ワークショップ報告書
CRDS-FY2024-WR-04

俯瞰ワークショップ報告書

コンピューティングアーキテクチャー

CRDS 研究開発戦略センター
Center for Research and Development Strategy

国立研究開発法人科学技術振興機構

エグゼクティブサマリー

本報告書は、2024年6月12日に開催した俯瞰ワークショップ「コンピューティングアーキテクチャー」の内容をまとめたものである。

これまでコンピューターの性能の劇的な進歩を引き起こしてきたムーアの法則に陰りが見え、昨今の生成AIに代表されるようにさらなる性能への要求、新しいワークロードへの対応が迫られている。また、自動運転におけるリアルタイムへの要求やエッジコンピューティングへの期待も高まっている。さらに、処理能力の高速化とともに省電力化は地球規模的に重要な課題となっている。

このような状況において、従来の考え方とは一線を画する新しいコンピューティングアーキテクチャーが必要ではないかと考え、今回のワークショップを企画した。本ワークショップにおいては、新たなコンピューティングパラダイムや、それを実現するプロセッサー、ドメインに特化したアーキテクチャーなどについて議論を行う。そして、今後の我が国の研究開発を強化する戦略となりうる方向性や研究開発テーマ、それらの課題などを明確にすることを目的としている。

まず、招聘した有識者6名から、現在の研究、注目される技術、学会などの動向について、ポジショントークがなされた。ついで、総合討議を行い、下記のような議論を行った。

1. Device ScienceとComputer Scienceの連携
   デバイスとコンピューティングは別物ではなく、相互に連携しながら、相互に発展するものである。したがって、両者の緊密な連携が必要である。また、デバイスとコンピューティングという分け方だけではなく、デバイス、回路、アーキテクチャー、アルゴリズムという構造もあるし、あるいはさらに別の新しい構造化の可能性もある。いずれにしても、複数の階層にまたがる連携が必要である。
   こういった複数の領域にまたがる連携を進めるにはファンディングが重要な役割を果たす。連携したチームを作るための施策が必要であり、事前段階としてチームビルディングの活動が必要である。
2. 基本ソフトウェアの重要性
   新しいデバイスとコンピューティングだけで何かが動くわけではない。地味ではあるが、コンパイラー、OS、EDA設計などがないと本当に動くものにはならない。周辺技術が重要である。
3. エネルギー効率の向上
   今回のワークショップの議論では、多くのコンピューティングは結局計算エネルギー効率の向上を目指している。そう言った状況においては、脳に学ぶということが一つのアプローチである。
4. 継続することが重要
   アーキテクチャーの研究者は数が減っており、このままでは絶えてしまう可能性がある。これまでにいくつか関連するファンディングがなされてきたが、これらの活動を継続することがこれまで以上に重要になっている。

本ワークショップにおける議論を受け、具体的な提案を行うための検討を進める予定である。

Executive Summary

This report summarizes the contents of the panoramic view workshop "Computing Architecture" held on June 12, 2024.

Moore's Law, which has brought about dramatic advances in computer performance, is showing signs of slowing down, and we are being forced to respond to demands for even greater performance and new workloads, such as the recent emergence of generative AI. In addition, there are growing demands for realtime operation in autonomous driving and expectations for edge computing. Furthermore, power saving along with the increase in processing speed has become an important issue on a global scale.

In light of this situation, we believe that a new computing architecture that breaks away from conventional thinking is needed, and that is why we planned this workshop. In this workshop, we will discuss new computing paradigms, the processors that realize them, and domain specific architectures. The purpose is to clarify the direction, research and development themes, and associated challenges that could serve as strategies to strengthen research and development in Japan in the future.

First, the six invited experts gave position talks on current research, noteworthy technologies, and trends in academic circles. This was followed by a general discussion, with the following points being discussed:

1. Collaboration between Device Science and Computer Science
   Devices and computing are not separate things; they work together and evolve together. Therefore, close cooperation between the two parties is necessary. In addition to the division into devices and computing, there are also structures such as devices, circuits, architectures, and algorithms, or even the possibility of other new structure. In any case, collaboration across multiple levels is necessary. Funding plays an important role in promoting collaboration across multiple fields like this. Measures are needed to create a cohesive team, and team-building activities are necessary as a preliminary step.
2. The Importance of System Software
   New devices and computing alone won't make anything work. It may seem mundane, but it won't actually work without a compiler, OS, EDA design, etc. Peripheral technologies are important.
3. Improving energy efficiency
   The discussion at this workshop highlighted that much of computing ultimately aims to improve computational energy efficiency. In such situations, learning from the brain is one approach.
4. Continuity is key
   The number of architecture researchers is declining, and if things continue as they are, they may become extinct. Several related funding efforts have been made to date, but it is more important than ever that these efforts continue.

Based on the discussions at this workshop, we plan to continue our studies in order to make concrete proposals.

目次

1 ｜ 開催趣旨、挨拶

1.1 木村上席フェロー

なぜ今コンピューティングアーキテクチャーなのかと思われる方も多いかと思うので、ここで開催趣旨を説明する。

まず、背景を図1-1-1に示す。コンピューティングを取り巻く状況について、いくつかポイントを挙げていく。皆さんご存じのように、ムーアの法則が陰りを見せており、生成AIなどで非常に高い性能が要求され、また電力の消費も増えている。また、メタバースを含む新しいワークロードが出現し、リアルタイム性が強く求められるようになってきている。このような背景から、アーキテクチャー自体がクラウドやエッジといった形で変革を迫られているように感じる。もちろん、省電力化は地球規模で重要な課題となっており、技術的な面でも環境が大きく変化していると感じられる。

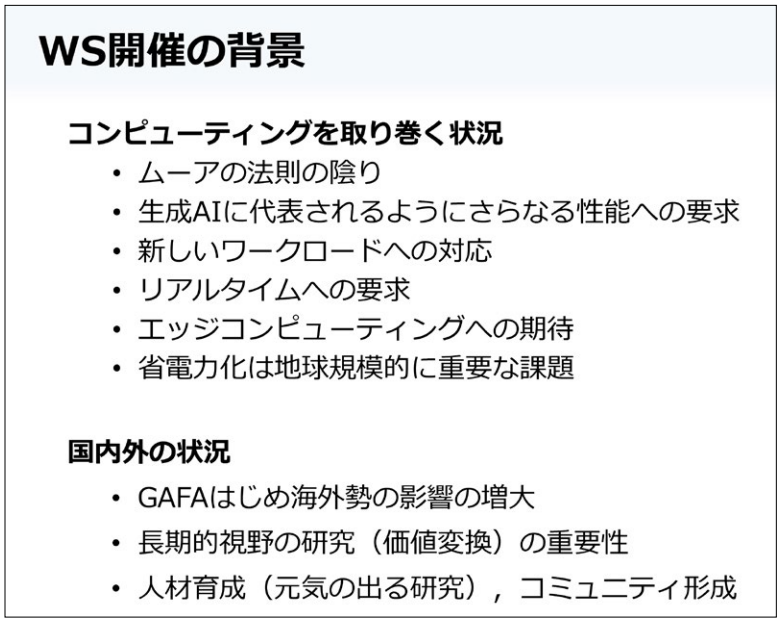


図1-1-1　　ワークショップ開催の背景

別の観点から見ると、多くの学会において日本の発表が少なく、GAFAをはじめとする外国勢に圧倒されている。研究を行うにも、資金を含めた多くのリソースが必要であり、ビッグサイエンス化している現状がある。

一方で、コンピューティングを取り巻く状況でも触れたように、見方を変えた長期的な視点で、地道ではあるが確実な研究を進める必要があると強く感じている。 GAFAに対抗するわけではないが、日本としてそういった取り組みが必要だと個人的に強く思っている。

このような経緯で、ワークショップを開催しようということになった。先生方のご意見を伺い、次のステップとして何ができるかを考えようというものである。人材育成を含め、元気が出る研究テーマにして、コミュ

ニティーを作りたいと考えている。学会に作るのか、他の場所に作るのかはさまざまな選択肢があるが、何かしらの形でこの取り組みを進めるべきだと考えている。そのため、皆様にお願いし、今回のワークショップを開催する運びとなった。

現状を見てみる。ムーアの法則がそろそろ限界に達していることや、スレッドあたりの性能が頭打ちになってきている。次に、AIの訓練コストが非常に大きくなり、計算リソースの要求が非常に高まっているという点が大きな変化である。特に、LLM（大規模言語モデル）が大規模化している。パラメーターが非常に大きくなってきており、性能が向上しているという側面もあるが、そのためのリソース要求も増大している。次に、消費電力に関しては、CPUの電力にはそもそも限界があるため、大きな変化は感じられないが、議論の余地はある。一方で、データセンターにおけるトラフィックやワークロードが大幅に増加しており、その変化は非常に大きいと感じている。

また、アプリケーションにおいて、たとえばドローンのようなものを全体としてどう制御するかといったリアルタイム性の課題があり、システムとしてエッジコンピューティングが出現していることにも注目している。さらに、メタバースのように、非常に大きな計算リソースとリアルタイム性を同時に要求するような新しい技術にも意識を向けている。

次に進むと、話が少し飛躍するが、こうした取り組みを行う際に、異なる観点からアーキテクチャーを考えてみたい。新たな計算原理、たとえばリザーバやスパイキングニューラルネットなどを研究している人だけでなく、利用する人やインプリメンテーションをする人たちとも協力し、アプリケーションも含めて一緒に考えてみたい。プロセッサーに関しても、ニューロモーフィックやインメモリーコンピューティング、布線論理などがある。こうした技術についても、様々な視点から多くの人が集まって議論する場を提供することが必要ではないかと考える。エッジコンピューティングや負荷分散といった課題も重要だと感じている。

このようなテーマを再度、長期的な視点で多くの人が集まって考える研究テーマとして取り上げてうまく進める仕組みを考えたい。平成29年度に「革新的コンピューティング」というプロジェクトを開始した。坂井修一先生のCRESTプロジェクトと、今日ご出席いただいている井上先生の「さきがけ」プロジェクトである。今日は坂井先生のご都合が悪く、代わりに入江先生にご出席いただいている。また、別の出口として、内閣府のSIPの2期におけるフィジカル空間データ処理もあった。さらに、NEDOのプロジェクトであるAIチッププロジェクトにも、当時のデバイス室長にインプットを行い、様々な議論をした。間接的ではあるが、貢献できたのではないかと自負している。

今回は少し長期的で面白いテーマに取り組み、将来を見据えた研究テーマとその実施方法や体制について考えたいと思い、皆様のお知恵をお借りするために本ワークショップを開催した。これが図1-1-2に示す今回のワークショップの狙いである。

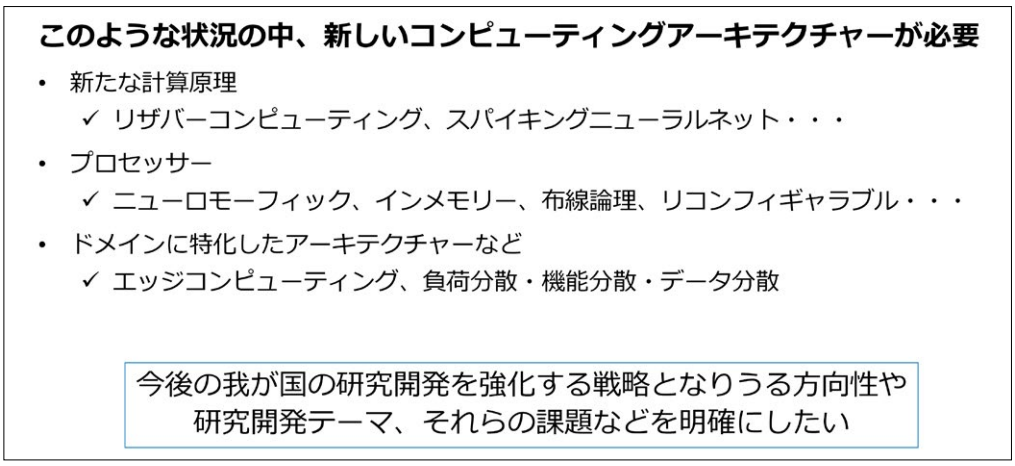


図1-1-2　　本ワークショップの狙い

1.2 徳田特任フェロー

私がお声がけいただいたのは、CRDSの俯瞰報告書のコンピューティングアーキテクチャーの区分総括を担当し、いろいろな方々に書いていただいた内容を一貫性が通るように形作りをしたからである。私自身は、現在、CRDSの特任フェローとJSTの次世代IoTの「さきがけ」研究総括をしているが、もともとはシステムソフトウェアの研究者で、リアルタイムOSに携わっていた。日本に戻ってきてからは、ユビキタスコンピューティングシステムの研究に取り組んできた。

現在私が所属している情報通信研究機構（NICT）では、Beyond 5G（6G）に向けた情報インフラの新しい研究がスタートし、すでに4年が経過している。生成AIの波がさまざまなレイヤーに影響を与えているため、今回の革新的コンピューティングアーキテクチャーを再度皆さんの知恵と工夫で議論することは非常に意義深いと思っている。

一方、NICTでは京阪奈に生成AIのチームがあり、NVIDIAのGPU、H100がまだ来ないためA100を800枚ほど使用しているが、それでも何週間も稼働させなければならない。最初は140億パラメーター、次に400億、1,790億、日本語に特化したモデルで3,110億パラメーターまで動いている。しかし、海外の企業では2万枚から3万枚のGPUを使用しており、すでに数兆パラメーターに達していると言われている。消費電力の問題は非常に深刻であり、個人的には情報通信のインフラをオール光にシフトすることが重要だと考えている。光電融合技術も進展しており、私は研究所内で「Computing and Communication for Carbon Neutral」というキーワードで研究者に呼びかけている。コンピューティングだけではなく、コミュニケーションとコンピューティングの両方のカーボンフットプリントを削減することが重要である。

また、研究所内には量子コミュニケーションのチームもあり、コンピューティングと併せて取り組んでいる。古典コンピューティングのパラダイムから完全に量子コンピューティングのパラダイムに移行するのは容易ではない。ハイブリッドで適材適所に利用することが重要であり、今日ご提案のあるさまざまなコンピューティングも、適材適所でハイブリッドに利用できる枠組みを育てることで、スムーズな移行が可能になるのではないかと思う。

今日はこのワークショップに参加させていただくことを非常に光栄に思っているので、どうぞよろしくお願いする。

2 ｜ ポジショントーク

2.1 浅井 哲也（北海道大学大学院情報科学研究院 情報エレクトロニクス部門 集積システム分野）

錯綜するコンピューティングとアーキテクチャーという言葉

最近、コンピューティングとアーキテクチャーという言葉が錯綜している気がするので、再定義したい。コンピューティングとは計算方式のようなもの、つまりアルゴリズムのようなもので、広義にはそのハードウェアも含む。一方で、アーキテクチャーは物理的なコンピューターの構成要素を設計するものである。そのような点で、最近の〇〇コンピューティングという言葉は、区別が曖昧である。たとえば、リザーバーコンピューティングは純粋なコンピューティングであるのに対して、インメモリーコンピューティングは、私の視点から言うと、積和演算をするためのアーキテクチャーであり、エッジコンピューティングもアーキテクチャーである。最近では、コンピューティングの専門家がアーキテクチャーに関する研究をするケースや、逆に、アーキテクチャーの専門家がコンピューティングまで背伸びをして研究するようなケースが結構ある。しかし、このままだと研究開発をする上で効率がよくないのではないか。やはり専門をきちんと分けたほうが良いと私は思う。

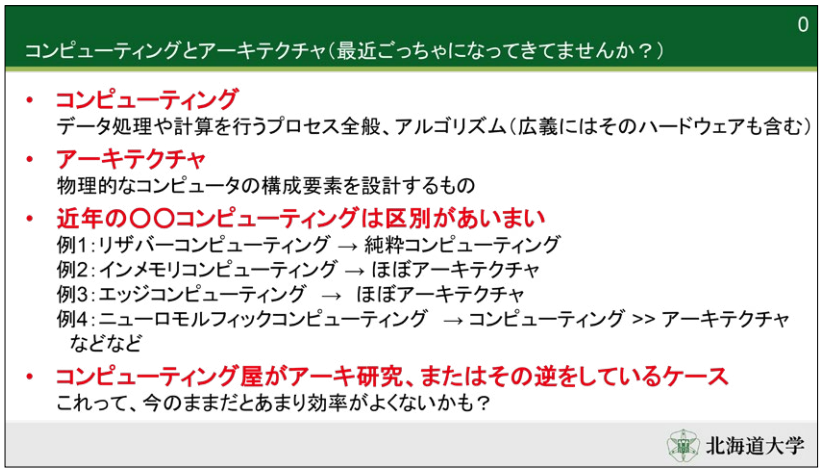


図2-1-1　　錯綜するコンピューティングとアーキテクチャーという言葉

コンピューティングに対する個人的意見および熱き思い

そのような意味で、もともとコンピューティングの研究者である私の個人的な見解としては、純粋なコンピューティングに注力したいし、その分野を育てたいという思いが非常に強い。コンピューティングに関する研究の問題は「何を計算すれば新しい価値が出るのか？」である。AI、ノイマン型、あるいは量子コンピューティングの分野は「何を計算すれば新しい価値が出るか」が分かっているので、お金もマンパワーも集中できる。しかし、「何を計算すれば新しい価値が出るか」が分からない未知の世界では「そもそも何を計算すればいいか」が分からない。新規コンピューティングを中心に研究されている研究者は、やはり欧米に偏っているが、全体としては少数派である。数学者フォン・ノイマン、物理学者ファインマン、そして、最近では、哲学者とか思想家と言われてもおかしくないような人たちが、純粋コンピューティングについて真剣に考えている。

一方で、〇〇コンピューティングを研究しているのは、やはりアーキテクチャーの研究者が非常に多く、多数派である。多くの研究者がいる理由は、もちろん計算要素・デバイスから新規なコンピューティングが生まれてくることもあるからだ。現行のノイマン型とかAIコンピューティングとかではない新しい価値を生み出すコンピューティング（計算方式）が固まれば、アーキテクチャーが後からすごい勢いでついてくるので、純粋なコンピューティングに対して、もう少し力を配分できないかと考えている。

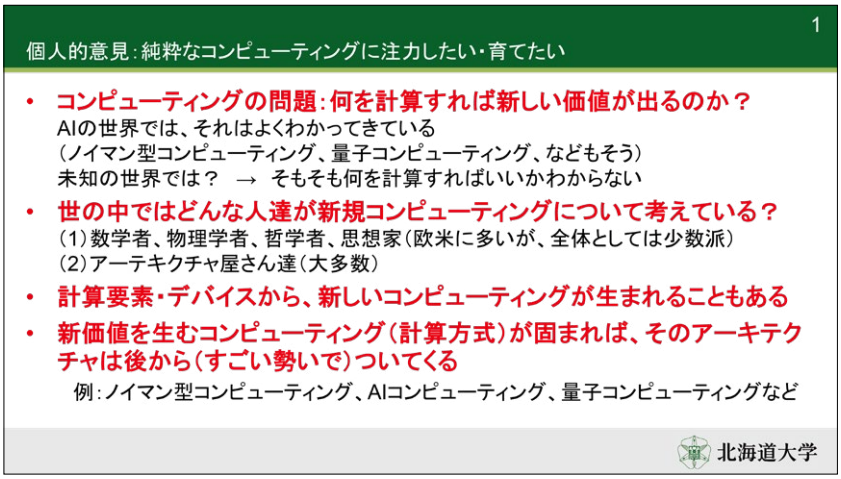


図2-1-2　　コンピューティングに対する個人的意見

次世代コンピューティングとして生き残るもの

では、次世代コンピューティングとして、どのようなコンピューティングが生き残っていくのか。やはり、汎用性は非常に重要で、歴史を振り返ってみても、今のノイマン型しかりで、汎用コンピューティングはずっと生き残ってくると思う。今の生成AIも、プログラムではなく学習によって汎用計算の仕方を獲得できるので汎用的であり、また、脳自体も汎用であると言える。一方、専用コンピューティングというのは、極端に現行のコンピューティングシステムに対して高い性能が出ないとなかなか生き残れない。量子アニーリングは圧倒的に高い性能を持つので生き残っていくと思うが、他にどんなものがあるのだろうかということを考えれば、やはり汎用性の高いものを目指すのが、純粋コンピューティングの研究にとって大事である。

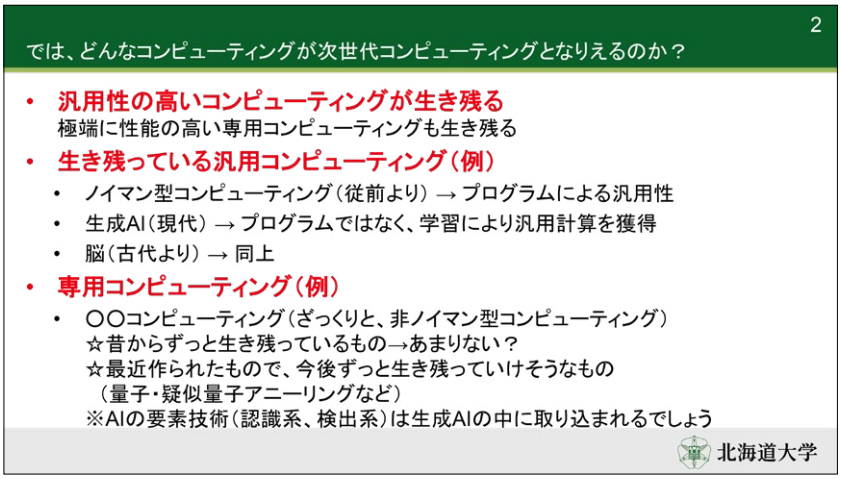


図2-1-3　　次世代コンピューティングとして生き残るもの

AIコンピューティングで非常に苦労していること

そのような枠組みで、私が今このAIコンピューティングで非常に苦労していることがいろいろある。一番苦労しているのは、せっかくニューラルネットワークで脳を真似て我々が楽できるような機械を作ろうとしているが、そこに有益な機能を持たせるためには、さらに我々の頭脳を相当使わないといけない。どのような構造にすればいいのか、どのような規模にすればいいのか、どのようなデータをどれくらいどのように学習させればいいのか、凄く頭を使って、そのアプリを探すことになるので非常に苦労している。たとえば、高性能の未学習AIを渡されて、これで何かやってくださいと言われたときに、ユーザー（研究者）は困ってしまう。これを使えば何かいろいろなことができるのは分かるが、それをどのように使えば新しい価値を生み出すことができるか、いろいろな人に使ってもらうことで半導体業界全体を含めてエコシステムをうまく回すようなモデルができるか、そのようなことを考えるのは非常に難しい。

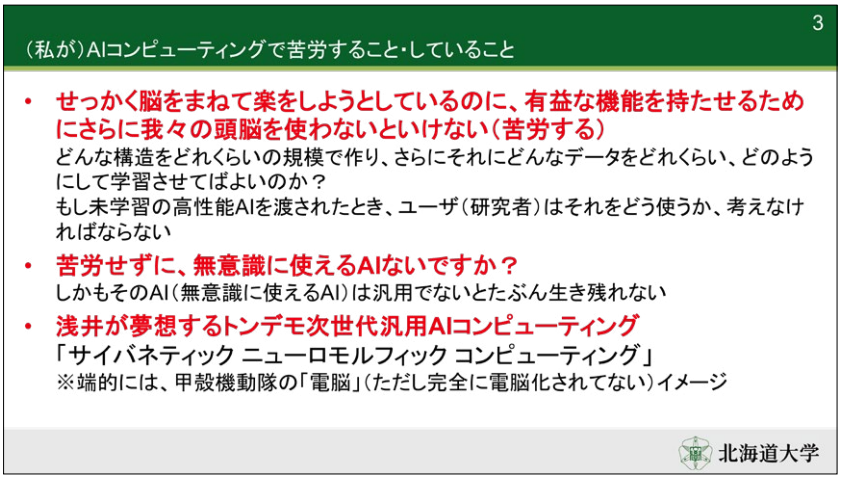


図2-1-4　　AIコンピューティングで非常に苦労していること

サイバネティック・ニューロモルフィック・コンピューティングのコンセプト

上記のような考え方で、今は多くのコンピューティングとそのアーキテクチャーの研究が進んでいると思うが、私はもう少し楽をしたいので、苦労しないで無意識に使えるAIについて何かないのかということをずっと考えていた。しかも、無意識に使えるAIは、汎用的でないと多分生き残れないので、何年か前からNEDOのFS研究でも少しやらせて頂いたが、「トンデモ次世代汎用AIコンピューティング」のようなものとして、「サイバネティック・ニューロモルフィック・コンピューティング」に関する研究を少しずつ進めている状況である。

脳の中には、学習を全くせずにひたすら外部から入ってきた入力を、次元を物凄く高く上げた状態へと複雑に変換するような可塑性を持たない領域が（NEDOのFS研究で）数多く見つかっている。そこでは、数多くの非線形素子が集まったような複雑ネットワークの機能として、複雑なダイナミクスが生み出される。実のところ、入ってきた情報を超高次元に上げる機能は、リザーバーコンピューティングにおけるリザーバーと全く同じである。

さらにもう少し調べてみると、そのような脳の組織の周りには可塑性を持つ組織が存在する。つまり、その複雑ダイナミクスを生成する部位の周りにそれを使って学習するような組織が存在している。そうすると、脳の中の一部はやはりリザーバーとして機能しているのではないかという「脳のリザーバー仮説」が立ち上がってくる。脳の中のリザーバーに相当する機能の一部は半導体回路として作ることが可能なので、脳のリザーバー相当の機能の一部を、たとえば半導体リザーバーで置き換えてしまうことができる。今日の半導体を使えば、脳組織よりも遥かに高密度で非線形性もかなり設計ができるので、おそらく脳のリザーバー相当の部位

よりも、リザーバーとしてはかなり性能の高いものを脳の中に埋め込むことができると考えている。それを埋め込むと、無意識のうちに、脳組織がそれを使うように学習をしていくはずである。

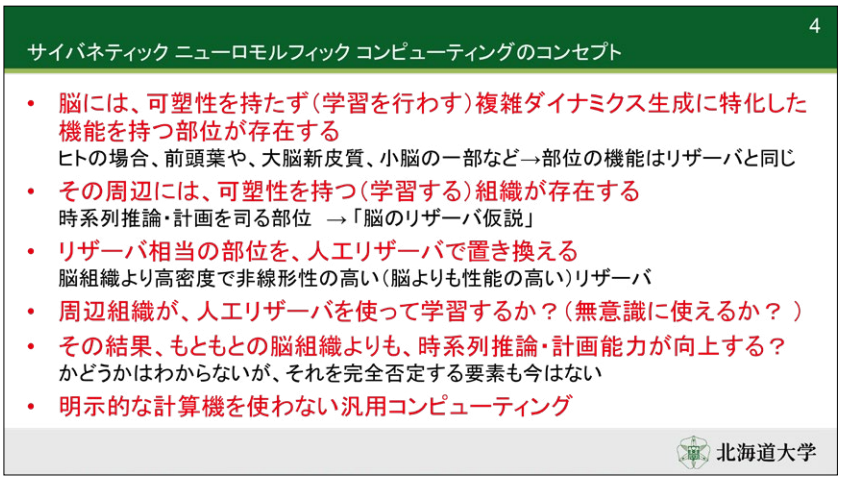


図2-1-5　サイバネティック・ニューロモルフィック・コンピューティングのコンセプト

元々脳の中にあるリザーバー相当の機能を持っている部位は、主に時系列推論や運動計画をつかさどっているので、そこの部分のリザーバー相当の脳の組織を半導体で作った（もう少し良い）リザーバーで置き換えれば、この時系列推論・計画能力が上がるのではないか。もちろん上がるかどうかは分からないが、それを完全否定する要素も今のところない。そうすれば、人の中に自然に入り込んで我々が意識することなく、気づけばいろいろな能力が上がっていたということになる。人の脳では基本的に汎用計算が行われているが、その汎用計算を加速するようなことにつながり、やがては（明示的な計算機を使わない）汎用コンピューティングにつながるのではないかと考えている。

4つのキーメッセージ

最後に、私から（ポジショントークとしての）4つのキーメッセージを送って結語としたい。まずは、（もちろん、アーキテクチャーも大事だと思うが…）純粋なコンピューティングの開拓は非常に大事である。そして、純粋コンピューティングの研究者を育てる土壌をつくらないといけない。日本では、この純粋コンピューティングの研究者がほとんどいないのが現状である。昔は10人くらいはいたが、今では多分3人くらいしかおらず、研究資金が乏しく変人扱いされるような状況にある。それは、今日の競争（選択と集中）の結果だと思うが、この状況は少しまずいと思う。できれば、純粋コンピューティングの知識を持つアーキテクチャーの研究者を育てていけるような土壌をつくらないと、状況はあまり変わらないような気がする。さらに、3つ目のキーメッセージとしては、汎用性の高いコンピューティングが生き残るということ。その結果、価値のある新規なコンピューティングが見つかれば、アーキテクチャーは後から物凄い勢いで付いてくるというのが最後のキーメッセージである。

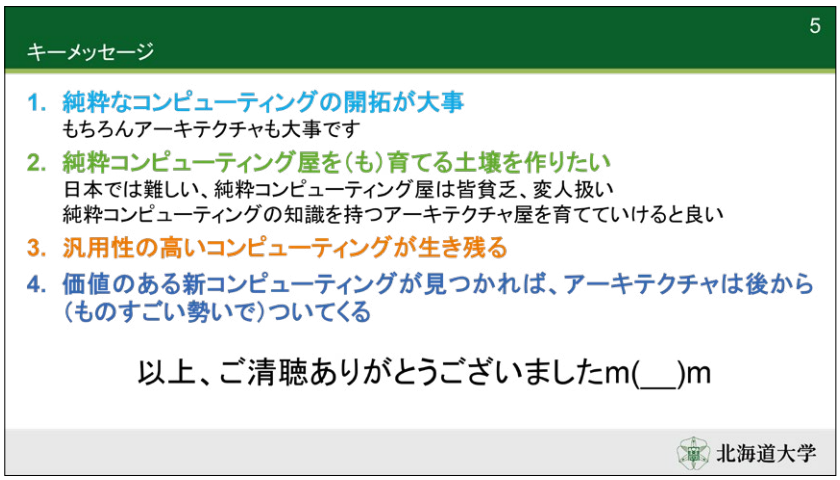


図2-1-6　　4つのキーメッセージ

【質疑応答】

高島：脳のリザーバー部分に関する質問だが、学習しないということは、先天的なもの、生まれついてのものと考えてよろしいか？

浅井：はい、もともとそのような構造になっている。

高島：そうすると、構造的にはどの人間でも大体同じという意味で、人工リザーバー部分を半導体で作ることに意味があると捉えてよろしいか？

浅井：はい。

徳田：実際に、脳の中にある部分にリザーバーコンピューティングのモジュールを組み込むと、時系列推論等の計画能力が向上するという話だったと思う。少しエンジニアリング的な質問で恐縮だが、私たちのウエットな脳の素子と人工的に作られた回路との接合部分に関して何か議論があるか？論理的な絵を描けば、おそらく計算論的にはつながるとは思うが、ウエットな脳の中に埋め込む点について、どのように考えているのか？

浅井：そこが、まさに我々が取り組んだNEDOのフィージビリティスタディであり、インターフェイスに関する研究である。まず、リザーバーが脳側の発したスパイク信号を受け取る必要がある。しかしながら、脳に電極を直接刺して受け取るのは良くないので、非侵襲なキャパシティブなカップリングを使って、神経１個のスパイク信号を拾うのではなく、ある程度のまとまった神経組織の平均活動を拾うような非侵襲なインターフェイスを作って、脳のスパイク信号を受け取らなくてはならない。受け口では、キャパシティーカップリングで得られるわずかな電圧を増幅して、リザーバーコンピューター側に回す。一方、出口の方では、安全に神経を刺激するための方法として、基本的には神経を電流刺激するために電流を流し込むことになる。しかしながら、電流を流し続けると、その部分がどんどんチャージアップして、その電位によって神経組織が発火してスパイクを出してしまう。現在は、チャージアップしないように、常にその周辺の電位をバックグラウンドレベルと監視をすることで、ある程度以上のチャージアップがあれば、それを引き抜くような機構を組み入れた集積回路を我々は開発している。

今のところ、電極はまだ平面電極であり、張り付けて電流を流すような形になっている。基本的には、非侵襲で脳の神経活動も読み取り、その後リザーバーコンピューターで処理した後の出力を非侵襲に脳側に電流刺激をするインターフェイスである。もともとこのような研究は昔からあったが、ほとんどが医療計測器であり、非常にサイズが大きく、医療機器なので精度が要求される。ところが、脳の中に埋め込んで使わせるには高い精度は要らない。しかしながら、１回手術して埋め込めば

100年間は電池で動いてほしい。我々は、そのような観点から、精度を犠牲にした代わりに超低消費電力で動作するようなインターフェイスを北海道大学で作っているところである。

2.2 井上 弘士（九州大学大学院システム情報科学府 情報知能工学専攻）

「さきがけ」の革新的コンピューティング技術（以下、革コン）で研究総括を担当した。まず革コンの活動を紹介し、その後普段思っているところを共有したい。

革コンの活動

これまでは半導体の微細化に基づいて量的なイノベーションを作り出してきたという歴史がある。一方で、その量的イノベーションが今後は難しくなるので、質的なアプローチに基づくイノベーションが必要になってくる。量から質へのパラダイムシフトが必要であり、そこをきちんと探究することが必要だろうと、革コンの領域が立ち上がったときに思っていた。

ムーアの時代は半導体の微細化が続く時代であり、MOSFETのデバイスや再構成可能デバイスとしてFPGAなどが使えるようになった時代である。そこでは、量を性能に転換してコンピューターシステムを作ることができて発展した。ムーアの時代が終焉を迎えると、量ではなく質を変えなければいけない。そうすると、図2-2-1に「デバイス多様性に基づく質的変化」と書いたが、ポストムーアの時代には新しいデバイスを創成し、それをいかにコンピューティングとして仕上げていくか、というアプローチが必要なのではないかと思っている。

新しいデバイスや計算原理・モデルを探求すると、新しいデバイスはピーキーな部分があり、それが非常に優れた部分につながっていることが往々にしてある。一方、CMOSはオールマイティーで、何でもある程度うまくいくデバイスである。今後は新しいデバイスを使ううえで、ピーキーなところを犠牲にして何かを解決しなければならないこともあるが、できるだけピーキーなところを生かしつつ、いかに欠点を隠すあるいは許容するかといったコンピューティングフレームワークを考えることが一つのポイントになってくるだろう。こういうことを考えながら、さきがけ「革新的コンピューティング技術の開拓」領域を立ち上げた。

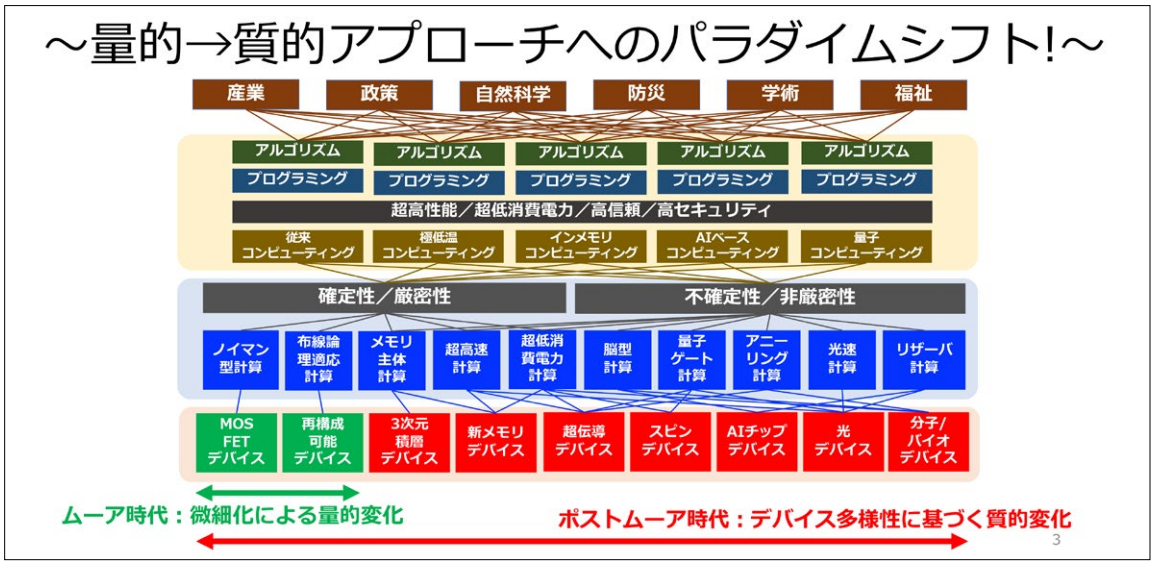


図2-2-1　　現状をどう打開するか？

図2-2-1の上部にあるアプリケーションはコンピューティングシステムにとって重要であり、アプリケーションの拡大ができないようなコンピューティング技術を作っては駄目だと思っている。新しいアプリケーションが次々と出てくるからこそ、社会は良くなっていく。したがって、アプリケーションを拡大できるようにしながら、

デバイス側の多様性を持たせる世界になってくるだろう。

そうすると、アプリケーションとデバイスの多様性の間をどうやってつなぐのかが非常に重要になってきて、ここがアーキテクチャーのやるべき仕事だと思っている。

私のアーキテクチャーの定義は、基本的にはインターフェイス、抽象化であり、上に異なるレイヤーがあったときに、インターフェイスをどう切って抽象化し、上に見せるべきところはどこか、下に見せるべきところはどこか、見せなくてよいところはどこかを考え、異なるレイヤーがインタラクションできるようにして、レイヤー間で協調最適化をできるような仕組みをいかに作るかということが、アーキテクチャーの最も重要な本質だと思っている。その最たるものが命令セットアーキテクチャーである。

アプリケーションが広がり、デバイスも広がったときに、どうインターフェイスを切ってシステム全体のレイヤーをうまく連携させればいいかを考えていくと、当然、空間がより広くなる。

こういった状況で革コンという領域ができたが、基本的には若手に頑張ってもらおうと考え、特に異分野を連携させることが重要だと思い、

①さまざまな分野の新進気鋭な若手研究者集団を構成する

②互いが切磋琢磨し、異分野連携を促進する

③世界に先駆けて新しいコンピューティング技術を開拓する

という三つのポイントに力点を置いた。

図2-2-2は、縦軸が理論、アルゴリズム、プログラミング、ネットワーク、システムソフトウェア、アーキテクチャー、回路、デバイスというシステムレイヤーで、横軸がいわゆる情報担体であるエレクトロニクス、光、超電導、スピン、化学、バイオなどだが、右端の理論は情報担体ではなく別の列だと思ってもらいたい。このようなマトリクスに、「さきがけ」革コンの研究者がカバーする範囲をまとめた。おおむね全体をカバーできる研究者の方々が集まってくれたことが分かると思う。この革コンの領域が良かった点は、化学、バイオ、分子をやっている人、さらに半導体設計、回路、スピン等をやっている人、OS、プログラミング、ネットワーク、物理をやっている人、理論をやっている人、アルゴリズムをやっている人、数学をやっている人といった、多方面の方々が集まって、自然発生的にいろいろな議論をして連携し合う場ができたということだと思う。

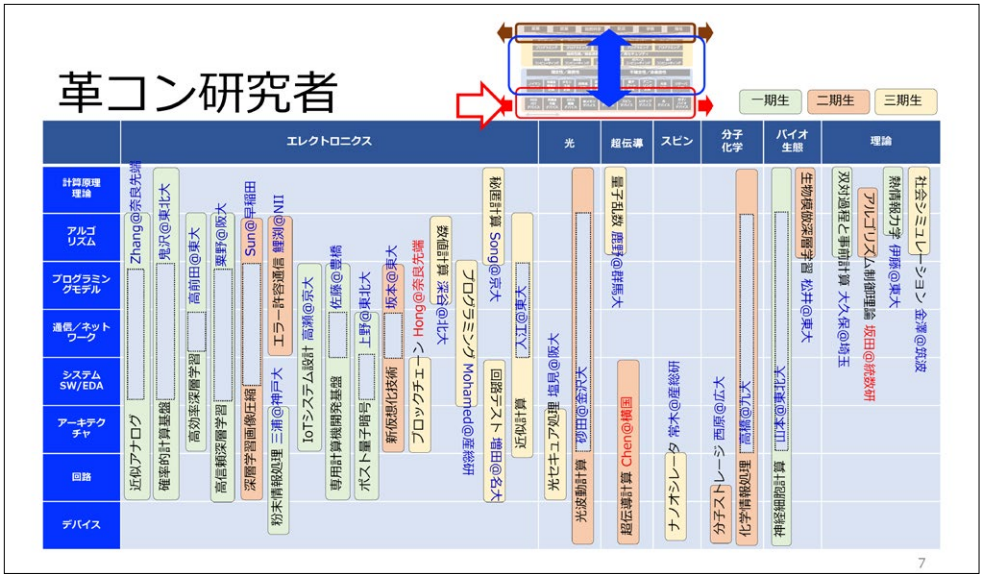


図2-2-2　革コン研究者

図2-2-3は革コンが今年の3月に終わったときの最終報告の資料からの抜粋である。この中で面白い成果

で非常に特徴的だと思ったのは、多分野に渡るトップレベル国際会議や論文誌、雑誌を通した発信である。これは一つの分野ではなく、IEEEのトランザクション系のデジタル、アナログ、セキュリティー、理論、超電導、メディア、分散といった分野のトップカンファレンスやジャーナル、それにシステム系の分野、さらに物理、数学、フィジカルレビュー、バイオ、化学、ネイチャーといったビッグサイエンス系などでの発信である。こういった場でトップの論文を出す人たちが集まることができたのではないかと思っている。

こういった場の刺激もあってだと思うのだが、革コンが始まる前の20年弱、日本から1本でも通ったら大騒ぎするような、コンピューターアーキテクチャーの分野で有名な国際学会であるISCAやMICROでの発表が、革コンが始まってからの5年ぐらいでかなり日本からの発表数が増えている。今日のポジショントークをする入江さんの発表は昨年ベストペーパー候補になっている。それは多分、日本から初めてだと思うが、そういったレベルの研究者が次々と出てきている。皆さんもよくご存じの、半導体関連のISSCC、スパコン関連のSC、設計技術関連のDAC、ICCAD、OS関連のUSENIXといったトップカンファレンスでも日本から若手が出始めたと感じている。

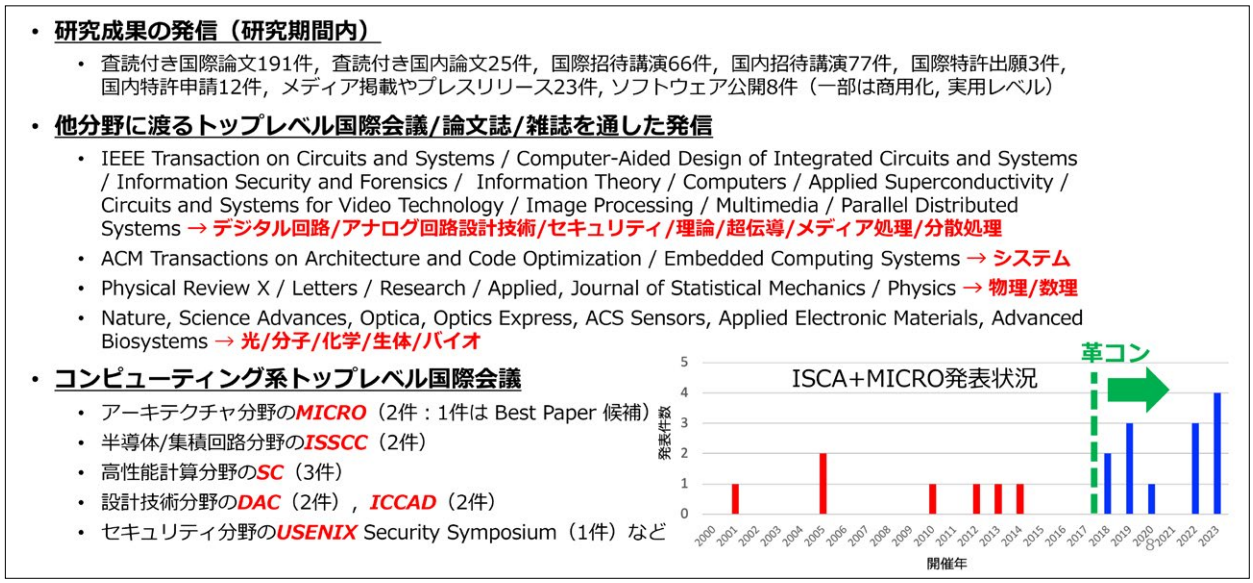
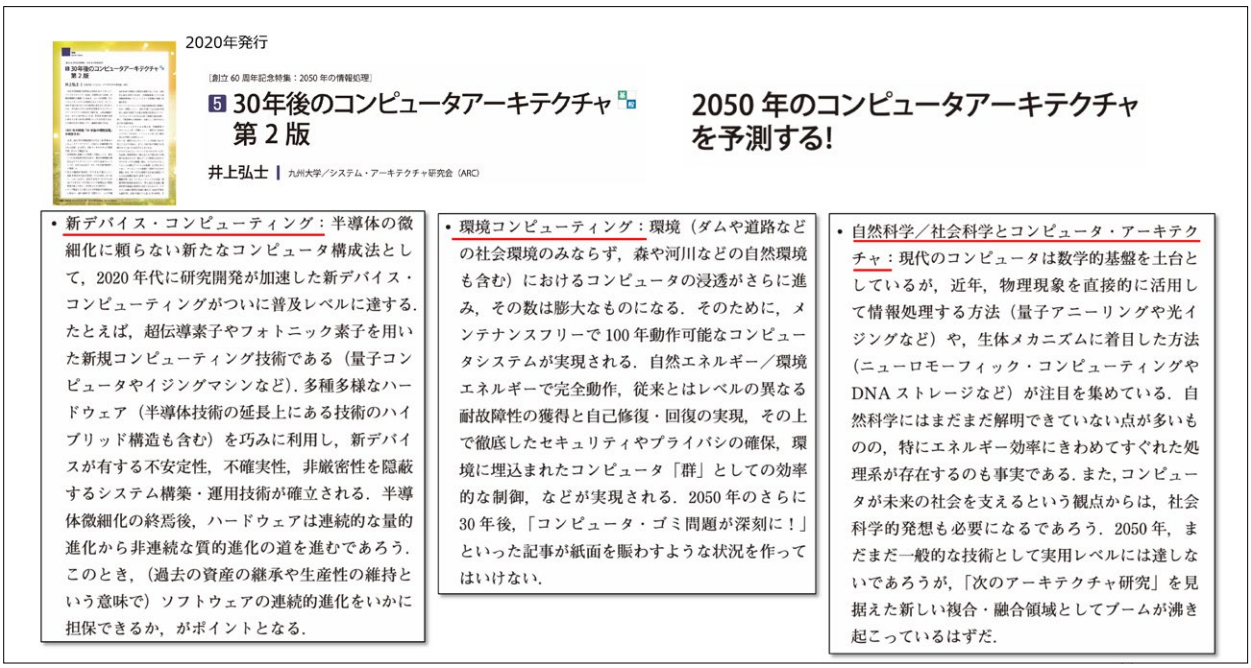


図2-2-3　　革コンの成果

今後の方向性

図2-2-4は私が2020年に情報処理学会のシステムアーキテクチャー研究会の主査を務めていたときに、「30年後の○○」という特集が組まれ、2050年のコンピューターアーキテクチャーを予測するために書いた記事である。テーマを三つ挙げており、一つ目は新デバイスコンピューティングであり、これは先ほど発表した内容である。次に環境コンピューティングを挙げているが、これは後で説明する。最後は、自然科学や社会科学とコンピューターアーキテクチャーをどううまく連携するかという話である。



2020年発行

［創立 60 周年記念特集：2050 年の情報処理］

5 30年後のコンピュータアーキテクチャ 第 2 版

井上弘士 | 九州大学／システム・アーキテクチャ研究会（ARC）

2050 年のコンピュータアーキテクチャを予測する!

• 新デバイス・コンピューティング：半導体の微細化に頼らない新たなコンピュータ構成法として，2020 年代に研究開発が加速した新デバイス・コンピューティングがついに普及レベルに達する．たとえば，超伝導素子やフォトニック素子を用いた新規コンピューティング技術である（量子コンピュータやイジングマシンなど）．多種多様なハードウェア（半導体技術の延長上にある技術のハイブリッド構造も含む）を巧みに利用し，新デバイスが有する不安定性，不確実性，非厳密性を隠蔽するシステム構築・運用技術が確立される．半導体微細化の終焉後，ハードウェアは連続的な量的進化から非連続な質的進化の道を進むであろう．このとき，（過去の資産の継承や生産性の維持という意味で）ソフトウェアの連続的進化をいかに担保できるか，がポイントとなる．

• 環境コンピューティング：環境（ダムや道路などの社会環境のみならず，森や河川などの自然環境も含む）におけるコンピュータの浸透がさらに進み，その数は膨大なものになる．そのために，メンテナンスフリーで 100 年動作可能なコンピュータシステムが実現される．自然エネルギー／環境エネルギーで完全動作，従来とはレベルの異なる耐故障性の獲得と自己修復・回復の実現，その上で徹底したセキュリティやプライバシの確保，環境に埋込まれたコンピュータ「群」としての効率的な制御，などが実現される．2050 年のさらに 30 年後，「コンピュータ・ゴミ問題が深刻に！」といった記事が紙面を賑わすような状況を作ってはいけない．

• 自然科学／社会科学とコンピュータ・アーキテクチャ：現代のコンピュータは数学的基盤を土台としているが，近年，物理現象を直接的に活用して情報処理する方法（量子アニーリングや光イジングなど）や，生体メカニズムに着目した方法（ニューロモーフィック・コンピューティングや DNA ストレージなど）が注目を集めている．自然科学にはまだまだ解明できていない点が多いものの，特にエネルギー効率にきわめてすぐれた処理系が存在するのも事実である．また，コンピュータが未来の社会を支えるという観点からは，社会科学的発想も必要になるであろう．2050 年，まだまだ一般的な技術として実用レベルには達しないであろうが，「次のアーキテクチャ研究」を見据えた新しい複合・融合領域としてブームが沸き起こっているはずだ．

図 2-2-4　　情報処理学会誌の記事

図 2-2-5 は、コンピューターアーキテクチャー分野の Top Picsと言って、年ごとに世界でトップ 12 のコンピューターアーキテクチャー研究を選んだものを 2018 年から 2022 年までの 5 年分まとめたものである。色分けがされていて、青が新デバイス、緑がメモリー、濃い茶色の文字が専用回路・アーキテクチャー、赤がセキュリティー、紫が AI、緑の白抜きが環境に関連したものである。これを見ると、2020 年にかけて新規デバイス関連が徐々に増えているし、セキュリティーやメモリーというトラディショナルな分野も依然として強い。最近は環境関連が二つ出てきているし、コンピューターの頭脳であるマイクロプロセッサーをきちんと考え直すものも出てきている。世界の動向はこういった具合である。今後のキーワードはデバイス多様性と情報社会インフラであり、デバイスに基づいた、新しいコンピューティング技術の開拓をする必要があると考えている。

ポストムーア時代の情報処理基盤技術を確立するためには、デバイス多様性に基づく方向に行かざるを得ないだろうと考えている。当然、デバイスには CMOS も含む。このときに一つ大事なのは、図 2-2-6 に示したように、コンピューターサイエンス（CS）とデバイスサイエンス（DS）が掛け算で効いてくるような新しい学問領域を作ることである。

こういった技術を作ったうえで、出口として情報社会がどう発展するかということと、コンピューティングが地球環境に与える影響がかなり深刻になるだろうという、この二つの観点からコンピューティングを探究することが大事だと考えている。

2018 (conference year)	2019	2020	2021	2022
Neural Cache: Bit-Serial In-Cache Acceleration of Deep Neural Networks	Unveiling the Hardware and Software Implications of Microservices in Cloud and Edge Systems	The Vision Behind MLPerf: Understanding AI Inference Performance	Overclocking in Immersion-Cooled Datacenters	PACMAN: Attacking ARM Pointer Authentication With Speculative Execution
Inside Project Brainwave's Cloud-Scale, Real-Time AI Processor	MAESTRO: A Data-Centric Approach to Understand Reuse, Performance, and Hardware Cost of DNN Mappings	Superconductor Computing for Neural Networks	Warehouse-Scale Video Acceleration	Hertzbleed: Turning Power Side-Channel Attacks Into Remote Timing Attacks on x86
Darwin: A Genomics Coprocessor	Energy-Efficient Video Processing for Virtual Reality	Leaking Secrets Through Compressed Caches	Practical and Scalable ML-Driven Cloud Performance Debugging With Sage	There's Always a Bigger Fish: A Clarifying Analysis of a Machine-Learning-Assisted Side-Channel Attack
A Hardware Accelerator for Tracing Garbage Collection	Towards General-Purpose Acceleration: Finding Structure in Irregularity	Understanding Acceleration Opportunities at Hyperscale	Chasing Carbon: The Elusive Environmental Footprint of Computing	Revizor: Testing Black-Box CPUs Against Speculation Contracts
Composable Building Blocks to Open Up Processor Design	Varifocal Storage: Dynamic Multiresolution Data Storage	Accelerating Genomic Data Analytics With Composable Hardware Acceleration Framework	Maya: Using Formal Control to Obfuscate Power Side Channels	Effective Mimicry of Bélády's MIN Policy
FireSim: FPGA-Accelerated Cycle-Exact Scale-Out System Simulation in the Public Cloud	AsmDB: Understanding and Mitigating Front-End Stalls in Warehouse-Scale Computers	uGEMM: Unary Computing for GEMM Applications	An Architecture to Accelerate Computation on Encrypted Data	Revisiting Residue Codes for Modern Memories
Breaking Virtual Memory Protection and the SGX Ecosystem with Foreshadow	Extending the Frontier of Quantum Computers With Qutrits	BabelFish: Fusing Address Translations for Containers	Characterizing and Mitigating Soft Errors in GPU DRAM	HeteroGen: Automatic Synthesis of Heterogeneous Cache Coherence Protocols
Context-Sensitive Decoding: On-Demand Microcode Customization for Security and Energy Management	Architecting Noisy Intermediate-Scale Quantum Computers: A Real-System Study	Characterizing and Modeling Nonvolatile Memory Systems	The Laplace Microarchitecture for Tracking Data Uncertainty	Online Code Layout Optimizations via OCOLOS
Security Verification via Automatic Hardware-Aware Exploit Synthesis: The CheckMate Approach	Speculative Taint Tracking (STT): A Comprehensive Protection for Speculatively Accessed Data	Temporal Computing With Superconductors	Systematically Understanding Graph Accelerator Dimensions and the Value of Hardware Flexibility	IOCost: Block Input–Output Control for Containers in Datacenters
Language Support for Memory Persistency	MicroScope: Enabling Microarchitectural Replay Attacks	A Next-Generation Cryogenic Processor Architecture	ILLIXR: An Open Testbed to Enable Extended Reality Systems Research	EyeCoD: Eye Tracking System Acceleration via FlatCam-Based Algorithm and Hardware Co-Design
Nonblocking DRAM Refresh	Creating Foundations for Secure Microarchitectures With Data-Oblivious ISA Extensions	Balancing Specialized Versus Flexible Computation in Brain–Computer Interfaces	Distributed Data Persistency	Toward Developing High-Performance RISC-V Processors Using Agile Methodology
Synchronized Progress in Interconnection Networks (SPIN): A New Theory for Deadlock Freedom	Trace Wringing for Program Trace Privacy	Virtual Logical Qubits: A Compact Architecture for Fault-Tolerant Quantum Computing	Vector Runahead for Indirect Memory Accesses	Architectural CO2 Footprint Tool: Designing Sustainable Computer Systems With an Architectural Carbon Modeling Tool

AI　Memory (except for Security)　Security
Emerging Device　Domain Specific　Environmentally Scalable　12

図2-2-5　　IEEE MICRO Top Pics

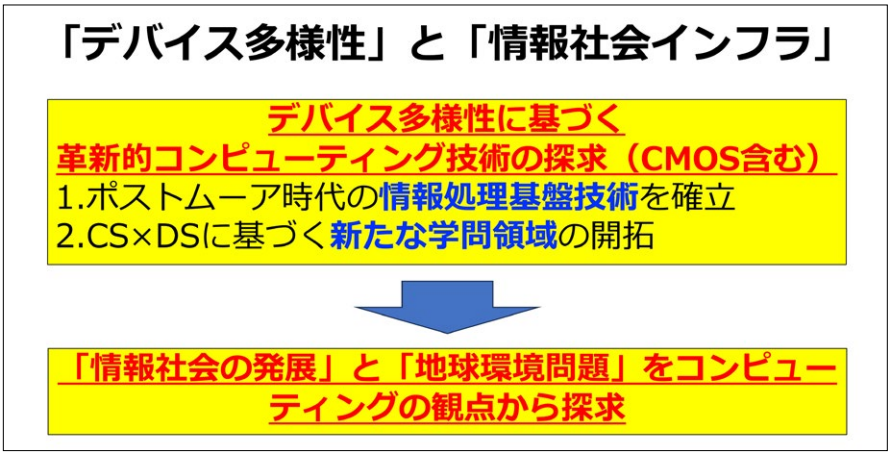


図2-2-6　　さらなる挑戦へのキーワード

デバイス多様性に基づく革新的コンピューティング技術を探求する際に大事なのは、プログラマビリティーとスケーラビリティーを絶対に忘れてはいけないということである。浅井先生も汎用性が大事と言っていたが、私も全くそのとおりだと思っている。それはプログラマビリティーに関係する。

もう一つは、コンピューターが現在社会を支える道具になっているので、一過的にすごいことができるというだけでは社会を支えられない。テクノロジーが進むとスケールし、より付加価値のあるコンピューティング技術を社会に提供し続けられるという意味でのスケーラビリティーという観点も考えておかないといけない。こういった観点を押さえた上で、方向性をどうするかを議論する必要がある。

コンピューターサイエンスとデバイスサイエンスは結構分断されているのではないだろうか。コンピューターサイエンス側は与えられたデバイスで何ができるかということしか探究できないし、デバイスサイエンス側はデバイスだけでさまざまな仕様を満足しようとして、トレードオフで利点を犠牲にしなければならないような状況がある。つまり、コンピューターサイエンスとデバイスサイエンスの間に壁がある。この壁を取り払って、私がCS×DSサイクルと呼んでいるデバイスサイエンスとコンピューターサイエンスの掛け算を回せるような仕組みや分野を作らなければならないと考えている。

アーキテクチャーの専門家側から言うと、アーキテクチャーの視点でデバイス研究の開発の途中段階からそのデバイスのポテンシャルを理解しなければならない。デバイスでしか解決できないことと、アーキテクチャーやもっと上流のソフトウェアやアルゴリズムを考えれば解決できることを切り分ける。そのうえで、デバイスでしかできないことにデバイスサイエンスにはフォーカスしてもらい、良いものを作るようにする。

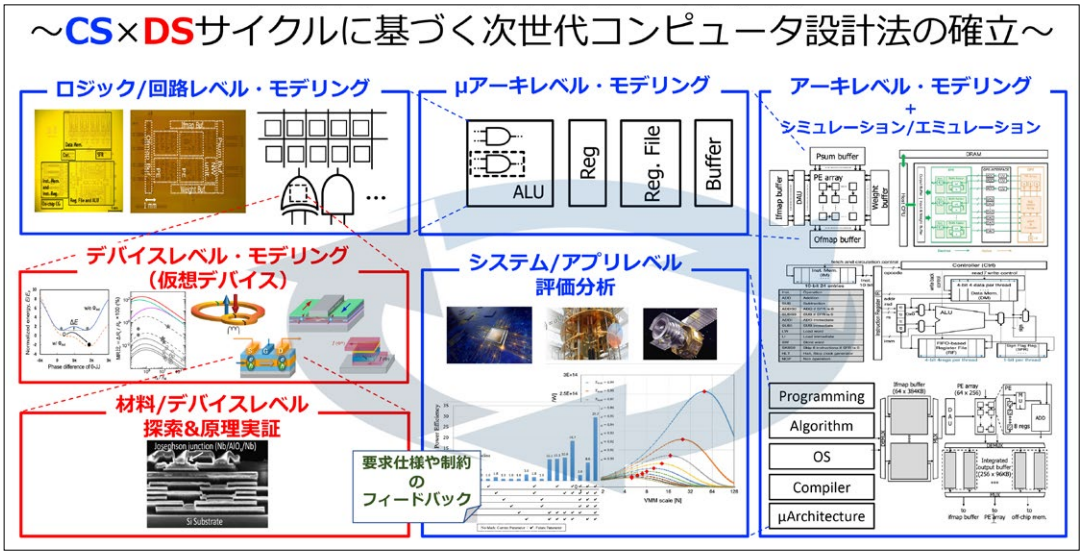


図2-2-7　　目指すべき方向性（私見）

こういった形で新しいデバイスが次々と生まれてくるような土壌を作って、デバイスをどうソフトウェアやコンピューティングにつなぐかという点をアーキテクチャーが頑張るべきである。このループの中に、システムソフトウェア、通信、アルゴリズム、アプリケーション、上位レイヤーも入ってくる必要がある。

もう一つの方向性として、環境スケーラブルなコンピューティングをどう作るかということが、今、われわれの分野で非常に重要視されつつある。カーボンフットプリント一つにしても、エンボディドカーボンと呼ばれる、半導体を作るときに出るカーボンフットプリント、それとオペレーショナルカーボンと呼ばれる、使用時のカーボンの2種類がある。半導体工場が出すカーボンも非常に大きいということが分かってきている。そうすると、半導体チップを再利用できる仕組みを持たせるなど、さまざまなことを考えなければならない。他にも、バイオなど、将来的にセンサーとしてさまざまなところで半導体が使われると、半導体ごみの問題が起きかねないので、地球環境を考えたうえでコンピューティング技術をどう発展させるかを考える必要がある。新しいデバイスを見ながら、地球環境を考えたコンピューティングをどうするかというのが今後重要ではないだろうか。

最後に、コンピューティング力はまさに国力だと考えている。コンピューティング力を伸ばし続けることは必要であり、若手を育てることも重要である。若手を育てるという観点からは、特に国際化と産学連携に関して言いたいことは山ほどあるが、今日はやめておく。

社会が求めるアプリケーションが変われば必要になるコンピューティング力、つまりどのようなコンピューティングの力が必要かということも変わってくる。新しい材料やデバイスが見つかれば、それを使ってどうやってコンピューターにまとめるのかという方法も変わってくる。先ほど挙げたカーボンのように、求められる制約が変わればコンピューターのあるべき姿も変わる。大事なのは、さまざまな制約や技術の進化、変化を世界に先駆けてきちんと察知して予測し、そのうえで最先端のコンピューティング技術を開拓し続けることである。それを実行するコミュニティーや活動が今後必要ではないだろうか。

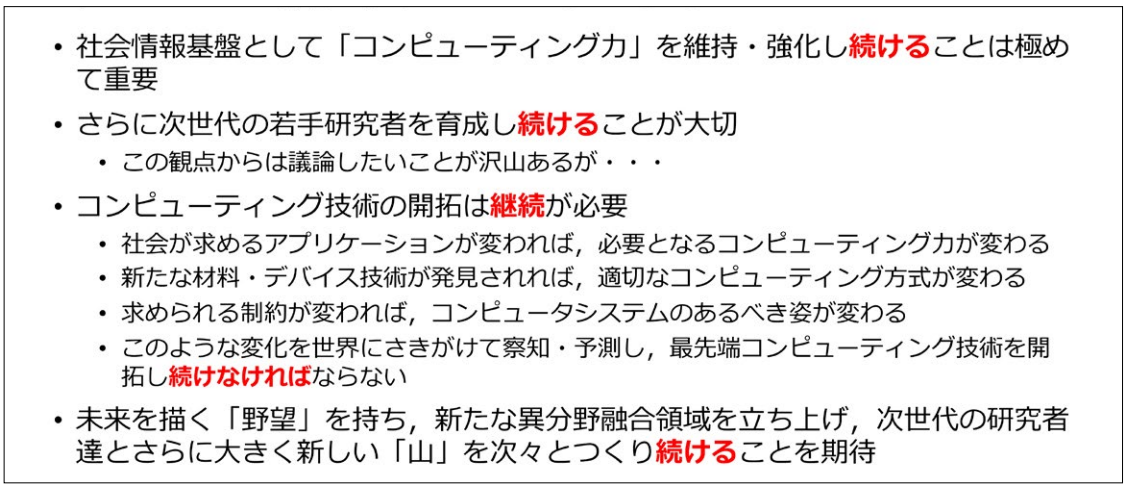


図 2-2-8　　革コンは続くよどこまでも！

【質疑応答】

高島：CSとDSの分断は重いテーマだと思うが、今回のワークショップはCS側がほとんどなので、別の機会に議論したいと思う。

馬場：コメントだが、デバイス側も上位のことがなかなか分からないので、コンピューターサイエンスの人たちと一緒にやるという方向が大事だろうと思う。学会をまとめるようなことを考えてもよいかもしれない。井上先生には、これからもいろいろと助けていただけるとありがたい。

井上：私は逆だと思っている。コンピューターサイエンス、特にわれわれはアーキテクチャーをやっているが、こちら側もデバイスのことを知らな過ぎるところがあるので、教えてもらわなければならないことが数多くあると思っている。学会を統合するような活動も必要かもしれない。

2.3 入江 英嗣（東京大学大学院情報理工学系研究科電子情報学専攻）

私はコンピューターアーキテクチャー、コンピューターシステムを専門としている。先ほども定義について少し話があったが、アーキテクチャーとは、機械と人間の界面であると捉えている。物理的な計算を行う機械と、人間が実現したいコンピューティングの間を、どのように命令セットを作成し、どのようにロジックを構成すればよいかを考える分野である。この関連で、ヒューマンコンピューターインタラクション、すなわち人間とコンピューターの間のインターフェイスのあり方についても議論しているが、特にCPUアーキテクチャーを主なポジションとしている。

さきがけの「革新的コンピューティングアーキテクチャー」に参加したり、距離指定型アーキテクチャーや形状自在計算機といったキーワードで研究を行ったりして、国際会議に出席している。また、研究の成果としてはあまり出していないが、2020年のコロナ禍以降、バーチャル空間でのメタバース的な授業を継続している。これまでの実施ノウハウも、おそらく私のポジションに含まれているかもしれない。

今日のポジショントークで話すのは、これからのコンピューターがどのように進化していくべきかという点である。そのシーズとして、たとえば光や量子、エレクトロニクスのシストリックアレイなど、様々な計算方法がある。一方で、コンピューティングを含むビジョンが存在する。このビジョンをどのようにアーキテクチャーに結びつけるかを考えていくべきだと思っている。

3つのビジョンについて述べる。また、CPUの高性能化は無視できないということで、地味ではあるし、終わったとも言われるが、シングルスレッド技術には新展開があるという話をする。

現状の問題意識

問題意識を図2-3-1に示す。

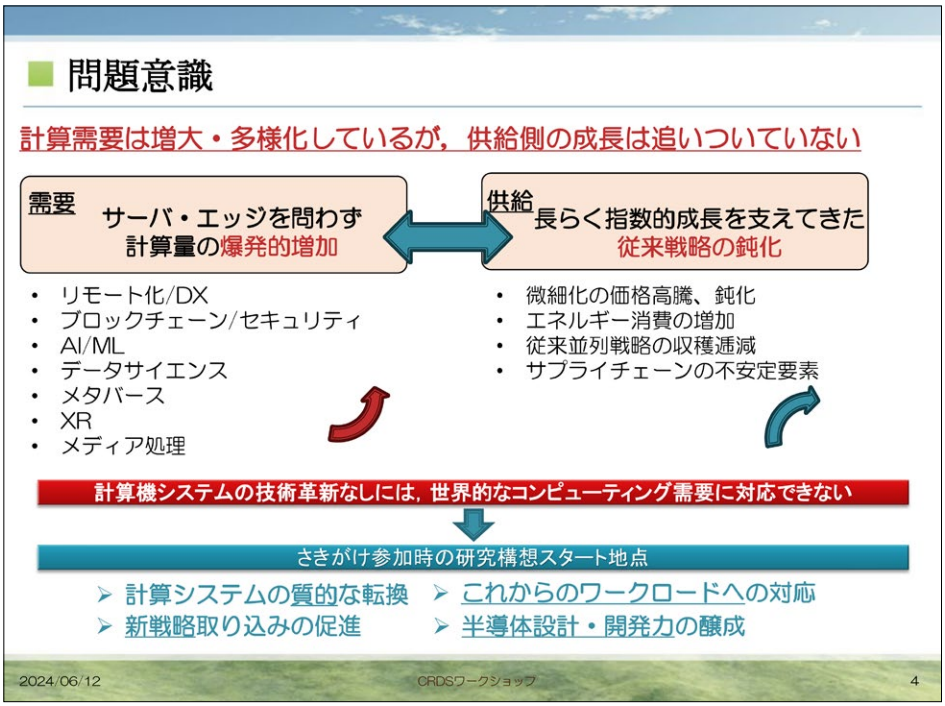


図2-3-1　　問題意識

需要のほうは爆発的に増加し、また多様性もどんどん増えている。それに対して、供給のほうは長らく質的成長を支えてきた従来の戦略が鈍化してしまっているという状況である。計算機システムの技術革新が必要だということである。やはり計算システムの質的な転換が必要だということ、そしてこれからのワークロードに対応する必要性がある。さらに、いろんな新戦略が特に回路側であれば、それはどんどん取り込んでいかなければいけない。先ほどアーキテクチャーのほうではデバイスのことをあまりよく知らないという話もあったが、どんな特性を持ったデバイスがあるのかを取り込んで研究を進めていくということもある。そして、研究によって半導体設計や開発力もきちんと維持していかなければいけない。

特に気になっているギャップを図2-3-2に示す。量的なギャップはあるが、潜在的に質的なギャップがあると感じている。それは新しいワークロードの傾向に関連している。計算機がこれからどのように使用されるかを考えると、人々に寄り添って支援するコンピューティングがますます増えていく。これらのアプリケーションには、既存の高効率化戦略では対応しきれないようなやっかいな性質が現れてくる。

たとえば、リアルタイム指向のように、即時に処理し、すぐに結果が欲しいという場合、精度よりもデッドラインに間に合うことや高いフレームレートを提供することがサービス品質に寄与する場合がある。これに対して、従来の高性能化は主にスループットを増やすことに重点を置いており、それによって高性能化してきた側面があるが、レイテンシ志向にされてしまうと対応しきれないということがある。

また、常時稼働（オールウェイズ・オン）のように、実際に使用するかどうかに関わらずプログラムを常に動作させている状態がある。モバイルデバイスではバッテリーの消耗が懸念されるが、ユーザーは明示的なオンオフ操作をせず、複数のプログラムが常時動作していることがある。従来のアプローチでは待機時間をターゲットにして省電力化を図っているが、ユーザーが利用しているかどうかわからないプログラムを常に稼働させ続けなければならないということもある。

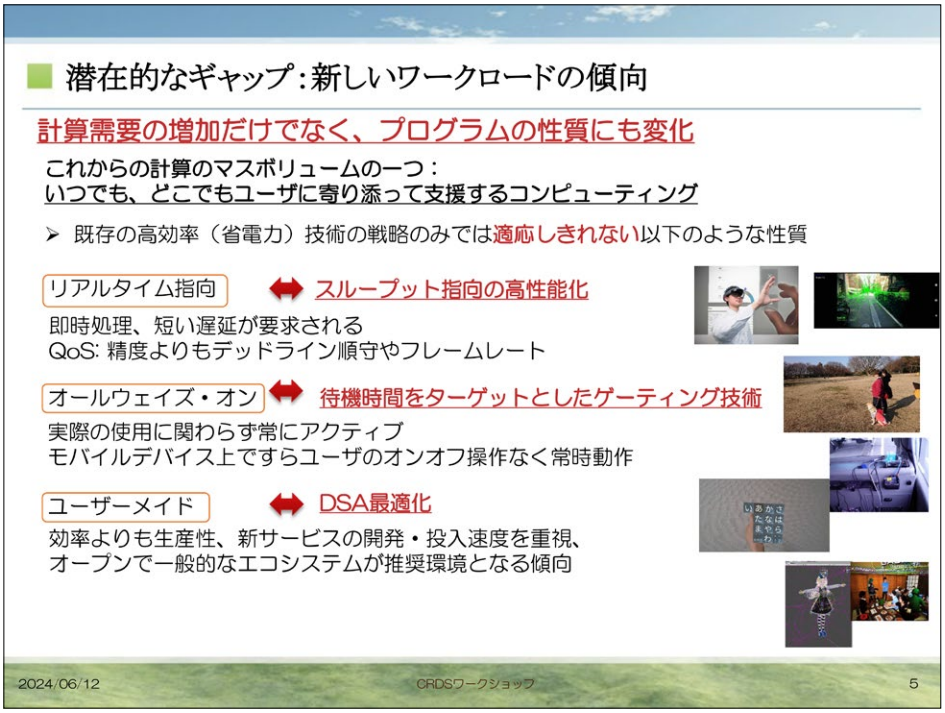


図2-3-2　　潜在的なギャップ

さらに、ユーザーメイドというものがある。効率よりも、とにかくユーザーがすぐに思いついたものをすぐに書ける生産性や、新しいサービスを始める際の開発投入速度が重視される。そうなると、最適なコード、

たとえばハンドアセンブルのような効率の良いコードを動かすよりも、オープンで一般的なエコシステムが推奨環境となる。具体的には、AIではPythonを使用したり、VRではUnityを使ったり、UnityのVisual Scriptingで適当に調整したりするようなプログラムをどんどん走らせる必要がある。これが汎用的なDSA（Domain Specific Architecture: ドメイン固有アーキテクチャー）に収まれば良いが、特殊なDSAを前提とする場合、その効率化には限界がある。このような傾向が今後ますます強まる中で、これに対応していかなければならない。

これらの問題意識に対する研究の進展を図2-3-3に示す。研究はどんどん進んでいるが、雑感としては、現在のところシルバーバレット（万能の解決策）は見当たらない。そうなると、いろいろな計算に関するシーズは出てきているが、かつてメニーコアアーキテクチャーなどのさまざまな技術が「自分の方式が一番良い」と主張してきた結果、技術が共存し、現在のプロセッサーがヘテロジーニアスマルチコア（異種混在マルチコア）になっているように、今後も資源があれば最も効率の良いところまで各技術が使われるようになるだろうと感じている。

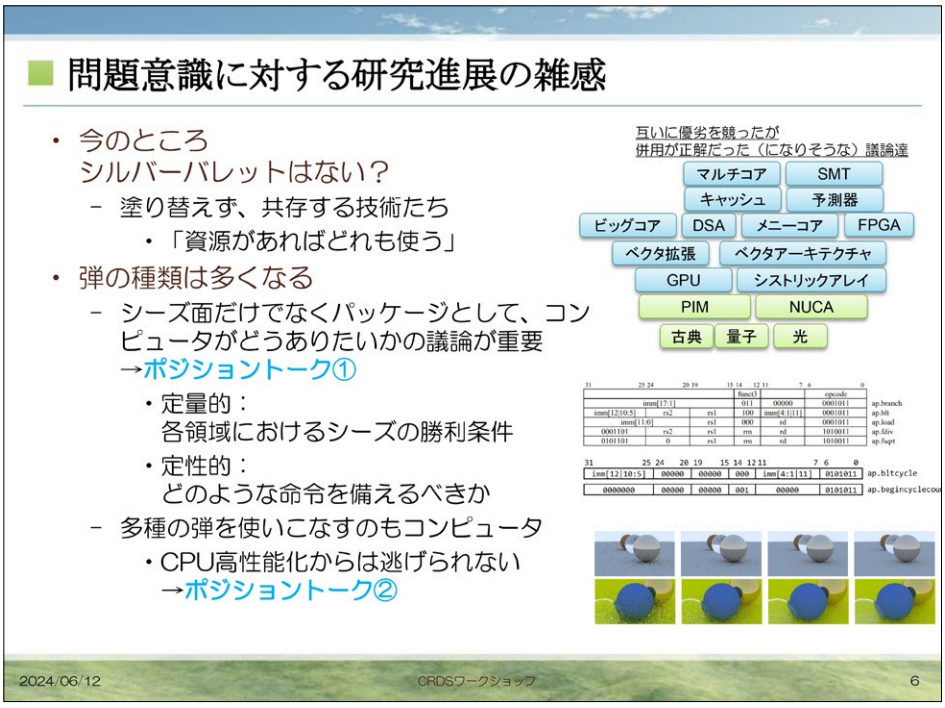


図2-3-3　　問題意識に対する研究進展

そうなると、弾の種類が多くなり、適材適所で適した弾を使っていくことになる。このような議論をする際には、恐らくシーズの側面、つまり「こういう計算方式がある」ということではなく、コンピューターのパッケージとしてコンピューターがどうあるべきかを考え、それに合わせてどの技術をどれだけ使うかという議論が重要になるだろう。

三つの改革

ここからポジショントークの1番に入る。たとえば、各シーズがどの程度、たとえば量子コンピューターがどれだけのスケーラビリティを達成すれば成功とみなせるかという勝利条件を決めるのは、コンピューターがどうあるべきかというパッケージによって決まるであろう。また、命令セットアーキテクチャーにどのような命令を備えていると望ましいかという定性的な議論も重要である。今まで存在しなかった命令クラスについても、

今後は新たな命令を追加していく必要があり、こうした議論を行うのがアーキテクチャーの仕事だと思っている。

もう一つは、多様な技術を使いこなすのもコンピューターの役割であり、そのためにはホストとなるCPUの高性能化が不可欠である。 CPUの高性能化が進まないと、逐次処理の部分でボトルネックが生じてしまうからである。これがポジショントークの２番目のポイントである。私が考えるビジョンとしては図2-3-4に示すように3つの改革がある。

図2-3-4　　3つの改革

1つ目は、成長戦略の改革である。これは需要と供給のギャップが崩れている問題に直接対応するものである。重要なのは、単に性能が向上すれば良いという話ではなく、成長戦略を見つけることである。つまり、時間とともに資源や要素が良くなり、それを性能に結び付けていくことが必要である。その意味で、現在、成長のリソースとしてスケールできるものは何かというと、CMOSである。ムーアの法則の陰りが指摘されつつも、結局CMOS素子の数は増え続けている。「終わる」と言われてきた中で、ITRS（International Technology Roadmap for Semiconductors 国際半導体技術ロードマップ）を見ると、スイッチングしないCMOSの数に関しては、まだ10年は堅調であるとされているし、10年以降にそれが終わるという保証もないと考えられる。ただし、スイッチングをすると熱密度の問題で動作しなくなる。電力が限られているパッケージの場合、成長に使える素子はあまりスイッチングしないものの、チップ内に置いておくと性能を担保してくれるというものである。このような用途でトランジスタを使い続ける限り、成長戦略はまだ有効である。 DSAやシストリックアレイ、あるいは私自身の研究である距離指定型アーキテクチャーのCPUなどは、コールドなトランジスタを置いておくだけで性能が向上するアーキテクチャーの改革になると思われる。一方で、量子や光などを実用の成長戦略とするためには、どこにスケーリングがあるのかを見通しを立てていくのがアーキテクチャーや回路の仕事だと思う。たとえば、光の場合、スケーリングしなくても性能が向上する別の成長戦略を見つけることが鍵となるだろう。

次に、形の改革についてである。システムのナノサイズ化が進む中で、1平方ミリメートルの範囲にベクター

ユニット付きの高性能プロセッサを作成することが可能であるが、そのパッケージはせいぜいクレジットカードのサイズである。これを相応に微細化し、ミリメートルやマイクロメートルのスケールに計算能力を付与することで、製造に関わる材料のコスト削減にもつながると考えている。センサーやコンピューター、アクチュエーターを微細なところに集積し、マイクロメートルスケールにして脳の髄液に流し込むようなコンピューターを考えることができるのではないかと思う。

次は働き方の改革である。これは「革新コンピューティング」で取り組んだこととも関連しているが、世界中で誰も必要としない計算が大量に行われおり、これを「ダークインストラクション」と名付けた。コンピューターが自律的に不要な計算や、セキュリティーに関わる危険な計算を省略するようなシステムを作れないかと考えている。この議論は、コンピューターに自律的な判断を任せることが危険な場合もあるため、まずは近似計算を対象にして、近似度を判断して、ユーザーからのフィードバックを取得できるようにするアーキテクチャーを考えるのが良いのではないかと思っている。

現在の技術予測やリソース、あるいは革新的な計算のシーズを踏まえ、これらをビジョンに合致させる新規アーキテクチャーの研究を進めていければ良いと考えている。

CPU研究の重要性

もう一つ、CPUの研究の重要性について話したい。

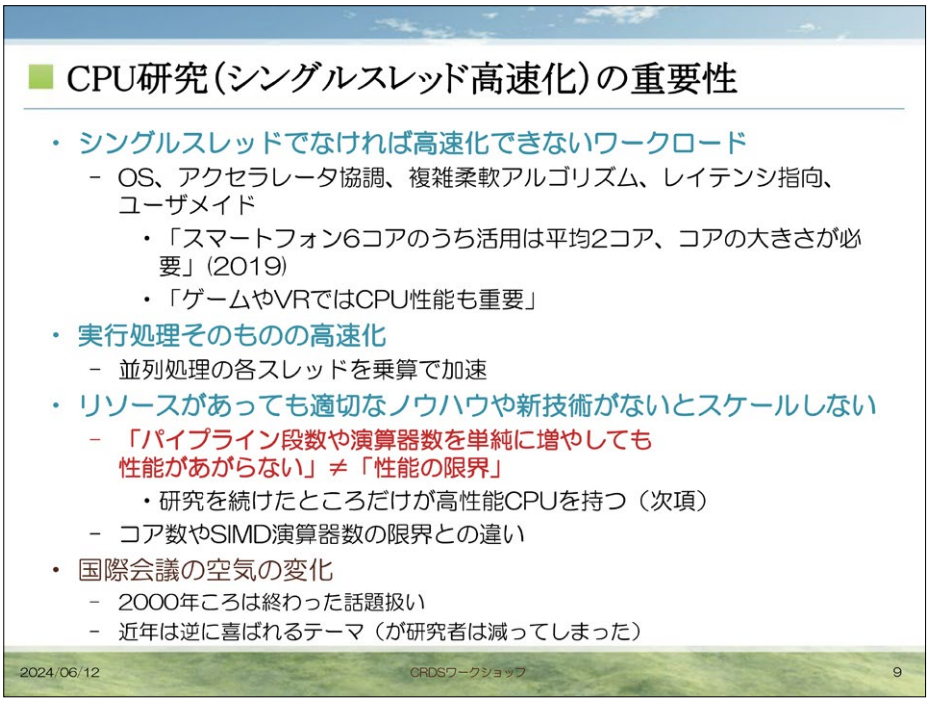


図2-3-5　　CPU研究の重要性

なぜCPUが重要なのかという理由は、大きく分けて2つある。1つは、シングルスレッドでなければ高速化できないワークロードが存在することである。たとえば、OSの動作やデバイスの協調、複雑で柔軟なアルゴリズムの処理、さらにはプログラムを素早く書き上げるといった状況がある。この点については少し古い話ではあるが、スマートフォンに6コアのプロセッサーが搭載されているものの、実際に活用されるのは平均して2コア程度であり、大きなコアが求められるという議論が2019年頃にされていた。依然としてゲームやVRなど、

ドメイン固有アーキテクチャーが重要とされるアプリケーションでは、CPUの性能が不可欠だと言われている。

もう一つ重要なのは、実行処理の高速化である。シングルスレッドの技術が向上すれば、それを並列処理に応用して、各スレッドの処理速度を加速することができる。たとえば、2000年ごろにはメモリー関係のマネージメント技術の研究がさかんに行われていたが、今ではDSAのアクセラレーターに搭載されようとしている。CPUは将来の様々なアクセラレーターが直面する課題を先取りして研究されてきている。そのため、技術の積み重ねが高速なアクセラレーターの品質に影響すると考えられる。

注意しなければならないのは、シングルスレッドの場合、リソースがあっても適切なノウハウや新技術がないとスケールしないことがある。そのため、パイプラインの段数や演算器を増やしても性能が向上しない場合があるが、それが性能の限界ではない、というところが難しいところである。

この研究を続ければ、高性能なCPUを作り続けられるということになる。これは、たとえばコア数を単純に増やした場合に直面するアムダールの法則による性能限界とは違うということである。

国際会議は先ほど井上先生からお話があったように、空気はやはり変化していると感じられる。2000年ごろにはもう終わった話題として扱われ、シングルスレッドなどと言うと門前払いされるような雰囲気があったが、2010年ごろから変わり始め、今はペーパーメリットの中にシングルスレッドなどと書くと喜ばれるテーマになっている。ここ10年から15年の間に、研究者が減ってしまったという感じもする。

現在のシングルスレッドアーキテクチャーの高性能化について図2-3-6に示す。

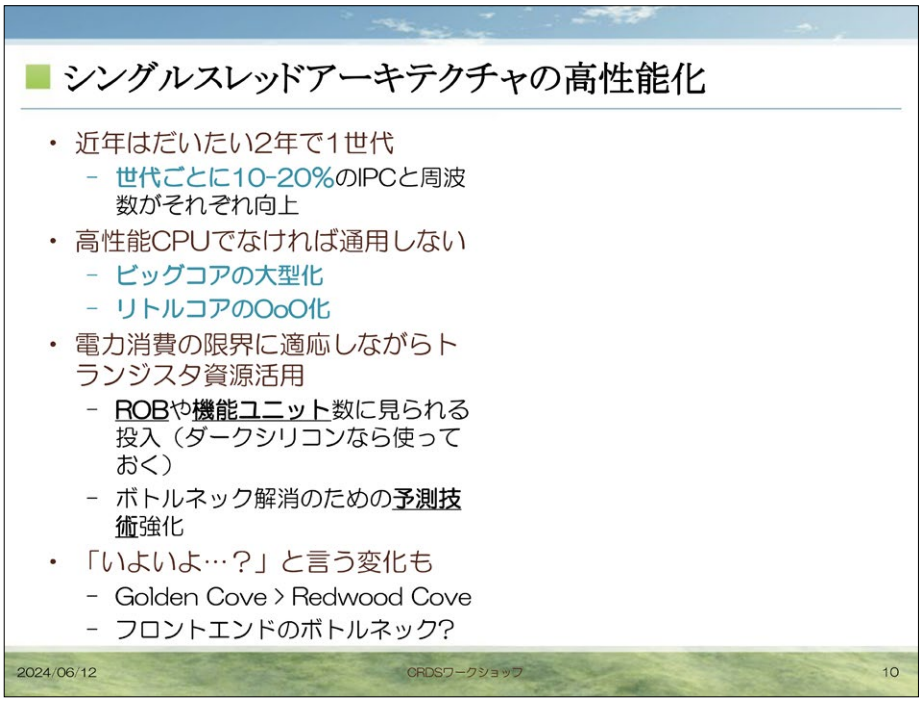


図2-3-6　　シングルスレッドアーキテクチャーの高性能化

要するに、大体2年ごとに1世代が登場し、それぞれの世代でIPC（Instruction Per Cycle サイクルあたりの命令実行数）と周波数が10%～20%ずつ向上している。研究者の立場から見ると意外と順調だなと感じる。通常、このペースで進めば、10年後には性能が約2倍になる。先ほど木村さんのスライドでも、シングルスレッドの性能が10年で約2倍になっている例があったと思う。高性能なCPUでなければ対応できないニーズが増えており、ビッグコアはますます大型化している。市場で売れるためには、性能を大幅に向上させる必

要があり、リトルコアも、低消費電力を売りにしているが、今では当たり前にアウト・オブ・オーダー処理を採用し、より複雑なタスクをこなせるようになっている。

また、先ほど述べたように、スイッチしないトランジスタを活用する必要があり、そのような資源に投入しながら進化している。

ただ、いよいよ大きな変化があるのではないかという感じもある。次世代のIntelのRedwood Coveについては、前世代から性能が向上しないとの懸念があり、本当にボトルネックに直面しているのかどうかが焦点になっている。

図2-3-7に、CPUの革新について、日本からの提案を示す。

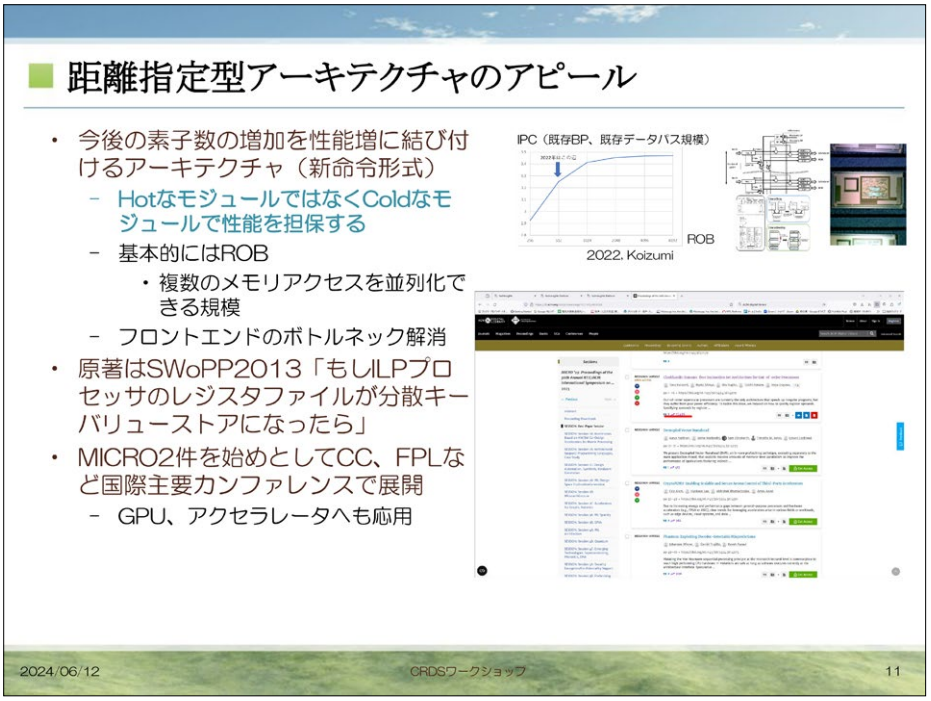


図2-3-7　　距離指定型アーキテクチャー

この提案は、2013年頃のSWoPP（Summer United Workshops on Parallel, Distributed and Cooperative Processing　並列/分散/協調処理に関するサマー・ワークショップ）で発表され、その後MICRO（IEEE/ACM International Symposium on Microarchitecture）などの主要なカンファレンスで展開している。去年のMICRO2023の発表がACM Digital Libraryで1万4,000近くのダウンロード数を記録し、他のベストペーパーが数百ダウンロードである中で非常に高い注目を集めている。コンピューティングの分野ではACM Digital Library内で2位になっている。研究としてはこれまで到達できなかったところに来ているという感じがする。

エネルギー資源の改革

また、スコープ外の話であるが、4つ目の改革としてエネルギー資源の改革が挙げられる。比較的無尽蔵で地球温暖化に寄与しない革新的なエネルギー開発が急務であると考えている。生成系AIなどがコンピューターのリソースを使って新しい価値を産み、その価値が次の資源獲得競争の原資となる世界に移行しつつある。リソースを使い、産み出した価値で次のリソースを競い合うという社会が形成されつつある。これまで省電力

と言えば主にチップやせいぜいウェアハウスの省電力に焦点が置かれてきたが、ここ2～3年で世界全体の計算電力についても議論が深まっている。そうすると、アジアでAIの使用が過剰だという議論も出ており、これが二酸化炭素排出量の問題と結びつく可能性がある。

したがって、われわれは省電力に尽力しているが、たとえコンピューターの効率を100倍に向上させたとしても、ユーザーが100倍以上の負荷をかけることが予想される。そのため、現在の余裕のある状況を利用して、新エネルギー創出にコンピューターの力を活用することが重要であると考えている。この目的のために、コンピューティングの支援も検討したいと考えている。

まとめると、アーキテクチャーではシーズだけでなくてビジョンのほうも考えたいということと、地味ながらCPUが新展開を迎えているということ、また、エネルギー技術についてもコンピューティングで支援したいということである。

2.4 竹内 健（東京大学大学院工学系研究科電気系工学専攻電子知能情報学講座）

半導体を巡る状況

自分は相対的にデバイス屋だと思うので、徹底的に半導体デバイスから身もふたもない話をする。先ほどからデバイスとアーキテクチャーとか、ソフトウェアも含めて融合しなければいけないというのは、まさにそのとおりだと思う。ただ、デバイスがいつブレークするかというのは、やはりマーケットが大きい。現状では、ご存知のようにGPUはかなり高価になっている。そのビジネスに乗ってしまい、さまざまな技術課題が吹き飛んでしまったという感がある。やはりマーケットが技術をけん引するというのはそういうことだとまざまざと感じている。

今はカーボンニュートラルということで低電力化もしなければいけなくなり、世の中の動向によってデバイスはどんどん進化する。まさに今がその変化の時期だと思っている。

学内のことで恐縮ではあるが、半導体を何とかしようということが言われており、本学でも半導体関係のプログラムをやっている。一つは、CTOやCOOなどの経営陣の方々に話をしていただく戦略的な経営の話と、技術系の方々に話していただくという講義を学部でやっている。これは新しい講義なので月・火の夜の5時～7時という学生から見るといまひとつな時間帯であるが、毎回150人ぐらい、電気系以外からも聴講してくれている。半導体は、大学ではにわかに盛り上がっている。

ご存じのようにAIチップというのもNVIDIA独り勝ちは許さじということでいろんなところから出ている。特徴的なのは従来の半導体メーカーだけじゃなくて、今日、Appleからも発表あったようであるが、MicrosoftやPreferred NetworksがAI半導体をやっていたり、TURINGという自動運転の車をやっているところがアクセラレーターを作っていたり、TIER IVも半導体を作っていたりする。日本でも新しい企業が半導体をやり始めている。中身を見ると私と同世代のような、電気メーカーで一度は夢破れた人たちが敗者復活戦みたいな形でやっていることが多いようである。今はそれでもいいが、若い人を育てないと今後はまずいのではないかと思っている。

学会の状況

図2-4-1にISSCCの論文投稿数を示す。

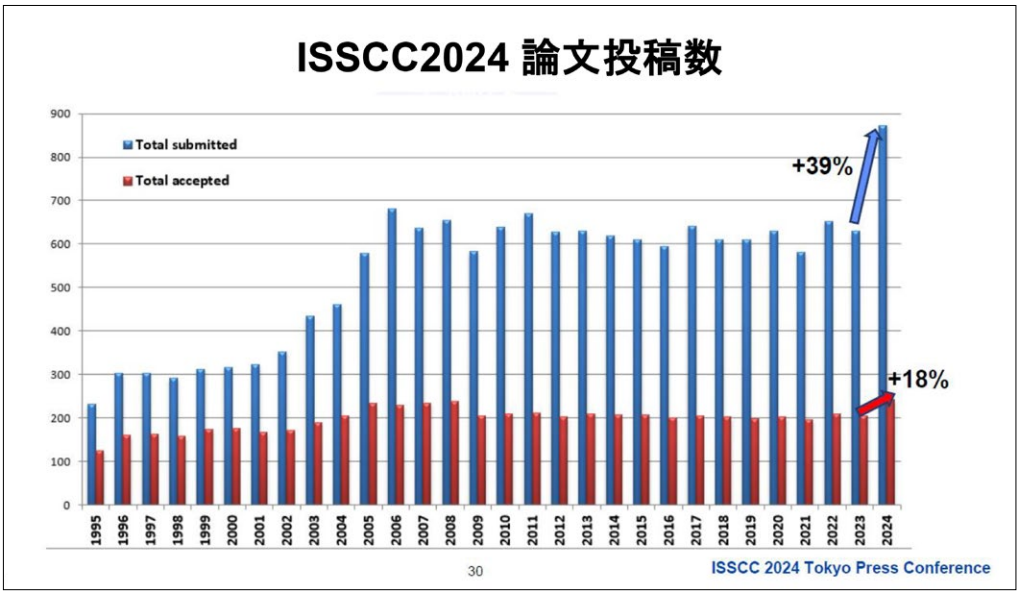


図2-4-1　　ISSCC論文投稿数

これを見ると、今、ものすごくフィーバーしており、2023年から2024年にかけて突然4割も論文投稿が増えている。 VLSIシンポジウムでも同時期に600件から1,000件近くに一気に増えている。 IEDMというデバイス系の学会においても、過去最高になっている。半導体の回路、デバイスはどこの学会も過去最高の投稿数ということで、大変な盛況になっている。

中身を見ていくと、IEDM2023では、図2-4-2に示すように、全体で30セッションのうち、10セッションくらいがコンピューティングに関連する。 ISSCCも似たようなもので、3分の1ぐらいがマシンラーニングであるとか、Compute in Memory、LLMなどが入っている。

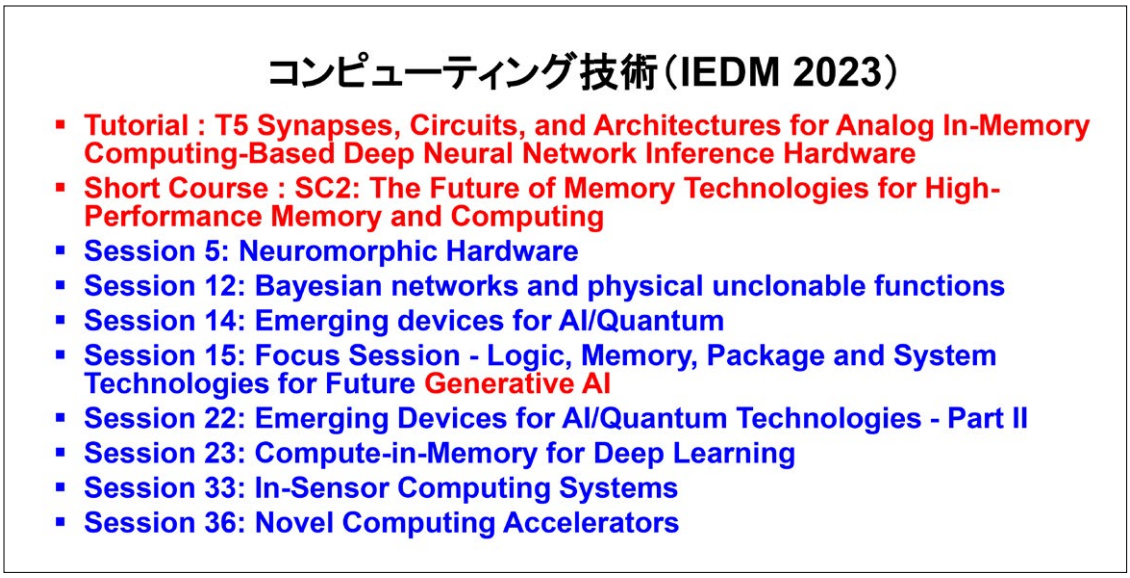


図2-4-2　　コンピューティング関連セッション

その中で統一的に議論をされているのは、デバイス機械学習のモデルであったり、あるいはISSCCのプレナリーでは、ベンチャーキャピタルの人がハードウェア、半導体までを一気通貫した開発が必要で、それをSoftware 2.0 Systemと呼んでいたりする。

似たようなことをIBMも言っており、彼らは上から下じゃなくて、下から上へ、デバイスからアーキテクチャー、マイクロアーキテクチャーまでcodesignが必要だと言っている。

デバイスからコンピューティングまでの大きな国プロをけん引しているスタンフォードのフィリップ・ウォンという先生が、マテリアルから回路、アーキテクチャー、ソフトウェア、アルゴリズムに至るcodesignが必要だということを盛んに言っている。難しいのではあるが、こういうところが共通の課題だと思われる。

こういう動向であり、学会も随分変わってきている。 VLSIシンポジウムには回路とテクノロジーのVLSI TECHNOLOGYとVLSI CIRCUITSがあったのが、VLSI SYMPOSIUMとして一つになってしまった。ヨーロッパでは、デバイスの学会であるESSDERCと回路の学会であるESSCIRCが、ESSERCという1つになってしまった。学会としてもやはりCross Layer、つまりレイヤーを取り払って全体を最適化しようという動きが顕在化している。

In-Memory Computing

最後に少しだけ自分のやってることに近いことを紹介する。図2-4-3にIn-Memory Computingを示す。

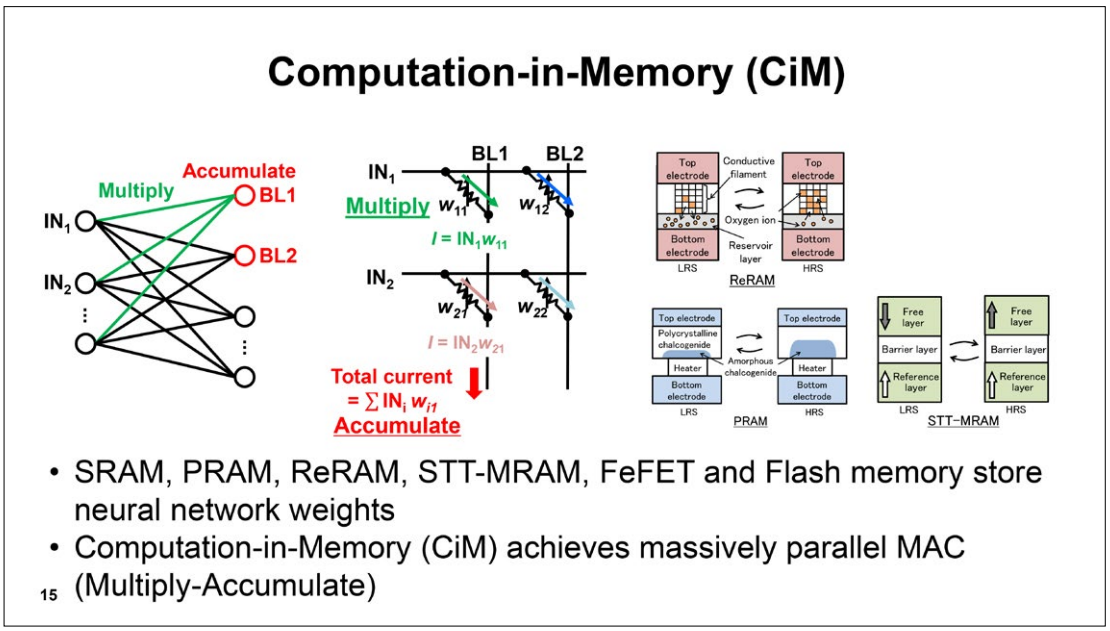


図2-4-3　　In-Memory Computing

メモリーを使ってMAC演算を並列で高速に処理するというものである。これもデバイスの発達と関連が強く、超並列、大容量にMAC演算をするということは、そもそもデバイスが大容量にできていないといけないということである。ここには何十年もの長い苦しい歴史があるが、”Emerging Memories Emerged!”と言って、抵抗変化型メモリーやMRAMなどが、TSMCやらUMCやらサムソンからもファウンドリーでできており、256ギガビットというものもあるし、3次元実装もできるようになってきた。ようやくデバイスの準備ができてきたので、アーキテクチャーをリアルに考えられるタイミングになってきたと言える。

次に、そういったIn-Memory ComputingでMAC演算などは高速にできるようになってきたが、MAC演算以外にも当然たくさんあるわけで、必然的にヘテロジーニアスなインテグレーション、トラディショナルなCPUとの組み合わせというのも必要になる。そうすると、さまざまなデバイスが組み合わされるということになり、その管理が必要になり、ソフトウェアのサポートやプログラミングモデル、あるいはコンパイラーなどといった技術も必要になる。デバイスだけ作ればいいというわけではない。

3次元集積化も可能になってきており、さまざまなチップを3次元に集積化できる時代にもなってきた。チップ上に色々なコンピューティングを実装できるという時代が来たと思う。

【質疑応答】

高島：井上先生がおっしゃっていた、コンピューターサイエンスとデバイスサイエンスが一緒になるというのは、もう学会レベルでは既に進んでいるということでしょうか？

竹内：そうです。たとえば学会へ行くと、本会議はともかくとして、たとえばショートコースとかフォーラムといったところではCSの人が発表する。要するに、本会議に行くとデバイスとか回路の話は幾らでも聞けるけれども、それを使うためのソフトウェアの話っていうのはなかなか本会議では聞けないので、それがショートコースとかフォーラムでたくさん議論されている。一番注目すべきは、デバイスの人とCSの人が組んでファンディングに対応するチームを作っているということである。論文の謝辞を見ると分かるのであるが、少なくともアメリカとか、あとヨーロッパでもCNETとかIMECとか、意識的にファンディングで結び付いてるということは感じる。日本においてそう言う動きが進まないのは、JSTにも責任がある。

井上：VLシンポジウムについて結構衝撃を受けたが、学会の進め方として一緒に議論をする場にしなければならないということを考えた動きなのか？

竹内：はい。実は、私はプログラムチェアだったのであるが、一つ大きかったのは、どちらに分けていいか分からない論文が提出されたということである。つまり、リザーバーコンピューティングなどはそうだと思うが、デバイスの人がコンピューティングをやっている。そうすると、本当にこの境界が曖昧になってくる。実は一緒になる前に、境界領域のサブコミッティーを作ったが、そこが非常に大きくなってくる。そうすると、もうこれは一緒でいいじゃないかということになった。実はIEEEのEDSとSSCSという違うソサイエティーであるので、結構微妙なところもあったが、そこを乗り越えて、要は聴講する人にとってみると一緒になったほうがいいということになった。この後には、たとえばコンピューターソサエティーがあり、ISCAとかMICROが一緒になったり、その前にDACとかICCADとかが合流したりというのもあるかもしれない。

井上：あり得ますね。最初はジョイントとか連続開催とか、そういうとこから始まるのでしょうね。

木村：そういうことを日本で起こそうという動きはないのか？

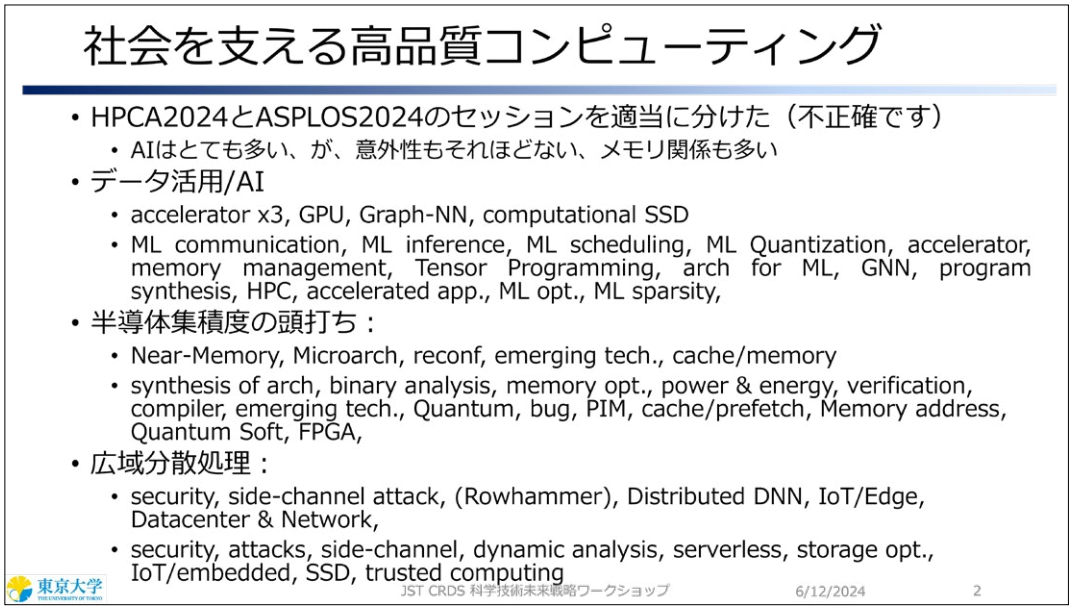
竹内：どうですかね。研究会レベルではあるとは思うが。自分も含めて多くの研究者は国際会議で頑張っているので、ちょっと日本までは手が回らないという状況である。ただ、応用物理学会はスコープを広げる方向に行っていると思う。

2.5 中村 宏（東京大学大学院情報理工学系研究科システム情報学専攻認識行動情報学講座）

コンピューティングは、より良い社会を実現するためのものになるべきであり、そのために必要となるこれからのコンピューティングアーキテクチャーについて話題を提供する。

社会を支える高品質コンピューティング

まず、社会を支える高品質コンピューティングについて、今年の国際会議HPCA 2024[1]とASPLOS 2024[2]のセッションから研究動向を説明する（図2-5-1）。全般的には、AIのセッションが多いが意外性もそれほどない。また、メモリー関係のセッションも多い。個別に見ると、まず、データ活用やAIに関係するセッションが多く、アクセラレーターやGPU、computational SSDをAIで活用をするといったセッションがある。半導体集積度の頭打ちという観点から見ると、メモリー階層やマイクロアーキテクチャー、リコンフィギャラブル、新しいテクノロジーの活用など、いわゆるアーキテクチャーのセッションがいろいろとある。広域分散処理の観点では、社会全体で使うための広域分散環境におけるエッジとクラウドの連携やセキュリティーのセッションがある。セキュリティーでは、ハードウェア関係のセキュリティーのセッション数が増えている。



注）各項目、上段の・がHPCA2024、下段の・がASPLOS2024

図2-5-1　国際学会HPCA2024、ASPLOS2024の動向

高品質コンピューティングの要件

次に、高品質コンピューティングの要件について説明する（図2-5-2）。

1点目は高スループットである。GPUや、Generative AIのための新規アーキテクチャーも出て来ている。

2点目は低消費電力であり、非常に重要である。データ移動（含read/write）を抑えるためにメモリー階

1 The International Symposium on High-Performance Computer Architecture, https://hpca-conf.org/2024/

2 The ACM International Conference on Architectural Support for Programming Languages and Operating Systems, https://asplos-conference.org/asplos2024/index.html

層の再考が起きており、たとえば、In-MemoryやNear-Memoryもこれに含まれる。最近は余りあるトランジスタが演算器ではなくキャッシュとして使われ、キャッシュも含めて演算器の稼働率を上げようとしている。そう考えると、演算器の稼働率向上は至上命題ではなく、非稼働時の消費電力がゼロになれば無駄な演算器を置いてもいいのではないかとなる。また、ドメインスペシフィックアーキテクチャーも重要であり、応用とアーキテクチャーの協調設計により、応用に応じてデータの移動を抑えるデータパスを構成することで高性能、低電力を目指すことが必要である。

3点目は、実時間性である。実時間性の要件の一つは低遅延であること、もう一つは低揺らぎであることである。今のCPUでは平均的な実行時間を短くしようとして設計しているが、リアルタイムアプリケーションには実行時間の揺らぎを抑えることが求められる。そのためには、今の入出力のソフトウェアスタックを考え直す必要がある。特にドメインスペシフィックになってくると、エッジデバイスも含めて、入出力デバイスの種類によっては保護機構を緩めユーザーが直接触ることを許容するといったことが考えられる。コンピューティングのカスタマイズではドメインスペシフィックアーキテクチャー、コミュニケーションのI/Oのカスタマイズではドメインスペシフィックコミュニケーション I/Oといった考え方が必要である。これまでの高スループットは、高揺らぎが前提の考え方だったが、今後はそうではない設計パラダイムが必要である。

4点目は、高セキュアである。サイドチャネルアタックや順同型暗号演算などは、これから研究すべきトピックである。

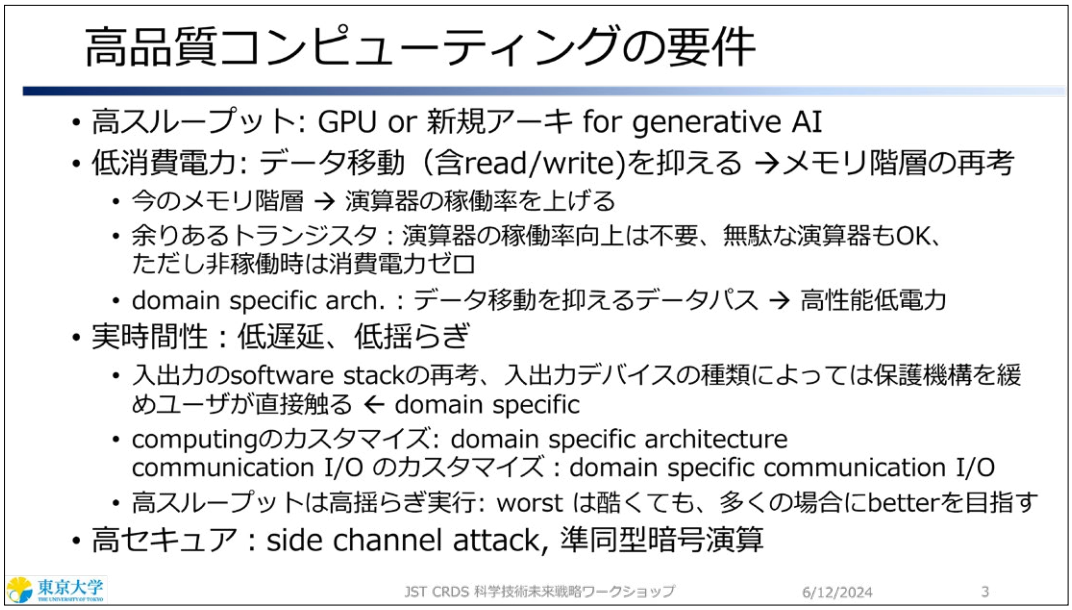


図2-5-2　　高品質コンピューティングの要件

高品質コンピューティングの実現に向けた課題

次に、高品質コンピューティングの実現に向けた課題を説明する（図2-5-3）。

一つ目は、ドメインスペシフィックを考えるのであれば、その応用展開を考える必要があるという点である。たとえば、自動運転や生成AIによるロボット制御などがある。生成AIによるロボット制御では、たとえば、「あそこの赤い服を持ってきて」といったあいまいな指示を言語で与える場合のエッジ側とサーバー側の処理の分担や通信に求められる時間を考える必要がある。エッジ側には、エッジ側が担う処理にカスタマイズした小さいLLMを置くとか、LLMを保持できるようなSRAMを置くことが必要になるかもしれない。ただ、エッジ側とサーバー側との通信が必要となるため、I/O通信が重要になる。このように、応用を考えてコンピューティ

ングはどうあるべきかを考えないといけない。そのためには、応用の研究者たちとのコミュニティーが重要になってくる。

もう一つは、新しいシステム設計ができる人材の育成である。人材の育成に関しては、指導教員だけではなく、コミュニティーなどのもう少し広い場での多様な横の繋がりが人を育てる。また、長いスパンでの人材育成や研究力の強化が必要である。たとえば、研究プロジェクトの評価は、終了時の成果評価だけでなく、研究プロジェクトに参加した研究者（学生も含めて）が、5年後、10年後に、どのように活躍しているのかまで評価しないといけない。さらに、ハードウェアを含むデジタルシステム設計力も重要である。アーキテクチャーには、命令セットアーキテクチャーを含めて抽象化レイヤー、あるいは設計階層という概念がある。CMOSでは設計階層が作りやすかったが、新規デバイスにはAdvantageとDisadvantageがある。 Advantageは1個上の抽象化レイヤーで簡単に使いこなせるが、Disadvantageはかなり設計階層上流で解決する必要があるという問題もある。抽象化レイヤー、設計階層という簡単な概念ではうまく扱い切れない。こういった問題に対処するためには、OSなどのソフトウェアの設計力も重要だが、ハードウェアの設計力がある人がアーキテクチャーまで考えることが必要である。

最後は、新しい計算原理を探求できる人材の育成である。物理や数学に加えて、プログラミング力が重要である。たとえば、新規デバイスを使う場合、全く新しいアルゴリズムを探索する必要が出てくる。プログラミング力がある人が量子のいろいろなアルゴリズムを考えたり、新規デバイスの使い方を考えたりするケースが考えられる。

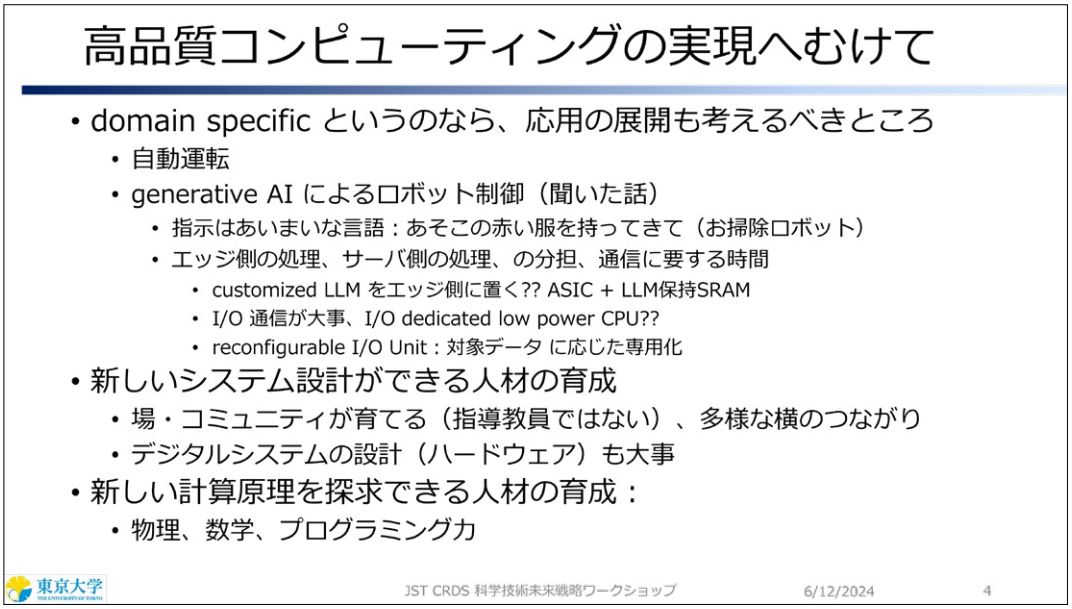


図2-5-3　　高品質コンピューティングの実現へむけて

エネルギー消費量、人材育成

最後に、参考としてエネルギー消費量、人材育成について説明する。

エネルギー消費量に関しては、IEAのレポート「Electricity 2024」[3]によると、2026年には1,000TWhの電力が必要となると予測されている。これは、日本の総エネルギー消費量とほぼ同じであり、何とかしないと

3　IEA, “Electricity 2024”, https://www.iea.org/reports/electricity-2024/executive-summary

いけない。Open AIのChatGPTの平均電力需要はGoogle Searchの10倍となることが示されている。データセンターのエネルギー消費量も全く楽観的ではない。

また、生成AIをGreenにすることが、今後、重要となる。生成AIのCO2排出機会は、大きく、学習、推論、ハードウェア製造に分けられる。生成AIの1回の学習のための計算量は、1回の推論よりもはるかに大きい。『Harvard Business Review』[4]で述べられているように、各自がそれぞれモデルに学習させるのではなく、1回学習させたモデルをファインチューニングして利用すれば、学習に要するエネルギー消費量を削減可能となる。また、製造時の見積もりは難しいが、生涯エネルギー消費量は、製造時のエネルギー消費量も含めて考えないといけない。

人材育成に関しては、東京大学では、さまざまな横断型教育プログラム[5]や半導体教育プログラム[6]を開始している。横断型教育プログラムの修了生は、数理・データサイエンス教育プログラムが一番多く、二番目がサイバーセキュリティー教育プログラムとなっている。半導体教育プログラムからも2年後には多くの修了生がでるだろう。こういった横断型教育プログラムには、他学部の学生も高い関心を持っており、工学部以外の他学部の学生も多く受講している状況である。人材育成はとても大事で、そのためには優れた研究を行い、皆さんをアトラクティブにする必要があると思っている。また、AIチップ設計拠点に関しては、NEDOによる「AIチップ開発加速のためのイノベーション推進事業[7]」があった。まだ試作に留まるものもあり、継続的に取り組む必要がある。

【質疑・討議】

高島：高品質という言葉を使っているが、高性能とはどう違うのか？

中村：高性能というとスループットや実時間性であり、消費電力やセキュリティーは高性能とは言わない。特に、低揺らぎは、遅くてもいいから揺らがないことであり、やはり性能ではない。社会を支えるクオリティーには何が必要なのか、コンピューティングにはどういうクオリティーが要求されるのかを考える必要がある。

4 Harvard Business Review , “How to Make Generative AI Greener”, https://hbr.org/2023/07/how-to-make-generative-ai-greener

5 東京大学, “横断型教育プログラム（University-wide Education Program”, https://www.u-tokyo.ac.jp/ja/students/special-activities/University-wideEducationProgram.html

6 東京大学, “半導体教育プログラム”, http://www.dlab.t.u-tokyo.ac.jp/Semiconductor/

7 NEDO, “AIチップ開発加速のためのイノベーション推進事業”, https://www.nedo.go.jp/activities/ZZJP_100142.html

2.6 本村 真人（東京工業大学 科学技術創成研究院）

革新的コンピューティングを領域とするCRESTプロジェクトを５年間やってきた。学習するモデルに基づく時空間展開型アーキテクチャーの創出と応用という内容で、機械学習、数理科学、アルゴリズムの研究者、そしてアーキテクチャーからチップ実装までを扱うグループとともに、機械学習や最適化に関するアルゴリズム、チップ実装を行った。社会実装に向けては、ISSCC（International Solid-State Circuits Conference）で発表したようなチップの応用開拓（実用化）コンセプトを出すといったことを続けていかねばならないが、それと並行して、次の研究テーマ（さらに将来に向けてどういう研究が大事かということ）を考えている。

脳とコンピューター

そのときに重要な視点のひとつが、図2-6-1に示すような、大脳とGPUの違いである。人の大脳とGPUの単体モジュールを比較してみると、大脳の演算性能は2 Tera Ops/secぐらいで、GPUはすでにそれを超えて数十倍もの開きがある。一方、メモリー容量を比べると、大脳の場合はシナプス結合で蓄えられているメモリー容量を測定したところPeta Byteレベルであるが、GPUはHBM（High Bandwidth Memory）で32Giga Byteということで数万倍の開きがある。消費電力でいうと、20Wと400W超で数倍の開きがあるという状況になっている。

メモリー容量を、GPUで人の大脳に合わせようとすると数万個並べる必要がある。そのときの演算性能は数Exa Ops/secになり、何桁も脳の性能を超えて、消費電力は数MWクラスの計算機ということになる。これは、いわゆるAIデータセンターで起きていることである。

この数字は、HPCA2024（International Symposium on High-Performance Computer Architecture）のキーノートでNir Shavitというmitの先生が話したところから持ってきた。この講演をきいて同感に思ったのは「We do not know the algorithm」で、どのように動いているのかを基本的に理解していない。逆にいうと、そこにチャンスがあるという読み方ができる。

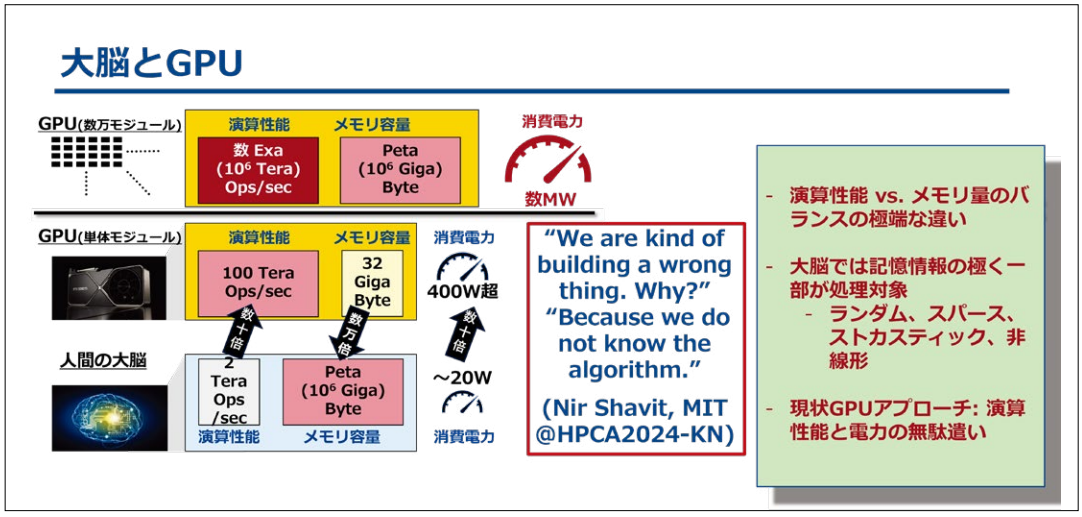


図2-6-1　　大脳とGPUの違い

この単純な比較から、演算性能とメモリー量のバランスが大脳とGPUとで極端に違うということがわかる。

大脳が、これだけ大量のメモリーを持ち、これだけ少ない演算性能で、認知機能を実現できているということは、どう動いているかは分からないにしても、記憶されている情報のごく一部だけを処理対象にしているであろうと考えることができる。つまり、ランダムに抽出されたスパースなネットワークに対して、確率的で非線形な情報処理が行われている。それらが何らかの形でキーになっている。そういう観点で見ると、今のGPUの演算性能と電力は無駄遣いであると言わざるを得ない。そう考えて、その先のコンピューティングを切り開いていくことが重要である。

コンピューティング基盤の構築

CRESTで研究開発した技術（ファウンデーション）の上に考えている内容のひとつが、ニューラルネットに関するもので、そこに強い宝くじ仮説という話がある（図2-6-2左）。これは、ニューラルネットの中のごく一部のネットワークを乱数で初期化したものをそのまま抽出してくるだけで学習ができてしまうという理論の枠組みである。こういったニューラルネット技術や離散系の組み合わせ最適化問題を解くようなアニーリング技術、特に情報を超高次元の分散ベクトルで表現し処理対象として解いている、というあたりの共通性に着目している。大量の超高次元分散ベクトル集合を統一アーキテクチャー思想の上で処理して知的価値を生み出す、これまでにないコンピューティング基盤の構築というような夢を掲げて研究を進めている。

その延長線上で、LLM（Large Language Model）のトランスフォーマーにおけるアテンション機構の話もこういう枠組みの中で見ると非常に興味深い（図2-6-2右）。たとえば、QKVという、入力からモディファイされてつくられる超高次元で非常にスパースなテンソル間の演算がある。要は、情報の超高次元な分散ベクトル表現が共通の鍵で、こういう抽象レベルに落とし込むことができ、これが今の情報処理の仕組みの中で非常に大きな位置を占めるようになっている。この超高次元で分散表現されたベクトルを如何に効率よく処理していくかが、今後のアーキテクチャーの大きな進歩につながっていくのではないか。

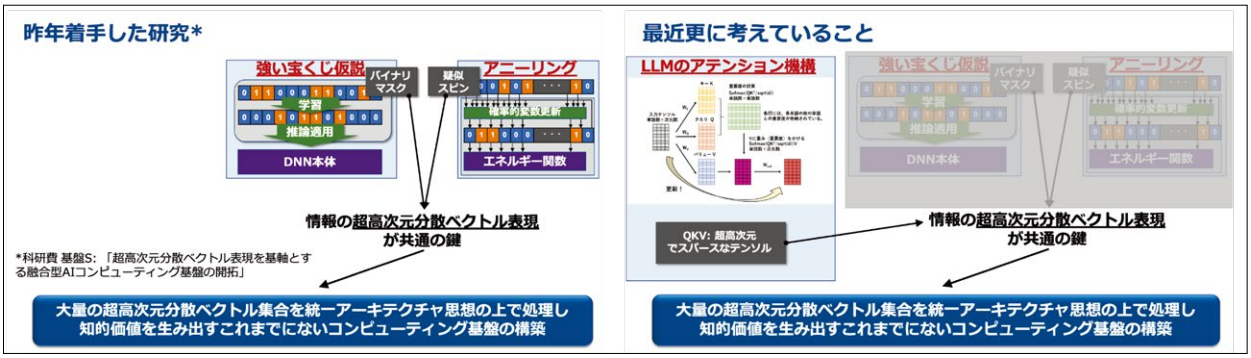


図2-6-2　　着手した研究と考えていること

まとめると、こういう超高次元データを対象として識別、検出、認識、予測といった知的な処理をする。図2-6-3にある構造型情報処理は、リコンフィギュラブル処理のようなもので、インメモリー的な要素も含む形で構想している。具体的には、大規模な超高次元データ群に対して、ランダムでスパースな、そして確率的でイベントドリブンな、しかもノンリニアな処理を行う。ノンリニアの要素が散りばめられた情報処理が重要で、それがデジタルに実行されるコンピューティングが非常に重要になっていく。ここでデジタルといっているのは、アナログで解こうとするといろんなことを変える必要があるため、まずはデジタルの枠組みの中で難しい課題を解いていくというのが研究のアプローチとして重要だと考えている。

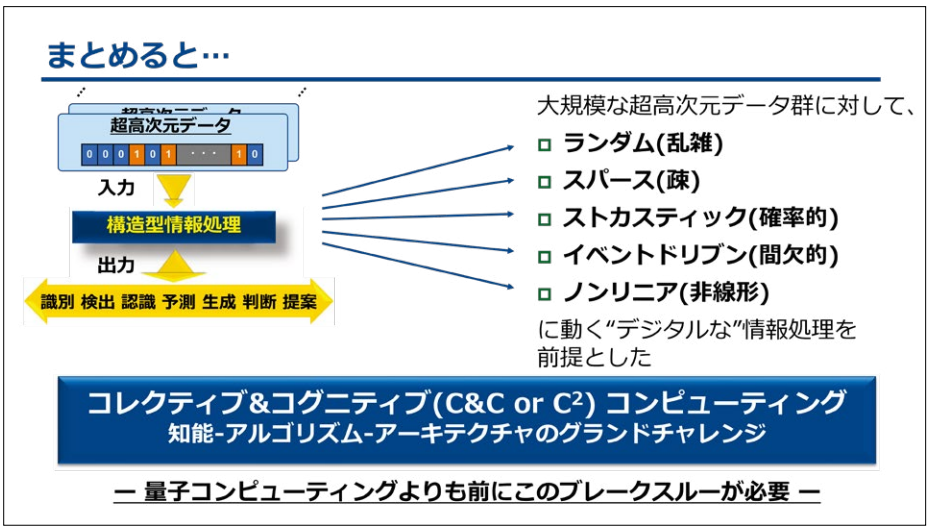


図 2-6-3　　構造型情報処理

コレクティブ＆コグニティブ

今回話題にするにあたり、付けた呼称が「コレクティブ＆コグニティブ」である。コレクティブ、つまり集合的な高次元のデータ群に対して、認知系の処理をするということで、コグニティブ。ブレインモルフィックやブレインインスパイヤードという言い方もあるが、脳の信号レベルの情報処理をまねるというよりも、抽象化された機能を目標にして現在の情報処理を見直していくというスタンスが重要であると考えて、この呼称にした。

こういうアプローチは実はなかなか難しい。たとえば、スパースの情報処理はどうしても非効率なやり方となるため、なるべくスパーシティーをなくす方向に変換をして処理する。これに対して、スパース性やランダム性が本質になると考える情報処理を、知能アルゴリズムアーキテクチャーのグランドチャレンジであると捉えてアプローチする方が、現状の社会が抱える電力消費量の問題等も想定したときに重要になると考えている。

アルゴリズム - アーキテクチャーの協創とチップ実証

もうひとつ重要な視点として、アルゴリズムとアーキテクチャーの協創とチップ実証がある（図 2-6-4）。アルゴリズムとしての AI モデルがあり、それを処理アーキテクチャーとして落とし込む。その間にワークロード特性があり、そこがランデブーポイントになる。ここを互いに共有しながら研究開発を行い、そこから生まれるのが革新的 AI モデル技術や革新的処理アーキテクチャーになる。それが、軽量化、疎結合化、エネルギー効率大幅向上へとつながる。これをチップで実証して応用展開していく。このようなサイクルを思い描いて研究を推進している。

アルゴリズム、アーキテクチャー、チップといった領域間の垣根をなくすことが重要である。現状は、アルゴリズムやアーキテクチャーから回路レイヤーが離れがちで、互いのことを見ない傾向がある。そこが、コンピューティングの効率化や新しいコンピューティングを生み出していくことに対する障壁になっている。

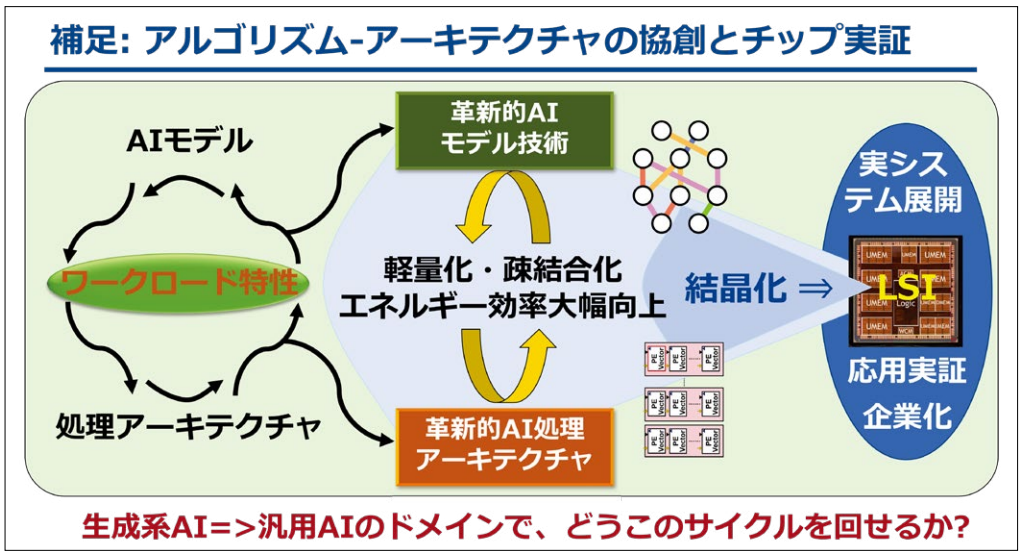


図 2-6-4　　アルゴリズム - アーキテクチャーの協創とチップ実証

【質疑応答】

高島：結局、アルゴリズムとチップがあり、そのあいだをアーキテクチャーが埋めるということで、やはり全部を同時並行でやらないといけないと多くの先生方が指摘されたが、本村先生もそういうことか。

本村：そうである。

高島：木村上席フェローは、このあたりの点をどう考えるか。

木村：全く同感である。他の先生も、JST のファンディング等で CS と DS を掛けたところをやる必要があると言われた。やるとすれば、アーキテクチャーの人とモデルやアルゴリズムの人とが共同でプロジェクト提案して欲しい、ということになりそうだが、それでうまくいくものなのかと思ってしまうところがある。

高島：結構大きな問題であり、この後の全体討論で議論したい。

木村：他の先生から、日本の学会は相手にしないというような発言もあった。そこは、まさに私が今回問題意識を持ったところでもある。そこをどうにかしないと、この分野が大きくなっていかない。また、世の中の期待にも応えられない。どのようにして実現するのかをずっと悩んでおり、そこを見出すことが、本ワークショップを開催した意図のひとつである。

3 ｜ 討議

コンピューティングの革新は何をめざすのか？

高島：コンピューティングの革新というのは何をめざしているのだろうか？

本村：まず、私が脳の話を出したのは、やはり計算のエネルギー効率を向上するためである。アーキテクチャーは究極的には効率に落ちる。応用を開いていくということも非常に重要ではあるが、それは既存の計算プラットフォームの上でソフトウェアは何でも書くことができ、その上でさまざまな応用が広がるということだ。しかし、あまり向いてないアーキテクチャーだとエネルギー効率が下がる。エネルギー効率を上げるというのは社会的な問題であり、エネルギー効率を上げるというのが計算上のメインの課題になると思っている。そのときに、脳の仕組みやそのやり方を横目で見ながら、それを計算プラットフォームの上に落とし込んでいくという考え方は重要になると思う。方法論として重要になる。したがって、脳は目的ではなく、あくまでも方法論である。

今のディープラーニングとか生成系AIの仕組み自体がアーキテクチャーにまで落とし込まれていない。モデル構造を考える際に、場当たり的にいろんな構造をテストしながら人の脳はこういうことをやっているのではないかという仮説が、若干の影響をしつつモデルができている。しかし、そのモデルが走っているアーキテクチャーはそれらのモデル構造の影響を受けずに開発された並列アーキテクチャーであり、それでは電力効率が上がらない。これが今の電力問題の一つの大きな要因になっている。そこが研究課題である。

木村：ディープラーニングが脳を模して成功したが、そもそもデジタルコンピューターという合わないものの上に無理やり載せたから、電力効率が非常に悪くなってしまった。もう一度その基本に戻って、もともと脳の話から始まったのであれば、脳というのは20Wで動くのだから、そこを参考にしながら、もう一度そのアーキテクチャーをリデザインしたら、もっと低電力でうまくいくような仕組みがあるのではないか。そうすると、今のエネルギー問題の解決にも資することになる。脳を使うことが目的ではなくて、それは一つのスターティングポイントであり、考え方の指針とすればもっとうまくできる。本村先生の指摘はこういう理解でよいか？

本村：そういう趣旨である。

中村：今のコンピューティングは連続であるとか局所性があるという場合に非常に有効であったが、データのスパーシティーというのは本質的に対象が持ってる特徴なので、それに対して本村先生がここに着目して、それをどうやってうまくやるのか、方法論として脳を持ち出されている。私はそこに本質的な問題があり、それは解かなければならない問題であると理解した。

脳とコンピューティング

木村：脳を参考にしてアーキテクチャーを考えるときに、リザーバーというのは有効なのか？ リザーバーコンピューティング的なものを持ってくると、人間の脳のようにエネルギー効率が良くなるのか？

浅井：リザーバー単体では脳になり得ない。全体が統合して初めて意味のあることができるようになるので、リザーバーだけを極めてもそうはならない。

学習の効率ということを考えると、どこの演算の効率化を考えるか次第であるが、リザーバーは入ってきた入力を高次元にマップするので、あとの学習が非常に簡単になる。1層なので学習のコストがかからないが、超高次元にマップするにはそれなりのエネルギーがかかるわけで、どちらを重視するかということになる。一般的には、学習というのは非常に重い計算であり、その学習を1層でできるというのは有利な点である。ただ、今のリザーバーコンピューティングのフレームワークでは、

学習のところがすごく単純化されており、その学習方式自体も今どんどん変わってきている。1層じゃなくて多層にするとか、リザーバー自体も多層にしてしまうという考え方がある。今のリザーバーコンピューティングの世界の研究は何が正解なのか分からないが、色々な組み合わせを試しているようである。つまり、低電力にするためにリザーバーでやろうというようなことではなく、リザーバー計算をするとどういうところでメリットが出てくるのかということを手探りで探しているような状態かなと思っている。

ただ、学会ではリザーバーコンピューティングはホットな時期を過ぎているような気もする。バズワードのようになっているところもあり、そこは注意が必要である。私としては、リザーバーコンピューティングの将来が実はあまり見えていない。

生の脳が勝手に学習するという方法を考え出すのが大変である。脳の本当の学習メカニズムを理解しないと無理である。これからのコンピューティングを考える上で、脳は非常に重要なお手本である。

徳田：我々は、2030年とか2050年の未来社会からバックキャストしながらどういうコミュニケーション環境になっていくのかということを議論している。コンピューティングと違い、デファクトだけでは無理がある。すなわち、どの周波数帯が地上系、HAPS系、宇宙系で使えるかということは国際コンセンサスなので、ワールド・ラジオ・コングレスにおいてどの周波数帯を使うかということを各国と協調しながら議論している。一方で、自分たちの国の強い力を出そうということも考えている。日本は主にテラヘルツのコミュニケーション、ノンテレストリアルネットワーク（NTN）、つまりHAPSなど、そして、もう一つが時空標準である。時空標準というのは、証券会社のタイムスタンプだけのビジネスではなく、我々のピコ秒レベルの精度を持ったタイムカードがMeta社でやっているオープンコンピューティングプロジェクトで採用された。データセンターの中のファイルの同期のワーストケースの遅延をバウンドしたいということで、PTPのプロトコルよりももう少し精度の高いクロックを使いたいということで、実証実験をやっている。それは、日本のデータセンターでもやっている。

また、西海岸の研究所においては、量子コンピューティングのチームで光デバイスの研究もしている。先ほど話が出たデバイスサイエンスとコンピューターサイエンスの融合は非常に重要だと思う。

Device Science × Computer Science

高島：では、次にデバイスサイエンスとコンピューターサイエンスのところのお話をお願いしたい。

竹内：これだけ生成AIが盛り上がってくると、デバイスメーカーがAIの専門家と話をしたいのでつないでほしいという希望が多い。たとえば本学だと社会連携講座というのがあり、そういうところにAIの専門家を呼ぶと結構盛り上がる。たとえばデバイス自身に内在する揺らぎを使うと、かえって推論の効率が上がるということがある（Variation Aware Training）。こういうことに機械学習の先生が興味を持つことがある。こういうことが他にもあると思う。

高島：そういうものをもっと広げるにはどうすればいいか？

竹内：ファンディングが一番シンプルである。あるいは、本村先生にCRESTのときにどうやってチームを作ったのかという話を伺えると一番参考になる。

本村：アーキテクチャー側の研究者がアルゴリズムに興味を持ち、新しいアルゴリズムをアーキテクチャーに落とし込むことに興味を持つこと。逆にアルゴリズム側研究者が新しいアーキテクチャーがあればもっとアルゴリズムは変わるのではないかという興味を持つこと。そういう視点からのチャレンジや相性などがあって初めて成立したような気がする。このようなことがより広がるようにするには、ファンディングというのは確かに一つ大きな手ではあると思う。

本村：別の話題ではあるが、付け加えたい。先ほどからの話で少し明確にした方がいいと思うことがある。

DS（Device Science）とCS（Computer Science）という話があったが、少し解像度を上げると、まずデバイスのレイヤーがあり、その上に回路、その上にアーキテクチャー、さらにその上にアルゴリズムのレイヤーがある。少なくともこの4階層はある。これでもまだ粗いかもしれない。しかし、今の議論で言うと、その4つは少なくとも定義しなければならない。井上先生の指摘は、今の4レイヤーのうちの下のデバイスと上から2番目のアーキテクチャーが連携していくのが重要だということであると理解している。そうすると、その間に回路のレイヤーが入っているので、そこも自然にその中に入ってくる。

竹内先生がおっしゃった学会の融合というのは、一番下のデバイスとその1つ上の回路のレイヤーの融合が起こっているという話であり、その上のアーキテクチャーとはまだ線があるように思える。

私が言った融合というのは、一番下のデバイスは置いておくというスタンスで、そこは井上先生と違う。スタンスの違いだけであって、どちらがいいという話ではない。私としては、回路とアーキテクチャー、アルゴリズムという上の3階層の融合が新しい情報処理を考える上で重要だと思っている。ということで、DSとCSという言葉が何を指しているのかよく分からないので、4階層に分けて話をしたほうがいいと思う。

いずれにしても、その分野のレイヤー間の融合を進めていくというのは、全体に対して有意義、あるいは必要だと思うので、そこに関しては賛成であるが、言葉の定義をしたほうがいいと思ったので補足した。

井上：本村先生がおっしゃることは全くそのとおりだと思う。ただ、レイヤーを昔ながらの切り方で定義するのは、違うような気がする。たとえば量子が非常に分かりやすいが、量子はデバイスをやっている人とアーキテクチャーをやっている人もかなり密にやっている。その切り方に、量子ならではの回路のつくり方があるし、ソフトウェアのつくり方もある。したがって、そもそもどういうレイヤーが必要なのかというところから、考えなければならない。そういう意味では、かなり先の話をしている。たとえばCMOSや今のトランジスタをベースにするということであれば、先人たちが築いた良い階層構造があるので、それを利用するということになる。しかし、その階層構造が本当に通じるのかどうかというところから考えなければならない。したがって、階層構造を定義する場合に、デバイスが何かということで、その構造が変わってくると思う。

ファンディング

井上：もう一点述べると、自分で連携が必要だと言っておきながら逆のことを言うが、連携というのはそう簡単ではない。自分も超電導や光コンピューティングの方などとさまざまなプロジェクトをやっているが、理解し合えるのに5年はかかる。そこから初めて問題は何かというのが見えてきて、そこから探り始めてやるので結構時間はかかる。

そのときに、ファンディングがこの連携をドライブする。自分たちの時はファンディングがなかったので、知的好奇心や将来良くなるという気持ち、ほかの人はやっていない、ということでやっていたが、ファンディングがあるとそういうことをドライブし始めると思う。

JSTに呼ばれたときによく言うことであるが、本当は、3年後にこういう戦略目標を立てるのでその間に皆さんきちんと準備してください、という形で公募して始めるというのがあるべき姿だと思っている。2月ぐらいに分野連携、融合をやるぞ、と言って5月に出てくる提案というのは、もちろんきちんとしたものもあるとは思うが、それに目がけてつくった連携になりがちである。だから、本当にどうすべきかと言うことを根本的に考え直さなければいけない。

先日、University of Southern Californiaの教授から、いわゆるDARPAのモデルとはちょっと違うが、かなりの決定権を持ってDARPAがファンディングをしているという話を聞いた。こうあるべきだというビジョンを持ち、かなりプッシュするらしい。そういうことも必要なのかもしれない。我々

も協力するが、JSTであるべき姿を描いて、その上で議論をしないと、今回の話も空中戦になってしまい、もやもやしたことの議論で終わってしまい、次はどうするかと言うようことが決まらなくなってしまう。

木村：本村先生の話も井上先生の話ももっともだと思っており、こういうことをやるとすると、ファンディング側としての考えをきちんと議論をした上で、2～3年の助走期間を設けて研究者のある程度の合意形成もしながらやっていくと言うのが必要かと思う。今は、年度の初めぐらいに目標をつくり、募集となる。ツボを押さえたものをやっていないのではないかと思う。

また、井上先生もおっしゃっていたが、そういう検討に対してもファンディングは要ると思われる。プロジェクトをやるための資金をつけると言うだけではなく、考えたり悩んだりすること自身にも資金を提供する必要がありそうだ。一見無駄なように見えるけど、実は無駄ではないと思う。

井上：まさにそのとおりである。たとえば2年ぐらいの探索期間があり、本提案になったときに大きなプロジェクトにして、そこから5年というようなセットにすることが、ある種リーズナブルな一つの実装かなと思う。

気を付けなければいけないのは、最初の2年間やってうまくいかなかったら駄目だというような言い方をしてはいけないということである。探索をすること自体に意味があり、それがまたいつか生きてくる可能性があるわけである。たとえば、新規デバイスの研究をやっていて一番難しいのは、結果が出るまでに長い時間がかかるので本当にこれに賭けていいのか、ということでずっと不安があるということである。もし、プロジェクトは失敗してはいけないという立て付けになると、誰もそんなことをやらなくなる。したがって、探索して駄目だったら駄目で仕方がないので、そこで得たものを成果にしてもらえればよいというようにしなければならない。もし、その探索でいいものが見つかったらそこから本格的にプロジェクト化するというやり方とか雰囲気の作り方が必要ではないだろうか。

中村：まず、分野融合するには結構時間がかかるというのは、井上先生のおっしゃるとおりだ。3年は短か過ぎるし、5年、あるいはもっと長いほうがいいかもしれない。

事例としては、スパコンにはフィージビリティスタディという期間がある。フィージビリティスタディはやった方が良いが、そこで各チームの採否を判定するのではなく、チームはシャッフルしてより良いチームを構成する。つまり、もう少しダイナミズムをチーム構成に持たせるということである。今のやり方だと、有望そうな人と組んでチームを作って応募するとなると、日本で3チームぐらいしかいいのが出てこないとか、若い人が育たないということになる。もう少し何かうまいスキームを作っていただく必要がある。

自分の考えでは、JSTはプロジェクトの成果を論文などだけでみていてはいけない。それも見る必要はあるが、その後、5年後、10年後にそのチームの学生や所属した人がどれだけ育ったかということで、その領域の成否を決めるべきだと思っている。長い目でも評価するべきである。それにはファンディング側の人も責任を持たなければいけないと思っている。

ファンディングの計画を立てるところにも大学の人が入っていって、その人は業績だけじゃなくて、そのファンディングがうまくいったかどうかでもその人を評価するとか、そういうふうにしないといけない。アメリカだとそういう評価も入っているように思える。このように言うと、大学の評価の仕方がよくないのではないかというように、大学にも批判が戻ってくるのかもしれない。

木村：日本の場合、皆さん真面目だからきちんとやろうとし過ぎて、白黒をつけ過ぎている感じもする。研究なんて何が成功するか分からないような世界でやっているので、そこであまり堅くやっても駄目で、もう少し余裕を持つべきだと思う。

コンパイラ、OSなどのシステム基礎技術の重要性

井上：コンピューティングに話を戻し、言っておきたいことがある。入江先生の話にもあったが、日本の中での懸念と思うのは、コンパイラー技術やOS、あるいは半導体でいうとEDAの設計技術であったりという、目立たないがシステムをつくるときに重要な技術の領域である。いわゆる汎用プロセッサーのマイクロアーキテクチャーもそうだと思うが、そういうところがあって初めて新しいデバイスができたらソフトウェアをコンパイルする新しい技術を作ったり、新しいデバイスができたときにそれ用のEDA作ろうというようになる。そこがないと基礎研究で終わってしまい、竹内先生の指摘のように、ビジネスが市場にまでつながるところまで行かない。

NVIDIAがすごかったのは、CUDAを持ったことだと思う。あの領域でソフトウェアのエコシステムを作ったというところがすごい。そこを見ずに、GPUを作って成功している、とだけ思うと危険だ。そういう、日は当たらないけど大事なところにもファンディングするというのが大事ではないかと思っている。

木村：大賛成だ。コンパイラーというのは、日の目を見ない。最適化をやって性能が5%上がりましたといっても相手にもされない。しかし、それは基幹の部分だから定常的に継続し、一定レベルの人材も育成しておかないといけない。急にやれと言われてもできない。一度やめてしまうと、そこで途絶えてしまう。

余談ではあるが、CUDAの話が出たのであえて言う。2007年か2008年にスパコンを始めたときに、いろいろなベンチャーとか企業がアクセラレーターの話をしにきた。「チップがこんな速い」という売り込みがあり、「どうやってソフトを書くの」と聞くと、「アセンブラで書いてください」と言う。「それから後はどうなるの」と聞くと、「全部書き直しです」と言う。しかし、NVIDIAだけは「CUDAというツールがあるので、コンパルすれば使える」と言う。この会社はちょっと違うなと感じた。それが今や日本の国家予算を超えるぐらいの会社の価値を持っている。そういうところのセンスが違う、そこが地力なのかなと感じる。まさに井上先生がおっしゃったように、NVIDIAは自分たちはソフトの会社であると言う。そういうところのセンスというのも含めて、日本にはそういう人材がなかなかいないと思う。コアの部分をやる人に加えて、そういうのを見通して何かをやるという、全体を設計できる人があまりいないという気がする。そこをやっていかないと全体としての日本の研究力というか、国力というか、そういうのが上がっていかないのではないかと思う。

井上：私が先ほど申し上げたCS、DSというのは、そういったコンパイラーなども含んだ話である。

入江：ちょうど基礎的研究の人材の話になったので話をしたい。国内のクライシスの話として見ると、もちろん研究のいろいろなリソースが足りないが、研究のマインドやマネジメントのリソースと受け皿が足りないと感じる。

たとえば国際レベルで通用するような研究をしたり、成果を出している人がいたり、あるいは、点だけれども、突き抜けた成果をだしている。しかし、総論として国内におけるアーキテクチャーの研究は遅れていると言われる。そうすると、お金の前に人が必要であるが、人は寄り付いてこないし、人が来るためにはレピュテーションが必要ということになる。そこがうまく行けばいいと思っている。実際、その遅れている研究で育った人材は、結局、国内には受け皿がなく、海外のトップ企業に行ってしまう。あっさりとライバルを強化し続けてしまう。したがって、危機感をあおって研究資金を調達するというのもあるが、国内のアーキテクチャー研究にはこんな種があってこんなに進んでいる、と主張しないと若者が入ってこない。それこそ「死馬すら且（か）つ之を買ふ」ではないけれど、どんなものでも褒めて盛り立てて勢いをよくしたいと思っている。

井上先生がおっしゃった「予算のためにチーミングする」ということに同感する。第一段階はチーミングなしで気軽に参加でき、学際に貢献するという集まりがあり、それで2年間で本申請のための

チーミングを考えながら交流する。そこからドロップアウトしてもそれは交流の結果が得られたということなので、それでよい、というような組み立てが良いと思う。

竹内：さきほどのコンパイラーもそうであるが、異分野連携を実現するにはプラットフォームが必要である。NEDOの資金で新しいAIアクセラレーターを機械学習のアルゴリズムを組み込むためのプラットフォームとして作っているが、実はそこが主戦場になっている。たとえば、Georgia Techが国プロで作っているNeuroSim、IMECのZigZag、IBMのAIHWKIT、CEA-LetiのArchSimなど、異分野が連携するためのコンパイラーというか、マッピングツール、あるいはプログラミングモデルと言うのかもしれないが、この一見地味なところに力を入れるべきだと思う。

我々はクローズドでやっているが、オープンでやっているものもある。井上先生のCADのお話にも近いが、こういったミドルウェアやプログラミングモデル、コンパイラーなどに対しても支援が必要である。これ自体ですぐに何らかの成果というわけではないが、連携のためのプラットフォームとして必要である。

最後に

栗原：この4月に着任した文部科学省の計算科学推進室長の栗原です。

先週、報道もされたように、ポスト富岳の検討が進んでおり、そのためのフィージビリティスタディには、理研、神戸大学、ＰＦＮやなどもいらっしゃって、量子計算機も含めて、また運用技術面の検討なども進めており、具体的なアーキテクチャーに関する議論をしていく段階にもある。

ポスト富岳に関しては遅くとも2030年運転開始ということで、今すぐということではないが、ＰＦＮのMN-Core、またその次などMIMDアーキテクチャーでの継続的な議論もされているし、もっと進んで今日も議論があったような非ノイマン型のシストリックアレイやデータストリーム型のプロセッサ、竹内先生からもお話のあったNeuroSimのようなアナログ回路をとりいれたもの、光や量子などが、ポスト富岳には間に合わないかもしれないが、ポストポスト富岳なども見据えた検討の対象になっている。計算のエネルギー効率の話とか、インターディシプリナリーの話であるとか、今日のお話は実にごもっともで、政府としても検討しなければいけないと思う。

私は2017年まで情報参事官におり、「ソサエティー5.0を支える革新コンピューティング」という戦略目標を担当し、文科省内の説明プロセスもやった。これに続くようなものを是非やりたいと考えている。

こちらの部署に着任したばかりであるが、先月ハンブルクのISCに行ってきた。NVIDIAのGrace Hopper、またAMDもGPUとCPUが層になった密結合のアーキテクチャーMI300Aなどを出していた。富士通やNECの展示もあったが、なかなか日本としての存在感を示すということには大きな課題がある。

文科省はご存じのとおり戦略目標を作ってJSTとAMEDに通知をして、そしてJSTとAMEDでファンディングをしていただくということであるが、私はこの分野の担当室長になったということで、文科省としても盛り上げて、一緒に二人三脚でやっていきたいと思っているのでよろしくお願いしたい。

ポスト富岳では実際に物をつくらなければならないが、今日議論があったようなファンディングの仕組みも、文科省として考えなければならない。 CREST・さきがけの仕組みもそうであるが、おそらく先生方がおっしゃっていたのはもう少し広い概念であり、科研費も含めて文科省としての全体のファンディングの問題であると捉えた。

私がいた時にはちょうどACT-Iとか、ACT-Xといった、若い人や新しい分野への対応を議論していた。創発事業もこの間にできたものであるが、研究を進めていく上での多様性とか余裕はとても重要であり、文科省も本格的に取り組まなければいけない。

システムソフトウエアとかミドルウエア、コンパイラー、エコシステムの議論もあったが、文科省の次世代計算基盤報告書最終取りまとめでもその議論、システムソフトウエアの改善と、その社会実装の話が出た。また、NVIDIAのCUDAの話があったが、特定の一社が市場を占有することは日本政府として好ましくない。並列プロセッサーをさらに超えたようなアーキテクチャーも今後出てくるかもしれないが、そこにおいてもそういったミドルウエアやコンパイラー、ソフトウエアエコシステム自体を一緒に育成していくというところが、今の日本の産業基盤でどこまでできるかというのは問題があるものの、皆さまのご協力も得ながら取り組んでいかなければいけない点である。先生方のご指導、CRDSの協力もよろしくお願いする。

井上：せっかく皆さんに立ち上げていただいた革新コンピューティングであるが、発表資料のタイトルにも書いたように、止めてはいけない。間髪入れてはいけない。3年、5年空くとその間にいい人たちがいなくなる可能性がある。継続が重要である。

中村：私も連続性が大事だと思う。先ほど栗原室長から次のスパコンは2030年という話であったが、次は線で考えていかないといけない。点としてアドバルーンをあげるのではなく、サポートは連続的にやっていくべきなのではないかと思う。とにかく連続的にサポートしていくっていうことが大事である。

徳田：絶滅危惧種と呼ばれており、連続的なサポートが必要である。

木村：最初はどうなるかなと思っていたが、いい議論ができたと思う。栗原室長のコメントを頂き、勇気百倍で、きちんと継続的にやっていくような形にしていきたいと思っている。今後とも皆さま方のご協力をよろしくお願いする。

今日はいろいろコメントいただいき、ありがとうございました。これを糧に次のステップに行きたいと思う。

付録　ワークショップ開催概要

日時：2024年6月12日（水）　13:00～16:00
場所：オンライン

（1）開会挨拶・開催趣旨説明　13:00～13:15（15分）
　開会挨拶　　徳田 英幸（NICT）
　開催趣旨説明　木村 康則（JST CRDS）
（2）ポジショントーク　　13:15～14:55（100分：発表10分、質疑5分×6名）
　浅井 哲也　北海道大学大学院情報科学研究院 教授
　井上 弘士　九州大学大学院システム情報科学府 情報知能工学専攻 教授
　入江 英嗣　東京大学大学院情報理工学系研究科電子情報学専攻 教授
　竹内 健　　東京大学大学院工学系研究科電気系工学専攻電子知能情報学講座 教授
　中村 宏　　東京大学大学院情報理工学系研究科システム情報学専攻認識行動情報学講座 教授
　本村 真人　東京工業大学 科学技術創成研究院 教授
（3）休　　憩　　14:55～15:05（10分）
（4）総合討論　　15:05～15:55（50分）　　司会: 高島 洋典（JST CRDS）
（5）閉会挨拶　　15:55～16:00（5分）　　木村 康則（JST CRDS）

参加者：
　関係部門に限定したクローズドな開催とし、本ワークショップを企画・運営するCRDSメンバーと登壇者のほかに24名の参加があった。
　その内訳は以下の通り。
　内閣府：2名、文部科学省：5名、産業技術総合研究所：3名、情報通信研究機構：2名、情報セキュリティ大学院大学：1名、個人：1名、科学技術振興機構（JST）戦略研究推進部：5名、科学技術振興機構（JST）研究開発戦略センター：5名

ワークショップ企画・報告書編纂メンバー

総括責任者　木村 康則
リーダー　高島 洋典
メンバー　的場 正憲
青木 孝
福井 章人
平池 龍一

俯瞰ワークショップ報告書　**CRDS-FY2024-WR-04**

コンピューティングアーキテクチャー

令和 6 年 8 月　August 2024
ISBN 978-4-88890-931-0

国立研究開発法人科学技術振興機構　研究開発戦略センター
Center for Research and Development Strategy, Japan Science and Technology Agency

〒102-0076 東京都千代田区五番町 7 K's 五番町
電話　03-5214-7481
E-mail crds@jst.go.jp
https://www.jst.go.jp/crds/

ISBN 978-4-88890-931-0

FOR THE FUTURE OF
SCIENCE AND
SOCIETY

CRDS

https://www.jst.go.jp/crds/