MPLS Japan 2023
超巨大チップを用いた高速AI学習

東京エレクトロン デバイス株式会社

2023年10月26日

中川 隆志

※資料内に掲載の会社名・団体名・商品名・サービス名またはロゴマークは、各社・団体の商標・登録商標・もしくは商号です。

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

1枚のウェハーから取り出せる最大サイズのチップ
果たしてこれがどのように使われているのでしょうか

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

今日お話しすること

- AIの現状
- 巨大チップWSEとCerebras CS-2の紹介
- 巨大チップWSEとCerebras CS-2の今後

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

あなたは誰ですか

- 東京エレクトロンデバイスのエンジニアです
  - 10年ほどData Warehouse、Hadoop、分散KVSなどのBig Dataに携わる
  - 最近は主業務をCerebras Systems社、Nvidia社などのAI製品にチェンジ
  - 昔はWindowsプログラミングもやっていました

- ~~東京エレクトロンデバイスとは？~~ 本日は会社紹介は行いません
  - MPLSの趣旨にそぐわないため

- ですが大事なことを1つだけ申し上げます。東京エレクトロンデバイスは**商社**です

- Cerebras Systems社と**代理店契約**を結び、製品販売に尽力しています

- 本日は**代理店の社員**が、可能な範囲で**中立的な話**をお届けしたいと考えています

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

AIの現状

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

AI で何ができるのか (ほんの一例)

GPT-4 visual input example, Chicken Nugget Map:

User Can you explain this meme?

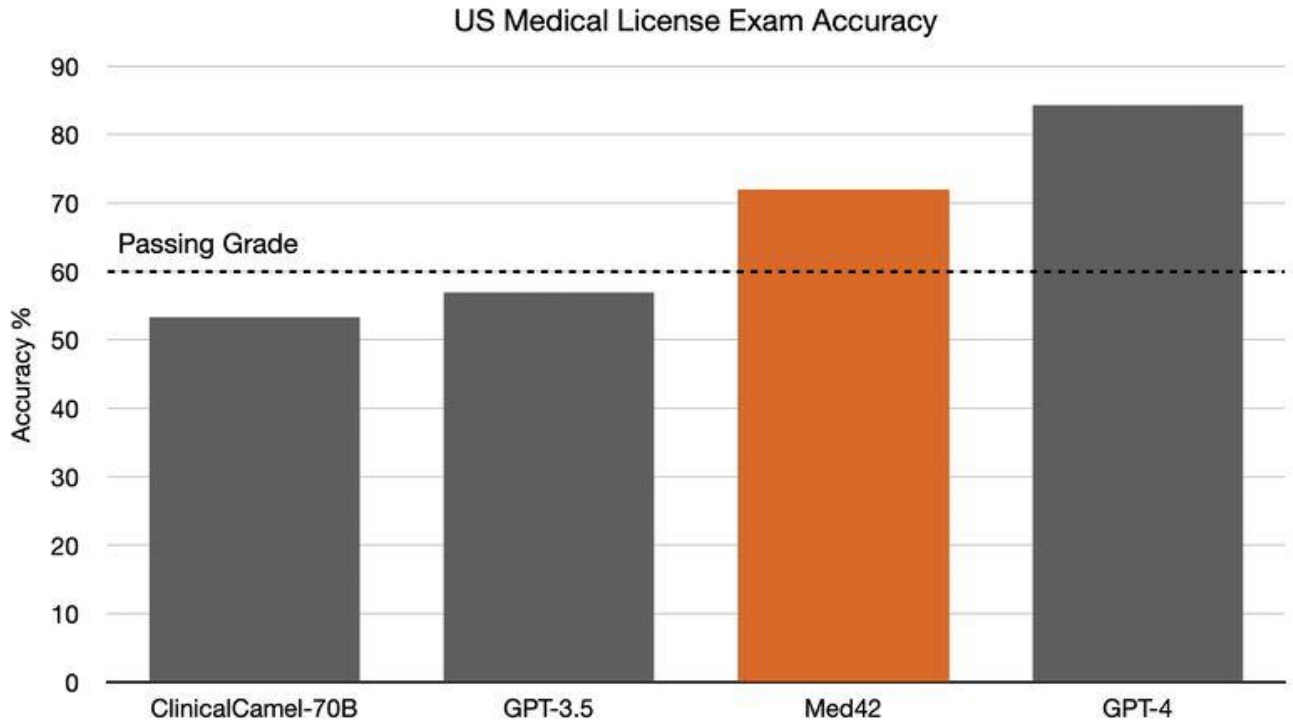
Sometimes I just look at pictur
the earth from space and I ma
how beautiful it all is.

この写真の面白いところは何？

GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.
The text of the meme suggests that the image be
earth from space. However, the image is actua
vaguely resemble a map of the world.
The humor in this meme comes from the unexpe
the image. The text sets up an expectation of a m
image is actually something mundane and silly.

チキンナゲットが地図みたいになっているところ

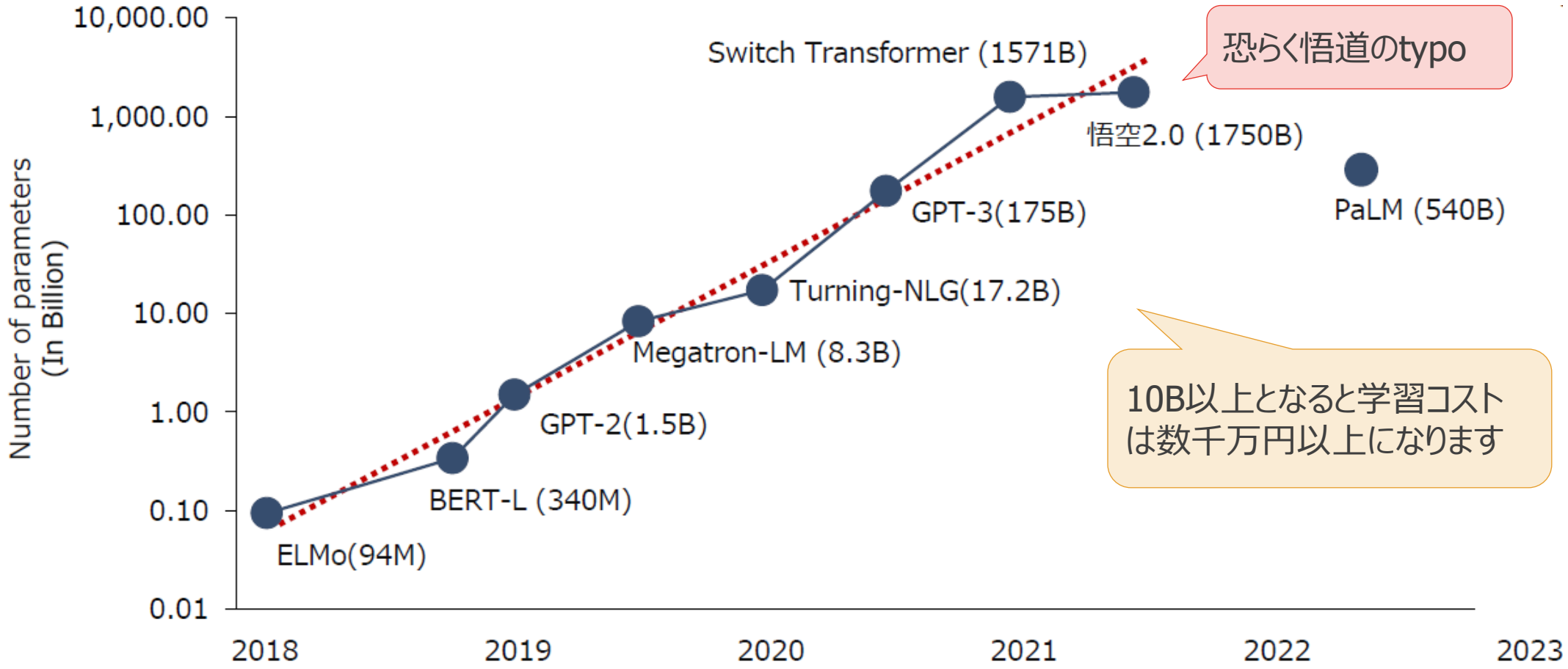
アメリカの医師国家試験に合格することができます
※ Cerebrasの機器を使ってファインチューニングした事例



https://www.linkedin.com/posts/cerebras-systems_we-are-pleased-to-announce-that-our-strategic-activity-7118282684660924416-j7QA/

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

昨今のモデルサイズのトレンド



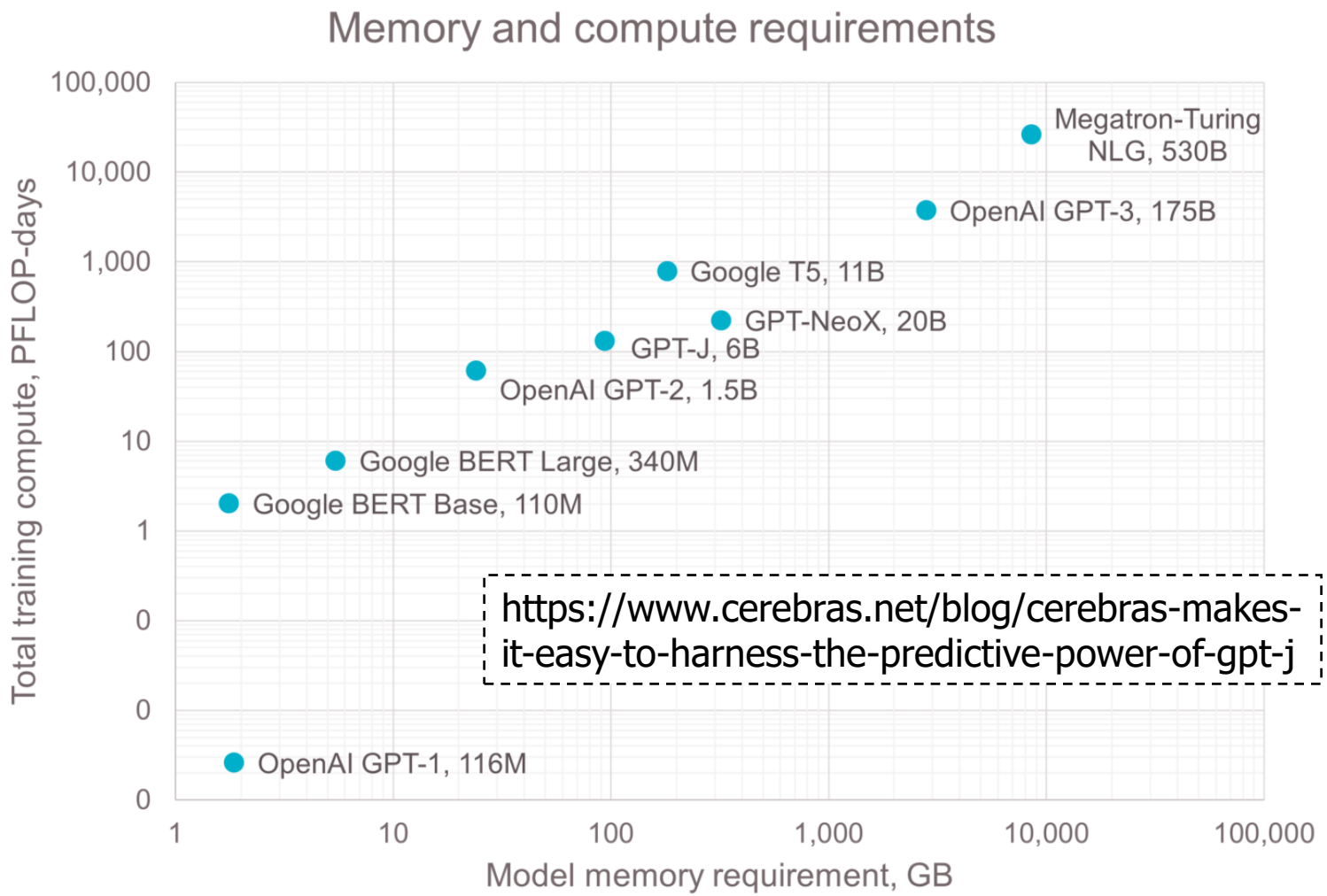
松尾研究室「AIの進化と日本の戦略」より抜粋
https://note.com/api/v2/attachments/download/a29a2e6b5b35b75baf42a8025d68c175

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

モデルサイズと必要なメモリ量

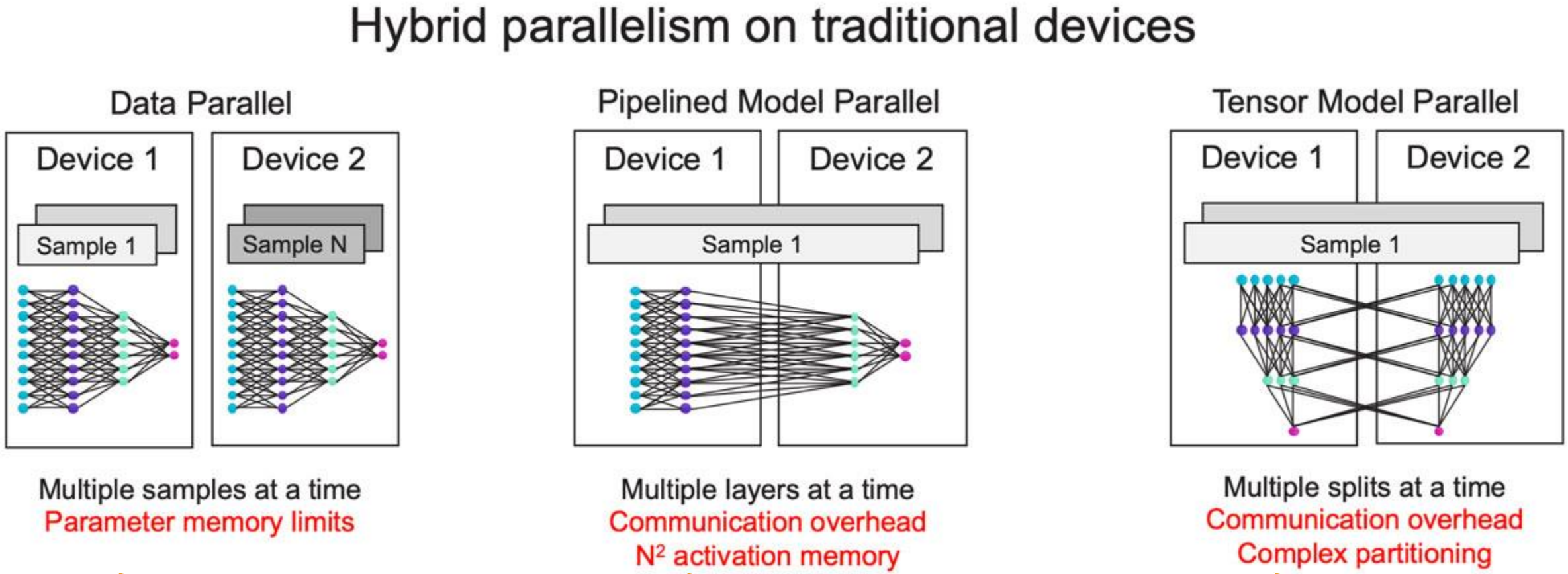
- パラメータ数が増えるにつれて必要なメモリ量は増えます (右図では1パラメータあたり16バイト必要と仮定)
- 一方で最新のGPUでも、**1枚あたり80GB**のメモリしかありません
- そのため通常これらのモデルを学習させるためには、**複数のGPUをまとめて使う**必要があります

※実際にはパフォーマンスの観点でも複数GPUの活用が必須です



Copyright © Tokyo Electron Device LTD. All Rights Reserved.

大きいモデルを複数のデバイスに分割する手法



デバイスごとに違うデータで学習する

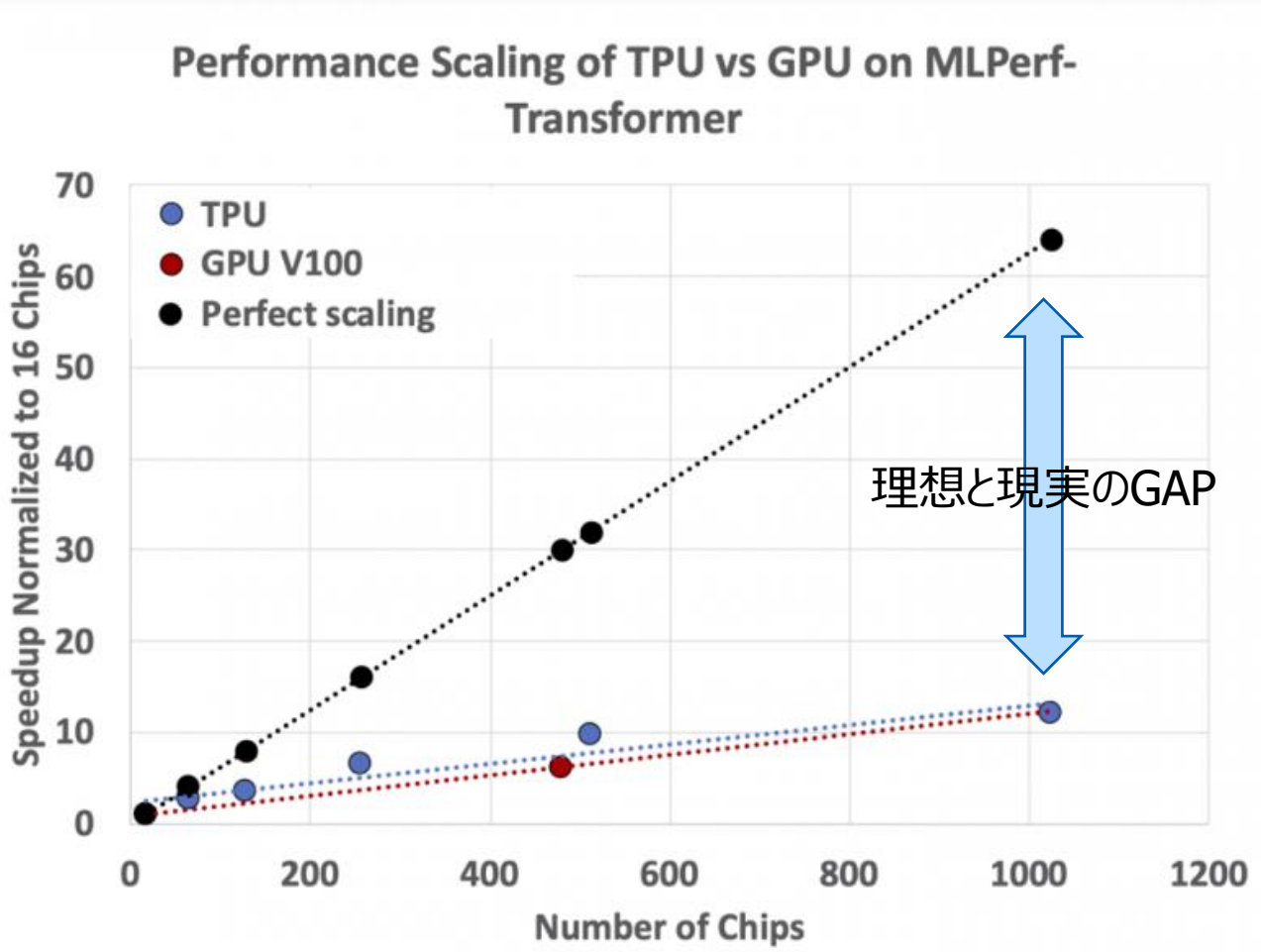
モデルの層 (Layer) を別のデバイスに配置する

1つの層を複数のデバイスに分割する

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

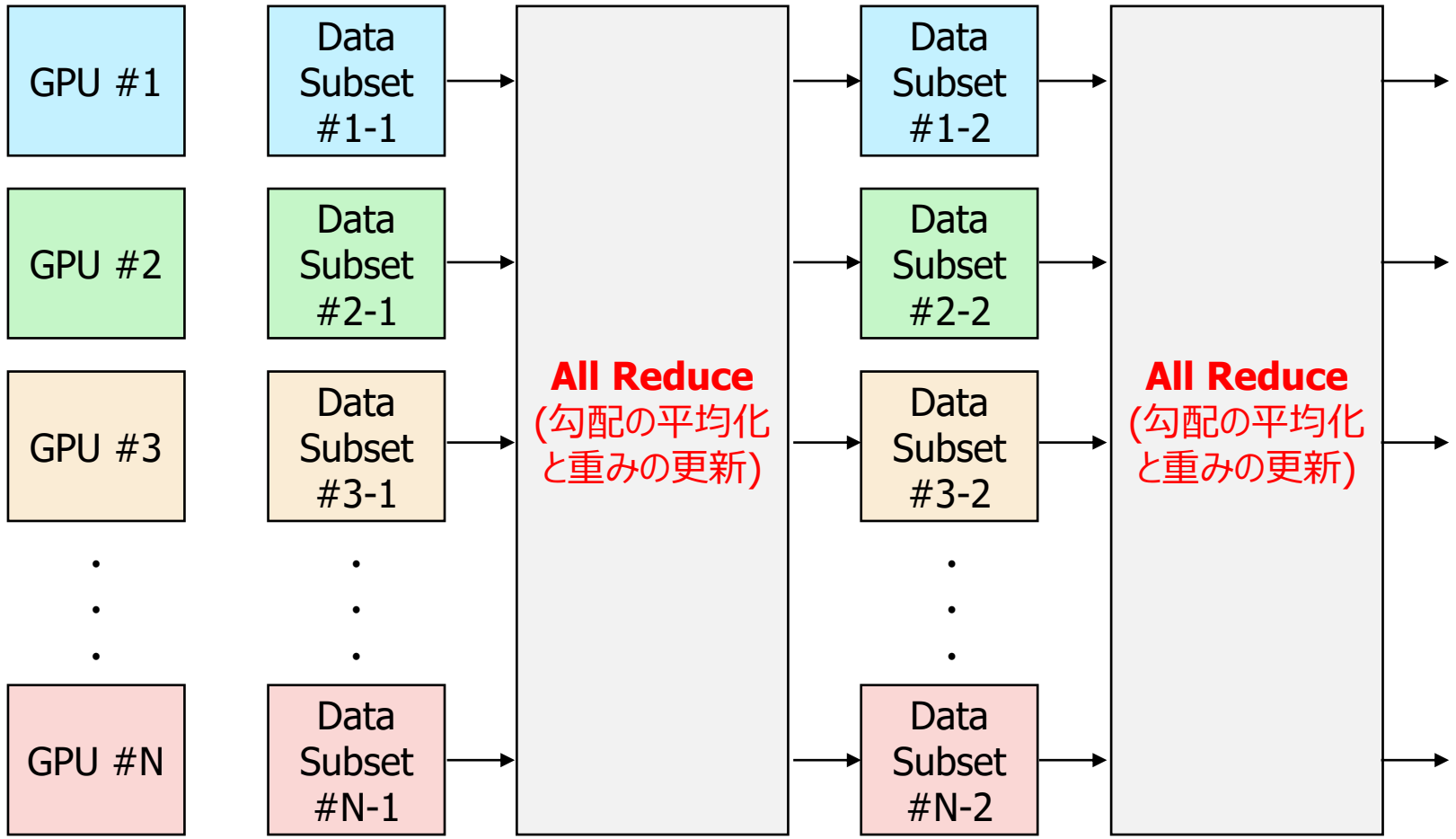
モデル分割時のパフォーマンス

- 右のグラフはMLPerfというベンチマークの結果を基にした、GPU枚数と性能の伸びを表したグラフです
- GPU/TPUを何百枚も使うことで、巨大モデルを学習することが可能ですが枚数を増やすごとに性能の伸びは**鈍化**していきます
- **チップ間 (デバイス間) 通信のボトルネック**がこの原因と言えます
- **GPU/TPUも万能ではない**ことが分かります



Copyright © Tokyo Electron Device LTD. All Rights Reserved.

Data Parallel方式の課題



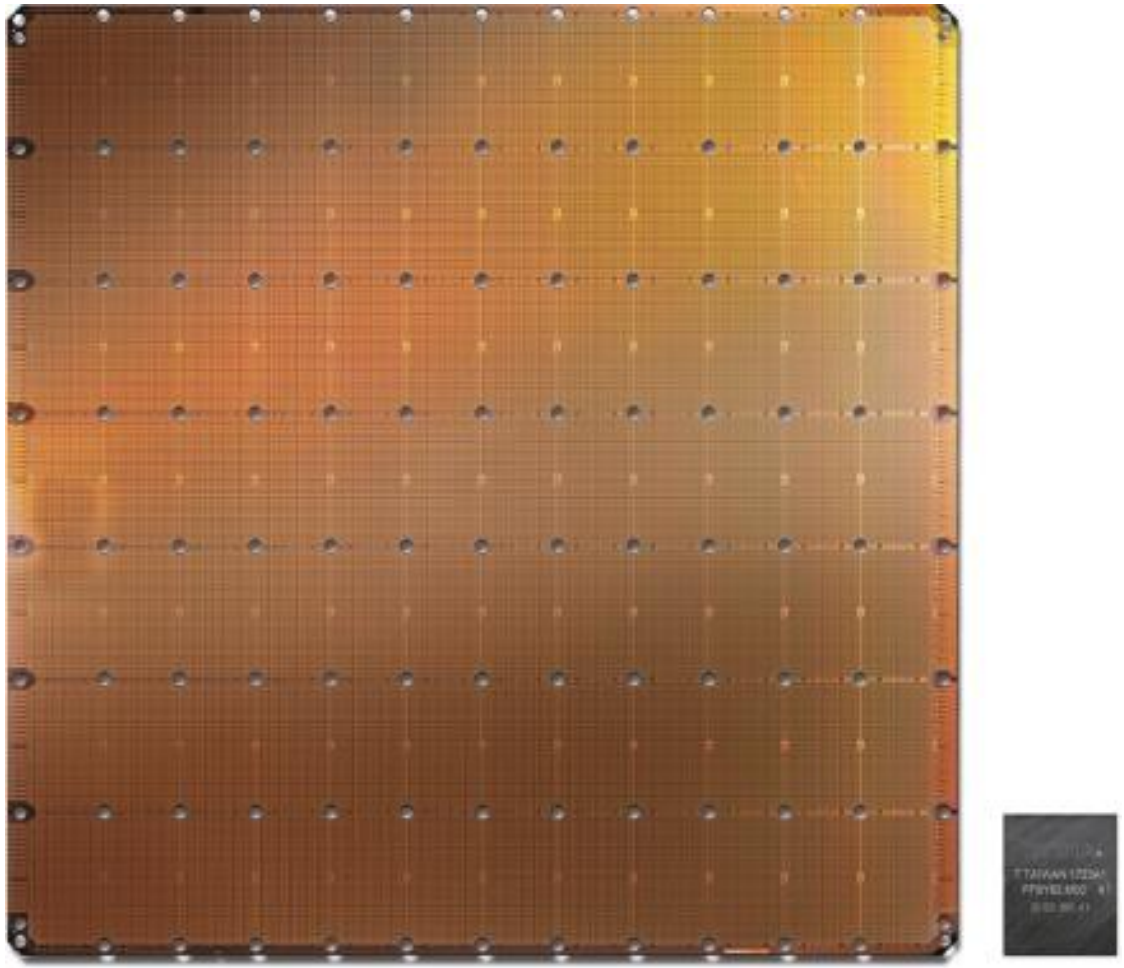
- 左図はData Parallel方式を簡易的に説明した図です
- 「各GPUで学習して全体で同期」という処理を繰り返す
- GPU数が増えるにつれて処理の複雑さ、待ち時間等が爆発的に増えていきます
- チップが大きければオーバーヘッドも減らせるかもしれません、、、

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

巨大チップWSEとCerebras CS-2の紹介

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

Wafer Scale Engine by Cerebras Systems社



46,225㎟ の面積
最大のGPUの56倍

85万コア
最大のGPUの123倍

2.6兆トランジスタ
最大のGPUの50倍以上

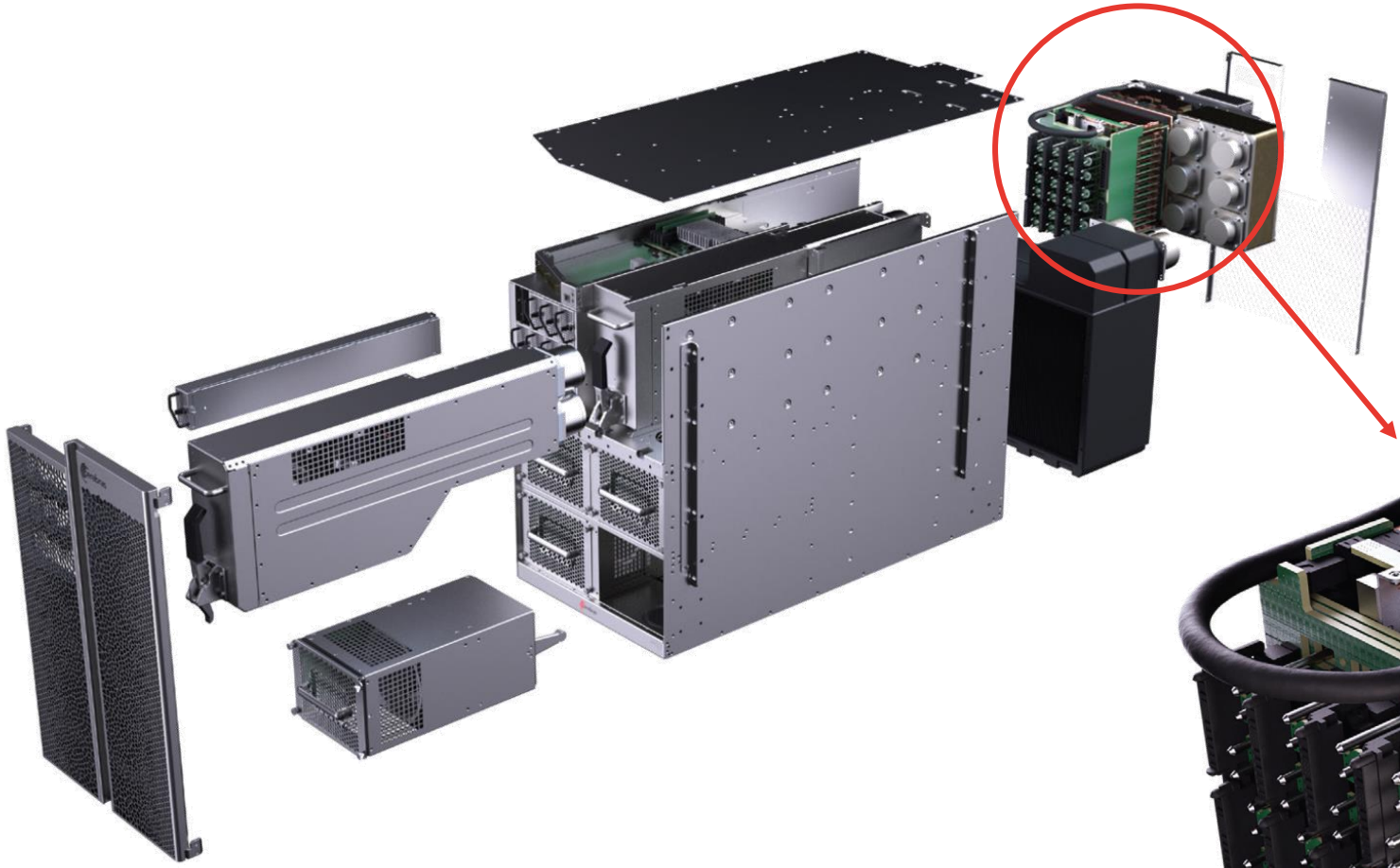
40 GB オンチップSRAM
最大のGPUの1,000倍

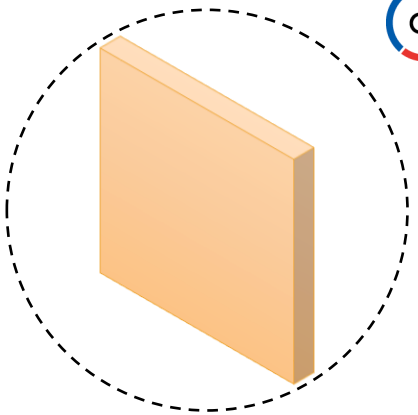
220 Pb/s コア間接続
最大のGPUの45,000倍以上

面積はたった56倍なのに、コア間接続の帯域は45,000倍となっています
※1世代前のGPUとの比較です

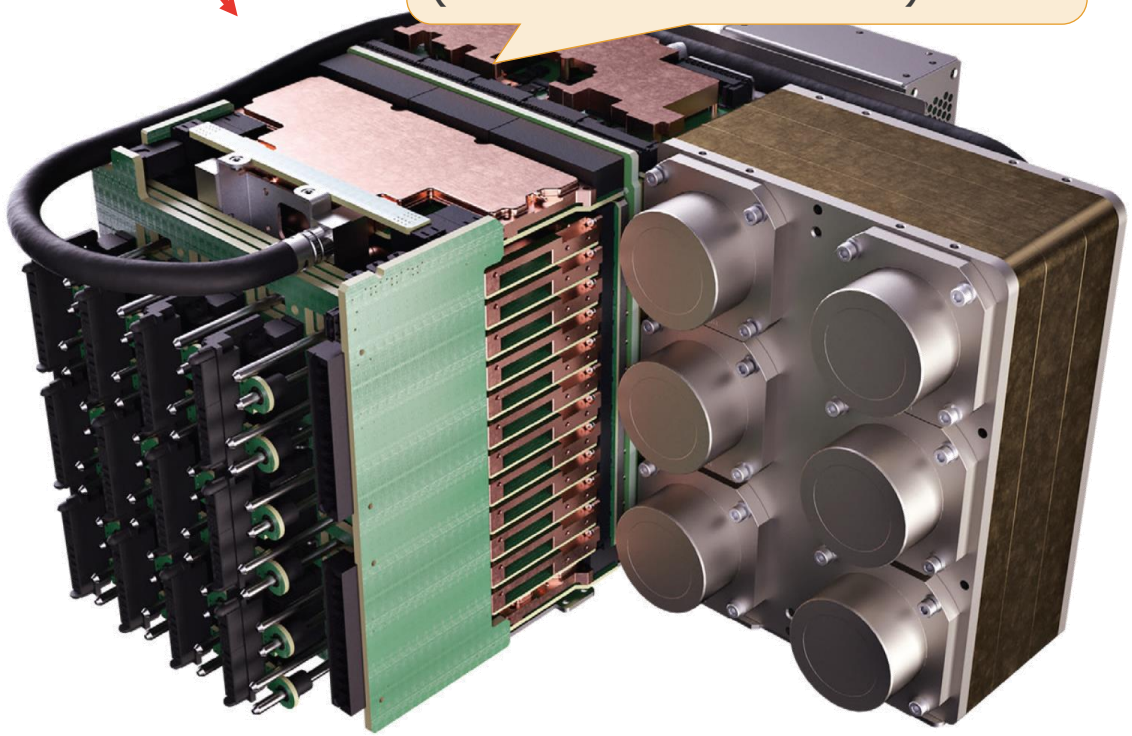
Copyright © Tokyo Electron Device LTD. All Rights Reserved.

筐体3Dイメージ



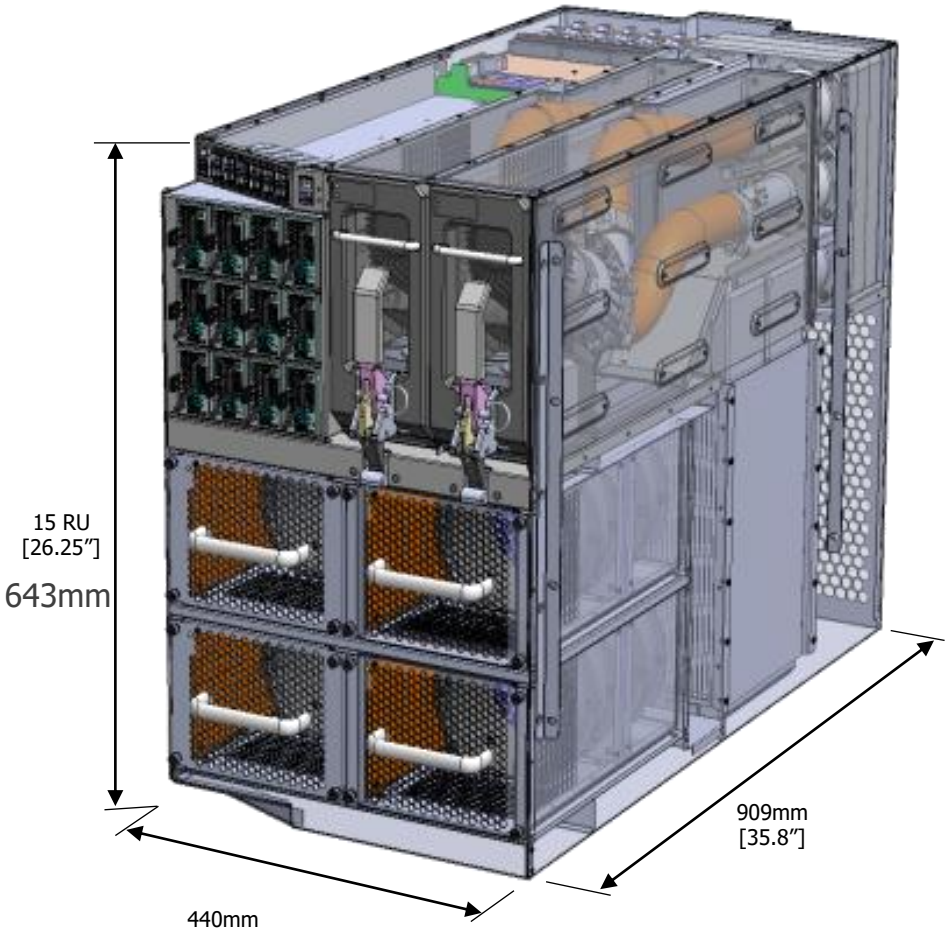


WSEは立った状態で設置される
(上の図形のような入り方)

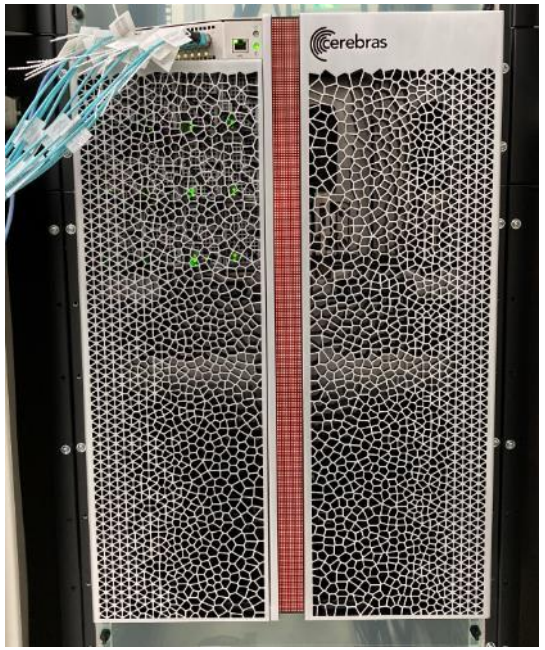


Copyright © Tokyo Electron Device LTD. All Rights Reserved.

CS-2製品外観と仕様



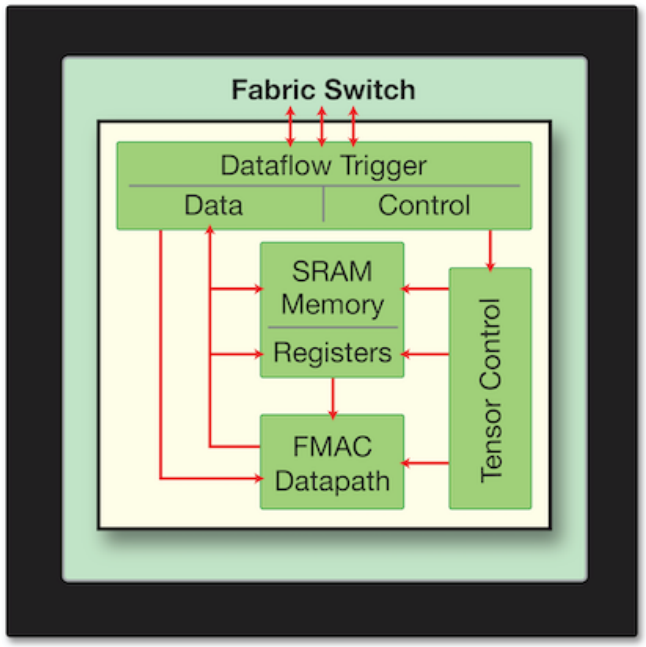
サイズ	15 Rack Units
重量	254 kg (≒17 kg/RU)
最大消費電力	23kW
Process	TSMC 7nm
冷却方式	水冷または空冷
システムIO	1.2 Tb/s (12 x 100 Gig Ethernet)



当社AIラボに設置されている実物(正面と背面)

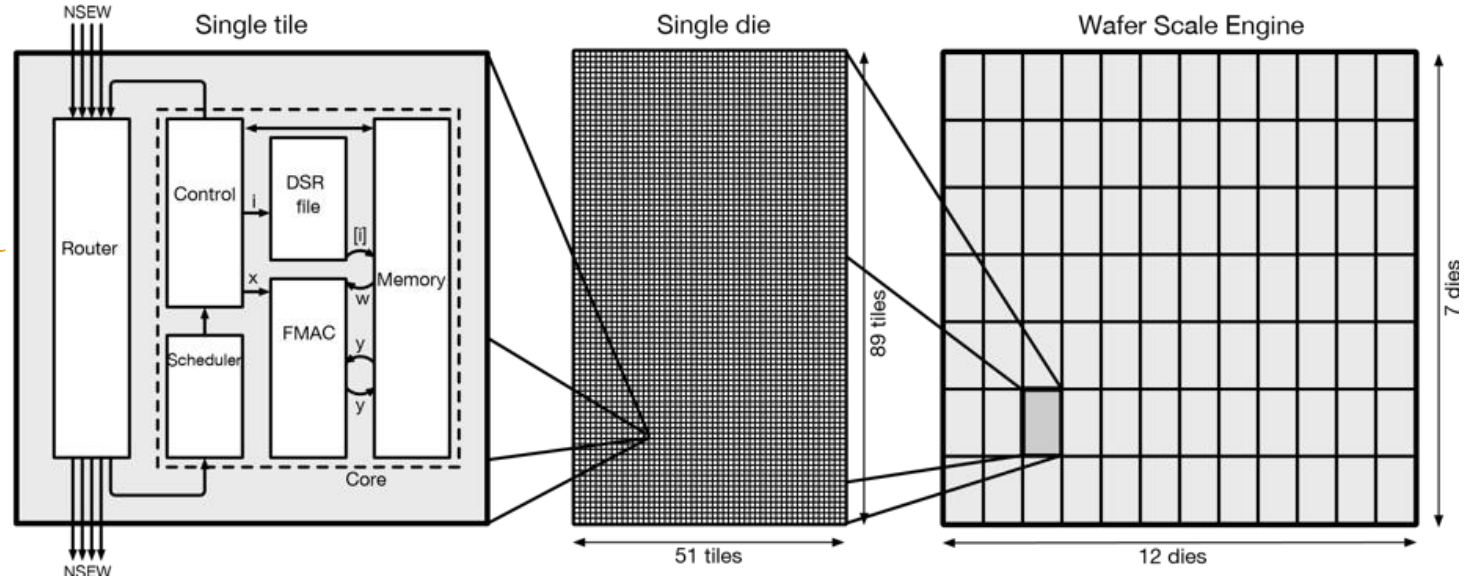
Copyright © Tokyo Electron Device LTD. All Rights Reserved.

WSEの内部アーキテクチャ概要



- ✓ データフローアーキテクチャー
- ✓ 全てのコアが2Dファブリックで相互接続
- ✓ 4方向隣接コアとは1クロックサイクルで双方向通信
- ✓ ゼロをスキップすることで疎行列計算も高速化
- ✓ オンチップメモリを使った高速アクセス
- ✓ 各コアに高速(1クロックサイクルでアクセス可能)SRAMを搭載

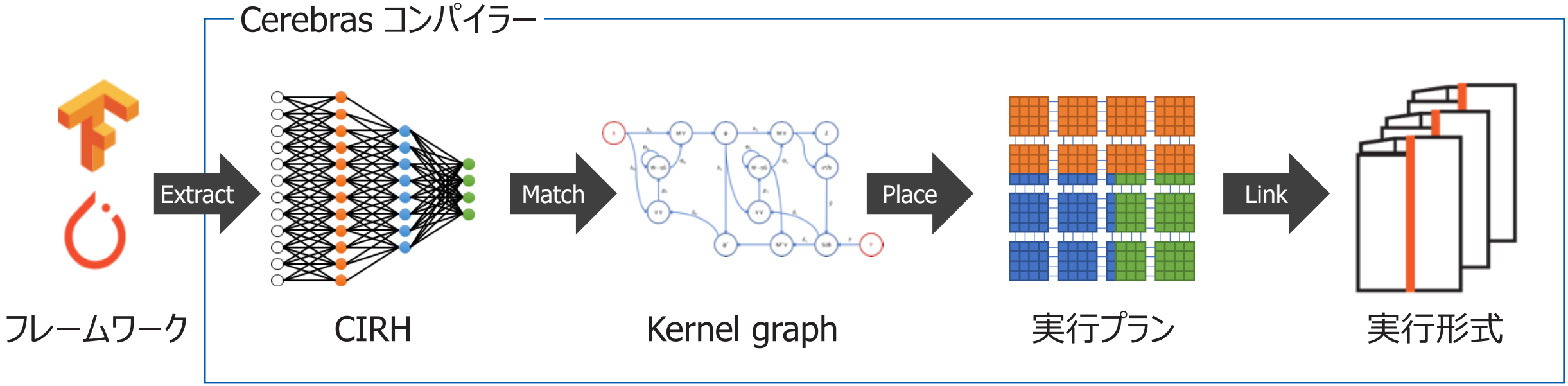
このコアが集まってWSEを形成する
右図はCS-1(第一世代)のケース



Copyright © Tokyo Electron Device LTD. All Rights Reserved.

Cerebras ソフトウェアプラットフォーム (Csoft)

フレームワーク上で書かれたコードがCS-2上でそのまま動作



Copyright © Tokyo Electron Device LTD. All Rights Reserved.

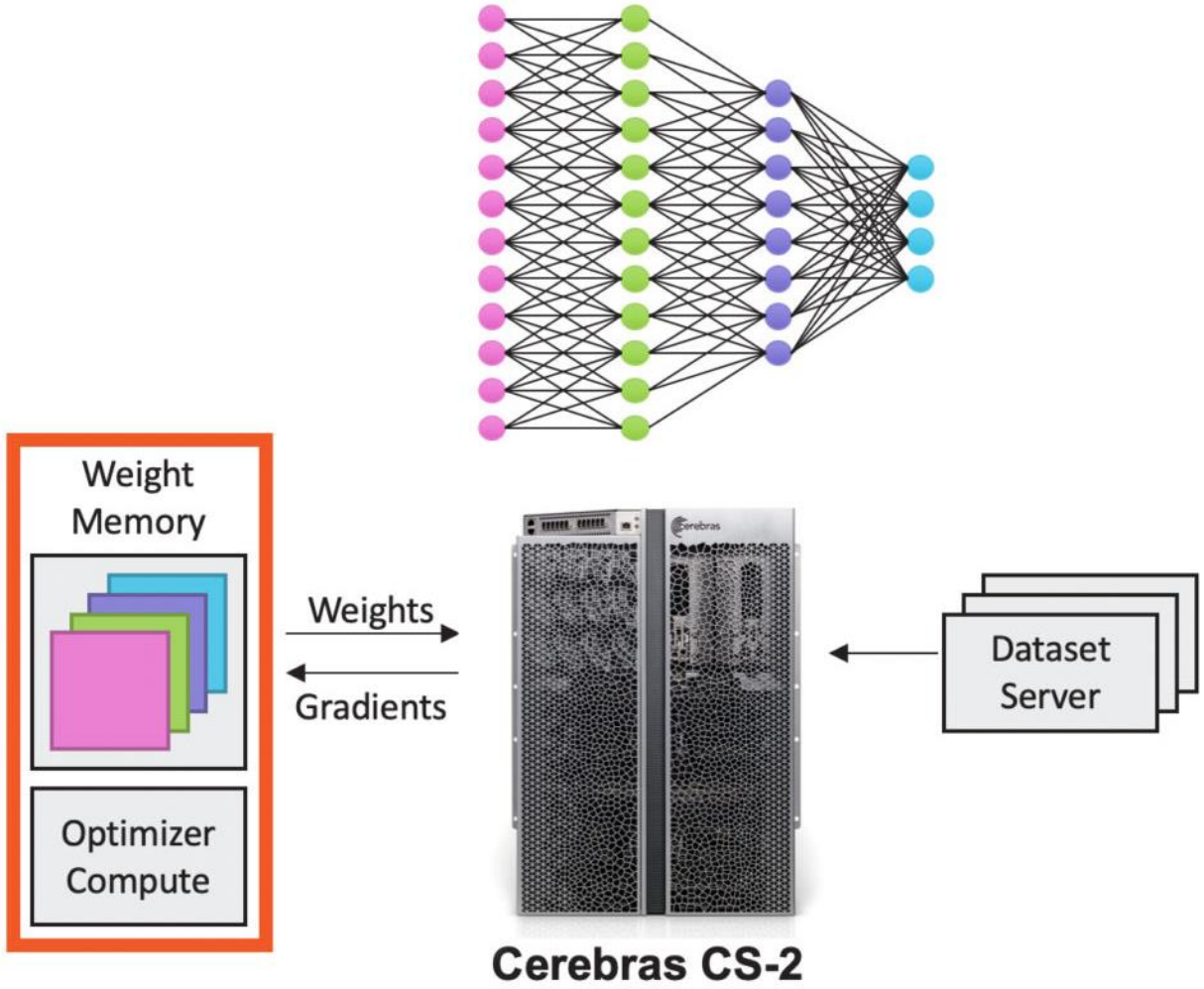
巨大チップを活かした使い方：Weight Streaming

- 通常はモデルの全部、若しくは一部をデバイス上に展開し、それを学習に使用します
- この使い方では自デバイスが、どのパラメータを保持するかは変わることはありません
  - 例えば100パラメータを2分割する場合は、GPU#2の受け持ちは#51～#100と固定されています
- これに対してCerebras Systems は **Weight Streaming** という技術を開発しました
- この技術では Weight を Layer ごとに分割し、分割した単位で Weight を CS-2 に行き来させます
  - Weight、重み、パラメータは同じ意味で使われます
- Weight は外部サーバーに置かれるため、全 Weight を同時にCS-2上に展開する必要がありません
  - 例えば100層のモデルの場合は、あるタイミングで必要なメモリは1/100程度になります
    ※ 実際にはWeight Streamin用の値を代わりに保持しますので、ここまで単純ではありませんが・・・
  - さらに言うと Layer が綺麗に並んでいるようなモデルは少ないため、このあたりに対する工夫もされています

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

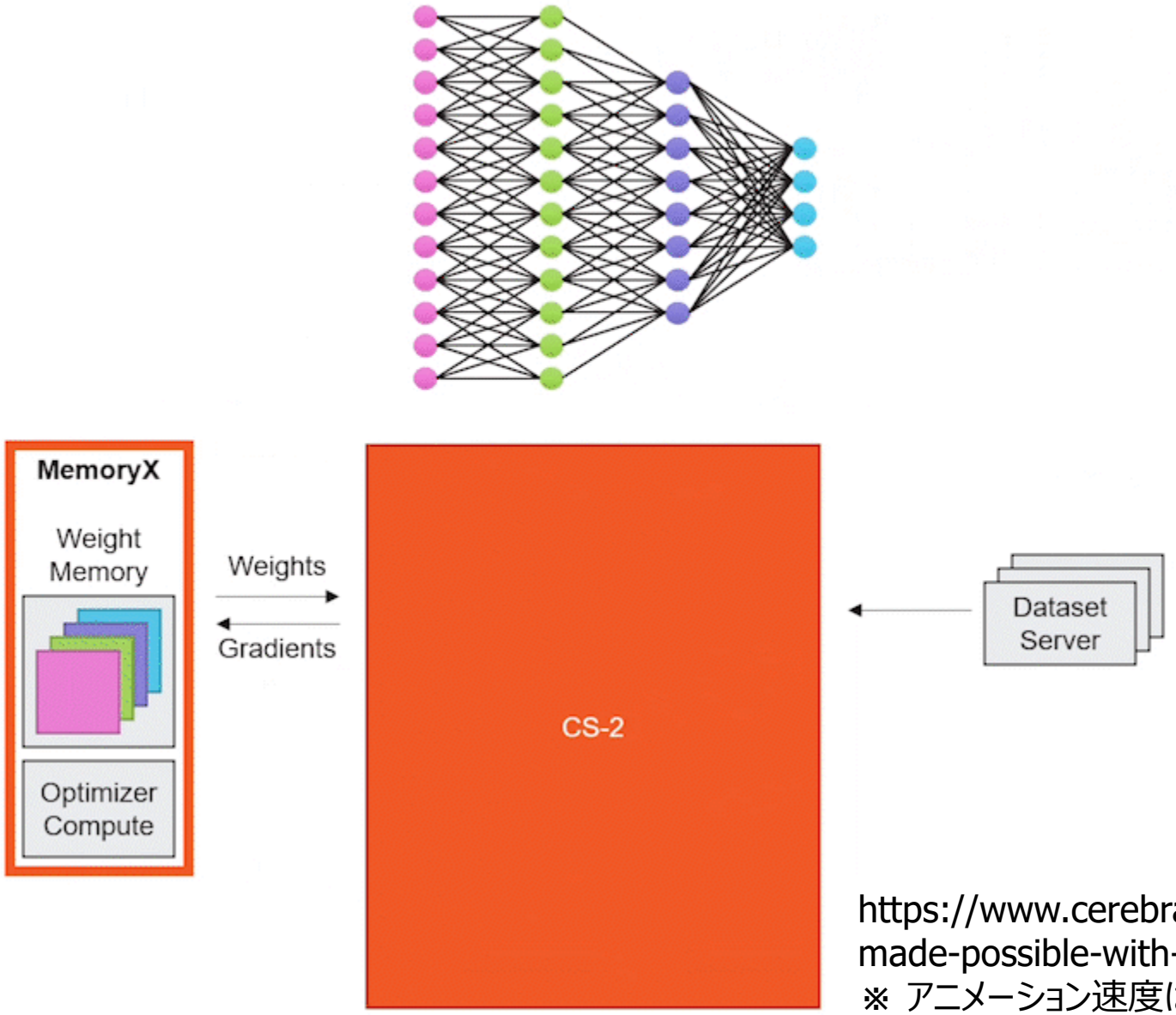
Weight Streamingのイメージ

- 右のような非常にシンプルな4層のニューラルネットワークを例にしてみましょう
  - 左下のMemoryXは専用のIAサーバーでWeightをサーバー内のRAMに蓄えます
  - 右下の四角は学習データを保持しているサーバーです
- Weight Streamingではこれらの4層のWeightを1層ずつ処理します
  - 順伝播：データをCS-2に流して、1層目を流して演算、2層目を流して演算・・・N層目を流して演算します
  - 演算結果についてはCS-2内に保持します
  - 逆伝播：N層目をCS-2に流して演算して結果を基にMemoryX上のWeightを更新、N-1層目を流して演算・・・1層目を流して演算します



Copyright © Tokyo Electron Device LTD. All Rights Reserved.

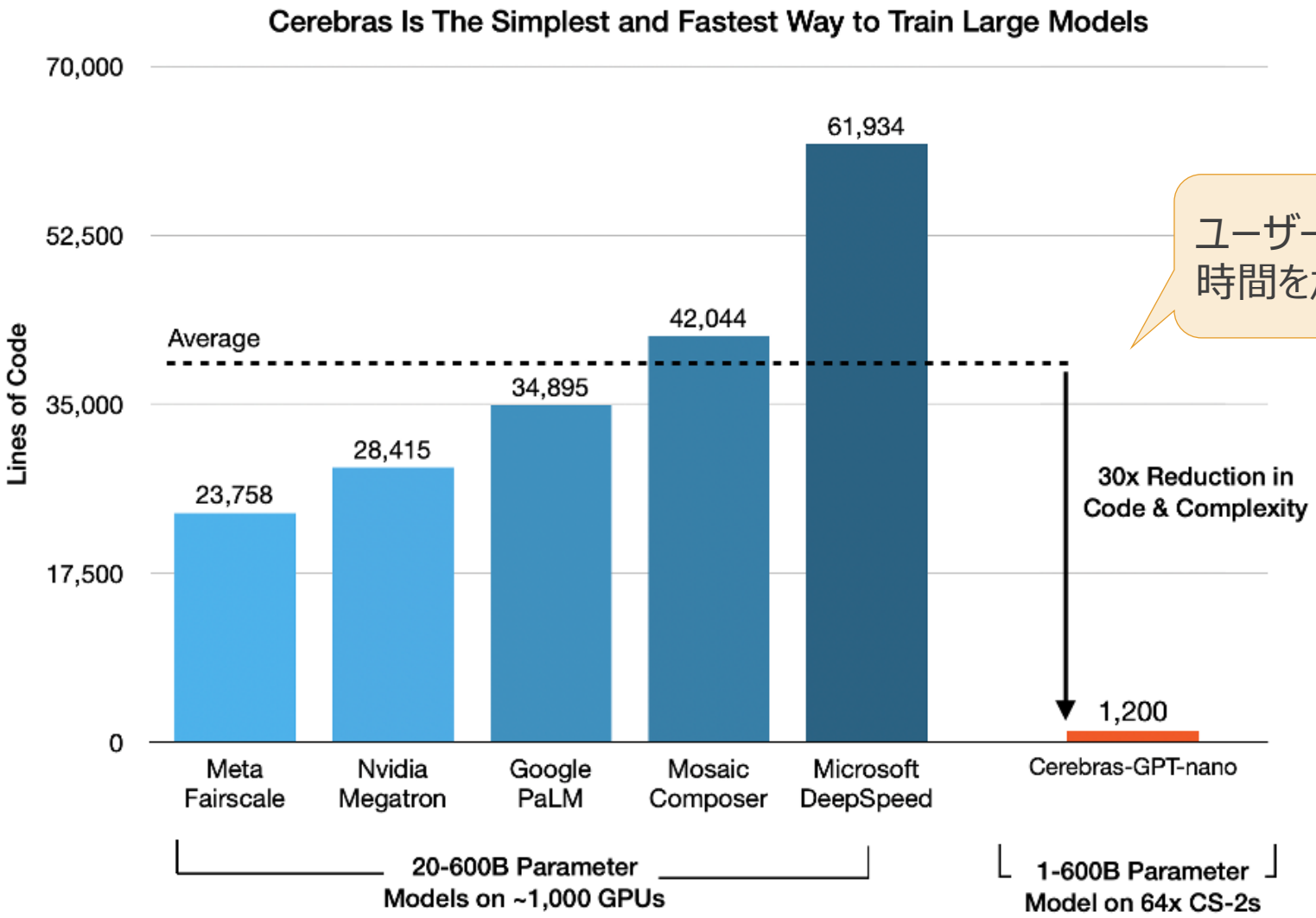
Weight Streaming In Action



注意
Gifアニメーションのため、資料配布時にはアニメはご覧いただけません

https://www.cerebras.net/blog/linear-scaling-made-possible-with-weight-streaming
※ アニメーション速度は調整しています

Cerebras-GPTと他のモデルとのコード量の比較



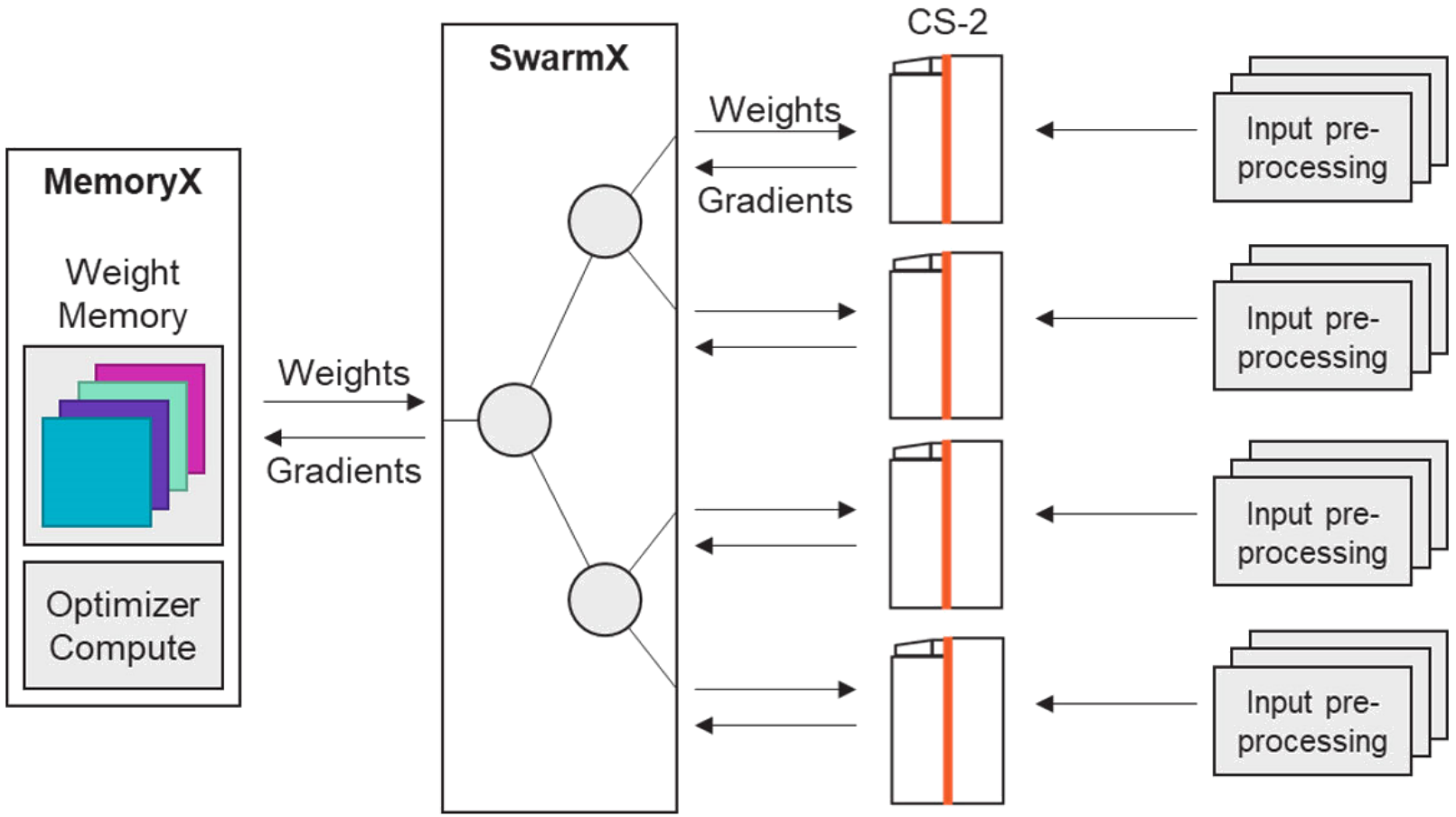
ユーザーは並列処理の実装に、
時間をかける必要がありません

Cerebras社　2023年7月20日投稿記事より転載
https://www.cerebras.net/blog/introducing-condor-galaxy-1-a-4-exaflop-supercomputer-for-generative-ai/

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

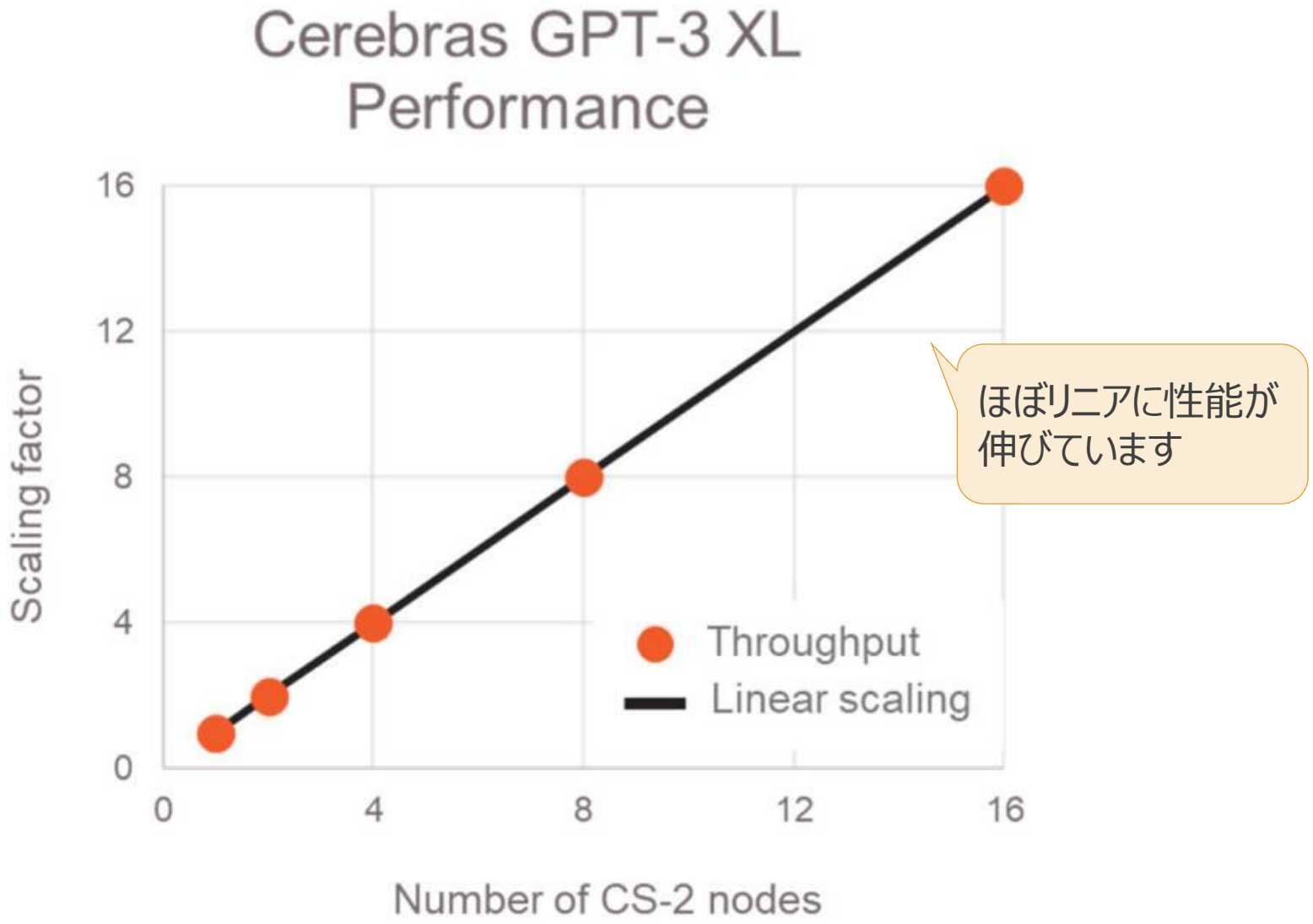
and more ... SwarmX によるクラスタ化

- CS-2を複数台使うことも可能です
- この場合はSwarmXというネットワークスイッチと組み合わせます



Copyright © Tokyo Electron Device LTD. All Rights Reserved.

Weight Streamingの性能



Copyright © Tokyo Electron Device LTD. All Rights Reserved.

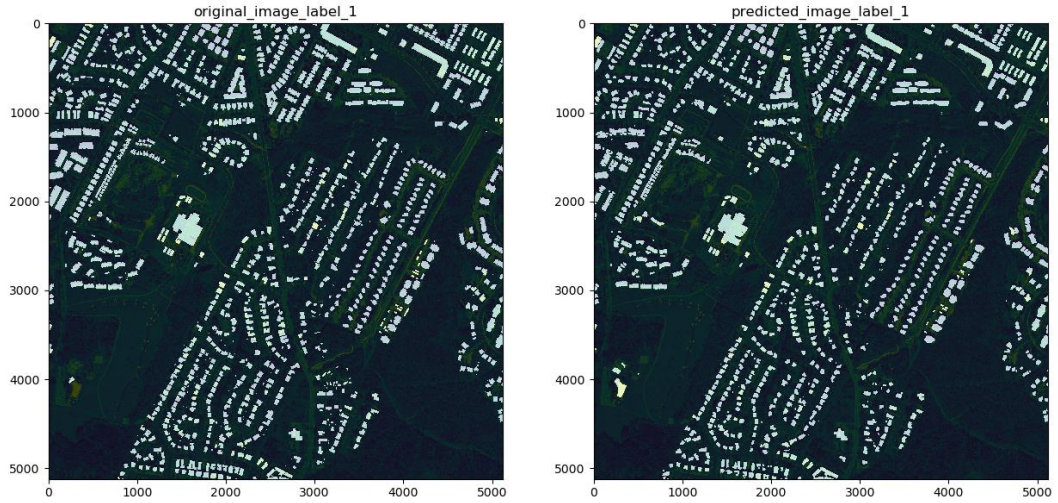
CS-2を用いた超高解像度に対するAI学習

- Cerebras CS-2を用いると、最大**50メガピクセル**のセグメンテーションが可能

  ※50メガピクセル=7,168 x 7,168 x 1 (グレースケールの場合)

- 左は7k x 7k の画像をセグメンテーションした結果
- 右はCS-2で5k x 5kの衛星画像に対し、画像トリミングやサイズダウン無しで学習した結果

左：1,300Pixels、右：7,000pixel



左：オリジナル衛星画像上に表示された正解ラベル
右：同じ画像の上に表示されたピクセル単位の予測値

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

巨大チップWSEとCerebras CS-2の今後

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

新しい歴史が生まれている最中です

August 3, 2022

In プレスリリース

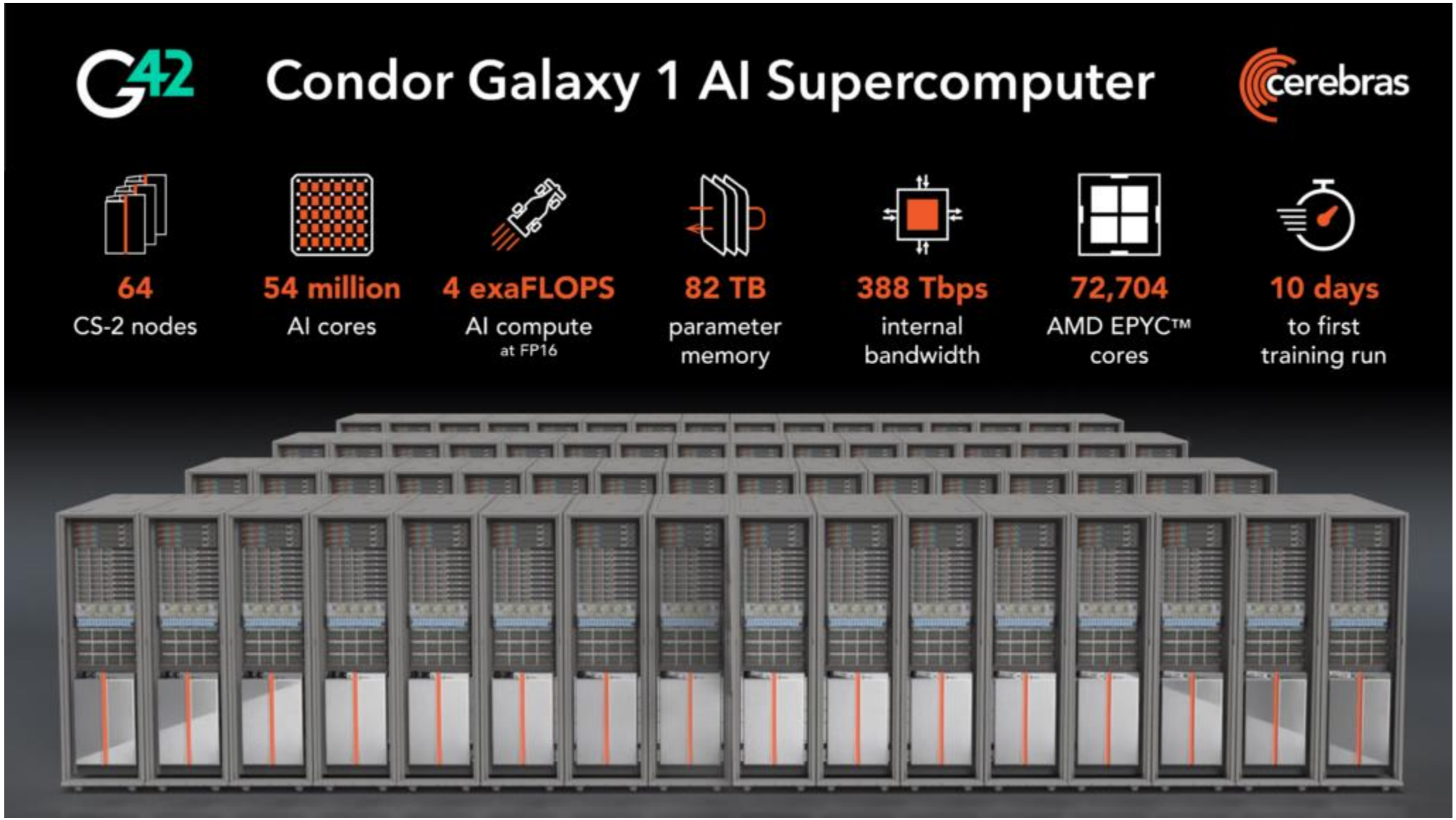
コンピュータ歴史博物館、ウェハースケール・エンジン用の新しい展示場所をCerebras Systemsに提供

The largest chip ever built, WSE-2 has transformed the artificial intelligence landscape, powering the industry's fastest AI supercomputer, the Cerebras CS-2

https://www.cerebras.net/press-release/computer-history-museum-honors-cerebras-systems-with-new-display-for-wafer-scale-engine/

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

CS-2を64台並べたクラスタシステムも提供開始



Copyright © Tokyo Electron Device LTD. All Rights Reserved.

Cerebras CS-2

世界で最もパワフルなAIコンピューター
一つの筐体システムにすべてのソリューションを満載

- **Wafer Scale Engine (WSE)** によって支えられたシステム
- TensorFlow、PytorchでDeep Learningの学習/推論が可能
- Cerebras SDKを用いるとHPC用のプログラムも作成可能
- 通常のサーバーラックに容易に設置
- 複数ラックの従来汎用クラスターサーバーを1ラックの単一システムに統合

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

まとめ

- NVIDIAを筆頭に多くの企業がAI市場の拡大に向けて活動中
- Cerebras Systems、SambaNova、Tenstorrent、Preferred Networks、その他多くの企業が打倒NVIDIAを目標に奮闘中
- その中でも **Cerebras Systems** は **Wafer Scale Engine** と呼ぶ超巨大チップを武器に世間を賑わせている

Copyright © Tokyo Electron Device LTD. All Rights Reserved.


Q & A