AIによる破壊：データセンター設計の課題と指針

ホワイトペーパー110

第 1.1 版

エネルギー管理研究センター

Victor Avelar
Patrick Donovan
Paul Lin
Wendy Torell
Maria A. Torres Arango

エグゼクティブサマリー

大規模なトレーニングクラスターから小規模なエッジの推論サーバーに至るまで、データセンターのワークロードに占める AI の割合が大きくなっている。これは、より高いラック電力密度への移行を表している。AI のスタートアップ、企業、コロケーションプロバイダー、インターネット大手は、データセンターの物理インフラの設計と管理に対するこれらの密度増大の影響を考慮する必要がある。本ホワイトペーパーでは、AI ワークロードについて関連する属性と傾向を説明し、その結果生じるデータセンターの課題について説明する。これらの課題に対処するための指針を、電源・冷却・ラック・ソフトウェア管理などの物理インフラカテゴリごとに示す。

はじめに

近年、人工知能（AI）の成長が驚くほど加速しており、私たちの生活・仕事・技術との関わり方に変革が起こっている。生成 AI（ChatGPT など）が、この成長の触媒となっている。予測アルゴリズムは、ヘルスケア [1] や金融から製造 [2]、運輸 [3]、エンターテイメントに至るまで、さまざまな業界に影響を与えている。AI に関わるデータ要件が、新しいチップとサーバー技術を推進し、その結果、ラックの電力密度が極度に高まっている。同時に、AI に対する大きな需要も存在している。これらが相まって、この需要をサポートするデータセンターの設計と運用に新たな課題を生み出している。

AI の成長予測

現在、AI は 4.3 GW の電力需要を生み出していると推定されており、これが 26%～36%の CAGR で成長し、2028 年までに総需要が 13.5 GW～20 GW になると予測されている。この成長は、データセンター全体の電力需要の CAGR 11%の 2～3 倍である。詳細については、表 1 を参照されたい。重要な知見の 1 つとして、新しくトレーニングされたモデルが本番環境に移行されるにつれて、推論 [4] による負荷が時とともに増加する。実際のエネルギー需要は、次世代のサーバー、より効率的な命令セット、チップパフォーマンスの向上、継続的な AI 研究などの技術要因に大きく依存する。

表 1

データセンターにおける AI ワークロードの概要。

シュナイダーエレクトリックによる推定	2023	2028
データセンターの総ワークロード	54 GW	90 GW
AI によるワークロード	4.3 GW	13.5～20 GW
AI によるワークロード（全体に占める割合%）	8%	15～20%

[1] Federico Cabitza, et al., Rams, hounds and white boxes:Investigating human–AI collaboration protocols in medical diagnosis, Artificial Intelligence in Medicine, 2023, vol 138 Federico Cabitza, et al., Rams, hounds and white boxes:Investigating human–AI collaboration protocols in medical diagnosis, Artificial Intelligence in Medicine, 2023, vol 138

[2] Jongsuk Lee, et al., Key Enabling Technologies for Smart Factory in Automotive Industry:Status and Applications, International Journal of Precision Engineering and Manufacturing, 2023, vol 1 Jongsuk Lee, et al., Key Enabling Technologies for Smart Factory in Automotive Industry:Status and Applications, International Journal of Precision Engineering and Manufacturing, 2023, vol 1

[3] Christian Birchler, et al., Cost-effective simulation-based test selection in self-driving cars software, Science of Computer Programming, 2023, vol 226

[4] 定義については、「AI の属性と傾向」セクションを参照のこと。

シュナイダーエレクトリックによる推定	2023	2028
AI のワークロード（トレーニングと推論）	20%トレーニング、80%推論	15%トレーニング、85%推論
AI のワークロード（セントラルとエッジ）	95%セントラル、5%エッジ	50%セントラル、50%エッジ

本ホワイトペーパーでは、電力、冷却、ラック、ソフトウェア管理などのデータセンターの物理インフラカテゴリごとに、課題を生み出す要因となる重要な AI の属性と傾向について説明する。次に、これらの課題に対処する方法についての指針を提供する。[5]最後に、データセンター設計が今後どう変化するかについての将来的な展望を示す。本ホワイトペーパーは、AI を物理インフラシステムに適用することについては取り扱わない。**次世代の物理インフラシステムでは、最終的にはさらに多くの AI が活用されることになるが、このホワイトペーパーでは、現在利用可能な既存のシステムで AI のワークロードをサポートすることに焦点を当てる。**

AI の属性と傾向

物理インフラの課題の根底には、次の 4 つの AI の属性と傾向が存在する：

- AI のワークロード
- GPU の熱設計電力（TDP）
- ネットワーク遅延
- AI のクラスターサイズ

AI のワークロード

AI のワークロードは、一般的にトレーニングと推論という 2 つのカテゴリに分類される。

トレーニングのワークロードは、大規模言語モデル（LLM）などの AI モデルをトレーニングするために使用される。このホワイトペーパーで言うトレーニングのワークロードのタイプは、今日のデータセンターに課題をもたらしている大規模な分散型ト

[5] この指針は、ハイパフォーマンスコンピューティング（HPC）などの他の高密度ワークロードにもあてはまる。HPC アプリケーションとの主な違いは、カスタム IT、電源、冷却、ラックソリューションを採用する 1 回限りのインストールとなる傾向があることである。対照的に、AI アプリケーションに対する膨大な需要には、拡張するための標準装備（IT およびサポートするインフラ）が必要となる。

レーニング（多数のマシンを並行して実行する形式[6]）である。これらのワークロードでは、アクセラレーターと呼ばれるプロセッサを備えた専用サーバーに大量のデータを供給する必要がある。グラフィックスプロセッシングユニット（GPU）はアクセラレーター[7]の一例である。アクセラレーターは、LLM のトレーニングで使用されるような並列処理タスクを実行する際に非常に効率的である。トレーニングにはサーバーに加えて、データストレージおよび、それをすべて接続するネットワークも必要になる。これらの要素は AI クラスターとして知られるラックのアレイに組み上げられ、実質的に単一のコンピューターとしてモデルがトレーニングされる。適切に設計された AI クラスターのアクセラレーターは、数時間から数か月にわたるトレーニング期間の大部分で、ほぼ 100%の使用率で動作する。これは、トレーニングクラスターの平均消費電力がピーク消費電力にほぼ等しいことを意味する（ピーク対平均比 ≈ 1）。

モデルが大きくなるほど、より多くのアクセラレーターが必要になる。大規模な AI クラスターのラック密度は、GPU のモデルと量に応じて 30 kW から 100 kW の範囲になる。クラスターの範囲は数ラックから数百ラックになることがあり、通常は使用されるアクセラレーターの数で表現される。たとえば、22,000 H100 GPU クラスターは約 700 ラックを使用し、平均ラック密度 44 kW で約 31 MW の電力を必要とする。この電力には、冷却などの物理インフラ要件は含まれていないことに注意されたい。最後に、トレーニングワークロードはモデルを「チェックポイント」として保存する。クラスターに障害が発生したり電源が失われたりした場合でも、中断したところから続行できる。

推論とは、新しいクエリ（入力）に対する出力を予測するために、以前にトレーニングされたモデルが運用環境で利用されることを意味する。ユーザーの観点からは、出力の精度と推論時間（つまりレイテンシー）の間にはトレードオフの関係が存在する。科学者であれば、高精度の出力を得るために、割増料金を払ってクエリ間の待ち時間を長くしても構わないと考えるであろう。一方で書き物のアイデアを探しているコピーライターであったら、即座に回答が得られる無料のチャットボットを重宝するであろう。つまり、ビジネスニーズによって推論モデルのサイズが決まるが、元のトレーニング済みモデル全体が使用されることはほとんどない。代わりに、モデルの軽量バージョンが導入され、精度の損失は許容範囲内で推論時間が短縮される。

推論ワークロードでは、モデルが大規模な場合にアクセラレーターを使用することが多く、アプリケーションによっては CPU に大きく依存する場合もある。自動運転車、レコメンデーションエンジン、ChatGPT などのアプリケーションには、要件に

[6] モデル内に多数のパラメーターとトークンがあるため、モデルのトレーニングにかかる時間を短縮するのに、処理ワークロードを多くの GPU に分割する必要がある。

[7] アクセラレーターの他の例としては、テンソルプロセッシングユニット（TPU）、フィールドプログラマブルゲートアレイ（FPGA）、特定用途向け集積回路（ASIC）がある。

合わせて「チューニング」された異なる IT スタックが搭載されていることが多い。インスタンスごとのハードウェア要件は、モデルのサイズに応じて、エッジデバイス（スマートフォンなど）から複数のラックのサーバーに及ぶこともある。これは、ラック密度が数百ワット～10 kW 以上までの範囲に及ぶ可能性があることを意味する。トレーニングとは異なり、推論サーバーの数はユーザー／クエリ数に応じて増加する。実際、よく用いられるモデル（ChatGPT など）では、クエリが 1 日あたり 100 万の単位となっており、推論にはトレーニングの場合よりもさらに多くの数のラックが必要になる可能性がある。最後に、推論ワークロードは多くの場合ビジネスクリティカルであり、柔軟で強靭でなければならない（UPS や地理的冗長性など）。

GPU の熱設計電力（TDP）

ストレージとネットワークがなければトレーニングや推論は不可能であるが、GPU が AI クラスターの消費電力の約半分を占めるため、GPU に焦点を当てる [8]。GPU の能力は、世代が新しくなるたびに向上する傾向にある。ワット単位で測定されるチップの消費電力は、通常、TDP で指定される。ここでは GPU について具体的に取り扱うが、TDP が増大するというこの一般的な傾向は他のアクセラレーターにも当てはまる。GPU 世代ごとの TDP の増加は、より短い時間とより低いコストでモデルをトレーニングし推論するために、操作数の増加に合わせて GPU が設計された結果である。表 2 は、TDP とパフォーマンスの観点から 3 世代の Nvidia GPU を比較したものである。[9]

表 2

さまざまな世代の Nvidia GPU の TDP とパフォーマンス

GPU	TDP（W）[10]	TFLOPS[11]（トレーニング）	V100 のパフォーマンス	TOPS[12]（推論）	V100 のパフォーマンス
V100 SXM2 32GB	300	15.7	1X	62	1X
A100 SXM 80GB	400	156	9.9X	624	10.1X
H100 SXM 80GB	700	500	31.8X	2,000	32.3X

ネットワーク遅延

ネットワーク遅延分散トレーニングでは、コンピューティングネットワークファブリックを確立するには、すべての GPU にネットワークポートが必要である。たとえ

[8] 400W では、NVIDIA V100 GPU がこのクラスターの 55%を占め、700W では H100 が 49%を占める。

[9] GPU がこれらのパフォーマンス向上の鍵となるが、メモリや GPU 間通信の増加など、改良された GPU を活用するために他のシステムの改善も行われている。

[10] V110, A100, H100

[11] TFLOPS - 1 秒あたりのテラ（兆）回浮動小数点演算 - テンソル float 32（TF32）精度での行列乗算のスループットの測定尺度、一般にトレーニングワークロードで使用される。 V110, A100, H100

[12] TOPS は 1 秒あたりのテラ（兆）回オペレーションであり、8 ビット整数（INT8）精度での整数演算スループットの尺度であり、通常は推論ワークロードで使用される。 V110, A100, H100

ば、AI サーバーに 8 つの GPU がある場合、そのサーバーには 8 つのコンピューティングネットワークポートが必要となる。このコンピューティングファブリックにより、大規模な AI クラスター内のすべての GPU が高速（800 ギガビット/秒など）で連携して通信できるようになる。モデルのトレーニングにかかる時間とコストを削減するためには、GPU の処理速度の向上に加えて、ネットワークの速度も向上する必要がある。たとえば、100 GB/秒のコンピューティングファブリックでメモリからのデータを 900 GB/秒で処理する GPU を使用した場合、GPU の次の動作を調整するためにネットワーク上で待機するので平均 GPU 使用率が低下する。これは、低速ネットワークを介して通信する一連の高速センサーを搭載した 500 馬力の自動運転車を購入するようなものである。車の速度はネットワーク速度によって制限されるため、エンジンのパワーを最大限に活用できない。

高速ネットワークケーブルは高価である。たとえば、InfiniBand 光接続のコストは銅線の 10 倍のオーダーである。したがって、データサイエンティストは IT チームと協力して、ネットワークのケーブル配線距離が許容可能な遅延内に収まるように、銅線を使用した AI トレーニングクラスターを指定しようと試みる。ラックあたりのポートを増やすとケーブル距離は短くなるが、ラックあたりの GPU の数が増加しラック密度が増加する。その結果ラッククラスターが非常に大きくなり、遅延により設計者はファイバーへの切り替えを余儀なくされ、コストが増加する。推論ワークロードの GPU を並列化するのはさらに困難であるため、このラック密度の関係は通常、推論には当てはまらないことに注意されたい。[13]

AI のクラスターサイズ

上で説明したように、大規模なモデルをトレーニングするには、何千もの GPU が連携して動作しなければならないこともある。GPU がクラスターの消費電力の約半分を占めることを考えると、GPU の数はデータセンターの消費電力を見積もるのに役立つ指標である。図 1 は、3 つの GPU 世代にわたる AI トレーニングクラスターの GPU の数の関数としてデータセンターの電力消費量を推定したものである（表 2 より）。大雑把に言うと 40,000 kW の発電所は、米国の平均的な家庭約 31,000 世帯に電力を供給することができる。これら 3 本のトレンドラインは同じ生産性を表すものではないことに注意されたい。つまり、H100 GPU を搭載したデータセンターの消費電力は V100 GPU を搭載したデータセンターを上回るが、H100 データセンターの生産性は消費電力の増大をはるかに凌ぐものである。

[13] Accelerating Deep Learning Inference with Hardware and Software Parallelism, 2020 年 4 月

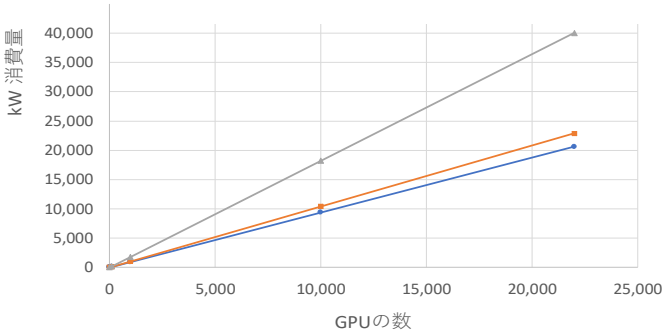


図 1：

GPU の数の関数によるデータセンターの電力消費量の見積もり

データセンターの PUE = 1.3

生産性の向上はこのグラフには示されていないことに注意

説明した 4 つの属性と傾向は、ラックの電力密度に直接影響する。大部分のデータセンターは現在、約 10～20 kW のピークラック電力密度をサポート可能である [14]。しかし、AI クラスター内にすべて 20 kW を超える数十または数百のラックを配置すると、データセンター運営者にとって物理インフラの課題が生じる。これらは電力に固有の場合もあれば、複数の物理インフラカテゴリにまたがる場合もある。これらの課題は克服できないわけではないが、運営者は IT だけでなく、物理インフラ、特に既存のデータセンター施設要件を十分に理解した上で作業を進める必要がある。施設が古いほど、AI トレーニングのワークロードをサポートすることが困難になる。以下の主要なセクションでは、物理インフラカテゴリごとにこれらの課題をより詳細に説明し、これらの課題を克服するための指針を示す。推奨される設計アプローチの中には、新しいデータセンターの構築にのみ適用されるものもあるが、その他のアプローチは新築とブラウンフィールド（改修）建物の両方に関連するものもある。

電力

AI ワークロードは、開閉装置・配電・ラック配電ユニット（rPDU）などのパワートレインに影響を与える 6 つの主要な課題を呈する。

- 120/208 V 配電の導入は実際的でない
- 小さな配電ブロックサイズは IT スペースを無駄にする
- 標準の 60/63 A ラック PDU の導入は実際的でない
- アークフラッシュの危険性の増大により作業が複雑化する
- 負荷の不等時性がないと、上流のブレーカーがトリップするリスクが増大する
- ラックの温度が高いと、故障や危険のリスクが高まる

120/208 V 配電の導入は実際的でない

北米のデータセンターで歴史的に使用されてきた電圧である 120/208 V は、密度が比較的低く（ラックあたり 2～3 キロワット程度）、サーバーに 120 V の電源コードが供給されていた時代にはその目的を果たすことができた。今日、AI クラスターなどの高密度負荷ではこの電圧は低すぎる。これらの負荷に 120/208 V で電力を供給することは依然として可能であるが、電力はボルトとアンペアの積に等しい（P = V x A）という関係にまつわる課題が生じる。この式が示すように、電圧が低いほど、

[14] Uptime Institute, *Rack Density is Rising*, 2022 年 12 月

同じ電力に対してより高い電流が必要となる。したがって、より大きな電流を安全に供給するには配線を太くする必要がある。

ここで、8 台の HPE Cray XD670 GPU アクセラレーションサーバーからなる AI トレーニングラックを考えてみる。ラック密度は合計 80 kW である。120/208 V では、1N 冗長性でラックに電力を供給するには 5 つの 60 アンペア回路が必要になる（各回路は 120 V x 3 相 x 60 A x 80% ディレーティング = 17,280 W = 17.3 kW に相当）。2N が必要な場合（ただし、AI トレーニングロードでは一般的ではない）、この数は 2 倍の 10 になる。ラックあたり 5〜10 回路が存在し、100 ラックの AI クラスターに分散された電源ケーブルで混雑する様子を想像されたい。その結果、電源ケーブルがラック上やラック付近にぶら下がり、その場しのぎの乱雑に張り巡らされる可能性があり、人的ミスやエアフローの制約などの問題が発生するおそれがある。これは望ましい状態ではない。さらに、回線の数が多すぎると、設置と管理にコストがかかる。

ガイダンス：電圧を 2 倍にすることは電力を 2 倍にすることを意味するため、120/208 V 配電を備えた既存のデータセンターは、配電設備を 240/415 V に改修することが望ましい。新しいデータセンターは、今から 240/415 V を念頭に置いて設計する必要がある。このトピックの詳細については、ホワイトペーパー128 *データセンター向けの高効率 AC 配電*を参照されたい。これには、240/415 V の電力をどのように配電するかという制約に関連する次の課題がある。

多くの国が AI ラックの電力需要を満たすのに適した 230/400 V の高電圧で電力を配電しているため、世界の大部分では同じ課題は生じていないことに注意されたい。

小さな配電ブロックサイズは IT スペースを無駄にする

データセンターの配電には、変圧器ベースの配電ユニット（PDU）、リモート電源パネル（RPP）、およびバスウェイの 3 つの主なタイプがある。配電ブロックサイズは、各配電ソリューションの容量（kW）を表す。240/415 V（230 V の IEC 諸国）というより高い配電電圧を使用しても、従来の配電ブロックサイズは今日の AI クラスターの容量をサポートするには小さすぎる。10 年前ならば、300 kW（120/208 V で 833 A）の配電ブロックで 100 ラック（平均ラック密度 3 kW/ラックで 20 ラック列を 5 列）に対応可能であった。現在その同じブロックは、NVIDIA DGX SuperPOD の最小構成（36 kW/ラックで 358 kW 10 ラック列を 1 列）をサポートすることさえできない。ラックの単一列に複数の分配ブロックを使用するのは、さまざまな理由から実際的ではない。たとえば、PDU と RPP の必要面積が少なくとも 2 倍になる。複数のブロックを使用すると、単一の高容量ブロックと比較してコストも増加する。

ガイダンス：高密度クラスターの要求を満たすには、配電ブロックサイズを大きくする必要がある。少なくともクラスター列全体を収容できる十分な大きさの配電ブロックサイズを選択することが推奨される。800 アンペアのブロックサイズが、240/415 V 配電の 3 つの配電タイプすべてで現在利用可能な標準容量サイズである。これにより、576 kW（461 kW ディレーティング）が提供される。

標準の 60/63 A ラック PDU の導入は実際的でない

より高い電圧であっても、標準の rPDU で十分な容量を提供することは依然として課題である。ほぼすべての意思決定者は、リードタイムが短く、すぐに入手可能で、コスト効率が高く、複数のベンダーから同様の構成で販売されているため、既製の rPDU を好む。

現在、最大容量の標準の既製品 rPDU の定格は 60 A（NEMA）/ 63 A（IEC）である。表 3 は、さまざまな電流定格および電圧における rPDU の使用可能な容量を示している。これに基づくと、60 A 定格と 63 A 定格では、単一の rPDU の容量がそれぞれ 34.6 kW と 43.5 kW に制限されることがわかる。このため、これを超えるラック電力密度をどのように処理するのが最適かというジレンマが生じる。

表 3

サーキットブレーカーのアンプ定格と電圧（ラインから中性点まで）に基づく、rPDU あたりの使用可能な三相電力密度

上部：NEMA（北米など）
下段：IEC（欧州など）

	標準	標準	カスタム	カスタム	カスタム	カスタム
NEMA	40 A	60 A	100 A	125 A	150 A	175 A
120/208 V	11.5 kW	17.3 kW	28.8 kW	36.0 kW	43.2 kW	50.4 kW
240/415 V	23.0 kW	34.6 kW	57.6 kW	72.0 kW	86.4 kW	100.8 kW

これらの値は、一般的なコード要件に基づいて 80%にディレーティングされることに注意されたい。

IEC	32 A	63 A	100 A	125 A	150 A	160 A
230/400 V	22.1 kW	43.5 kW	69.0 kW	86.3 kW	103.5 kW	110.4 kW

ガイダンス：ラック密度が 34.6 kW（NEMA）および 43.5 kW（IEC）を超える場合、2 つのアプローチがある。

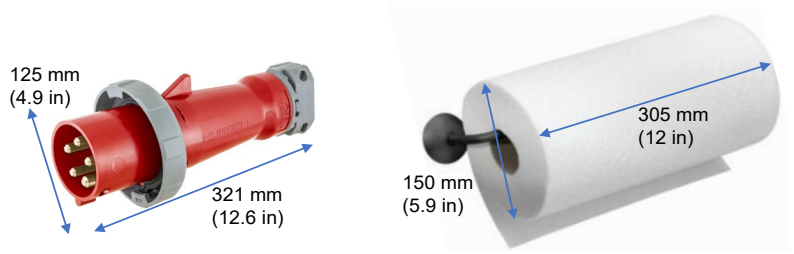
1. 複数の標準既製品 rPDU
2. 60 A および 63 A を超えるカスタム rPDU

現在のほとんどのゼロ U rPDU の高さはおよそ 2 メートル（80 インチ）である。これらの標準オファーでは、単一の空冷ラックに最大 4 つの rPDU を収容できる可能性がある（たとえば、4 x 60/63 A rPDU は 138 kW/174 kW）。あるいは、液体冷却マニホールドが必要な場合は、単一ラックに 2 つの rPDU を設置する（たとえば、2 x 60/63 A rPDU は 69 kW/87 kW）。これらの rPDU を組み合わせて容量を増やしたり、冗長性（2N など）に対応したりすることができる。

rPDU の数量によるスペースの制約がある場合は、カスタム rPDU を推奨する。たとえば表 3 に示すように、北米では 175 A rPDU、欧州では 150 A の rPDU を使用して 100 kW ラックに電力を供給できる。カスタム rPDU には、ピンおよびスリーブコネクターが付属することや、配線接続されることもあり、コンセントの数やタイプを柔軟に選択できる。定格電流が高くなると、ピンおよびスリーブコネクターの物理的なサイズにより、ラックへの取り付けと給電にさらに多くの作業が必要になる（図 2 を参照）。定格電流が 60 A を超える場合、設置と運用には電気技師が必要になる場合があることに注意されたい。

図 2：

240/415 V 125 A ピンおよびスリーブコネクター（ペーパータオルのロールのサイズとの比較）
このような大型のコネクターのペアを嵌合するのは、一人では困難である



アークフラッシュの危険性の増大により作業が複雑化する

ホワイトペーパー194 *データセンターIT スペースにおけるアークフラッシュの考慮事項*によると、「アークフラッシュ」という用語は、短絡電流が空気中を流れるときに起こる現象のことを表している。アークフラッシュでは、電流が文字通り、ある一点からもう一点まで空気中を伝わり、入射エネルギー[15] として知られる大量のエネルギーが 1 秒以内に放出される。このエネルギーは、熱・音・光・爆発的な圧力の形で放出され、これらはすべて怪我を引き起こすおそれがある。具体的な怪我としては、火傷・失明・感電・難聴・骨折などがある。

rPDU の電流定格を増やすと、ワイヤの直径が大きくなり、より大きな故障電流が rPDU に流れる。rPDU に流れる故障電流が、1.2 カロリー/cm2 以上の入射エネルギーになる場合、適切な訓練を受け個人用保護具（PPE）を装着した作業員でなければそのエリアに入ることができない [16]。rPDU の電流定格が増加すると、リスクが増加する。データセンター職員の安全は、対処が必要な課題である。

ガイダンス：非常に多くの変数が関係するため、アークフラッシュのリスク評価から始めて、利用可能な故障電流を分析することを推奨する。これにより、特定のサイトに最適なソリューションが得られる。この解析は、中電圧機器からラックレベルに至るまで実施することが重要である。ソリューションの例を次に挙げる：

[15] NFPA 70E（2015）によると、入射エネルギーとは「電気アーク現象中に発生し、発生源から一定の距離にある表面に加えられる熱エネルギーの量」である。

[16] 詳細については、ホワイトペーパー13、「*通電機器交換時の電気的リスクの軽減*」および 194、「*データセンターIT スペースに関するアークフラッシュの考慮事項*」を参照のこと。

- より高いインピーダンスを持つ上流トランスを指定する
- ラインリアクトル（つまりインダクター）を使用して短絡電流の流れを妨げる
- 電流リミッターブロックを使用する
- 電流制限ブレーカーを使用する

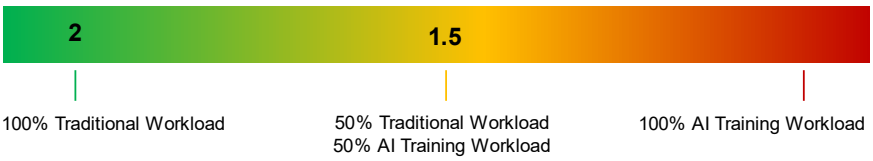
アークフラッシュの危険への対処の詳細については、ホワイトペーパー*アークフラッシュの軽減*およびホワイトペーパー253 *大規模データセンターにおける MV 短絡電流制限の利点*を参照のこと。

負荷の不等時性がないと、上流のブレーカーがトリップするリスクが増大する

さまざまなデータセンターのワークロードの電力消費は、通常、不規則なタイミングでピークに達する。統計的に言えば、これらすべてのピークが同時に発生する確率は非常に低い。したがって、個々のワークロードすべてのピークを合計し、この値を平均消費電力の合計で割ると、一般的な大規模データセンターでは、ピーク対平均比が 1.5 から 2 以上になることがわかる。これにより、設計者は電源および冷却システムを「オーバーサブスクライブ」することができる。しかし、「AI の属性と傾向」セクションで説明したように、AI トレーニングの負荷には不等時性がない。これらのワークロードは、ピーク電力で数時間、数日、さらには数週間にわたって実行される可能性がある。その結果、上流の大型ブレーカーがトリップする可能性が高まる。これは、家庭内で多数の大きな負荷が同時に動作し、メインパネルのブレーカーが落ちた場合に起こる状況と似ている。図 3 は、データセンターの負荷が 100% AI 負荷に移行するときのピーク対平均比（不等率とも呼ばれる）の典型的なスペクトルを示す。

図 3

100%従来の混合負荷から 100% AI トレーニングワークロードまでの一般的なピーク対平均比のスペクトル



ガイダンス：AI トレーニングワークロードの 60～70%を超える新しいデータセンターホールの場合は、下流の個々のフィーダブレーカーの合計に基づいてメインブレーカーのサイズを決定することが推奨される。言い換えれば、ピーク対平均比が 1 であると仮定する。すなわち平均消費電力がピーク消費電力に等しいとする。オーバーサブスクライブして、不等時性に依存するという慣習は推奨されない。

既存のデータセンターの場合、上流のブレーカーがサポートできる AI 負荷の合計を計算する。たとえば、AI ワークロードクラスターの上流に 1,000 A のメインブレーカーがある場合、AI 負荷の合計が 1,000 A を超えないようにする。

ラックの温度が高いと、故障や危険のリスクが高まる

密度の上昇と運用効率重視の間で、IT 環境はますます高温になっている。動作温度が高くなると、冷却システムの効率が向上するが、コンポーネントへのストレスも増大する。コンポーネントが定格外の温度にさらされると、次のような結果が生じる可能性がある：

- **コンポーネントの早期故障** – システムは初日には期待どおりに動作するが、仕様範囲外の条件にさらされるとコンポーネントの期待寿命が大幅に短くなるおそれがある。
- **安全上の問題** – 動作範囲がコードの定格外である場合、コードの溶融などの安全上の問題が発生するおそれがある。

IEC 60320 は電源コードの接続に関して、世界中のほぼすべての地域で認知され使用されている国際規格である。高温に耐えられる特定の IEC コネクターがある。表 4 は、標準 C19/C20 コネクターと高温 C21/C22 コネクターを比較したものである。

表 4

250 V および 16/20 A 用 IEC 60320 標準コネクターと高温コネクターの比較

	メス	オス	限界値	備考
標準			65˚C	C20 は一般にジャンパーケーブルとして使用され、ラック PDU から高出力 IT デバイスに電力を供給する。
高温度			155˚C	C21 は、C22 または C20 コネクターと嵌合し、温度が C19 定格を超える場合に使用される。

ガイダンス：AI クラスター内のすべての負荷を分析して、適切なコネクターとコンセントが使用されていることを確認することを推奨する。AI サーバーのように計算負荷が高密度になると、C21/C22 コネクターがより一般的になってきている。AI サーバーは多くの場合、これらの高温定格コード/コンセントを使用して構成されるが、ラック内の他のデバイス（トップオブラックスイッチなど）はそうでない場合がある。機器が動作する環境を理解し、ラック PDU とそのすべてのサブコンポーネントを含むすべてのデバイスの定格が適切に定められていることを確認することが重要である。

ラック PDU を指定するときは、電圧・アンペア数・コンセント数だけでなく、温度定格も考慮することが重要である。このタイプのアプリケーション向けに、高温定格の rPDU が市販されている。通常コストは高くなるが、その追加コストは一般的に、発生するおそれのある潜在的な障害コストを補ってくれる。動作条件が想定どおりであることを検証するために、ラックの背面に温度センサーを配置する（DCIM によって監視を行う）ことも推奨される。

空調・冷却

AI トレーニングサーバークラスターの高密度化により、増加する TDP に対処するために空冷式から水冷式への進化が余儀なくされている。密度の低いクラスターや推論サーバーでは、引き続き従来型のデータセンター冷却が使用されるが、データセンター運営者は次の 6 つの主要な冷却課題に対処する必要がある：

- 空冷は 20 kW/ラックを超える AI クラスターには適さない
- 標準化された設計手法がなく設置場所に制約があるため、液冷の改修は複雑となる
- TDP が将来どうなるか不明であるため、冷却設計が陳腐化するリスクが高まる
- 経験不足のため、設置・運用・メンテナンスが複雑になる
- 液体冷却により IT ラック内の漏れのリスクが増加する
- 液体冷却を持続的に動作させるための流体の選択肢は限定的である

空冷は 20 kW/ラックを超える AI クラスターには適さない

IT 用の液体冷却は、特殊なハイパフォーマンスコンピューティングのために半世紀以上にわたって使用されてきた。空冷は主流の選択肢であり、ホットアイルの封じ込めを適切に設計すれば、約 20 kW の平均ラック電力密度に対応できる。単一の 8〜10U AI サーバーで 12 kW を消費すると、この 20 kW のしきい値を簡単に超えてしまう。この課題に加えて、遅延の制限により、大規模な AI クラスター内のサーバーは（ラック密度を下げるために）分散化できない。AI トレーニングサーバーの水冷バージョンがますます利用可能になり、TDP の増加により水冷のみを使用したものもある。

ガイダンス：ラックあたり 20 kW 以下で構成された小型の AI クラスターと推論サーバーラックは空冷で対応可能である。これらのラックでは、より効果的かつ効率的な冷却を確保するために、適切なエアフロー管理手法（例：ブランクパネル、アイルコンテインメントなど）を適用することが望ましい。空冷システムに依然として制約がある場合は、AI サーバーを複数のラックに分散することでラック密度を低減するという戦略がある。たとえば、クラスターが 20 kW/ラックの 20 ラックである場合、サーバーを 40 ラックに分散すると、ラック密度は 10 kW/ラックに減少する。ネットワークのケーブル配線距離が長くなることにより AI クラスターのパフォーマンスが低下する場合、ラックの拡張は不可能になる場合があることに注意されたい。

AI ラック密度が 20 kW を超える場合は、水冷サーバーを強く検討する必要がある。液体冷却技術とアーキテクチャには数種類のものが存在する。ダイレクトチップ（DTC）（導伝プレートまたはコールドプレートとも呼ばれる）と浸漬が、2 つの主要なカテゴリである。浸漬と比較して、既存の空冷との互換性が高く、改造アプリケーションも容易であるため、現在はダイレクトチップが好まれている。データセンター運営者は選択肢がある場合、パフォーマンスを向上させ、エネルギーコストを削減

するために水冷サーバーを選択することが望ましく、これにより投資の増加分を相殺できる。たとえば、HPE Cray XD670 GPU アクセラレーションサーバーの消費電力は、ファンの電力要件が低減され、シリコンのリーク電流が低減されるため、空冷の場合は 10 kW であるが、液冷の場合は 7.5 kW である。液体冷却の詳細については、ホワイトペーパー279、液体冷却を採用する 5 つの理由、およびホワイトペーパー265、データセンターおよびエッジアプリケーション向けの液体冷却テクノロジを参照されたい。

液体は単位体積あたりの熱を吸収する能力がはるかに大きいため、液体冷却技術は空冷よりも効率的に熱を除去できることに注意されたい。ただし、流体の流れが停止すると、チップ温度は空気の場合よりもはるかに速く上昇し、シャットダウンにつながる時間も短くなる。ポンプに UPS を配置することは、この問題の解決に有効である。

標準化された設計手法がなく設置場所に制約があるため、液冷の改修は複雑となる

従来の冷水システムと比較して、ダイレクトチップ液冷サーバーには、水温、流量、化学的性質に関してより厳しい要件がある。これは、チラーシステムからチップのコールドプレートを通じて水を直接流すことができないことを意味する [17]。データセンターを液体冷却に改修する際に水質が課題の一部となることは確かであるが、最大の問題は、この規模（つまり、数百ラック）の AI 負荷に対する標準化された設計手法が存在しないことである。冷却水分配ユニット（CDU）にはいくつかの取り付けオプションと位置があるという事実を考えてみていただきたい [18]。部屋の周囲や列の端に床設置にすることも、各サーバーラックにラック設置にすることもできる。配管をラックに分配するにはいくつかの方法があり、冷却システム機器の設置場所には多数候補があり、温度制御には複数のアプローチがある。液体冷却システムのコンポーネントを視覚化するために、図 3 にさまざまな水ループと CDU を示す。

実稼働するデータセンターにとって、液体冷却のための改修も大規模工事が必要になり、床面積が限られていたり水配管を通すのに上げ床の高さが不十分であるなど物理的制約に遭遇する可能性がある。サーバーの 100%がダイレクトチップ水冷式であっても、ネットワークスイッチなどの他の機器を冷却するための補助的な空冷機能や、水冷サーバーからの熱伝導機能が依然として必要とされる。つまり、設計の組合せが多く、解析が十分になされておらず参考にできる大規模な液冷導入例がほとんどないため、改修には課題が多い。一部のデータセンターには冷水システムがないため、改修がさらに困難になることに注意されたい。

[17] 未処理水をサーバーのコールドプレートに流すと、腐食・生物増殖・汚れが生じるおそれがある。

[18] CDU は、冷水ループをサーバーに供給する「きれいな」水のループから物理的に分離する。

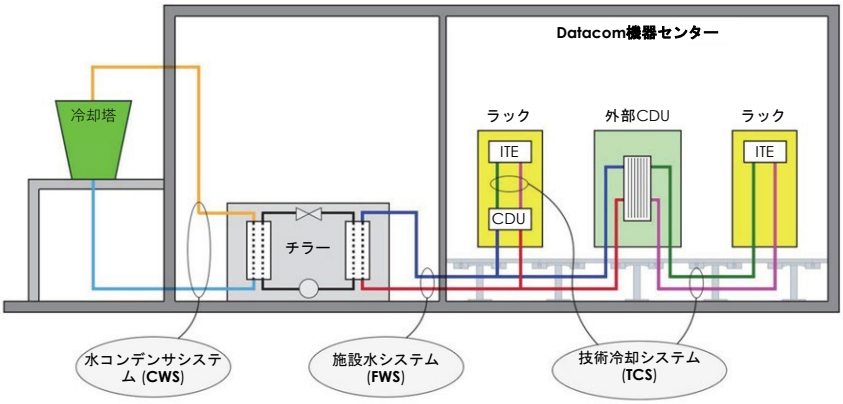


図 3

データセンターにおける CDU を使用した液体冷却

出典： ASHRAE, Water-Cooled Servers:一般的な前提条件、10 ページ

ガイダンス：データセンター運営者には、液体冷却を導入する前に、提案されている液冷負荷と施設の既存の条件の設計評価を実施することを推奨する。可能な設計を評価し、予期せぬ建築上の制約によりコストへの影響が出るのを避けるには、専門家のレビューが不可欠である。たとえば、パイプが上げ床の下の空気の流れを妨げたり、電源トレイやケーブルトレイに干渉したりする可能性がある。詳細については、ホワイトペーパー133、水冷 AI ワークロードを統合するためのデータセンター設計の実践を参照のこと。

TDP が将来どうなるか不明であるため、冷却設計が陳腐化するリスクが高まる

AI テクノロジーは非常に速いペースで進化しているため、次世代 GPU では TDP が高くなり、冷却要件も増加する可能性がある。たとえば、8 つの GPU を備えた現在のサーバーが、次世代では 16 の GPU を備える可能性がある。その結果、今日の負荷に合わせて設計されたデータセンターの冷却配分が、明日の負荷をサポートするには不十分になる可能性がある。

ガイダンス：空冷と液冷に対応し、必要に応じてスケールアップし、さまざまな世代のアクセラレーターに対応できるように冷却システムを設計することが推奨される。たとえば、現在空冷に高温チラーを使用していれば、明日はより高温の液体冷却に簡単に切り替えることができる。もう 1 つの推奨される方法は、将来の CDU 用にタップオフを備えた冷水配管システムを設計することである。これにより、補助空冷用のリアドア熱交換器と組み合わせた 100%ダイレクトチップ液冷負荷への移行が可能になる。

経験不足のため、設置・運用・メンテナンスが複雑になる

データセンター運営者は、空冷システムが使用されてきたため、空冷システムに精通している

データセンター運営者は、何十年も使用してきた空冷システムには詳しいが、ほとんどの運営者にとって液体冷却は初めてである。液体冷却では、コールドプレート、マニホールド、ブラインドメイトバルブなどの部品が使用される。これらの部品は、運営者にとって馴染みのない追加の設置・操作・保守手順を伴う。たとえば、ダイレクトチップコールドプレートサーバーの細い配管は汚れの影響を受けやすいため、運営者は水の化学的性質を管理するために、新しい操作やメンテナンス手順を学ぶ必要が生じる可能性がある。別の例としては、水をサーバーに送り込む際に水漏れのリスクが生じる。

ガイダンス：液体冷却設計は、設置・運用・保守の作業を最小限に抑える上で重要な役割を果たす。水冷サーバーの保守に慣れていないデータセンター運営者は、専門家に設計の徹底的なレビューを依頼し、日常の操作に関する詳細な標準操作手順（SOP）と手順方法（MOP）を作成してもらうことを推奨する。これにより、特に漏れに関する故障や人的エラーが最小限に抑えられる。

液体冷却により IT ラック内の漏れのリスクが増加する

ダイレクトチップ技術では、サーバー内のコールドプレートで水（脱イオン水、アルコールベースの溶液など）を使用する。水漏れは安全性と信頼性に関する懸念事項であり、設計と調達の段階で考慮する必要がある。

ガイダンス：信頼できるプロバイダーと協力して、システムが厳格な圧力テストに合格していることを確認して、漏れのリスクを最小限に抑えることが推奨される。さらに、サーバーおよびラックレベルでの漏れ検出は、漏れがさらに深刻になる前に発見するのに役立つ。従来の CDU ポンプシステムの代わりに、革新的な漏れ防止システム（LPS）を備えた CDU を検討してみる。LPS は水ループをわずかな真空（負圧）に維持し、IT 機器内の漏れのリスクを排除する。浸漬液体冷却では誘電性流体を使用するため、サーバー内での水漏れのリスクも解消される。これらは、AI サーバーまたは統合ベンダーのオプションとなっている場合がある。最後に、漏洩が発生した場合に対処するための緊急作業手順（EOP）を作成しなければならない。

液体冷却を持続的に動作させるための流体の選択肢は限定的である

従来の空冷式 IT と比較して、液体冷却にはエネルギー消費と水の使用量の両方を削減するという点で、環境の持続可能性に関する利点がある。これは、サーバーファンのほとんどまたはすべてが不必要となり、水温が上昇することでエコノマイザーの稼働時間が増えるので、IT サーバーと冷却システムの両方のエネルギー効率が向上するからである [19]。ただし一部の水冷システムでは、環境に有害な人工化学物質が使用されている。たとえば、フルオロカーボン流体は、その伝熱性能により、液浸冷却技

[19] 外気温が水温より低い場合に節約となる。ダイレクトチップサーバーからの戻り水の温度は、従来の冷水の戻り温度よりもはるかに高くなる。このように高温になると、水が自然冷却される時間が年間で長くなる。

術 [20] の誘電性流体として広く使用されている。残念ながら一部のフロン類の地球温暖化係数（GWP）は 8,000 程度である。比較のために、冷蔵庫でよく用いられる HFC-143a 冷媒の GWP は 1,430 である。さらに、社会的圧力により、製造業者は冷媒などの製品から PFAS（パーフルオロおよびポリフルオロアルキル物質）を（環境への影響を軽減するために）排除し、GWP の低い冷媒に移行している。ほとんどのデータセンター運営者にとって持続可能性は最優先事項となっており、選択肢は少なくなってきている。

ガイダンス：フルオロカーボン流体の使用は避けることが望ましい。以前これらの流体は、ダイレクトチップ冷却システムと浸漬冷却システムの両方で使用されていた。現在、チイレクトチップ冷却には水を使用するため、問題とならない。液浸冷却を導入する場合は、GWP がゼロの油ベースの誘電性流体を使用することが推奨される（2 相機能性流体とは異なる）。ただし、油ベースの誘電性流体は水を使用したダイレクトチップ冷却の熱伝達ほど効果的ではないため、今日ではダイレクトチップ冷却が好ましい液体冷却構造となっている。ベンダーがフルオロカーボン流体に代わる持続可能な代替誘電体を開発中であることに留意されたい。これにより、液浸冷却の熱除去効率が大幅に向上し、おそらく冷却構造の変化が促進される可能性がある。詳細については、ホワイトペーパー291、*IT 機器の浸漬液体冷却用誘電性流体の比較*を参照のこと。

ラック

前のセクションで説明した電源と冷却の課題の一部は、IT ラック（IT キャビネットまたはエンクロージャなど）にも最終的に影響する。ラックシステムには、AI ワークロードに関連して次の 4 つの課題がある：

- 標準的な幅のラックには、必要な電源および冷却装置を設置するスペースが足りない
- 標準的な奥行のラックには、奥行がある AI サーバーとケーブル配線用のスペースが足りない
- 標準的な高さのラックには、必要な数のサーバーを収容するスペースが足りない
- 標準的なラックには、AI 機器を収容するのに十分な耐荷重能力が足りない

標準的な幅のラックには、必要な電源および冷却装置を設置するスペースが足りない

AI サーバーは奥行きが深くなっているため、ラック PDU や液体冷却マニホールドを取り付けるためのラック背面のスペースが狭くなっている。サーバーの電力密度が増加し続けると、標準幅のラック（つまり、600 mm / 24 インチ）の背面に、必要な

[20] 浸漬冷却では、チップ全体またはサーバーも誘電性流体に浸漬させる。

電力の配線と冷媒の配管を収容することが、不可能ではないにしても非常に困難になる。さらに狭いラックでは、電源ケーブルやネットワークケーブルにより、ラックの後方で排気の流れが停滞するおそれがある。

ガイダンス：ラック PDU を収容するには、少なくとも 750 mm（29.5 インチ）幅のラックを使用し、液体冷却の場合は液冷サーバー用のマニホールドを使用することが推奨される。これらのラックは、標準の 600 mm ラックのように幅 600 mm の上げ床タイルと揃わなくなるが、これは意味のある制約ではなくなった。これは、空冷 AI サーバーには高流量の空気流が必要であり、上げ床は通常、空気の分配には使用されず、配管やケーブル配線に使用されるためである。

標準的な奥行のラックには、奥行がある AI サーバーとケーブル配線用のスペースが足りない

AI ワークロード用に最適化されたサーバーは、一部の標準ラックの取り付け最大奥行を超える奥行に達する可能性がある。奥行きのあるサーバーを浅いラックに取り付けることができる場合でも、通気を十分に確保しながらネットワークケーブルを収容するには、背面に十分なスペースが必要となる。

ガイダンス：IT ラックには、さまざまな IT 機器の奥行きに対応できる調整可能な取り付けレールが付いているが、取り付け可能な最大の奥行きは異なる。ラックの奥行きは少なくとも 1,200 mm（47.2 in）、取り付け可能な最大奥行きは 1,000 mm（40 in）を超えることを推奨する。

標準的な高さのラックには、必要な数のサーバーを収容するスペースが足りない

AI サーバーの高さにもよるが、一般的な 42U の高さのラックでは、すべてのサーバー、スイッチ、その他の機器を収容するには低すぎることがある。たとえば、64 ポートのネットワークスイッチは、ラックに 8 台のサーバーがあり、それぞれに 8 つの GPU が搭載されていることになる。この密度では、サーバーの高さを 5U と仮定すると、サーバーだけで 40U を占めるため、他のデバイスを収容できる残りのスペースは 2U だけになる。

ガイダンス：データセンターの出入口の高さが十分にあることを前提として、AI トレーニングクラスターは 48U ラック以上に設置することを推奨する。1U は 44.45 mm（1.75 in）である [21]。

[21] たとえば 48U では、機器に利用可能な内部垂直スペースが 2.13 m（84 in）ある。

標準的なラックには、AI 機器を収容するのに十分な耐荷重能力が足りない

重い AI サーバーを使用すると、高密度ラックの重量は 900 kg（2000 lb）を超えることがある。これにより、静的耐荷重と動的（転動）耐荷重の両方の観点から、IT ラックとフリーアクセスフロアに大きな負荷がかかる。これらの重量に対応していないラックでは、フレーム、水平調整脚、キャスターが変形するおそれがある。さらに、上げ床ではこれらの重いラックをサポートできない場合がある。

ガイダンス：IT ラックの耐荷重は、静的および動的荷重として指定される。静的耐荷重とは、ラックが静止しているときに支えることができる重量を指す。動的耐荷重とは、ラックが動いているときにサポートできる重量を指す。静的耐荷重が 1,800 kg（3,968 lb）を超え、動的耐荷重が 1,200 kg（2,646 lb）を超えるラックを指定することを推奨する。これらのラック耐荷重は、独立したサードパーティの検証を受けることが望ましい [22]。現時点の AI 導入が小規模で、まだこれらの耐荷重を必要としていない場合でも、ラックの耐用年数は IT 機器よりも長くなることが多い。おそらく次世代の AI 導入では、これらのラックに関する推奨事項の一部またはすべてが必要になると考えられる。最後に、場合によっては、IT ラックが現場以外で事前構成されてからデータセンターに輸送されることがある。これらのラックは、輸送中に発生する動的な力に耐えることができなければならず、関連する梱包もラックとラックがサポートする高価な IT 機器を保護する必要がある。

データセンターの床、特に上げ床は、AI クラスターの重量に耐えられるかどうかを評価する必要がある。これは、データセンター内で重いラックを移動する際のフリーアクセスフロアの動的耐荷重に関して特に重要である。

ソフトウェアツール

データセンターの設計と運用をサポートする物理インフラソフトウェアツールには、DCIM、EPMS、BMS、デジタル電気設計ツールなどがある。従来の空冷 IT と並行して、高出力密度の液冷 IT のクラスターが存在するということは、特定のソフトウェア機能がより重要になることを意味する。一部の AI トレーニングワークロードでは高可用性が必要ない場合でも、設計と監視が不十分であると、ビジネスクリティカルである可能性が高い隣接するラックやテナントにダウンタイムのリスクが生じるおそれがある。次の 2 つの課題は、高密度 AI トレーニングワークロードのコンテキストにおいてより関連性が高く重要な管理ソフトウェア機能に焦点を当てたものである：

- 極端に高い電力密度と AI クラスターの需要が設計の不確実性につながる
- エラーに対するマージンが減少すると、動的な環境での運用リスクが増加する

[22] Underwriters Laboratory（UL）および International Safe Transit Association（ISTA）の利用を推奨する。詳細については、ホワイトペーパー201、「IT ラックの選び方」を参照されたい。

AI クラスターの極端に高い電力密度と電力需要が設計の不確実性につながる

新しい AI クラスターに対応するために既存のサイトを改修する前に、十分な電力と冷却能力、およびその能力を新しい負荷に分配するために必要なインフラがあることを確認するためのフィージビリティスタディが必要である。ラックの電力密度が 10 kW をはるかに下回り、大容量電力と冷却能力が過剰な一般的なケースでは、標準 IT の追加は比較的簡単で、それほど詳しい調査や検証は必要とされない。ポイントインタイムの電力および冷却の測定を、使い慣れた一般的な配電コンポーネントや既存の冷却ユニットと併せて使用できる。この自動化の程度が低く大雑把な改造設計アプローチでは、大規模な高密度 AI トレーニングクラスターには不十分である。数百キロワットの電力を消費する AI クラスターでは、設計ミスを犯した場合（つまり、実際のピークから平均までの消費電力がわからない、どの回路にどのような負荷がかかっているかがわからない、など）、影響がより大きくなる。設計に関して不明点や不確実性があってはならない。また、AI クラスターの設計は非常に特殊であるため（標準ではない高アンペア数の rPDU/バスウェイ、液体冷却の使用など）、クラスターが起動時にどのように動作するかについて不確実性が大きくなる。

ガイダンス：現在の電力容量とその傾向を大容量電力レベルと IT スペース内の配電レベルの両方で正確に把握するには、EPMS と DCIM を使用することを推奨する。これらのツールは、長期間にわたる実際のピーク電力消費量を示してくれる。不意にブレーカーが落ちることがないようにするには、これを理解することが重要である。この容量評価は、AI 負荷を受け持つ能力を判断するのに役立つ。これは、必要な電力メーターが設置されていることを前提としていることに注意されたい。次に、変更の前に、容量分析・保護調整・アークフラッシュ検討・短絡およびデバイスの評価などの安全性および技術的な検討を実施することを推奨する [23]。電気設計（別名、電力システムエンジニアリング）ソフトウェアツールを使用すると、データ収集や計算量が簡素化される。

評価後、AI クラスターを追加するには電力ネットワークの変更が必要になる可能性がある。この場合、電気設計ソフトウェアツールを使用すると、IT 分野で電気ネットワークの作業および保守を行う際に、最適な電気機器の選択、電気的故障の防止、効果的な手順の開発、適切な安全プロトコルの実装に必要な適切なデータを確実に入手できる。

デジタル化された単線図（iSLD）[24]のある既存のデータセンターでは、上記の評価プロセスが簡素化される可能性があることに注意することが重要である。正確でインテ

[23] つまり、容量および kA 定格・特定の設計への適合性に関するその他の仕様を評価する

[24] 一部のベンダーは、iSLD の作成と保守をサービスとして提供している。

リジェントな iSLD を使用すると、データの収集と計算の実行に必要な時間と専門知識が大幅に削減される。iSLD は、専用のソフトウェアで保存および管理されたより高度な単線図であり、高度な機能を持ち、デバイスの特性と運用動作を考慮している。物理的な電気ネットワークのデジタルツインを作成する。実質的に、この 1 つのソフトウェア プラットフォームを使用して、電気回路の設計、SLD の作成と保守、すべての技術検討と安全性評価を実施することができる。

エラーに対するマージンが減少すると、動的な環境での運用リスクが増加する

第 1 の課題のガイダンスに従って最適なデータセンター設計を実装したとすれば、「初日」の運用はスムーズに実行されるはずである。しかし他のタイプの施設と比較すると、データセンターは、IT 機器の移動、追加、変更が頻繁に行われる動的な環境である。大規模な AI クラスターの追加では容量の安全マージンが小さくなることが多く、IT 空間内の負荷が時間の経過とともに変化し、ブレーカーが落ちたり、ホットスポットが形成されたり、リソースが孤立したりするリスクが増加する。リスク増加の根本的な理由は、前述したように、ラック密度が高く、AI クラスターのピーク対平均比が低い（1 に近い）ことである。エラーの許容範囲が狭くなるということは、ダウンタイムを防止し、データセンターの耐用年数全体にわたって利用可能なリソースを効率的に使用するためには、運営者がますます状況を認識する必要があることを意味する。

ガイダンス：上記の課題を最小限に抑えるか、防ぐためには、IT 空間全体（ラック内の機器や VM を含む）のデジタルツインを作成することを推奨する。このレイアウトを、長期間にわたって維持する必要がある。DCIM のプランニングおよびモデリング機能により、ルールベースのツールを使用して効果的な IT 空間のフロアレイアウト運用が可能である。IT 負荷をデジタル的に追加または移動することで、それらをサポートするのに十分な電力容量・冷却能力・床耐荷重があることを検証できる。DCIM は、IT スペースのデジタルツインを作成し、リソースに対するすべての機器の依存関係を記録する。これにより、リソースの滞留を回避し、ダウンタイムにつながる可能性のある人的エラーを最小限に抑えるための決定を行うことができる。EPMS と DCIM を併用すれば、すべての PDU、UPS、rPDU などの電力容量を監視し、電力しきい値の超過に関する早期警告を受け取り、ダウンタイムを回避できる。DCIM ソフトウェアは、電源・冷却・冗長レベル要件・利用可能な U スペース・ネットワークポート・耐荷重に基づいて、新しい機器を配置する最適な場所について助言する。これは、AI 以外の機器や AI 推論サーバーにさらに当てはまる。推論負荷とは異なり、AI トレーニング負荷には、事前に設計された構成が必要である。この構成は、たとえ変更されたとしてもめったに変更されることはない。

多くの DCIM プランニングおよびモデリングソフトウェアツールには、機器の物理的なレイアウトと熱負荷を考慮して適切な空気の流れを確保するのに用いる数値流体

力学（CFD）ツールが含まれる。DCIM を使用すると、インフラと負荷の最適な配置と構成を通じて利用効率の低かった冷却能力をなくし、冷却能力を最適化できる。AI 負荷の移動・追加・変更に関しては、ユーザーの需要（クエリなど）を満たすためにより多くのサーバーが追加されるため、CFD は AI 推論負荷により多く適用される。場合によっては、AI トレーニングまたは推論クラスターが独自の電源セグメントと冷却アーキテクチャに分離されることがあるのに注意されたい。このような場合、非 AI 負荷は AI クラスターの影響を受けにくくなる。ただし、どちらの場合でも、これらのスペースのデジタルツインを確立することは有益である。

AI を支える物理インフラの将来展望

この指針ではここまで、現在利用可能な技術と設計アプローチに焦点を当ててきた。このセクションでは、提示された課題の解決にさらに役立つと考えられるいくつかの将来技術と設計アプローチについて簡単に説明する。

- **AI に最適化された標準 rPDU** – フォームファクターは、より少ない電力密度のサーバーに対応し、孤立したコンセントが少なくなるように変更されるであろう。不要なコンセントを排除することで、各ラックにさらに多くの rPDU を搭載したり、単一の高容量 rPDU（240 V、86 kW で定格最大 150 アンペア）を搭載することが可能になる。これらの rPDU は、スイッチなどの低密度機器用のコンセントとしても機能する。
- **技術／IT スペースにおける中電圧から 415/240 V への変圧器** – 中電圧（たとえば 13 kV）で電力を配電すると、銅線が削減され、必要な導体が減り、設置時間が短縮される。たとえば、IT 配電では 2 MW の変圧器を使用して 3,000 A の母線路に 415/240 V で電力を供給し、これにより AI クラスター全体または 2 MW を超えるクラスターの一部に電力を供給する。この配電アーキテクチャにより、IT 配電の上流にある従来の 13 kV から 480/277 V への変圧器や開閉装置も不要になる。これにより、480 V 配電装置のサプライチェーンの制約も緩和される可能性がある。
- **半導体変圧器** – これは本質的にパワーエレクトロニクス変換器である。半導体コンポーネントを使用して、一次電圧を二次電圧に変換する。これらは、一次側と二次側を電気的に絶縁する中周波トランス（MFT）を使用する。C:\Users\SESA395242\AppData\Local\Temp\3b549b9a-15f7-4943-ba55-f7cd9b5b3e89_WP110_V1.1_日.docx.zip.e89\www.se.com\jp 従来の変圧器は重く、交流（AC）のみで動作するが、半導体変圧器は小型・軽量で、AC 電圧と DC 電圧の間で変換を行う。
- **ソリッドステートサーキットブレーカー** – これらのサーキットブレーカーは、電流の流れをオンまたはオフにするために半導体を使用する。これは、故障個所への電流の流れを遮断する場合に特に重要である。ただし、サーキットブレーカーとみなされるには、ガルバニック絶縁を提供することが必要で、半導体と直列に機械的スイッチを使用しなければならない。ソリッドステートブレーカーを使用すると、より高速な動作が可能になり、故障電流をより厳密に制御

できるようになる。これは、高密度 AI ラックでのアークフラッシュエネルギーを削減するのに非常に有益である。

- **持続可能な誘電性流体** – これは、熱伝達効率を高め、より高いチップ TDP が実現できれば、ダイレクトチップ冷却の最新の選択肢として水を代替する可能性がある。
- **奥行きが極めて深い IT ラック** – 奥行きの深いアクセラレーターベースのサーバーが導入されると、ラックの奥行きがさらに深ければ、サーバーだけでなく、ネットワークケーブル配線、水道配管、およびラック PDU も収容できるようになる。
- **グリッドとのインタラクション／最適化の強化** – ワークロードのスケジューリングをマイクログリッドの状態に基づいて電力会社とともに行うことで、グリッドのバランスをとり電力を節約するのに貢献する。ワークロード管理の例として、負荷を別の冗長ゾーンに移行したり、UPS をバッテリーで動作させたりすることが挙げられる。

結論

AI の急速な成長と応用により、データセンターの設計と運用が変化している。AI ワークロードは、2028 年までにデータセンターの総エネルギーの 15%～20%を占めると私たちは推定している。推論ワークロードは、トレーニングクラスターよりもはるかに多くの電力を消費すると予想されるが、さまざまなラック密度で動作する。一方で AI トレーニングワークロードは、ラックあたり 20～100 kW 以上の非常に高い密度で一貫して動作する。ネットワークの需要とコストにより、これらのトレーニングラックはクラスター化される。これらの極めて電力密度が高いクラスターは、データセンターの電力・冷却・ラック・ソフトウェア管理設計に根本的な課題を課す。このホワイトペーパーでは、課題にどのように対処するかについて指針を提供した。それらを以下に要約する：

電力：120/208 V 配電（NAM 内）の使用はもはや十分ではなく、代わりに 240/415 V 配電を使用して高密度ラック内の回路数を制限することが推奨される。より高い電圧であっても、標準の 60/63 アンペアラック PDU で十分な容量を提供することは依然として課題である。たとえば、液体冷却ラックは 69/87 kW を提供する 2 つの rPDU に制限される。作業員の安全を確保するために、アークフラッシュのリスク評価と負荷分析を実施し、暴露温度に基づいて適切なコネクター、コンセント、rPDU が使用されているのを確認することを推奨する。上流の配電ブロックサイズは、AI クラスターの単一列をサポートするのに十分な大きさでなければならない。

空調・冷却：空冷は近い将来もまだ存在するが、AI クラスターを備えたデータセンターにとって推奨されるまたは必要となるソリューションとして、空冷から液冷への移行が予測される。空冷と比較して液体冷却には、プロセッサの信頼性とパフォーマンスの向上、ラック密度の向上によるスペースの節約、配管内の水による熱慣性の増大、エネルギー効率の向上、電力使用率の向上（IT に投入される電力が増加）、水の使用量の削減など、多くの利点がある。データセンター運営者は、私たちが提案する指針をもとに、空冷から液冷への移行を成功させて AI ワークロードをサポートすることができる。

ラック：AI クラスターでは、サーバーの奥行きが増し、電力需要が増大し、冷却がより複雑になる。そのため、より大きな寸法と耐荷重のラック、特に幅 750 mm（29.5 in）、奥行き 1,200 mm（47.2 in）、高さ 48U、取り付け奥行き 1,000 mm（40 in）、静的耐荷重 1,800 kg（3,968 lb）超、動的耐荷重 1,200 kg（2,646 lb）超のラックを使用することが推奨される。

ソフトウェア管理：AI クラスターの管理においては、DCIM、EPMS、BMS、デジタル電気設計ツールなどのソフトウェアツールが重要になる。これらにより、複雑な電気ネットワークで予期せぬ動作が発生するリスクを軽減できる。また、データセンターのデジタルツインを提供して、制約となる電力および冷却リソースを特定し、レイアウトの決定に情報を提供できる。

著者について

Victor Avelar は、シュナイダーエレクトリックのエネルギー管理研究センターの上級研究アナリストである。データセンターの設計とオペレーションズリサーチを担当し、リスク評価と設計手法についてクライアントのコンサルティングを行い、データセンター環境の可用性と効率性を最適化する方法を検証している。レンセラー工科大学で機械工学を専攻し学士号を取得しているほか、バブソンカレッジで MBA を取得した。AFCOM の会員である。

Patrick Donovan は、シュナイダーエレクトリックのエネルギー管理研究センターの上級研究アナリストである。シュナイダーエレクトリックのセキュアパワービジネスユニットの重要な電源および冷却システムの開発とサポートに 27 年以上の経験を持ち、電源保護・効率・可用性ソリューションの分野で受賞歴を持つ。多数のホワイトペーパー、業界記事、テクノロジー評価の著者である Donovan 氏は、データセンターの物理インフラの技術と市場に関する研究を行っており、データセンター施設の計画・設計・運用のベストプラクティスに関して、ガイダンスとアドバイスを提供している。

Paul Lin は、シュナイダーエレクトリックのエネルギー管理研究センターの研究ディレクター兼エジソンエキスパートである。データセンターの設計と運用調査を担当しており、データセンター環境の可用性と持続可能性を最適化するためのリスク評価と設計実践についてクライアントにコンサルティングを行っている。Lin 氏は著名な専門家であり、データセンター業界のイベントで頻繁に講演者やパネリストを務めている。シュナイダーエレクトリックに入社する前、LG Electronics で R&D プロジェクトリーダーとして数年間勤務した。登録プロフェショナルエンジニアでもあり、10 件以上の特許を取得している。Lin 氏は吉林大学で機械工学の学士号と理学修士号を取得している。 また INSEAD から Transforming Schneider Leadership Program の証明書も取得している。

Wendy Torell はシュナイダーエレクトリックのデータセンターサイエンスセンターの上級研究アナリストである。この職務において、彼女はデータセンターの設計と運用におけるベストプラクティスを調査し、ホワイトペーパーと記事を発表し、クライアントがデータセンター環境の可用性・効率・コストを最適化できるように支援するトレードオフツールを開発している。また、データセンターのパフォーマンス目標を達成できるように、Availability Science のアプローチと設計実践についてクライアントにコンサルティングを行っている。ニューヨーク州スケネクタディのユニオン大学で機械工学の学士号を取得し、ロードアイラン

ド大学で MBA を取得した。ASQ（米国品質協会）の信頼性技術管理士の資格を有している。

Maria A. Torres Arango は、シュナイダーエレクトリックのエネルギー管理研究センターの研究アナリストである。この職務において、彼女は意思決定に情報を提供するための技術的な戦略的トピックを調査しており、現在はエネルギー貯蔵システムと持続可能性に重点を置いている。コロンビアのポンティフィシアボリバリアナ大学で航空工学の学士号を取得し、ウェストバージニア大学で航空宇宙工学の修士号、および材料工学の博士号を取得している。

Resources

High-Efficiency AC Power Distribution for Data Centers
White Paper 128

Data Center Design Practices for Integrating Liquid-cooled AI Workloads
White Paper 133

Arc Flash Considerations for Data Center IT Space
White Paper 194

Benefits of Limiting MV Short-Circuit Current in Large Data Centers
White Paper 253

Liquid Cooling Technologies for Data Centers and Edge Applications
White Paper 265

Five Reasons to Adopt Liquid Cooling
White Paper 279

Comparison of Dielectric Fluids for Immersive Liquid Cooing of IT Equipment
White Paper 291

Arc Flash Mitigation
White Paper

Browse all white papers
whitepapers.apc.com

Browse all TradeOff Tools™
tools.apc.com

Note: Internet links can become obsolete over time. The referenced links were available at the time this paper was written but may no longer be available now.