2024 年度成果報告書

グリーンイノベーション基金事業／次世代デジタルインフラの構築／次世代グリーンデータセンター技術開発に関する調査

2025 年 3 月

国立研究開発法人新エネルギー・産業技術総合開発機構

委託先　株式会社日経ビーピー

目次

まえがき

国立研究開発法人新エネルギー・産業技術総合開発機構（以下、NEDO）は、「エネルギー・地球環境問題の解決」や「産業技術力の強化」実現に向けた技術開発の推進を通じて、経済産業行政の一翼を担う、国立研究開発法人である。

本調査は、NEDOが進める「グリーンイノベーション基金事業／次世代デジタルインフラの構築／次世代グリーンデータセンター技術開発」プロジェクトの研究開発成果を着実に社会実装につなげるために、データセンターを取り巻く環境に関する情報を収集し整理するとともに、効果的なプロジェクト管理に資する分析・考察を行うことを目的として実施する。具体的には、2023年度に行った「データセンターのアプリケーション及び当該アプリケーションの要求性能に係る動向」に係る質的観点（技術面と用途面）および量的観点からの変化についての調査結果（仕様書別紙）を一歩進めて、それらアプリケーションの変化が2028年時点のデータセンター全体に及ぼす影響と変化について情報収集・分析・考察するとともに、その影響と変化を更に深掘りし、データセンター構成要素への要求事項（変化）の予測も行う。

1. 研究開発の成果と達成状況

1.1 要約

（1）和文要約

2028 年度末時点において、データセンターサーバを構成する要素には、どのような事項が求められるのかを明らかにするため、2023 年度に行った調査結果を一歩進め、まずはアプリケーションの変化がデータセンター全体に及ぼす影響と変化について情報収集・分析・考察を行った。データセンターに影響を及ぼすと思われる項目は大きく三つに分けた〔①データ利用（処理）の変化②データセンター・アーキテクチャの変化③データセンターの在り方の変化〕。そのうえで、データセンターサーバを構成する主要な要素として、CPU、アクセラレータ、ストレージ、ネットワークデバイスの四つを取り上げ、それぞれに対して要求される事項について情報収集・分析・考察を行った（④）。

それぞれの技術の進捗が見通せる識者に対してヒアリングを行うとともに、日経 BP が発信するメディア等を介した情報を元に、日経 BP 総合研究所の研究員が分析・考察を行った。

また、データセンターの業務経験者、およびデータセンターのユーザーにアンケートを実施。上記、データセンターに影響を及ぼすと思われる項目についてコメントをもらうとともに、データセンターサーバを構成する主要な 4 要素への要求事項についてもコメントをもらった。

①のデータ利用（処理）の変化における情報収集・分析・考察では、具体的には「エッジ／クラウドの分担進展」「量子コンピューティング技術」がデータセンターにどのような影響を与えるのかを明らかにした。エッジ／クラウドの分担進展に関しては、データセンターの地方分散を政府も後押ししていることから、間違いなく技術進歩が進む。一方で、量子コンピューティング技術に関しては期待が高いものの、2028 年までの広がりは大きく期待できない。

②のデータセンター・アーキテクチャの変化における情報収集・分析・考察では、具体的には「データセンター・ネットワーク構造の進化」「リソース共有機構の導入」がデータセンターにどのような影響を与えるのかを明らかにした。ネットワークには高速、低電力化が求められ、IOWN をはじめとする光回線への期待が大きい。また高速なネットワークを利用して大容量メモリなどの共有が一部で進んでいく。

③のデータセンターの在り方の変化における情報収集・分析・考察では、具体的には「レジリエンス要求」「カーボンフットプリント要求」がデータセンターにどのような影響を与えるのかを明らかにした。レジリエンスでは、特に経済安全保障の観点から高度なデータ管理技術が求められるようになりそうだ。またカーボンフットプリントの観点からは、再生エネルギーの活用がデータセンターにおいてもカギを握る。

④のデータセンターサーバを構成する主要な要素に対して要求される事項についての情報収集・分析・考察においては、CPU の機能の先鋭化や AI（人工知能）向けアクセラレータの開発に勢いが加速することなどを明らかになった。

上記の傾向はデータセンターの業務経験者、およびデータセンターのユーザーに実施したアンケートからも確認できる。

(2) 英文要約

In order to clarify what items will be required for the elements that make up a data center server as of the end of fiscal year 2028, we took a step further from the survey results conducted in fiscal year 2023 and first collected, analyzed, and considered information on the impact and changes that changes in applications will have on the entire data center. Items that are expected to affect data centers were broadly divided into three categories: (1) changes in data usage (processing), (2) changes in data center architecture, and (3) changes in the nature of data centers. Based on this, we took up four main elements that make up a data center server: CPU, accelerators, storage, and network devices, and collected, analyzed, and considered information on the requirements for each of them(4).

We interviewed experts who can foresee the progress of each technology, and researchers at Nikkei BP Research Institute analyzed and considered the information provided by Nikkei BP through media and other sources.

We also conducted a survey of people with experience in data center work and data center users. We asked for comments on the items that are expected to affect data centers as mentioned above, as well as comments on the requirements for the four main elements that make up a data center server.

In the information gathering, analysis, and consideration of (1) changes in data usage (processing), we specifically clarified how "advancing edge/cloud division of roles" and "quantum computing technology" will affect data centers. The government is also supporting the decentralization of data centers, so the division of roles between edge and cloud will undoubtedly progress. On the other hand, although there are high expectations for quantum computing technology, we cannot expect it to spread significantly until 2028.

In the information gathering, analysis, and consideration of (2) changes in data center architecture, we specifically clarified how "evolution of data center network structure" and "introduction of resource sharing mechanisms" will affect data centers. Networks will require high speed and low power consumption, and there are high expectations for optical lines such as IOWN. In addition, the sharing of large-capacity memory and other devices will progress in some areas using high-speed networks.

In the information gathering, analysis, and consideration of (3) changes in the nature of data centers, we specifically clarified how "requirements for resilience" and "requirements for carbon footprints" will affect data centers. With regard to resilience, advanced data management technology is likely to be required, especially from the perspective of economic security. Also, from the perspective of carbon footprint, the use of renewable energy will be key in data centers as well.

In (4), the information gathering, analysis, and consideration of the requirements for the main elements that make up data center servers revealed that there will be accelerating momentum in the advancement of CPU functions and the development of accelerators for AI.

The above trends can also be confirmed from surveys conducted with people with experience in data center work and data center users.

1.2 調査の実施内容と結果の概要

I. データ利用（処理）の変化がデータセンター全体に及ぼす影響と変化についての情報収集・分析・考察

調査目的

エッジ／クラウドの分担進展など、データ利用（処理）の変化が 2028 年時点でのデータセンターにどのような影響や変化を及ぼすのか、情報収集・分析・考察を行う。

調査方法

2028 年時点の当該技術の進捗が見通せる識者に対してヒアリングを行うとともに、日経 BP が発信するメディア等を介した情報を元に、日経 BP 総合研究所の研究員が分析・考察を行う。

実施期間

2024 年 12 月 2 日～2025 年 3 月 31 日

結果

①エッジ／クラウドの分担進展による影響

まず、以下で論じるエッジに対する理解は専門家・業務経験者の間で多少幅があるが、「ユーザー側およびデータ発生源側に設置するコンピュータ資源」という意味でほぼ合致している。

特に識者の間で「ニーズが高く、今後伸張する」との声が強かったのは、次の四つである。

一つ目は、コロケーションサービスを提供しているデータセンター施設への、エッジ的な位置づけとしてのサーバの設置である。例えば、グローバルでサービスを提供するハイパースケーラーの間では、日本国内でいわゆる「自前」のデータセンター施設の建設を進める動きが見られている。今後、こうした「コアとなるクラウドデータセンター」の整備と並行して、自社サービスへのアクセス品質を高める取り組み〔コローケーション先にサーバを設置し、CDN（コンテンツ・デリバリー・ネットワーク）的な機能を持たせる〕も積極的に進めると見込まれる。

これに応じて、自社データセンター施設の主要展開地域（IX へのアクセスなどを考慮したデータセンターの主要設置区域）において、国内データセンター事業者が提供するコロケーションサービスをさらに求めるようになると考えられる。

二つ目は、日本国内の事業者が、セキュリティやプライバシーを含めた「データ主権」の考え方に沿って自社データを国内に保持するためのデータセンター施設だ。ソブリンクラウドもこれに該当する。

三つ目は、生成 AI のために設置する AI データセンターである。

四つ目は、主に災害対策として、東京圏および大阪圏以外の地域に設置するデータセンター施設へのニーズだ。現状この 2 圏にデータセンターが集中しているが、日本の政策としてデジタルインフラ強靱化のために、データセンターを地方へ分散させる意向が示されている。その核を成すプランの一つが、データセンター中核拠点として東京・大阪のほかに北海道や福岡を位置づけ、光ファイバーおよび海底通信ケーブルを追加的に整備して接続を強化するというものだ。今後 3 年、民間側の旺盛なニーズと国家プランの両方に支援される形で、分散化の流れはさらに強まる可能性がある。電力供給の分散要求や再エネ活用の方向性も、これを後押しする。

こうしたデータセンターの設置を加速させる要因は、バックボーンネットワークだ。有識者らの話を総合すると、現状の利用形態においては、バックボーンネットワークの遅延は問題として顕在化してはいないようだ。しかし、今後データセンター数が増加し、さらには立地も分散化することを考慮すると、センターの役割を果たすデータセンターとエッジデータセンター、あるいはエッジデータセンター同士における低遅延性の確保が欠かせなくなる。この点、NTT グループの次世代ネットワーク構想「IOWN（アイオン、Innovative Optical and Wireless Network)」の主要要素である APN（オールフォトニクス・ネットワーク）への期待は高いとの声があった。従来のデータセンター間バックボーンよりもさらに低遅延性が確保できる可能性があるためだ。

エッジコンピューティングの理想型として、クラウドサーバ、エッジサーバ、エッジ端末の 3 者を連携させてレスポンスを高めるシステム構成が語られることが多い。このシステム構成については、低遅延性の確保という面からその有効性が取り上げられている。想定されるアプリケーション分野としては、スマート工場や自動運転支援など。実際、このシステム構成に基づいた実証実験プロジェクトが各所で実施されている。しかし、2028 年までというタイムスパンにおいては、実用サービスが続々と生まれるという姿までには至らないと予想される。

理由はいくつかあるが、主なものの一つは、事業としての経済性や合理性の観点から見た場合に懸念が残るため。識者からは次のような意見が挙げられた。「エッジ側でリアルタイム性が求められるケースは、エッジ端末側に必要な機能を集約させ、必要な時にだけサーバ（クラウド）側にアクセスする形態として組んだほうが合理的だ」「エッジサーバなどのインフラ側への投資分を回収できるビジネスの構図が描けなければ、投資に踏み込みにくい」。

ただし、実用サービスが生まれていない理由として、エッジコンピューティングのシステム構成の有効性を目に見える形で示した実例や、それを支える基盤サービスが世に流通していないことだとする意見もある。その観点から、実用化を見据えた技術面での準備も進められている。これが進めば、諸条件に対する解像度が高まり、実用サービスの開発が促進される可能性がある。

準備の一例が、産総研とソフトバンクが取り組む「CloudEdge Fusion」プロジェクトだ。コアのデータセンター、エッジ側のデータセンター、5G 回線でつながるエッジ端末の 3 要素を組み合わせ、「時間確定性」の高さが求められるアプリに必要なシステム基盤要素の研究開発を進めている。題材とするユースケースは自動運転支援と物流最適化。これにより、低遅延・高信頼性が求められるエッジコンピューティング・アプリの社会実装に向けたマイルストーンを構築する。事業終了年度は 2027 年度の予定。

また、ソフトバンクが推進している「AI-RAN」は、モバイル通信会社の本業としての合理性があるため、順当に発展していく可能性があると評価する声もあった。AI-RAN は無線アクセス基地局に GPU（Graphics Processing Unit）搭載サーバ（＝エッジサーバ）を設置し、通信網の効率化をはじめとした AI の処理ができるようにする構想のこと。

②量子コンピューティング技術による影響

量子力学を応用した量子コンピュータは、画期的な処理能力の向上とエネルギー効率を実現しうるポテンシャルを持つ。そのため、AI をはじめとする様々な用途での応用に多くの期待が集まり、量子コンピュータの研究開発が世界で精力的に進められている。量子コンピュータを実用化

するうえで特に重要な量子誤り訂正技術についても、この 1 年で新しい研究開発成果が複数発表された。この技術領域をリードしているとみなされているグーグルが 2024 年 12 月に発表した量子チップ「Willow」はその一例である。これらの研究開発成果により、量子コンピュータの実用化時期が当初の予想より大幅に早くなると考えられている。

しかし、2028 年までのレンジで見ると、量子コンピュータが実用的な商用サービス・製品として登場する見込みは非常に小さく、データセンターに与える影響はほとんどないとみられる。既に量子コンピュータを利用するクラウドサービスは国内にもいくつか存在するが、それらは主に量子コンピュータのアプリケーションを探索・研究開発するためのサービスであると認識している。商用クラウドサービスに必要な機能が全くなく、データセンター事業者から見ると「昔の研究所にあった大型コンピュータを、ただシェアしているだけという状況」に映る。

量子コンピュータが商用サービス・製品として実用化されるには、少なくとも量子誤り訂正技術が確立され、処理結果に対する信頼性が担保される必要がある。この技術が研究開発の途上であるため、量子コンピュータに対する大手 IT 企業の動きは鈍い。

2028 年に商用化するテクノロジーであれば、現時点で IT 企業はその設計をしている段階にあるはずだという。しかし、現在の量子コンピュータには、ハンドリングの難しさ、データの信ぴょう性をすべて確認しなくてはいけない煩雑さがあり、今のアプリケーション資産がそのまま使えるどうか分からない点も大きな課題となっている。大手サーバベンダーの立場では、一般的な事業領域で、従来のサーバを提供するように量子コンピュータを提供できるようになるとは現時点で想像できない、というのが実情のようだ。また、マイクロソフトは 2016～2017 年ごろから、すでに量子コンピューティングにあまり力を入れなくなっているという。今は生成 AI に全力を注いでいる。

このようにハードウェアとソフトウェア／サービスの両面で大手 IT 企業の開発があまり進捗していないため、実用的かつ量産可能な量子コンピューティングの製品・サービスが 2028 年までにデータセンターに届く可能性は非常に低いと考えられる。

今回、量子コンピュータについてヒアリングした識者約 10 人のうち 1 人は、用途を AI に限定すれば 2028 年までに量子コンピュータが実用化されると回答した。生成 AI 用途では必ずしも正確な結果が要求されないため、まだ誤り訂正が不十分な量子コンピュータであっても使えるだろうという。

一方、現状の量子コンピュータに生成 AI は向いていないと見る向きもある。生成 AI のソフトウェアレベルの誤回答にハードウェアレベルの誤処理が加わり、とんちんかんな回答ばかりしていたら意味がないからだ。また、生成 AI で扱うデータ量が膨大すぎて、量子コンピュータに処理させるにはまだ遠いという。

量子コンピュータの用途として最も注目されているのは、化学分野（マテリアルインフォマティクス）だという。従来の GPU には難しい化学計算を効率的にできるようになるとしたら、量子コンピュータの重要なアプリケーション分野になると考えられている。

③その他のデータ利用（処理）の変化による影響

データ利用（処置）という観点で、近年、データセンターにもっとも影響を与えたのは AI、と

りわけ生成 AI の登場だ。対話型 AI「Chat（チャット）GPT」を米オープン AI が公開したのが 2022 年 11 月末。2 年以上が経過した今でも、生成 AI は世界を巻き込みながら進化を続けている。現在、データセンターが建設ラッシュを迎えようとしているのも、生成 AI の処理を行うニーズの拡大に対応するためだ。

そして、生成 AI の活用が次のフェーズを迎えようとしている。あらかじめ特定の仕事を完了することを目的にデザインされている「AI エージェント」の広がりだ。生成 AI は、自分がほしい結果を得るためには、人間が考えて細かい指示を出す必要があった。一方で AI エージェントは、人間が細かい指示を与えなくても、目標を達成するために AI が自律的に考えてタスクの完了までを実行する。例えば、出張を支援する AI エージェントであれば、希望する行先と予算などの情報を与えれば、AI エージェントが Web 検索から適切な移動手段を提案するだけでなく、予約サイトに接続してチケットの手配や宿泊予約までも行ってくれる。AI エージェントは、人間から依頼された任務を実行する際、自分で判断する過程で生成 AI を利用する。

マイクロソフトや NTT データなど IT 大手各社は 2024 年末から、相次いで企業向けにサービスの提供を開始。インドの調査会社マーケッツ・アンド・マーケッツによれば、世界の AI エージェントの市場規模は、2024 年の 51 億米ドルから 2030 年には 471 憶米ドルと、6 年間で 9 倍を超える成長を予測している。

2025 年は AI エージェント元年ともいわれ、多くの企業で AI エージェントの活用が始まる。当初は用途を特化させた AI エージェントからはじまり、その後はより複雑で高度なタスクをこなせる AI エージェントが登場するだろう。

そして、そう遠くない未来には、AI エージェント同士がやり取りする世界が訪れる。取引先との交渉を AI エージェントが行うようになり、また取引先の交渉相手も AI エージェントになれば、AI エージェント同士が商談を成立させる。不合理な要素は排除されることになり、商談のスピードは圧倒的に早くなるだろう。

その場合、AI エージェントに必要とされるのは、より迅速な判断であり、処理をする部分は距離が近い方が望ましい。エッジコンピューティングの応用として最適な分野の一つだ。AI エージェントの広まりは、データセンターの地方分散を促す可能性がある。

II. データセンター・アーキテクチャの変化がデータセンター全体に及ぼす影響と変化についての情報収集・分析・考察

調査目的

データセンター・ネットワーク構造の進化など、データセンター・アーキテクチャの変化が 2028 年時点でのデータセンターにどのような影響や変化を及ぼすのか、情報収集・分析・考察を行う。

調査方法

2028 年時点の当該技術の進捗が見通せる識者に対してヒアリングを行うとともに、日経 BP が発信するメディア等を介した情報を元に、日経 BP 総合研究所の研究員が分析・考察を行う。

実施期間

2024 年 12 月 2 日～2025 年 3 月 31 日

調査結果

①データセンター・ネットワーク構造の進化による影響

データセンター・ネットワークを論じるうえで、ここではネットワークを二つに分ける。一つはデータセンター間のネットワーク、もう一つはデータセンター内のネットワークだ。

データセンター間ネットワークの技術としては、NTT の次世代ネットワーク構想である IOWN の一環である APN の動向が注目される。IOWN では、現状のネットワークと比較して「電力効率 100 倍」「伝送容量 125 倍」「遅延時間 200 分の 1」を目指している。

現状、データセンター間接続には光ファイバー通信を使用するケースが多いが、途中で複数のルーターを経由する。ルーターでは、光信号を電気信号に変換し、処理が済んだら再度光信号に戻しており、揺らぎや遅延が生じやすい。これに対して、APN を利用すると、光のままデータを送ることができる。専用の光伝送装置が必要になるが、すでに NEC や富士通などが開発・提供を進めている。2024 年 12 月 1 日から NTT 東西が「All-Photonics Connect powered by IOWN」として、最大 800Gbps の専用線サービスを始めた。

NTT は、分散化したデータセンター間を光ファイバーで直接接続する「データセンターエクスチェンジ（DCX）サービス」の開発にも注力している。現在の DCI（データセンターインターコネクト）はデータセンター同士が 1 対 1 で接続し、その接続関係は固定的である。しかも一般的に 50km 圏内でなければ遅延が大きくなって、一体的な運用が難しくなる。これに対し、DCX は多対多でデータセンターを接続する。キャリアコントローラとスイッチなどを利用することで、任意のデータセンターとの接続を可能にする。データセンター間は APN で接続する。キャリアコントローラは、各データセンターの接続を制御するソフトウェアで、データセンター間の接続に合わせて指示を実行する。データセンター間には光スイッチなどで構成する DCX 網があり、そこで接続関係を調整する。

DCX を適用したデータセンター同士であれば、距離が 100km 以下で往復遅延 1ms 以下に抑えられるという。DCX が実用化されれば、遠距離に配置されたデータセンター群を、あたかも 1 つのデータセンターのように利用できるようになる。NTT は 2027 年度前後に顧客提供を目指すとしている。現状でもすでに、東京と大阪の近郊では、データセンターを運用するために必要な電力や土

地を確保するのが難しくなっている。2028 年には、都市部に集中しているデータセンターの分散化が進む可能性がある。

APN の課題の一つは料金が高いことである。ただし、初期のインターネットのように、ユーザーが広がれば低価格化していくと見られ、IOWN の普及時期である 2030 年くらいには安価にできる可能性がある。

もう一つの課題は標準化である。いくら技術的に優れていても、ハイパースケーラーはグローバル標準ではないと採用しない。そこで、NTT は IOWN 構想を実現するために、「IOWN Global Forum」を 2020 年 1 月に発足した。日本だけではなく世界の企業と一緒に技術を開発し、ビジネス展開につなげ、グローバルなエコシステムを構築することを目指す。NTT、インテル、ソニーの 3 社で設立し、約 160 の企業や研究機関、大学が参画している。

2024 年 10 月に台湾で開催された第 7 回 IOWN Global Forum 総会では、IOWN を使ったユースケースとして、金融システムにおけるデータセンターの分散化、放送コンテンツ制作のリモート化、AI データセンターの分散利用という三つの事例について 2025 年から 26 年にかけて実現できるメドが立ったと、ユースケース WG の議長が明らかにした。2028 年にはほかにも様々なユースケースが出てくると見られる。

一方、データセンター内のネットワークの動きとして注目される一つは、「ウルトライーサネット」である。AI 処理の増加によってこれまで以上に高速・大容量のネットワークが求められるようになった。

AI 用途では、GPU サーバのクラスター構成をとるケースが多い。大量のデータを高速でやり取りする必要があり、ネットワークの遅延はシステム全体の処理性能に直結する。このため、ネットワークとしては RDMA（リモート・ダイレクト・メモリ・アクセス）をベースとする「InfiniBand」を使うことが多い。RDMA は OS を介さずに GPU と NIC（Network Interface Card）が直接、データをやり取りする仕組みである。IP プロトコルでデータを送受信する場合、通常はアプリケーションのプロセスが OS カーネルを呼び出し、OS カーネルがソケットを作ったり、メモリにデータをコピーしたりといった処理が必要になる。データの転送に逐一 OS を介するために、性能のボトルネックになる。RDMA を使えればこの問題を解消できる。ただし、InfiniBand は専用のアダプターやスイッチなどを使い、ネットワークを別途形成する必要があるためコストや手間がかかる。

これに対し、ウルトライーサネットは、イーサネット及び IP ベースのネットワーキング技術を基盤としており、既存のプロトコルと共存できる点が特徴である。ウルトライーサネットは、2023 年に設立された「Ultra Ethernet Consortium（UEC）」で規格検討が進められている。ウルトライーサネットでも、OS を介さずに GPU と NIC がデータをやり取りする RDMA を有効活用する方向で議論されている。イーサネットでパケットロス・低遅延といった要件に応えるには、GPU メモリ間での直接転送を実現する RDMA と、輻輳通知・制御を行う DCQCN（Data Center Quantized Congestion Notification）を駆使することになる。だが、RDMA も DCQCN も AI 基盤を想定して作られたものではなく、UEC ではそれを解決するための新機能が検討されている。

UEC には、AMD やシスコシステムズ、マイクロソフトなど約 100 社が参加している。エヌビディアの独占状態にある InfiniBand に対抗するイーサネット規格を作るのが UEC の狙いと言われている。AMD やブロードコムが UEC のサポートをすでに発表している。「現状、AI 用途ではエヌビデ

ィアフルスタックを受け入れなければならないが、ウルトライーサネットによって制約から解放されるのではないか」（大手ネットワークインテグレータ）といった見方がある。2028 年に向けては実用化が進むと見られる。

もう一つ注目すべき動きは、IOWN 構想で NTT が研究開発を進めている「光電融合デバイス（PEC）」である。光電融合デバイスとは、電子回路と光回路を一つのシステムに統合したもの。NTT の開発ロードマップでは、データセンター間の通信向けデバイスを開発（PEC-1）、コンピュータ内のボード間通信（PEC-2）、半導体パッケージ間通信（PEC-3）、半導体パッケージ内のダイ間通信（PEC-4）と発展させていく計画である。PEC-2 は 2025 年から、PEC-3 は 2029 年以降となっている。

PEC-1 の一部は既に製品化されている。APN で使われ、伝送速度の向上と低電力化に寄与している。PEC-2、PEC-3 になると、光電融合型スイッチによってディスアグリゲーテッドコンピューティングが可能になる。現状の AI/HPC では、CPU やメモリ、GPU がサーバに搭載され、演算処理する。これに対しディスアグリゲーテッドコンピューティングでは、CPU やメモリなどリソースをプールし、光で内部接続を行うことで箱の単位を超えて最適化できる。既に光通信ネットワークでは多くの光電融合デバイスが利用されており、今後の進展が期待されているが、量産化に向けての課題は多く、2028 年での実用化は難しいという見方が強い。実際、「光電融合には期待しているものの、早くても 2030 年以降と見ており、むしろウルトライーサネットのほうが気になっている」（大手インターネットプロバイダー）といった声が聞かれた。

②リソース共有機構の導入による影響

データセンターのリソースを効率的に共有する技術としては、すでに仮想化やコンテナ化が存在している。仮想化においては、一つのサーバ上に複数の仮想サーバを構築し、またコンテナ化ではアプリケーションの実行環境を構築する「コンテナ」を複数持つことで、サーバが持つ CPU やメモリ、ストレージなどのリソースを共有する。

仮想化やコンテナ化では、一つのサーバ上で複数の仮想サーバ、あるいは複数のコンテナを稼働させることを基本とする。一方で、サーバ単位に閉じず、データセンターのすべてのリソースをワークロードによって最適に配分し、論理サーバを構築しようという考えが、上記の「データセンター・ネットワーク構造の進化による影響」においても述べたディスアグリゲーテッドコンピューティングだ。わかりやすく説明すれば、A サーバから CPU を、B サーバからアクセラレータを、C サーバからメモリを使用して論理サーバを構築、その論理サーバ上で顧客のワークロードを実行するといったイメージだ。

ディスアグリゲーテッドコンピューティングの構想は、決して最近のものではない。というのも、リソースを本当に効率よく利用したければ、サーバに閉じる必要がないからだ。これまでの一般的なデータセンターでは、同じ仕様のサーバを複数台ラックに並べて、それぞれのサーバを仮想化。その上でアプリケーションを稼働させていた。ただし、あるサーバでは CPU やメモリを多く使用するアプリケーションが稼働している一方で、別のあるサーバではアクセラレータが得意とする計算処理を必要とするアプリケーションが稼働したりなど、構成要素の使われ方が不均一な状態が存在することが課題としてあった。その解決策としてディスアグリゲーテッドコンピューティングが有効であることは認識されていた。

そして、このディスアグリゲーテッドコンピューティングの必要性が急速に議論されている。理由の一つがムーアの法則の疲弊である。各種アプリケーションが必要とするインフラへの要求は年々上がっている。ムーアの法則が健在だった頃は、高まるインフラの要求に、ハードウェアの向上によって対応できていた。サーバを交換すれば、各サーバの利用のされ方が不均一でもアプリケーションはなんの不都合もなく稼働できるという状況だった。また、性能を向上させた各種のデバイスも比較的安く大量に生産されていたため、課題が顕在化しにくかった。これが、ムーアの法則が疲弊し、ハードウェアがアプリケーションの要求を容易に上回ることが難しくなり、ディスアグリゲーテッドコンピューティングの必然性が浮上し始めた。

もう一つの理由が、急速に拡大する AI 処理への対応だ。アプリケーションが必要とするインフラへの要求は年々上がっていると上述したが、AI 処理が必要とする要求は甚大なものだ。最近では「AI データセンター」といった言葉が新聞誌面などでも賑わすように、AI 処理に特化させたデータセンターの構築が進められているほどである。そのような AI 処理を効率よく行うためには、少しのリソースも余らせることなく有効活用したい。

ディスアグリゲーテッドコンピューティングが現実のものとなれば、現在のハードウェアの構成ががらりと変わる可能性がある。これまでのサーバは、CPU やアクセラレータ、メモリ、ストレージなど、基本的にコンピュータを構成する要素が一つのブレードに搭載され、複数台のブレードサーバがラックに収納されているスタイルが一般的である。一方、ディスアグリゲーテッドコンピューティングが可能になれば、CPU 置き場、アクセラレータ置き場など個々のデバイスを個別に集中して設置しておけばいいことになる。

こうすることで、これまでハードウェアの能力が足りなくなった場合はサーバの増設が必要だったのと比較して、CPU だけを増強するといったデバイスごとでの対応が可能となる。また、デバイスごとに置き場を変えられるので、例えば多大な冷却を必要とするアクセラレータは集中的に冷やす、ストレージは振動をシビアに制御するといった、個々のデバイスごとの設置環境対応が可能になる。

非常に理にかなった考え方であるディスアグリゲーテッドコンピューティングが、2028 年度末までに普及するかどうかのカギを握るのが、標準化の進展だ。前述したように、ディスアグリゲーテッドコンピューティングでは必要となる個別のデバイスだけを利用して、論理サーバを構成すればよいが、そのためにはこれらのデバイスが接続できるような業界標準のネットワークを規定することが望ましい。また、デバイス間の距離が離れることを想定すると、距離が離れてもシビアなレイテンシーに応えられる必要がある。

これまで、サーバ内のデバイスを接続する標準規格としては PCI Express がある。2002 年に策定された PCI Express は、その後、パフォーマンスを改善するために幾度も改定がなされ、2022 年にリリースされた PCI Express 6.0 では 64G 転送／秒のデータ転送速度を可能とし、さらに 2025 年にリリースが予定されている PCI Express 7.0 では 128G 転送／秒のデータ転送速度になる予定だ。

一方で、PCI Express はディスアグリゲーテッドコンピューティングに必要とされる、例えばネットワークを仮想的に複数の「スライス」に分割して、仮想的なネットワークとして運用するための仕様などを備えていない。そこで PCI Express の物理層で動作し、ディスアグリゲーテッ

ドコンピューティングを可能とする相互接続プロトコルが「Compute Express Link」（CXL）だ。

CXL はメモリや CPU、周辺デバイスを接続するための規格。このようにサーバの機能を強化するためのネットワークをスケールアップネットワークと呼ぶが、スケールアップネットワークの標準化を狙ったものには、ほかにもスケールアップネットワークの標準化を狙ったものには「UALink（Ultra Accelerator Link）」などもある。ただし、ディスアグリゲーテッドコンピューティングを見据えた標準仕様としては CXL が先行している。

CXL は、2019 年 3 月に CXL1.0 がリリースされ、その後、大きく仕様を拡張している。昨年 12 月には最新の CXL3.2 をリリースしており、メモリのディスアグリゲーションを可能とする仕様などが拡張されている。仕様の充実度合いは、1.0 と比較にならないぐらいだ。

この CXL をサポートするデバイスメーカーも、登場し始めている。インテルが 2025 年 2 月に発表した、データセンター向け CPU「Xeon6 プロセッサ」では CXL2.0 に対応。また、CXL にいち早く対応してきたサムスン電子は、2023 年 5 月には CXL2.0 に対応する 128G バイトの DRAM モジュールを開発。韓国 SK ハイニックスや米マイクロン・テクノロジーも CXL2.0 対応の DRAM モジュールを開発済み。一方でキオクシアは、NAND 型フラッシュメモリによる CXL 対応を進めると表明している。前述の IOWN 構想においても、CXL の仕様を考慮に入れている。

このように CXL 対応のメモリ製品の市場は拡大している。メモリ容量を拡張できるという CXL の特徴が、AI 用途を狙ったデータセンター向けに適していることがその要因だが、マイクロンは CXL メモリの市場が 2025 年に 20 億米ドル（約 3000 億円）、2030 年には 200 億米ドル（約 3 兆円）規模に達すると予測している。より仕様が充実した CXL3.2 をサポートしたデバイス群の市場投入は 2～3 年後になるかもしれないが、その頃には、メモリに関しては多数の CPU や GPU の負荷に応じて構成を動的に変えられるディスアグリゲーテッドコンピューティングが実用化となっている可能性もある。

こういった標準化の議論は進む一方で、「現在は若干停滞している感がある」という声も聞かれる。現在は AI バブルともいう状況で、AI の学習などに最適なハードウェアを早く揃えたいという点に関係者が注力しているということがその要因だ。それぞれのベンダーがとにかく製品を納めるようと自分たちがやりやすいように進めるため、標準化に意識は向いていないという意見だ。AI バブルが長引くほど、標準化が進みにくくなる可能性はある。

また、ここまではスケールアップネットワークの議論を説明してきたが、真のディスアグリゲーテッドコンピューティングを実現するためには、サーバ同士の接続、すなわちスケールアウトネットワークの議論もしなくてはならない。ただし、2028 年に向けてはスケールアップネットワークの整理は進むが、サーバをまたいでのリソースの共有は、そこまで逼迫したニーズもないことから、現実化しないと想像できる。スケールアウトネットワークの議論が本格化するのはそれ以降であろう。

③その他のデータセンター・アーキテクチャの変化による影響

今回の調査において、その他のデータセンター・アーキテクチャの変化に関して、最も多くの意見が聞かれたのは冷却方法の変更についてだった。CPU やアクセラレータの進化は目覚ましいものがある一方で、その発熱量も甚大なものになっている。発熱を吸収するためには、冷却方法

を現在の主流である空冷から水冷に置き換えることが望ましい。現在、建設を予定されている新しいデータセンターの冷却方法は、多くが水冷を採用している。2028 年に向けては、既存のデータセンターに関しても、水冷方式への置き換えが進んでいくと予測できる。

それ以外に、データセンター・アーキテクチャとして特徴だった技術を挙げる声はなかったが、データセンター・アーキテクチャ全体が用途に応じて専用化する流れが生じるのではないかという指摘を複数人の識者からうけた。データセンターでは、ユーザーの要望に応じて、様々な処理を行っている。これらの処理に対して、これまでは比較的均一なハードをそろえたデータセンターで対応してきた。今後は、特色を持ったデータセンターを複数揃え、その中からユーザーに合ったものを組み合わせて提供していく流れが加速するという。

分かりやすい例が、多くの建設ラッシュが見込まれている AI データセンター。多くの AI 処理をこなせるように、他のデータセンターと比較して GPU を多く備えるなどの専用化を行っている。またソブリンクラウドもその一つ。ソブリンクラウドを可能とするデータセンターには、暗号化、アクセス制御、監視などの高度なセキュリティ対策が必要とされる。

こういった高度なデータセンターの専用化だけではない。用途によっては、これまでのデータセンターがハイスペックということもあった。例えば、ほとんどのデータセンターは電源の冗長化構成が取られているが、電源が落ちても大きな問題がない用途もある。こういった用途のユーザーは、冗長化構成がなくても、安価にサービスを受けたいと考えるであろう。また例えば再エネ 100%で稼働するデータセンターを構築すれば、環境意識の高いユーザーが好んで使用するかもしれない。

GAFAM をはじめとするハイパースケーラーは、グローバルに画一的なメニューを提供している。その一方で、彼らとの差別化を図る意味で、価格や環境配慮に特徴を持たせることは、データセンター事業者の戦略として十分にあり得ることだ。

III. データセンターの在り方の変化がデータセンター全体に及ぼす影響と変化についての情報収集・分析・考察

調査目的

レジリエンス要求など、データセンターの在り方の変化が 2028 年時点でのデータセンターにどのような影響や変化を及ぼすのか、情報収集・分析・考察を行う。

調査方法

2028 年時点の当該技術の進捗が見通せる識者に対してヒアリングを行うとともに、日経 BP が発信するメディア等を介した情報を元に、日経 BP 総合研究所の研究員が分析・考察を行う。

実施期間

2024 年 12 月 2 日～2025 年 3 月 31 日

調査結果

①レジリエンス要求の影響

一般にレジリエンス（Resilience）とは、「回復力」や「耐久力」という意味だが、産業分野においては、「外部からの脅威や障害に直面しても、迅速に適応し、機能を維持または回復する能力」のことを指す。データセンターにおけるレジリエンスとは、サーバ、電源、通信などデータセンターの機能を支えるシステムやインフラが様々なリスクに直面しても、安定した運用を維持し、障害が生じても迅速に復旧する能力のことを意味する。

データセンターにまつわるリスクは、自然災害、サイバー攻撃、電力供給の障害、ネットワーク障害、人的エラーなど多岐にわたる。こうしたリスクへの対策を強化し、レジリエンスを高める取り組みは、従来からなされているが、「デジタル化」のトレンドを背景に、あらゆる分野で大規模データやクラウドの利活用が拡大するとともに、その要求は年々厳しさを増している。2025 年以降も、その傾向は続く。

特に、2022 年暮れに米オープン AI 社が、大規模言語モデル（LLM:Large Language Model）を公開したのを契機に、生成 AI を導入する動きが、様々な分野で急速に広がっており、これとともにデータセンターを利用して大規模なデータを扱うユーザーが、今後数年間の間に急増する。これによってデータセンターのレジリエンスに対する要求が一段と高度化する。

その影響は、主にデータセンターの運用管理の技術に及ぶ。その一つが、運営管理の自動化である。これまでレジリエンスを高めるアプローチとして、データセンターのシステムを、複数の階層に分けて運用する「多層化」の手法が導入された。さらに、システムを多重化あるいは冗長化して、故障が発生したときに代替システムに即座に切り替える手法なども、広く導入されている。ただし、これらの手法の場合、現状では人手により作業が欠かせないので、ヒューマンエラーによる事故が発生する可能性がある。これを排除するために、運用管理を自動化して人手をかけないようにする仕組みの導入が広がる。ここに AI を利用する動きも進む。「すでに、米国の大手 IT 企業など先行する企業は、AI を利用してデータセンターの運用管理を自動化することでレジリエンスの向上を図る手法を取り入れている。この動きは着実に広がり、2028 年ころには、一般的になる」（マイクロソフトの勤務経験者）という声も聞かれた。

主にデータセンターの運用管理にまつわるレジリエンスを左右する今後の大きな問題として浮

上しているのが、経済安全保障上のリスクである。特定の国や地域の法律・規制に従い、その国や地域内でデータを保存・処理するといった、データのローカライズを強化するための新たな仕組みの導入が進む。同時により高度なセキュリティが求められる。この結果、データセンターの管理運用が複雑化する。

近年、「プラットフォーマー」と呼ばれる米国の大手 IT 企業に、世界中のデータが集中していることを問題視する機運がグローバルな規模で高まり、これを背景に、データに関する規制を強化する動きが世界各地で進んでいる。例えば、2016 年に EU（欧州連合）が、EU を含む欧州経済領域（EEA）域内で取得した個人データを EEA 域外に移転することを原則禁止する「EU 一般データ保護規則（General Data Protection Regulation：GDPR)」を 2016 年 5 月に発効した。さらに、2022 年以降、「データガバナンス法（DGA：Data Governance Act)」「デジタルサービス法（DSA：Digital Services Act)」「デジタル市場法（DMA：Digital Market Act)」「データ法（DA：Data Act)」など次々とデータに関する法整備を進めている。中国も、国家安全保障の強化を目的に、2017 年に「中国サイバーセキュリティ法」が制定しており、これ以降データや個人情報、サイバーセキュリティ関連の法令やガイドラインなどが次々と制定・改正されている。2021 年 9 月には「中国データセキュリティ法」、同年 11 月には「中国個人情報保護法」が施行されている。これらのようなデータに関する規制を強化する動きは、さらに広がる可能性がある。

さらにデータの保護だけでなく、ある国や地域で生成・収集・保存されるデータに対して、その国・地域が主権を持ち、その管理や利用を自国の法律や規制に基づいて統制する権利、いわゆる「データ主権」を保有していることを前提に、企業・組織の枠を超えて安全にデータを共有するクラウド上の仕組み、「データスペース」を構築する動きも欧州を中心に進んでいる。具体的には、非営利組織 IDSA（International Data Spaces Association）が、ドイツを拠点にデータ共有のための標準やルール、アーキテクチャを策定する活動を 2016 年から展開。2019 年 10 月にドイツ連邦経済・エネルギー省（BMWi）が発表した欧州を網羅する連邦型データ・インフラストラクチャ「GAIA-X」の構築を目指す非営利団体 Gaia-X European Association for Data and Cloud AISBL(GAIA-X)も 2020 年から活動を開始した。ドイツの製造業革新プロジェクト「Industrie4.0」の推進団体である Plattform Industrie 4.0（PI4.0）の傘下でも「Catena-X」や「Factory-X」など多数のデータ連携基盤構築プロジェクトが進んでいる。IDSA によると、GAIA-X が策定し採用しているデータ連携基盤構築プロジェクトの数だけで、2024 年 3 月時点で 145 にも及ぶ。

こうした国・地域によってデータに関する規制・法律を制定したり、新しい仕組を構築したりする動きは拡大しつつある。そうなると、データセンターの運用管理は、ますます複雑化する。

例えば、国・地域ごとに異なる規制・法律に対応することによって、国や地域を超えてデータを移動することで問題が生じる。「少し先の話になるかもしれないが、データが保管されている場所を把握できる仕組みや、データの動きをトレースする仕組みが求められるようになる」（ハイパースケーラーでキャパシティプランニングを担当していた勤続経験者）という意見もあった。データ主権の強化に伴い、特定の国や地域でのデータ保管を求められると、それに応じてインフラの配置や運用の変更を強いられることにもなる。データの安全性やプライバシー保護に関する新たな技術的要件や規制順守の負担も増える。扱うデータの重要性によって運営管理のレベルが変わることから、求められるレジリエンスのレベルも多様化する。「国のデータと民間のデータでは、

レジリエンスに関する要求の度合いは異なる。多様化するレジリエンス要求に対して、どう対応するかはデータセンターの大きな課題の一つになる」（江崎浩氏　東京大学大学院 情報理工学系研究科 教授）。

②カーボンフットプリント要求の影響

現在、企業のカーボンフットプリント対策としては主に、LCA（ライフサイクルアセスメント）を通じた CO2 削減対策の工程単位での把握、自社設備やプロダクトを含めたエネルギー利用効率の向上、原材料の調達からプロダクトの廃棄まで含めた CO2 削減対策の実施、CO2 排出量を相殺するカーボンオフセットの利用などが進められている。データセンターについてもこれらに沿った対策の必要性が指摘された。データセンターでは特に「PUE（電力使用効率）」指標が重視されており、理論上最も効率が高い 1.0 を目指した対策がいっそう進む。

特に顕著に進むのは再エネ活用だ。日本を含めた世界各地のデータセンター事業者の間で再エネ利用が活発化する。リーディングカンパニーの 1 社として米アマゾンが挙げられる。同社は 2030 年までに自社の事業を 100%再エネでまかなうと表明している。今同社が各地で建設を進めている自社データセンターは全て、再エネ利用の準備がなされていると考えてよいだろう。同社は 2025 年 1 月 29 日、日本において新規に 4 件の大規模太陽光発電所への投資を発表した。

Scope3（バリューチェーン全体における CO2 排出量）対応や ESG（環境・社会・ガバナンス）投資の拡大を踏まえて、近い将来は顧客企業側がデータセンター事業者に再エネ証書を積極的に求めるタイミングが来ると予測される。併せてデータセンター事業者が顧客の代理で証書を調達するサービスも各所で行われるようになるだろう。

データセンターが備える冷却システムの重要度がさらに高まる。AI 活用の進展により CPU および GPU の発熱に対していかに冷却性能を高めるかが、CO2 削減のカギを握るようになるためだ。

冷却水を送って熱を除去する水冷技術が注目を集めている。また、一部の先端ユーザーは電子機器を専用の液に浸ける液浸冷却に注目している。液浸は現状、初期コストの高さと運用ノウハウがないという問題があるが、これらの点がこなれてくれば冷却関連の電気代が安く抑えられると見込まれている。まずは既存のデータセンター設備に導入しやすい水冷システムから始まり、2028 年には液浸が一部で導入が進められ、それほど珍しくないものになっていると予想される。

データセンターの統合管理も進む。2028 年頃までには、デジタルツイン技術が適用されるなど、AI 活用を含めたより洗練された管理システムが登場すると予想される。その先には、発電所を含めたデータセンターの電力供給、冷却、サーバの統合管理を可能にする取り組みも注目を集めるだろう。性質の異なる設備同士の統合管理の難易度は低くないが、これが進展すれば、消費電力の平準化と低減、さらには中長期的に見た設備のコスト削減が可能になる。

以下では、ヒアリングした識者から聞かれた、データセンターのカーボンフットプリント対策を考えるうえで重要と思われる視点二つを挙げる。(1) 消費電力の削減に向けたソフトウェア面でのインパクトの大きさ、(2) データセンターが備える他産業への CO2 削減効果の計測について。

(1) の電力消費量の削減には、ソフトウェア面での対策が世間で思われている以上に効くという声があった。2016 年の発表のため参考情報としてとどめるべきだが、米グーグルは、同社傘下のグーグル・ディープマインド（Google DeepMind）による AI をデータセンターの空調機の制御

に適用したことで、空調向けのエネルギー効率が約 4 割向上し、データセンター全体の PUE は 15%改善したとする。つまり、ソフトウェア側で工夫すれば消費電力量が大きく削減できる可能性がある。

AI では学習が要となるが、必ずしもパラメータ数が全てではないという指摘があった。日本語の LLM について見ると、大規模なパラメータのモデルよりも、チューンナップした小型のモデルのほうが良い成績を収める場合がある。現状の LLM はデータ量が増えると良い結果が出る傾向があるが、用途に適した方式が複数出てくる可能性がある。こうした取り組みが各所で進むと、データセンター全体が要求する消費電力の伸びが一定レベルで落ち着くシナリオも考えられる。

（2）については、データセンターが顧客および社会に提供しているアプリケーションの役割を評価し、CO2 排出量削減への貢献度について、何らかの形で可視化するべきではないか、という意見があった。

データセンターだけを見ると、増大する電力消費量に対して「いかに下げるか」の視点に限られてくる。だが実際には、データセンターが提供するアプリケーションによって、物流や人の移動が効率化されたり、製品の設計が最適化されたりすることで社会のムダが減っている。

ここから考えると「データセンターで動かしているアプリがどのようにカーボンフットプリント対策に貢献しているのか」という観点でデータセンターがもたらす CO2 削減効果を評価することが欠かせない。

従来の「グリーン IT」にも、「by IT」という「IT によっていかに効率化できたか」を見る観点がある。日本では 2007 年から経済産業省がプロジェクト「グリーン IT イニシアティブ」を立ち上げ取り組んできた。グリーン IT の by IT の考え方を用いてデータセンターを適切に評価することで、デジタル分野の健全な発展が促される可能性がある。

③その他のデータセンターの在り方の変化による影響

GAFAM と言われるハイパースケーラーは、自社のクラウドサービスを利用している世界中の顧客のニーズに対応するべく、世界で統一的な調達基準に沿って製品を調達している。世界中の膨大な技術企業のなかから選んでいるが、ハイパースケーラーでの業務経験者らに話を聞くと、日本企業が一事業者として売り込みに行って採用されるという姿は現実的ではないという。

そうした現状から考えると、「いかにハイパースケーラーが望む製品をつくるか」という観点で、マーケティングの視点も含めた技術の企画・開発アプローチが望まれる。識者からは、「いかに売れる技術をつくるかという姿勢で、研究者や技術者だけでなくマーケティング専門家など多様なメンバーによるチームを形成し、国としてもその活動を支援する取り組みが望まれる」との声が寄せられた。

ハイパースケーラーは急拡大しており、IT の主要なイノベーションはそこからしか生まれないという状況になりつつある。日本企業および産業界としては、そのような巨人であるハイパースケーラーの動向を見つつ、自国の発展のためにいかにそれらと「共存」するか、という観点が望まれそうだ。

IV. データセンター構成要素への要求事項（変化）予測の調査

調査目的

データセンターの構成要素となるCPUなどのデバイスについて、2028年時点での要求事項が現状からどのように変化しているかを、情報収集・分析・考察を行う。

調査方法

2028年時点の当該技術の進捗が見通せる識者に対してヒアリングを行うとともに、日経BPが発信するメディア等を介した情報を元に、日経BP 総合研究所の研究員が分析・考察を行う。

実施期間

2024年12月2日～2025年3月31日

調査結果

①CPUに対する要求事項

データセンターが設置、活用されるようになった当初から、情報処理の頭脳としてCPUが活用されていた。その処理対象は、アプリケーションに関わる情報処理、ストレージの管理・制御、ネットワーク処理、施設内のインフラ管理処理など多岐にわたっていた。そして、データセンター活用の需要が高まるにつれて、より高性能なCPUを搭載したサーバが大量に導入・運営されるようになった。

ただし近年、CPUを取り巻く環境が、ニーズ側とシーズ側の両面で大きく変わり、あらゆる処理をCPUで実行することはなくなってきている。特定処理に特化した、別のタイプのプロセッサで処理させた方が、高い演算性能が得られ、効率的に実行できるようになったからだ。

本調査では、まず、CPUの進化の方向性に大きな影響を及ぼす、データセンター向けサーバに搭載するCPUに対する需要側での要求と技術側で抱えているリスクの現況をまとめた。ここでまとめた、現状分析は、CPUのみならず、他の章で取り扱うAIアクセラレータ、メモリを含むストレージ、ネットワークデバイスにおいても同様に適用できる。

調査は、まず、データセンター向けサーバに搭載するCPUと、個人用パソコンに搭載するCPUそれぞれに求められる要件の違いを明確にして、調査・分析・洞察を実施する際のベースとした。そして、収集した情報を、現時点から2028年までの短期的な動きと、10年後の2035年の中長期的な動きに分別してまとめた。

●サーバ用CPUを取り巻く、需要面・技術面での環境変化

現在、データセンターに導入する情報処理システムのハードウェアでは、技術を継続的に進化させていくうえでの、技術革新で解決すべき深刻なジレンマを抱えている（図1）。従来の技術開発トレンドの延長線上での進化では、ハードウェアを供給する側で抱えているリスクを解消しながら、需要側からの要求に応えることができなくなってきていると言える。ジレンマを抱えながらの、従来実績の継続にこだわらず斬新なアイデアに基づく技術開発を進める動きは、CPUおよびAIアクセラレータ、ストレージなどあらゆるハードウェア領域において散見される。まず、CPUをはじめとするデータセンターのハードウェア開発で抱えているジレンマについて、技術の需要側と供給側それぞれの状況をまとめ、洞察した。

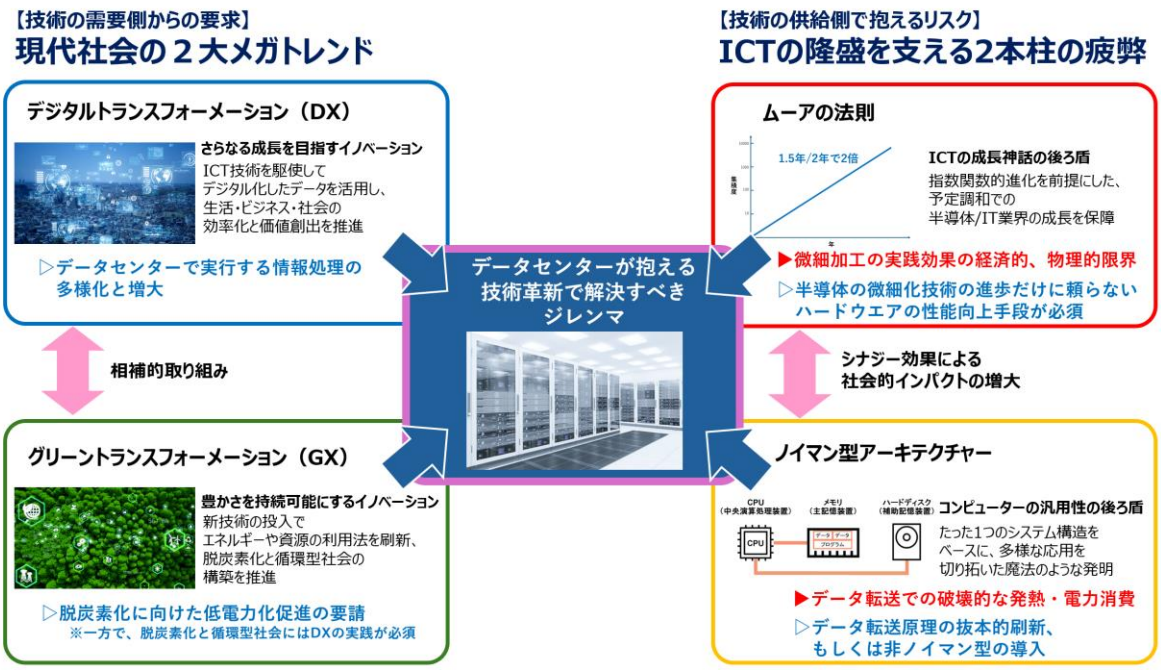


図１　データセンターに導入するハードウェアの技術開発で抱えているジレンマ
出所：日経 BP 総合研究所

●技術の需要側での環境変化

データセンターには、現代社会の改革に関わる二つのメガトレンドに対応した進化が求められている。そこで利用されている CPU をはじめとするハードウェアにも、これに準じた技術革新が必須になる。

メガトレンドの一つは、あらゆる業界・業種における業務改革である「デジタルトランスフォーメーション（DX）」である。既にデータセンターは、DX を推し進めるために、必要不可欠な社会インフラとなっている。過去にも企業の基幹業務や科学計算など多様な応用の情報処理を実行していた。これが DX が活発化した現在では、SNS やネット通販、動画配信など民生応用から、工業製品の開発や生産ラインや社会インフラ、産業プラントの管理・制御、さらには行政業務など、より多彩な応用に関わる処理を実行するようになった。近年ではさらに、生成 AI や深層学習、機械学習などの学習・推論処理の需要が増大している。

こうしたデータセンターが担う役割の拡大に伴って、そこで実行される情報処理は、量的に増大するだけでなく、質的に多様化してきている。ハードウェアには、情報処理すべきデータ量の増大に対応する性能向上と共に、多様化に対応可能な進化が求められている。

もう一つのメガトレンドは、2050 年のカーボンニュートラル達成を目指した社会の仕組みや行動様式の改革である「グリーントランスフォーメーション（GX）」である。GX を推し進めるためには、再生可能エネルギーの活用などと共に、世の中の人やモノ、事象の動きをキメ細かく管理・制御・管制することで効率化させる必要がある。その際にはデジタルデータによる管理が必須になる。大量の情報を集約し、AI など高度な情報処理技術を活用して包括的かつ一元的な処理を可

能にするデータセンターの活用が欠かせない。ところが、世界のデータセンターの総電力消費量、及びサーバ 1 台あたりの消費量が急増している。このためあらゆるハードウェアには、DX や GX の実践に貢献しながら、データセンター自体が GX の妨げにならない、低消費電力技術の創出・投入が求められるようになった。

●技術の供給側での環境変化

DX も GX も、データセンターの継続的進化をあてにして、取り組みを構想している面がある。データセンターの進化なくして、DX と GX の成就はありえないとも言える。ただし、データセンターでの情報処理能力を継続的に高めていくためには、現在、ICT 技術と半導体技術の進化において直面している二つの技術的リスクへの対応が必須になる。

技術的リスクの一つは、「ムーアの法則」が疲弊してきていることだ。これまで CPU をはじめとする半導体は、60 年間以上にわたって指数関数的な進化を遂げてきた。すなわち 1 チップに集積可能な素子数が 2 年ごとに 2 倍に増える、ムーアの法則が維持できていたからだ。チップに搭載可能な素子数の増加は、演算器の並列度増大や命令・データの管理・制御機能の高度化、メモリなど付随回路との密な連携を可能にする。これによって、CPU もまた、指数関数的高性能化を継続させることができた。ところが近年、このムーアの法則の終焉が、いよいよ目前に迫っていることを示す現象が見られるようになった。

ムーアの法則の終焉は、「原子より小さな素子は作れない」といった“物理的限界”を論点に語られることが多い。しかし、実際には、「微細パターンが描けない」「素子特性を左右する不純物濃度の精密制御が困難」といった“工学的限界”や、「ビジネスとして成立可能な投資額や歩留まりで量産できない」といった“経済的限界”の方がより早く訪れる。既に、モノリシックな大規模 CPU は、チップレットといった補助的技術の併用なく最先端製造技術を適用できなくなってきており、事業的限界や工学的限界が目前になってきたことの証左となっている。このため、素子の微細化に頼らない“ポストムーア型”ハードウェア技術の導入が必須になっている。

もう一つの技術的リスクとは、「ノイマン型アーキテクチャ」の欠点が顕在化して、その極めて有用な特徴を発揮できなくなりつつあることだ。データセンターのサーバを含む主要なコンピュータは、80 年間以上にわたって、たった一つのシステム構造、すなわちノイマン型アーキテクチャを採用し続けてきた。CPU とメモリをバスでつなぎ、演算処理するたびにプログラムに記された手順・対象の命令とデータを逐次読み出すことで多様なタスクをこなせる汎用性を実現してきた。ところが、転送速度の頭打ちと消費電力の増大といった、ノイマン型であるがゆえのコンピュータの性能向上と安定動作を妨げる現象が顕在化してきている。こうした現象が見られる領域は、より長いバスから短いバスへと年々拡大している。

通信技術の世界では、一層の高速・大容量化、高効率化を実現するために、高度な通信技術が投入される。そして、より高度な通信技術が、長距離転送の領域から短距離転送の領域へと適用範囲が逐次拡大していく傾向が見られる。ノイマン型アーキテクチャが抱えるリスクを解消するための技術開発においても同様の動きが進む。これまでの情報処理システムでは、バスを主に金属配線で形成し、電気信号で命令やデータを転送していた。ただし、そのままでは、CPU などプロセッサとメモリをつなぐボード上のバスを高性能化できなくなるノイマンボトルネックが顕在化

してきた。そこで、長距離通信で導入されている光転送技術が、ボード間転送においても導入される例が出てきた。早晩、CPU の入出力においても光転送技術が導入されることになる可能性が高い。さらなる高性能化を追求するため、チップ内への光転送の導入、もしくは頻繁なデータ転送を必要としない非ノイマン型アーキテクチャへの移行が進む可能性も出てきている。

●サーバ用と PC 用の CPU に求められる要件の違い

ここで、データセンターに導入するサーバ用 CPU と、家庭やオフィスなどで活用するパソコンに搭載されている CPU の差異を明確にしておきたい。

データセンター向けサーバ用の CPU では、不特定多数のユーザーのタスクを処理している。特定ユーザーが専ら利用するパソコン用とは、利用者像に大きな違いがある。ただし一般に、データセンターでは、ユーザーが異なっていたとしても、タスクの内容には、それほど大きな差異がない。確かに、データセンター業界全体でみれば、DX の進展によって、あらゆる業界・業種の多様な情報処理を担うようになったが、個々のデータセンター、あるいは個々のサーバの単位でみれば、処理しているタスクの内容が大きく変わることはない。企業の基幹システムを動かしているサーバが、動画配信サービスや AI の学習処理に利用されるようなことはほぼない。同じ CPU を、ネット閲覧やゲーム、オフィスソフトなど多様な用途に利用するパソコン向けとは、CPU に求められる機能・性能面での要件が異なる。

一般に、サーバ用 CPU には、多数の細かなタスクを同時処理するのに向く、マルチスレッド性能と並列処理に最適化した構造が採用される。このため、1 チップに 100 個以上の CPU コアを搭載したチップが多用されている。こうしたチップでは、チップとメモリをつなぐ入出力がノイマンボトルネックとなる可能性が高まるため、より多くのメモリチャネルを持ち、大容量メモリに対応可能な構成を備える。一方、パソコン用では、高周波数動作による汎用性の高い単一スレッド性能向上に最適化した構造が採用される傾向がある。

また、稼働時間も大きく異なる。サーバ用では 24 時間 365 日、稼働し続ける高い信頼性が求められる。しかも、不具合や故障が起きれば社会的・経済的に甚大な影響が及ぶ処理を実行している場合が多い。パソコン用は、1 日に 8 時間程度の稼働を想定している場合が多い。データセンター全体でも、仮想環境の導入などによって信頼性を高める仕組みが導入されている。チップレベルにおいても、サーバ用では高度なエラー検出・訂正機能や ECC (Error-Correcting Code) メモリへの対応、より堅牢な熱管理システムの導入など、CPU 自体に長期にわたる信頼性を保障できる機構の導入が求められる。

消費電力とコストに対する要求に関しても、サーバ用の方が、パソコン用よりもより厳しい。導入費用に直結するコストだけでなく、電力料金を含めた運営コストを最小化することが徹底される。さらには脱炭素化に対する社会的貢献の観点からも、要求が年々高まってきている。このため、サーバ用の低消費電力化を推し進めることが可能なアーキテクチャや機能・機構の導入が進められている。近年では、消費電力の増加による発熱量の増大が、CPU の安定動作を妨げる状況が顕著になってきているため、信頼性向上の観点からも低消費電力化技術の重要性が高まっている。

●今後、3 年間で想定される CPU 開発のポイント

直近、3 年間で想定される CPU 開発の動向のポイントをまとめた（図 2）。

図中、横方向には、情報処理システムの機能・性能を具現化・詳細化していく流れに沿って、具体的な技術開発の動きを「システム設計」「チップ設計」「チップ製造」「チップ実装」それぞれの段階に位置付けて示した。縦方向には、技術開発の動きを、適用対象となるシステム範囲を「チップスケール」「モジュールスケール」「ボードスケール」「ラックスケール」「データセンタースケール」の各領域に分類して示した。

図中のオレンジの枠内に示した動きが、直近 3 年間で想定される技術動向のポイントである。赤の矢印は、各技術動向間の依存関係を表している。一例を挙げると、「プロセス技術の継続的進化」は、「チップレットの適用」が前提となっていることを示している。一方、緑の枠内に示した動きは、10 年後までの長期的視野に立って、実用化が進められる動きである。緑の矢印で、同様に依存関係を示した。なお、直近 3 年間で想定される動きは、10 年間の長期的視野に立って見ても継続的に見られると予想される。

直近、3 年間で想定される CPU 開発の動向のポイントの中から、「CPU で実行する処理のオフロード化」、および「チップレットの適用」を起点とした、大きな技術革新の潮流について、詳細を示したい。

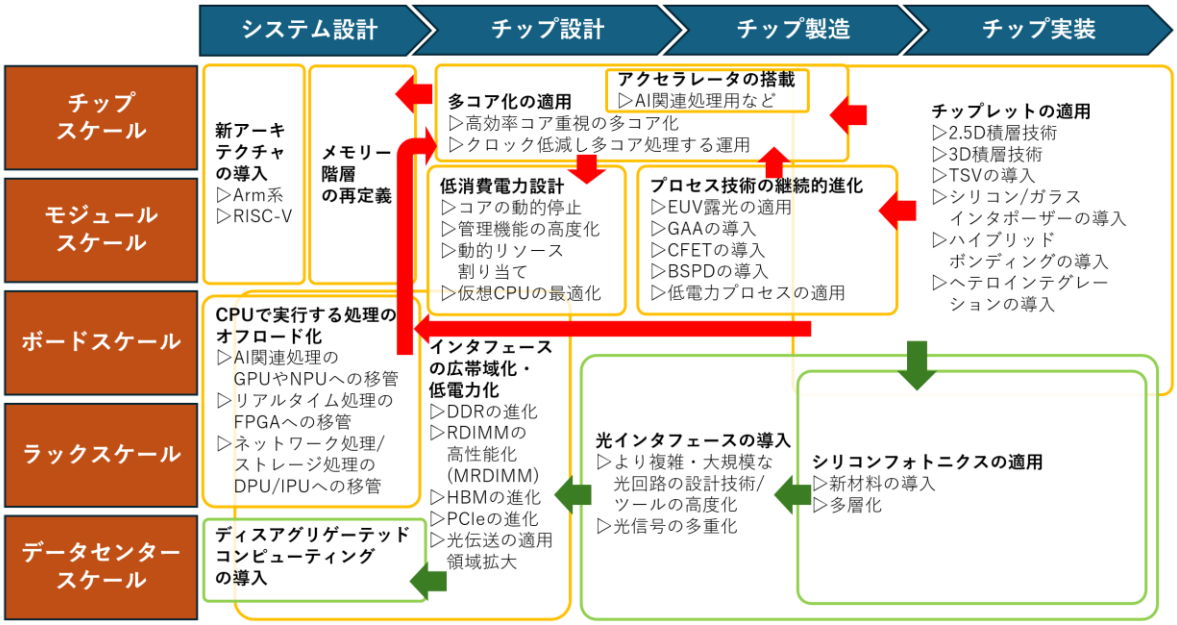


図 2　CPU 開発のポイント

出所：日経 BP 総合研究所

●CPU で実行する処理のオフロード化を起点にした技術革新

CPU は、極めて汎用性の高いプロセッサである。このため、かつてのデータセンターでは、システム内で処理すべき多様なタスクを、すべて同一の CPU で処理していた。その時点で、サーバ用とパソコン用では、性能差こそあれ、チップの機能と内部構造には大きな差異はなかったと言え

る。

ところが現在、かつては CPU で実行していたタスクの処理を、よりタスクの内容に適した機能・構造を持つプロセッサに移管する動きが顕著になってきた。最も代表的な例が、AI 関連処理を GPU に移管するというものだ。生成 AI の発展などに伴って、現在では、サーバ単位で CPU と GPU を使い分けるだけでなく、データセンターの施設自体を、一般クラウドサービス用と AI 用に分別する動きさえ出てきている。

そして、データセンター内では、GPU 以外にも、データセンターで処理すべきタスクの大規模化・多様化に伴って、さらに多彩なプログラマブルデバイスに CPU で実行していたタスクを移管していく動きがみられるようになった（図 3）。そして、CPU は、CPU でなければ効率的に実行できない、逐次処理や分岐が多い処理にフォーカスして適用し、より効率的な運営を可能にしている。

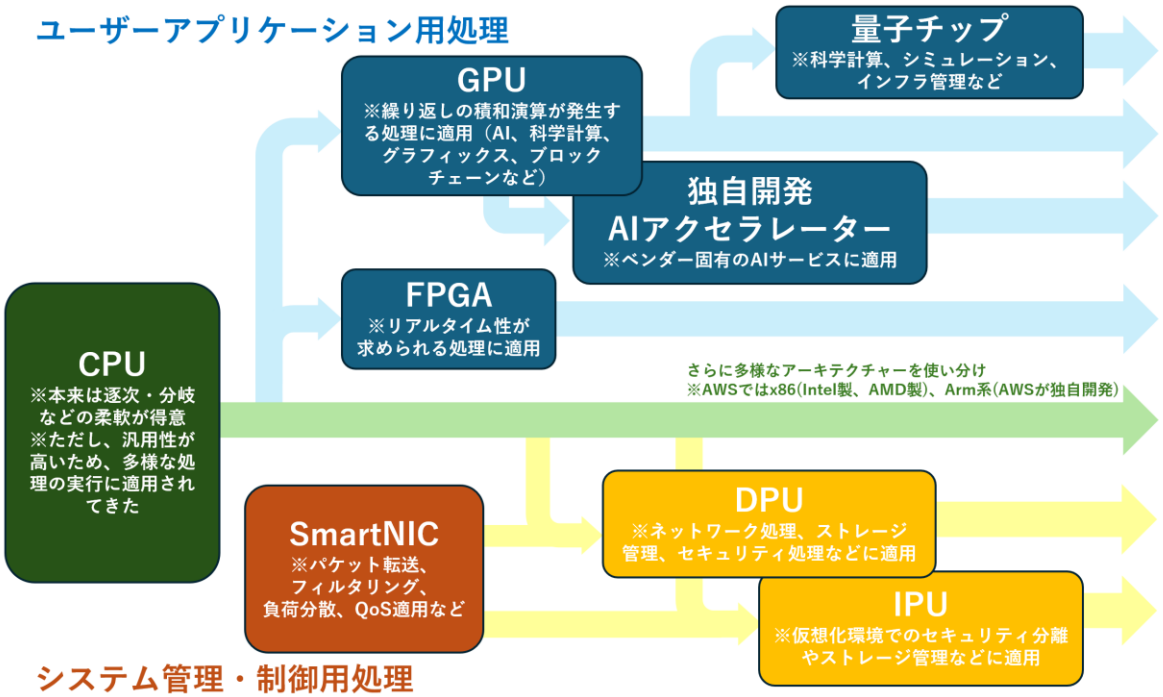


図 3　CPU で実行する処理のオフロード化

出所：日経 BP 総合研究所

かつてのデータセンター向けサーバに搭載されていた CPU では、ユーザーアプリケーション用処理以外にも、データセンター内を管理・運営するためのシステム管理・制御用処理も実行していた。つまり、サーバの利用者は、顧客とデータセンター運営者という立場が異なる 2 者がいた。この状況は、データセンターの危機管理、ガバナンスの観点から見て問題があった。現在、CPU のタスクのオフロード化は、この 2 者を明確に区別して、それぞれ進められている。

●ユーザーアプリケーション用処理でのオフロード化

ユーザーアプリケーション用処理では、先述したように、まず GPU が導入されて負荷分散された。GPU には、繰り返しの積和演算が発生する処理が割り当てられている。具体的には、AI 関連

処理や科学計算処理、グラフィックス処理、ブロックチェーン関連の処理などである。

詳しくは、アクセラレータの動向を示す章に記すが、GPU で実行していた AI 関連処理を実行するアクセラレータを、AI を活用したサービスを提供するクラウド事業者が独自開発する例が増えている。米国の GAFAM（グーグル、アップル、Facebook（メタ）、アマゾン、マイクロソフト）、中国の BATH（バイドゥ、アリババ、テンセント、ファーウェイ）は、例外なく独自 AI アクセラレータを開発し、自社サービスなどに適用している。

また、現在 GPU で実行しているタスクのうち、科学計算などは将来的には量子コンピュータで実行される例が増える可能性が高い。グーグルが量子チップ「Willow」を開発するなど、5～10 年先の実用化を見据えて、この領域での技術開発が急激に進む可能性が高まっている。ただし、GPU や科学計算専用のスーパーコンピュータで実行している厳密なシミュレーションが、量子コンピュータへとオフロード化される方向に向かうとはみられていない。双方は、併用される方向に向かう可能性が高い。

一方、リアルタイム性が求められる処理に関しては、プロセッサとは異質なプログラマブルデバイスである FPGA（Field-Programmable Gate Array）が適用されている例が多い。例えば、マイクロソフトは、「Bing」の検索エンジンの高速化やサーバ間のネットワーク構成の最適化などに利用している。FPGA は、プログラムを書き込むことによって専用回路を構成することができるため、非ノイマン型のメモリアクセスを最小化した情報処理が可能である。プログラムする専用回路の設計次第では、プロセッサを利用するよりも低消費電力化が可能なこと、さらには近年プログラムを動的に書き換えて多様な用途に柔軟適用可能になってきていることから、利用例が増えている。

●システム管理・制御用処理でのオフロード化

システム管理・制御用処理のオフロード化に向けて、DPU（Data Processing Unit）や IPU（Infrastructure Processing Unit）と呼ばれる新しいタイプのプロセッサが開発・導入されるようになった。

DPU と IPU は、いずれもスマート NIC と呼ばれる、ネットワーク処理を実行していたチップの発展チップであり、ほぼ同じ目的で利用されるチップである。ただし、DPU よりも IPU の方が、より多機能化している傾向が見られる。DPU には、エヌビディアや AMD、Marvel Technology が開発・外販、マイクロソフトと AWS が独自開発して自社利用している。中でもエヌビディアは、GPU における CUDA（Compute Unified Device Architecture）にあたる、DPU 向けの開発プラットフォームとして DOCA（Datacenter-on-a-Chip Architecture）を提供しており、DPU と同様のベンダーロックインの状態を現出させる戦略を実践している。一方、IPU はインテルが開発・外販しており、同社は Google Cloud と共同開発してグーグルのデータセンターに導入している。

DPU で実行する処理には以下のようなものがある。IP、TCP、UDP、HTTP などの通信プロトコル処理。パケットのルーティング、トラフィック管理、Open vSwitch（OVS）の実装のためのパケット解析・照合・操作といったネットワーク処理。ストレージにデータを読み書きする際の転送、圧縮、管理に関わる処理。暗号化/復号化、パケット検査、ファイアウォール機能などセキュリティ関連処理などである。それぞれの処理に適した、専用回路を搭載している例が多い。

一方、IPU は、DPU の機能の一部に加え、ストレージやネットワークの仮想化に関連した処理を実現。仮想化関係の専用処理回路を、FPGA や ASIC で搭載し、効率的な管理・運営を可能にしている。

●オフロード化によって、機能が先鋭化しつつある CPU

これまで多様なタスクを実行してきた CPU から、多くの処理が他のプロセッサやプログラマブルデバイスへと移管されたことで、CPU の機能・構造の先鋭化が進みつつある。これからのデータセンターでは、CPU で実行する処理は、企業で顧客管理や受発注管理などを行う基幹システムや、ネット通販やネットバンキングなどの処理が中心になってくる。AI 関連処理のように、大量のデータを対象にして同じ演算を繰り返すようなタイプの処理ではなく、複雑だが小規模なタスクが膨大に発生する処理が中心になる。このため、一つひとつの CPU コアは非力であっても、多数のコアを並列動作させるような構造が向いてくる。

実際、こうした傾向が、CPU メーカーの製品開発に明確に現れるようになった。例えばインテルの例を挙げると、応用先で実行するタスクの内容を決め打ちして、そのタスクに向いたコアを多数集積したチップを開発し、製品化するようになった（図 4）。

図 4　処理対象となるタスクのオフロード化が進み、CPU の機能と構造が先鋭化してきた
（左）インテルの高性能コア（P-Core）と高効率コア（E-Core）それぞれの仕様、（右）P-Core もしくは E-Core のどちらか一方だけを多数集積してデータセンター用チップを構成
出所：インテル

近年、パソコン向け CPU では、単一スレッドの処理性能が高い“高性能コア”と電力効率に優れコア面積が小さい“高効率コア”を一定の割合で混載させて、用途に応じて柔軟に最適なコアを活用して、要求性能と低電力化を両立させる構成が採用されている。arm が導入した big.LITTLE と呼ばれる技術である。ただし、この構成は、パソコンのように多目的で利用される場合に効果を発揮する。

一方、サーバ用では、同一世代で高性能コアと高効率コアの 2 種類を用意する点は同じだが、チップに集積する際には、どちらか一方だけを集積してチップ内では使い分けない構成にすることが多くなった。注目できる点は、データセンターでは巨大な演算能力が求められるため、高性能コアを適用するというわけではなく、むしろ高効率コアをより多く集積したチップを作るようになってきた。そして、高性能コアを集積したチップでは演算負荷の高い HPC やデータサービス

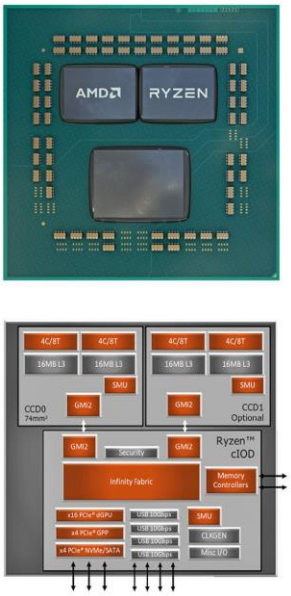
などのワークロードに最適化し、高効率コアを集積したチップでは低負荷でアイドル時間の割合が高いワークロードに最適化した構成を採用している。高性能コア版では、用途を想定した専用演算器の搭載などを進め、より機能の先鋭化を図っている。今後、同一世代で用意するコアの種類がさらに増えたり、特定用途限定の命令セットや専用回路の導入などが進められる可能性を指摘する声が聞かれる。

クラウドサービスを提供する企業でも、CPU の活用において特徴的な動きが見られるようになった。AWS では、利用可能な CPU として、Intel 版 x86、AMD 版 x86、arm 系の自社開発 CPU など、多様な CPU をあえて用意。コストパフォーマンスを最適化し、顧客の特定ニーズに最適なインスタンスを選択できるようにしている。CPU の選択は、技術的な優劣だけで行われるわけではなく、ユーザー側のサービス利用目的や技術資産、ライセンス取得状況など多様な側面から総合的に決められる。今後は、AWS の CPU「Gravition」のように、提供サービスの仕様とチップ仕様を擦り合わせた独自開発が増えてくる可能性がある。

●チップレットの適用を起点にした技術革新

先端半導体の多くが、チップレット技術を活用して設計・製造されるようになった。チップレットとは、これまで 1 チップに集積した大規模な回路をあえて複数の小さなチップに個片化し、インタポーザー上に乗せて大規模化して 1 パッケージに収める技術である。

回路集積技術としてのチップレットの有効性が広く知られるようになった契機となったのが、2019 年に発売された AMD の 7nm 世代の CPU「Ryzen 3000 シリーズ」。同社は、競合であるインテルを上回る多コア化を実現しながら、採算が取れる歩留まりでの生産を可能にするための技術としてチップレットを導入して成功を収めた。ムーアの法則の守護者を自認するインテルは、当時、大規模回路の 1 チップ化にこだわりを持っており、10nm 世代の立ち上げに失敗して、発売延期を繰り返してしまった。AMD は、チップレットを活用することによって、よりコア数の多いチップを市場投入することに成功し、市場での同社の競争力を劇的に高めた。チップレットは、チップ面積が大きくなりがちなサーバ用 CPU で特に有効である。市場でのシェアは、インテル製サーバ用 CPU「Xeon」が AMD 製「EPYC」を上回っているが、これは AMD が確保できる台湾 TSMC の生産枠が限定されているからであり、市場での製品競争力は常に AMD 製が勝っている状況である。現在では、インテルもチップレット化を推し進め、この技術の活用は先端チップを作るうえでの大前提となった。



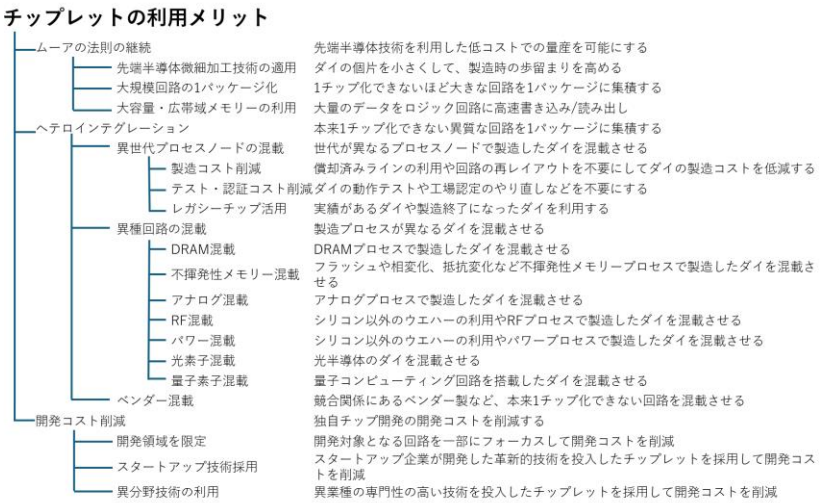


図 5　チップレットを半導体の開発・製造に適用する際のメリット
(左)AMD の Ryzen 3000 シリーズとチップレットの構成、(右)チップレットの利用メリット
出所：日経 BP 総合研究所

チップレットの利用メリットは極めて多様である（図 5)。先端チップを開発する際には、技術面とビジネス面の両面で様々なしがらみ・課題から 1 チップ化が困難な事情が生じてくる。例えば、先述した AMD やインテルが直面したように「大規模すぎる回路の集積」「プロセスノードが異なる回路の集積」、DRAM やアナログといった「製造技術が異なる回路の集積」。あるいは「自社開発できない回路の集積」などもそうした例に挙がる。チップレットは、こうした技術面・ビジネス面でのしがらみ・課題を包み込み 1 つにまとめることが可能である。今後、こうしたチップレットの特徴をフル活用した CPU が続々と登場してくる可能性が高い。

既に、必ずしも最先端プロセスノードを適用する必要がない、インタフェース回路などのチップレットを成熟したラインで生産し、最先端ノードで作ったプロセッサコアと集積した製品が投入されるようになった。また、5nm や 3nm といった先端プロセスノードでは、SRAM の回路面積が小さくなりにくいといった課題を抱えており、こうした回路のチップレットを成熟したラインで生産することで、低コストで、より容量のキャッシュメモリを搭載可能になってきている。近年では、チップレット間、もしくはチップレットとインタポーザーの間を、バンプを介することなく直接金属接合するハイブリッドボンディング技術が量産適用されるようになり、1 チップ集積する場合と遜色のない性能を実現できるようになった。

●今後 10 年間を見据えた CPU 開発動向のポイント

かつて CPU で実行していた処理のオフロード化やチップレットの適用は、今後 10 年間を見据えた長期的視野からの CPU 開発においても極めて重要な技術である。

オフロード化では、さらに処理対象のタスクの内容を見極めて専用化を推し進めたプロセッサを開発し、使い分けていくことになる可能性が高い。ムーアの法則が疲弊してきたことで、シス

テム開発の手法が大きく変わっていく可能性があるからだ。もはや、最先端プロセスノードを適用すれば、CPU などプロセッサの性能向上が保証され、後はソフトウェアで味付けすれば市場や顧客の要求に柔軟対応できる時代ではなくなりつつある。そして、ソフトウェアとハードウェアを、仕様を擦り合わせながら、同時並行開発していく時代が到来する公算が高い。その際、チップの仕様に差異化要因を盛り込む必要がある。また、チップをすべて設計し直すことは困難であるため、仕様変更が必要な部分だけにフォーカスしてハードウェアを更新していくための手段としてチップレットが活用されることになりそうだ。

加えて、今後の 10 年間を見据えた場合、いよいよノイマン型アーキテクチャの疲弊に対する対策を本格的に講じていく必要がありそうだ。チップレットおよび付随する実装技術を有効活用すれば、従来、別モジュール、別基板に実装されていた異種回路を、高速・広帯域のバスを通じて高密度実装できるようになる。CPU とメモリを集積することで、ノイマン型であることによるボトルネックに対策することができる。実際、パソコン用での DRAM と SoC を集積したアップルの独自チップ「M シリーズ」や、大容量・広帯域の DRAM である HBM（High Bandwidth Memory）を集積したエヌビディアの GPU のように、成功例が出てきている。今後は、より高度な実装技術を投入することで、さらなる高性能化と低消費電力化を目指すことになる。

ただし、それでも１パッケージ、１モジュールには集積できない回路の組み合わせや、より高速・大容量のデータ転送が求められる用途では、信号の転送手段を電気信号によるから光信号へと刷新する必要が出てくる。今後 10 年間は、ボード上でのデータ転送、パッケージの入出力において、光インターコネクトが求められてくることになりそうだ。

②アクセラレータに対する要求事項

AI を活用したサービスや応用分野の拡大、さらにはサービスの利用者の急増によって、データセンター内で処理する AI 関連処理の負荷が劇的に高まっている。データセンター内で実行すべき処理に占めるその負荷の割合は高まる一方だ。その結果、もともとデータセンターの頭脳として利用されていた CPU を活用して AI 関連処理を実行したのでは、非効率かつ需要の増大に応えることができなくなってきた。

こうした状況に対応するため、近年、アクセラレータ、特に GPU をはじめとする AI アクセラレータを導入・活用する動きが広がっている。現在では、専ら AI 関連処理を実行する AI データセンターが設置されるようにもなってきた。GPU 最大手のエヌビディアが、一時期、株式時価総額世界 1 に躍り出た背景には、モデルの学習に莫大な演算能力を求める生成 AI の利用拡大に伴う同社 GPU の爆発的需要増がある。もはやデータセンター向けチップの領域では、存在感と成長期待において AI アクセラレータが CPU を凌駕する状況になってきている。

その一方で、AI 関連処理の実行に伴う GPU など AI アクセラレータでの消費電力増大と莫大な発熱が、可及的速やかに解決すべき大問題になってきている。一般に、AI 関連処理を実行する際の GPU の稼働率は、CPU とは比べものにならないほど高い。このため、AI アクセラレータには、高性能化と共に、低消費電力化と発熱への対策技術も同時投入した進化が求められてくる。

本調査では、データセンター向けAIアクセラレータの技術開発動向について調査し、現時点から3年後の2028年の短期的な動きと、10年後の2035年の中長期的な動きに分別してまとめた。

●今後、3年間で想定されるAIアクセラレータ開発のポイント

直近、3年間で想定されるAIアクセラレータ開発のポイントをまとめた（図6）。

図中、横方向には、情報処理システムの機能・性能を具現化・詳細化していく流れに沿って、具体的な技術開発の動きを「システム設計」「チップ設計」「チップ製造」「チップ実装」それぞれの段階に位置付けて示した。縦方向には、技術開発の動きを、適用対象となるシステム範囲を「チップスケール」「モジュールスケール」の各領域に分類して示した。図中のオレンジの枠内に示した動きが、直近3年間で想定される技術動向のポイントである。赤の矢印は、各技術動向間の依存関係を表している。一方、緑の枠内に示した動きは、10年後までの長期的視野に立って、実用化が進められる動きである。緑の矢印で、同様に依存関係を示した。なお、直近3年間で想定される動きは、10年間の長期的視野に立って見ても継続的に見られると予想される。

直近、3年間で想定されるAIアクセラレータ開発の動向のポイントの中から、データセンター向けGPUの技術開発を先導するエヌビディアに見られる技術開発の潮流、「ユーザー企業によるAIアクセラレータの独自開発」の動きとこの領域での技術開発を先導するグーグルに見られる技術開発の潮流、および演算性能と電力効率の両立に向けた究極的な取り組みと言える「IMC（In Memory Computing）の適用」に関する動きについて、詳細を示したい。

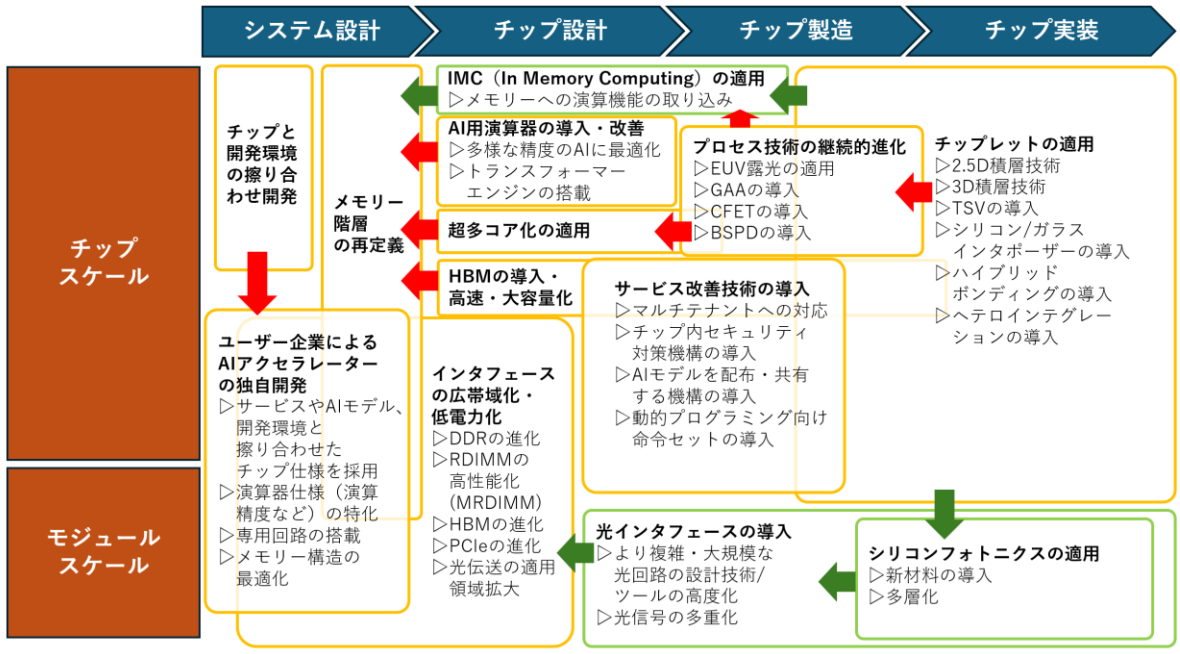


図6　AIアクセラレータ開発のポイント

出所：日経BP 総合研究所

●AI関連処理向けGPUの開発を主導するエヌビディアに見られる技術開発の潮流

現在、データセンターにおける AI 関連処理を効率化するために利用されているアクセラレータの主流は GPU である。なかでも、エヌビディアの GPU は、データセンター向け市場において 2024 年時点で 92%ものシェアを占める、ほぼ独占状態となっている。GPU では、AMD やインテルもデータセンター向け製品を保有しているが、同社に割って入る状況にはない。

エヌビディアがデータセンター向け GPU で強みを発揮してきた背景には、GPU と、そこで動かすソフトウェアの開発プラットフォーム「CUDA」の両者を擦り合わせ開発して、両者の歩調を合わせながら進化させてきたことがある。同社の技術・商品戦略を言い換えれば、パソコンにおけるインテルの役割とマイクロソフトの役割を 1 社で兼ね備える戦略であると言える。そして、CUDA は、2006 年に、グラフィックス処理むけだった GPU を非グラフィックス用途に活用するための開発環境としてリリース。高速な積和演算処理の手段を求めていた黎明期の AI 研究者が好んで採用し、その後の AI ブームを生み出すことになった。

CUDA を利用した AI の研究開発は、研究者とエンジニアの中に深く根を下ろした状態が続いている。エヌビディアは、GPU の世代代わりを続ける中で、時々の AI 技術トレンドを取り入れて機能・性能を向上させていくのと同時に、研究者やエンジニアが新たな技術革新を生み出す余地を残すためにあえて汎用性を維持したアーキテクチャを採用し続けている。こうした汎用性の高い GPU の進化と、進化した GPU の機能・性能を効果的かつ効率的に活用する CUDA の改善を両輪とすることで、現在の同社の独占的競争力を維持している。エヌビディアに対抗する AI アクセラレータを開発・投入して成功するためには、同社が築いた、ユーザーの開発文化に根差したビジネス体制を崩す論点を生み出すことが必要条件となる。単に、技術面での優位性を示しただけでは対抗できない。

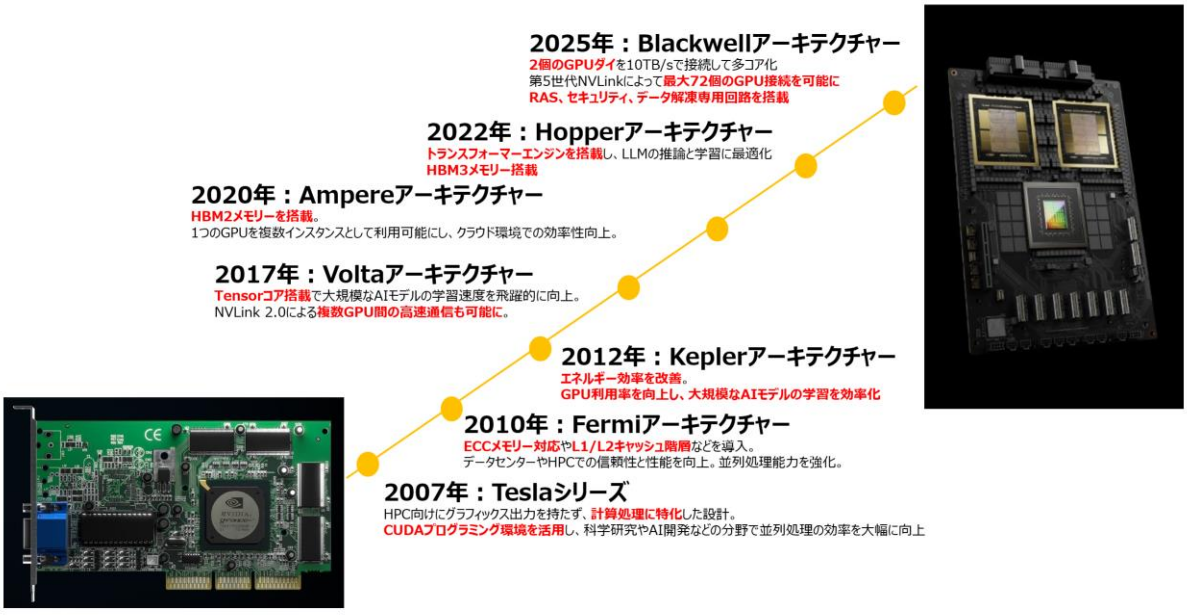


図 7　エヌビディアによるデータセンター向け GPU の進化の流れ
出所：日経 BP 総合研究所、図中の写真はエヌビディア

エヌビディアによる、データセンター向けGPUの進化の流れを図7に示す。同社は、1999年に世界初のGPU「GeForce 256」を市場投入し、応用市場の拡大を狙って、2007年にはグラフィックス向けの出力機能をあえて外したGPU「Teslaシリーズ」をCUDAと共に市場投入した。先述したように、GPUのチップだけでなく、開発環境も同時投入した点が、結果的に現在のAI市場を生み出す足がかりとなった。もっとも、エヌビディア自体は当初からAI向けを狙っていたわけではなく、スーパーコンピュータで対応していた科学計算をもっと手軽にできる環境を作ることが狙いだったようだ。

その後、同社は、第1世代のTeslaシリーズから、2025年に製品を投入する第7世代の「Blackwell」に至るまでの間にアーキテクチャの刷新を進めていった。同時にCUDAの機能更新も進めている。その間、同社はチップに搭載するGPUコア数を増やすことによる高性能化を進めてきたわけではない。同時に、HBMの早期採用や複数チップ間の連携に向けた独自高速通信技術の導入などによるシステムレベルでのAI関連処理の性能向上、Tensorコアやトランスフォーマーエンジンなど AIモデルの進化と歩調を合わせた専用回路の導入、およびGPU利用率の向上やRAS（Reliability, Availability, Serviceability)、セキュリティ、データ解凍専用回路の搭載などの、データセンターでの使い勝手と付加価値を向上させる機能の搭載などを、一貫して推し進めてきた。こうした基本方針に関しては、将来に同社が投入することになるデータセンター向けGPUでも大きな変わりがないものと思われる。ただし今後は、チップレットを導入して、さらなる大規模化が進む可能性が高い。

エヌビディアの目下の課題は、高性能化に伴う、GPUでの消費電力増大とそれに付随する発熱への対策である。5年前ぐらいの製品は、1サーバ当たりの消費電力が約3kVAであった。しかし、現在は10kVAを超えている。これほど大電力・高電力密度になると、既存の一般的なデータセンターでは1ラックにサーバが1台も入らないという事態になりかねない。大電力・高密度への対応が必須になる。特に、発熱の増大はGPUの安定稼働を妨げる大問題である。同社は、オープンなパートナープログラムと特定企業とのクローズな連携などを通じて、この領域での技術革新に取り組んでいる。

AMDやインテルなど、エヌビディアの競合企業が投入しているGPUは、基本的な進化の方向性に関してはエヌビディアのチップと大きな違いはない。ただし、インテルに関しては、データセンター向けGPUビジネス自体が後発であることから、AI以外のクラウドゲームやメディア配信などのワークロードへの適用を想定したチップ設計を進めている。開発環境に注目すると、AMDは、ROCm(Radeon Open Compute)と呼ぶオープンソースのプラットフォームを提供している。ROCmは、CUDAと同様に開発者がGPUの力を引き出すためのツールやライブラリを提供しているが、現時点はCUDAほどの普及やサポートがない。一方、インテルは、OneAPIという統一プログラミングモデルを提供し、CPUやGPU、FPGAなどの異なるプログラマブルデバイスを連携させたシステム開発の容易化を目指している。OneAPIは、CUDAやROCmと同様に並列処理を効果的に実行するための機能を備えているが、まだ開発者コミュニティの規模やサポートは他のプラットフォームに比べて小さい。

スタートアップ企業では、Tesstorrent が、AI アクセラレータと CPU を密に統合したチップ「Grendel」を開発し、市場投入を目指している。AI プログラムの一部処理が CPU で行われることがあることから、この部分でより効率的なデータ通信を実現するために、RISC-V アーキテクチャの CPU と AI アクセラレータをチップレットとして統合する。

また、半導体メーカーが提供する AI アクセラレータとして、GPU とは別に、FPGA が AI 関連処理に適用される例もある。実際、マイクロソフトは、リアルタイム性が重視される推論処理に FPGA を適用している。

●ユーザー企業による AI アクセラレータの独自開発の動き

半導体メーカーによるエヌビディアへの対抗は、現時点では今ひとつ実績を上げられない状況である。ただし、別の角度から AI アクセラレータ市場の切り崩しが進められてきている。AI アクセラレータを活用し、多様なアプリケーションに向けた学習や推論などの AI 関連処理を実行する場を提供するクラウドサービス事業者による、AI アクセラレータを独自開発・自社利用する動きである。米国の GAFAM、中国の BATH は、例外なく独自 AI アクセラレータを開発し、自社サービスなどに適用。ハイパースケーラーにとっての競争の論点となっている（図 8）。

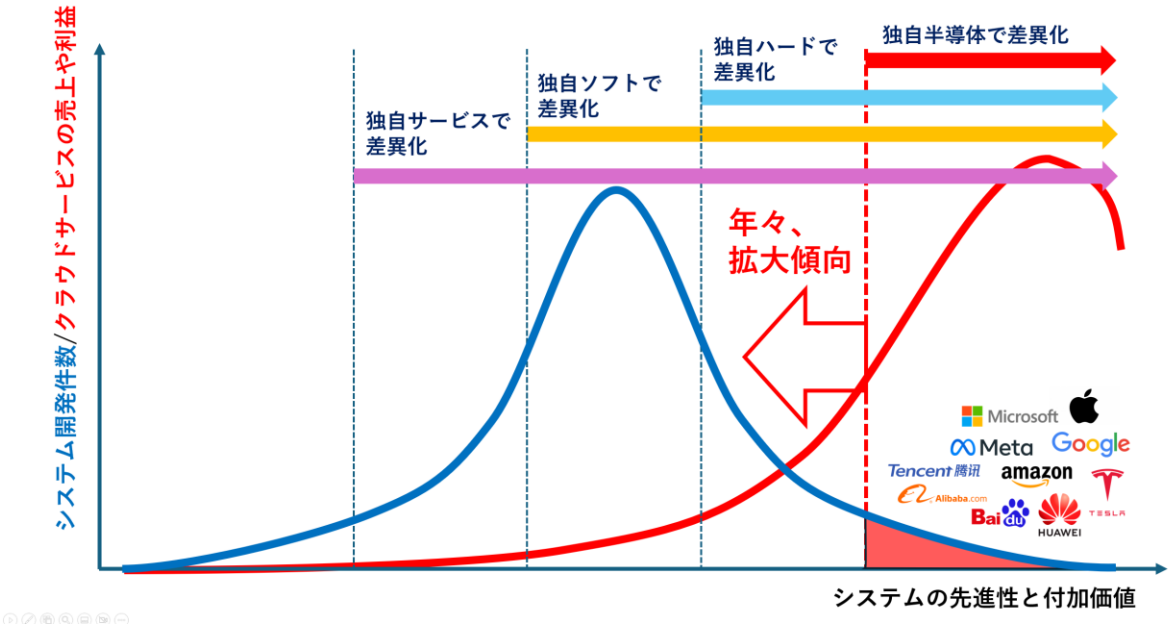


図 8　ハイパースケーラーなどは、AI アクセラレータの独自開発でシステムの差異化を目指す
出所：日経 BP 総合研究所

ハイパースケーラーが AI アクセラレータの独自開発を進めている背景には、大きく二つの要因がある。

一つは、エヌビディアのデータセンター向け GPU が極めて高価で、しかも安定調達が困難なこと。同社の第 6 世代 GPU である H-100 は、1 基あたりの価格が約 3 万 5000 米ドル（日本円で約 500

万円）となり話題になった。2025 年に投入される Blackwell アーキテクチャの GPU は、約 7 万米ドル（同 1000 万円）になるとみられている。さらに、エヌビディアは工場を持たないファブレスの半導体メーカーであり、製品製造は TSMC などに委託しているが、TSMC の生産枠の獲得が需要に追いつかない状態である。このため、クラウドサービス事業者が想定した量のチップを確保することが困難である。自社開発して、製造委託する先を自由に選べれば、調達が安定する余地が出てくる。

もう一つは、エヌビディアの GPU と CUDA の組み合わせで AI 技術やサービスを開発しても、他社に対する差異化が困難であること。特に、ムーアの法則が疲弊してきた昨今では、半導体の微細加工技術だけに頼ったハードウェアの進化が望み難くもなってきている。このため、AI システムのハードウェアの中核を独自開発し、自社のサービスの形態とそこで使われる AI モデル、ソフトウェアなどの仕様と擦り合わせれば、AI アクセラレータとして、システム中で極めて効率的に機能するチップが出来上がる。ただし、こうして出来上がった独自チップは、基本的に自社だけしか使えないため、一定の開発・生産コストを投じて、なおかつ一定量を利用することで量産効果が見込める巨大企業のみが独自開発に踏み切ることができる。現時点で、日本でクラウドサービスを提供する企業のなかでは、プリファード・ネットワークス以外に同様の手段を導入できているところはない。今後は、チップレットなどを適用して、開発すべき回路規模を限定し、より多くのクラウドサービス企業が独自アクセラレータを開発できる環境、仕組みづくりが望まれる。

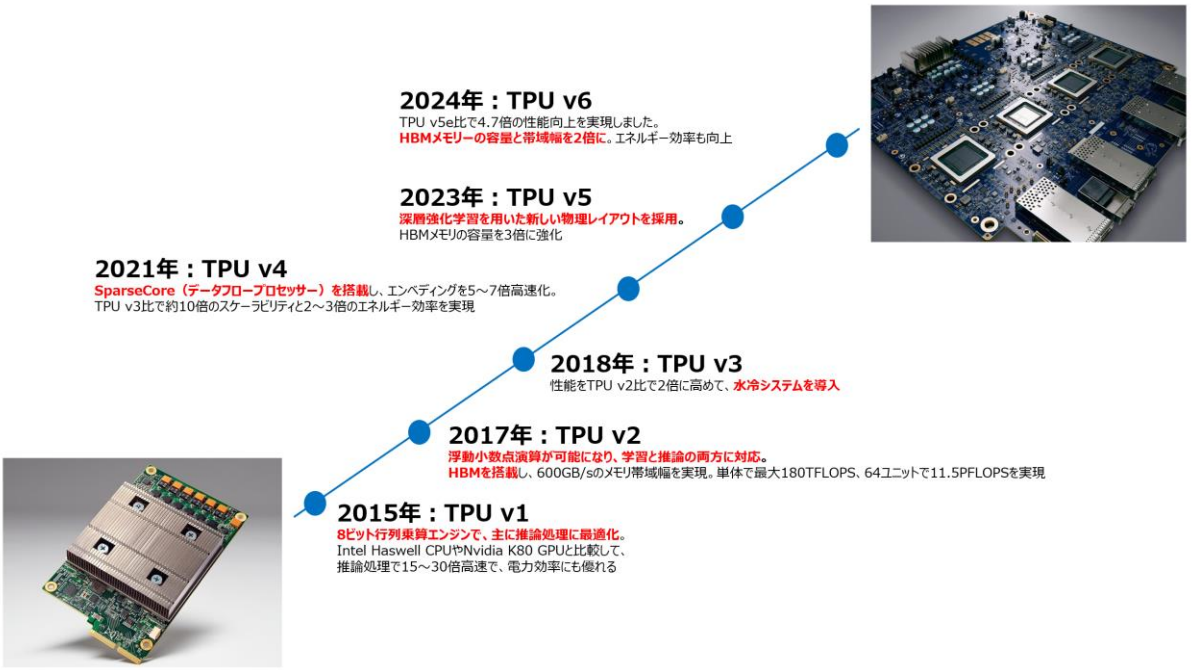


図 9　グーグルによるデータセンター向け GPU の進化の流れ
出所：日経 BP 総合研究所、図中の写真はグーグル

AI アクセラレータの独自開発に最も早く着手し、この領域での技術開発と実戦投入をリードしているのがグーグルである。同社による、データセンター向け GPU の進化の流れを図 9 に示す。

グーグルは 2013 年からデータセンター向け独自 AI アクセラレータ「TPU(Tensor Processing Unit)」の開発に着手し、2015 年に第 1 世代となる「TPU v1」を完成させて、自社データセンターに配備した。TPU v1 は、推論処理に特化したチップであり、演算精度を 8 ビット整数に落として演算器を簡素化した分、より多くの演算器を並列搭載することで演算性能を高めた行列乗算エンジンの構造を採用した。これは、推論処理では、演算精度を向上しても推論結果の精度向上には大きくは寄与しないことを見極めた手法である。高精度の演算器を利用するエヌビディアの GPU に対する勝機をそこに見出して、開発された。実際、当時製品化されていたエヌビディアの第 3 世代 GPU「K80」比で、推論処理に関しては 15～30 倍高速で、なおかつ電力効率に優れたチップを実現している。

その後、同社は、2017 年に 32 ビット浮動小数点演算を可能にして、学習と推論の両方に対応する第 2 世代「TPU v2」を投入。いよいよ、エヌビディアの GPU と全面競合していく体制を整えた。TPU v2 では、HBM をエヌビディアよりも早く投入するなど、先回りした技術開発を進めた。その後、2024 年に投入した「TPU v6」に至るまでに 6 世代、自社内での適用先を明確にしながらチップを進化させ続けている。現時点で、グーグルは、エヌビディアからも GPU を購入して利用しており、TPU に 1 本化しているわけではない。ただし、サービスユーザーの選択肢の一つとして着実に技術と実績を蓄積しつつある。

GAFAM や BATH さらには、自動車メーカーである Tesla などが同様にデータセンター向け AI アクセラレータを開発しているが、それぞれのチップの基本的な構造に大きな違いはない。学習用と推論用の両方を開発している点もほぼ同様である（ただし、アリババとバイドゥに関しては推論用のチップ開発の動きしか見えていない）。それぞれ、AI 技術を保有する企業もしくは半導体設計企業などを買収して、独自チップ開発を加速させている。

ハイパースケーラーによる独自チップ開発とは真逆のアプローチを実践するスタートアップ企業も出てきている。米グロック（Groq）は、大規模言語モデル（LLM）向けの推論専用アクセラレータを開発。それを活用したクラウドサービスを自社提供している。同社のチップは、外部の DRAM とのデータのやり取りはせずに、チップ内部の SRAM だけで低消費電力のデータ転送を可能にしている。ただし、現時点ではこうしたアプローチのチップを利用するためには、特殊なプログラミング技術が求められ、自社利用し、クラウドサービスのかたちで提供している。

●今後 10 年間を見据えた AI アクセラレータ開発のポイント

直近で展開されている AI アクセラレータの技術開発のトレンドは、今後 10 年間を見据えた長期的視野から見た際にも大きな変わりはない。ただし、チップレットの活用による大規模化（多コア化による高性能化）の進展とシリコンフォトニクスによる光インターコネクトの導入、冷却技術での液冷・液浸などを含む技術革新を模索する動きは一層加速する可能性が高い。

さらに、現時点では、ハイパースケーラーにおいても、エヌビディアの GPU の使用がデータセンターでの AI アクセラレータの中心だが、独自チップの技術開発と応用実績が蓄積していくことで、どの程度の置き換えが進むかが注目点になってくる。ハイパースケーラーによる独自チップの開発が中心になってくると、大市場であるデータセンター向け AI アクセラレータ市場に半導体専業のメーカーが入り込む余地がなくなってくる。

ただし、今後 10 年間の間に起きる可能性が高い技術革新に、AI アクセラレータとメモリの融合がある。「イン・メモリー・コンピューティング（In Memory Computing：IMC)」と呼ぶ新たなハードウェア構成を採用したチップの投入である。IMC は「Processing in Memory（PIM)」や「Computation in Memory（CIM)」と呼ばれる場合もある。

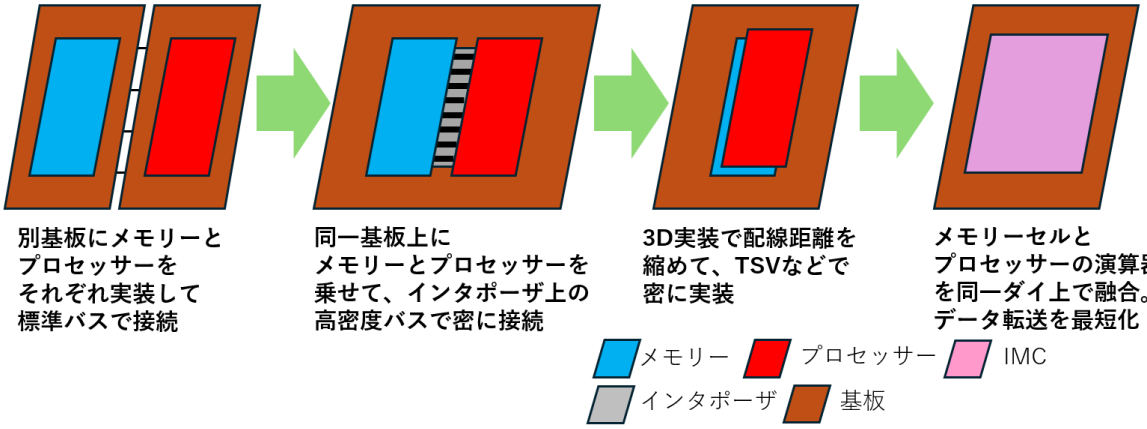


図 10　AI アクセラレータとメモリの融合によって、ノイマン型固有の消費電力増大要因を除外
出所：日経 BP 総合研究所

IMC は、ノイマン型アーキテクチャ固有の課題である、データ転送バスでの消費電力増大に対する抜本的解決策である。特に、ニューラルネットワークや機械学習の演算で消費する電力の大部分が、バスを介した莫大なデータ転送で発生しているとされている。チップレットとそれに付随する高度な実装技術の活用によって、AI アクセラレータとメモリの間の距離を、ジリジリと近づける方向へと技術開発が進められている。究極的には、双方を融合させてしまえば、データ転送が不要になり、そこで発生する消費電力を最小化できるという発想の技術である。IMC の構造は、記憶と処理（判断、知覚、感情など）の機能が物理的に一体化されている脳の構造に酷似している。そして、人間の脳は、極めて高度で汎用性の高い知的処理が可能でありながら、その活動で消費しているエネルギーは電力換算で約 20W に過ぎないと言われている。

IMC は、既に、IBM や NEC など国内外の IT 企業、さらにはインテルやサムスン電子、ルネサスエレクトロニクスなどの大手半導体メーカー、多くの大学・研究機関が、それぞれ特徴のある IMC の構造を考案し、実用化を目指して技術の研究開発を進めている。そして、今後 10 年の間には、何らかの製品と応用が社会実装されるとみられている。

③ストレージに対する要求事項

データセンターの役割は、大きく二つある。DX や GX の実践に伴う多様な情報処理を実行する役割と、社会の中で生み出される情報・データを蓄積し、共有・有効利用する役割である。前者の主役の主役となる半導体チップは CPU や AI アクセラレータ。そして、後者の主役が、DRAM やフラッシュメモリなどメモリデバイス、さらにはそれらを導入した SSD などになる。

ノイマン型アーキテクチャに基づく情報処理では、CPU などの演算器とメモリは同等に重要な構成要素である。システムを高性能化する際にも、低消費電力化していく際にも、それら両方を同時進化させていく必要がある。特に、データ駆動型社会において発生する多様な情報処理、特に AI 関連のアプリケーションで実行される情報処理を進化させていくためには、メモリやストレージ周りの技術の進歩が特に重要になってくる。具体的には、低コストでの莫大なデータの蓄積手段と、効果的かつ効率的に演算器へとデータを転送する手段の両方が必須になる。

メモリやストレージの技術は、年々、より高速・広帯域かつ大容量な方向へと進化している。ただし一般に、高速・広帯域でデータを読み出し/書き込みできるメモリやストレージほど高価である。このため、あらゆるデータを最も高速なメモリに蓄積することはできない。その一方で、蓄積しておくべきデータは極めて莫大。このため、プロセッサチップ内に搭載される SRAM のキャッシュメモリ、DRAM、SSD 内のフラッシュメモリ、さらには HDD など特徴が異なるデバイスを適材適所に階層配置して、システム要件に合った仕様のメモリシステムを構成・活用している。そして、メモリシステムの構成要素の一部に技術革新が起きれば、メモリ階層全体を再定義していくことになる。

CPU や AI アクセラレータにおいて顕在化してきているムーアの法則とノイマン型アーキテクチャの疲弊は、メモリ/ストレージの分野でも見られる。微細加工技術の進歩の困難さは、むしろプロセッサを構成するロジック回路以上であるとも言える。実際、特に集積度の向上が求められるフラッシュメモリでは、他の半導体チップよりも早く 3 次元化が進んだ。また、メモリ/ストレージは、演算器との間での円滑なデータのやり取りが求められるため、ノイマン型が疲弊しつつある現在、インタフェース領域にも技術革新が強く求められている。

本調査では、データセンター向けメモリ/ストレージの技術開発動向について調査し、現時点から 3 年後の 2028 年の短期的な動きと、10 年後の 2035 年の中長期的な動きに分別してまとめた。

●今後、3 年間で想定されるメモリ/ストレージ開発のポイント

直近、3 年間で想定されるメモリ/ストレージ開発のポイントをまとめた（図 11）。

図中、横方向には、情報処理システムの機能・性能を具現化・詳細化していく流れに沿って、具体的な技術開発の動きを「システム設計」「チップ設計」「チップ製造」「チップ実装」それぞれの段階に位置付けて示した。縦方向には、技術開発の動きを、適用対象となるシステム範囲を「チップスケール」「モジュールスケール」「システムスケール」の各領域に分類して示した。図中のオレンジの枠内に示した動きが、直近 3 年間で想定される技術動向のポイントである。赤の矢印は、各技術動向間の依存関係を表している。一方、緑の枠内に示した動きは、10 年後までの長期的視野に立って、実用化が進められる動きである。緑の矢印で、同様に依存関係を示した。なお、直近 3 年間で想定される動きは、10 年間の長期的視野に立って見ても継続的に見られると予想される。

直近、3 年間で想定される AI アクセラレータ開発の動向のポイントの中から、メモリ階層の再定義に大きな影響を及ぼす、チップレットの適用に基づく「L3 キャッシュの高速化・大容量化」の動きと「HBM の導入とその高速化・大容量化」に関わる潮流、および新たな技術の導入が相次ぐ「インタフェースの広帯域化・低電力化」に関する動きについて、詳細を示したい。

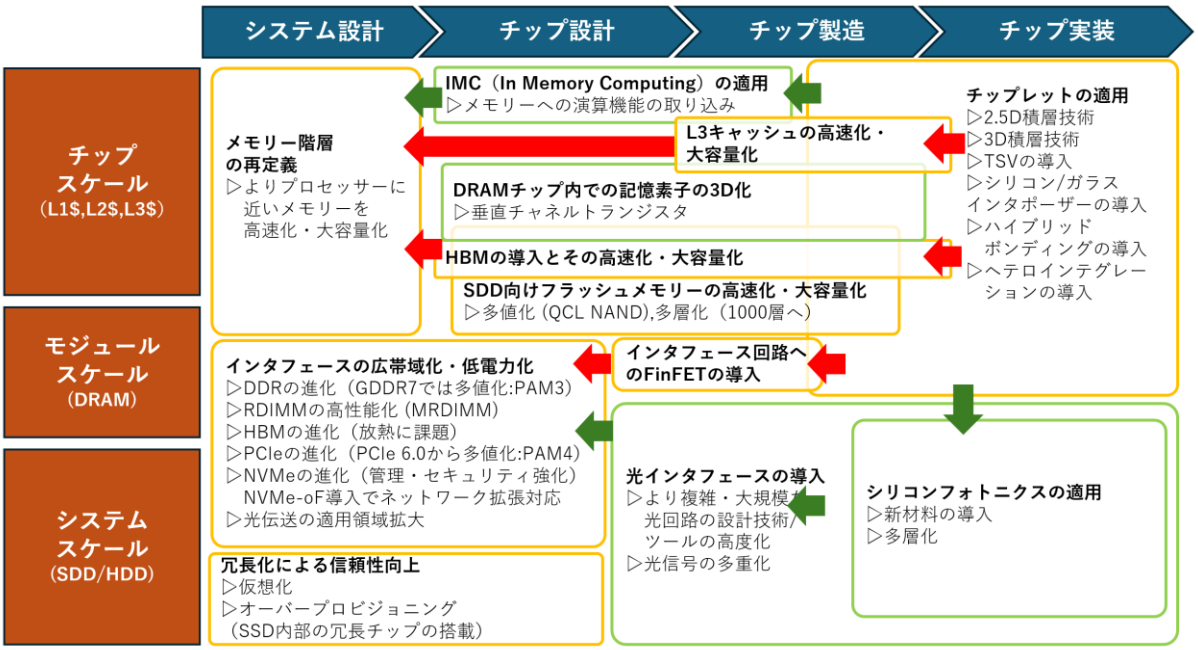


図 11　メモリ/ストレージ開発のポイント

出所：日経 BP 総合研究所

●チップレットの適用に基づく、メモリ階層の再定義

チップレットの適用によって、CPU や AI アクセラレータでは、チップの大規模化やより効果的で効率的な機能集積が可能になった。同様の効果は、プロセッサに集積されていたキャッシュメモリ（SRAM）や DRAM、フラッシュメモリにも及ぶ。既に、チップレットを活用して、これまで実現できなかった速度・帯域幅・容量のメモリを実現して、データセンター向け CPU や AI アクセラレータと合わせて利用する例が増えてきている。こうした動きの狙いはほぼ同一であり、プロセッサとメモリの間のデータ転送において高速化・低消費電力化・転送頻度の最小化し、ノイマン型アーキテクチャ固有の課題であるバスでの性能向上と電力削減のボトルネックを解消することにある。そして、メモリシステムを構成する要素にチップレットを適用することによって、メモリ階層の再定義が起こりつつある（図 12）。

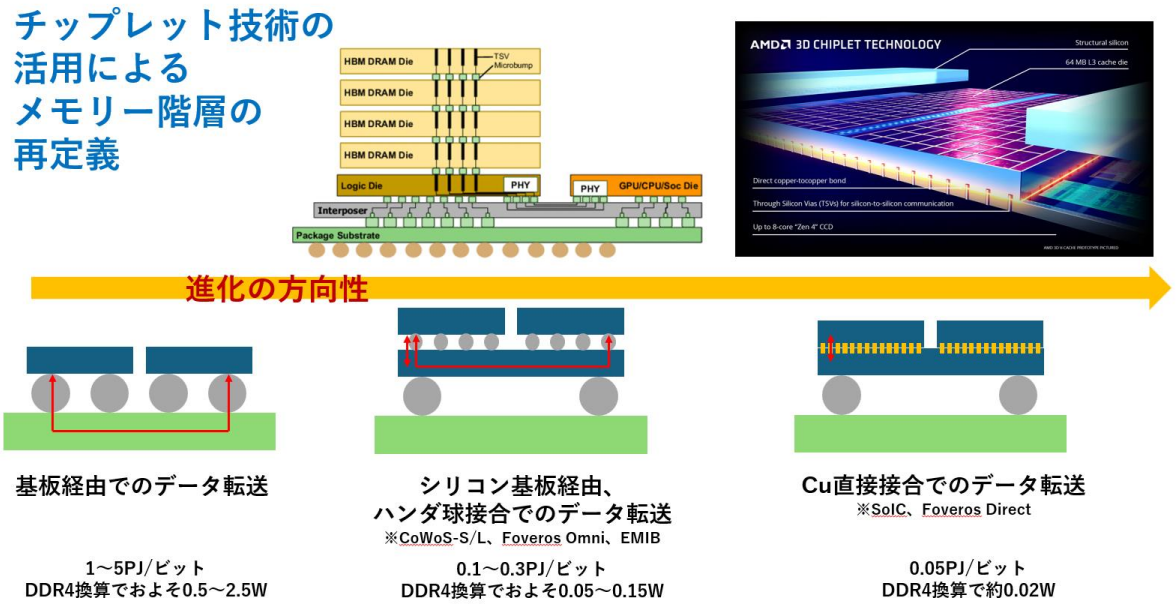


図 12　チップレット技術の活用で、データ転送時の性能・消費電力のボトルネックを軽減
出所：日経 BP 総合研究所

ここでは、チップレットをメモリに適用して効果をあげている代表例である、L3 キャッシュの高速･大容量化と､主に AI アクセラレータで活用されるようになった HBM(High Bandwidth Memory)における動向をまとめる。

近年、微細加工技術を進化させても、キャッシュメモリ（SRAM）の密度が、ロジック回路よりも高めにくい傾向が見られるようになった。このため、性能向上に向けて求められる大容量のキャッシュメモリを搭載すると、思いのほかチップ面積が大きくなり、同時に歩留まりが低下してしまう状況が生じてきている。チップレットを活用すれば、キャッシュメモリの部分を、成熟した前世代プロセスで製造し、最先端プロセスで製造したロジック部分と集積することが可能になる。近年では、チップ間を抵抗値の高いハンダバンプではなく、金属配線を直接接続するハイブリッドボンディングが実用化している。このため、ロジックとメモリを 1 チップ化した場合と遜色のない高速・高密度な接続が可能になってきている。

AMD は、チップレット技術に基づく「V-Cache」と呼ぶ 3D 積層キャッシュメモリ技術を同社のサーバ用 CPU「EPYC」に導入し、大容量の L3 キャッシュを搭載した製品を市場投入した。この製品では、プロセッサダイの上に前世代のプロセスで製造した L3 キャッシュを積層し、ハイブリッドボンディングで金属配線同士を直接接続してプロセッサとメモリ間の配線長を最短化している。例えば、2022 年に市場投入した第 3 世代 EPYC では、V-Cache を適用しない場合の L3 キャッシュの容量は CPU1 個当たり 256MB だったが、V-Cache を適用することで 768MB と 3 倍に増やした。L3 キャッシュを大容量化することで、特にメモリ集約型のアプリケーションでキャッシュヒット率が向上し、外部の DRAM へのアクセス頻度を大幅に減少させて、性能向上と消費電力の削減が実現した。ただし、V-Cache は万能の高性能・低消費電力化技術ではない。主に、EDA ツールや有限要素法を利用するシミュレーション、3D モデリングとレンダリングなどのキャッシュヒット率が高

まるアプリケーションで効果を発揮する。AMD は、キャッシュをプロセッサチップの下に配置して熱抵抗を削減し、冷却効率を向上させる第 2 世代 V-Cache も投入しており、この技術をさらに改善していく方向である。

HBM は、TSV（Through silicon via）を形成した DRAM チップを薄化し、3 次元積層した DRAM である。AI アクセラレータを中心に広く活用されるようになった。シリコンインタポーザーを介してプロセッサと接続することで、従来 DRAM よりも広いデータバスと高密度配線が実現できる。AI 関連処理や科学計算のような高頻度でメモリとプロセッサ間のデータ転送が発生するアプリケーションの高速化に効果的な技術である。2013 年に SK ハイニックスと AMD が共同で最初の HBM(1GB、4 積層、128GB/s、1.2V 駆動）を開発して以来、帯域幅を高め、積層枚数を増やし、容量を増大させ、駆動電圧を下げて低消費電力化を図る方向へと規格を更新し続けている。2026 年には、第 4 世代の「HBM4」(48GB、16 積層、1650GB/s、0.8V 駆動）に対応した製品が市場投入される見込みである。HBM は、JEDEC によって「JESD235（HBM)」および「JESD238（HBM3)」という規格で標準化されている。

SK ハイニックスは、既に HBM5 対応のチップの開発をしており、エヌビディアに供給することを明らかにしている。その一方で、米マイクロン・テクノロジーは、HBM3 および HBM4 を進化させた、「HBMNext」と呼ぶ容量を 36GB～64GB に増やし、帯域幅を 1.5～2TB/s に高めた独自仕様のチップを開発。より大規模なデータセットの高効率な処理を実現する製品を 2026 年に投入する計画である。HBM は、チップを 3D 積層して、AI のようにデータのアクセスが頻繁に発生する高負荷アプリケーションで利用される。このため、規格の進化と同時に、冷却技術の高度化が必須になる。この点も、今後の技術開発の焦点になってくる。

●難易度の高い技術が必須になるメモリ/ストレージのインタフェース高速化

メモリ/ストレージでは、データ転送の高速化・大容量化に向けて、チップ、モジュール、システムそれぞれのスケールでインタフェースの高速化・低消費電力化に向けた技術開発が進められている（図 13)。ただし近年、既に高速化・大容量化が進み切り、従来技術の延長線上での改善では、ノイマン型アーキテクチャ固有のボトルネックが解消できなくなってきている。このため、これまでとは異なる発想からの技術の導入が求められるようになってきた。

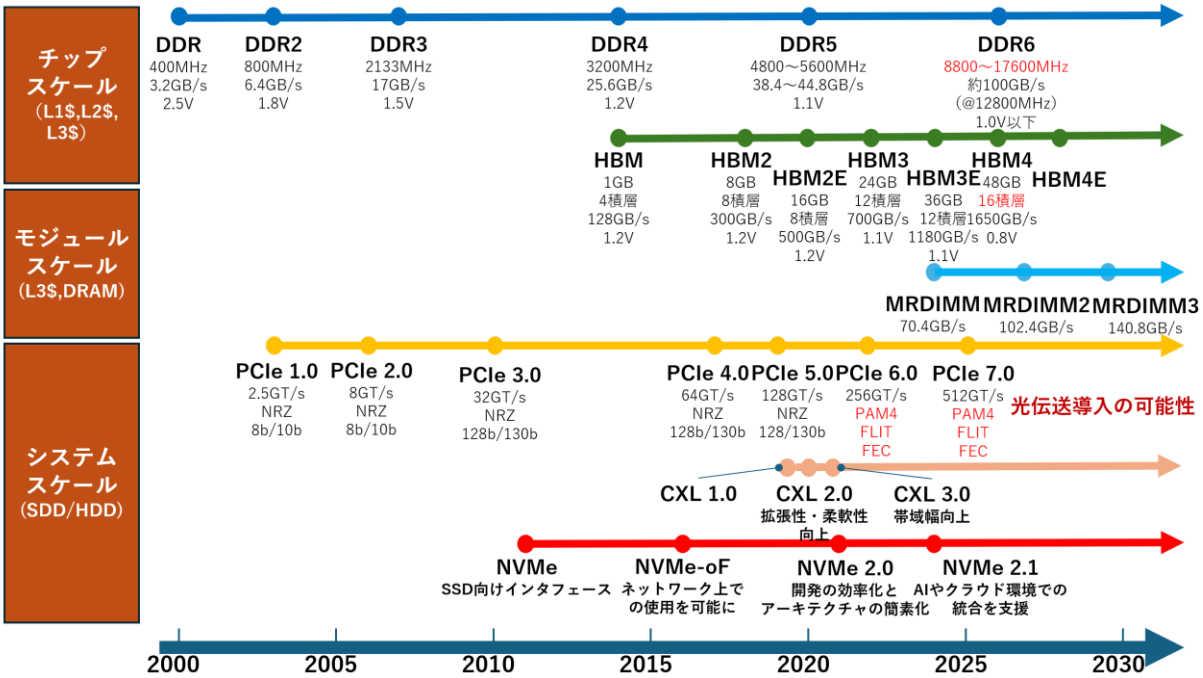


図 13　データセンター向けメモリ/ストレージに関連するインタフェース規格の進化
出所：日経 BP 総合研究所

DRAM のインタフェースでは、JEDEC によって標準化されている DDR の規格に準拠した製品として、2026 年には DDR6 対応品が投入される見込みである。これまで、動作周波数、帯域幅を向上させ続け、駆動電圧を低下させてきた。DDR6 以降の世代では、高速インタフェースを実現するために、インタフェース回路を構成するトランジスターとして FinFET を投入する必要性が生じてきている。その際、セル領域の回路のプロセスとは大きく異なるプロセスを導入が求められる。このため、チップレットを導入して、セル領域とインタフェース回路を別に製造して、後工程で組み立てる手法が標準的になりそうだ。ちなみに、AI 関連処理に向けたシステムでの DRAM のインタフェースでは、ユーザーからのカスタム仕様での対応が求められ、総じて標準規格よりも厳しい要件が出されることが多いという。このため、早い時期に量産レベルでのチップレットを導入した DRAM プロセスの確立が求められることになる可能性が高い。

なお、データセンター向けサーバでは、メモリーモジュールのインタフェースとして、通常の DIMM ではなく、ホストとの間にバッファーを挟み、なおかつ単一モジュール内に複数のランクを組み込んで容量と帯域幅を大幅改善する MRDIMM を利用するようになる可能性が高い。MRDIMM は、2025 年に市場投入される予定である。既に、対応する CPU も市場投入されている。データセンターでは、メモリインタフェースの高速化に対する需要が高いため、DDR6 対応の製品投入が手間取れば、その間、モジュールで MRDIMM を積極活用するという可能性があるという。HBM の動きに関しては既に言及しているため、ここでは割愛する。

システムスケールでは、SSD や HDD などストレージ向けインタフェースでは、PCI Express（PCIe）の広帯域化が進められている。規格自体は PCIe 6.0 まで策定されているが、データセンターで現在主に利用されているのは PCIe 5.0 に対応したストレージである。PCIe 6.0 以降には、信号方

式として、4 レベルのパルス振幅変調である PAM4（Pulse Amplitude Modulation 4）が採用された。従来の NRZ（Non Return to Zero）変調技術に比べて、同じボーレートで 2 倍のデータ転送速度を実現できる。その半面、信号レベルが密集しているため、ノイズ耐性が低く、エラー率が増加される可能性がある。このため、FEC（Forward Error Correction）技術が同時導入される。PAM4 は複雑な信号処理を必要とするため、受信側の回路がより複雑になるが、信号検出のための回路の複雑さは軽減される場合もある。

AI アプリケーションの処理能力向上を目指して、CPU とデバイス間での高速かつ低遅延の接続を実現する技術である CXL の利用が始まる。PCIe 5.0 の物理レイヤーを活用し、最大 64GB/s のデータ転送速度を実現。さらに CPU とデバイス間でキャッシュコヒーレンシーを維持することで、メモリアクセスの効率化を図り、複数のデバイスが共通のメモリ空間を利用できるようになる。加えて、異なるデバイス間でリソースを動的に再割り当てできるため、AI ワークロードの変化に柔軟に対応可能である。CXL を活用すれば、LLM の学習に必要な大容量メモリを効率的に提供することが可能になる。インテルやマイクロン・テクノロジーなどの企業が、CXL を利用したメモリを開発している。

データセンター向け SSD のインタフェースには、高速性と拡張性を改善する NVMe（Non-Volatile Memory Express）規格に対応したストレージが求められている。NVMe は PCIe ベースの SSD に最適化されたプロトコルの規格であり、既に NVMe 2.1 までバージョンが進んでいる。データセンターでは、冷却効率が高い EDSFF（Enterprise & Datacenter Storage Form Factor）規格に対応したフォームファクターで利用されている。NVMe over Fabrics（NVMe-oF）と呼ぶ、NVMe をネットワーク上で使用するための規格も用意されている。

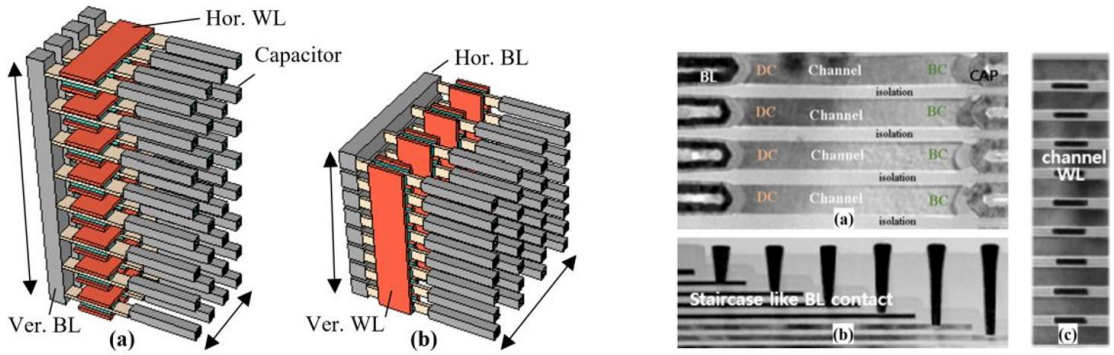
●今後 10 年間を見据えたメモリ/ストレージ開発のポイント

直近で見られるメモリ/ストレージの技術開発トレンドは、今後 10 年間を見据えた長期的視野から見た際にも同様に進む。ただし、チップ内部の設計・製造技術にしても、標準化されるインタフェース技術にしても、年々、実現技術の難易度が高まっている。これまでにも技術が継続的に進化してきたのは確かだが、これからは、2D から 3D への飛躍、モノリシックからチップレットへの移行、マルチレベルな信号の取り扱い、メモリとプロセッサの融合、電気信号から光信号への移行など、技術の質自体が変化していくことが見込まれる。特に、チップレットの活用とシリコンフォトニクスによる光インターコネクトの導入、さらには冷却技術での技術革新を追求する動きが活発化する可能性が高い。

特に大きな技術的変化が起きそうなのが DRAM の設計・製造技術である。SSD で利用されている NAND 型フラッシュメモリは、2 次元方向での微細化による高集積化が限界に到達。2007 年に 3 次元化する技術が開発され、2013 年に 24 層の 3D NAND が商品化された。その後、層数は着実に増加し、現在各社は 1000 層のチップを目指して技術開発を進めている。

その一方で、これまで 2 次元方向での微細化を推し進めてきた DRAM も、いよいよそのままでは高集積化が困難になってきた。現在の DRAM 技術は、最大 620 億個のセルを基板上に水平に配置しているため、トランジスターを高密度に配置すると電流干渉の回避が困難になってくる。サムスン電子の洞察によると、N+4（ロジックの 9nm ノードに相当）世代以降には、キャパシターの形成

が限界に達するとしている。セルを垂直に積み重ねれば、トランジスター間のギャップが広がり、干渉を減少させて、セル密度の向上が可能になる。そして、この技術を応用すれば、セルを垂直に積層して、単位面積あたりのメモリ容量を飛躍的に増加させることも可能になる。同社はこうしたコンセプトに基づく「Stacked DRAM」の開発を推し進め、2030 年の量産化を目指している（図 14）。既に同社は、Stacked DRAM のロードマップも公表。単位面積あたりの容量を 3 倍に高め、100 G ビット品が実現できるとしている。サムスン電子は、Stacked DRAM を実現するための基礎技術となる垂直チャネルトランジスター技術を導入した DRAM を 2025 年に市場投入する予定である。



Stacked DRAMの構造の候補　　開発段階での断面TEM写真

図 14　サムスン電子が開発している Stacked DRAM

3D DRAM には、次に示す二つの方式があることを Fig.4 で説明している。左図の（a）ビット線を垂直にする構造、（b）ワード線を垂直にする方法。短冊のように見える部分がキャパシター。右図の（a）チャネル付近の構造、（b）垂直ワード線における階段状の水平ビット線、（c）垂直ビット線構造のチャネルとワード線の積層構造。

出所：J.W. Han et al. (Samsung) “Ongoing Evolution of DRAM Scaling via Third Dimension-Vertically Stacked DRAM -”, 2023 Symposium on VLSI Technology and Circuits Digest of Technical Papers, TFS1-1.

④ネットワークデバイスに対する要求事項

IT 時代のインフラともいえるデータセンターが建設ラッシュを迎えている。DX の言葉に代表されるように、あらゆる人、物があらゆる場所からデジタルでつながる世界になってきており、データセンターの出番は増えるばかり。さらに拍車をかけたのが、AI 市場の急速な発展だ。特に 2022 年末に注目を集めた生成 AI は進化を続けており、多くの調査会社が、今後数年は AI 市場が CAGR30%以上で成長していくと予測している。当然、データセンターを支えるネットワーク機器にもこれまで以上の高機能化が求められる。

本調査では、データセンター向けネットワークデバイスの技術開発動向について調査し、現時点から 3 年後の 2028 年の短期的な動きと、10 年後の 2035 年の中長期的な動きに分別してまとめた。

●今後、3 年間で想定されるネットワークデバイス開発のポイント

直近、3 年間で想定されるデータセンター向けネットワークデバイス開発の動向のポイントをまとめた（図 15）。

図中、横方向には、情報処理システムの機能・性能を具現化・詳細化していく流れに沿って、具体的な技術開発の動きを「システム設計」「チップ設計」「チップ製造」「チップ実装」それぞれの段階に位置付けて示した。縦方向には、技術開発の動きを、適用対象となるシステム範囲を「チップスケール」「モジュールスケール」「ボードスケール」「ラックスケール」「データセンタースケール」の各領域に分類して示した。

図中のオレンジの枠内に示した動きが、直近 3 年間で想定される技術動向のポイントである。赤の矢印は、各技術動向間の依存関係を表している。直近、3 年間で想定されるデータセンター向けネットワーク機器開発の動向のポイントの中から、「広帯域幅への対応」、および「シリコンフォトニクスの適用」を起点とした、大きな技術革新の潮流について、詳細を示したい。

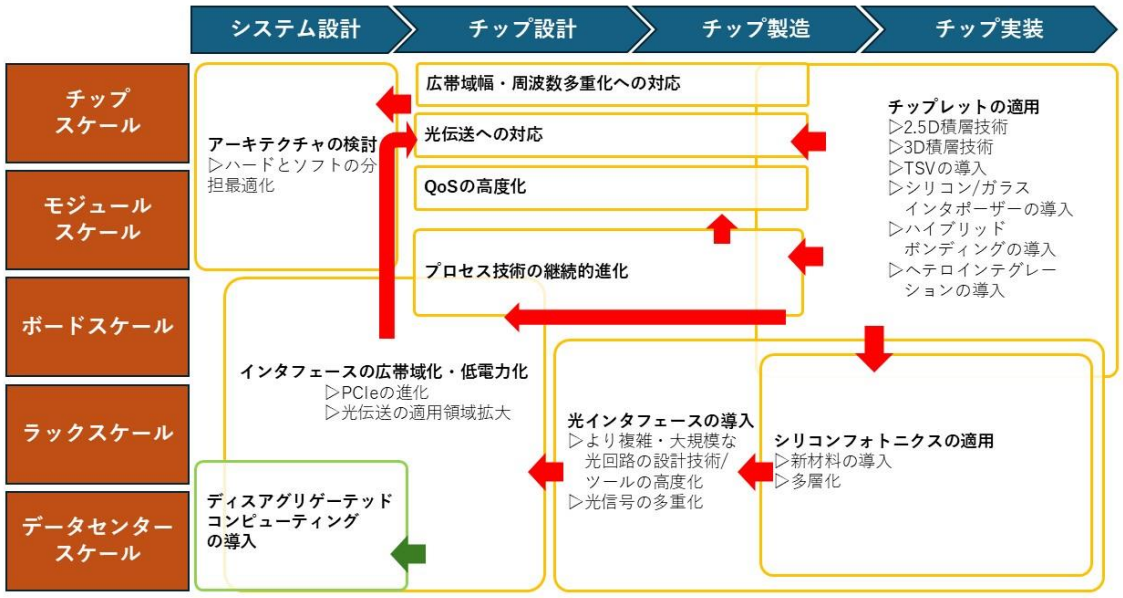


図 15　ネットワークデバイス開発のポイント

出所：日経 BP 総合研究所

●広帯域幅への対応を目指した技術革新

これまで、データセンターの技術革新は、主にはハードウェアのベンダーがリードしていた。ところが、現在はハイパースケーラーが主導するようになってきており、CPU やアクセラレータでも述べてきたように、特に生成 AI が出てきてからの数年は顕著になっている。AI 処理を中枢とするデータセンターにおいては、ベンダーの既製品では満足に至らず、処理が追いつかない状況になってきている。

ネットワーク機器などにおいてもその構図は同じ。生成 AI などから発生する多くの演算需要を

より高速で処理するためには、スイッチ IC 等のデバイスなどは購入するものの、ハイパースケーラー自らがスイッチなどの仕様を決める。自分たちのデータセンターが求められる用途に対して本当に効率的な処理を考え、既製品とはデバイスなどの配置が変わった構造のスイッチを製造し、利用している。このように。ハイパースケーラーのデータセンターでは、カスタムな世界になってきているのが現状だ。

では現状は、スイッチなどにどの程度の能力が求められているのか。ここでは現状を知るために、スイッチの能力の中枢を決めるスイッチ IC で、ハイパースケーラーに多大な影響力を持つ米ブロードコムの例を取り上げて説明する。ブロードコムはファブレス半導体メーカーで、サーバやネットワーク向けプロセッサ、ネットワークスイッチを主力にしており、特にネットワークスイッチで大きなシェアを持つ。最近では、米 Vmware（VM ウェア）買収の話題が、IT 業界を賑わしている。

そのブロードコムのスイッチ IC の主力製品が、「StrataXGS Tomahawk 5 シリーズ（BCM78900 ファミリー)」。5nm 世代の半導体プロセスで製造することで、1 チップで 51.2Tbps の帯域を実現した。これを使うと、64 ポートの 800Gbps スイッチや 128 ポートの 400Gbps スイッチ、256 ポートの 200Gbps スイッチを構成できることになる。

実は Tomahawk 5 シリーズが発売を開始した 2022 年 8 月時点では、多くの顧客はここまでの性能を必要としていなかった。ところが、生成 AI の普及によりデータセンターでのトランザクション処理が爆発的に増大。Tomahawk 5 シリーズの需要は一気に拡大し、業界で奪い合いとなった。より広帯域を実現するニーズは、ハイパースケーラーを中心に確実に拡大している。

加えて Tomahawk 5 シリーズには、今後のデータセンターに必要とされる機能を搭載していることもポイントだ。電気接続と光接続の双方に対応が可能。また、コグニティブルーティングや共有パケットバッファリング、プログラム可能なインバンドテレメトリー、ハードウエアベースのリンクフェイルオーバーといった機能を備え、AI や機械学習のワークロードのジョブ完了時間が最短化できる。

一方で、800Gbps の性能でもまだまだ AI 用途では足りないという声もある。GPU の進化は著しく、800Gbps のバックエンドネットワークでも、もはやその性能は十分に活かしきれていない。これまでのコンピュータシステムにおいては、ネットワーク帯域には比較的に余裕があり、ハードウェアの性能がそれを追っている状況だったが、今ではこの構図が逆転。800Gbps の製品が品薄のことを鑑みると、次の世代の 1.6Tbps に関しても、製品が出れば飛びつく向きは多くあるだろう。

ブロードコムが、3.2Tbps の帯域を持つ Tomahawk の最初のシリーズを発売したのが 2014 年。以来、ほぼ 2 年に 1 回の頻度で新シリーズを投入しており、102.4Tbps の帯域を持つ Tomahawk 6 シリーズもすでに発表されている。このような製品へのニーズは、当初はハイパースケーラーだけのものであろう。ただし、数年経てば量産効果も伴って、それらの技術が他の用途に降りてくると予測できる。

●シリコンフォトニクスの現在

もう 1 点、ここ 3 年のネットワーク機器で注目すべきは、シリコンフォトニクスである。シリコンフォトニクスとは、シリコンは波長が 1.2μm～8μm の光では透明であることを利用して、半

導体プロセスで培った微細加工技術を活用し、シリコン基板上に転送路や受光器、光変調器などの素子を形成する技術である。光ファイバー通信は超帯域である 1.3μm 帯（O バンド）、1.55μm 帯（C バンド）での転送が可能であり、光トランシーバーの基板やオンチップ光インターコネクトの作成への適用が期待されている。研究レベルでは、2000 年代初頭に登場し、2010 年代に急速に発達した技術領域である。現在、半導体チップの生産に利用されている 300mm ウエハーを使って、キッチリと設計されていれば、想定通りに機能する 1 万個レベルの大規模光回路を作ることが可能なまでに技術が確立されてきている。しかも、シリコンであるため、通常の電気回路を併せて作ることも可能だ。

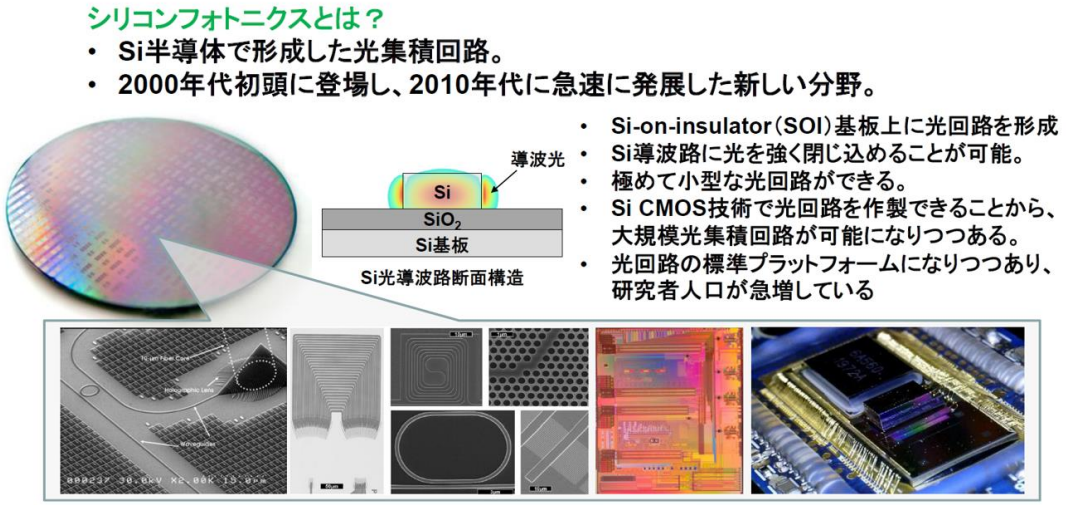


図 16　シリコンフォトニクスとは
出所：東京大学 竹中充教授

シリコンフォトニクスは、光電融合を実用化するための中核技術と目されている。従来の銅線を用いた電気によるデータ伝送は、周波数が高くなるほど消費電力が大きくなる。つまり、大容量のデータを送ろうとすれば、それだけ消費電力が増大する。膨大なデータの伝送が求められる中、このままではネットワークにおいてデータ伝送にかかる電力が激増するという状況が生まれている。そこで光電融合では、電気で信号をやり取りする部分の一部を光に置き換えることで、消費電力の大幅な削減や伝送速度の向上を狙う。

現在、光信号と電気信号を変換する装置として一般的なのは、プラガブル光トランシーバーだ。プラガブル光トランシーバーでは、プリント配線基板（PCB）上に光信号と電気信号を変換する光学エンジンを配置。IC チップと光エンジンの間は基板上の電気的な配線でつなぐ。AI 処理のようなものを必要としない場合は、プラガブル光トランシーバーでも問題ない。ただし、高速の転送が求められると、この配線が問題となる。配線が複数本必要となるので、スペースに制約ができてしまうことに加え、電気信号の遅さがせっかくの光電融合のメリットを消してしまう。

そこで、IC チップと光エンジンの距離を近づけようとする取り組みが進む。現在、実用化が進んでいるのが、Co-Packaged Optics（CPO）である。シリコンフォトニクスを用いて、光学エンジ

ンと IC をパッケージング基板にまとめて実装するために、両者の距離は極めて近い。CPO 製品に関しても、やはりブロードコムが先陣を切り、2024 年 3 月に「Bailly（ベイリー）」の市場投入を発表した。

現在は、商用も含めたシリコンフォトニクス向けの回路試作を請け負うサービスを提供する工場が世界中にできつつある。具体的には、シンガポールの AMF や米グローバルファウンドリーズなどがサービスを提供している。TSMC も、特定ユーザー向けに同様のサービスを行っている。日本では、産業技術総合研究所が R&D 用ファウンドリーの役割を担っている。加えて、イスラエルの Tower Semiconductor 傘下のタワーパートナーズ セミコンダクターが、旧パナソニック魚津工場で 2024 年からシリコンフォトニクスのファウンドリーサービスの提供を開始している。

なお、光回路の設計では、電気回路とは異なり曲線パターンが多くなり、しかも基本的にはアナログ回路であるため、既存の設計ツールが対応していない場合が多い。現在、電気回路向け半導体設計ツールの大手である、米ケイデンス・デザイン・システムズや米シノプシスなどが、フォトニクスも設計可能なツールの整備を開始している。今後、さらに光回路が複雑になれば、設計後の回路検証などが難しくなる。そうした作業をなるべく自動化すること、エレクトロニクスとの協調設計がより重要になってきている。光回路の設計固有の現象を考慮した設計手法の確立も必要になってくるが、AI などを活用することによって、作業を効率化できる見通しも見えてきている。

フォトニクスは、歴史的に日本が結構強い分野である。化合物半導体を使ったレーザーや、それに関連したフォトニクスの領域での競争力は高い。また、IOWN 構想を推し進める NTT が、技術開発を着実に進めているため、技術的な基盤は固い。

ただし、商用化の観点から見ると、決して世界の中で優位に立っているとはいい難い状況である。国内にシリコンフォトニクス適用製品の量産を担う記号・工場がないからだ。さらに、サーバやネットワーク機器など、ハードウェアで強いアプリケーションを持っている企業が少なくなってきている点も商用化が進みにくい要因になっている。例えば、エヌビディアは TSMC と連携して、シリコンフォトニクスの開発を進めているが、そこに日本が入り込む余地はあまりないのが現状である。

●今後 10 年間を見据えたネットワークデバイス開発のポイント

直近で見られるネットワークデバイスの技術開発トレンドは、今後 10 年間を見据えた長期的視野から見た際にも同様に進む。この先、ネットワークを流れるデータ量は拡大することはあれど、減少することは考えられない。考慮すべきはその拡大の勢い。三菱総合研究所によれば、2040 年のデータ量は 2020 年に比べて約 350 倍に拡大するとしている。果たして、この予測が正しいかどうかはわからないが、こういった点を鑑みれば、この先、ネットワークの帯域幅はより広いことが求められるのは明白だ。

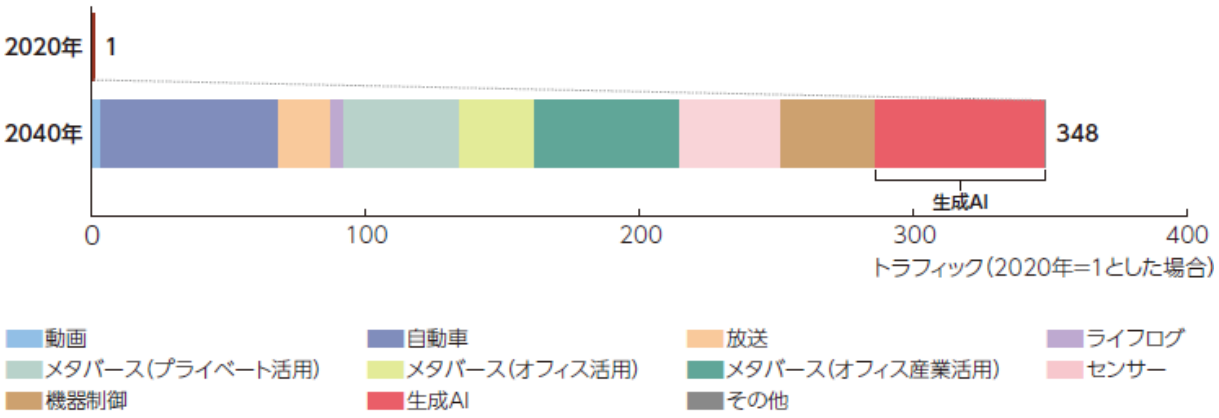


図 17　2020 年と 2040 年のデータ流通量の比較
出所：三菱総合研究所

こういった点を鑑みると、あらゆるシーンで活用が広がるのは光技術だ。そしてデータセンター間ネットワークとしては、ここまでも繰り返し述べているが、IWON の APN の広がりに期待が集まる。IWON で目指すのは「伝送容量 125 倍」に加えて「電力効率 100 倍」「遅延時間 200 分の 1」。また、APN を利用すればデータセンターを多対多で接続することも可能とする。当然、拡大するデータ量に対しては、それに対応するデータセンターの拡大も必須となるが、APN はデータセンターの分散化を間違いなく後押しすることになりそうだ。

こういった APN の拡大を受けて期待されるのが、リアルタイムの電力需要に応じたデータセンターの使い分けだ。2 月に閣議決定された「エネルギー基本計画」によれば、2040 年度には国内の電源構成において 4 割～5 割を再生可能エネルギーとすることを目標とする。すなわち太陽光発電や風力発電などの再エネが、2040 年度には主力電源化するということだ。

一方で、日本においては再エネの供給地にバラつきが見られる。北海道や九州では比較的再エネが豊富に供給できるのに対して、都心部での供給は少ない。需要と供給のバランスを考えるのであれば、電力需要が旺盛である都心部に再エネの供給があることが望ましいが、それがままならないという現実がある。現在、データセンターの分散化が叫ばれているのもこの文脈、すなわち莫大な電力を使用するデータセンターだから、その需要は再エネが賄う方がいい、という論理からなのだが、電力需要と同様に、データ処理に対する需要が少ない地方にデータセンターを建設することが、そもそも正しいのかという議論もある。

この課題を一気に解決する可能性があるのが APN だ。広帯域、低遅延のネットワークをもってすれば、もはやデータ処理をする場所はどこでも構わない。できる限り再エネを使いきるように地方のデータセンターでデータ処理をすることを前提にするといった運用が可能だ。また、出力が不安定であるという再エネの課題解決をも可能とする。要は、データセンターの処理をバッファーにできるという考えだ。今後、再エネの利用を拡大していくという観点において、APN は不可欠なインフラとなりそうだ。

V. データセンター業務経験者・ユーザーへのアンケート調査

調査目的

2028 年時点でデータセンターに対して大きな影響を与えるトピックスについて、そのトピックスがどのように進展し、どのようにデータセンターに影響を与えるか、さらにはそれらの影響を鑑みたうえで、データセンターを構成する要素にはどのような要件が求められるのかを分析する目的で、データセンター事業の経験者、およびデータセンターのユーザーからの意見を募る。

調査タイトル（回答者に提示した調査名）

データセンターの業界未来動向調査

調査方法

BtoB 向けナレッジプラットフォームを運営するビザスク（実名登録有識者数 56 万人超）の登録者の中から、データセンター事業の経験者、およびデータセンターのユーザーを抽出。その対象者に向けてインターネットを通じてアンケート調査。

有効回答数

データセンター事業の経験者：19 名

データセンターのユーザー：33 名

調査機関

ビザスク

実施期間

2025 年 1 月 19 日～2025 年 1 月 30 日

調査結果

①エッジ／クラウドの分担進展による影響について

アンケート回答者には、エッジ／クラウドの最適な分担は今後どのように変化し、そしてその変化はデータセンターにどのような影響を与えるのか、自由記述で考えを尋ねた。

回答の多くは、エッジデータセンターの広がりを予想している。理由の一つは、厳しいレイテンシーを求めるソリューションには近い場所で処理をすることが理想的であり、その役割をエッジデータセンターが担えるという点。代表的な回答がクルマの自動運転への適用だが、その他にも工場へのプロセス処理などの用途も迅速な応答が必要であり、エッジデータセンターが求められる領域だ。また、巨大な電力を必要とするデータセンターだけに、比較的再エネが入手しやすい地方に、エッジデータセンターを構える必要があるとの意見も多い。加えて災害対策の観点からも、バックアップ用途としてのエッジデータセンターは有効な解決策だ。もちろん、これらを実現するためには、ネットワークの進展がセットであると考えている。

一方で、エッジデータセンターの用途は限定的であると、多くの回答者は考えている。エッジデータセンターには計算リソースという点では多くを望めないため、高度な処理にはクラウドの中枢を成すデータセンターに頼らざるを得ないということがその根底にはある。また、エッジデータセンターを地方に分散させる懸念点として、人材不足を指摘する声もある。

セキュリティの観点からは、相異なる意見が寄せられた。中枢のデータセンターの方が機能を充実させられるため、秘匿性の高いデータの格納には適しているという意見。一方で、エッジデ

ータセンターの方が、限定的にセキュリティを高めやすいという意見もあった。もちろん、データセンターの使い方などによると思われるが、対局した意見は注目に値する。

特徴的だったコメントを以下に列挙する。

・エッジでは、データ取得/管理、リアルタイムで簡便な BI（ビジネスインテリジェンス）処理、決められたデータクレンジングと加工処理が限界であると推察する。クラウド環境側に求められるのは、大規模な解析であり、解析とは相関と因果の関係の創出である。エッジ側でデータに何らかの処理を行い、選別し、削ぎ落としてしまうと、もしかしたら因果関係があるかもしれない可能性を捨てることと同義となってしまう。エッジ分散処理により、多少のコンピュータリソースの消費は減れど、データ量は恐らく変わらず、大規模データ解析によるコンピュータリソースの多くはやはり、クラウド側に必要なままであるものと推察する。

・一次的にエッジコンピューティングの流れは出るが、最終的には中央に集中したほうがコスト効率/電力効率があがるため、クラウド集中型に収れんすると考えている。また、通信回線の大容量化/低遅延/低電力化については、NTT の IOWN 構想等により、解決される見込みである。

・現段階でのエッジ／クラウドの分担は、データの前処理（ノイズ除去や特徴量抽出）をエッジで実行したあとに、本処理以降をクラウドで実行する、という形がよく見られる。今後は、本処理の一部もエッジで実行されるような形が増えてくると考えられる。その結果、各地のデータセンターにエッジ処理を担う計算リソースを置かれるケースも出てくるであろうが、それを運用する担い手が不足していることもあり限定的だとみている。

・今後、事業分野ごとの最適な形が個別化していくと考える。ビッグデータの処理自体は、クラウド上でなければコンピューティングリソースが不足するため､継続してクラウドが利用される。一方、応答性はクラウドを使うほど遅くなることから、クラウド上の最新データをキャッシュとして保管しておき、IoT や GPS などのデバイスから収集したデータを基に結果を返すための計算を、キャッシュであるデータセンター内で処理する形に変わっていくことが予想される。その結果、グローバルデータベースからいかに高頻度かつ高速、低価格でキャッシュデータを同期しておけるか、というのがデータセンターの役割になると考えられる。インメモリ系のサービスとメジャークラウドへの大容量高速回線の接続状況がキモになってくる。

・これまでもコンピューティングは、集中と分散を繰り返しており、現在はクラウドへの集中から、エッジへの分散の過渡期にあると考えている。具体的には、処理、とくにリアルタイムないし揮発性のものはエッジ､一方で記録およびその全体分析が必要なものはクラウドとなることで、主にデータセンターでは記録を担うことからより堅牢性を求められるのではと考える。

・AI の進展に伴い、アウトプットを作るために参照する情報の量が爆発するとともに、アウトプットに至る時間を短縮するための事前学習の比率が高まっている。これらの情報は一般的に入手可能なものが多く、あらかじめ処理を走らせておくことが馴染みやすい。こうしたことから、これらの一般的な処理の格納、事前学習、アウトプットプロセスの一部まではクラウドで処理するという分担になる。他方で、個人情報や秘密情報、環境依存の情報などは事前に学習するインセンティブも低いし、処理量もさほど大きくない。そのため、こうした情報の格納や処理は、エッジで行われる。こうしたことから、データセンターではより大規模かつ汎用的な処理が求められ

るため、データセンターに対して、こうした処理が得意な回路やそのための応用ソフトウェア（ベクトルデータベースなど）を効率的に運用することが求められる。
・短時間に大量のデータが発生する製造工程などではエッジが進展し、現場での解析スピードが向上。この大量データをデータセンターに転送するには時間を要するため、サマリーデータがデータセンターに転送される。データセンターでは、他のデータと合わせた解析を担当する為、従来以上に大規模な解析に対応する必要が生じる。
・通常の基幹業務システムでも、今後 5 年は爆発的に AI 機能を取り入れることになると想定される。また、AI もより高度になり、クラウドでの処理ではレスポンス遅延が発生するため、エッジ/クラウドのニーズは高まると想定される。エッジサーバを拠点に設置することは、運用事業者のメンテナンスコストが上がるため、拠点のエッジサーバの保守をサポートするサービス（エッジサーバのバックアップ、バージョンアップ、監視機能等）が必要ではないかと考える。
・よりエッジでの処理を重視する方向に変化していくと思われる。その結果、データセンターに対してはデータ発生場所からの距離は問われず、むしろ再生エネルギー発電所の近くの立地等による環境配慮をしているかを重視する。
・必要なデータのみをクラウドに送り込む絞り込みは一部では実施しているものの、基本的にはほぼ全ての情報をクラウド上に上げている実態がある。そもそもデータ分析をかける前に、関連する情報が洗い出し切れているとは限らない。様々な情報から関連性を導く業務は今後も増えていくと想定され、まだまだ通信量は増加すると見込む。

②量子コンピューティングによる影響について
アンケート回答者には、量子コンピューティング技術が 2028 年時点のデータセンター全体にどのような影響を及ぼしていくのか、自由記述で考えを尋ねた。
多くは、量子コンピューティングの効果を評価するものの、2028 年時点ではデータセンターには影響をほとんど及ぼさない、もしくは影響を及ぼすとしても限定的であるという答えだった。量子コンピューティングの量子誤り訂正技術はまだ発展途上であること、安定的に稼働させるためには非常にデリケートな環境設定が必要であるために設備コストが高くなることなどの理由を上げる向きが多い。
また、量子コンピューティングをどうしても必要とする応用範囲が少ないことも、導入がそれほど進まないと考える理由の一つだ。導入を促進するための人材不足も課題である。また量子コンピューティングは計算を行うだけなら、古典的コンピューティングよりも少ないエネルギーで済む可能性があるものの、莫大な冷却装置など周辺機器を含めると、省エネにも貢献できる可能性はまだ少ない。本格的な普及は、2030 年以降になるという意見が多くで見られる。
特徴的だったコメントを以下に列挙する。
・量子コンピューティング技術による成果の一部は既存の方法でも同程度の成果が出せることが示されるなど、まだ発展途上の技術であるように見受けられる。2028 年ではまだ大きな影響を及ぼさないと考える。
・量子コンピュータはノイズを嫌う。この部分の改良がブレイクスルーになりうると考えると、データセンター建築方法に影響を与えると考えている。これまでは簡易な建物でも可能だったデ

ータセンターとは差別化した建屋が必要になる。また、コネクティビティについても、IOWN のような低遅延のネットワークが重要になってくる。
・基礎技術の発展はあるが、Application Framework が発展するかどうかの課題と、そもそも量子コンピュータによって解く必要のある問題設定に課題があるため、データセンターへの利用が進まないと考えている。キラーアプリが必要である。一瞬で解く必要がある問題で費用対効果が見られるものは、気象予測？地震予測？軌道計算？災害？宇宙？防衛？まだわかっていない。
・データハンドリングの最適化とエラー修正におけるアナリスト育成が充分と言えず、一般的な ICT コンサルティングとは異なるアプローチが必要となる。そのため多くのクライアントニーズを消化できる状態となるには、さらに数年の経験と知識の蓄積が必要と考える。
・直近 5 年程度で確かに実用化レベルに至ると言われている。ただし、その領域はあくまで研究領域での話であると考える。量子コンピュータリソースのクラウドサービス提供がある程度普及したときにそのリソース費用を投じてまでやりたい解析が、一般のビジネス用途（経営や事業分析）にまで落ちてくるかは少し疑問が残る。
・量子コンピューティング技術は、今後誤り耐性ありの方にシフトし、より複雑なものが短時間で解が導けるようになると思われる。ただデータセンターとしては特にゲート方式は巨大な冷却設備が必要だったり、そもそも大型なので、場所や電力の問題が残り続けるので、2028 年ではまだそこが課題だと思われる。
・主に AI 分野や分子シミュレーションなどの集約した計算能力の捌け口となる。またレガシーなコンピュータシステムと組み合わせたハイブリッドな計算環境を提供する事で、物流や金融といった分野での活用が進む。
・量子コンピューティングは、未だ開発段階にあり、最適なリスクを許容または分散した活用シナリオが確立されていないため、2028 年時点では従来通り、顧客のニーズに応じてデータセンターに導入されていく形態が維持されると考えている。2028 年頃から、デメリットを解消するソリューションではじめて実用化がエンタープライズレベルに落ちてくると思われるため、サービスとして大量の量子コンピュータを配置したデータセンターを構える投資も 2030～2032 年頃になると考えている。
・特定分野の処理能力を飛躍的に向上させるが、すべての処理を量子コンピュータで行うのは難しく､従来型とのハイブリッド運用が主流に｡量子エラー訂正技術の進展が実用化のカギとなり、新たなアルゴリズムや専門人材の育成が求められる。
・量子コンピューティング技術は、ハードそのものの貸し出しというよりは、パッケージングされたサービスとしての貸し出しといった汎用的な形態で一般ユーザーへ開放される形になっていくと考える。その結果、データセンター事業者側としては、ある意味どれだけのサービス量を自社で抱えられるかといったことが差別化要因になってくると考える。
・量子コンピューティング技術の発展は、従来のデータセンターの運用と設計において新たな機会と課題をもたらす。しかし、その全面的な実用化にはまだ多くの技術的課題が残されており、2028 年までには限定的な導入が最も現実的なシナリオとなる。
・現在のところ古典的なアルゴリズムなどの計算を早く行うことが苦手で、いわゆる進歩したアルゴリズムを記述できる新しい量子コンピュータの計算アーキテクチャが進展していく。その実

現には5年以上の時間がかかると思われるため、データセンター技術への影響はさほど出ないはずだが、一部AIの需要などが取り込まれたり、量子計算が元々得意なセキュリティに関わる分野の一部が量子コンピュータ側に取り込まれるといった影響があると考えられる。
・量子デコヒーレンス問題の部分的解決により、特定領域に特化した量子処理モジュールがデータセンター内に「協働プロセッサ」として組み込まれると考える。例えば最適化計算や暗号処理などの特定ワークロードをオンデマンドで量子ハードウェアにオフロードする形態が普及し、グーグルのTensor Processing Unitのような専用チップの進化形として、Quantum Processing Unitラックが導入され、従来サーバとのハイブリッドアーキテクチャが標準化すると考える。
・量完全な汎用量子コンピュータ（FTQC: Fault-Tolerant Quantum Computer）の実現には依然として時間がかかると思う。特定の用途に調律された「ハイブリッド・コンピューティング環境」がデータセンター内で形成される動きになるかなと推察する。

③その他のデータ利用（処理）の変化が及ぼす影響について

アンケート回答者には、エッジ／クラウドの分担進展、量子コンピューティング技術以外のデータ利用（処理）の変化が、2028年時点のデータセンター全体にどのような影響を及ぼしていくのか、自由記述で考えを尋ねた。

データ利用の変化という点では、まず避けて通れないのが、AI利用の動向だ。現在はテキストがメインではあるが、今後は画像を含む非構造データをAIで扱う頻度がより多くなってくる。データ量が増大する非構造データを効率よく処理、管理する仕組みがデータセンターにも必要になってくる。

また、データ利用の変化として、デジタルツインを挙げる回答も目立った。現実世界にあるモノの形状や状態、機能を仮想世界に再現するデジタルツインは、リアルのデータを元に社会課題を的確に解決できることから、「Society 5.0」においても意欲的に取り組まれている。このデジタルツインの広がりが急速に進むという前提で、データセンターにもその構築を支援する取り組みが求められるようになるという意見である。

特徴的だったコメントを以下に列挙する。

・生成 AI 技術の発展とともに非構造データの活用方法が増えている。非構造データを利用したいというニーズを満たすためにストレージ、計算リソースともにさらなる増強を求められるのではないかと考える。
・資産としてのデータが見直されたとき、自社の持ちモノとして保管する金庫としてのデータ保管をする場所としてデータセンターを利用するケースが増えてくるのではないかと考えている。
・経産省曰く、日本の経済を推し進める一つのドライバーはCPS（サイバーフィジカルシステム）だそうである。海外ではデジタルツインといったキーワードの方が通じることもあるそうだが、結局はリアルのデータをどう取得し、可視化し、解析して、未来を予測するか（何をどう作るべきか？未来に対してどう意志決定すべきか？）が、この辺りのテクノロジーに求められるものであると感じている。結局はアプリケーションが重要であり、コンピュータリソースはある程度技術が普及すれば規模の経済でコストを落とせたところが、ニッチを除き総取りとなると感じている。

・サイバーセキュリティ、情報セキュリティ、個人情報保護といった領域は、データセンター業務における敏感な対象であるため、データセンター事業者の態度如何では、歯牙にもかけられない対象になり得る。その際の水準や基準は、オープン系クラウド事業者となるため、彼らよりも高度化されたものを目指す姿勢が無いと判断された場合、撤退もあり得る。次に仮想化技術でのコスト要因を懸念しており、当該技術者の確保をデータセンター側技術者が、どのように認識するかで変化するとの見立てになる。
・量子コンピュータの発達により、現在主流の公開鍵暗号を破る可能性が出てくるため、ポスト量子暗号技術への移行が進む。特に暗号鍵を多く扱うデータはデータセンターで管理する方針となるため、セキュリティ戦略の変換と実装を求められ過大な投資ニーズが懸念される。
・これまで画像処理 AI では、簡単な画像処理をエッジ側ですることが多かったが、これからはフォトグラメトリのような高負荷の画像処理をクラウド側でリアルタイムで行うニーズが出てくると予想している。
・IoT デバイスに近い地域のデータセンターニーズが増えてきている。そのため、エッジコンピューティングを行うために小規模でも各所に点在するデータセンターを活用するニーズ（コンテナデータセンター）が増えてきている様に思われる。一方で、この傾向は 5G が普及しきれておらず、データの伝送が遅いことも理由の一つであるため、5G の普及がされた場合には、従来通りに地域毎のメガデータセンターに戻ると予想している。
・技術的な観点とは別に、デジタル化が進み、特に日本の様な小国の場合には、データセンター建設による住民の反発が各地方での悩みの種にもなりつつある。このことから、国として対策を進めていく可能性が高く、2028 年時点では普通のデータセンターで許容されていたとしても、2030 年頃には、市民のためにもなる体育館や図書館、カフェやランドリーなど、データセンター+α で計画された土地活用（地域に優しいデータセンター）でなければ、進出ができなくなってくることが予想される。データセンター事業者は、この点も視野に入れた戦略パートナーの確保が必要になってくると考えている。
・データ利用の変化として、「デジタルツイン・メタバースの浸透」が重要なトピックスになると考えられる。現実世界の精密な仮想モデル構築と、その上でのシミュレーション・分析ニーズが高まる。
・個人情報の取り扱いについては法令も含めて厳しくなっていく一方であるため、クラウドなどに上げる前の処理段階でマスク処理などの加工を行うことが必須になっていくと思われる。
・現在でもだが、複数のデータを統合、また不要データの削除など ETL と呼ばれる仕組みが今後とも重要になっていくと考えられる。データセンターやサーバなどはあくまでも箱であり、利用者側からすると、簡単に ETL などを扱えるか？が観点となり、ハードウェアだけの観点ではなく、ソフトウェアを含めたデリバリーを検討するが良いと思う。
・従来のデータセンターは構造化データが中心であったが、今後は非構造化データを利用した処理がますます増加すると予想できる。その為、従来とは桁違いのデータ量を高速に処理する設備への転換が必要になってくると考える。

・プライバシーの管理徹底が求められることから、エッジ機能はインターネットを介さない、データ発生元に近い場所での接続を求められて行くと考える。例えば、特定の県内で保有する製薬や育苗にかかるデータはエッジで県内の閉鎖網で処理したいなど。
・データの範囲がより個人情報に近ければ近いほど、利用価値が増えるのは明らかで購買履歴から医療データ、主義思想等のデータを扱う必要がある中で、匿名化や仮名化がここに対しての有効な対策になり得ていない現状で、不正アクセスにとどまらない新しいセキュリティの観点が出てくると思う。
・AI 主導のデータ処理の自律化が進んでいき、 データセンターの AI 最適化や、インフラの自律管理が入っていくと思う。また、データ主権と分散ストレージの進展が入っていき、ネットワーク構築が確実であることを念頭にして、データセンターの分散化が加速し、国ごとの規制に適応したストレージ戦略が必要になるとは感じる。業界的にも、プライバシー情報の担保と強靭なセキュリティが求められることから、分散化されていてもデータセンターのセキュリティ設計をより気にしていくと思っている。

④データセンター・ネットワーク構造の進化による影響について
アンケート回答者には、データセンターに関連するネットワーク構造は、2028 年時点でどのように進化し、データセンター全体にどのような影響と変化を及ぼしていくのか、自由記述で考えを尋ねた。
膨大なデータを短期間に処理することを今後はより一層強く求められるデータセンターには、デバイス間、サーバ間、データセンター間をつなぐ各ネットワークに、高速・低遅延で低消費電力であることが求められる。このような問題を、一気に解決し得る手段が電気信号ベースから光信号ベースへと情報伝送を置き換えることだ。この光を中心としたネットワークに期待する向きは多い。一方で、実際にインフラを変更することには、コストも伴うために、一気に光伝送を利用してネットワーク構造に変化するとも考えにくい。2028 年は、その黎明期と位置付ける記述が多くみられた。
特徴的だったコメントを以下に列挙する。
・3 年後であれば、APN の期待と現実が明確になる頃と思う。データセンター間では使われる可能性はあると思うが、データセンター内部への適用はデバイスが出来上がっていないので、3 年後では難しいと思う。データセンター間は、今でも光接続なので、利用者から見ればあまり違いは感じられないと思う。
・技術的にはネットワーク構造の進化は可能かもしれない。しかし、現在の利用方法の延長線上では、光伝送を必要とされるユースケースが少ないように感じている。グローバル企業でデータの同期を要求されるようなユースケースがどの程度あるかが不明と感じている。宇宙規模からの衛星データ転送などを送受信するケースなどがあれば要求されるかもしれない。データセンター内に関しては、コスト次第（投資対効果など）で導入可否が決定されると思っている。故障率、転送速度、構築費用など。
・今後センター間の速度、安定性、セキュリティを考慮すれば NTT 等のメガキャリア回線に対する期待とニーズが高まると考える。しかしながら光の道構想による全国的なファイバー通信網は

進展しており、自治体や道路、線路インフラ企業も回線を保有していることからキャリアに頼らないデータセンター・ネットワーク網の構築も可能な状態と考える。首都圏、大規模ユーザーに偏りがちなキャリア回線とは異なるこうした回線の相互接続によるデータセンター分散対応に期待している。
・光通信設備は既存の技術でも非常に高額である。APN 等も軍事用途での確立も完全ではないなか、28 年までに民間用として普及している可能性は極めて低いのではないかと推察する。
・よりソフトウェアで柔軟に制御ができるようになり、低遅延、高速化ができるように変化する。専用のアプライアンスよりも Linux 等のサーバ上でソフトウェア処理で行うようになり、ネットワーク機器は減少し、サーバの台数が増加する。
・利用者から見てデータセンター間の接続はより意識されないシームレスなものとなっていく（電気や水道のように）ことから、物理的には低遅延な光化、論理的には仮想化が前提となると考える。
・利用者側の視点で回答をすると、デバイスやサーバ、データセンターといった物理的なハードウェア単位で捉える時代は終わり、物理的なハードウェアは意識せずにサービスといったソフトウェアの単位で検討を行なっていくだろう。その際、物理的なハードウェアの制約がサービスのレスポンスに影響を与えないように、デバイス間、サーバ間、データセンター間でのネットワークの高速化・冗長化をしていくのは必須機能になると考えている。
・データセンター内のネットワークは、間違いなく光化すると考えているが、サーバ間をつなぐものは LAN 経由が続くと考える。現在でもカテゴリ 7 などの規格を活かしきっているシステム、サーバは少なく、今後活かしていくものと思う。
・光通信の導入により、エネルギー効率が向上するが、それに加えて、データセンター全体の冷却システムや電力供給も環境に優しい方法に切り替えられる動きが加速するだろう。光通信技術の全面的導入と、AI 処理など特定のタスクに特化したサーバと汎用処理を担うサーバとの役割分担、そしてこれら変化を支える SDN の本格導入により、異なる種類のトラフィックを効率良く処理できるように進化していくと思われる。
・データセンターのネットワーク構造は、単一障害点を持たない形で冗長性を進展させ、ゼロトラストな考えをベースとしたアーキテクチャへの移行が進んでいくと考えられる。その結果として、Starlink のようなモバイル回線などを組み合わせたファシリティに依存しないネットワーク構造や、NTT などの事業者が提供する新しいマネージド機器の導入や提供が行われてくる可能性があると考えている。
・IOWN APN の発展が進むことで、距離と通信遅延の関係が変化し、遠距離での処理に対する遅延の懸念が最小限になると予想される。その場合、場所の確保が難しい都心のデータセンターから地方の大型データセンターに需要が移りデータ格納場所の自由度が上がると考えられる。
・データセンターに関連するネットワーク構造は、利用者がコストを抑えられるよう、ネットワーク・ポートや帯域などを柔軟にスケールアウトできる仕組みが求められる点を想定する。
・今後ネットワーク装置の仮想化の進展を見せ、ネットワーク装置のソフトウェアとハードウェアの分離が進む方針に変化していくと思われる。その結果、データセンターに対しても、NOS 導入を求められて行くといった影響を与える。

⑤リソース共有機構の導入による影響について

アンケート回答者には、リソースを共有する仕組みは 2028 年時点でどのように進化し、データセンター全体にどのような影響と変化を及ぼしていくのか、自由記述で考えを尋ねた。

CPU やメモリ、ストレージなどのリソースは、情報システム内で複数のユーザーやアプリケーションが自由に利用できることが望ましい。仮想化やコンテナ化などの技術によって、各種リソースを効率的に配分・管理することは現状でも行われている。このため、多くの回答では、今後さらにコスト高が予想されるデータセンターにおいて、ユーザーメリットが大きいことから、リソースを共有する仕組みはさらに広範に普及すると考えている。

一方で、リソースを共有することに対して、セキュリティの観点から好ましいと思わない層もいる。こういった層にも十分に受け入れられるセキュリティ技術などがセットになって技術進化することが、リソース共有機構の導入を一般化させるポイントともいえる。

特徴的だったコメントを以下に列挙する。

・CPU やメモリ、ストレージなどのリソースの共有は長年研究されていて、2028 年時点では、仮想化とコンテナ化が主流になるであろう。全体最適ではあるが、個々の利用ユーザー、企業からみれば最適とはいえないケースもあり、AI を利用して利用実績を分析して、よりユーザー側のビジネスに沿った最適化がもとめられる。

・今のところは「Kubernetes」を中心とした、コンテナベースでリソースを管理するソフトウェアが今後も発展すると考えられる。業界にもよるが、動かしているソフトウェアはほぼすべてコンテナベースの技術で管理しているということになるかもしれない。データセンターはコンテナベースのリソース最適化への特化が求められてくるのではないかと思われる。

・AI による最適なリソース配分などが可能になると考えられる。データセンターでは CPU、メモリ、ストレージの最適配分を考えられるような設備設計が必要になる。

・低遅延のネットワークによって仮想化の仕組みがより高度化することができると考えている。現在の仕組みは遅延による影響によって十分にその機能を活かせていないケースが多く、その部分において改善されてくると考える。個人的には共有する仕組みでの本当の課題は認証認可の仕組みの必要性が高いと考えている。

・リソース共有をする仕組み自体はすでに普及しているが、さらに現在エッジコンピューティングでも求められているような、Security を担保した Sandbox 化が求められる。SaaS 企業などが Noisy Neighbor で苦労しているように再分配の仕組みは民主化・解放される必要がある。

・データセンターでの保管データが極めて保護されるべき個人情報、極秘情報となる傾向は高まり、その処理を分散共有リソースで処理することを可とするユーザーが増加するとは考えにくい。完全なシンクライアント構築とデータ通信増加のリスクは避けられず共有化には否定的なクライアントが多いのではないかと考える。極めてシンプルな 1to1 ネットワークのような構成による業務処理、保管体制を持つことができれば、そこからのシンクライアントによる分散処理体制なら受け入れられると考える。

・リソースを共有する仕組みは、大容量になる一方なのでよりソフトウェアで仮想化され、プール化されたり、逆に I/O をソフトからダイレクトにハードへパススルー仕組みへ移行すると思わ

れる。そういった場合、隠蔽された部分が増えるのでセキュリティなどで問題が発生するので注意が必要となる。
・リソースを共有する仕組みは、常に学習タスクや推論タスクを抱えている AI 事業者にとっては待望されるものであるが、そのようなタスクを抱える事業者数は限られると思われる。その結果、データセンターとしてはリソース共有機構の導入はそれほど進まないと予想している。
・ソフトウェア化が非常に進んでいることから、コンピューティングのリソースが大半を占め、データは局所集中型のデータセンター内構造になっていくことが予想される。そうなると、データセンターは最も熱を発するコンピューティングリソースは専用のフロアーなどを設けて冷却効率を上げる必要が出てくるため、データセンター内のフロアレイアウトや熱効率の良い役割型の部屋を準備していく必要性が増すと考えている。
・サーバ機能のディスアグリゲーションは開発は進むものの 2028 年での実用化は困難ではないかと考える。理由は光電融合デバイスの実用化の開発、導入にはまだ課題が大きいと考えるからだ。
・AI 専用チップやメモリを含むあらゆるリソースがプール化され、ワークロードに応じて動的に割り当てられる構成が主流となる。これにより、リソース利用効率が 4%程度向上すると予想される。異なる性能特性を持つプロセッサやメモリを、タスクの特性に応じて最適に組み合わせる自動調整機能が実現する。
・現在は管理者によってリソースの割り振りがなされることが多く、自動で割り当てする機能を行うと請求が増えたり、全体が重くなったり弊害を産むため、各ユーザーのリソース消費を最適化するツールと合わせて展開されていくべきと思う。
・リソースを共有する仕組みは、今後仮想化が技術の発展により、物理的な機器への依存が無くなり、プールされた能力から効率良くリソースを分配する仕組みが取られると考えられる。その結果、データセンターに対しては仮想化されたリソース群を持つ機器保管場所、または仮想化出来ない専用装置を置くための場所としてのニーズが高まると考えられる。
・リソースを共有する仕組みは、今後光結合型ディスアグリゲーションと AI 駆動の動的配分が進展し、リソースプールの遅延 1ms 未満・高利用率を実現していくと思われる。これにより、データセンターは需要変動に即応する「自己組織化リソースクラウド」へ進化し、投資効率が向上する。
・政府機関や銀行などの機密性の高い情報を取り扱う業態は引き続きプライベートクラウドなど、ハード面でも独立した環境の利用を要求されると思われるが、それ以外の業態では仮想化された環境を利用することは当たり前となってくると思われる。データセンターとしてもパブリック・プライベートの両方の環境を柔軟に提供できることが必須となってくる。
・現在はサーバ単位の仮想化技術が中心だが、2030 年頃には Composable Disaggregated Infrastructure（CDI）技術が普及する。データセンターはサーバプールではなく、CPU、GPU、メモリといった単位のリソースプールに変化する。
・おそらく、仮想化・ディスアグリゲーションを超えたコンポーザブルインフラとして進化するとは思っている。必要なリソースを柔軟に割り当てられる環境が機械レベルからデータセンターそれ自体に売りになっていくということもわかってはいる。だが、リソース共有が進むことで、

複数のユーザーが同じ物理インフラを利用する機会が増えるため、マルチテナント環境のセキュリティ強化の可視化が不可欠になる。コンフィデンシャルコンピューティングなどをどう見せていくか？がユーザーとしては必要になってくると思っている。

⑥その他のデータセンター・アーキテクチャの変化が及ぼす影響について

アンケート回答者には、データセンター・ネットワーク構造の進化、リソース共有機構の導入以外のデータセンター・アーキテクチャの変化が、2028 年時点のデータセンター全体にどのような影響を及ぼしていくのか、自由記述で考えを尋ねた。

非常に大きな話として挙げられるのは、データセンター自身の専用化だ。今、AI 処理を得意とする AI データセンターの建設に関する話題がかまびすしいが、これも専用化の一つである。様々な用途においてデータセンターのニーズが高まるとしたら、用途に応じて特化されたデータセンターが構築される可能性がある。用途ごとに必要とされる機能を実現するアーキテクチャを備えたデータセンターが構築される。

特徴的だったコメントを以下に列挙する。

・データセンター・アーキテクチャは従来、データセンターの物理的な要素となる、サーバやネットワークなどをもとに設計されてきたが、今後は利用されるソフトウェアを定義したインフラ設計（SDI)が重要となる。このデータセンターは、〇〇のソフトを利用する場合には便利、あるいは効率が良いなどとの色分けがでてくると考える。

・AI によるデータセンターの全体制御が進み、電力量のコントロールやリソース配分などを最適化する技術が現れると考える。

・今後のデータセンターは専用化していくものと考えている。データ保管用、AI 学習用、認証認可など高セキュリティなど。すべてに共通して必要なのは低遅延のネットワークサービスで、このネットワークをどう組み合わさるかでアーキテクチャは変わってくる。

・Developer Platform の進化を感じていて、より Nocode で Copilot 的にシステムを構築できる世界の中で、データセンターリソースについてもオンデマンドで利用可能な状況が得られたり、コラボレーションした上でリソースを共有・共用できることが求められるかもしれない。

・例えば、CDN の Akamai では、詳細なサービス仕様は公開されないように契約で縛っており、実態の技術的優位性は、先行者であるというブランドにぶら下がっているため、データセンター間の連携が中心テーマになると考えられるが、GDPR、経済安保、個人情報保護法等規制との折り合いのつけ方になる。データセンター事業者が、どこまでグローバルな利害関係者から受ける規制に対しての認識となる。

・リソース共有型の進歩が進むと、データの保護などはソフトウェア側の役割になり、ハードウェアは如何に効率的に機械の故障を検知して、迅速に交換できるか、という点に注目される様になるため、データセンターに配置した機器類の監視や、交換対象にのみアクセス可能な貸金庫型ラックなどの整備や、人による管理を不要にした極論で言えば配達員が自由に軽快に交換するくらいの無人化を行わないと、維持費を下げることができなくなってくると予想される。

・「モジュール型データセンターの標準化」が重要なトピックスとなると考えられる。標準化された機能ブロックを組み合わせてデータセンターを構築する手法が主流となると予想される。

・データセンターのアーキテクチャで言えば、機能レベルでの部品化・冗長化がどれだけ対応できるかが、一つのポイントになってくると考える。ハード部品は壊れるモノである前提を踏まえると、簡単に交換でき、冗長化が図れる仕組みが求められると考える。
・ネットワーク回線の共通化、共有化、イントラネット、専用線のクラウド化は間違いなく進む。このため、データセンターで最も大きな変化は、ラックごと、ベンダーごとに引き込んでいたネットワーク回線が共有化されることと考えている。また、それを期待している。
・動画ファイルなどの大容量のデータは、今日でもメディアに書き込んでアップロードするほうがネットワークを利用するよりも速い場合があり、インターネットで高速にマスタデータを安価に伝送する仕組みができるとイノベーションに繋がると思う。
・データセンター・アーキテクチャの変化として、エッジコンピューティングは処理を分散させ、低遅延を実現しつつ中央データセンターの最適化を促進すると思われる。また、ハイパーコンバージドインフラはソフトウェア定義でリソースを統合し、物理構成の簡素化と動的スケーリングを可能とする。AI による自動運用で効率化とダウンタイム削減を推進していくと思われる。これらの変化は、データセンターを「自律的で柔軟なプラットフォーム」へ進化させていくと思う。
・データセンターを必要とする利用企業は、セキュリティ要件が厳しい利用者が多くなると考える。従って、高セキュリティを担保するアーキテクチャが重要になる。これまで、データセンターは災害対策に重点が置かれていたが、今後は、サイバーセキュリティの堅牢性も重要な要素と考える。
・データセンター・アーキテクチャの変化という視点で、従来の進化となるならば、ニューロモルフィック・コンピューティングの導入で AI 処理の超低消費電力化・リアルタイム推論の最適化がおこるのと、DNA ストレージによるアーカイブが今までとは変わると思っている。特に、長期保存ストレージの概念を根本的に変え、低コストでのデータ保持を可能にできる点は、管理上要検討項目。もう少し大きな目線で、保管やアウトソースされたファシリティというよりは、超巨大な AI 的な脳機能の実現のようなイメージを考えている。

⑦レジリエンス要求が及ぼす影響について

アンケート回答者には、危機に対するレジリエンス要求は、2028 年時点でどのように変化し、データセンターにはどのような変化と要求が及ぼしていくのか、自由記述で考えを尋ねた。

レジリエンスと一言でいっても、その種類は数多くある。自然災害、サイバー攻撃、電力供給の障害、ネットワーク障害、人的エラーなどリスクの数は増えるばかり。多くの回答者は、このような多様化するリスクに対して、データセンターはレジリエンスをより高める必要があるとした。特に地震のリスクは、日本人独特の悩みであろう。レジリエンスの対策として有効な手段として、自動化を挙げる声も多くあった。より高度化する AI が、レジリエンス要求にも威力を発揮するという意見だ。

特徴的だったコメントを以下に列挙する。

・データセンターは 24 時間、閉じられた空間ということから、コロナなどが一度、蔓延したときのリスクが高い。地形（地震が少ない）、電力供給などが立地条件であったが、感染時の人の手配

の課題をとりあつかう必要がある。また、グローバルに分散した場合の、カントリーリスクにも備える必要がある。
・人件費の高騰、各データセンターが管理するリソースの増強を背景に、レジリエンスへの対応についても自動化が求められるのではないかと考える。
・グローバルでビジネスをしている企業にとっては、さまざまなリスクが発生する。特に国を跨いだシステムを管理するデータセンターでは、自然災害やサイバー攻撃へのレジリエンスのみならず、政治的な判断による遮断などに対するレジリエンスが求められると考えられる。
・災害対策としてのデータセンター分散化は既に多くのニーズが存在し、実施されています。攻撃を受けやすいネットワークのセキュリティ強化も急務となっており回線の冗長化やサイバーテロ対策はさらに強化を求められると考えます。また電源喪失対策としての太陽光発電装置併用などの対策も重要視されてくると考えます。
・AI の発達により、より自動制御が容易となり、レジリエンスを人がコントロールするのではなく、データセンターの機器が判断を下し、人間にアドバイスできるようになる。
・災害に対するレジリエンスは、BCP 観点でより求められ、分散化していくと思われます。環境に関して再生可能エネルギーの採用などでより地方などで分散していく。サイバー攻撃もより、トラブルが起こった際に障害範囲を狭めるという意味で、分散させたり、見える化できるマネジメントシステムの導入が進むと思われます。
・レジリエンスは単なる防護性の強化ではない。起こりうると想定される危機の発生管理と、想定以上の事象に対する回避システムが有機的に起動する仕組みが大事になる。データセンターに要求される自然災害への対応能力強化は必要だが一点集約して無尽蔵には増やせない。2028 年段階では負荷分散と BCM で乗り切る。
・2028 年では、災害によるレジリエンスが最も求められる状況が維持されると思われます（サイバー攻撃はソフトの世界なため、自動化が最重要のハード面での防御策になると思います）。その上で、環境変化の面では、技術の進歩速度が確実に上がっていることから、設備面においても、今後は解体、再興などを容易にするために、ビルディングブロック型、モジュール型のデータセンターが求められる様になってくると考えています。
・マルチクラウド環境を前提とした分散型アーキテクチャが標準となり、地政学的リスクや自然災害への耐性が強化される。その際、データ主権に配慮した地域分散が重視される。AI を活用した自律型防御システムにより、複雑化するサイバー攻撃への即応性が向上する。特に量子暗号通信の実装が進み、セキュリティ基盤が強化される。再生可能エネルギーの大規模導入と電力調達の多様化により、エネルギー供給の不安定性に対する耐性が向上する。
・データセンターという物理的な地理的制約がある中での話で言えば、地震大国である日本国においては災害に対するレジリエンスは多く求められていくだろう。首都圏直下型地震や南海トラフ地震を想定すれば、少なくとも東西拠点でのバックアップやデータセンター単位での冗長化、この機能装備は必須となってくるのかもしれない。
・回線やハードウェアの冗長化は、オンプレ時代は大企業だけが行う高コストなものであったが、現在はクラウドを使うことで安価に実現できる。そのため、2028 年には中小企業ふくめデータセンター内でのレジリエンスは一般化するだろう。

・データセンターのレジリエンス要求は気候変動による災害多様化で地理分散型エッジ配置が加速し、AI 予測制御で事前復旧が可能になったり、サイバー攻撃対策では量子耐性暗号の実装と自動修復 AI が標準化していくのではと考えます。
・自然災害やパンデミックに対するレジリエンシーはある程度対策されてきている。今後はセキュリティインシデントに対するレジリエンシーが更に重要になってくると考える。ランサムウェアなどのインシデントが発生した場合のリカバリータイムをいかに短くできるかが求められる。
・地球環境への変化に対する強靭性は引き続き求められているが、より多様化すると思う。例えば関東圏に位置するデータセンターで、火山噴火を想定したデータセンターがどれくらいあるのか疑問。ユーザー側がどこまで求めるのか、ユーザー側によるセンターの選別が始まる認識。
・場所としてはコストの安い僻地や後進国地域への移転が加速してくると考える。その際にカントリーリスクの低い安全な東欧やスイス、経済的にも厳しいカナダなどが自国産業化を狙ってくるのではないか、と想像する。
・2028 年には、災害・地政学リスク・サイバー攻撃へのレジリエンス要求がさらに強化される。分散型データセンター（エッジデータセンターや多地域冗長化）の導入が進み、事業継続計画対応が強化されると考えます。また、量子暗号・ゼロトラストセキュリティを活用した高度なサイバー防御が求められ、AI を活用したリアルタイム監視・自動防御が標準化する。これにより、データセンターはより柔軟で分散型のアーキテクチャへと進化すると考えます。
・2028 年時点での普及は難しいところはあるものの、データセンターのレジリエンスは、単なるバックアップや災害対応などの状況ではなくて、自律的に問題を予測、修復し、自己完結できるインフラへと進化する形に向かっていると思います。超分散化により、単一障害点を排除し、エネルギー自律性により、グリッド依存からの脱却も見られると思っています。これは日本国内では難しいので、どこか別の国で行われるサービスに目を向けると思っています。AI と量子暗号によるセキュリティ機能の強化からデータ保護も進んだ形のサービスを提案されると思っています。データセンターは様々な環境の変化に対応し、自己修復で存続できるレジリエント・データセンターへと進化することが求められると思っています。

⑧カーボンフットプリント要求が及ぼす影響について

アンケート回答者には、カーボンフットプリントへの取り組み要求は 2028 年時点でどのように変化し、データセンターにはどのような変化と要求を及ぼしていくのか、自由記述で考えを尋ねた。

カーボンフットプリントへの対応は、企業として強めていかなくてはならない一つの項目であることは、回答者全員の認識として一致しているだろう。このため、環境対応が新たな競争軸になるだろうという意見がある一方で、環境対応は表面的なものにとどまり、企業として利益を優先せざるを得ないという現実的な意見もあった。また、カーボンフットプリント要求を満たすための一つの手段として再生可能エネルギーの有効利用を挙げる声が多く、こうした声からもワット・ビット連携の議論が一層盛んになる可能性があると予測できる。

特徴的だったコメントを以下に列挙する。

・マイクロソフトへの電力共有のために、米国のスリーマイル原発が再稼働したことは重要な動きと思っており、このような動きが進むような気がします。一方、再生可能エネルギーも導入が進むと思います
・膨大な CO2 を排出している米国の大統領は環境問題への意識が薄いため、欧州やアジア諸国が中心となってカーボンフットプリントが主導されると考える。そのため、米国以外の国が持つ新技術などがデータセンターに反映される可能性がある。
・カーボンフットプリントに関しては、世界企業に求められる姿勢（ポーズ）だと感じている。必須というニュアンスがあるが本来はなるようにしかならないものであり、政治マターである。守る守らないではなく、実現可能か？達成可能か？を含めて検討すること。そしてそれを政府やロビー活動を通じて交渉していくものであると考えている。
・2028 時点でまだ実現していないかもしれないが、小型核融合技術を用いた発電が一つの解となると考える。利用者の多い都市部のデータセンターに電力供給しやすいからである。
・過去データセンター分散化へのニーズが高まりインターネット IX 構想に元づく高速ネットワーク網の構築が都市部以外の地域で多く導入されたが、現実的に実用化されている IX は少ない。しかしながら設備基盤は残存しているため、そこに再生可能エネルギー利用を付加する事で高度に再生した IX 構想エリアが出現する可能性はあると考えます。
・基本的に Scope2（消費電力による排出量)及び、Scope3 カテゴリ 15（調達商品・サービスに関する排出量）が関係し、データセンター全体の再エネ転換もしくは、データセンターにおけるリソース電力および、空調電力の低減に終息します。いずれにせよ極論、調達する電力ソースの転換と機器及び設備の冷却コストとなります。これらの要求はここまで確立されたカーボンニュートラルのデファクト化により、落ち着くことはないと思われます。それを考慮し、今すぐデータセンターに投資するのであれば、現時点の最適解は寒冷地によるデータセンター設立に落ち着くと思います。
・すでに大規模事業者の中には、「CO2 排出の実質ゼロ化」に向けた取り組みを始めている企業は存在する。2028 年時点ではこのような事業者が増えるとは予想される。また、利用者のデータセンターに対する要求として、カーボンフットプリントの可視化が盛り込まれる確率は確実に高まる。計算リソースの電力効率の向上や冷却効率の向上を各社が図るとみられるが、効果は限定的である。
・再生可能エネルギーの豊富な地方と、需要の多い都市部を高速光ネットワークで結ぶハイブリッド型データセンター構成が主流となる。AI による電力需給予測と連動した動的なワークロード分散により、再生可能エネルギーの変動に対応した運用が実現する。データセンターの環境性能が入札要件として一般化し、PUE に加え、CO2 排出量が重要な評価指標となる。
・現場の実務レベルでは、環境に対する興味は全くと言っていいほどなく、対外的に環境を意識したアピールをしているだけである。そのため、「カーボンフットプリント」を意識するのは、本質的なことではなく表面的なレベルに留まり、あくまでも利益の追求が最優先という現在の企業方針は変わらないだろう。
・IT システムには膨大な電力消費が必要となる。社会全体で考える必要があり、IT システムには従来の原子力や火力発電を利用し、一般家庭に再生可能エネルギーを使う手を考える必要がある。

・カーボンフットプリントの要求はさらに厳格化が進み、データセンターをグリーンエネルギー活用に対する取り組みは選定基準として、さらに要求度が高まると予想される。具体的にはデータセンターにおけるグリーンエネルギーの割合などが想定される。
・データセンターは炭素最適化が必須要件に進化し、再生可能エネルギー比率の高い地方拠点と都市近郊エッジ施設を AI で連携制御し、ワークロードに応じて CO2 排出量とレイテンシーを最適配分するハイブリッド運用が主流化していくと思われます。
・データセンターと利用企業とのロケーションは近傍である必要はなくなってきている。理由は通信費の低下と高速化であり、リモートで連絡ができれば既にデータセンターの地域国に拘る必要はないと考える。そのため環境配慮への意思決定がより柔軟なロケーション、エリアを選ぶことができると考える。
・環境負荷ベースのインフラから環境負荷を減らすインフラへと進化し、地球環境を再生する役割を担う時代へと進むと思いますが、どちらにしろコストに反映されて、我々としてはこの点の重視からコスト増への対応を急務としていくことになると思います。そうなると、各サービス会社から見せてもらえる評価軸と、そのデータの正しさをより証明してもらいたいと思っています。その正しさやそこにかけるコストまで反映するようならば、国を超えて別のクラウド的なモノに乗せ換えたほうがよく、正しいとは思っていませんが、規制の緩いところに移動しようという動きが活発化していくと思っています。

⑨その他のデータセンターの在り方の変化が及ぼす影響について

アンケート回答者には、レジリエンス要求、カーボンフットプリント要求以外のデータセンターの在り方の変化が、2028 年時点のデータセンター全体にどのような影響を及ぼしていくのか、自由記述で考えを尋ねた。

回答なしも多かった中、カーボンフットプリント要求にも重なる部分はあるが、地域への配慮を挙げる声が一部にあった。データセンターは、地域の雇用も創出するなどの点から、より分散化していった際には地域社会との共生が、求められる重要なファクタの一つとなるかもしれない。

特徴的だったコメントを以下に列挙する。

・より巨大なインフラポイントとなり、電気、水、ガスと同じように、情報という観点からの基盤になる。しかも、情報を集積して出し入れに対応していく基盤であるため、様々な業務やサービスなどが、より正しく行えるというインテグリティに対するマインドセットの革新が必要になる。ハードではなく、AI を使うとはいえ人間が管理するならば、その管理者を薄給で業務として従事させるようにすることは危険であり、そういう点ではコストとの兼ね合いやビジネスとしてのモデル構造がより変わっていく。
・地政学リスクやプライバシーに関する関心の高まりから、国内の処理は国内で完結させることが求められるようになると考えている。国内のデータセンターは安全性と効率の両方の追求を求められるかもしれない。
・製造業で主に話題になっているトレーサビリティはデータセンターにも影響するかもしれません。建設資材や消費材、データセンターで利用している電力が何に起因するかを明らかにして説明できるようにする、といった要求が発生する可能性はあります。

・コスト以外ない。技術の内容に置けるコスト有意性は、ハード領域からソフトへ流れるため、米国のように中性子爆弾にも対応可能なデータセンターと、日本のレベルの低い防犯カメラ装備のデータセンターでは自ずと比較する基準が違いすぎる。
・多くのユーザーが多種多様な AI 開発を進めていくと、データセンターのカスタマイズ自由度はますます重要になっていく。
・データセンター単一での事業から、環境配慮、地域配慮を伴うデータセンター事業へと変化していく事が予想されます。データセンターには警備が必要なため、地域警備との連携なども検討されている様です。
・データ主権、量子コンピューティング、メタバース、AI の進化など、様々な要因によってデータセンターは変化していくが、これらの変化に対応するため、地域分散化、ハイブリッド化、高性能化、自律化を進め、より柔軟かつ高度なインフラへと進化していくと予想される。
・日本において言えば、地方創生といった社会課題を解決する一助としてデータセンターの立地先として地方で検討され、そこで雇用の創出や環境影響への分散化が図られていくと良いと考える。
・企業が集中管理する必要がない情報を扱う場合は、ブロックチェーンを活用したプラットフォームを活用することも今後はあり得、その場合はデータセンターを利用しないため、一部需要は減る可能性がある。
・サイバー攻撃や自国での情報保護の観点が挙げられる。自国の情報は自国で守るということが、ハイパースケーラーの事業やデータセンター事業者のあり方を規定していく可能性がある。
・地方データセンターのニーズが高まる中、作業員の確保が困難になる為、運用を完全に AI と作業ロボットによる無人化されたデータセンターの構想が必要になると考えられる。
・各自治体での災害対策には限界がある一方、データセンターは非常に優れた災害対策を実施された建造物になります。顧客との調整は必要となりますが、災害時の拠点として活用できることも自治体より求められると予想しております。
・データセンターを運営していく上で、悪意のある従業員によるテロ的な行為で、データやサーバの消失が起きる事故も過去に発生している。そのため、どのように運用業務が統制されているかの開示なども要求されるようになると考える。

⑩データセンターを構成する CPU に対する要求事項について

アンケート回答者には、これまで挙げてきたような影響や変化をデータセンターが受けた際に、データセンターを構成する CPU にはどのようなことが要求されるかについて、自由記述で考えを尋ねた。

CPU に求められるのはまず高速化。そのうえで、低電力化や発熱量が少ないこと。このトレンドは従来と変わらないであろう。ただし、ムーアの法則が終焉を迎えつつあることで、画期的な技術への期待も高い。その一つとして IMC を挙げる声も多いが、2028 年というタイミングでは、その実用化へは懐疑的だ。一方で、技術の進展とは少し離れるが、昨今の半導体不足を経験した回答者からは、供給の安定化を望む声もあった。

特徴的だったコメントを以下に列挙する。

・CPU の高速化、発熱量が少ないということで、2028 年には 2 ナノのチップ主流になるかがポイントである。台湾のメーカーが先行しているが、米国、日本、韓国企業もどこまで、おいつけるかが課題である。インメモリは、高速化には必須であるが、比較的、対応が早いと思われ、同じ会社がしなくてはならない理由はみあたらない。
・消費電力に対する制約が強くなっており 、CPU に対する要求は消費電力に対する処理能力の最大化が強く求められる。消費電力を考えて、処理能力が最大ではないモデルを選択することが更に増えるのではないかと思われる。膨大な消費電力を少しでも抑制する技術を備えた CPU を導入する必要が出てくる。
・CPU 業界全体と市場ニーズの関係性としてみると、性能要求はムーアの法則にあったような曲線で求められるのは変わらないと思う。ただ、その内部アーキテクチャについてはユーザー側として強い意識はしておらず、トレンドを押さえつつ、サービスとして成立しているかが最も重要である。
・IMC の実現にはまだまだ時間を要すると考えます。それ以前に変換効率の低い交流電源利用型から再生可能エネルギー活用形直流電源流通型のコンピュータ導入による高効率電源量子型の構成促進が効果的であると考えます。
・いくら CPU とメモリの連携がよくても、排熱ができなければ性能は出ません。個人的には IMC による性能向上より、物理設備の冷却による性能劣化の防止の方が、需要が延びていくのではないかと推察してます。
・より専門性の高い CPU、ストレージ処理、ネットワーク処理、AI 処理などが開発され、細分化、専門性が要求される。あわせて省エネ化、小型化の研究も進む。問題は安定した供給ができるよう産業が安定するかである。
・IMC 技術の進展により、CPU とメモリの一体化が進むことで、データ処理の高速化とエネルギー効率の向上が求められる。
・半導体不足により、CPU とメモリは需要に対して供給が追いついておらず、調達コストがかかる印象が増したと思う。安価でリードタイムが少なく提供できることと、できるだけ CPU とメモリを消費しないプログラムの両軸で考える必要があると感じる。
・CPU のアーキテクチャの前に、その上で動くアプリケーション・ソフトウェアありきになる。大規模パブリッククラウドを提供するアマゾンやグーグルレベルでない限り、CPU アーキテクチャに関しての優先度は高くないと感じる。
・データセンター向けの CPU には光-電子融合アーキテクチャにより、演算密度とメモリ帯域を同時最大化する 3D 積層構造が必要となるのではと思われます。IMC 実現のため、TSV 技術を超える原子層接合によるメモリ―CPU 融合体が広がるのではと思われます。
・CPU については並列処理などの技術を利用して性能アップはしてきているものの、性能そのものついては頭打ちの感が否めない。仮想化によるリソースの分配などをより効率的に実施する必要性が高まってくると思われる。
・一時期よりはましになりましたが、価格高騰、納期遅れは未だに引きずっているので、新規要求よりは足元の改善が続くと思います。

・どちらかというと、CPU はハードウェアの中枢ではなく、データセンタービジネスを構築するエコシステムの一部として、リソース全体の最適化を担う存在へと進化する。2028 年のデータセンターでは、CPU という概念すら変わっていき、計算リソースとしてはより柔軟で持続可能な形に再構築されるとおもっています。意識的にそう持てるかが、課題だと推察します。

⑪データセンターを構成するアクセラレータに対する要求事項について

アンケート回答者には、これまで挙げてきたような影響や変化をデータセンターが受けた際に、データセンターを構成するアクセラレータ―にはどのようなことが要求されるかについて、自由記述で考えを尋ねた。

関連する記述が多かったのがやはり AI 用途。GPU をはじめ、アクセラレータの機能として、AI 関連の処理をサポートするものが登場するという声が多かった。ただし、アクセラレータの機能を論じるためには、どのようなアプリケーションを求めるのかを見定めるべきだという指摘があり、そういった意味では、2028 年は AI 以上のアプリケーションが出現しない限り、現在の延長上でアクセラレータの機能向上が進むのかもしれない。

特徴的だったコメントを以下に列挙する。

・データセンター側からのアクセレータへの要望ははっきりしており、高性能、エネルギー消費が少ない、スケーラビリティがあるということである。技術的に一長一短はあるが、FPGA の方が、上記の要求に対して柔軟性があると思われる。

・GPU 以外のアクセラレータについても有望なものがいくつか出てきはじめるのが 2028 年ごろではないかと考えている。まだ決定的なものは決まっていない状態で、いくつかのアクセラレータを試しに使ってみることができるという状況が求められると考えられる。

・基本的には GPU・TPU の AI・LLM 向け演算処理が中心であろうと考えている。FPGA を採用することで専用チップの最適処理ユニットの書き換えが可能になった場合にどのくらいメリットが出るかわからないが、ロマンはあると感じている。数が要求され、準専用チップである状態が良しとされる場合がどの程度あるかが鍵。

・セキュリティ強化のための暗号化アクセラレータは即効性があると考えます。演算系についてはデータセンターよりもエッジ配備が効果的でありデータセンターからエッジまでの伝送に暗号化を噛ませる運用が増加すると考えます。″

・データセンターとある意味アプリケーションに近いアクセラレータを関連付けて考えることは非常に難しいと思います。本当に垂直統合的にデータセンターからマシンリソースの提供、アクセラレータやアプリケーションまでビジネスで提供を考えるのであれば影響はあるでしょうが、そこまで分散してリソースと資金を投資する意味はあまりないと考えます。

・現状の生成 AI や認識ソフトウェアは基本的にベクトル/行列演算が多いのですが、今後アルゴリズムの進化によって、別の演算処理が脚光を浴び、それに合ったアクセラレータが登場する可能性があると思います。

・エッジ側での AI 処理要求拡大によりアクセラレータ進化は継続するが、データセンターでの GPU/TPU アクセラレータは熱効率の課題で、2028 年以前に限界が来るかもしれない。

・ここ数年、AI の台頭により GPU のニーズが劇的に高まった。今後も引き続き、GPU のニーズが

高い状態は続くであろう。エッジで動作するGPUの種類が少ない状況であることから、これを補強する動きが2028年にかけてはみられると考えられる。
・量子コンピューティングとの連携を視野に入れ、量子コンピュータとのインタフェースを持つアクセラレータも登場する可能性がある。
・アクセラレータについても、生成AIに代表されるAIサービスの普及に伴い、その役割はこれまで以上に求められてくると考えている。その上でそれを実現するために、どのような製造方法を採用していくのかは、サービス提供側の課題と考えている。
・ユーザー視点では、劇的な高速化が求められるシステムはかなり限られているため、安価で提供できないのであればアクセラレータは、広く需要を捉えるものではないという印象を受ける。
・アプリケーション依存だと思う。GPUもAI処理に特化するように，アプリに依存した製品提供となっている。2028年に求められるアプリがなにかを議論する方が建設的だと思う。
・データセンター向けアクセラレータは、AI/量子ハイブリッド演算に対応するため、プロセスの改善とチップレット設計で演算密度・メモリ帯域を増やしつつ、動的リコンフィギュラブル回路によりワークロード最適化を実現するのではと思われます。

⑫データセンターを構成するストレージに対する要求事項について

アンケート回答者には、これまで挙げてきたような影響や変化をデータセンターが受けた際に、データセンターを構成するストレージにはどのようなことが要求されるかについて、自由記述で考えを尋ねた。

回答者の多くが挙げたのが、大容量化への対応。特にAI用途では大量のデータを元に学習を行う必要があるために、その大量データの収容先としてストレージの進化を望む声が大半を占めた。容量が現在の1000倍以上になるという指摘もあった。また、データが大量になるだけに、データの鮮度や重要度によって、自動で管理方法を変えていくべきではという意見も見られた。

特徴的だったコメントを以下に列挙する。

・ブロックストレージとオブジェクトストレージのように、データの特性に応じて特化していくことが考えられると思います。また、分散ファイルシステムのような方向性も、再検討される可能性があるかもしれないと思います。
・データ量はこれまで以上に増えていき、減ることはないと考えられる。特に、生成AIの学習のためにとりあえずデータをためておきたいというニーズが増えるのではないかと考えており、磁気テープレベルの値段でもう少し使いやすい、といった大容量で安いストレージが求められると思っている。
・キャッシュを要求するアーキテクチャパターンは多くのクラウド事業者が提供しているため、システム全体としてのIO性能については要求を満たしてきているように感じる。キャッシュとして利用できているような感覚でありながら、非同期で書き込み・Flushするような処理を高速でできるようになると期待している。
・データセンターにとってストレージは永遠の課題。データ鮮度を勘案したデータ分散と陳腐データの圧縮化によるストレージリソースの有効活用が求められると考えます。
・大容量、読み書きの低遅延、低価格に加え、ストレージ側で自動的にクラスターを組み、デー

タのレプリケーション等を自動的に実施してくれるようなストレージが求められるのではないか。
・LLM や生成 AI は、大量のデータを使用して学習する必要がある。精度向上のためにモデルのパラメータは増加傾向にあり、それに伴いストレージ容量のさらなる拡充が求められる。しかし、消費電力の増大が課題となるため、省電力化の実現が不可欠である。
・結局、ストレージに求められるのは大容量でかつ高速な IO スピードというのは変わらないので、より汎用的大容量のものとエッジ側に軽く小 IO のものを置く分散モデルとにわかれていくと思われます。
・処理ニーズの拡大は止まらないが、溜め込んだデータを捨てていく仕組みが必要。データそのものに自己の有効性や寿命を管理する概念が生まれるかも知れない。
・オールフラッシュストレージ（SSD）でないと GPU リソースの性能を引き出せない。基本的にはオールフラッシュであることが要件になるケースが増えるであろう。
・圧倒的な容量と高速なアクセス速度が求められる。次世代不揮発性メモリや階層型ストレージシステムの普及などが予想される。
・生成 AI に代表される AI サービスの普及に伴い、ストレージ要件は必須の検討事項となってくるだろう。しかしながら全てのデータに画一的にサービスを提供するのは非効率となるため、データライフサイクル管理（DLM）の考え方を取り入れ、頻繁にアクセスされるデータとそうではないデータの保管先を分けるなどの工夫が必要になってくると考える。
・ストレージ速度ももちろん大事だが、一時データをオンメモリ化することで高速化を図ることは今でも可能だと思っている。それよりも、ストレージの場合キャパシティプランニングが一番の課題となる、それを解消できるソリューションが求められていると感じている。
・AI の進展に伴い、ベクトルデータベースなど新しいデータベースアーキテクチャが進展してくると考えられる。そのための専用アーキテクチャを持ったサーバが導入できるような要請が強まると考えられる。
・量子化圧縮/分散 AI による効率的なストレージ管理技術の普及、CXL メモリープールと SCM を組み合わせた階層型アーキテクチャ、暗号化データへの直接演算可能なプライバシー保護ストレージの実装が求められるのではと思います。
・昨今のデータ容量の増加率は膨大なため、コスト面で有利な従前の HDD の需要は引き続き高まっていくものと思われる。一方でよく利用するデータはスピードが求められるので、SSD などの媒体との使い分けがデータセンター側の運用として求められる。
・現在は異なる階層のストレージ間のデータ移動は人手で行われていたり、ルールベースで自動化する程度になっている。2028 年頃にはストレージ階層の自動化技術が発展し、最適なストレージを自動判断してデータが保管されるようになる。
・ペタバイト・エクサバイト級のストレージの普及が加速。また、フラッシュストレージの高密度化が進む。データアクセスの高速化と大容量のストレージは引き続き求められると思われるが、ただ拡大していくだけでなく、細分化、圧縮化に向けた動きはマストだと考える。

⑬データセンターを構成するネットワークデバイスに対する要求事項について

アンケート回答者には、これまで挙げてきたような影響や変化をデータセンターが受けた際に、データセンターを構成するネットワークデバイスにはどのようなことが要求されるかについて、自由記述で考えを尋ねた。

ネットワークデバイスというよりは、その利用の前提となるネットワークへの要求事項を挙げている回答が多く、その意味では、やはり高速・低遅延で低消費電力への要望が目立った。加えて、セキュリティの堅牢性を重視している意見も多かった。

特徴的だったコメントを以下に列挙する。

・これまで同様に East-West 通信が更に増え、データセンター内部のネットワーク帯域及びトポロジの強化・増強が求められると考えられる。内部通信に対する耐久性はデータセンターの選定時に強く確認されるかもしれない。

・より大容量で、低遅延を求められるし、必須であると考えています。IOWN は時間軸で間に合わないかもしれないが、必要とされているもので、何かしら前倒しでサービス化されるのではないかと考えています。

・ネットワーク専用デバイスは減少し、サーバでネットワーク処理を行うようになる。ハードウェアではなくソフトウェアでより高速に処理できるようになる。

・より高帯域なものでかつ特殊なシーケンスやセッションを分散したり、集中させるようなインテリジェンスをもったトラヒック制御を行うネットワーク機器が AI を活用して開発され、実装されると思われます。

・高速大容量時代を迎えて、伝送デバイスたる光ファイバーの導入本数とその集積度がより向上して、その為の設備の改修・増強が求められます。

・ストレージエリアには、100Gbps 以上のネットワークが必要とされるケースが急増する。具体的には、Infiniband に対応していることが必須要件となる。それ以外のネットワークについても、柔軟にデータ転送ができるように 100Gbps 化されるケースが増える。

・今まで物理的なセキュリティはデータセンターのネットワークに求められていなかった。昨今のサイバー攻撃増加に伴い、データセンター内のネットワークケーブルや配線など、物理セキュリティ対策の程度を見られるケースが増えていることから、管理方法の可視化と堅牢化が今後は求められると考える。

・高速大容量化が必須。400G/800G といった高帯域幅なネットワーク機器の普及や、ソフトウェア定義ネットワーク（SDN）の活用による柔軟なネットワーク運用などが予想される。

・データセンターを構成するネットワークに求めることは、高速であること、止まらない（冗長性を持っている）こと、暗号化など安全にやりとりができること、こういった当たり前に求められていることを、より完全に実現することが必要になってくるだろう。

・セキュリティの強化が最も重要と思われ、中露を中心としたグローバル対応などマクロな視点が必要。

・データセンター・ネットワークは、超高速通信・AI 自律制御・量子耐性暗号が必須になるのではと思われます。光技術と機械学習の融合でレイテンシー1μ秒未満や電力効率 10 倍改善を実現し、メタバース対応 3D 最適化や光スイッチ動的再構成により、大量データと環境規制に対応していく。エネルギー効率とセキュリティが技術競争の軸となり、ネットワークの自己意識化も

考えられます。
・物理的なネットワークインフラの増強をするとともに、利用帯域やユーザーの利用料を詳細に監視し、必要に応じて課金体系を変えるなど、通信を抑制するサービスも必要かと感じている。
・より安価で安全な光通信網の導入と、ワンチップ化が加速してくると考える。
・低レイテンシー、高帯域であるだけでなく、転送距離も必要になる。リソースが複数建屋に分かれても同じワークロードを実行したいという要件が増えるため。
・新技術に対応するようなデバイスであることもそうなのですが、より高速処理とメンテナンスがしやすい機材が求められていると思っています。一番怖いのは、機械が全体的に落ちるということよりも、妙に体力があって劣化したパフォーマンスで動こうとするところです。機械同士の連携もさることながら、流れ的に滞っていることの検知を急がせて、すぐに負荷の分散を行えるようなデバイスと管理体制が求められていると思います。

2. 研究発表・講演、文献、特許等の状況

なし。

契約管理番号：	24001635-0