開発	○	→	・ARM社がドメインスペシフィック分野に事業戦略のかじを切り始めた。 ・英国のGraphCoreやイスラエルのCEVAなど、技術的な特徴が鮮明なスタートアップ企業も存在する。
中国	基礎研究	◎	→	・アーキテクチャー分野では優れた研究成果を発表し続けており、集積回路分野でもそのレベルに達しつつある。巨額の国費を投資して技術開発を振興しており、清華大学や中国科学院の成果が目立っている。
	応用研究・開発	◎	↗	・米国には劣るものの、BAIDUやAlibabaなどのプラットフォーマーを有し、積極的にアーキテクチャー分野に投資している。30社以上のAIハードウェアスタートアップ企業が生まれているといわれており、例えば、その一つであるユニコーン企業のCambricon社の技術がHUAWEI社のスマートフォンに搭載されている。2020年のHotChipsでは、BAIDUとAlibabaのAIチップ（DNN推論）が注目を集めていた。
韓国	基礎研究	○	→	・アーキテクチャー分野でも集積回路分野でも、日本よりもはるかにプレゼンスが大きい存在となっている。 KAISTやソウル大学がメインプレーヤー。
	応用研究・開発	△	→	・Samsung社の研究開発動向が垣間見える程度で、産業界全体の状況は不明。
台湾	基礎研究	○	→	・国立清華大学や台湾大学などで、メモリー回路をベースにしたインメモリー・コンピューティング技術の研究開発が盛んに行われている。
	応用研究・開発	△	→	・TSMCがLSIファウンダリーとして世界的に1強の地位を占めているが、プロセッサー分野でのプレゼンスはあまりない。

（註1）フェーズ

基礎研究：大学・国研などでの基礎研究の範囲

応用研究・開発：技術開発（プロトタイプの開発含む）の範囲

（註2）現状　※日本の現状を基準にした評価ではなく、CRDSの調査・見解による評価

◎：特に顕著な活動・成果が見えている　　○：顕著な活動・成果が見えている

△：顕著な活動・成果が見えていない　　×：特筆すべき活動・成果が見えていない

（註3）トレンド　※ここ1～2年の研究開発水準の変化

↗：上昇傾向、→：現状維持、↘：下降傾向

関連する他の研究開発領域

・革新半導体デバイス（ナノテク・材料分野 2.3.1）

参考文献

1）John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6th ed.（Morgan Kaufmann, 2017）.

2）Toshitsugu Yuba, et.al., “The SIGMA-1 dataflow computer,” in ACM '87: Proceedings of the 1987 Fall Joint Computer Conference on Exploring technology: today and tomorrow（IEEE Computer Society Press, 1987）, 578-585.

3）Yoshinori Yamaguchi, Shuichi Sakai and Yuetsu Kodama, “Synchronization Mechanisms of a Highly Parallel Dataflow Machine EM-4,” IEICE Transactions E74-D, no.1（1991）: 204-213.

4）Xiao-Ping Ling and Hideharu Amano, “WASMII: a data driven computer on a virtual hardware,” in Proceedings IEEE Workshop on FPGAs for Custom Computing Machines（IEEE, 1993）, 33-42., https://doi.org/10.1109/FPGA.1993.279481.

5）本村真人, 他「新世代マイクロプロセッサアーキテクチャ（後編）：3. 実例4. 動的再構成プロセッサ（DRP）」『情報処理』46 巻 11 号（2005）: 1259-1265.

6）村岡洋一, 古谷立美『知的連想メモリマシン』（東京: オーム社, 1989）.

7）Parthasarathy Ranganathan, “More Moore: Thinking Outside the（Server）Box,” 44th International Symposium on Computer Architecture（ISCA）, 24-28 June 2017, https://iscaconf.org/isca2017/doku.php%3Fid=wiki：main_program.html,（2023年2月5日アクセス）.

8）David Patterson, “50 years of computer architecture: From the mainframe CPU to the domain-specific tpu and the open RISC-V instruction set,” in 2018 IEEE International Solid - Sate Circuits Conference（ISSCC）（IEEE, 2018）, 27-31., https://doi.org/10.1109/ISSCC.2018.8310168.（講演ビデオ: https://www.youtube.com/watch?v=NZS2TtWcutc）.

9）Norman P. Jouppi, et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit,” in Conference Proceedings of 44th Annual International Symposium on Computer Architecture（New York: Association for Computing Machinery, 2017）, 1-12., https://doi.org/10.1145/3079856.3080246.

10）Yu-Hsin Chen, Joel Emer and Vivienne Sze, “Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks,” in 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture（ISCA）（IEEE, 2016）, 367-379., https://doi.org/10.1109/ISCA.2016.40.

11）Thomas Norrie, et al., “Google’s Training Chips Revealed: TPUv2 and TPUv3,” in 2020 IEEE Hot Chips 32 Symposium（HCS）（IEEE, 2020）, 1-70., https://doi.org/10.1109/HCS49909.2020.9220735.

12）Chris Nicol, “Wave Computing: A Dataflow Processing Chip for Training Deep Neural Networks,” 2019 IEEE Hot Chips 29 Symposium（HCS）, 20-22 August 2018, https://hc29.hotchips.org/,（2023年2月5日アクセス）.

13）Jeremy Fowers, et al., “A Configurable Cloud-Scale DNN Processor for Real-Time AI,” in 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture（ISCA）（IEEE, 2018）, 1-14., https://doi.org/10.1109/ISCA.2018.00012.

14）Shouyi Yin, et al., “An Ultra-High Energy-Efficient Reconfigurable Processor for Deep Neural Networks with Binary/Ternary Weights in 28NM CMOS,” in 2018 IEEE Symposium on VLSI Circuits（IEEE, 2018）, 37-38., https://doi.org/10.1109/VLSIC.2018.8502388.

15）中原啓貴「FPGAを用いたエッジ向けディープラーニングの研究開発動向」『人工知能』33 巻 1 号（2018）: 31-38., https://doi.org/10.11517/jjsai.33.1_31.

16）Kodai Ueyoshi, et al., “QUEST: A 7.49TOPS multi-purpose log-quantized DNN inference engine stacked on 96MB 3D SRAM using inductive-coupling technology in 40nm CMOS,” in 2018 IEEE International Solid-State Circuits Conference（ISSCC）（IEEE, 2018）, 216-218., https://doi.org/10.1109/ISSCC.2018.8310261.
17）Taro Fujii, et al., “New Generation Dynamically Reconfigurable Processor Technology for Accelerating Embedded AI Applications,” in 2018 IEEE Symposium on VLSI Circuits（IEEE, 2018）, 41-42., https://doi.org/10.1109/VLSIC.2018.8502438.
18）小島郁太郎「ルネサスがAI推論使うビジョン処理MPU、動的変更DRPと専用MACで電力効率急上昇」日経XTECH, https://xtech.nikkei.com/atcl/nxt/news/18/08080/,（2023年2月5日アクセス）.
19）岡林凛太郎「PFNのスパコン「MN-3」が世界1位に、消費電力性能ランキングのGreen500で」日経XTECH, https://xtech.nikkei.com/atcl/nxt/news/18/08188/,（2023年2月5日アクセス）.
20）Kazutoshi Hirose, et al., “Hiddenite: 4K-PE Hidden Network Inference 4D-Tensor Engine Exploiting On-Chip Model Construction Achieving 34.8-to-16.0TOPS/W for CIFAR-100 and ImageNet,” in 2022 IEEE International Solid- State Circuits Conference（ISSCC）（IEEE, 2022）, 1-3., https://doi.org/10.1109/ISSCC42614.2022.9731668.
21）Mohamed M. Sabry Aly, et al., “Energy-Efficient Abundant-Data Computing: The N3XT 1,000x,” Computer 48, no. 12（2015）: 24-33., https://doi.org/10.1109/MC.2015.376.
22）Paul A. Merolla, et al., “A million spiking-neuron integrated circuit with a scalable communication network and interface,” Science 345, no. 6197（2014）: 668-673., https://doi.org/10.1126/science.1254642.
23）Luping Shi, et al., “Development of a neuromorphic computing system,” in 2015 IEEE International Electron Devices Meeting（IEDM）（IEEE, 2015）, 4.3.1-4.3.4., https://doi.org/10.1109/IEDM.2015.7409624.
24）Mike Davies, et al., “Loihi: A Neuromorphic Manycore Processor with On-Chip Learning,” IEEE Micro 38, no. 1（2018）: 82-99., https://doi.org/10.1109/MM.2018.112130359.
25）D-Wave Systems Inc., “The D-Wave 2000Q™ Quantum Computer Technology Overview,” D-Wave, https://dwavejapan.com/app/uploads/2019/10/D-Wave-2000Q-Tech-Collateral_1029F.pdf,（2023年2月5日アクセス）.
26）山岡雅直「組合せ最適化問題に向けたCMOSアニーリングマシン」『IEICE Fundamentals Review』11巻3号（2018）: 164-171., https://doi.org/10.1587/essfr.11.3_164.
27）富士通株式会社「AIと量子コンピューティング技術による新時代の幕開け：デジタルアニーラが未来を切り拓く」FUJITSU JOURNAL, 2017.
28）Kasho Yamamoto, et al., “7.3 STATICA: A 512-Spin 0.25M-Weight Full-Digital Annealing Processor with a Near-Memory All-Spin-Updates-at-Once Architecture for Combinatorial Optimization with Complete Spin-Spin Interactions,” in 2020 International Solid-State Circuits Conference（ISSCC）（IEEE, 2020）, 138-140., https://doi.org/10.1109/ISSCC19947.2020.9062965.
29）本村真人「AIエッジコンピューティングへの期待と展望」『Okiテクニカルレビュー』8巻2号（2019）: 4-7.
30）Defense Advanced Research Projects Agency（DARPA）, “DARPA Electronics Resurgence Initiative,” https://www.darpa.mil/work-with-us/electronics-resurgence-initiative,（2023年2月5日アクセス）.
31）小島郁太郎「中印激増で25年にRISC-V搭載IC累計600億個、最大手から車載向けコアも」日経

XTECH, https://xtech.nikkei.com/atcl/nxt/column/18/01537/00403/,（2023年2月5日アクセス）.
32）Hisa Ando「Esperantoの低電力メニーコアMLサーバプロセッサ「ET-SoC-1」、Hot Chips 33」TECH+, https://news.mynavi.jp/techplus/article/20210825-1955136/,（2023年2月5日アクセス）.