低炭素社会の実現に向けた
技術および経済・社会の定量的シナリオに基づく
イノベーション政策立案のための提案書

情報化社会の進展がエネルギー消費に与える影響（Vol.4）

－データセンター消費電力低減のための技術の可能性検討－

令和４年２月

Impact of Progress of Information Society on Energy Consumption (Vol. 4):
Feasibility Study of Technologies for Decreasing Energy Consumption of Data Centers

Proposal Paper for Policy Making and Governmental Action
toward Low Carbon Societies

国立研究開発法人科学技術振興機構
低炭素社会戦略センター

LCS-FY2021-PP-01

概要

前報において、情報化社会の進展に伴う世界の情報量（IP トラフィック）は 2030 年には現在の 30 倍以上、2050 年には 4,000 倍に達すると予想され、現在の技術のまま、全く省エネルギー対策がなされないと仮定すると、データセンターだけで 2030 年には年間 3,000 TWh、2050 年には 400 PWh という膨大な消費電力が予測された。

本提案書では、このような IT 社会の消費電力の増大をどのような技術で抑制することが可能か、データセンターの省エネルギー技術について、その消費電力の大部分を占めるサーバを中心にできるだけ定量的に検討した。特にサーバを構成するデバイスの中でもエネルギー消費の大きいプロセッサーに重点を置き、次いでメモリとストレージについて、改善幅の小さい場合（Modest ケース）と大きい場合（Optimistic ケース）について検討した。

2030 年における Modest ケースでは今後 5 ～ 10 年における現行技術の改善、特に CPU におけるマルチコア技術、微細化技術、統合化技術などを、また Optimistic ケースではアクセラレータ技術の進展を織り込んで検討した。この結果 2030 年のデータセンターの消費電力は Modest ケースで国内 24 TWh、世界 670 TWh、Optimistic ケースで国内 6 TWh、世界 190 TWh と推定された。このうちサーバ消費電力は Modest ケースで国内 17 TWh、世界 510 TWh、Optimistic ケースで国内 5 TWh、世界 140 TWh と推定された。

2050 年は遠い将来のため、その予測の信頼性は高くはない。Modest ケースは 2030 年までと同等の改善率で進捗するとした。Optimistic ケースでは、新計算原理としての非ノイマン型 CMOS コンピューティング、量子アニーリングが寄与するとした。この結果 2050 年のデータセンターの消費電力推定として Modest ケースで国内 500 TWh、世界 16,000 TWh、Optimistic ケースで国内 110 TWh、世界 3,000 TWh と推定された。このうちサーバ消費電力は Modest ケースで国内 330 TWh、世界で 11,000 TWh、Optimistic ケースで国内 50 TWh、世界 1,600 TWh と推定された。

結局、2030 年までは現行技術の改善によりデータセンターの消費電力は許容不可能な状態にまでは達しないと考えられる。一方で、今後 10 年程度で現行技術は限界に到達すると考えられることから、AI の社会への浸透がさらに進んで自動運転なども実用化されるであろう 2050 年を見通すときには革新的な新技術の開発が求められ、特に CPU および CPU を補完する計算機能の研究開発が極めて重要である。このためには基礎研究、応用研究、人材育成への長期的継続的投資が必要である。

Summary

In the previous report, it was predicted that the amount of information (IP traffic) would surge by a factor of 30 by 2030 and 4,000 by 2050 due to the development of the information society. Assuming that no energy conservation measures are taken at all with the current technology, data centers alone would be projected to consume an enormous 3,000 TWh per year in 2030 and 400 PWh in 2050.

In this report, we quantitatively studied the energy saving technologies for data centers, focusing on servers, which account for most of the power consumption, to evaluate technologies that can be used to suppress the increase of power consumption in the IT society. In particular, we focused on processors, which consume the largest amount of energy among the server devices, and then on memory and storage, for the case of small improvement (Modest Case) and large improvement (Optimistic Case).

For the year 2030, in the Modest case, improvements in current technologies over the next five to ten years, particularly multi-core, miniaturization, and integration technologies in CPUs, while in the Optimistic case, developments in accelerator technologies were taken into consideration. As a result, the power consumption of data centers in 2030 was estimated to be 24 TWh in Japan and 670 TWh worldwide in the Modest case, and 6

TWh in Japan and 190 TWh worldwide in the Optimistic case. Of this result, server power consumption was estimated to be 17 TWh in Japan and 510 TWh worldwide in the Modest case, and 5 TWh in Japan and 140 TWh worldwide in the Optimistic case.

Since 2050 is in the distant future, the reliability of that prediction is not high. In the Modest case, the improvement rate will be the same as in 2030, and in the Optimistic case, non-von Neumann CMOS computing and quantum annealing as new computational principles were expected to contribute. As a result, the power consumption of data centers in 2050 is estimated to be 500 TWh in Japan and 16,000 TWh worldwide in the Modest case, and 110 TWh in Japan and 3,000 TWh worldwide in the Optimistic case. Of this, server power consumption was estimated to be 330 TWh in Japan and 11,000 TWh worldwide in the Modest case, and 50 TWh in Japan and 1,600 TWh worldwide in the Optimistic case.

These results show that the power consumption of data centers will not reach an unacceptable level until 2030 due to improvements in current technology. On the other hand, current technologies are expected to face their limits in the next 10 years, therefore, when we look ahead to 2050, when AI will become more widespread in society and autonomous cars will become a practical reality, the development of innovative new technologies will be required. In particular, research and development of CPUs and computational functions that complement CPUs will be extremely important. For this purpose, long-term and continuous investment in basic research, applied research, and human resource cultivation is required.

目次

概要

本提案書の作成にあたり、貴重なご助言を賜りました理化学研究所計算研究センター　チームリーダー　佐野　健太郎氏、奈良先端科学技術大学院大学教授　中島　康彦氏、群馬大学研究協力員　中谷　隆之氏、東京工業大学特任教授　西森　秀稔氏、電気通信大学准教授　三輪　忍氏に心より感謝申し上げます。

1. はじめに

インターネットの利用は通信、娯楽、ビジネスに大きな利便を人々にもたらし、一方でこれらの要請にこたえるための膨大な計算と通信をデータセンターとネットワークが支えていること、そして現在の増加率が将来にわたって維持され、かつ技術が現状のまま固定されるとしたら、著しいエネルギー需要が予測されることを前報までに報告した [1-3]。

現実には、今までは計算需要の著しい伸びにも関わらず、半導体微細化技術を中心とする技術進歩により、その消費電力増加は比較的抑えられていた。しかし、AI、ビッグデータ、機械学習、自動運転、ブロックチェイン、シミュレーションなどの技術により、今後も爆発的な計算需要が予想されている。一方で半導体微細化技術による省電力と計算能力の増大は、ほぼ限界を迎えているといわれている。

この消費電力増加の抑制が重要課題であるが、計算の増加を抑制して消費電力を抑制することは社会の要請と相いれない。したがって、情報処理分野における課題は計算需要に対応する計算能力の増大とその時の消費電力の抑制の 2 つになる。

本提案書では、そのなかでデータセンターについて消費電力低減のための課題と低減効果を検討した。通信ネットワークについては、vol.5 で報告する予定である。

データセンターには現在 44 ZB のデータが存在する。2021 年時点で 80-300 億台の端末がつながり、10% / 年の成長で世界の 1%のエネルギーを消費している [1, 4]。データセンターの消費電力のうち、サーバの占める割合が 50%以上と極めて大きく、次いでストレージ、メモリ、ネットワークスイッチとなる。そして今後さらにサーバの電力消費の割合が増大すると予想されている [2]。そのサーバの消費電力の大部分がプロセッサー、次いでメモリである。したがって省電力検討の対象技術も、特にプロセッサーに重点を置き、次いでメモリとストレージについて検討する。

ネットワークスイッチについても、上述 Vol.5 の中で報告予定である。

また、電源や冷却システムも一定程度の電力消費をしていて、その効率は PUE（Power Usage Effectiveness）で表される。この部分は中小規模データセンターでは数十%に上るがハイパースケールデータセンターでは 10%以下程度まで低下していること、また、IT 機器の消費電力が低下して発熱量が低下すれば、それに準じて低下する性質のものなので、本報告では検討を省略する。

なお、データセンターのエネルギー消費の計測やその低減については多くの文献があるが、以下を参考にした [5, 6]。

本提案書では、これら機器につき過去の技術変遷と現在の実態、将来技術等について計算能力と省電力に注目しながら消費電力低減のための課題と低減効果を検討した。もっとも、将来技術については現段階の確度の高い評価は困難であり、専門家の意見を伺いながらまとめたが、不確実性が大きいことを申し添えておく。

2. サーバ

2.1 はじめに

さまざまな計算をする装置をコンピュータと呼ぶ。サーバもコンピュータで、そのなかでデータを処理するプロセッサー、データを記憶するメモリが主要デバイスである。データセンターにはサーバが数千ないし数万台設置される。プロセッサーには主となる CPU（Central Processing Unit）のほかにも GPU（Graphic Processing Unit）、TPU（Tensor Processing Unit）などがある。同様の用途に用いられる FGPA（Field Programmable Gate Array）、ASIC（Application Specific Integrated Circuit）もある。現在使用されているコンピュータの中心プロセッサーは CPU であるが、膨大な計算や画像処理は CPU のみでは効率が悪いため、GPU や TPU など（アクセラレータといわれる）が受け持つことが多くなっている。

CPU については、微細化を中心とする技術進歩により、計算能力の向上をしつつ省エネルギーが達成されてきた。しかし、微細化による計算能力の向上と省エネルギーについては 2005 年頃にほぼ限界に到達し、その後は多数のコアを導入した並列計算などにより、計算能力の向上と省エネルギー化が図られている。この方式がいつまで可能か、その次の技術としてどのような技術が考えられていて、それぞれどの程度の計算能力増大と省電力が期待されるのかを検討した。

2.2 プロセッサーの種類

過去数十年から現代まで、ほとんど全てのプロセッサーはノイマン型コンピューティングといわれるアーキテクチャを採用していて、プログラムが格納されたメモリから命令を順番に読み込み、次に指定されたメモリ番地にあるデータを読み込んで指定された命令を実行するという逐次実行型といわれる方法で演算をする。この読み込み速度を制御するのがクロック速度であり、クロック速度が高いほど演算は速くなる。

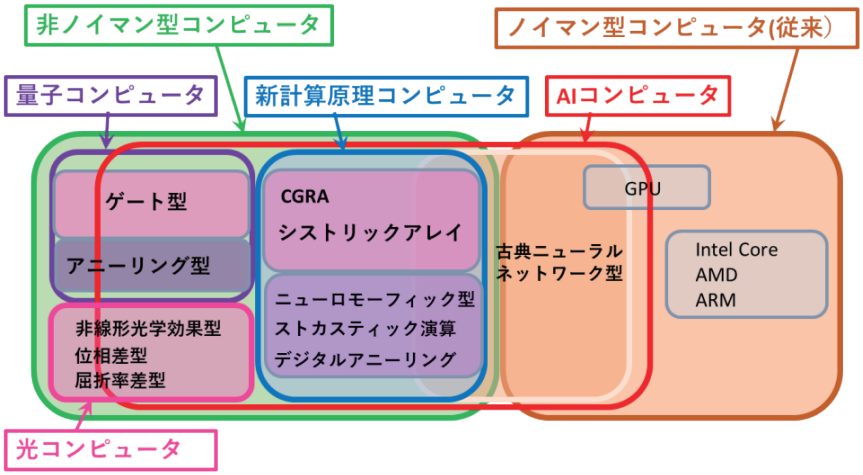

図１　様々なコンピュータ
（文献 [7] などを参考に LCS が作成）

ところが、近年、特に AI などによる膨大な計算需要が増加して、従来のコンピュータの限界が指摘されている。図 1 に示すように次の技術としてはノイマン型コンピュータのまま脳神経回路を模倣するニューラルネットワーク型、さらにその先の技術として非ノイマン型コンピュータの研究がなされている。非ノイマン型コンピュータは多種多様であり、詳しくは後述するが、大きくは計算精度を犠牲にして高速化を図る手法と、計算精度を保ちながら省電力を図る方法に分けられる。さらに CMOS 型、量子コンピュータ、光コンピュータに分けられる。このような非ノイマン型コンピュータは、大規模化に必要な安定性やプログラム容易性に問題があるなど、使いにくいといわれていて、今後もノイマン型の CPU を中心として、その CPU を仲介して使用するようになるだろうといわれている [7]。

2.3 CPU の進化の歴史

図 2 にインテルのデスクトップ用 CPU の 1980-2020 年の性能向上の歴史を示した。また、近年（2000-2020 年）の性能向上については、図 3 にインテルのサーバ用プロセッサーについて示した。両図とも文献値から LCS が作成した [8-11]。

具体的には、製造上の技術ノード（デザインルール（nm））、クロック周波数（GHz）、TDP（設計消費電力（W））、演算速度（Gflops）、消費電力性能（Gfolps/W）について縦軸を対数にとり、経年的に示している。

技術ノードは、かつてトランジスタのゲート長といわれていたが、現在は具体的なサイズとは直接関係がなく、小さいほうが微細である。演算速度は 1 秒間当たりの倍精度浮動小数点演算回数（flops）について 10 億回を単位として表示、設計消費電力は連続使用した場合の最大許容消費電力を W で表している。消費電力性能（Gflops/W）は 1 W 当たりの演算回数について 10 億回を単位として表示している。

これから分かることは、技術ノードは 2015 年以降 14 nm で停滞していること、クロック周波数は 2005 年頃から 3 GHz 程度で停滞していること、TDP も 2005 年頃から停滞していること、消費電力性能も 2015 ～ 6 年頃から改善幅が小さく、ほとんど停滞していることである。

すなわち、後述のような微細化に関する Dennard 則や Moore の法則は、CPU についてはすでに成立していないことが分かる。

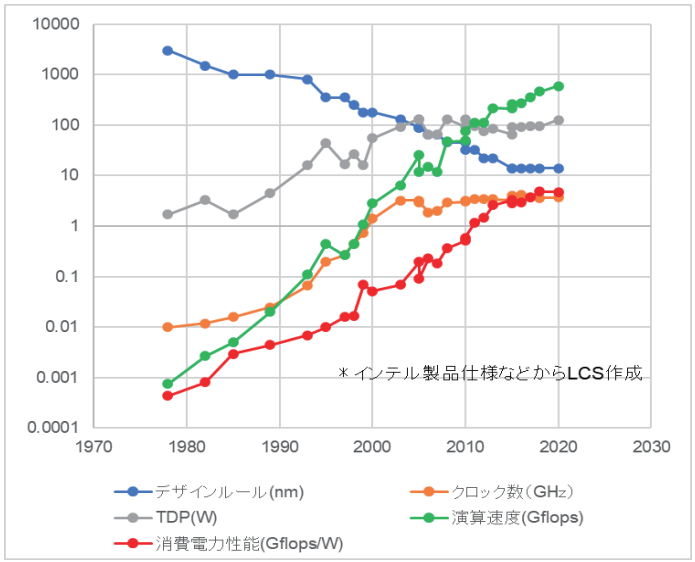

図 2　intel CPU の性能向上

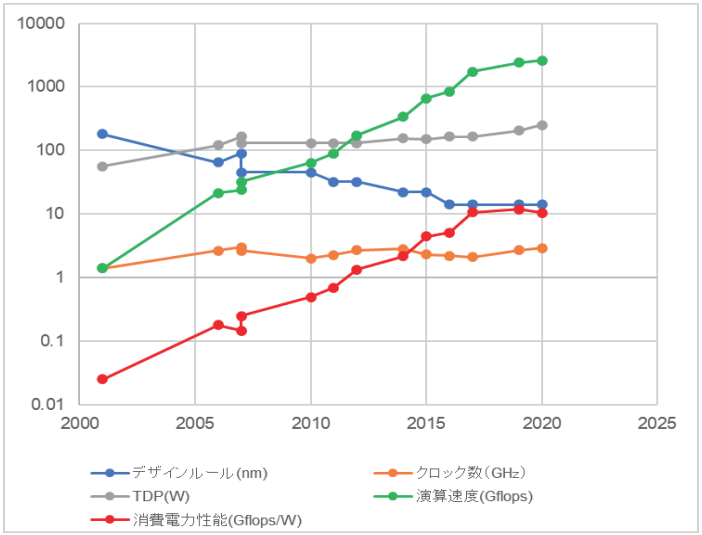

図 3　intel サーバ用 CPU の性能向上

3. CPU の演算速度と消費電力

現在ほとんどのコンピュータに採用されているノイマン型アーキテクチャを前提にして CPU の演算速度と消費電力の電源電圧、クロック周波数、配線容量、トランジスタ数、コア数との関係について整理しておく。

3.1 CPU 演算速度

CPU が計算を実行する場合にはメモリ上にある命令をプログラムの順序に従って読み込み、その命令を実行するためにメモリ上のデータを演算器に読み込んで演算を実行する。この命令の実行やデータの読み込みはクロックのタイミングで行われる。

したがって、CPU の演算速度はミクロ的には演算器の速度、データをメモリから演算器に移動する速度、キャッシュヒット率や分岐予測ヒット率等ソフトウェアから生成される機械語命令列の特性などで決まる。マクロ的な取り扱いとしてはメモリの帯域幅を考慮したルーフラインモデルが知られている [12]。ただ、このモデルではパラメータが多く、取り扱いが複雑になる。実際にメモリの帯域幅が律速とされてはいるが、まずは演算器に着目し、メモリの帯域幅は十分であり（演算器が律速段階）、またソフトウェアは多種多様なので、その効果は後述の（3）式の定数 A に含まれることにする。

CPU の演算器は多数の論理ゲートからなる。論理ゲートは CMOS トランジスタで構成されている。そして CPU の演算速度（S: flops）は CMOS 論理ゲートのスイッチング速度で決まる。スイッチング速度は、負荷容量（C: nF）を電源電圧（V: V）から V/2 に放電する時間で決まる。MOS トランジスタのドレイン電流（I_d: nA）、スイッチ時間（t_s: s）とすると

$$t_s = 0.5 \times C \times V/I_d \qquad (1)$$

となる。そして、最大クロック周波数（f_{max}: Hz）はスイッチ時間の逆数に比例するので

$$f_{max} \propto 1/t_s = 2 \times I_d/(CV) \propto V/C \qquad (2)$$

したがって、同じ回路であればクロック周波数 f_{max} は電源電圧に比例する [13, 14]。なお、上記関係については、クロック周波数一般について成立するものではない。しかしクロック周波数を高くするためには高電圧が必要になることには異論はない。

演算速度はクロック周波数の 0.67 乗に比例するという記述もあるが、簡単にするためにここでは比例すると考える。また、ポラックの法則により演算速度は論理ゲート数（N_g）の平方根にも比例するから、ソフトウェアの効果も含めて A を定数として（3）式となる。

$$S = A \times N_g^{1/2} \times f \qquad (3)$$

次に、マルチコア [15] と呼ばれるプロセッサーを多数連携させた最近主流の CPU を考える。マルチコアの効果についてはソフトウェアの構造に依存し、逐次処理の部分は高速化されない（Amdahl の法則）[16]。同一コアにより構成されるマルチコアの場合、非並列化率 x としてコア数 N_{core} とするとシングルコアの演算速度 S、マルチコア演算速度 S_m として

$$S_m = S/(x+(1-x)/N_{core}) \qquad (4)$$

となる。

x は 0 ではないから、演算速度はコア数に応じて増大するものの比例しては増大しない。

以上をまとめると、CPU の演算速度は、CPU のクロック周波数、論理ゲート数（トランジスタ数）の平方根に比例し、またコア数に応じて増大する。そして最大クロック周波数は電源電圧に比例する。

3.2 CPU 消費電力

3.2.1 CPU 消費電力と電圧、クロック周波数、トランジスタ数の関係

CPU の消費電力（E_c）について、単純化して、計算量に比例する部分（$E_{c,p}$）と、トンネル効果によるリーク電流 J による部分（$E_{c,leak}$）の和と考える。リーク電流は全消費電力の 30 ～ 40% といわれている [14]。

$$E_c = E_{c,p} + E_{c,leak} \qquad (5)$$

CPU 計算量に比例する部分は、Dally らによると、演算器による部分（$E_{c,r}$）、Overhead（$E_{c,o}$）と Localization（$E_{c,l}$）によると解説されている [17]。Overhead とは命令を読み込み、プログラム全体を高速で実行するために命令の分岐を予測してスケジューリングして実行する部分である。Localization は命令を読み込むときに、命令を格納したメモリの距離に関するものである。すなわち、

$$E_{c,p} = E_{c,r} + E_{c,o} + E_{c,l} \quad (6)$$

この各部分の考察は後述するが、ここでは簡略化して、$E_{c,p}$ を演算器のスイッチングに比例するとする。CMOS 論理ゲートの出力の 1 回のスイッチに伴う消費エネルギー（E_{sw}）は論理回路の負荷容量（C）、電源電圧（V）として

$$E_{sw} = 1/2 \times CV^2 \quad (7)$$

CPU の持つ論理ゲートの数を N_g、クロック周波数 f とすれば、CPU の演算消費エネルギー（$E_{c,p}$）は、

$$E_{c,p} = f \times N_g \times E_{sw} \quad (8)$$

すなわち電圧が低いほど、負荷容量が小さいほど、CPU の消費電力は小さくなる。またクロック周波数 f とトランジスタ数に比例して消費電力は大きくなる。

（8）と（7）から、CPU のスイッチングによる消費電力は

$$E_{c,p} = 1/2 \times f \times N_g \times C\,V^2 \quad (9)$$

上記から、CPU 消費電力（$E_{c,p}$:W）は負荷容量、クロック周波数、トランジスタ数に比例し、電源電圧の 2 乗に比例する。

3.2.2 リーク電流

リーク電流は電界強度 F とすると

$$J \propto F^2 \exp(-K/F) \quad (10)$$

ここで $K=4 \times (2m(q\Phi b)^3)^{-1/2}/3h)$、m は電子質量、h はプランク定数、qΦb は障壁高さ [18]。
電界強度は、電源電圧 V、絶縁層誘電率 ε とすると

$$F = V/(\varepsilon d) \quad (11)$$

（12）と（11）からリーク電流による CPU の消費電力は、exp のべき数が負のため、この項の寄与は小さいとすれば

$$E_{c,leak} = N_g JV \propto N_g V^3/(\varepsilon d)^2 \quad (12)$$

よってリーク電流による消費電力は電界の 2 乗、電源電圧の 3 乗に比例して増大する。

3.2.3 マルチコア

マルチコアの消費電力（E_{cm}: W）についてはコア数に比例する。

$$E_{cm} = N_{core} \times E_c \quad (13)$$

CPU の演算速度を上げるとき、配線容量など他の要素が変わらなければ、最大クロック周波数を用いることになる。この場合、(9) 式と（2）式から、

$$E_{c,p} \propto (V/C)\, N_g\, C\, V^2 = N_g\, V^3 \quad (14)$$

最大クロック周波数 f_{max} を 2 倍にすると電圧は 8 倍となり消費電力が増大し、結果として単位面積当たりの発熱量が許容範囲を超えてしまう（後述図 5 参照）。一方で（13）式によりクロック周波数を変えずにコアを 2 倍に増加させれば、消費電力は 2 倍にとどまり、演算速度は（3）式のようにポラック則により 1.4 倍程度と推定される。このため、消費電力の増加を抑えて演算速度をあげるためにマルチコアまたはメニーコアアーキテクチャが導入された [15]。

3.3 微細化の効用（Dennard 則）

微細化の効果については、デナードのスケーリング則が知られている [19]。

すなわち、寸法パラメータを 1/α 倍したときに回路密度は α^2 倍、電圧は 1/α 倍、回路当たり消費電力は $1/\alpha^2$ となる。その結果、単位面積当たりの消費電力はほぼ変わらず、速度はα倍になる。また（5）式の負荷容量（C）も概ね線幅（技術ノード）に比例するために微細化により減少し、消費電力を減少させる。

この経験則は電圧が低下できることが前提となっていた。しかし微細化が進むにつれて電界強度によるリーク電流の増大が無視できなくなった。これを抑えるために、絶縁層を高誘電率にする必要があり、絶縁層の材料にも種々工夫がされてきた [20]。また後述のように電源電圧をデナード則に従って下げることができなくなり、2005 年頃からこの経験則は成立しなくなった。また微細化が進んでも消費電力は全て熱に変わるため CPU の冷却が問題となる。

有名な Moore の法則（1 年半ごとにトランジスタ数が 2 倍）も、このデナード則に支えられていたため、次章で述べるように論理 IC では既に成立していないといわれている（一方で立体化技術などにより継続しているという見方もある）。

4. 現行プロセッサーの技術進化

最初に述べたが演算能力の向上と省エネルギー技術の推進がプロセッサーの課題である。これら、およびこの双方に対して様々な技術が検討され実現されてきた。

まず、この両課題に今まで対応してきた主要技術である微細化技術について検討し、次に演算能力の向上、消費電力の向上、および双方の解決課題としてのアクセラレータについて述べる。

4.1 微細化技術の現状と将来

4.1.1 現状

表 1 のように技術ノード 10 nm までは ArF 液浸＋多重露光（SADP（自己整合型ダブルパターニング））という製造プロセスがとられている [21]。

表 1　各社の論理 LSI 技術ノードの詳細 [21]

	Intel	GLOBALFOUNDRIES	TSMC
公称世代寸法	10nm	7nm	7nm
発表した学会	IEDM 2017	IEDM 2017	IEDM 2016
基本技術	バルクFinFET	バルクFinFET	バルクFinFET
フィンピッチ	34nm	30nm	公表せず
ゲートピッチ	54nm	56nm	公表せず
金属配線ピッチ（最小値）	36nm	40nm	40nm
ゲートピッチ×コンタクトピッチ	1944平方nm	2240平方nm	不明
SRAMセル面積（最小値）	0.0312平方μm	0.0269平方μm	0.027平方μm
リソグラフィ技術（すべてArF液浸）	SAQPとSADP	SAQPとSADP	マルチパターニング

微細化技術は技術的に難易度が高く歩留まりが低くなり、また設備投資も巨額なため、技術ノードのステージが上がるごとにプレーヤーが減少してきた。現在最先端の技術ノードでは、装置メーカーが ASML 1 社、製品メーカーは TSMC、IBM、サムソン、インテル以外にプレーヤーがいないという寡占的状態になっている。微細化技術で最先端といわれている TSMC は技術ノード 10 nm、7 nm で商用化済といわれている。Intel の 10 nm は他社より配線が微細といわれているが、2015 年から主要 CPU は基本的には 14 nm にとどまっていた。しかし、2021 年から再度微細化に向けた方針を発表している。

微細化技術が追及されてきた理由はICの高性能化とコンパクト化、コストダウンが同時に達成できたからである。一般に単位面積当たりのプロセスコストは微細になるほど高くなるが、トランジスタ数の増加割合がそれを上回るために、1トランジスタ当たりのコストは低下し続けてきた。

しかしTSMCでも16 nmから10 nmの移行時にトランジスタ当たりのコストは増加した。今後も多重露光によるプロセスコストの増加のため、微細化によりコストは上がるといわれている[22]。

4.1.2 今後

技術ノードについてIMECによるロードマップを表2に示すが、ここで分かるように技術ノードのサイズと実際のプロセスの解像度とはすでに乖離していて、技術ノードの数値自身には物理的意味がないとされている。なお、この予定のようには必ずしも進んでいないようである[23]。

表２　技術ノードのロードマップ[23]

技術ノード	7nm	5nm		3nm			2nm
量産開始時期	2019年	2020年～2021年		2022年～2023年		2024年	2026年？
解像度（ハーフピッチ）	19nm	16nm		12nm		12nm	9nm？
開口数（NA）	0.33	0.33		0.33		0.55	0.55
プロセス係数(k1)	0.46	0.39	0.46	0.29	0.39	0.46	0.46	0.37
露光回数	シングル	シングル	ダブル（LELE）	シングル	ダブル（LELE）	トリプル（LELELE）	シングル	シングル

技術ノード7 nmについてはTSMCのみが量産化に成功しているといわれている。このレベルでは微細化により性能を確保するためにさまざまな技術課題が生じ、例えば配線材料でもCu配線は反応層の問題があり、ビアとコンタクト形成にコバルトの採用が検討されている。

このように、単純な微細化技術だけでなく付帯した材料技術の開発も必要になり、技術的難度は上がっている。パターニングについても7 nm以降ではEUV露光、さらに5 nmでEUV露光＋多重露光が必要になるといわれている。

構造についても技術ノード5 nm以降では、単純な従来技術の延長では駆動力不足になったりセル面積が減少したりしないので、FinFETからナノシートFET、フォークシートFET、CFET（Complimentary FET）などという複雑な微細構造が提案されている。

EUVは技術的にも難易度が高く、装置も極めて高価のため10 nm以降のチップコストの高騰が指摘されていて、コスト面からも汎用CPUでの微細化は限界に近付いているといわれている。微細化によるトランジスタ密度の向上を通じた演算速度の向上は期待されるものの、微細化による漏れ電流も増大しているため、今後は大きな消費電力低減効果は期待しがたいといわれている。

さらなる微細化追求の研究は技術難易度が高く量産化が難しいわりに効果が小さいと考えられ、EUVや多重露光を利用しない微細回路技術のような新しい発想の研究が期待される。

4.2 微演算能力の向上

（3）、（4）式で示したようにCPUの演算速度はクロック周波数に比例し、さらに1クロック当たりの演算回数（演算を行うトランジスタの数）に比例する。また、2章では述べなかったが、データを演算器―キャッシュメモリ間で移動する速度（メモリ帯域幅）にも関連する。ここでは、デバイス技術としてのクロックの高速化、トランジスタ数の増加、回路技術としての微細化、アーキテクチャと関連したマルチコア、アクセラレータについて検討する。

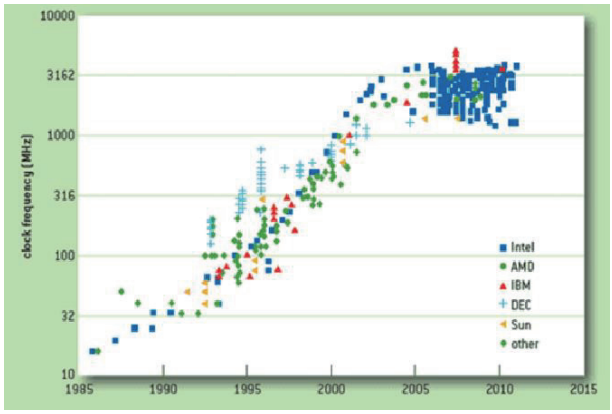

図４　クロック速度の向上 [25]

4.2.1 クロックの高速化

式（2）のようにf_{max}はVに比例するため、動作電圧が高くなるので、消費電力はクロック周波数に比例して増加する。実際には微細化による配線容量の低下と、電源電圧を下げながら、周波数を上げることができていた。しかしながら2005年頃から微細化による漏れ電流の問題もあり、周波数高速化によるエネルギー密度が過大となり技術上の限界に達した [24]。図4のようにクロック周波数で3-4 GHz、図5のように発熱密度で1 W/mm^{2}が限界と思われる [25]。

今後ともクロック周波数の増加は消費電力増加に結び付くと考えられる。

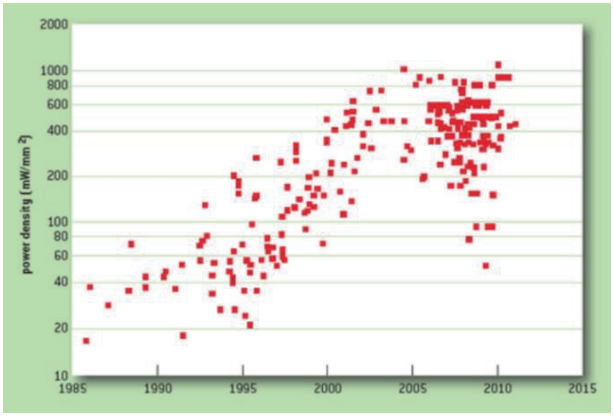

図５　電力密度の増加 [25]

4.2.2 トランジスタ数の増加

式（3）のように演算速度はトランジスタ数の平方根に比例して高速化する。一方で式（7）、（12）のようにトランジスタ数に比例して消費電力が増加する。したがって単純にトランジスタ数を増やすと演算速度が1.4倍で消費電力は2倍となり、エネルギー効率は低下する。ただ、従来は前述のデナードのスケーリング則により、電圧が低下したためにトランジスタ数の増加は消費電力の増大につながらなかった。

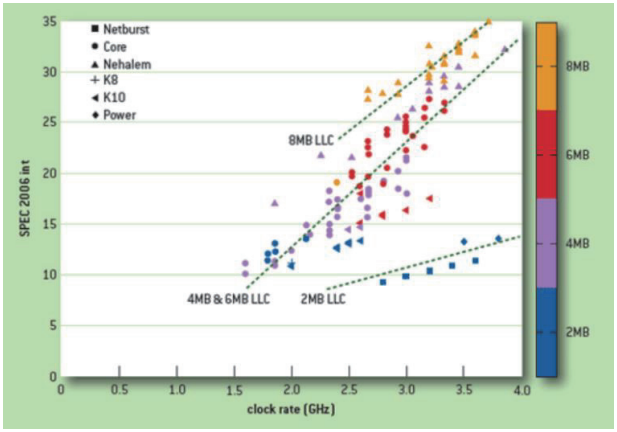

図６　キャッシュメモリの大容量化 [25]

もっとも、電圧は45 nm以下でスケーリング則にしたがって低下しなくなり、デナード則は破綻した [26]。

今後はトランジスタ数の増加に応じて消費電力は増大すると考えられる。

4.2.3 キャッシュメモリの大容量化

キャッシュメモリの構造化（1 次キャッシュ、2 次キャッシュ、3 次キャッシュなど）やキャッシュメモリの大容量化で演算速度は向上する [25]。もっとも、データ移動に伴う消費電力の抑制という意味での効果は不明である。

4.2.4 データ転送速度向上

4.2.3 で示したようなデータ処理量の増加に伴うコアの増加やメモリの大容量化などにより、コア間、キャッシュメモリとレジスタ間、メインメモリとキャッシュメモリ間などのデータ転送速度が問題になっている。その解決策の一つとして配線の短縮や配線材料の改良、光通信の短距離間への導入などが挙げられる。

具体的には垂直に何層も積み重ねる HBM、HBM2 などの広帯域メモリが提案され、商用化されているが、コスト高のようで普及はしていない [27]。

もっとも、CPU はすでに発熱密度による上限の状態のため、コアを積層しても発熱の問題から十分に稼働できず、省電力という観点からはあまり意味がないと思われる。ただ、高コストを負担してでも高速計算をしたいスーパーコンピュータなど HPC 向けには意味のある技術と考えられる。

4.3 省電力

4.3.1 低電圧化

前述の（14）式のように f_{max} において消費電力は電圧の 3 乗に比例するので、低電圧は、省電力には大きな効果がある。一方で（2）式のように f_{max} と電圧は比例するので、低電圧にすると高速化が困難になる。これを解決するためにメニーコアアーキテクチャが採用されている。

3.3 で述べたように低電圧化は停滞していて、技術的に難しいといわれている。一方で、実際には電界の問題なので素子構造の工夫により、さらなる低電圧化は可能といわれている。この場合には、後述のメニーコアとの併用になると思われる。

4.3.2 微細化

4.1 で詳述したように微細化によって得られる省電力効果に対する高プロセスコストの観点から汎用用途ではほぼ限界ではないかと考えられ、また、エネルギー効率も従来ほど大きな改善は見込めないと考える。

4.3.3 配線抵抗の低減

消費電力が 1/2 になるほどの効果はないと思われるが、配線距離短縮、配線断面積増大（埋め込み）、新規配線材料（Cu → Co といわれている）などが検討されている。配線抵抗の低減は、特に微細化による配線断面積の減少による配線抵抗の増大の問題に対処するために研究されている。

4.3.4 キャパシタンス低減

論理ゲートのキャパシタンスは配線などによるものも含まれるため、実態は必ずしも明らかでない。ただ、極めて微弱な電流は検出に負担がかかること、また、熱的励起などでも電流は流れて S/N が悪化するために冷却が必要になったりするなど、微細現象を追求するほど電力を必要とすることがある。液体冷却するトランジスタはそれらへの冷却ラインの煩雑さから高コストになると思われ汎用 CPU では使えないように思われる。

4.4 演算能力と省電力の同時解決（アーキテクチャ）

前述のように電圧、クロック、微細化、トランジスタ数などによる高速化は限界を迎えたために省電力かつ演算能力の向上を目的とした新しい設計思想が必要となった。その一つが CPU 内に多数のプロセッサー（コア）を配置して並列計算を行うマルチコア技術、もう一つは負担となる演算部分を CPU の外に置くアクセラレータ技術である。

4.4.1 マルチコア技術 [15]

インテルのサーバ用 CPU などは、2005 年頃にクロックの高速化をあきらめて、コア数の増加で計算能力を稼いでいる。コア数の増加は特に並列処理に威力を発揮するといわれている。一方で、順次計算のみの場合にはメリットがないともいわれている。コア数はすでに 20 ～ 30。GPU では数千に達している。この効果は 2.2 で述べたが、すでに技術的には成熟していて現在のアーキテクチャのままでは今後の改善の余地はあまりないといわれている。

コア数の増加の効果が今後はあまり見込めない原因の一つとしてコア間、コア―メモリ間のデータ転送による遅延と電力消費が指摘されている。演算器とメモリ間などのデータ転送の頻度を少なくするなど、データ転送による消費電力を低減する新しいアーキテクチャの開発の可能性はあると考えられる。

また、コア間連携技術としては異種のコアを実装するヘテロジニアスマルチコアアーキテクチャと呼ばれるものが注目されている [28]。例えばモバイル向けの big.LITTLE アーキテクチャにおける、電力消費が大きく高速演算が可能なコアと、電力消費が小さく演算速度が遅いコアを処理に応じて使い分ける技術が挙げられる。また、負荷が下がった時には不要なコアやメモリに給電しないスリープ技術などもある。これらにより消費電力は従来の 1/2 から 1/3 になり得るという。

4.4.2 アクセラレータ

各種用途への応用を考えると、プログラマビリティが高く、汎用性の高いプロセッサーは必須と考えられるので、従来のノイマン型の CPU は今後も継続使用されるといわれている。AI 用途やビットコインマイニングなど、演算能力が不足する場合に、計算部分を GPU や ASIC などのアクセラレータに任せることでコンピュータ全体としての演算能力を上げる方法が採用されている。この場合、Man-Machine Interface は使い勝手の良い CPU が使えるために、アクセラレータの選択に関しては自由度が高いので、今後この方向に技術が進むと考えられる。これはアーキテクチャに関わるので、詳細は次章で検討する。

(1) GPU

アクセラレータとして近年盛んに用いられているのが GPU である。これは元々グラフィック用のプロセッサーであり、多数のコアを用いて並列計算する。GPU は精度の異なるさまざまな計算モードをもち、桁数の小さい数で演算速度が高い。このため特に AI など膨大なデータを処理する用途には必須といわれていて、CPU の数倍から整数モードなどでは 2 桁大きい演算速度を持つ。さらに GPU 自身の演算速度の向上も NVIDIA で 4 年に 6 ～ 9 倍と著しく、A-100 では 20-300 TFlops、消費電力 400 W という仕様である [29, 30]。

ただ、かなり技術的には成熟しつつあり今後は桁違いの速度向上は困難との見方もある。エネルギー効率的にはあまり向上しておらず、V100 でも A100 でも 25 Gflops/W 程度である。したがってエネルギー効率の向上は CPU と同様に 2030 年には悲観的にみて 2 倍、楽観的にみて 10 倍程度と推測する。

(2) ASIC

ASIC（Application Specific Integrated Circuit）は、特定の用途に向け複数機能の回路を一つにまとめた集積回路である。定まった計算処理を高速で実行するためにプログラミングの機能の一部を

ハードウェア上で実現することができる。長所としては消費電力の低減、動作速度の向上、量産時に低コストであり、短所としては開発費が高く開発期間が長いこと、回路設計の修正が困難であること、目的用途以外に使用できないことである。

有名な応用例はビットコインのマイニングである。専用機として威力を発揮するが、多様な業務をこなす部分には向かず、標準バス制御やIoT向けなど定型業務向けに実用化されているが、汎用技術の中心となるとは考えにくい。

4.5 省電力技術とその効果のまとめ

以上をまとめると、現行アーキテクチャのままの回路、デバイス技術の改善、既存アクセラレータ（GPU）の改良の効果を考える場合、クロックの高速化やトランジスタの増加は省エネルギーに結び付かない。微細化による省エネルギー効果は、CPUコストが跳ね上がってコストパフォーマンスが悪化することを考慮すると、汎用のサーバなどでの大量導入には疑問があり、特別な用途を除けば、せいぜい数倍の改善にとどまると推測される。メニーコア技術も省電力効果はほぼ頭打ちに近づいてきているようでダークシリコン対策やヘテロジニアス技術などによる改善効果は数倍～10倍以内程度と思われる。アクセラレータについては、将来技術の部分は次章に譲るとしてGPUの効果としてはおよそ現在の5～20倍程度の省電力が可能と推定される。また、データ移動に伴う消費電力の低減も3D化、メモリとプロセッサーの一体化などの統合技術で配線距離の短縮が検討されていて、一定の効果は上げつつある。これらも数倍程度の効果はあるものと思われる。

上記種々技術は相互に関係する部分もあるので、それぞれの改善による省エネルギー効果が各々独立に達成されるかは不明であり、今後5～10年における現行技術の改善は省電力効率としてはCPUでおよそ2～10倍程度、GPUで5～20倍程度のオーダーではないかと推測する。

5. プロセッサー将来技術

逐次処理によるノイマン型コンピュータはプログラマビリティや精度の長所から今後も主流と考えられる。一方で、ディープラーニング、AI、量子化学計算、気候予測など膨大なデータを短時間で計算する需要が今後増大する。しかしその計算のための電力は無尽蔵ではない。

ここでは、現行のアーキテクチャによるプロセッサーの消費電力の内訳から、消費電力低減の課題がどこにあるかを検討し、ついで現在提案されている種々の技術について解説する。

詳細は以下に述べるが、概略の見通しを述べると、計算需要は今後爆発的に増大すると考えられるため、現在よりも消費電力性能が2桁レベル高い計算機が必要となる。ノイマン型コンピュータはメモリにデータを蓄えてプロセッサーで演算を行うため、データ量が膨大になるとデータバスのメモリ帯域が律速になり、演算速度が上がらないという本質的な問題（ノイマンボトルネック）がある。

この課題に対応するものが、非ノイマン型コンピュータであり、後述するようにCMOS技術の非ノイマン型コンピュータであるニューロモーフィックコンピュータ、CGRA、デジタルアニーリングとか、さらに将来の技術としての量子コンピュータ、光コンピュータなどがある。いずれも現在は研究段階であるが、CMOSプロセスを利用できる方式が実用に近いと考えられる。ただ、確率的な計算機やハードウェアとソフトウェアを明確に区別できないデバイスについてどのような検査で動作保証できるのかといった問題点が指摘されている[31]。

5.1 CPU 消費電力の内訳

3.2.1 で述べた E_c の内訳については、図 7 に示すようにプロセッサーの消費電力を細分化した報告がされている [17]。In-order はプログラムに従った逐次命令実行で省電力な計算方法である。これによればプロセッサーの消費電力の内、純粋な計算に消費される部分（ALU）は、4 ～ 6%であり、残りは計算の順番の制御やデータの移動に関連する。

3.2 で述べたように CPU の消費電力の内、純粋に計算に消費される電力は 10%以下といわれている。

これに関するデータとして図 8、表 3 のような報告がある。

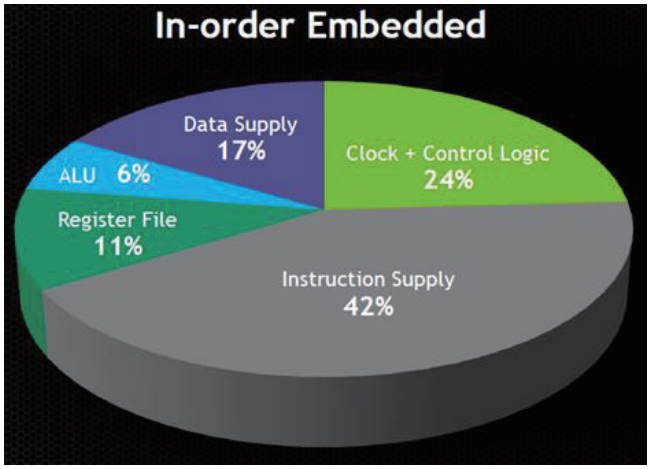

図 7　CPU 内での電力消費内訳 [17]

図 8 のようにデータのプロセシングには 64 bit のデータ処理に 20 pJ 要するが、それをチップ内の近くから読み込むと 26 pJ、遠くからは 256 pJ、チップ外からは 500 pJ-1 nJ と、データの移動がエネルギーを消費する [17, 32, 33]。

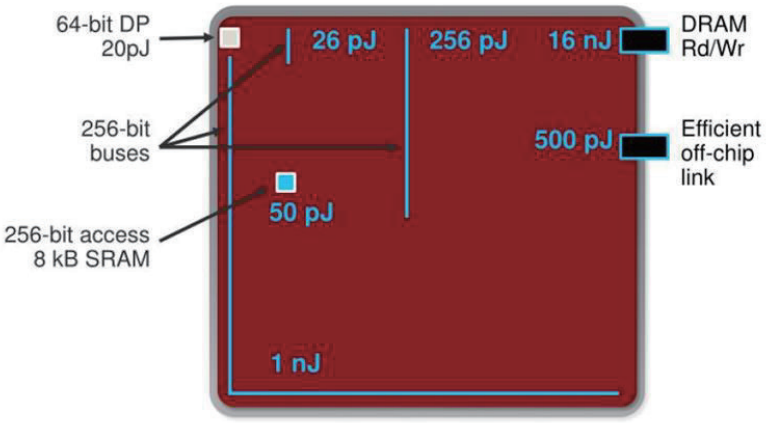

図 8　CPU の計算とデータ移動の電力消費 [33]

また表 3 はデザインルール 40 nm と 10 nm のときの CPU の消費エネルギーの比較で、10 nm では動作電圧が 0.65 V と 0.75 V の場合がある。微細化により演算エネルギーや SRAM からの読み出しエネルギーは低下するが、データの伝送損失はあまり小さくならないことが示されている。SRAM も大量のデータを扱うために現在は MB のオーダーになり、面積の増大とともに消費電力は増大しているといわれている。

上記から、消費電力の観点からは演算速度を犠牲にしても計算の順番の制御に電力を使わない高効率 CPU（コア）が考えられ、演算速度を重視する従来型の CPU と組み合わせて、4.4.1 でも述べたヘテロジニアスアーキテクチャが出現している。消費電力の観点からはプロセスによるものは数倍程度で、計算アーキテクチャと回路の改善によって、一桁程度の余地はあるとの主張がある。

表 3　CPU の消費エネルギー [32]

			40nm	10nm	10nm
Vdd	電源電圧	V	0.9	0.75	0.65
Frequency	周波数	GHz	1.6	2.5	2
DFMA（Double-precision Fused Multiply Add）	演算器	pJ	50	8.7	6.5
64bit read from 8KB SRAM	読み出し	pJ	14	2.4	1.8
Wire Energy(256bits,10mm)	伝送損失	pJ	310	200	150
Wire Energy per transition	伝送損失	fJ/mm	240	150	116

5.2 非ノイマン型 CMOS コンピューティングアーキテクチャ

前章でも述べたように非ノイマン型コンピュータアーキテクチャを現在の CMOS 技術によるハードウェアで実現しようとするさまざまな提案がなされている。前章で述べたビットコイン専用 ASIC や AI 用の TPU など実用化されているものもある。ここではより詳細に述べる。

このなかでニューロモーフィック、デジタルアニーリング、ストカスティックなどの方法は精度を上げるのか、これらの全てが実用化されるのか、一部にとどまるのか、省エネルギーの改善率がどの程度か、などは現時点で見通すことは難しいが、量子コンピューティングや光コンピューティングよりは早期の実用化が見込まれ、今後 10 年程度で研究が進展すると期待される。なお、実際にコンピュータとして使用するためには、CPU のような定型的なものでないため、検査方法の確立も必要といわれている。

(1) FPGA

FPGA（Field Programmable Gate Array）はビット単位の機能をプログラムできるハードウェアであり、様々なコンピュータアーキテクチャを実現する。うまく使えば電力効率が高く、特定の用途で実用化されている。ただし、高級言語を直接ハードウェアに変換するため、合成技術は発展途上であり、高効率を達成するためには、まだハードウェア記述言語によるプログラムが必要である。また ASIC に比べるとクロック周波数が低く、大容量の FPGA は高価であるなど、応用範囲を広げるための障壁は多い。もちろん FPGA は、今後再構成可能な粒度を上げることで、前述のストカスティック演算や CGRA などを実装するハードウェアとしても利用可能と考えられている。今後のアーキテクチャやソフトウェア研究の進展によっては汎用用途のコンピュータで大幅な効率改善が可能といわれている [34, 35]。

(2) CGRA

CGRA（Coarse Grained Reconfigurable Array）は、再構成単位をビット単位ではなくワード単位で行うハードウェアであり、消費電力効率を 10 ～ 100 倍高くする可能性があるとされて現在研究段階ではあるが、特定の計算ドメインに対しハードウェア・アーキテクチャの一定の開発が行われて実用化も始まりつつある。CGRA は高速化したいアルゴリズムに合わせたハードウェアのチューニングが重要であり、汎用品の普及を待って利用を開始するプラットフォームではない。このため、課題はプログラミングで、既存の汎用 CPU と同じような言語によるプログラミングは未だに困難な状況である。AI も含めた応用分野の開拓など今後の改善が期待される。デバイス技術としては現行の CMOS プロセスを利用するために実用化に近い位置にあると考えられ、10 年以内程度の実用化が期待される [36, 37]。

(3) ニューロモーフィックコンピュータ

神経回路を模擬したシナプス素子とニューロン素子により積和演算を模倣するハードウェアで、厳格な演算精度が要求されない AI の計算に用いられている。研究レベルで利用されている

が商用化レベルではないようである。やはり既存の製造プロセスが適用可能なため、実用化に近い位置といわれている [38, 39]。

(4) デジタルアニーリング

CMOS プロセッサーで量子コンピュータシミュレーターを作って確率的処理を行うデジタルアニーリングについては、実用化に近付いているといわれており、特定の課題で既存コンピュータより数桁高速と報告されている [40, 41]。

(5) ストカスティック演算

0 と 1 の乱数系列を使って確率的に計算を行うストカスティック計算が省エネルギー型計算アーキテクチャとして提案されている。これも CMOS 技術が適用可能とされるものの、演算精度は落ちることから、応用分野の開拓も含めて、実用レベルには 10 年程度を要するといわれている [42]。

(6) シストリックアレイ

シストリックアレイは行列演算に特化した計算アーキテクチャで、演算器の中に入力データを残しながら演算を行う。このためにメモリと演算器の間のデータ移動が抑えられて高速かつ低電力に計算できるという特徴があり、グーグルの TPU にそのアーキテクチャが採用されている。

デメリットとしては行列演算しか実行できず、汎用性は CPU/GPU に劣るため、アクセラレータとしての使用が考えられている [34, 43-45]。

5.3 量子コンピュータ

量子コンピュータは量子ビット（qubit）という、0、1 およびそれらの重ね合わせ（同時実現）をとり得るビットを用いて計算するコンピュータで、従来の 0 と 1 を用いる 2 値演算とは全く原理が異なり、量子力学特有の重ね合わせ原理を用いて計算できるため、膨大な演算を短時間でできる可能性があるとされている。量子コンピュータとして、大きく分けて量子ゲート方式と量子アニーリング方式とが提案されていて、難解な理論に基づくので詳細には文献を参照されたい [46-49]。

量子ゲート方式は、将来プログラム可能な汎用機となることを目指している。量子ビットの作り方としていくつかの方式があり、超電導量子ビット、イオントラップ、光量子ビット、量子ドットなどが提案されている [47]。量子コンピュータは確率的な処理を行うために何回も演算を行う必要があり、最終的な完成形では誤り訂正機能がついたエラー耐性量子コンピュータであるといわれている。現在は、その手前の誤り訂正機能のない量子コンピュータ（NISQ: Noisy Intermediate-Scale Quantum device）、例えば 27qubit の超電導量子コンピュータプロトタイプ（IBM Quantum System One）が試作されている段階である。まだこの程度では実用的ではなく、単純な材料シミュレーションの実行に必要な量子ビット数は最低でも数百量子ビットといわれていて、極めて限定された用途では今後 10 年以内程度での実用化が期待されている。

実用的な複雑なシミュレーションには数百万量子ビット以上が必要と予想されていて現状との乖離は大きい。もしも完成した場合には桁違いの計算速度が可能とされているために期待は大きいが、この方式のハードウェアの製造は技術的に極めて難しく、実用化には相当長期的な研究が必要と考えられている [46]。

量子アニーリング機はイジングマシンと呼ばれる装置の一種であり、磁気スピンのエネルギーが最低状態になることを利用した計算方法で、最適化問題の解決など特定の応用に使用される。これは前述の量子ゲート方式よりも実用化が近いと考えられていて、すでに試作機が稼働している。

なお、イジングマシンには FPGA などの既存の CMOS 技術で構築したものもあり、この方が安定した大規模化の可能性が高いため量子アニーリング機へのステップとして検討されている。もっとも既存の技術によるイジングマシンは通常の CMOS プロセッサーに対して理論的な高速

性は示されていないとされている [47]。

現段階では、量子ゲート方式の本格な量子コンピュータについては、大規模な計算機が製作できるのか、また超電導方式の場合、動作環境が -273 ℃ という極超低温のため、装置のサイズや、冷却のための方法、必要エネルギーなど種々の検討が必要と考えられ、実用化時期はかなり遠い将来といわれ、2050 年に実用化されるかは不明である [48, 49]。

5.4 光コンピューティング

光コンピューティングは、光トランジスタとして、非線形光学効果を利用したもの [50]、位相差を利用したスイッチ [51]、光変調器による波長変換技術 [52, 53] などいくつか種類がある。光による演算機能は高速、かつ微弱エネルギーで可能だが、メモリ機能が原理的に小さく作れない [50-52]。そこでメモリ機能は電子的に、演算部分は光を使う光電融合技術が提唱されている [51-53]。

光による演算は電子の 1/1,000 以下のエネルギーで良いといわれているので、未来の省電力プロセッサーとして期待がある。課題として電気信号を光に変換、またその逆変換時にそれぞれ 30％のエネルギーを消費するとされている [50]。しかし NTT のグループは微小エネルギーでの変換が可能という発表をしている [52, 53]。この技術はシリコン半導体技術の転用が可能と思われ、素子作製技術、量産技術など将来的なスケールアップの実現性があり、今後が期待される。

もっとも、光コンピューティングの実現性に疑問を投げかける専門家もいるため、その技術的可能性についてはもう少し詳細な調査が必要と考えられる [50]。

5.5 その他

そのほかに分子コンピュータ、バイオコンピュータ、スピントロニクスコンピュータなども提案されているが、現在その消費電力低減効果を評価するには時期尚早と考える。

5.6 将来技術のまとめ

非ノイマン型コンピュータについては、実用化の可能性の高いものから未知のものまで幅広い。今後 10 年以内程度に実用化の可能性があるといわれている技術としては、CMOS 技術の適用が可能で製造技術的な不安が少ない非ノイマン型コンピューティングが挙げられ、ハードウェア、ソフトウェアの開発が必要ではあるが実現可能性は高いと思われる。次に実用化の可能性があるものとして量子アニーリングマシンが挙げられる。これは組み合わせ問題など極めて限定された用途に対して既存コンピュータの 1/1,000 以下という計算時間が可能とされていて、消費電力は下記量子ゲート方式と同等といわれている。

さらにその先の技術として、量子ゲート方式コンピュータや光コンピュータも高速性に加えて現在の 1/100、1/1,000 の電力消費という画期的な省エネルギーの可能性のために注目を集めている。量子ゲート方式コンピュータについては、NISQ が 2030 年くらい、誤り耐性形が 2050 年頃に実用化するのではないかという意見もあるが、まだ基礎研究段階で、多数のプロセッサーを一定の体積に収めるようなデバイス技術、回路技術、それらの量産技術など課題も多く、そもそも実用化されるかどうかも明確とはいえない。

この章で取り上げたどのコンピュータも利用方法としては CPU の代替というよりは、CPU の補助として膨大な計算部分を受け持つ利用法が考えられている。今後 10 ～ 20 年間のプロセッサーの省エネルギーは CMOS ベースの非ノイマン型コンピュータアーキテクチャと量子アニーリングマシンが中心となると考えられる。省電力効果としては、10 ～ 100 倍といわれている。また、量子コンピュータや光コンピュータは今後 20 ～ 30 年、その実現に向けた研究がなされていくと考えられるが現在のところその省電力効果や演算速度については評価できない。

6. メモリ技術

6.1 はじめに

メモリには大きく分けて揮発メモリと、不揮発メモリの 2 種類がある。

揮発メモリは RAM（Random Access Memory）と呼ばれ、プロセッサーの中に設置されて高速な読み書きが可能な SRAM とプロセッサー外に設置される大容量の DRAM に大別される。SRAM はキャッシュメモリとして利用され、DRAM はプログラムやデータの一時的格納をする。

メモリの 1 ビットはトランジスタとキャパシタの組み合わせで構成され、キャパシタの電荷量で 1 と 0 を判別する。DRAM においてキャパシタの電荷は時間とともに減少するため、一定時間（64 ms）ごとに電流を流して記録を更新することによって記録が保たれる。このために電荷を更新するための電力が必要となる。またキャパシタの電荷の有無を判別するための最低電荷量が存在する。

不揮発メモリは、読み出し専用の ROM（Read-Only Memory）、書き換え可能な EPROM（Erasable Programmable Read-Only Memory）がある。EPROM の中に電気的に書き換え可能な E2PROM（Electrically Erasable Programmable Read-Only Memory）があり、フラッシュメモリは、この 1 種である。ROM には書き換える必要のないプログラムなどが保存される。フラッシュメモリは、SD カード、USB メモリ、中期的に保存するデータ用メモリとして利用される [54]。

6.2 現行技術によるメモリの省電力

メモリの消費電力は、DRAM のインターフェースが 20 pJ、アクセスエネルギーが 20 pJ/bit 程度と言われる [55] 。したがって 1 バイトの読み出しに 300 pJ 程度が消費され、クロック周波数 2 GHz で動作するメモリでは、メモリの帯域幅の制限がなければ最大 0.6 mW/kB 程度の消費電力となる。実際にはメモリボードがメモリ帯域幅による制限を受けるため、メモリの容量によらずメモリボード 1 枚当たり約 5 W とされる。

メモリの省電力のためには、キャパシタの電荷量を小さく、すなわちプロセッサーと同様に微細化するほど省電力になるはずだが、熱雑音との区別のための最低電荷量が必要なことや、メモリの帯域幅の増加は高コストになること、メモリにかけられるコストなどから、微細化による高コストの吸収は難しく、ほぼ限界といわれている。

なお、不揮発メモリでは電荷を安定に保持するためには 60 k_BT 程度のエネルギーバリアが必要といわれていて（k_B : ボルツマン定数）、書き込みエネルギーも少なくともそれ以上が必要となる [54]。

メモリの省電力はメモリバス（経路）の短縮の効果が大きく、3D メモリなど立体的な配置が提案されている。また、DRAM などは常時リフレッシュするために電力を使用する。そこで直ちに使う必要のないデータを不揮発メモリに退避させ、不要メモリを電源から切り離す技術などが開発されている。

6.3 将来技術

メモリについては様々な方式が提案されている。磁気抵抗メモリ（MRAM）、相変化メモリ（PCRAM）、クロスポイントメモリ（XPoint）、抵抗変化メモリ（ReRAM）、強誘電体メモリ（FeRAM）、ナノチューブメモリ（NRAM）などである。

表４　次世代メモリの特性比較

	3DXPoint	MRAM	ReRAM	FRAM	DRAM	NANDFlash
サイクル寿命	2*10^5	10^9	10^6	10^18	10^15	10^5
記憶容量	128Gb	256Mb	4Mb	64Kb	16Gb	1Tb
書込み速度	10-100ns	10ns	100ns	16ns	10ns	100000ns
書込みエネルギー	Medium	Medium	Medium	Medium	Low	High
記録保持時間	300年間			10年間	0	
2018年価格	$0.5/Gb	$10-100/Gb	$100-1000/Gb		$1/Gb	$0.03/Gb
参考文献	[56,57]	[56,57]	[56,57]	[58]	[53,54]	[53,54]

これらのいずれもDRAMとは異なり､不揮発性メモリである。このためリフレッシュが不要で、省電力が期待でき、消費電力は1/10になるといわれる。もっとも、書き込みエネルギーが大きいと省電力にはならない可能性もある。さらに書き込み、読出し速度が遅いと、全体の演算速度が遅くなるため、DRAMの代わりには使えないが、安価であればストレージとしての利用が考えられる。

これらのうち、3DXpointはインテルからOptaneという商標で発売されていて、商業ステージであるが、その他のメモリは現在開発段階である。主な次世代メモリの種類と特性を表4にまとめた[54, 56-58]。

各種LSIメモリの詳細は参考文献[54]を参照されたい。

(1) 相変化メモリ（PCRAM、3DXPont）

相変化メモリは材料の相変化による物性の変化を記録として利用するメモリである。PCMは、カルコゲナイドガラスの結晶相が低抵抗でアモルファス相は高抵抗である事をデータの記録に利用する。相変化は電流によるジュール熱によるがレーザーを利用することも可能である。特徴は高密度化が可能であること、書き込み速度が速いこと、サイクル寿命が長いこと、記録保持時間が長いことであるが、相変化の加熱のため、書き込みエネルギーはDRAMに比べて大きいとされている。

3DXpointはインテルからOptaneという商品が発売されていて、相変化メモリであるといわれている。すでに64 GBメモリが市販されているので商用化されている技術である。DRAM並みのアクセス速度はないが、用途によっては実用上大差ないともいわれる。リフレッシュが不要で省電力であり、コストについてもDRAMと同等または低価格といわれているので、今後普及する可能性がある。

(2) 磁気不揮発メモリ（MRAM）

MRAMは磁性層（Ni,Fe,Coなどの合金）の磁化の向きを0と1に対応させるもので、書き込み方式として、(a) 磁界書き込み, (b) スピン注入書き込み、(c) 磁壁電流駆動書き込みが提案されている。磁界書き込みは既に実用化されている。MRAMの特徴は繰り返し寿命であり、ほぼ無制限といわれている。また、磁化は長期間の保持が可能であり、短時間で書き込みができる。問題点は高集積化と価格であると考えられる。

(3) 強誘電体メモリ（FeRAM, FRAM）

強誘電体メモリの研究開発の歴史は長く、既に30年以上にわたっている。基本は強誘電体の双極子モーメントの向きを記録として利用するものである。以前は短寿命とか、書き込み速度が遅いとか、集積化が難しいとかなどの難点があったが、近年改善が見られ、表4のように書き込み速度、記録保持時間、寿命の問題はないように思われる。また64 Kbitのメモリが試作されている段階であるが、今後集積化やコストの問題がクリアされれば実用化されると思われる。

(4) 抵抗変化メモリ（ReRAM）

基本的なセルは抵抗変化材料である金属酸化物を電極で挟んだキャパシタ構造であり、セルにパルス電圧を印加することによりセルの抵抗値を変化させて情報を記憶する。電圧で書き換えるため（電流が微量で）消費電力が小さく、比較的単純な構造のためセル面積が約 6 F^2（F は最小加工寸法で、数十 nm 程）と小さく、高密度化が可能で、電気抵抗の変化率が数十倍にものぼり、多値化も容易、かつ、読み出し時間が 5 ns 程度と、DRAM 並に高速といわれている。

現在 MB クラスのメモリが製造されていて、ほぼ商用化といえる段階である。今後 DRAM と競争し得るようなコストダウンが可能になれば普及するものと思われる。

情報をセルの抵抗値の違いとして記憶する不揮発性メモリで、電気的に書き込み・消去が可能である。

6.4 まとめ

現在の DRAM はコストも勘案すると微細化の余地はあまりないといわれている。このため容量増大は 3D 化、省電力はメモリーバスの距離短縮などが検討されている。また、直ちに使用しないデータを不揮発メモリに退避させ、不要メモリの電源を切るなどの対策が進められている。今後10年程度ではおそらく現在から2～5倍程度のエネルギー効率の改善にとどまると思われる。

その先であるが、将来技術としての DRAM 代替不揮発メモリは現在研究開発段階であり、省電力のポテンシャルとしてはあるが、実用化段階での性能の推定はまだ難しい。もっとも、相変化メモリの 1 種として実用化されている Optane は消費電力が低く、コスト的にも低下してきているという。より広い分野で DRAM と競合するために今後のさらなる研究開発が望まれる。

7. ストレージ

7.1 はじめに

現在サーバ用ストレージとしては、HDD（Hard Disk Drive）と SSD（Solid State Drive）とが、長期保存用には磁気テープがある。ストレージは電源が遮断された場合でもデータを保存できることに最大の目的があり、さらに膨大なデータを一定期間安定に蓄え、比較的高速にメモリに読み込める必要がある。なお磁気テープは大量データ長期保管用で消費電力が少ないがアクセスに時間がかかる。通常のサーバストレージとは用途が異なり消費電力も少ないため本報告では検討を省略する。

HDD は長い間ストレージの主流で、回転する磁気ディスク上にデータを保存する形式のため、一定期間の使用で機械的に損耗すること、機械的振動に弱いこと、アクセス時間が SSD に比較して長いことが短所であるが、ビット当たりのコストが非常に低いという特徴がある。

SSD は NAND 型フラッシュメモリで、半導体メモリのため消費電力が少なく、高速アクセス可能という特徴がある。ただし、近年価格低下しているとはいえ、ビット当たりのコストは HDD に比較して高い。ストレージは常時稼働するわけではないので、省電力への寄与は計算コアほどではないが、故障や小型化も考えると、SSD の比重が増加するが、増加するデータ量から安価なストレージに対する需要も根強く、SSD と HDD は当面は併用されていくと思われる。

7.2 現行技術

7.2.1 現行 HDD のエネルギー消費

HDD を定常の回転状態にまで上げる電力は通常 5.25 W、アイドル時は 5 W といわれているので、回転に必要な電力はほとんど変わらない [59]。これは前回の基準 HDD の 8 W よりも低下しているが、今後の改善は見込みにくい。これに対して、データの読み書きの消費電力は、読み出しが約 13.3 μW/KB、書き込みが 15 μW/KB となる [59]。HDD のデータ転送速度はディスク回転

数と線密度で決まり回転数はほぼ上限のため、物理的に上昇の余地は小さいとされている。シーケンシャルな読み書きは 200 MB/s 程度であるが、実際に使用されるランダムアクセスでは読み出し、書き込みともに 2 MB/s 程度というデータがある [60, 61]。そうすると 2 MB/s のデータ転送では、読み出しの消費電力は 26.6 mW、書き込みが 30 mW となる。結局機械的な回転に消費される電力が圧倒的に大きく、データ処理のための消費電力は無視し得る程度に小さいため、一台当たりの記録容量を大きくすると、容量当たりの消費電力は低下する。

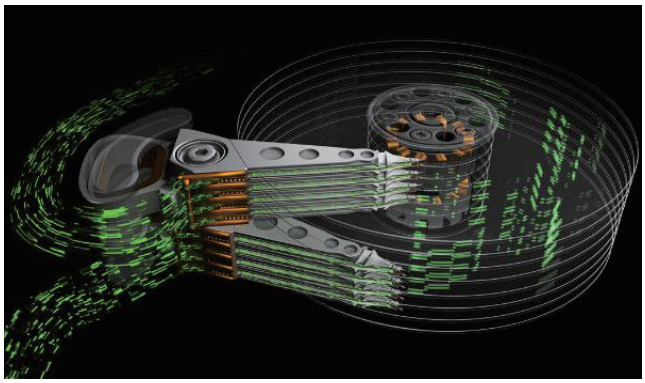

図 9 複数プラッターの HDD の構造
（日本シーゲイト HP より [62]）

また、記録容量を増加するためには HDD の記憶メディア（プラッターと呼ばれる）を複数枚（例えば 8 ～ 10 枚）重ねている（図 9）[62]。これにより物理的な記録面積は増加している。

7.2.2 SSD

従来型（2D）NAND 型フラッシュメモリの 16 nm/15 nm を超えた微細化は経済的ではないとされる。そこで、各社ともチップ内 3 次元積層化により、ビット密度（Gb/mm^2）を継続的に増加させようとしている。今後も図 10 に示すように 3D による密度増加が続き、既に 1 テラビット SSD がある [63]。ただしデータ容量と価格の関係は残る。

読み込み・書き込み速度はシーケンシャルで 0.5 ～ 4 GB/s、ランダムで 0.05 ～ 2 GB/s 程度というデータがある。

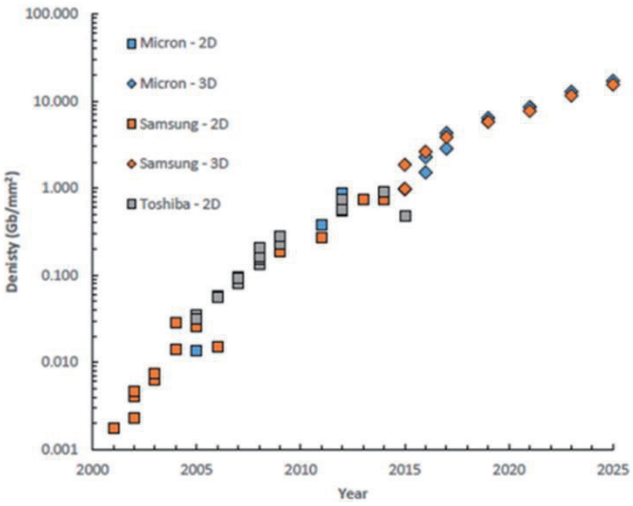

図 10　Micron、Samsung、東芝の NAND 型フラッシュメモリの記憶容量の変遷（Gbit/mm^2）[63]

7.3 将来技術

7.3.1 HDD

HDD の記録密度向上については「エネルギーアシスト磁気記録」技術がある。これは、書き込み時に記録媒体に外部からエネルギーを与えることで抗磁力を一時的に小さくし、磁化反転による書き込みを容易にするもので、「熱アシスト磁気記録（HAMR: Heat-Assisted Magnetic Recording）」方式と「マイクロ波アシスト磁気記録（MAMR: Microwave Assisted Magnetic Recording）」方式がある [64]。さらに、自己組織化媒体やパターンドメディアなど、記録密度向上の研究がされている。これらにより記録容量は前回基準の 2 ～ 4 TB から、現在は 8 ～ 10 TB、今後 10 年程度で 50 ～ 100 TB まで向上すると予測されているが [63, 65]、面積的には技術の限界に近いといわれている。また機器の構造上、消費電力のほとんどが回転エネルギーであることから、一台当たりの消費電力低減は難しいが、電力効率（W/Byte）は今後 10 年で 10 ～ 20 倍程度は効率化されると推測される [63]。

7.3.2 SSD

3D NAND 型メモリのロードマップ（Intel/Micron 連合の場合）を表 5 に示す。

表 5　NAND メモリのロードマップ例 [63]

Year	2016	2017	2019	2021	2023
Stacks	1	2	2	2	3
Layers/stack	32	32	48	64	64
Bits/cell	2/3	2/3	3	3	3
Channel type	Poly	Poly	Poly	High mobility	High mobility
Peripheral logic	CMOS under	CMOS under	CMOS under	CMOS under	CMOS under
Gbit/mm2	1.52/2.28	2.86/4.29	6.4	8.54	12.8

最左欄上から、西暦（年）、スタック数（積層したチップ数）、スタック当たりの層数、セル当たりのビット数、チャネルの構成材料（poly-Si あるいは高移動度材料）、周辺ロジック回路の位置（メモリアレイの下）、密度（Gb/mm^2）である [63]。メモリ密度としては 4 年で 2 倍程度が見込まれていて、仮に 1 枚当たりの消費電力は従来と変わらないとすれば電力効率は 4 年で 2 倍程度の向上が期待される。

図 11　3D NAND Flash Memory[66]

CMOS 周辺回路をメモリアレイの下に配置するとダイ上のメモリアレイの面積の割合を増加できるが、ストレスを増加させるなどもあり、一部の採用にとどまっているという。積層レイヤが増えるにしたがって、チャネル移動度が小さくなるので、高移動度材料に置き換える必要も指摘されている。

2D および 3D NAND フラッシュメモリを製造するのに必要な設備投資額のプロセスカテゴリ別割合（%）を表 6 に示す [63]。3D の場合は、リソグラフィの負担は減るが、100 層に及ぶ多層膜の体積のための CVD（Chemical Vapor Deposition: 化学的気相堆積）、深堀りエッチング用のドライエッチング、SiN 膜を除去してタングステン柱状電極を埋め込むための ALD（Atomic Layer Deposition: 原子層堆積）などで、これらの比率が倍増する点が注目される。

表 6　NAND Flash Memory 製造に必要な設備投資のプロセス別割合 [63]

	16nm 2D NAND	32L 3D NAND
ALD/CVD/dry etch	22%	47%
Inspection/ metrology	21%	15%
Lithography	38%	18%
Other	19%	20%

7.3.3 その他

現在は基礎研究の段階であるが、分子メモリとか、不揮発性メモリの項で述べた磁気メモリなどもかつての NAND 型フラッシュメモリと同様に価格次第ではストレージに応用される可能性はある。

7.4 まとめ

HDD は回転の電力消費が大部分という構造の特性上、機器自身の大幅な省電力は見込みにくく、容量増による容量当たり消費電力の向上が主となる。これは今後 10 年で 50 ～ 100 TB/ 台まで向上すると予測されるが、その先の技術見通しは不透明である。さらにソフトウェアで不要な HDD の電源を切るなどの節電対策も併せて、今後 10 年で消費電力効率（GB/W）向上は現在の 10 ～ 20 倍程度と推定され、2050 年に 30 倍以上になるかどうかは見通せない。

NAND 型フラッシュメモリは大容量化と価格の低下がまだ進行すると予測される。ビット当たりの消費電力も低下し、今後 10 年で 5 ～ 6 倍程度の電力効率の向上も期待できるため、今後ストレージにおいて SSD の比率が増加すると考えられる。一方で今後も膨大なデータ量の流通が予測されるため、低コストの HDD の需要も残り、比較的高速アクセス用の SSD と低頻度アクセスで長期データ保管用の HDD の併用、詳細には触れなかったがさらに低頻度アクセス向けかつ低消費電力のテープ型ストレージも含め、コストとパフォーマンスの点から選択されそうである。

省電力の観点からは、SSD は読み出し、書き込み電力の低減も課題である。ただ、キャパシタに低電圧で電荷を注入するということは低い熱エネルギーで電荷が失われる可能性も高くなり、データの保存性との兼ね合いで限界があると思われる。

今後解決すべき課題としては、半導体メモリ共通の課題としてデータの書き込みと読み出しのための配線が各セルに必要で、この微細化が難しく、配線断面積を確保するためのトレンチ構造とか新配線材料の探索などが行われている。これから考えると、今後は配線の問題で微細化の進展は困難と思われ、また消費電力の低下はほぼ限界に近いと思われる。

この課題は将来技術として提案されている微細メモリ（分子メモリなど）にもあてはまり、素子を小さくしても配線が小さくできなければ微細化も省電力もあまり期待できないことになる。

8. まとめ

データセンターの省エネルギー技術について、その消費電力の大部分を占めるサーバを中心に検討した。特にサーバの中でもエネルギー消費の大きいプロセッサーに重点を置いて検討し、ついでメモリとストレージについても検討した。

8.1 機器の省エネルギー技術の検討

（1）プロセッサーに関しては、今後 5 ～ 10 年における現行技術の改善、特にマルチコア技術、微細化技術、統合化技術などにより、エネルギー効率は、改善程度が低い Modest ケースで CPU が 2 倍程度、GPU が 5 倍程度、予想の上限程度に進捗する Optimistic ケースではアクセラレータ等も利用され、CPU が 10 倍程度、GPU が 20 倍程度のオーダーで改善するのではないかと推測する。（表 7, 8 の“Assumed Power consumption efficiency”参照）

さらに、次世代技術として、おそらく今後 10 ～ 20 年にわたって新計算原理コンピュータとしての非ノイマン型 CMOS コンピューティングが、ついで量子アニーリングがともにアクセラレータとして AI などの計算に使用されると予想される。これら技術による現在からのエネルギー効率改善幅は 10 ～ 100 倍といわれていてこれは実用化される見込みが高い。

以上より、現状実用化の可能性が高いと見込まれる技術を組み合わせると、既存技術改善部分で CPU は 2 ～ 10 倍、GPU は 5 ～ 20 倍、新計算原理部分で 10 ～ 100 倍であるから、これらによ

り、CPU で 20 ～ 1,000 倍、GPU で 50 ～ 2,000 倍程度まで達成される可能性があると推定する。

光電融合を含む光コンピューティングについては素子作成技術が現在のシリコン半導体技術の転用が可能とされているところから、その次の世代の技術と考えられる。量子ゲート型コンピューティングはさらにその後と推測され、実用化は 2040-2050 年頃といわれている。これら技術についてはまだ集積化される段階ではなく、ポテンシャルとして 100 ～ 1,000 倍以上ともいわれるが、実用化時のエネルギー効率改善の程度は現段階では予測できない。

（２）メモリについては現行技術では、微細化もほぼ限界といわれ、エネルギー効率の改善余地は少なく、主にソフトウェアで不要なデバイスの電源を切るなどの省電力対策が考えられ、今後 10 年では、おそらく現在の水準から 2 倍から 5 倍程度の電力効率の改善にとどまるものと推測する。

将来技術に関してメモリは不揮発メモリの方向で研究が進んでいる点でストレージと共通点がある。既述のように速度や省電力の点で可能性のある技術が見出されているが、現段階での評価は難しい。

（３）ストレージについては、HDD はエネルギー効率としては現行技術の延長で 10 ～ 30 倍程度の向上が期待できると考えられるが、その先の向上は現在のシステムでは難しく全く新しいシステムが必要と思われる。SSD については今後 10 年で 4 ～ 8 倍程度の効率向上が期待されるが、その先はメモリにも用いられる新方式の記録技術の開発が期待される。

（４）なお、本文中では触れなかったが、データセンター内では電圧の変換部分、交流と直流の変換個所などが数多く存在し、各変換部分にはインバータやコンバータといったパワーデバイスが用いられている。使用箇所が膨大なため変換ロスは無視しえない程度になるため、パワーデバイスの研究はデータセンターにとっても重要である。同様に設備冷却システムも重要であるが高効率データセンターの PUE が 1.1 レベルまで下がっていること、および改良技術的側面が強く基礎研究的な部分が少ないことから本提案書では触れなかった。

8.2 消費電力の推定

前回の報告書 [2] で AI 業務の消費電力の計算において DL（ディープラーニング）用チップセット出荷台数とサーバ出荷台数を等しいとした。その後専門家から必ずしもサーバに 1 チップとは限らず ASIC も多すぎるとのご指摘、また、AI サーバは DL 用が中心ではあるが、それに限られるものでもないことなどを勘案し、世界の消費電力については、DL 用チップセット 1.5 台につき AI サーバ 1 台として計算しなおすことにした。結果を下記予測と併せて国内を表 7 に、世界を表 8 に示した。また、後述のような Modest ケースと Optimistic ケースについてのデータセンター消費電力予測を、現行技術基準（橙線）と比較して、国内については図 12 に、世界については図 13 に示した。

（1）2030 年消費電力

消費電力予測は 2018 年を起点として考える。基準とした CPU（Xeon Gold 6242）のエネルギー効率は 6.67 Gflops/W、GPU（TeslaV100）のそれは 18 Gflops/W であった。現在（2021 年）最上位機種はもう少し高性能化が進んでいるが、消費電力の計算にはサーバに広く使用される中位グレードで考え、かつ、前回のレポートとの整合性も考慮して、前回レポートの値を基準とする。

2030 年においてサーバのベース業務部分消費電力の内 CPU の寄与分を 50%、GPU による部分を 50%と仮定し、AI 業務部分は CPU 寄与分を 10%、GPU（含むアクセラレータ）寄与分を 90%と仮定した。

メモリのエネルギー効率（記憶容量当たりの消費電力）は 2 ～ 5 倍程度、ストレージのエネルギー効率（記憶容量当たりの消費電力）は現行の 10 ～ 30 倍程度の改善と推定して、それぞれ表下段に示した値を用いて表にまとめた。また、ネットワークスイッチについては今回検討しなかったが、全く効率が改善しないと仮定するのも妥当でないので、メモリと同等（改善率としては低め）

と仮定して計算した。全体に占める割合は低いのでこの程度の仮定でもあまり結果に影響はないと考える。その他については機器類の合計の 20%とした。

2030 年のエネルギー効率は表下段にまとめたように 2018 年を基準として Modest ケースでは CPU：1/2、GPU やアクセラレータ：1/5、メモリ、スイッチ：1/2、ストレージ 1/10 とした。また Optimistic ケースでは、CPU：1/10、GPU やアクセラレータ：1/20、メモリ、スイッチ：1/5、ストレージ 1/30 とした。

この結果 2030 年のデータセンターの消費電力推定として Modest ケースで国内 24 TWh、世界 670 TWh、Optimistic ケースで国内 6 TWh、世界 190 TWh と推定された。このうちサーバ消費電力は Modest ケースで国内 17 TWh、世界 510 TWh、Optimistic ケースで国内 5 TWh、世界 140 TWh と推定された（図 12, 13）。

（2）2050 年消費電力

2050 年の消費電力予測は困難で、信頼性に乏しい。ここでは Modest ケースは 2030 年までと同等の改善率で進捗するとした（10 年間では CPU で 1/2、GPU で 1/5、メモリ、スイッチで 1/2 など）が、ストレージの効率向上は根拠に乏しいので 2018 年比 1/30 とした。Optimistic ケースは CPU で 1/200、GPU で 1/1,000、メモリ、スイッチ、ストレージで 1/100 まで改善すると仮定した（表下段参照）。

新計算原理による種々のアクセラレータ等の技術が進むと考えられるので、それを GPU 分にカウントしてベース業務部分の消費電力の内 CPU の寄与分を 20%、GPU による部分を 80%と仮定した。AI 業務部分は計算部分が増加するために CPU 寄与分がさらに低下するとして 5%、GPU（含むアクセラレータ）寄与分を 95%と仮定した。

この結果 2050 年のデータセンターの消費電力推定として Modest ケースで国内 500 TWh、世界 16,000 TWh、Optimistic ケースで国内 110 TWh、世界 3,000 TWh と推定された。このうちサーバ消費電力は Modest ケースで国内 330 TWh、世界で 11,000 TWh、Optimistic ケースで国内 50 TWh、世界 1,600 TWh と推定された（図 12, 13）。

この結果において日本の消費電力が世界と比べて相対的に低く推定されていると思われる。世界では消費電力の伸びが大きい AI 業務部分が大きく、基準時（2018 年）のベース業務と AI 業務の比率の差がより拡大したためである。この分野の出荷統計は公式のものがなく、また AI の定義も不明確であるので、この程度は誤差の内と思われるが、あるいは現状、日本が AI 分野で出遅れ気味なのかもしれない。

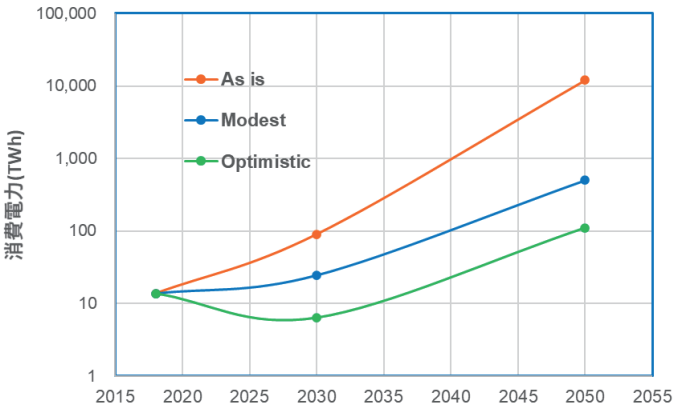

図 12　国内データセンター消費電力推定 (TWh)

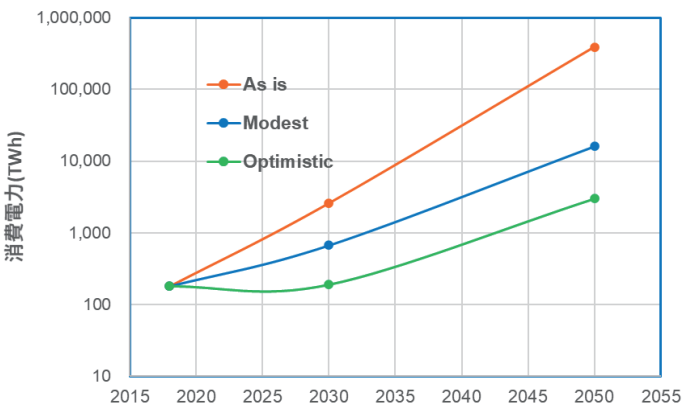

図 13　世界データセンター消費電力推定 (TWh)

表７　国内データセンター消費電力推定 (TWh)

				Japan						
			Year	2018	2030	2030	2030	2050	2050	2050
					As is	Modest	Optimistic	As is	Modest	Optimistic
IP traffic			ZB	0.7	11	11	11	1,400	1,400	1,400
power consumptions of data centers			TWh	14	90	24	6	12,000	500	110
power consumptions of servers		basic task	TWh	6	30	13	3	3,500	229	39
		AI task	TWh	0.7	16	4	1	3,000	97	14
		total	TWh	7	46	17	5	6,500	330	50
	CPUs	basic task	TWh	4	20	7	2	2,200	75	24
		AI task	TWh	0.5	12	3	1	2,300	37	8
		total	TWh	4	32	10	2	4,500	110	30
	memories	basic task	TWh	1	7	4	1	890	116	9
		AI task	TWh	0.1	2	1	0	340	44	3
		total	TWh	1	9	4	2	1,200	160	12
	power supply etc	basic task	TWh	1	3	2	1	410	38	7
		AI task	TWh	0.1	2	1	0	400	16	2
		total	TWh	1	5	3	1	810	54	9
power consumtions of storages			TWh	2	29	3	1	3,700	110	40
power consumptions of switches			TWh	0.1	1	1	0	70	9	1
power supply, cooling, etc			TWh	5	11	4	1	1,500	90	20
Assumed power consumption		CPU		1	1	0.5	0.1	1	0.13	0.05
efficiency		accelarators(GPU etc)		1	1	0.2	0.05	1	0.01	0.001
		memories		1	1	0.5	0.2	1	0.13	0.01
		storages		1	1	0.1	0.03	1	0.03	0.01
		switches		1	1	0.5	0.2	1	0.13	0.008

表８　世界データセンター消費電力推定 (TWh)

				World wide						
			Year	2018	2030	2030	2030	2050	2050	2050
					As is	Modest	Optimistic	As is	Modest	Optimistic
IP traffic			ZB	11	170	170	170	20,200	20,200	20,200
power consumptions of data centers			TWh	180	2,600	670	190	390,000	16,000	3,000
power consumptions of servers		basic task	TWh	90	450	190	50	53,000	3,400	590
		AI task	TWh	15	1,200	320	90	221,000	7,200	1,000
		total	TWh	105	1,700	510	140	270,000	11,000	1,600
	CPUs	basic task	TWh	60	280	100	20	33,000	1,100	360
		AI task	TWh	11	880	200	50	167,000	2,700	580
		total	TWh	71	1,200	300	70	200,000	3,800	940
	memories	basic task	TWh	16	110	55	22	13,000	1,700	130
		AI task	TWh	2	130	65	26	25,000	3,300	250
		total	TWh	18	240	120	50	38,000	5,000	380
	power supply etc	basic task	TWh	14	60	31	8	7,000	600	100
		AI task	TWh	2	150	53	15	29,000	1,200	170
		total	TWh	16	210	80	20	36,000	1,800	270
power consumtions of storages			TWh	27	430	43	13	51,000	1,500	510
power consumptions of switches			TWh	2	20	10	4	3,400	440	30
power supply, cooling, etc			TWh	43	400	110	30	66,000	2,600	430
Assumed power consumption		CPU		1	1	0.5	0.1	1	0.13	0.05
efficiency		accelarators(GPU etc)		1	1	0.2	0.05	1	0.01	0.001
		memories		1	1	0.5	0.2	1	0.13	0.01
		storages		1	1	0.1	0.03	1	0.03	0.01
		switches		1	1	0.5	0.2	1	0.13	0.008

9. 政策提言

まとめの章で示したように、2030 年までは現在の需要の伸び率が続くと仮定すると、現行技術の改善によりデータセンターの消費電力は許容不可能な状態にまでは達しないと考えられる。一方で、今後 10 年程度で現行技術は限界に到達すると考えられることから、AI の社会への浸透がさらに進んで自動運転なども実用化されるであろう 2050 年を見通すときには革新的な新技術の開発が求められる。

今後の技術として AI は膨大な計算資源を消費すると考えられるため、特に CPU および CPU を補完する計算機能の研究開発が極めて重要である。CPU の開発においては現在の延長線上にある微細加工技術や 3D、チップ化などの技術の推進は直近の課題解決として重要である。将来を考えると、人的、資金的資源の制約を考慮して実用化の可能性の高い技術に重点を置いて展開することが必要で、今後 10 年間では CPU も含めたデータ移動が少なく簡素な制御部からなるアーキテクチャの開発、および CMOS 技術を利用した非ノイマン型コンピューティングアーキテクチャ、やや遅れて量子アニーリングマシンの実用化が期待され、重要テーマと考えられる。次の 20 ～ 30 年で量子ゲート型、光コンピュータなどの技術の実用性や電力効率が評価可能なレベルに達すると考えられる。

これら技術はアーキテクチャ、ハードウェア、ソフトウェアなど広範な分野にまたがり、研究開発期間も長期にわたり、同時にこれを支える人材も当該専門領域に加えて、生物学、物理学など広範な分野の専門家が長期間大量に必要とされる。これに伴い異分野の知識・技術を融合して独創的な技術開発につなげるような人材育成・体制構築も極めて重要と考えられる。

したがって基礎研究、応用研究、人材育成に重点的に資金を継続的に投入する必要がある。必要な人材の確保のためには学生の教育に加えて産業構造の転換を伴う当該分野への人材の再配置のための社会人再教育も必須と考えられる。

メモリは、現行メモリ技術についてその改善幅が限られること、また設備産業のために既存大企業の競争力が強いと考えられることから、基礎研究の成果が上がる余地は少ないと思われる。むしろ相変化型や磁気メモリなど次世代不揮発メモリの研究開発が重要と考えられる。

ストレージについては HDD や磁気テープは容量、コストなどから将来も残っていくものと考えられるが、画期的な不揮発メモリがストレージ用途にも用いられる可能性はあり、メモリ用途と合わせて不揮発メモリの研究開発に期待される。

また、電力変換効率向上という点ではパワーデバイスの研究は電力、一般産業のみならずデータセンターにとっても重要である。

10. 謝辞

本提案書の作成にあたり、貴重なご助言を賜りました理化学研究所計算研究センター　チームリーダー　佐野　健太郎氏、奈良先端科学技術大学院大学教授　中島　康彦氏、群馬大学研究協力員　中谷　隆之氏、東京工業大学特任教授　西森　秀稔氏、電気通信大学准教授　三輪　忍氏に心より感謝申し上げます。

参考文献

[1] 低炭素社会の実現に向けた政策立案のための提案書，“情報化社会の進展がエネルギー消費に与える影響（Vol.1）”，科学技術振興機構低炭素社会戦略センター，2019 年 3 月.
[2] 低炭素社会の実現に向けた政策立案のための提案書，“情報化社会の進展がエネルギー消費に与える影響（Vol.2）”，科学技術振興機構低炭素社会戦略センター，2021 年 2 月.
[3] 低炭素社会の実現に向けた政策立案のための提案書，“情報化社会の進展がエネルギー消費に与える影響（Vol.3）”，科学技術振興機構低炭素社会戦略センター，2021 年 2 月.
[4] Paul Wiener, “As Data Centers Focus on Density, GaN Looks Increasingly Attractive”, Industry Perspectives, mar.31,2021 (2021), https://www.datacenterknowledge.com/industry-perspectives/data-centers-focus-density-gan-looks-increasingly-attractive, (アクセス日 2021 年 6 月 21 日).
[5] European Commission, Brussels, 15.3.2019 SWD (2019) 106 final, “COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT” (2019), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52019SC0106&from=EN, (アクセス日 2021 年 6 月 21 日).
[6] L. A. Barroso, et al., "The Datacenter as a Computer: Designing Warehouse-Scale Machines, 3rd ed.", Morgan & Claypool Publishers (2018).
[7] 大窪 宏明，“人工知能を支えるハードウェア分野”，平成 30 年度 NEDO「TSC Foresight」セミナー（第 2 回），(2018).
[8] 佐野 正博，“Intel 社が開発したマイクロプロセッサーの技術的スペックの歴史的変遷（詳細版）”，https://www.sanosemi.com/history_of_Intel_CPU_techspecs0.htm, (アクセス日 2021 年 6 月 16 日).
[9] 笠原 一輝，“この 10 年の Intel CPU 進化の歴史をベンチマークとともに振り返る”，https://pc.watch.impress.co.jp/docs/topic/special/1262523.html, (アクセス日 2021 年 6 月 16 日).
[10] インテル ,“インテル製マイクロプロセッサーの歴史”, https://www.intel.co.jp/content/www/jp/ja/innovation/hof.html, (アクセス日 2021 年 6 月 16 日).
[11] インテル製品仕様, https://ark.intel.com/content/www/jp/ja/ark.html#@Processors, (アクセス日 2021 年 6 月 16 日).
[12] Wikipedia, “Roofline Model”, https://en.wikipedia.org/wiki/Roofline_model, (アクセス日 2021 年 9 月 21 日).
[13] Hisa Ando, “コンピュータアーキテクチャの話 CMOS における消費エネルギーと動作速度の関係”，https://news.mynavi.jp/article/architecture-327/, (アクセス日 2121 年 6 月 16 日).
[14] 安藤 壽茂，“半導体技術とコンピュータ技術の動向。低電力，高性能なコンピュータの実現に向けて「次世代 HPC を支える基盤技術」”，サイエンティフィック・システム研究会 2006 年度科学技術計算分科会，(2006).
[15] Wikipedia，“マルチコア”，https://ja.wikipedia.org/wiki/%E3%83%9E%E3%83%AB%E3%83%81%E3%82%B3%E3%82%A2, (アクセス日 2021 年 9 月 21 日).
[16] Wikipedia, “Amdahl の 法 則 ”，https://ja.wikipedia.org/wiki/%E3%82%A2%E3%83%A0%E3%83%80%E3%83%BC%E3%83%AB%E3%81%AE%E6%B3%95%E5%89%87, (アクセス日 2021 年 9 月 21 日).
[17] William Dally, “Deep Learning and HPC”, Keynote Speech, NVIDIA Deep Learning Institute 2017, (2017), https://images.nvidia.com/content/APAC/events/deep-learning-institute-jp/2017/pdf/keynote-nv-bill- dally.pdf, (アクセス日 2021 年 6 月 15 日).
[18] 前沢 宏一，“量子効果デバイス　第 11 回”， http://www3.u- toyama.ac.jp/maezawa/QED11/QED11.

pdf,（アクセス日 2021 年 6 月 16 日）.
[19] Wikipedia,“デナード則”,
https://ja.wikipedia.org/wiki/%E3%83%87%E3%83%8A%E3%83%BC%E3%83%89%E5%89%87,（アクセス日 2021 年 6 月 16 日）.
[20] 須黒 恭一，江口 和弘,“高性能トランジスタ技術”，東芝レビュー，vol.59, No.8（2004），https://www.global.toshiba/content/dam/toshiba/migration/corp/techReviewAssets/tech/review/2004/08/59_08pdf/a04.pdf,（アクセス日 2021 年 6 月 17 日）.
[21] 福田 昭,“福田昭のセミコン業界最前線　EUV を使わずに微細化の極限を目指す半導体製造技術 ”，PCWatch,（2018），
https://pc.watch.impress.co.jp/docs/column/semicon/1099661.html,（アクセス日 2021 年 10 月 21 日）.
[22] 中谷 隆之,“2020 年版半導体技術の概要と動向”，第 452 回 アナログ集積回路研究会講演,（2021），https://kobaweb.ei.st.gunma-u.ac.jp/lecture/20210129_nakatani-sensei.pdf,（アクセス日 2021 年 10 月 21 日）.
[23] 福田 昭,“福田昭のセミコン業界最前線 2020 年も半導体はおもしろい（前編）”，PCWatch,（2020），https://pc.watch.impress.co.jp/docs/column/semicon/1232236.html,（アクセス日 2021 年 10 月 21 日）.
[24] 大島 篤,“Core マイクロアーキテクチャに迫る”，ITmedia PC-USER,（2007），
https://www.itmedia.co.jp/pcuser/articles/0701/24/news002.html,（アクセス日 2021 年 6 月 21 日）.
[25]“知られざる CPU の過去 40 年における性能向上と進化の歴史”，Gigazine（2013），
https://gigazine.net/news/20130725-40-year-cpu-history/,（アクセス日　2021 年 6 月 21 日）.
[26] 平 洋一,“将来のコンピューティング”, エレクトロニクス実装学会誌, vol.15, No.6, 430-4（2012）.
[27] 後藤 弘茂,“広帯域と大容量にフォーカスした“第 2 世代”の HBM2 メモリ”, PC Watch,（2018），
https://pc.watch.impress.co.jp/docs/column/kaigai/1112390.html,（アクセス日 2021 年 8 月 19 日）.
[28] Takashi NAKADA, et al.,“An Energy-Efficient Task Scheduling for Near Real-Time Systems on Heterogeneous Multicore Processor”, IEICE TRANS. INF. & SYST., VOL.E103-D, NO.2, 329-38 (2020).
[29] NVIDIA ウェブサイト，https://www.nvidia.com/ja-jp/data-center/a100/,（アクセス日 2021 年 8 月 20 日）.
[30] elsa-jp ウェブサイト，https://www.elsa-jp.co.jp/products/detail/nvidia-a100/?tab=specification,（アクセス日 2021 年 8 月 20 日）.
[31] 山道 新太郎,“次世代コンピュータの中核となる新発想のチップを日本企業とともに創る”, TELESCOPE Magazine, No.011, (2016).
[32] Stephen W. Keckler,“GPUs and the Future of Parallel Computing”, IEEE Micro 31(5): 7-17 (2011).
[33] William Dally,“From Here to Exascale Challenges and Potential Solutions”, Slide Serve Online Presentation (2014) ,
https://www.slideserve.com/brendy/from-here-to-exascale-challenges-and-potential-solutions,（アクセス日 2021 年 8 月 20 日）.
[34] 石村 俊介 , 早川 雄貴，金杉 昭徳,“動的再構成可能なシストリック・アレイの一構成法と FPGA 実装（専用システム，ネットワーク技術及び一般）”，電子通信情報学会研究発表会, 2008 年 12 月,（2008）.
[35] 佐野 健太郎,“高性能計算に革新をもたらす非ノイマン型 FPGA オーバーレイアーキテクチャの創出”，科学研究費助成事業研究成果報告書，課題番号 17H01706 基盤研究（B），(2020)，
https://www.semiconportal.com/archive/blog/insiders/hattori/210401-intelidm2.html,
https://www.tel.co.jp/museum/magazine/artificial_intelligence/160729_interview/index.html,（アクセス日 2021 年 8 月 20 日）.
[36] Y. Nakashima, EMAX6/ZYNQ64 (IMAX2) Architecture Handbook (2021)

- In Memory Accelerator eXtension -,
http://arch.naist.jp/proj-arm64/doc/emax6/emax6j.pdf , (アクセス日 2021 年 8 月 20 日).
[37] T. Kojima, et al., “Real Chip Evaluation of a Low Power CGRA with Optimized Application Mapping”, Proceedings of the 9th International Symposium on Highly-Efficient Accelerators and Reconfigurable Technologies , (2018).
[38] Sally Ward Foxton, “SNN を加速するニューロモーフィック AI チップを開発”, EEtimes , Japan , https://eetimes.itmedia.co.jp/ee/articles/2108/04/news066.html, (アクセス日 2021 年 8 月 18 日).
[39] 福田 昭, “不揮発性メモリが切り拓く超低消費の AI ハードウェア”,
https://pc.watch.impress.co.jp/docs/column/semicon/1189183.html, (アクセス日 2021 年 8 月 18 日).
[40] 富士通ホームページ, “デジタルアニーラとは”,
https://www.fujitsu.com/jp/digitalannealer/superiority/, (アクセス日 2021 年 8 月 17 日).
[41] Cnet Japan, “8 億年分の計算を 1 秒で処理：量子のパワーをデジタルに転換した「デジタルアニーラ」の衝撃”,
https://japan.cnet.com/extra/fujitsu_201803/35115699/, (アクセス日 2021 年 8 月 17 日).
[42] 鬼沢 直哉他, “ストカスティック演算に基づく省エネルギー脳型設計技術”, IEICE Fundamentals Review, Vol.11, No.1, (2017) ,
https://www.jstage.jst.go.jp/article/essfr/11/1/11_28/_pdf/-char/ja, (アクセス日 2021 年 8 月 20 日).
[43] Wikipedia, “Systolic Array” , https://en.wikipedia.org/wiki/Systolic_array, (アクセス日 2021 年 8 月 20 日).
[44] 伊藤 元昭, “AI チップを巡って競い合う巨人たち”, TELESCOPE Magazine, No.15, (2017),
https://www.tel.co.jp/museum/magazine/015/report02_02/02.html, (アクセス日 2021 年 8 月 18 日).
[45] 山野 龍祐他, “時分割多重実行によるシストリックリングの面積効率向上手法”, 電子情報通信学会技術研究報告, Vol.117 ,No.44 , (CPSY2017 1-15), 27-32, (2017),
https://jglobal.jst.go.jp/detail?JGLOBAL_ID=201702260521632338, (アクセス日 2021 年 8 月 20 日).
[46] NTT データ, “量子コンピューティングガイドライン”, 量子アニーリング, (2021).
[47] 楊他, “量子コンピュータの基礎から応用まで”, https://speakerdeck.com/qunasys/quantum-summit-2019, (アクセス日 2021 年 8 月 20 日).
[48] JST-CRDS, 戦略プロポーザル, “みんなの量子コンピューター　～情報・数理・電子工学と拓く新しい量子アプリ～”, CRDS-FY2018-SP-04,
https://www.jst.go.jp/crds/report/CRDS-FY2018-SP-04.html, (アクセス 2018 年 12 月).
[49] EETimes Japan ホームページ, https://eetimes.itmedia.co.jp/ee/articles/2009/29/news050.html, (アクセス日 2021 年 10 月 5 日).
[50] Wikipedia, “光コンピューティング”, https://en.wikipedia.org/wiki/Optical_computing, (アクセス日 2021 年 10 月 5 日).
[51] Hisa Ando, “Lightmatter の光推論 AI アクセラレータ「Mars」を読み解く”, Hot Chips 32(1) Tech+, (2020),
https://news.mynavi.jp/article/20200911-1294262/, (アクセス日 2021 年 9 月 24 日).
[52] 野地 秩嘉, “光コンピューティングの夢”, NTT 技術ジャーナル, 2021.2, (2021), https://journal.ntt.co.jp/article/10409, (アクセス日 2021 年 9 月 24 日).
[53] “光変調器を超省エネ化し, 高速高効率な光トランジスタを実現～光電子融合型の超低消費エネルギー・高速信号処理へ前進～”, ニュースリリース, NTT,
https://group.ntt/jp/newsrelease/2019/04/16/190416a.html?_ga=2.133527117.2130210742.1632472815-1629277088.1632472815, (アクセス日 2021 年 9 月 24 日),
https://www.rd.ntt/research/CT99-348.html, (アクセス日 2021 年 9 月 24 日).
[54] 電子情報通信学会, “知識ベース”, 10 群 4 編メモリ LSI, (2010),

https://www.ieice-hbkb.org/portal/doc_468.html,（アクセス日 2021 年 9 月 24 日）.
[55] Thomas Vogelsang, “Understanding the Energy Consumption of Dynamic Random Access Memories”, 43rd Annual IEEE/ACM Int'l Sym. on Microarchitecture, MICRO ‘10 (2010).
[56] 福田 昭，“MRAM の市場規模，2024 年には 2018 年の 40 倍へと急伸 〜”，MRAM 開発者デー 2019 レポート，PC Watch (2019), https://pc.watch.impress.co.jp/docs/news/1200644.html,（アクセス日 2021 年 10 月 1 日）.
[57] 福田 昭，“福田昭のストレージ通信（157）半導体メモリの技術動向を総ざらい（16）”，EE Times Japan (2019), https://eetimes.itmedia.co.jp/ee/articles/1908/02/news023.html,（アクセス日 2021 年 10 月 1 日）.
[58] 福田 昭，“福田昭のセミコン業界最前線：富士通とソニー，IMW 2021 で次世代不揮発性メモリの開発成果を披露”，PC Watch（2021），
https://pc.watch.impress.co.jp/docs/column/semicon/1332748.html,（アクセス日 2021 年 10 月 1 日）.
[59] A. Lewis, et. Al, “Run-time Energy Consumption Estimation Based on Workload in Server Systems”, Proceedings of Workshop on Power Aware Computing and Systems, HotPower 2008, (2008).
[60] PC 自作 .COM,“究極ガイド：ストレージの選び方 ”，（2020），
https://pcjisaku.com/articles/blogs/storage,（アクセス日 2021 年 10 月 1 日）.
[61] CMAN インタネットサービス，WEB 便利ノート，“SSD/HDD の速度比較・選び方”，
https://note.cman.jp/hdd/ssd_hdd_speed/,（アクセス日 2021 年 10 月 1 日）.
[62] SEAGATE BLOG, “Multi Actuator Technology: A New Performance Breakthrough”,
https://blog.seagate.com/craftsman-ship/multi-actuator-technology-a-new-performance-breakthrough/,（アクセス日 2021 年 11 月 19 日）.
[63] 服部 毅，“どうなる次世代の半導体プロセス技術（3）10nm 以降の超微細化における問題点と注目すべき点”，TECH+ （2016），https://news.mynavi.jp/article/icknowledge-3/ ,（アクセス日 2021 年 10 月 3 日）.
[64] 堀内 義章，“2020 年のストレージと HDD の業界展望”，IDEMA Japan News，（2020）.
[65] Seagate's 2021 Virtual Analyst Event (2021), https://s24.q4cdn.com/101481333/files/doc_downloads/2021/2/2021-Seagate-Analyst-Day.pdf,（アクセス日 2021 年 12 月 24 日）.
[66] 福田 昭，“福田昭のセミコン業界最前線”，東芝 -WD 連合の 3D NAND，製品の量産に Samsung の技術を採用”，（2019.12），https://pc.watch.impress.co.jp/img/pcw/docs/1223/976/html/photo008.jpg.html,（アクセス日 2021 年 12 月 21 日）.

低炭素社会の実現に向けた
技術および経済・社会の定量的シナリオに基づく
イノベーション政策立案のための提案書

情報化社会の進展がエネルギー消費に与える影響（Vol.4）

－データセンター消費電力低減のための技術の可能性検討－

令和 4 年 2 月

Impact of Progress of Information Society on Energy Consumption (Vol. 4):

Feasibility Study of Technologies for Decreasing Energy Consumption of Data Centers

Proposal Paper for Policy Making and Governmental Action
toward Low Carbon Societies,
Center for Low Carbon Society Strategy,
Japan Science and Technology Agency,
2022.2

国立研究開発法人科学技術振興機構　低炭素社会戦略センター

本提案書に関するお問い合わせ先

●提案内容について・・・低炭素社会戦略センター　上席研究員　三枝 邦夫（SAEGUSA Kunio）
●低炭素社会戦略センターの取り組みについて・・・低炭素社会戦略センター　企画運営室

〒102-8666　東京都千代田区四番町5-3　サイエンスプラザ 8 階
TEL :03-6272-9270　FAX :03-6272-9273
https://www.jst.go.jp/lcs/

② 2022.6