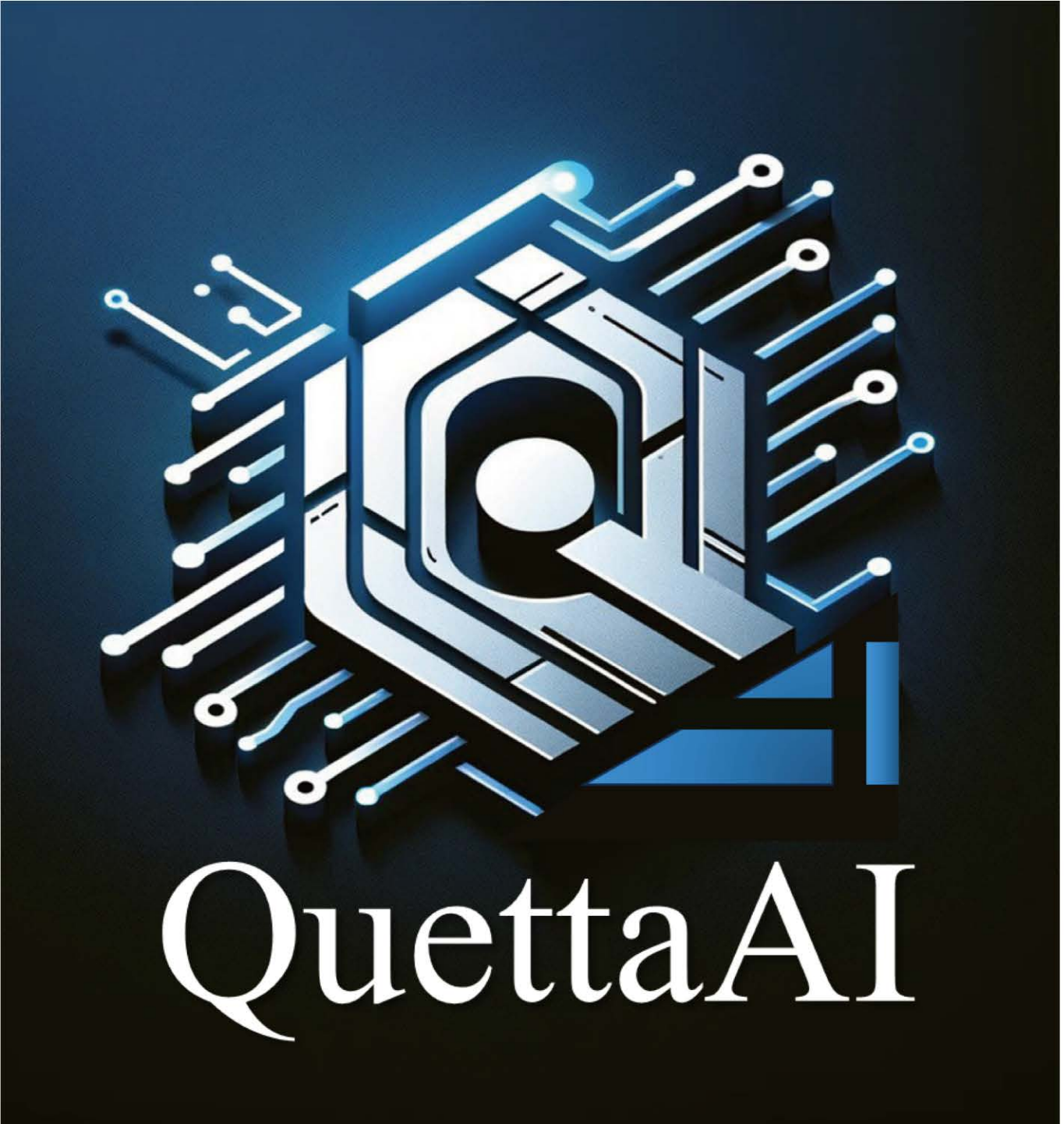


「国産GPGPU」DCファンド

日本の将来を救う唯一の現実解である「国産GPGPU」による水浸冷却AIデータセンター事業専用ファンドの御案内

2024年10月～11月

ZEXAVERSE

Executive Summary（概要）

・日本を国家破綻危機から救い、生成AI敗戦を避ける現実解が必要です

・そのための唯一の具体策として、開発が完了している「国産GPGPU」を活用し、世界最高の冷却手法である「水浸冷却」技術により構築する自社製AIデータセンターによるクラウドサービスを行うためのファンドを組成し、来春からの稼働を行います

・通常は500億円をも超える資金を要する、5nmの最先端半導体の大規模設計開発を行うことが可能な企業体は、日本には2社しか存在しません。
その1社と、4年間の先端半導体事業での協業実績を有するZEXAVERSEは、この度、最新の「国産GPGPU」の量産製造権・販売権を有する新設法人に出資を行って、AIデータセンター事業を行う権利を獲得致しました

・来週からのAIデータセンター稼働後、日本での需要を全く満たせないクラウドサービスを、潤沢に確保可能で十二分な性能を持つ国産GPGPUのクラウドサービスからは、年率50％ものファンド配当を見込めます

・金銭的なリターンに加えて、本ファンドが立ち上がり、広がっていくことで、生成AI革命の最中において、計算機資源であるGPGPUが全く確保できない状況のために、AI敗戦国に甘んじなくてはならない危機的な状況から、日本を救うことが可能となります

ファンド組成の背景と趣旨

現在、日本は半導体で台湾、米国、韓国に大きな後れを取り、自動運転やロボット技術でも、残念ながら最先端国からは小さくない後れをとってしまっています。
一方で、少子高齢化と人口減少が急速に進んで、大多数の地方都市町村が消滅可能性都市の認定を受けるなど、「課題先進国」とも揶揄される状況です。

そこに加えて、先の産業革命を大きく上回る規模と速度で、「生成AI」革命が始まり、「デジタル小作人」と称される日本の「デジタル赤字」が、巨額となることが確実な情勢です。

そして、2030年には20-30兆円にも上るデジタル赤字を超える規模となるのが、新しく生じてくる「ロボット赤字」です。
多くの日本人がスマホを複数所有する様に、近い将来、我々はロボットを複数台、多い人では数十台を所有して仕事を行い、生活をすることになります。しかし残念ながら、日本に10億台規模で稼働するロボットの殆どは米国製となる見通しです。

やがて生成AI革命は、「AGI（汎用人工知能）」を産み出し、世界は全く新しい時代に突入して新しい世界覇権が生じます。
ネットサービスをGAFAMに依存する状況と同じく、全ての生活と社会基盤を支えるであろうAGIサービスも、このままでは国外依存となり、「AGI赤字」発生と国家破綻が不可避です。

この既定路線を、何とか打破しなくては、日本の将来が危うい状況であることが明らかです。幸い、現時点ではまだ、この日本にそのための現実的、かつ具体的な回避策が存在します。

急速に進行する生成AI革命には、先の産業革命時の蒸気機関に相当する、非常に強力な計算機資源としての「GPGPU（汎用グラフィックス演算ユニット）」が必要とされます。

「GPGPU」の有無が、生成AI革命での勝者と敗者を峻別することになりますが、残念ながらその100％は米国及び中国製です。現在、日本が輸入できているGPGPUの数量は、米国が製造する約300万台に対して1万台程度で、僅か0.3％に留まります。

米国製のGPGPUの90％超は、時価総額が500兆円に達して世界1位にもなったNVIDIA社で、残る10％はAMD社で製造されます。しかし、そのNVIDIAに先んじて「GPGPU」の原型を開発したのも、構成部品として必須である「HBMメモリ」の原型を開発したのも何れも日本企業であり、まだ研究開発は進んでいます。

現在、NVIDIA社の主力製品に対抗可能であり、安価かつ潤沢に国内に供給が可能な国産GPGPUの設計開発が完了しており、これを用いたAIデータセンターの稼働が心待ちにされています。
その事業化に特化したファンドを、ZEXAVERSEが組成致します。

急増するデジタル赤字

我が国の「失われた30年間」は、むしろこれから非常に深刻な事態を、令和時代にもたらすことになります。

日本の技術基盤の喪失は、特に情報通信分野において危険領域に入っており、特に米GAFAM企業に日本企業と個人が支払う金額、「デジタル赤字」は指数関数的に急増していることが経産省資料でも指摘されています。

しかし現実は、その予測をはるかに上回る速度で進行してしまっているばかりか、ここには今年から本格的な影響を及ぼす、「生成AI革命」の甚大な影響はまだ含まれていません。

2030年には、2022年に経産省が予測した8兆円規模のデジタル赤字額は、このままでは20-30兆円にも達する可能性が高いですが、これは現在、日本が輸入する全ての鉱物動力燃料（原油、天然ガス、石炭他）の合計金額の約25兆円に匹敵します。

従って、この状況を早急に打開しない限りは、日本の国家財政と国富流出は、数年で取り返しがつかない事態となり国家破綻が避けられない道を進みつつあります。

この危機的状況を早急、かつ確実に回避するために、今直ぐに、具体的な対応策を講じなくてはなりません。

「失われた30年」を如実に表すのが、技術基盤の喪失です

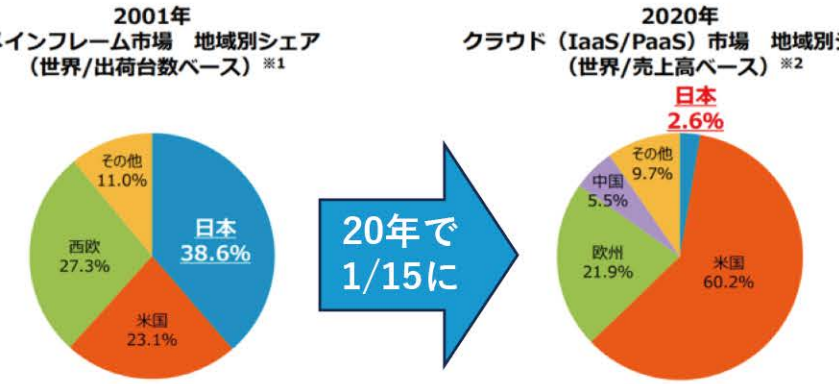
グローバルシェアの著しい低下には目を覆うばかりであり、IT分野シェアは直近の20年間で1/15に激減しています

経産省の2022年の当初予測は2023年に2兆円、2030年では8兆円のデジタル赤字でした

ところが2023年は既に5兆円を超過し、生成AI革命の本格化により2030年には20 - 30兆円規模のデジタル赤字が予想される状況です

グローバルシェアの低下　～技術基盤の喪失～

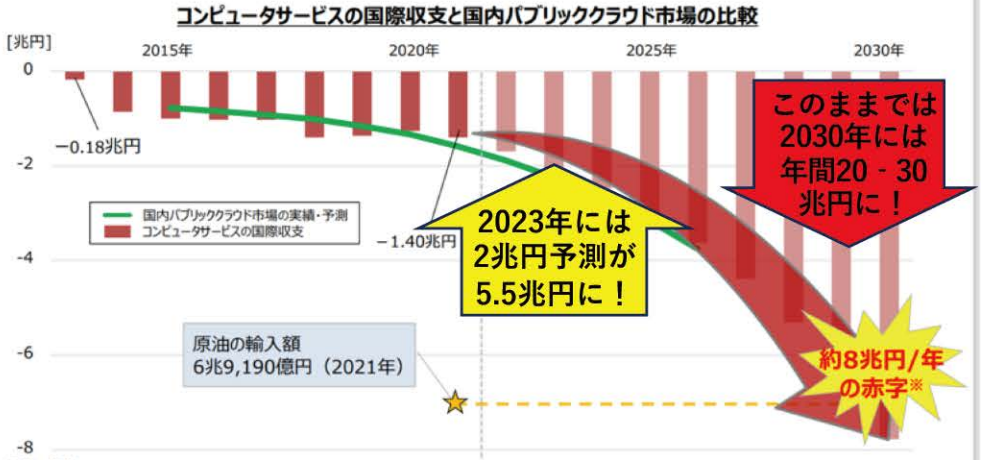
- かつて社会を支えたメインフレームの世界市場において、日本が高いシェアを誇っていたものの、現在はシェアを落とし、急速に拡大するクラウドサービス市場においては、**日本のシェアは極めて小さい状況**。
- このままでは、社会を支える**情報処理に関する技術的知見を失ってしまうおそれ**。



（出典）
※1：「@IT」IT Market Trend 第14回 問われる情報システム産業の構造（前編）―日本はメインフレーム大国のままでいいのか？―
※2：Cloud Services Global Market Report 2021: COVID-19 Impact And Recovery To 2030（The Business Research Company, August 2021）

海外への支出の拡大　～技術ギャップに伴う国富の流出～

- 足下では、コンピュータサービス領域における貿易赤字が大きく拡大。
  現在のペースでいくと、**2030年には貿易赤字が約8兆円に拡大するおそれ**。



※積算の根拠
コンピューターサービスの国際収支の赤字額について、実績ベースで、国内パブリッククラウド市場の規模に近似していると見なし、今後、国内パブリッククラウド市場の民間予測に基づく成長率と同程度に拡大すると仮定すると、2030年には年間約8兆円の赤字額になると推計。

（出所）貿易統計（財務省）、国際収支関連統計（日本銀行）及び 国内パブリッククラウドサービス市場 売上額予測（IDC Japan）を基に経済産業省作成

次は「ロボット赤字」

SONYのAIBO（アイボ）や、2000年にお披露目された2足歩行のHondaのASIMO（アシモ）が世界を驚かせてから四半世紀が経過しました。その「人型ロボット」市場には現在、もはや日本企業は見る影もない状況です。

人型ロボットは最新の生成AIを用いて開発、最適化されることで、そして生成AIによる「対話」機能と「世界モデル」認識機能を搭載するため、我々の生活の場と労働環境に、予想以上に急速に溶け込むこととなります。

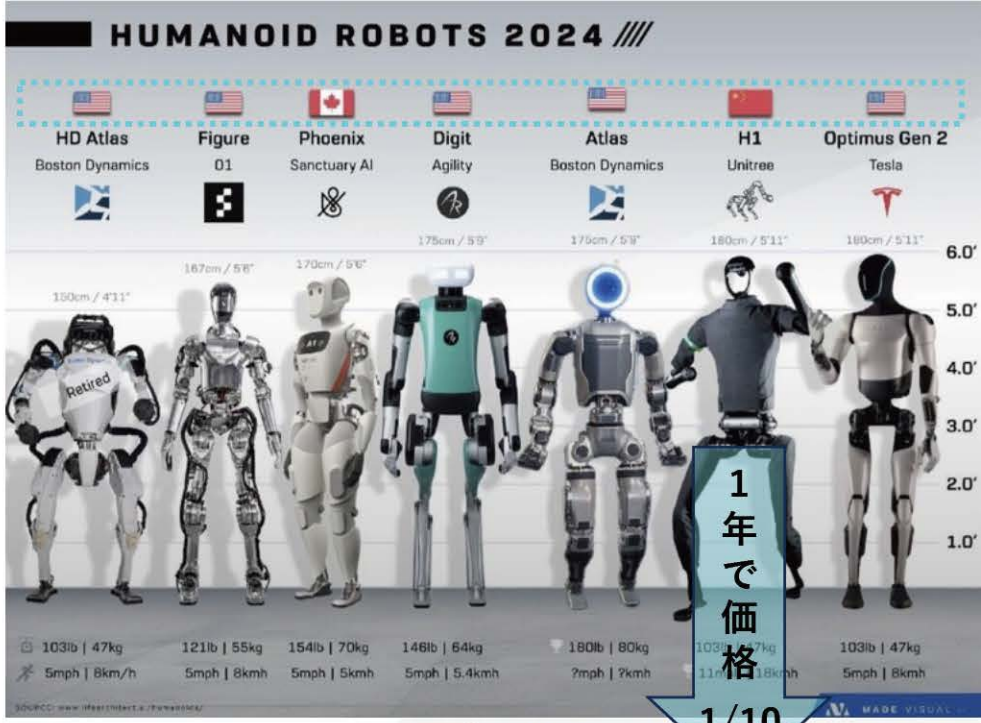
直ぐに職人技や匠の技を上回り、人間には行えない危険な作業や繊細・精密な作業も、難なく担い始めます。
既に人材難の介護分野を含めて、少子高齢化で労働力不足が深刻化する日本や中国では、「人型ロボット」は社会維持に不可欠な、最重要「インフラ」となります。

テスラ社は、人型ロボットの「Optimus」を最先端AIを搭載して100億台製造する計画です。OpenAI社も3月に1,000億円を調達したFigure社を含めて、3社に投資をしつつ自社でのロボット開発も再開したところです。

そこでも、現在のGAFAMと同様の状況が生ずることが確定的です。一人で複数台を保有するロボット需要の殆どを、生成AIとロボット技術を持つ海外に依存して、デジタル赤字以上の巨額のロボット赤字が生じます。

2024年人型ロボットの一覧では、7社を掲載していますが、日本企業の姿は皆無

中国Unitree社は1年で価格を10分の1にした大幅機能強化版を、僅か250万円で年内に販売開始予定



- 中国は2025年までに「上級レベル」の人型ロボットを大量生産するという大胆な計画を明らかにした。
- 中国工業情報化部は先週、この計画のロードマップを公表した。
- 詳細の多くはまだ明らかにされていないものの、中国は自国が開発するロボットの「破壊的な」力を強調した。

中国は人型ロボットを大量生産する野心的な計画を明らかにした。
ロボットはスマートフォン同様、「破壊的な」ものになると中国は考えている。

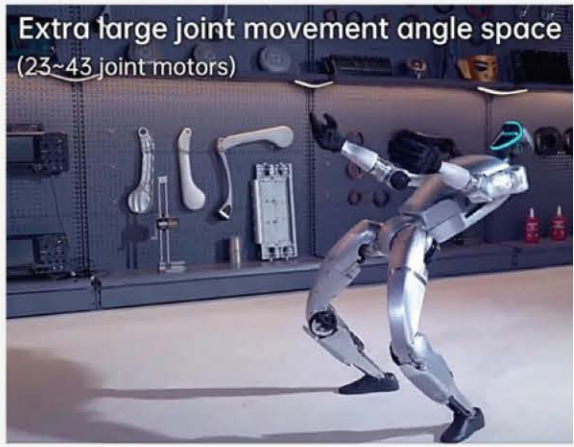
先週発表された野心的な計画書の中で、中国工業情報化部はロボットが「世界を塗り替える」としている。

中国が有する正しい理解と覚悟

工業情報化部では2025年までに中国が開発するロボットが「上級レベル」に達し、大量生産されるようになると考えている。ロードマップに掲げられた開発目標の中で、同部は「（人型ロボットは）コンピューター、スマートフォン、新エネルギー車に次ぐ破壊的製品になると期待されている」としている。

ビジネス 2024/6/2 [SUN]

ヒューマノイド型ロボット「Unitree G1」を発表　お手頃？価格の約250万円



2024年5月、中国のロボット開発メーカー「Unitree」が、新たなヒューマノイド型ロボット「Unitree G1」の発売を発表した。このロボットは、価格が1万6000ドル（約250万円）からで、他の競合製品と比較してリーズナブルな設定だという。

生成AI用「GPGPU」の起源は日本

飛ぶ鳥を落とす勢いで急成長を遂げて、時価総額が500兆円に達し、一時は世界1位の企業となったNVIDIAですが、その勢いは留まる気配がなく、遠からず時価総額が1,000兆円を超えて、東証の時価総額合計を超えるとも目されています。

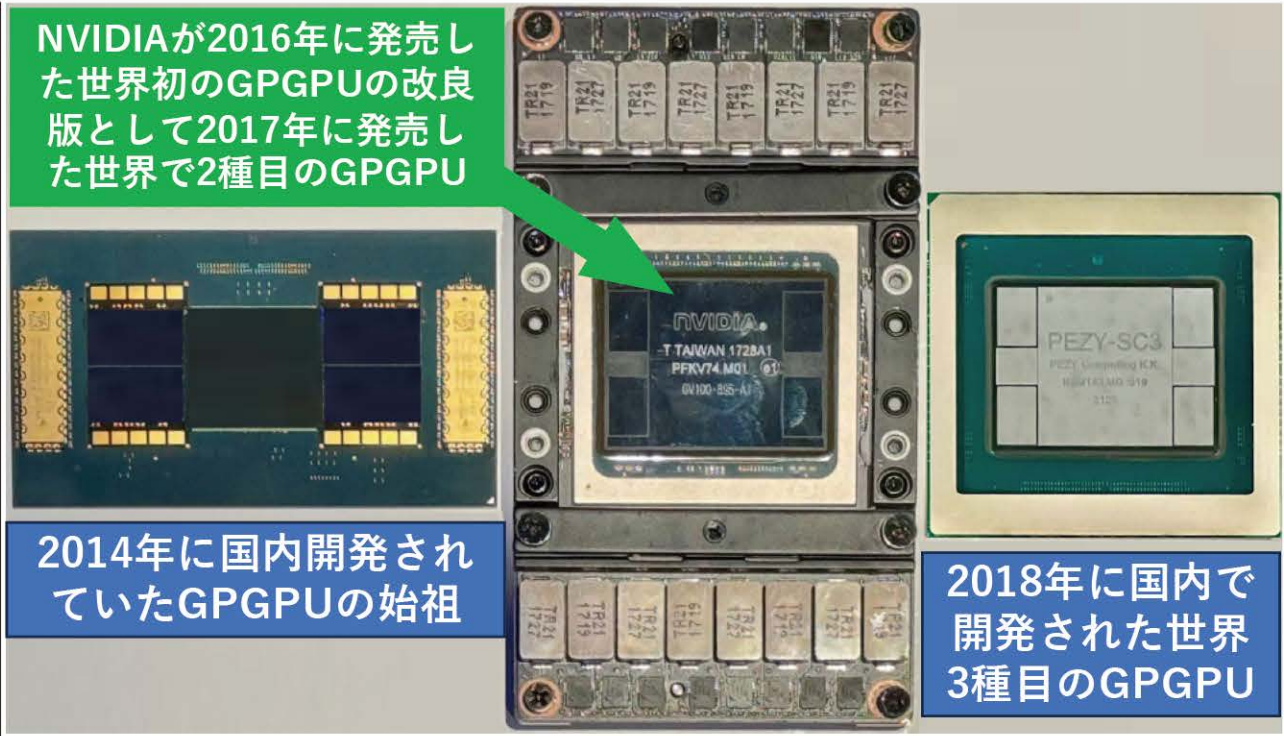
その急成長の源泉は専ら、生成AIデータセンター向けGPGPU製品である第3世代の「A100」と、第4世代の「H100」ですが、これらは世界中で熾烈な争奪戦が繰り広げられ続けています。

製造原価が3,000-3,500ドルとされる両製品ですが、販売価格は30,000-50,000ドルであり、90-93%の前代未聞の高粗利率です。しかし製造量が限られ、日本での入手は極めて困難な状況です。

現在は「GPGPU」と称される半導体は、当初は消費電力効率世界一をNVIDIA社と争っていた日本企業がスーパーコンピュータ向けにその基本構成を2014年に実現したのが、その起源です。

NVIDIA社はその後、2016年になってようやく初代のGPGPU製品である「P100」を第4四半期に発売し、翌年に改良版である「V100」を発売しましたが、3代目製品で現在も生成AI処理用に大量に使用されている「A100」の発売は2020年でありました。

その間、その日本企業は国産GPGPUを2018年に開発完了しており、世界で3種目のGPGPUは同製品ということになります。



左）2014年開発の国産GPGPU：8積層超広帯域DRAMを2個ずつを中央の大きなロジックチップの左右に配したパッケージの構成は、「GPGPU」の「始祖」とされるべき開発成果です

中央）NVIDIA社「V100」：2016年10月発売開始の初代GPGPUの「P100」も同一構成で、中央の大きなロジックの左右に8積層広帯域DRAMである「HBM」を2個ずつ搭載している

右）2018年に開発完了したHBMを4個搭載した国産GPGPUは、NVIDIA社以外では初めて完全なGPGPU構成をとった製品

「HBM」メモリの起源も、実は日本

NVIDIA社がGPGPU製品を販売するためには、自社で独自に設計開発を行い台湾TSMC社に製造を委託するロジック半導体の他にも、高速に生成AI処理を行うために必須であり、世界三大メモリベンダーしか製造できない、超広帯域・大容量の3次元積層DRAMである「HBM」メモリ半導体が必要です。

このHBMメモリは、生成AI用の演算処理に巨大な需要が生じたことで、世界三大メモリベンダーの韓国のSamsungとSK Hynix、米国のMicronは、今年に入り軒並みの好業績となって大きな恩恵を得ています。

NVIDIA社は、この三大メモリベンダー製造のHBMメモリを、少なくとも3年間分は前金で買い切っているため、競合他社がどんなに優れた製品を開発し、試作できたとしても、対抗するGPGPU製品を量産することは事実上不可能です。この強固な参入障壁が築かれていることで、NVIDIA社の市場独占が長期継続されています。

HBMメモリもその元を辿れば、2009年にエルピーダメモリが開発していた世界初の画期的な8積層DRAMが起源であることには、異論を挟む余地がありません。
2012年にエルピーダメモリが破綻した際、今日のこの状況を正確に予測した日本企業は、同製品を担当していた20名の設計開発チームを、同社のCTOを含めて丸ごと引き取り、グループ内のメモリ設計法人として今日まで開発を継続してきました。

その結果、GAFAMといえども確保することが全く叶わない希少なHBMメモリを、安価かつ潤沢に入手してGPGPU製造に利用することが出来る状況が整っています。

（右下写真：元エルピーダメモリ開発チームが2022年12月に製造成功した独自HBMメモリ）

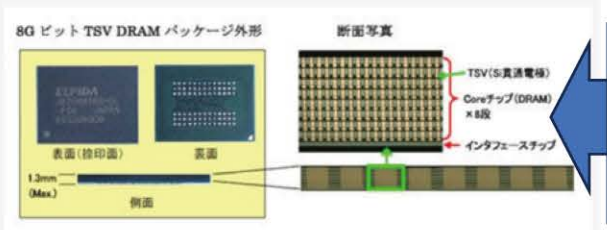
エルピーダ、1GビットDDR3 SDRAMを8枚積層した世界初のCu-TSV DRAM

エルピーダメモリによる世界初の8積層DRAM（HBMの起源）の開発成功を報じる、日経クロステック誌の記事

掲載日 2009/08/28 23:09

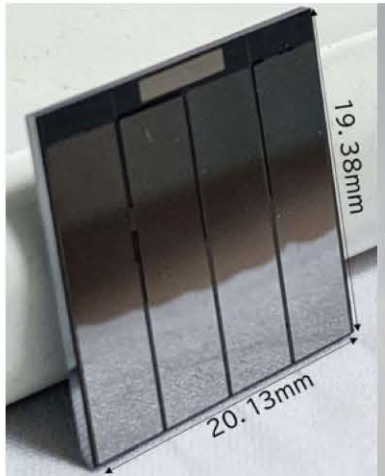
エルピーダメモリは、Cuを用いたSi貫通電極(TSV:Through Silicon Via)による積層8GビットDRAMの開発に成功したと発表した。

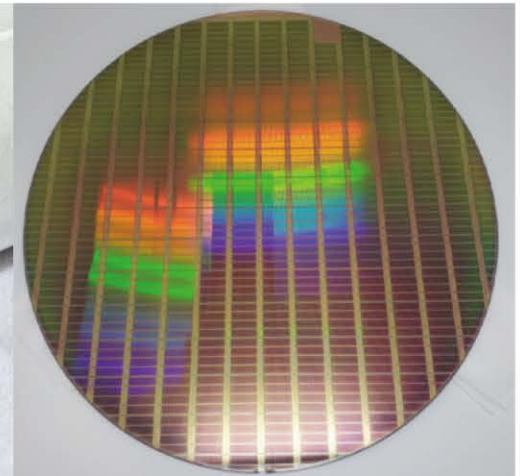
今回開発されたDRAMは、1GビットのDDR3 SDRAM(1600Mbps)を8枚積層したもので、インタフェース層と合わせると合計9層構成で8Gビットを実現している。従来のMCP(Multi-Chip Package)やPoP(Package on Package)に比べ、待機時の消費電力は1/4。パッケージ厚は1.3mm(Max)で、コア間の接合端子数は1030ピン(インタフェースを含み意パッケージ当たり総計8357ピンをバンプ接続)となっている。



今日の最新HBMメモリと殆ど同一の内部構造

今回開発された8GビットTSV DRAMパッケージ





NVIDIAに対抗可能な国産GPGPU (1)

2014年から2022年の間、日本企業はNVIDIA社とスーパーコンピュータ分野で世界一位の座を賭けて、鎬を削る直接の対決を行ってきました。

特に2015年から2018年には、世界1-3位の独占を2回も達成するなどして、NVIDIA社と互角以上の開発成果を上げた世界的に有数の実績を有します。

現在、TSMC社の5nm世代の製品が設計開発を完了しており、量産製造を行える状態にあります。

この今回開発品が有する性能は、NVIDIA社の主力GPGPU製品である第3世代品の「A100」よりもむしろ第4世代品の「H100」に近い性能を有していながら、これを第3世代の「A100」の半額で販売可能な程の原価率の低さを実現します。

そして何よりも、世界的に需給の逼迫が長期間に渡り継続している、NVIDIA社のGPGPU製品とは異なり、これも日本企業が独自に発明し、開発に成功した画期的な「水浸冷却」技術と併用することで、加えて独自開発したHBMメモリも併用することで、量産性の問題を完全に払拭することに成功しています。国産GPGPUの安価で潤沢な供給が、日本の将来を救済することになります。

GPGPU比較表	NVIDIA-A100	今回開発品 (Available in Q1/2025)	NVIDIA-H100
製造半導体プロセス	7nm	5nm	4nm(5nm)
演算コア数	6,912/3,456	2,048	16,896/8,448
動作周波数	1,410MHz	1,800MHz	1,830MHz(@BF16) 1,980MHz(@FP64)
半導体チップ面積	827mm^2/54B	556mm^2/49B	814mm^2/80B
FP64 for HPC	9T/18T	29T	30T/60T
FP32	19T	58T	60T
TF32(Tensor)	156T/312T	-	495T/990T
BF16 for G-AI	312T	768T	990T
FP8,I8(Tensor)	-	-	1,979/3,958
PCIe	Gen4x16	Gen5x16	Gen5x16
チップ間通信	NVLink	-	NVLink
HBMメモリ種別	HBM2e	HBM3/3e	HBM3/3e
HBMメモリ帯域	1,555GB/s	3,200GB/s	3,352GB/s
HBMメモリ容量	40	64	80
消費電力	400W	500W	700W

スパコンの性能値

生成AI処理の性能指標

NVIDIAに対抗可能な国産GPGPU (2)

日本で唯一、世界最長の履歴を有する半導体設計

当該日本企業は、当初は医療分野での開発を行い、CT装置等の超高速画像処理システム用の半導体開発に着手してから、30年間の期間に32種類もの大規模先端半導体の開発を行って来ました。

同一の開発メンバーで半導体の設計開発に取り組んだチームは、米国で最古で56年の歴史を誇るインテル社内にも、25年を超えるものは存在しないことから、恐らくは世界最長の開発履歴を誇るチームであります。そこに集積された膨大な経験値、ノウハウ、実績は、途轍もない価値を有していると言えます。

失われた30年間により、日本では先端半導体の設計開発チームが完全に散逸し、消失してしまいました。当該日本企業が率いてきているチーム以外には、日本で大規模半導体を先端プロセスでまともに開発できるところは、既に皆無となってしまいました。

今回開発品は、当該日本企業が過去4年間を費やして設計開発を行ってきた成果であり、これをゼロから大手企業が開発しようとしたら、少なくとも100人以上の開発体制と、5年間の期間と、500億円を超える費用を要することが明らかです。

その開発成果を、少額の資金で事業化することが可能な特殊な状況をもって事業法人が設立され、ZEXAVERSEは株主として名前を連ねることで、本AIデータセンター事業を行うことができます。

Before PEZY Computing K.K.

Processor	Year	Process	Die Size (mm)	Clock	Gates	Architecture	Core number	FLOPS	Power	Memory
Version 1.0	1997	600nm	8.0*8.0	50MHz	1.2M	VLIW+SIMD	1 Core/8 ALU	Fixed Point	6W	SDR
Version 1.5	1999	350nm	7.3*7.3	80MHz	1.5M	VLIW+SIMD	1 Core/8 ALU	Fixed Point	3W	SDR
3DVR Version 1.0	1999	350nm	13.65*13.65	133MHz	0.8M	Hardwired Pipeline	2 Pipeline	-	32W	DDR
Version 2.0	2001	250nm	8.1*8.1	80MHz	1.8M	VLIW+SIMD	1 Core/8 ALU	160M	2W	SDR
3DVR Version 2.0	2001	160nm	15.6*15.6	250MHz	3.2M	Hardwired Pipeline	4 Pipeline	-	20W	DDR
Version 2.0 shrink	2003	180nm	6.5*6.5	167MHz	1.8M	VLIW+SIMD	1 Core/8 ALU	333M	1W	SDR
Version 2.5	2003	180nm	6.5*6.5	167MHz	2M	VLIW+SIMD	1 Core/8 ALU	333M	2W	DDR
DBF Version 1.0	2003	180nm	11.5*9.6	40MHz	2.5M	Hardwired Pipeline	-	-	10W	-
Version 3.0	2005	130nm	16.5*12.0	333MHz	34M	RISC+VLIW+SIMD	8 Core/40 ALU	13.3G	19W	DDR
Version 3.0 B	2005	130nm	9.5*12.0	250MHz	20M	VLIW+SIMD	1 Core/8 ALU	8G	6W	DDR
3DVR Version 3.0	2008	130nm	10.5*10.5	333MHz	5.5M	Hardwired Pipeline	2 Pipeline	-	10W	DDR2

At PEZY Computing K.K.

Processor	Year	Process	Die Size (mm)	Clock	Gates	Architecture	Core number	FLOPS Double/Single	Power	Memory
PEZY-1	2012	40nm	21.0*16.8	533MHz	220M	RISC+SMT (MIMD)	512 Core	166/333G	35W	DDR3/Wide IO
PEZY-SC	2014	28nm	21.1 *19.5	733MHz	580M	RISC+SMT (MIMD)	1,024 Core	1.5/3.0G	70W	DDR4/Custom Ultra-Wide IO
PEZY-SC2	2016	16nm	25.8*24.1	1GHz	2.4G+	RISC+SMT (MIMD)	2,048 Core	8.2/16.4T	100W	DDR4/Custom Ultra-Wide IO
PEZY-SC3	2020	7nm	30.6*25.7	1.2GHz	5G+	RISC+SMT (MIMD)	4,096 Core	20/40T	500W	HBM2
PEZY-SC3s	2021	7nm	10.9*10.0	1.2GHz	960M	RISC+SMT (MIMD)	512 Core	2.5/8.0T	80W	HBM2
PEZY-SC4s (under development)	2022	5nm	TBD	1.5GHz	10G+	RISC+SMT (MIMD)	2,048 Core / Tensor Core	25 /50T	400W	HBM3

At ZettaHash/PEZY Computing K.K.

BTC Mining Chip	Year	Process	Die Size (mm)	Clock (MHz)	Gates	Architecture	Pipeline number	Hash Power GH/s	Power (Watt)	Power Efficiency (W/GH)
ZH-V0	2017	12nm	12.1*10.4	400	853M	Hardwired Pipeline	1024	410	80	0.195
ZH-V1	2018	7nm	7.2*7.3	500	592M	Hardwired Pipeline	1,024	512	80	0.156
ZH-V2	2018	7nm	4.7*5.1	500	294M	Hardwired Pipeline	524	289	30	0.104
ZH-V3 (Shuttle TEG)	2020	7nm	1.4*1.3	400	20M	Hardwired Pipeline	22	8.8	0.5	0.055
ZH-KD1	2022	12nm	-	800	-	Hardwired Pipeline	384	307.2	40	0.130
ZH-KD2	2022	6nm	-	900	-	Hardwired Pipeline	1,024	921.6	35	0.038
ZH-KD3 (in Plan)	2022	5nm	-	1,200	-	Hardwired Pipeline	768	921.6	25	0.027

同一の開発チームによる、30年間の先端大規模半導体の開発履歴：

32種類の開発品のうち、30種類については無修正での量産動作を得ており、残る2例についても、3か月程の小修整のみで完全動作を得ていて、世界最高の成功率と思われる。

NVIDIAに対抗可能な国産GPGPU (3)

NVIDIAを凌駕する、画期的な「水浸冷却」技術

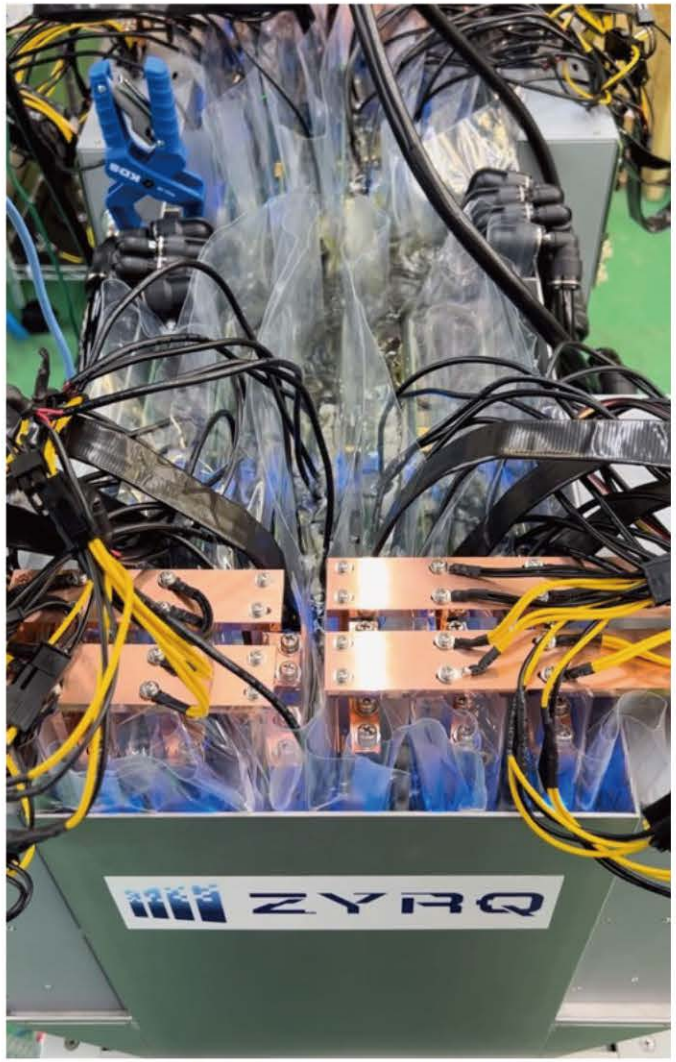
当該日本企業は昨年2月、革新的な冷却技術を新発明して莫大な熱量を25℃に安定冷却する手法を世界で初めて確立しました。

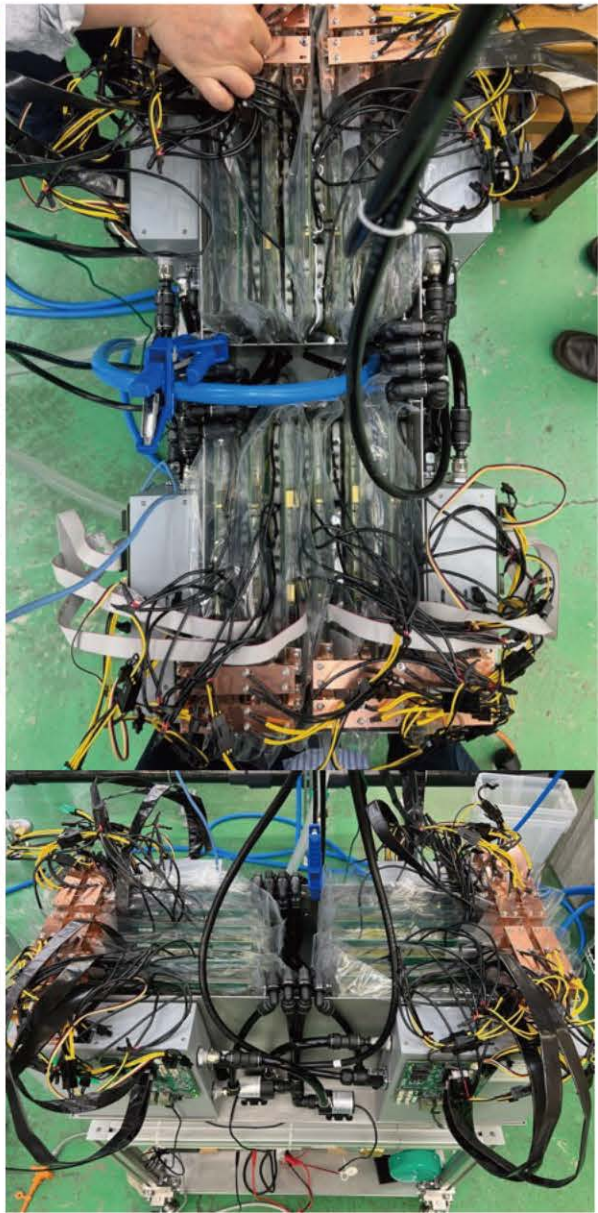
2014年に発明していた、それまでの液浸冷却技術も、合計14台もの独自液浸スーパーコンピュータで利用されて、消費電力効率の世界一を何度も獲得するなど、世界的な実績を多数残しました。
しかし、それまでに使用してきたフッ化炭素系冷媒は、永遠の化学物質「PFAS」として使用が禁止されることとなって、急遽、全く新しい冷却手法を考案する必要に迫られました。

当該日本企業は、再びゼロから液浸冷却の開発を余儀なくされた結果、むしろそれまでの4倍もの非常に強力な冷却性能を実現した上で、設備費用も、運営費用も大きく削減が可能となる全く新しい冷却手法を、一般的な「水」のみを使用して実現する画期的な方法を確立することができました。

NVIDIA社は、次世代GPGPU製品が1,000Wを超える消費電力と発熱になることから、空冷を放棄して液冷への移行を迫られ、米Vertiv社との提携により液冷システムを採用するに至りました。

しかしながら、一般の液冷方式の限界は明確であります。今回の発明による水浸冷却は、その10倍超の冷却能力を提供します。





非常に小さい体積で、大熱量を極めて低温に冷却し、消費電力を半減させ、半導体動作速度の大幅な高速化が可能

世界初の画期的な「水浸」冷却技術

- 25℃に冷却（80-105℃）
- 200kW/m^{3}の熱密度（30-40kW/m^3）
- 超小型で安価

AIデータセンターの現実解
（随一の強力な冷却能力提供）

超小型AIデータセンター
（安価、短納期、狭小土地）

AIデータセンター省電力化
（50-60%もの削減）

GPGPU不足の抜本的解消
（国産品の安価で潤沢な提供）

半導体設計の歴史的革新
（小型化、高速化、消電力化、開発短期化）

猛暑でデータセンター安定稼働が絶望的状況に

WIRED BUSINESS CULTURE GEAR MOBILITY SCIENCE WELL-BEING OPINION SZ MEMBERSHIP

CHRIS STOKEL-WALKER BUSINESS 2022.08.17

記録的な猛暑でデータセンターまでダウン。温暖化の影響を回避する現実的な対策とは？

欧州が記録的な猛暑に襲われるなか、グーグルやオラクルのデータセンターが冷却装置の故障により相次いで停止する"事件"が起きた。温暖化による気温の上昇が続くなか、生活のインフラでもあるデータに影響が出ないようにするには、どのような対策が求められているのか。

英国が記録的な猛暑に襲われた2022年7月19日、ロンドンにある「Google Cloud」のデータセンターで冷却装置が故障し、半日あまりにわたってサービスの提供が停止した。影響はこのデータセンターが管轄する米国や太平洋地域のユーザーにも及び、グーグルの主要サービスへのアクセスが数時間にわたり制限される事態が起きている。

このとき、同じロンドンにあるオラクルのクラウドサービス用データセンターでも暑さによる障害が発生し、米国のユーザーがサービスを利用できなくなった。オラクルは「季節外れの気温」が障害の原因だと説明している。

英国の気象庁は今回の記録的な猛暑について、今後の傾向を示す現象であるとの見解を出している。つまり、データセンターは"ニューノーマル"に備える必要があるということなのだ。

上昇を続ける気温に備えはできているのか？

世界気象機関（WMO）によると、2022年から26年のいずれかの年が観測史上で最も暑い年になる確率は93%だという。それも、その年に限った話ではない。

だが、これは欧州に限った話ではない。デジタルサービスの規格や認証サービスを提供する民間機関「Uptime Institute」の調査によると、米国内のデータセンターの45%が異常気象により運用に支障をきたしかねない事態を経験したと答えているという。

将来の気温上昇を考慮した設計が重要に

そこで問題が持ち上がる。このデータはあくまで過去のデータであり、英国で40°Cを記録することなどなかった時代の話なのだ。「われわれは気候が変わりつつある過渡期にいます」と、ハリスは言う。

「少し前まで、冷却装置を設計する際は最高外気温を32°Cに設定していました」と、英国のデータセンター専門コンサルティング会社Keysourceのジョン・ヒーリーは説明する。「いまは設計時の想定より8°Cも高くなっているのです」

設計条件はますます高度になっているが、データセンター運営会社もそのクライアントも、利益主導で動く企業だ。コンサルティング会社Turner & Townsendのデータによると、データセンターの建設コストは近年どの市場でも基本的に上昇しており、建設会社はコスト削減を言い渡されている。

データセンターの設計基準が、温暖化の影響によって、外気温「32度」から「40度」に、8度も引き上げなくてはならない状況に

（外気温の設計基準温度が1度上がる毎に冷却用設備コストが約1割増加する試算があり、8度の上昇では設備コストが**1.8倍**に）

昨夏の酷暑を経て、今後は外気温「45度」への対応も必要になる可能性も

（13度の上昇では、設備コストが**2.3倍**にも）

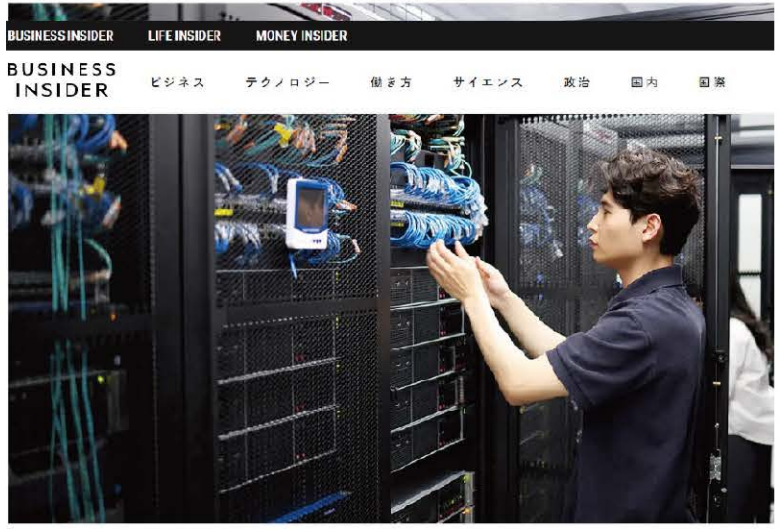
一方、大量のデータセンター新設のためのコスト削減要求は非常に厳しく、当然に抜本的な解決策が求められることに

データセンターを、超小型化した上で効率的かつ強力に冷却するしか解決法がないことが明白に

生成AIの突然の勃興で莫大な冷却需要が出現

データセンター「消費電力3倍増」問題、生成AIブームで一気に深刻化。マイクロソフト元幹部が懸念訴える

Ellen Thomas［原文］（翻訳・編集・情報補足：川村力）
Jul. 18, 2023, 07:20 AM | テクノロジー



人工知能（AI）ブームを背景に、膨大な処理能力を求められるデータセンターのコストや立地条件、運営手法にも変化の兆しが見えてきている。画像は韓国・サムスン電子の水原（スウォン）本社キャンパス内にあるデータセンター。
REUTERS/Kim Hong Ji

アメリカ国内のデータセンターが集中する「データセンター・アレイ（Data Center Alley）」は2022年、電力不足寸前にまで陥った。

米バージニア州アッシュバーンを中心とするこのエリアに電力を供給するエネルギー大手ドミニオン・エナジー（Dominion Energy）は、急増する需要に対応できない恐れがあるとの警告を発した。

電力不足は以前から問題化していたが、対話型AI「ChatGPT」のリリースに端を発するAIブームの到来により、次世代データセンターが消費するエネルギー量は2倍あるいは3倍に膨れ上がりそうだ。

業界の経営幹部や研究者、アナリストらは、ドミニオン・エナジーが2022年に経験したような非常事態がおそらくこれから日常茶飯事になると予測する。

生成AIで変わるデータセンター、冷却は外気から液体へ

根津禎 日経クロステック 2023.08.03

データセンターで液体による冷却方式が今後本格化する。写真はIIJの「白井データセンターキャンパス」にある2期棟のサーバー室（出所：IIJ）

多大な計算資源が求められる生成AI（人工知能）の登場で、データセンターにおける消費電力の増加に拍車がかかっている。それに伴い、冷却方法を見直す動きが盛んになってきた。これまでは空冷方式が中心だったが、今後は液体による冷却（液冷）方式が本格化する見込みだ。

千葉県白井市にあるインターネットイニシアティブ（IIJ）の「白井データセンターキャンパス」。2023年7月に2期棟の運用が始まったばかりだが、需要が旺盛なことから早くも3期棟の検討に着手している。

3期棟で目指すのは、生成AIや大規模言語モデル（LLM）といった多大な計算資源を求められる最新のAIに対応したデータセンターである。IIJによれば、AIの学習に必要な計算能力は、2012年以降、3～4カ月ごとに倍増してきたという。生成AIの登場で、一層のペースアップが予測される。

生成AIが「データセンター」に及ぼす影響

- 生成AIで用いられる大規模言語モデルを学習させるために多大な計算資源が求められる
- 計算資源を増やすために、高性能なプロセッサーを採用するので消費電力が増大する
- 消費電力が大きいプロセッサーを冷やして正常に動作させるために、冷却システムの性能向上が求められる

今後予想されるトレンド

- 演算処理性能が高いプロセッサーに対する需要が一層高まる
- 消費電力削減のために新しい半導体や通信技術、電力制御技術の導入が進む
- 外気冷却に加えて、直接液体冷却や液浸冷却の導入が進む

生成AI需要の出現で、旧来の300W程のCPUを主体とするデータセンターでは、電力供給と冷却性能が全く追い付かないことが日に日に明らかに

CPU:300W + GPU:700W***1台** = 1,000W（**3.3**倍）
CPU:300W + GPU:700W***2台** = 1,700W（**5.7**倍）
CPU:300W + GPU:700W***4台** = 3,100W（**10.3**倍）

上記の計算からは消費電力3倍増でも最低限で、実際の一般的な構成では5倍増、10倍増になり、低消費電力化（サーバー、冷却共に）と、強力な冷却性能がセットで必要に

生成AI市場急伸で、現在の20倍のDCが必要に

生成AI市場は10年以内に1.3兆ドルまで成長…ブルームバーグが試算

Sawdah Bhaimiya［原文］（翻訳：Makiko Sato、編集：井上俊彦） Jun. 13, 2023, 07:30 AM | テクノロジー

- 生成AI市場は、2022年の400億ドルから2032年は約1.3兆ドルにまで成長すると見られている。
- 同市場は年42%の複利で成長すると、ブルームバーグ・インテリジェンスは報じている。
- OpenAIが2022年11月に発表したChatGPTによって、AI関連ツールのブームが過熱している。

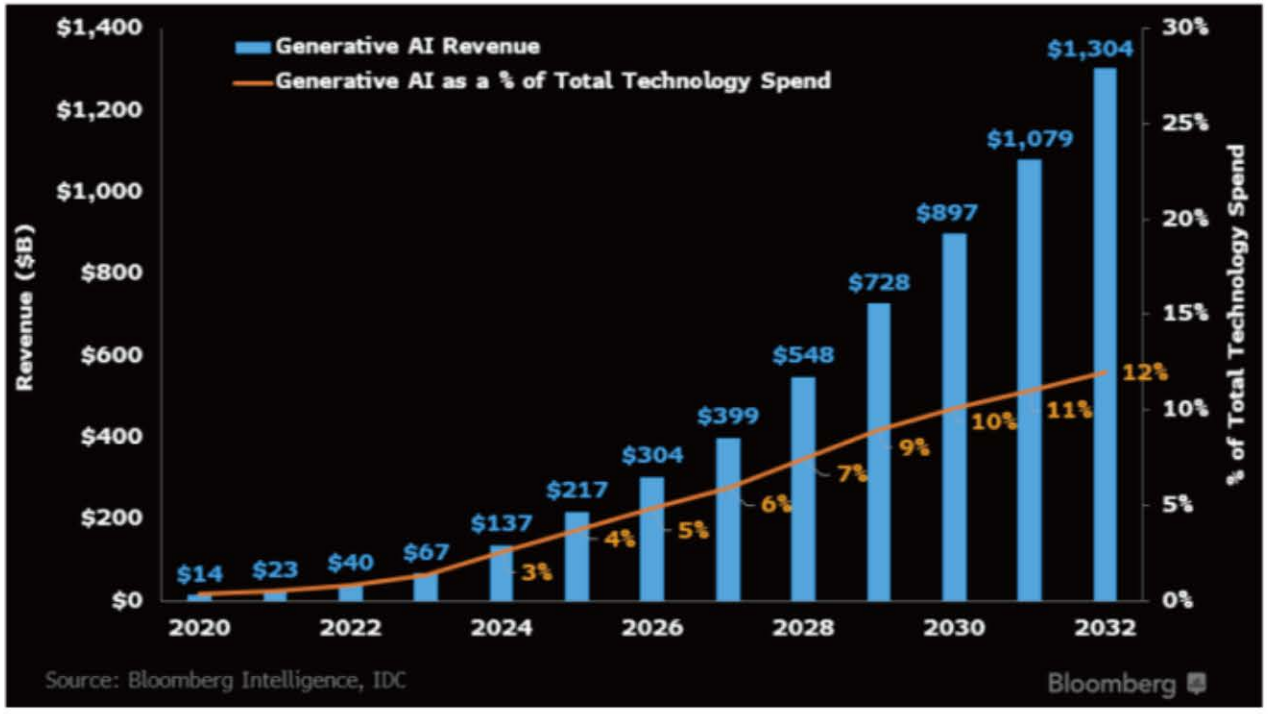
Insiderが確認したブルームバーグ・インテリジェンスの記事によると、ChatGPTやグーグルバードなどのAIツールの登場で、生成AI市場への注目が爆発的に高まり、今後10年間で約1兆3000億ドル（約181兆1850億円）規模まで成長する可能性がある。

ブルームバーグの記事では、2022年の生成AI市場は約400億ドル（約5兆5750億円）だったと見ている。それが2032年までに、年42%の複利で成長し、1兆3200億ドル（約184兆円）になる可能性があるという。

分野別の内訳では、AIアシスタント、インフラ関連、コーディングを短縮するプログラムなどのAIソフトウェアは年69%で成長し、2032年までに2800億ドル（約39兆円）まで伸びると見る。

しかし1.3兆ドルの大部分を占めるのはハードウェアで、2032年までに6410億ドル（約89兆3000億円）になるという。6410億ドルのうち、1680億ドル（約23兆4000億円）はデバイス、4730億ドル（約65兆9000億円）はインフラだ。

ハードウェアの中でも、AIサーバー、AIストレージ、コンピュータービジョンAI製品、AI会話ツールなどの売り上げが1080億ドル（約15兆円）になると試算されている。



単体のデータセンターの電力と冷却需要が3-10倍にもなることに加え、予測されている生成AI市場の急成長を支えるためには、**2027年までに現在の約6倍**の、**2030年までには現在の20倍超**ものデータセンター数が必要となる計算です。

これらの新設AIデータセンターの全てで、旧来の空冷や液冷では十分な冷却性能を得ることができずに電力コストが嵩むことに。更には莫大な初期投資の回収も、もはや正当化ができないレベルに。

DC消費電力は、2030年に全米消費の8.1%に

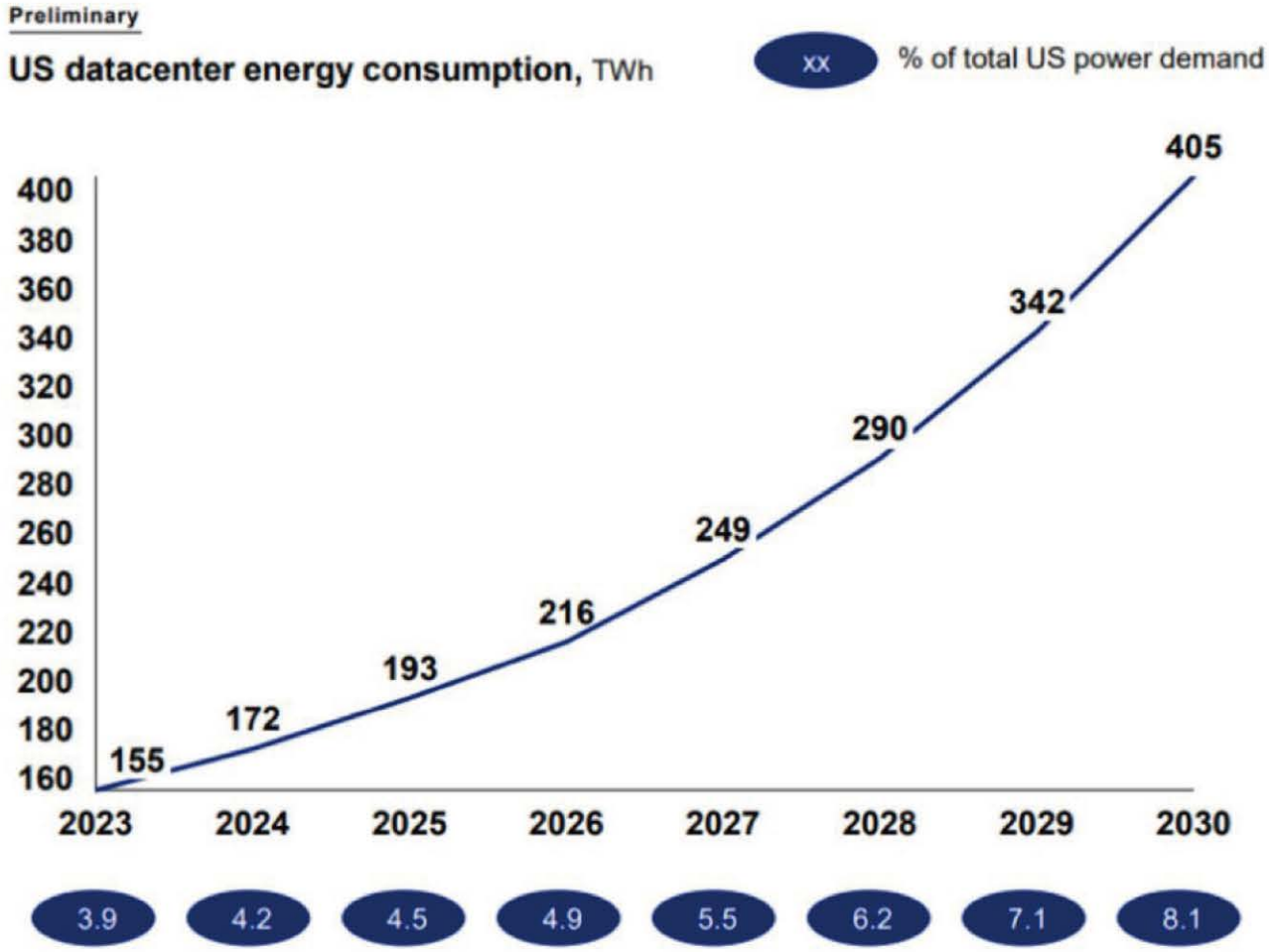
マッキンゼーグループの昨年時点の試算によると、2030年には、全米のDC消費電力が2023年から250TWh増加して405TWhとなって、**全米消費の8.1%**に達することに。

この資産には、生成AI勃興による急速なDC当たりでの電力消費増加が十分には反映されておらず、ここから更に上振れる可能性もあるものの、それだけの**電力需要を満たすための発電所と送電網が整備できない**可能性が高いことから、DC当たりの電力消費量を大幅に削減する必要性が明らかに。

GPGPUの電力消費量の削減に加えて、**DC自体の冷却手法の抜本的刷新**による電力消費削減の、双方の実現が急務。

（米国におけるDC冷却でよく使用される、河川水を使用したクーリングタワー冷却では、大量の河川水を蒸発させることから、規制が強化され今後の運用が困難に）

Exhibit 2: Data center power demand increasing to 8% of total US power demand by 2030
~250TWh of new electricity demand through 2030 driven by data centers



Source: McKinsey Energy Solutions Global Energy Perspective 2023; McKinsey datacenter demand model

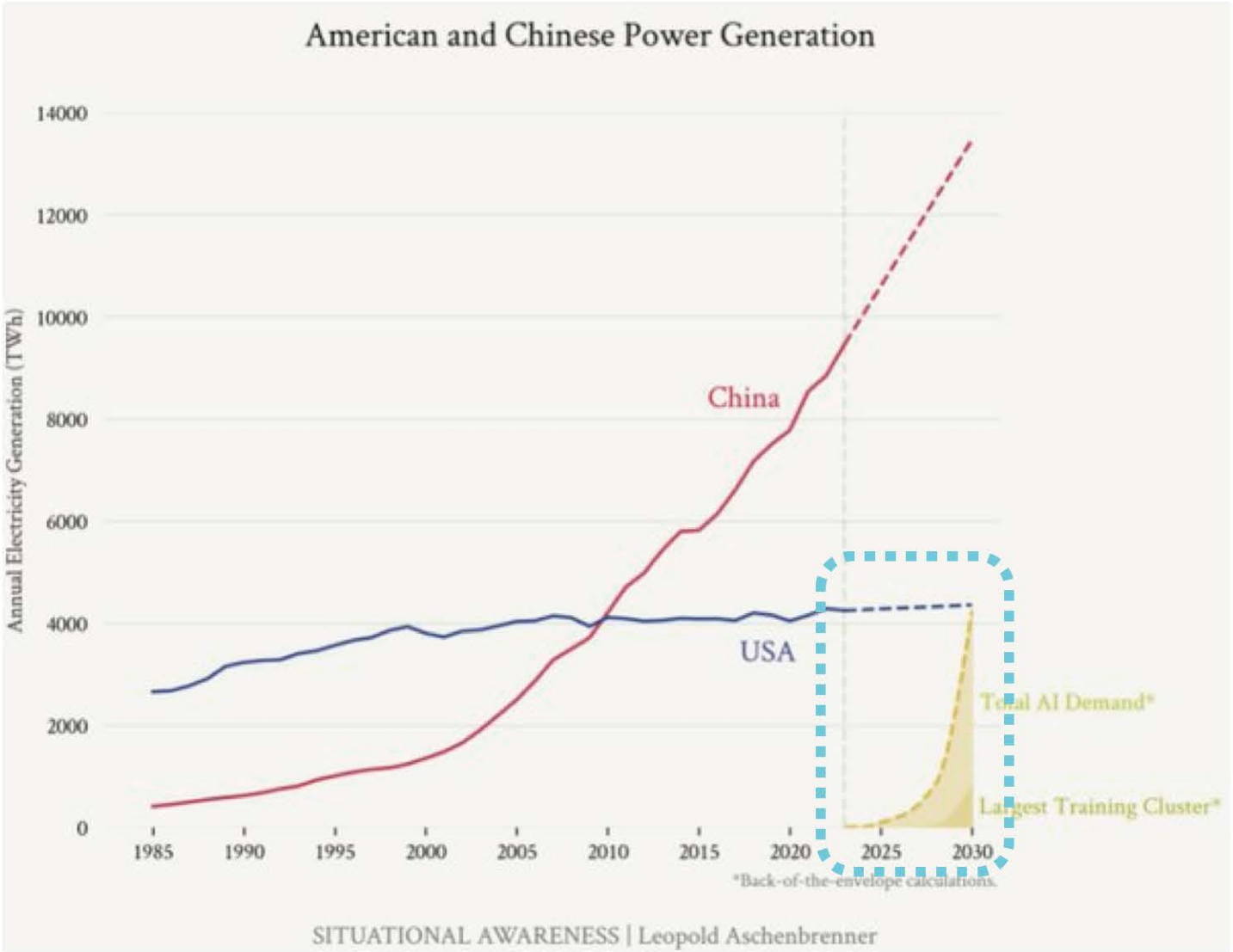
BofA GLOBAL RESEARCH

DC消費電力は、2030年に現全米消費と同等に

今年6月までOpenAI社に所属していた優秀なAI技術者のLeopold Aschenbrenner氏が作成した、非常に詳細なレポート「Situation Awareness」は、トランプ大統領候補の演説でも引用されて注目を集めています。

そこで論じられる、米中の電力消費予測は極めて衝撃的な内容であり、既に米国の2倍超の電力を消費している中国は、2030年には現在の米国の3倍を超える電力を消費することになると予想されています。
そして、米国のAIデータセンターの電力需要は、2030年には2024年現在の米国の総電力重要にも匹敵する程になることが、明確に指摘されています。

現実的には、これから6年間で、それだけの発電所を新たに建築することも、必要となる電力網を増強することも、何れも不可能でありますことから、実際は、GPGPU自体の低消費電力化と、AIデータセターの電力削減を徹底的に進める必要があります。
そのための唯一の現実解が、新開発された「水浸冷却」技術となります。



2）既存のデータセンター冷却技術

- 空冷

- ラック内部空冷＋ラック外気液冷（リアドア方式）

- 液冷（DLC (Direct Liquid Cooling)/DTC(Direct to Chip)）

- 液浸冷却
  伝熱（一相･開放）冷却方式
  沸騰（二相･密閉）冷却方式

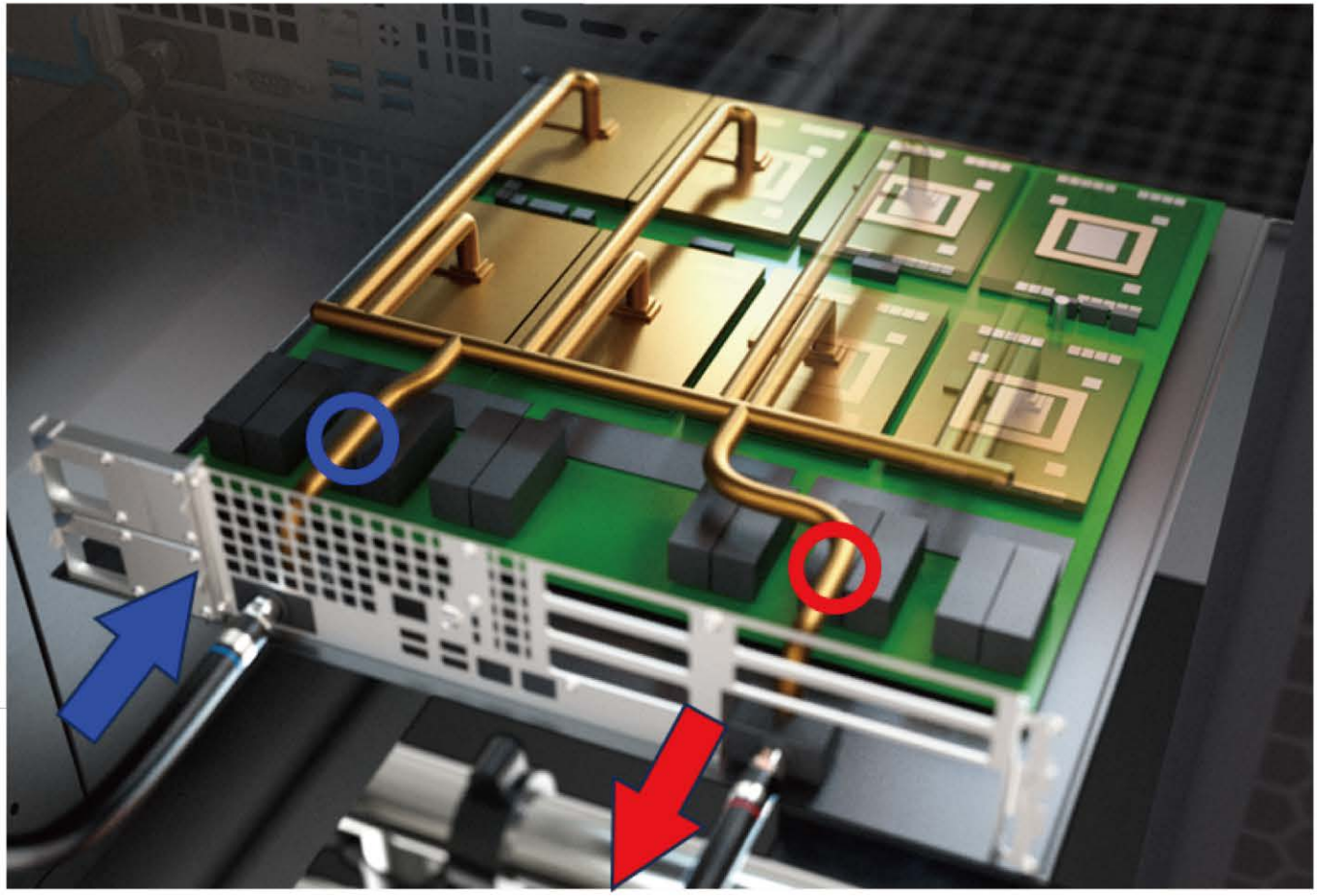
2）既存技術（空冷から液冷）

「空冷」

「ラック内部空冷
＋ラック外気液冷」
（リアドア方式）

部分「液冷」
（サーバー内空冷）

2）既存技術（部分液冷）



DLC: Direct Liquid Cooling（直接液体冷却）

DTC: Direct To Chip （直接半導体冷却）

部分液冷

：多分岐配管（マニフォールド）構造の細い金属管でGPGPU表面まで水を誘導し、温水を集めて回収する。

問題１）

：多分岐配管と細い金属管により、水の流量が十分には確保できない

⇒ GPGPU冷却は80-100℃まで

問題２）

：クーリングプレートと配管を有しない他の多くの電子部品は、空冷のまま

⇒ 液冷と空冷の両構造が必要

4）配管を排し、局所強制流水を提供

全く新しい水密構造

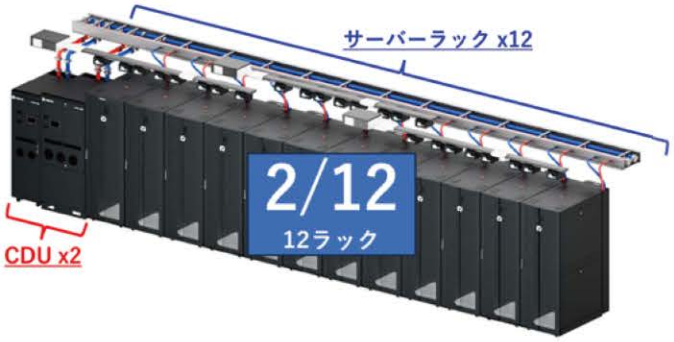
：基板全体は特殊な伝熱性の高い絶縁フィルムで被覆しつつ、GPGPU表面は水が直接的、かつ高効率に接触して伝熱する構造を持たせ、しかも水密性能を大幅に高めた構造を考案

25℃を実現する局所強制水流

：配管を排除して圧力損失を最小化した上、強力なカスタム水中ポンプ2台づつで局所強制対流を提供して水流を最大化

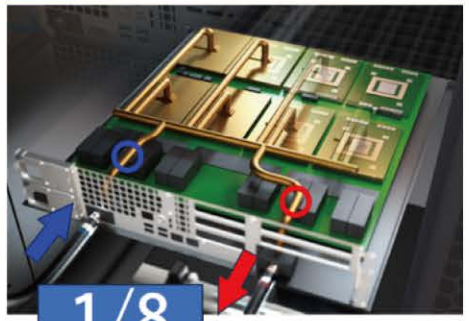
最新液冷（非液浸）と水浸冷却の水流量比較

最新の生成AIデータセンター用液冷システム事例
（12ラック＋CDU2系統で1.4MW）



CDU流量の1/6に

CDU流量の1/96に



GPGPUは100℃程度までの冷却に留まる

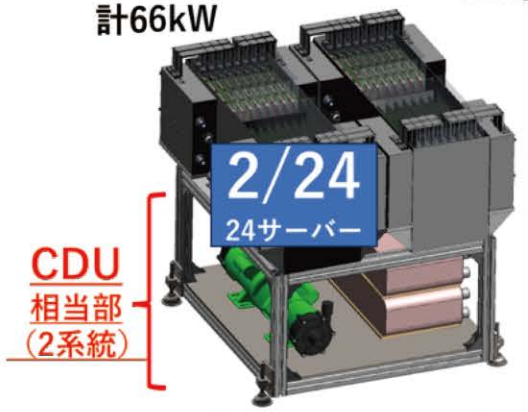
CDU流量の1/768に

20倍超

ZYRQ社試作第4世代機
24水浸槽構成例（1.6MW）

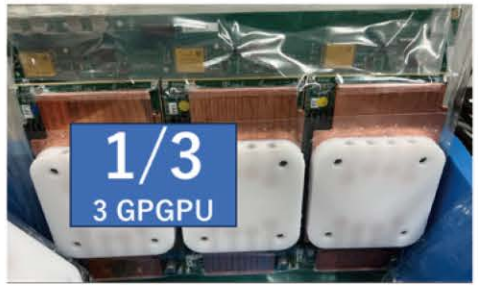
CDU流量の全量利用

2液浸槽で
計66kW



CDU流量の1/12に

配管の無い、水浸冷却と局所強制流水



GPGPUが25℃での安定低温稼働が可能

CDU流量の1/36に

個別CDU統合によりGPGPU当たり20倍超の冷却用流水量を確保して、更に局所強制水流を追加

その結果としてGPGPU温度は、100℃から25℃までに低下して電力消費の大幅削減が可能

5）水浸冷却の現在と近未来

Gen6：2024 H2～
ITRIと共同開発、熱容量密度倍増

Gen5：2024 H1～
油浸部を排し、完全な水だけの冷却実現

Gen4：2023 H2～
1,000W級CPU/GPUを25℃以下で安定駆動

Gen3：2023 H1～
自然水による冷却方式（水浸＋油浸）、地下水等を用いた熱交換

Gen2：2022～
フッ素系不活性液体による冷却、**地下水を用いた熱交換**

Gen1：2014～
フッ素系不活性液体による冷却方式、空冷式冷凍機を用いた熱交換

5）水浸冷却の現在と近未来

ZYRQ

- システムの小型化を突き詰め、同時に熱容量密度は可能な限りまで上げていく
- 第6世代 システム2は400kW/m³ (w/CDU)を目標として、ITRIと共同開発を推進する

水浸冷却システム	システム外形寸法 (cm)			熱容量	重量		熱容量密度 (kW/m³)
	横	縦	高		乾燥時	冷媒込み	
第3世代 System 3	48	50	95	30KW	115kg	180kg	132kW/m³
3ヶ月							25℃冷却
第4世代 System 1	80	84	108	64KW	345kg	485kg	88kW/m³
3ヶ月	体積1/3			熱容量1/2			熱密度x1.5
第5世代 System 1	**80**	**41**	**75**	**33KW**	**164kg**	**224kg**	**134kW/m³**
1か月	体積2/3						熱密度x1.5
第5世代 System 2	**74**	**42**	**53**	**33KW**	**170kg**	**250kg**	**200kW/m³**
3ヶ月				熱容量x1.5			熱密度x1.5
第6世代 System 1 (開発中)	**74**	**42**	**53**	**50kW**	**180kg**	**260kg**	**300kW/m³**
3-6か月				熱容量x1.33			熱密度x1.33
第6世代 System2 (計画中)	**74**	**42**	**53**	**67KW**	**170kg**	**250kg**	**400kW/m³**

5）データセンターの冷却方式別比較

冷却方式	冷媒	冷却熱密度（kW/m^3）			冷却可能温度	
		サーバー部のみ	CDU（冷媒分配器）含む	空気流路等全て含む	CPU/GPU 100-300W	GPGPU 700- 1,00W
空冷方式	空気	10	10	3	60- 100℃	冷却不可
液冷方式	水+空気	100	86	30	30- 60℃	90-100℃
液浸方式	合成油他	60	50	50	30- 60℃	80-100℃
水浸方式（第5世代）	水	331	200	200	15-20℃	25-30℃
水浸方式（第6世代）	水	662	400	400	15-20℃	20-25℃

データセンターへの寄与　「冷却能力」不足を解消　「電力」不足を大幅緩和

5）同等性能スパコン事例での比較

富岳の設置イメージ

432ラック（158,976ノード）

20,000㎡／30MW

28.2kW/m^3

（CDUと空気流路を含まず）

最新水浸冷却データセンター

160槽（2コンテナ）

50㎡／10MW

設置面積：1/400

※富岳のコンピュータルームと空調・配電盤ルームのスペースは6,600㎡であり、この数値との比較では、1/133

消費電力：▲67%削減