

[知能コンピューティングー AI とハードウェアの出会いー]

1 AI は新しいハードウェアを欲しているか？

—知能と計算とアーキテクチャの新しい関係—

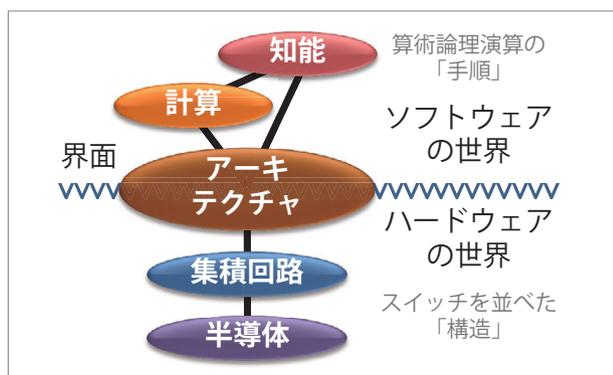
本村真人 東京工業大学



計算機に求められる中心的な処理は、人間がより知的な作業に集中できるように補完したり、人間のより良い判断を助けたりする「知能コンピューティング」へと急速に変貌しつつある。これに呼応する計算機のアーキテクチャ技術は、従来通りの正常進化で良いのだろうか。それともその処理の特質を生かして、より高度で効率的な処理の実現に向けて変化するべきなのだろうか。

知能，計算，アーキテクチャ

この問いについて考える前に、まずこれら3つの言葉の間に存在する関係性について論じてみたい(図-1上)。ただし、ここでは知能という言葉をも、主に工学的に実現される疑似的な知能、すなわち人工知能(AI: Artificial Intelligence)ないしは機械知能(Machine Intelligence)を指す言葉としても用いている。



知能と計算

当たり前のことではあるが、その急速な技術展開により社会応用が広がりつつあるAI技術は、すべて何らかの計算機の上で実行されている。すなわち、人工的な知能は、すべて何らかの計算過程に分解され、計算アルゴリズムとして再構築され、計算機上のアプリケーションプログラムとして実行されている。この在り方は、人間における知能の本来的な発現形態とは大きく違うものの、現在の情報処理技術を踏まえた工学的アプローチとしては至極まっとうである。21世紀は、まさしくデータの世紀であると同時にアルゴリズムの世紀である。

計算という言葉は、ComputationとCalculationの両方の意味を持つが、前者の解釈に絞ったとしても、その本質は何かという問いに真正面から答えることは案外難しい。計算機を頭に思い浮かべた上で一般的な見立てとしては、算術論理演算を主たる構成要素とする一連の手順だとは言えるが、それは果たして知能を支える計算技術としては適切と言えるであろうか。たとえば、AIの中心技術として君臨する深層学習(Deep Learning)ないしは深層ニューラルネットワーク(DNN: Deep Neural Network)分野は、大量の積和演算を必要とすることは知られているものの、認識・推論などの機能を実現しているカギは、実は積和演算の間々に埋め込まれた非線形演算であると言われている。また、古くて新しい計算概念として近年注目されるようになったリザー

特集

Special Feature

バ計算^{☆1}の世界では、非線形性がその中心的位置を占めている。さらに、生体神経回路網に学ぶニューロモルフィック^{☆2}な知能アプローチの場合、その基本的な仕組みは、時間的な信号系列の蓄積と確率的な発火として定義されている。しかし、従来の計算機には非線形演算は定義されていないし確率的な挙動も織り込まれていない。これらはいずれも、アルゴリズム上の工夫により、計算機が備える計算機構で実行しやすいように変換ないしは近似され、代替されている。

さらに視野を広げて、自然計算という言葉で知られているような、自然界の現象や物理現象を使って等価的に計算を行うモデルや、自然界のエネルギー最小化原理を模擬して最適化問題や条件充足問題に対する解を探索するアニーリング計算の概念まで考えに入ると、果たして計算とは何か、知能を支える適切な計算技術はいかなるものか、という問いの奥深さはさらに増す¹⁾。少なくとも、現代の計算機が備えている計算機構は必ずしも知能を実現するために適したのではなく、逆に、知能を実装するためのアルゴリズムが計算機の現実に寄り添い、その計算過程になじむように知能を分解・再構築してきたのだと言えそうである。

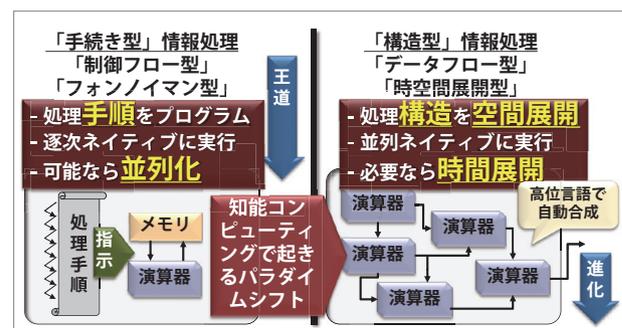
計算とアーキテクチャ

現在の計算機は、ほぼすべてフォンノイマン型アーキテクチャに則っているとって過言ではない。その本質は、処理対象を基本演算ステップに分解し、これを手続き（すなわち制御フロー構造）として表現して、その指示に従って、メモリとALU（Arithmetic Logic Unit：算術論理演算器）の間でデータをやり取りしつつ「手続き型処理」の計算を行うことである（図-2左）。この成り立ちを背景に、これまでの計算機アーキテクチャの主

流の研究は、永らく、1) その場の判定で制御フローを動的に変える条件分岐を無害化する予測・先読み等の技術や、2) 頻繁なメモリアクセスが性能律速要因となるフォンノイマンボトルネックを緩和するためのメモリーシステム階層構築の技術の2つの分野に集中してきた。

一方で、フォンノイマン型に代わる「非ノイマン型」のアーキテクチャ思想を打ち立てる挑戦的な研究も古くから続いている。1980～1990年代を中心として、その一派としてデータフロー型計算機が盛んに研究されていた。電総研（現産総研）を始めとして日本でも有力な研究がいくつも進められ（SIGMA-1, EM4, Impp等）、第5世代コンピュータプロジェクトでも「知識情報処理を支える新時代の計算機アーキテクチャ」として注力対象として取り上げられた。これは、主に処理並列化の観点から、制御フローに基づく実行制御を捨て、演算が可能になったデータから処理すること（すなわちデータフローに従って並列計算すること）を是とする計算パラダイムであった。

データフロー型計算機概念は、一時期の研究の盛り上がりした後、フォンノイマン型プロセッサにおけるマルチスレッド実行や順序外命令並列実行、あるいはマルチプロセッサ環境でのスレッド並列実行機構などの形で技術展開・貢献しながら、アーキテクチャ流派としては静かに主流派に吸収されていった。その理由の1つとして、制御フローのくびきから離れて並列実行することにより、分



■ 図-2 手続き型と構造型

☆1 非線形素子の相互結合網による時系列データの学習を主眼とする再帰型ニューラルネットの一種。

☆2 ニューラルネット分野と比べ、より忠実に生体神経回路網の動作・構成を忠実に模擬することを志向する知能処理アプローチ。

岐実行やメモリアクセスの局所性が崩れ、結果的には実行効率が下がってしまうという問題があった。すなわち、データ駆動並列実行による時間的・空間的局所性の棄損というマイナス面の顕在化である。その技術展開の経緯は、計算機アーキテクチャが制御フローの呪縛からは離れられず、制御フローの統治の中での民（計算ステップ）の自由として、データフロー並列が生き残った、という見立ても成り立とう。また、後述のムーア則に基づく単体プロセッサ高速化の波に飲み込まれたという面や、そもそも当時の計算機の処理対象ワークロードが手続き処理型向きだったという面も指摘できる。これらについては後述する。

知能とアーキテクチャ

脳の情報処理については分からないことがまだ多いと言われているし筆者の専門からも遠い。ここでは、知能的な処理の種類と効率という観点で、大脳と計算機のアーキテクチャをきわめて大雑把に比較してみたい。

2015年、AlphaGoが最高位の棋士を囲碁で破ったことが話題になった。AlphaGoが稼働していた計算サーバの消費電力は250KW程度であったと言われている。一方、人間の脳の消費電力は20W程度と言われており、当時の囲碁で競っていたということは、このゲームの情報処理の電力効率としては、大脳の方が10K倍以上高かったということの意味しよう。これを以て大脳のエネルギー効率の良さを喧伝する向きもあるが、それは事柄の一面しか捉えていない。別の極端な例として、四則演算をひたすら続けるような計算課題を考えてみよう。ざっくり勘定してみると、GHzを超える高速動作クロックと数千以上に及ぶ並列演算の威力により、この場合は計算サーバの方が大脳よりも1億倍以上エネルギー効率が高いことが分かる。

この簡単な目の子勘定から明らかになることは、求められる情報処理の種類により、最適なアーキテ

クチャが異なるということである。大脳はより高度で直感的な知能の実現に、計算機はより低位で機械的ないしは論理的な情報処理には向いており、その向き不向きが処理効率に直結しているという仮説が成り立ちそうである。多量のALUを核に持つ計算サーバのアーキテクチャにとって、四則演算はネイティブに実行できる演算である。一方、囲碁等の知能的なプログラムに関しては、その基本的な計算機構の上に、仮想的な計算機構を上部構造として装着（オーバーレイ）して実行していると言える。

すなわち、この例は、知能を実現しそのエネルギー効率を上げるためには、その目的に沿ったアーキテクチャを構築してなるべくオーバーレイ計算構造を避ける、すなわち、極力ネイティブ実行可能なアーキテクチャを構築する戦略が必要なことを物語っている。

アーキテクチャと集積回路と半導体

前節では、知能の実現に向けて、その処理の特徴に合わせて計算の基本要素やアーキテクチャの構成を最適化することの重要性を論じた。計算機の世界において、アーキテクチャとはソフトウェアとハードウェアを結ぶ界面であり、両者の間の約束事を定義するものである。今度はハードウェアの世界からアーキテクチャを眺めてみたい（図-1下）。

半導体と集積回路

昔、「コンピュータ、ソフトがなければただの箱」という言葉があった。それに倣えば「半導体、回路が載らなきゃ高い石」であり、集積回路すなわちシリコン半導体上に大量のスイッチを並べ接続した構造物は、その価値の源泉である。

電子工学の世界で最も有名なムーアの法則、すなわち「集積回路の密度は1.5年で2倍になる」は、1960年近くにわたり続いてきた。近年の新型コロナウイルス感染症により指数関数による爆発的な増大の

脅威は人類共通の認識となったが、情報処理—電子工学の世界は、60年間にわたってこのムーアの指数関数則とともに生き、その威力を味わってきたと言える。

ムーアの法則の根っことは、加工最小線幅を1.5～2年で0.7倍にするという半導体製造技術上の努力目標であり、それが長年達成し続けられたことで、集積回路の継続的規模拡大・速度向上・電力効率向上が続いてきた。その意義は、半導体の物理的な技術則の観点よりは、むしろ集積回路の経済則として見たときに正しく理解できる。その意味するところは、大まかに言えば、同一機能・性能を実現するための集積回路の価格は、数年で1/10になる、ということである。この強烈で永続的なコモディティ化が情報処理技術の発展とそれによる社会変革を支えてきた²⁾。

集積回路とアーキテクチャ

そのムーアの法則が続くかどうか、ここ数年にわたって議論が尽きない。ただ、半導体微細化観点での技術則としてはまだ続こうとも、集積回路経済則としての役割は終えようとしている。それは、あまりにも高難度化・高製造コスト化が進んだ最先端技術が半導体製造の寡占構造を生むとともに、その生命線であったコスト低減によるコモディティ化を生み出さなくなっているからである。

これまでの集積回路の性能向上の主たる貢献は実はムーアの法則によるものであり、純然たるアーキテクチャによる貢献分は実は少ないと言われている。その背景には、ムーアの法則によるチップ集積の境界線の絶え間ない更新を性能向上に反映する作業にアーキテクチャ側が忙しかったからであり、時の経過とともに性能が向上する、という流れの中で大胆なアーキテクチャ更新を行うことが難しかったからでもある。

今やその流れは変わろうとしている。ムーアの法則による半自動的な集積回路の性能向上・効率向上・コスト低減に期待を持てなくなった現在、今後の継

続的な情報処理の発展のためには、アーキテクチャ技術を前面に押し出したポストムーアの技術革新が真に求められている。

知能を支える新しいハードウェアの形

ここまで、図-1上部に関する前半の議論の指し示すところは、知能コンピューティングの実現や効率化に向けて計算とアーキテクチャの変革が必要であるという、ソフトウェア側からのプルの存在であった。一方、図-1下部に関する後半の議論が導き出すのは、ムーアの法則の減衰に伴い今こそアーキテクチャの貢献が求められるというハードウェア側からのプッシュの存在である。

このプルとプッシュの2つの意味で、より進んだ知能を実現するためのアーキテクチャ技術の研究が、より重要になっている。世界的な「アーキテクチャ黄金時代」の到来は（残念ながら日本のコミュニティが十分にその中に入れていたとは言いがたいが）、至極当然の帰結である。

知能コンピューティング技術

福島によるネオコグニトロンに源流を発する畳み込みニューラルネット（CNN）が画像分類精度で従来手法を大きく超えることが2012年に報告³⁾され、大規模学習データ、高性能計算機、さまざまな学習手法の改善も相まって、CNNに代表される深層ニューラルネット技術（DNN）は一躍脚光を浴びることとなった。この知能コンピューティング技術の代表的系譜が対象としているのは、計算機の「入力」となる大量のデータ（データ爆発）からいかに人間にとって役立つ知見を得るか、という課題である。

一方、DNNに少し遅れて、種々の組合せ最適化問題をスピン格子のエネルギー最小化問題に置き換え、その近似解を並列に求めるアニーリング計算の分野も広く注目を集めている。その発端は量子ア

特集
Special Feature

ニーリングに対する興味の高まりだが、今やアニーリング計算は複雑化する社会に必要な技術として多面的な研究が進みつつある。この智能コンピューティング技術の系譜は、計算機の「出力」である計算結果の爆発に関係する。大量に存在し得るさまざまな組合せ結果（組合せ爆発）の中から、良い解を選択して提供するという課題である。技術的には、DNNの学習はニューロン間結合重み係数空間のエネルギー最小化問題であるが、アニーリング計算モデルはその逆問題（重み係数を固定して、スピン[バイナリニューロンとも見なせる]値空間のエネルギー最小状態を探索する問題）であり、深層学習との関連性は深い。

さらに、DNNの興隆の影響を受けて、生体神経回路網の動作を出来る限り精密に模擬することを通して知能の実現を探るニューロモルフィック分野も活性化している。生体模倣の目的や工学的な意義については慎重に考える必要がある（鳥を忠実に模倣しても飛行機は実現できない）が、先に述べた計算の基本構成要素の違いを探求し、その本質を「智能コンピューティング」に向けて融合していくことができるならば、それは大きな意味を持ち得ると言えよう。

構造型と時空間展開型

筆者は、この変革のチャンス（プル）を活かし差し迫った要求（プッシュ）に応えるために、「構造型」情報処理の拡大と発展を支える「時空間展開型」アーキテクチャ技術をいかに構築するか、が重要な課題であると考えている。

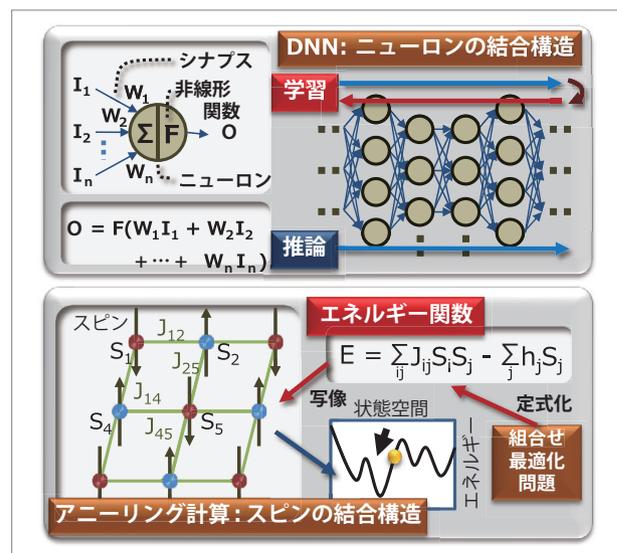
DNNは、大量・多層に並べられたニューロン間の複雑な結合網という「構造」の中に入力データストリームを流し込んで学習や推論を行うという意味において、「構造型」の情報処理課題である。一方のアニーリング計算も、相互作用を持つスピン（すなわちバイナリニューロン）の状態を、系のエネルギー状態に応じて次々に更新していくという点において、同

様の意味で「構造型」の情報処理課題である（図-3）。

従来の情報処理の中心は「手続き型」課題であったため、制御フローを中心に据えるフォンノイマン型が王道であった。構造型の情報処理課題に対してはどのようなアプローチがネイティブであろうか。ここでいう「構造」とはデータフロー構造そのものであり、そこに制御フローはほとんど存在しない。当然の帰着として、それはデータフロー型であろう。すなわち、計算ステップに分解して手続き的な構造をオーバーレイして実行するのではなく、データフロー構造をネイティブに実行できるアーキテクチャが求められていると言える（図-2右）。

これをハードウェア技術の観点で見た言葉が、時空間展開型アーキテクチャとなる。その基本的なコンセプトは、

- 処理対象のデータフロー構造を空間展開してハードウェア要素にマッピングし、処理対象に内在する並列性をそのまま活かして並列実行する
- ハードウェア量の物理制限が存在するため、そのままでは解ける問題のサイズが制約されてしまう。そこで、いったん空間展開された構造を時間展開する。すなわち、空間構造を時間的に切り替えながら実行する



■ 図-3 構造型情報処理

ことであり、これは、動的再構成（ダイナミックリコンフィギュラブル）ハードウェアとして知られている技術コンセプトそのものである。この分野は、実は AI 分野と並んで日本の研究コミュニティによる技術蓄積が豊富な分野であることは指摘しておきたい⁴⁾。

研究事例と課題

筆者らのグループはこれまで新しい智能コンピューティング向けハードウェア技術の確立に向けた取り組みとして、いくつかの研究成果を発表してきた⁵⁾。代表的な例をいくつか簡単に紹介する（より最新の成果はほかの稿に譲る）。

2016 年から、バイナリ DNN のアルゴリズム（重み係数／ニューロン値を 2 値で表現して DNN の演算を大幅に軽量化するアルゴリズム）に一早く注目してアーキテクチャと実行方式の研究を進め、2017 年の国際会議 VLSI シンポジウムで世界初のバイナリ DNN 推論エンジン LSI-BRein Memory を発表した（図-4[1]）。メモリに密結合した並列回路

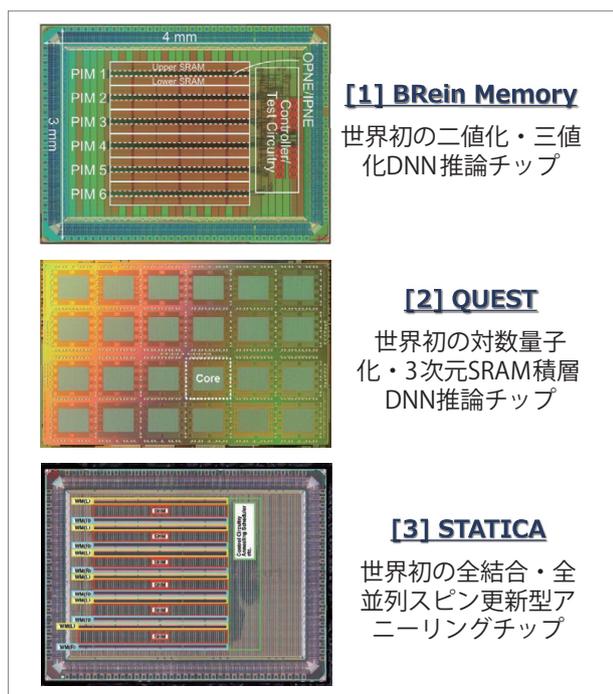
でバイナリ DNN 処理をニアメモリ・リコンフィギュラブル型で処理するとともに、容易に多段接続拡張して DNN のレイヤ数増大に対応可能な HW 方式が最大の特徴であり、CPU/GPU/FPGA に比べてそれぞれ 30K 倍／3K 倍／1K 程度の高いエネルギー効率を実現可能なことを実証した。

さらに、1 ビットから最大 4 ビットまでの対数量子化（2 のべき乗表現の重み係数／ニューロン値の指数を量子化）アルゴリズムをスケーラブルにカバーする DNN 推論アーキテクチャを提案し、用途ごとに推論精度と HW 量をバランスすることが可能な新たなメモリ密結合型・リコンフィギュラブル型 DNN 処理エンジン LSI-QUEST を実現した（図-4[2]）。短レイテンシ・高バンド幅の SRAM と 3 次元積層することで DNN 推論時のメモリボトルネック問題も解消した上で、畳み込み層・全結合層を含む DNN 全般に広く適用可能な柔軟なビットシリアルプロセッサアレイアーキテクチャを実現しており、2018 年の国際会議 ISSCC で発表している。

また、先述の、アニーリング計算におけるスピン系とバイナリ化された DNN の間の類似性に着目し、上記のバイナリ DNN チップのアーキテクチャの系統を踏みつつ、アニーリング計算のアクセラレータアーキテクチャにも踏み込んで研究を進めてきた。並列にスピンを更新できる新しいスピン更新モデルとして、新しく確率的セルラーオートマトンモデルを提案するとともに、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 LSI-STATICA を国際会議 ISSCC2020 で発表している（図-4[3]）。

Hope or Hype?

DNN の認識能力が人間を超えつつある現在でも、先に述べたような、脳と DNN のエネルギー効率の大きな違いはどこから（マクロレベルの情報処理方



■図-4 筆者らのグループの研究事例

特集

Special Feature

式、メゾなアーキテクチャ、マイクロな計算ステップ等)どのように生まれるのか、まだ解明されていない。今後、構造型情報処理と時空間展開型アーキテクチャを出発点としながら、計算—アーキテクチャ—回路の多層にまたがる協創により、大脳レベルに迫る高エネルギー効率ハードウェアの実現を目指した総合的な知能コンピューティングの研究が必要であろう。

本特集の著者らは、ここまで述べてきた問題意識を踏まえ、1) DNN 処理エンジンのアーキテクチャ技術やその画像処理応用技術を中核として、2) 組合せ最適化問題をエネルギー最小化問題に変換して並列に近似解を導出するアニーリング計算機技術、3) より忠実に生体神経回路網を模倣することを目指したニューロモルフィック HW 技術、4) 非線形性からの知能の発現を探索するリザーバ計算技術、など関連分野の新たな知見・研究進展を総合的に結集し、既存 DNN 処理の枠を超えて発展する知能コンピューティングを支えていく革新的アーキテクチャ基盤技術の構築を目指す研究を展開している⁶⁾。ここで特に重要となる課題は、個々の知能処理に特化したアプリケーションスペシフィックなアプローチではなく、知能コンピューティングの領域(ドメイン)を過不足なく俯瞰しつつその共通要求を定義し、以てドメインスペシフィックな基盤アーキテクチャ技術を確立していくことであると感じている。

DNN の勃興に呼応したアーキテクチャ研究の隆盛や新しい DNN アクセラレータの相次ぐ提案は、一過性のブームではないかという否定的な見方もある。しかし、過熱している面は否めないが、筆者にはとても一過性のムーブメントとは思えない。それは、本特集の他稿にも現れているように、知能コンピューティング技術自体がまだ発展途上で次々に新しい知見が見つかっている状況だからであり、現時点で優れたハードウェア技術がたとえば十年後も最適である保証などどこにもないからである。

この重要な分野は、国際的な競争の最前線でもある。昨今、日本でも経済安保的な側面での「半導体戦略」が取りざたされているが、それに加えて(あるいはそれ以上に)、知能コンピューティングに向けた国際技術競争と協創の「集積回路戦略」が必要なのではないか、と強く思う。先人たちの努力により、AI 分野、データフローアーキテクチャ分野、動的再構成ハードウェア分野など、コアとなる技術領域に、それぞれの冬の時代と言われた期間にも継続されてきた日本の技術開発の強みが集積されている。これらが今後の知能コンピューティングを支える強固な技術基盤となって結実することを願ってやまない。

参考文献

参考文献

- 1) 丸山 宏: 計算の未来と社会, https://japan.cnet.com/blog/maruyama/2019/08/21/entry_30022971/
- 2) 池田信夫: 過剰と破壊の経済学, アスキー新書(2007).
- 3) 岡野原大輔: ニューラルネットの逆襲, <https://tech.preferred.jp/ja/blog/deep-learning/>
- 4) 天野英晴 他: FPGA の原理と構成, 8 章, オーム社(2016).
- 5) 本村真人 他: 深層ニューラルネットワーク向けプロセッサ技術の実例と展望, 信学会論文誌 C, Vol.J103, No.5, pp.288-297.
- 6) 科研費 基盤 S, <https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-18H05288/>

(2021 年 12 月 6 日受付)

■本村真人(正会員) motomura@artic.iir.titech.ac.jp

1987 年京大修士(物理学), 1996 年同博士(工学). 1987 年~2011 年 NEC 研究所および NEC エレクトロニクス, 1992 年 MIT 研究員. 2011 年 北大教授, 2019 年 東工大教授. AI コンピューティングの研究等に従事. 1992 年 IEEE JSSC Best Paper, 1999 年 本会最優秀論文, 2011 年 信学会業績賞, 2018 年 ISSCC Silkroad Award を各受賞. 信学会, 人工知能学, 日本工学アカデミー正会員, IEEE Fellow.