ジェネレーティブAIの時代に求められるコンピュート基盤とは

東京エレクトロン デバイス株式会社

2023年6月6日

中川 隆志

イントロダクション

はじめに

(ChatGPTなどの)

大規模言語モデルを ・・・

自分達で作りたいと思いませんか？

自分達で作る必要性はありませんか？

ChatGPTをはじめとするとジェネレーティブAIのサービスが我々の生活に根付いていくことでしょう

そうなってきたときに、果たして核となるAIモデルは借り物でよいのでしょうか

セキュリティは？学習データは？基盤は？などなど多くの課題があることでしょう

本日は**超巨大AIモデルの構築**に必要となる**コンピュート基盤**についてお話しします

今日お話しすること

- イントロダクション
- ジェネレーティブAIの台頭
- 超巨大AIモデルを学習するということ
- Cerebras CS-のご紹介

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

自分の職業を一言で表すと **データエンジニア** が近いのかな

プログラマー

- Visual C++
- Visual Basic

Big Data

- Data Warehouse
- Hadoop

★転職

IoT

- 予知保全
- プロトタイプ開発

AIアクセラレータ

- Cerebras
- TED AI Lab

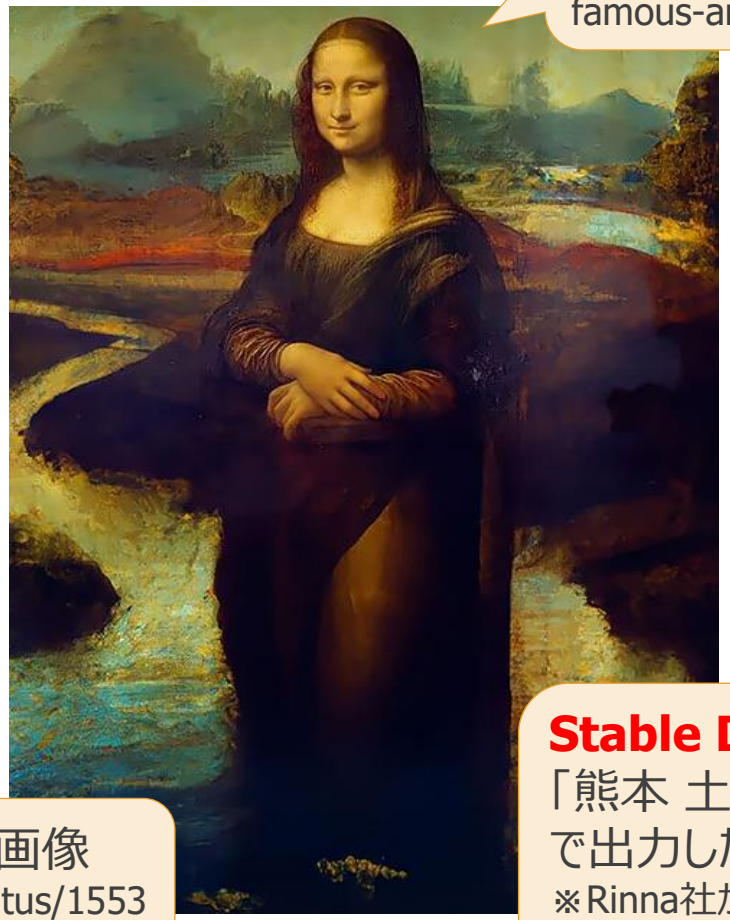
200X年 → 2010年 → 2016年頃 → 2019年頃 → 現在

ジェネレーティブAIの台頭

画像出力は既に実用レベル

Midjourney でバズった画像
https://twitter.com/8co28/status/1553611630252277760

DALL E2
https://designyoutrust.com/2022/06/the-dall-e-2-neural-network-has-redrawn-classic-paintings-by-famous-artists-while-preserving-their-style/



Stable Diffusion に対して「熊本 土産」というキーワードで出力した画像
※Rinna社が日本語用に学習して公開したモデルを使用

※Dall E2、Stable Diffusion は **ViT** = Vision Transformer を用いている

ChatGPTの衝撃

- 2022年11月30日に公開され、その性能の高さから大きな話題になっています
- これもまた**Transformer系列**のモデルであり、GPT-3.5 をファインチューニングしたものです
- 公開から**1ヶ月強で1億人**のアクティブユーザー数を記録し、他と比べるとそのスピードは凄まじいものです
  - TikTok：9ヶ月
  - Instagram：2年半
  - Facebook：4年半

- 将来的にはGoogle検索が不要になるかもしれません

ChatGPTのリリースでGoogleは「コードレッド」を宣言、AIチャットボットが検索ビジネスにもたらす脅威に対応するためにチームを再割り当て

https://gigazine.net/news/20221223-google-code-red-against-chatgpt/

ChatGPTの進化は止まりません

- 2023年3月にはGPT4がリリースされました
  - 有償版のChatGPTでGPT4を使用できます
- GPT3よりもさらにすごいことができます
  - 画像を見てその面白さを説明できます
  - 国家試験で合格点を取れます
- 残念ながらパラメータ数は公開されていません
  - 5000億とも100兆とも言われています

  (でも学習時間を考えると数兆が限界では？)

GPT-4 visual input example, Chicken Nugget Map:

User　Can you explain this meme?

Sometimes I just look at pictures o
the earth from space and I marvel a
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 below is a b
earth from space. However, the image is actually of chicke
vaguely resemble a map of the world.
The humor in this meme comes from the unexpected juxtap
the image. The text sets up an expectation of a majestic imag
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

ChatGPTを業務利用してみたくないですか？

カスタマーサポート

- 顧客の問い合わせやサポート要求に対して即座に応答
- 自然言語理解と迅速な情報処理能力を活用した問題解決

タスクの自動化（処理）

- メールのフィルタリングや分類、文書の要約、データの整形などの作業の自動化

マーケティング/セールスサポート

- 顧客との対話を通した製品やサービスの情報提供、QAの実施、個別のニーズに合わせた提案

知識管理と教育

- 企業内の知識管理システムと統合することによる社内の情報へのアクセスや検索
- 新入社員のオンボーディングや研修プログラムへの活用

業務に導入する企業、多数

BUSINESS INSIDER　ビジネス　テクノロジー　働き方　サイエンス　政治　国内　国際

パナソニックコネクトの「社内ChatGPT」全社導入。1カ月使い倒して見えてきた成果とは

ChatGPT　パナソニック（Panasonic）

太田百合子［テクニカルライター］
Apr. 12, 2023, 08:05 AM | テクノロジー　138,604

https://www.businessinsider.jp/post-268299

日本経済新聞　朝刊・夕刊　LIVE

トップ　速報　オピニオン　経済　政治　ビジネス　金融　マーケット　マネーのまなび　テック　国際　スポーツ　社

ソフトバンク、生成AI活用の新会社　ChatGPT生かす

ネット・IT　＋フォローする

2023年5月10日 20:14

保存

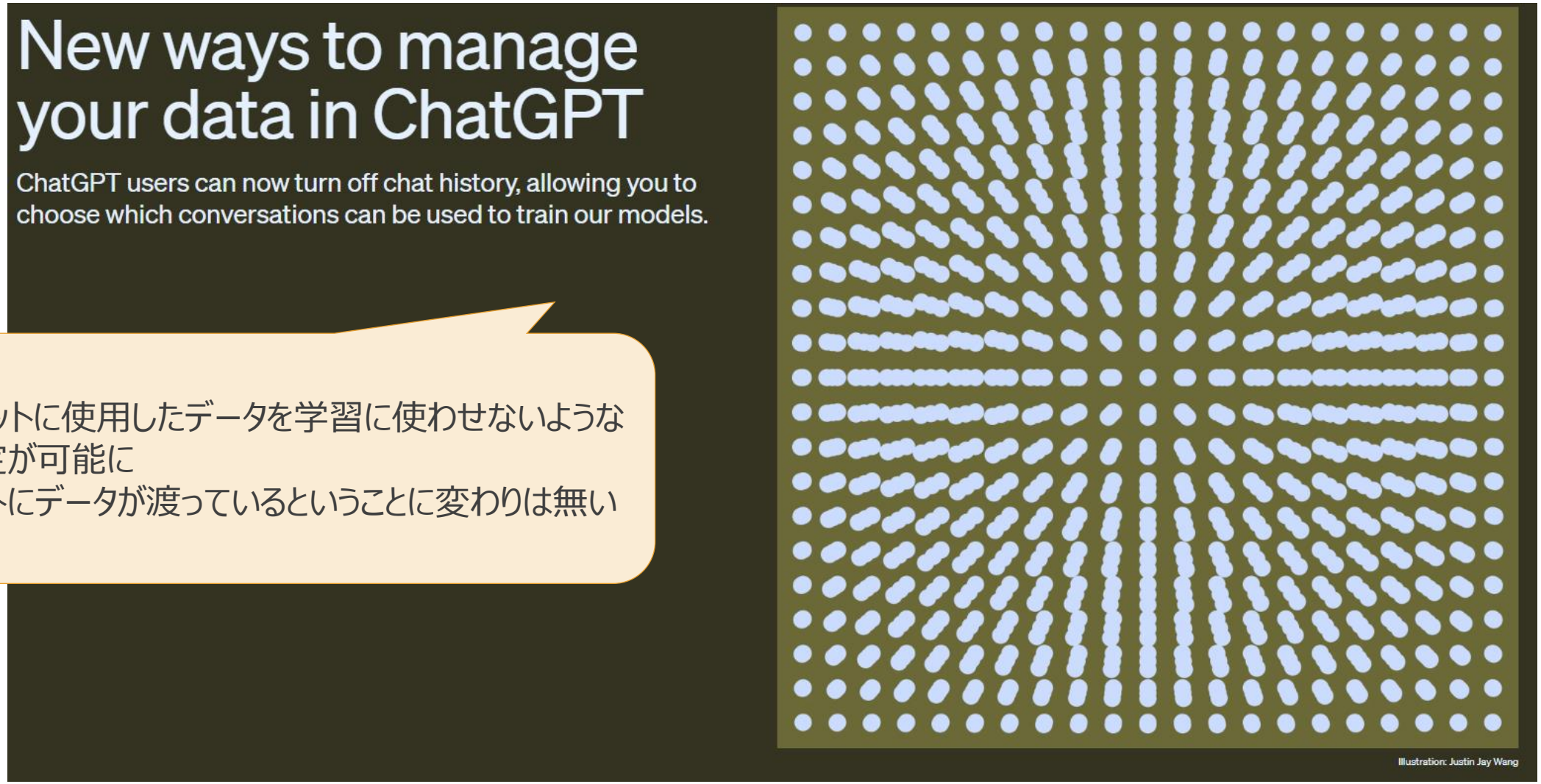
https://www.nikkei.com/article/DGXZQOUC10AQS0Q3A510C2000000/

とはいえ、手放しで使用することは危険

- データのセキュリティ
  - SamsungのエンジニアによるChatGPTに社外秘コードを貼付け、会議音声のアップロード
    - https://gigazine.net/news/20230410-samsung-chatgpt-security-leak/
    - **会社はChatGPT使用を許可していた**
    - ただし、「社内情報セキュリティには注意し、私的な内容は入力してはならない」と通告済みであった・・・
  - 気を付けないと、チャット内容が勝手に学習に使われてしまう

- 回答の正確性や特定の分野での精度への懸念
  - 誤った情報をあたかも正常であるかのように答える
  - オープンデータで学習されているため、特定のドメインでは頓珍漢な回答をする

入力データをChatGPTに学習させない機能追加



- チャットに使用したデータを学習に使わせないような設定が可能に
- 海外にデータが渡っているということに変わりは無い

https://openai.com/blog/new-ways-to-manage-your-data-in-chatgpt

日本国内での動き

政府 ChatGPTなど有効活用に向け新たに検討チーム設置へ

2023年4月14日 21時12分

「ChatGPT」など、文章や画像を自動的に
野で有効に活用していくためには、課題へ
省庁による検討チームを設ける方針を固め

https://www3.nhk.or.jp/news/html/20230414/k10014039071000.html

G7デジタル相会合「信頼できるAI」へ ルール作りが課題

2023年5月1日 7時13分

4月30日に閉幕したG7＝主要7か国のデジタル・技術相会合は「信頼できるAI」の普及に向けた取り組みを進めることで合意しました。ただ、AIをめぐる具体的なルール作りはこれからで、G7として議論を深めていけるかが課題となります。

https://www3.nhk.or.jp/news/html/20230501/k10014054471000.html

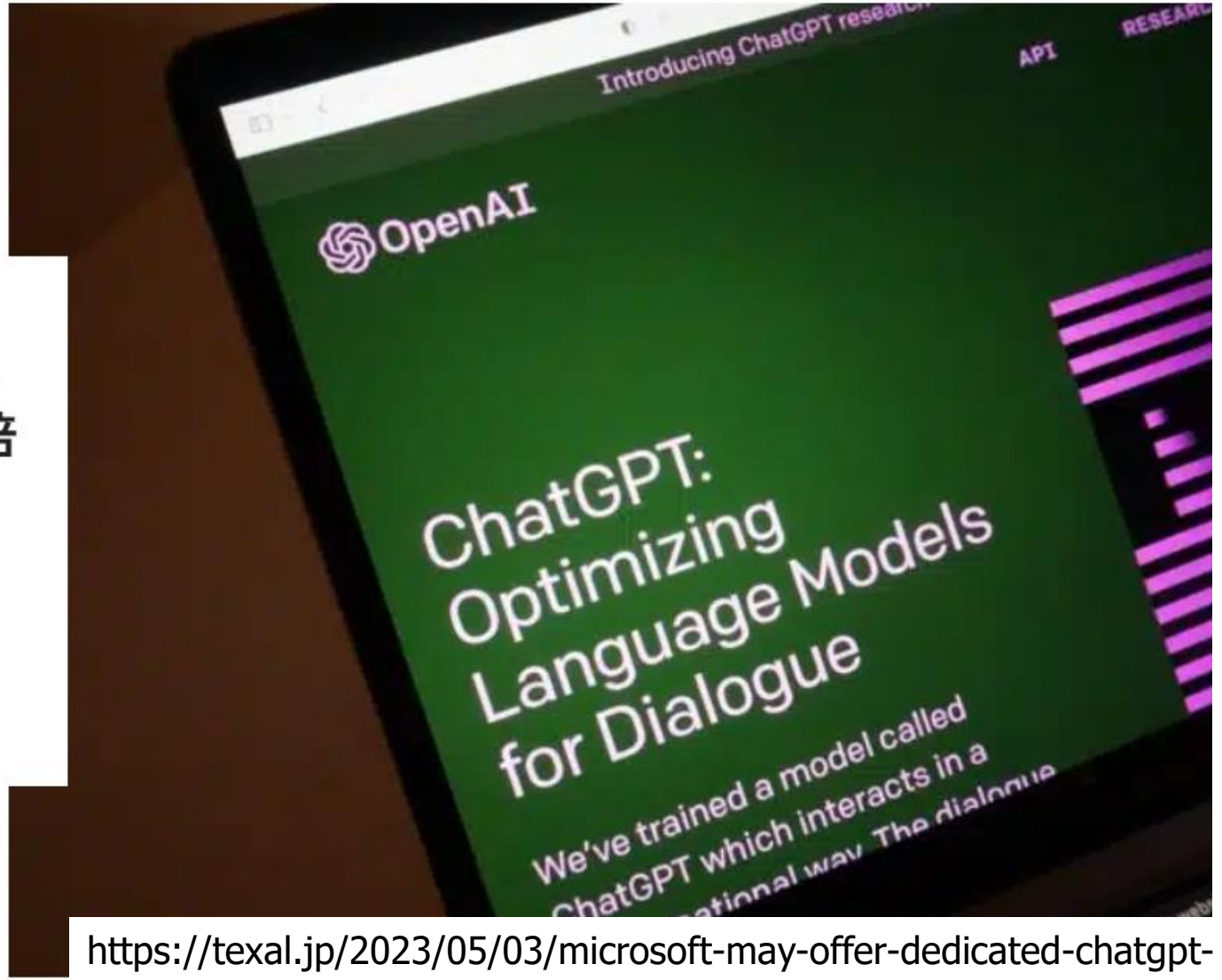
え、マイクロソフトさん？

～からの、↓

Microsoft、データ漏洩リスクに対処した「専用ChatGPT」を企業向けに“10倍のコスト”で提供する可能性

masapoco | 投稿日：2023年5月3日 6:31

テクノロジー



https://texal.jp/2023/05/03/microsoft-may-offer-dedicated-chatgpt-to-businesses-at-10-times-the-cost-to-address-data-breach-risks/

自分達のための安全安心なモデルを作ろう

・

・

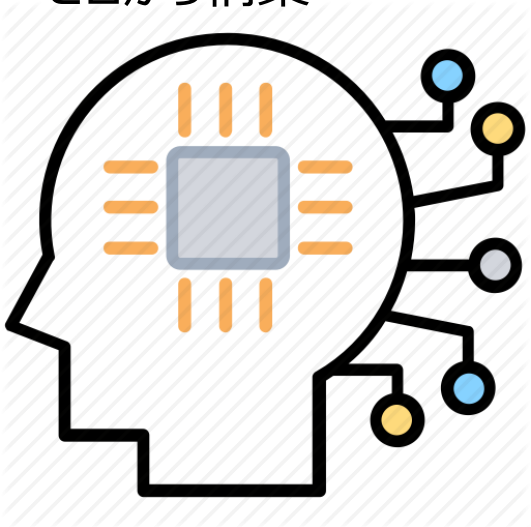
・

でもどうやって？

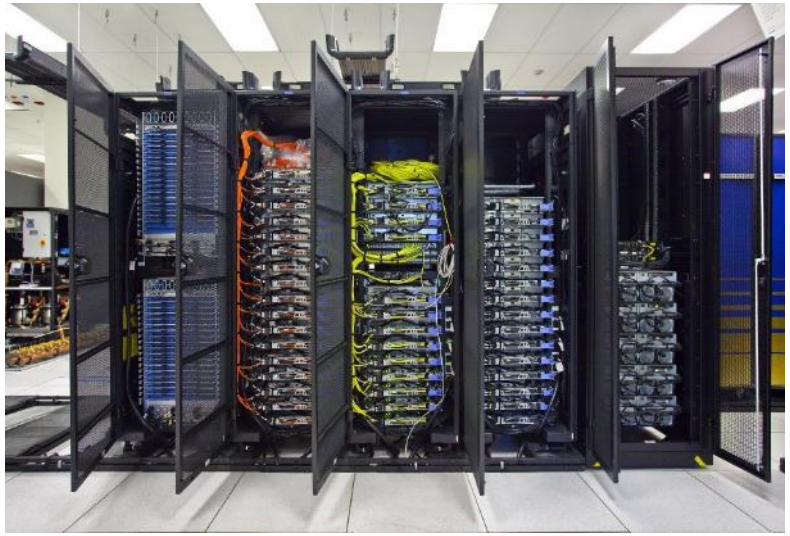
ジェネレーティブAIを作るということ

大規模モデルの作成に必要なもの

1. モデル (アルゴリズム)
- 学習済みモデルの活用
- OSSの活用
- ゼロから構築



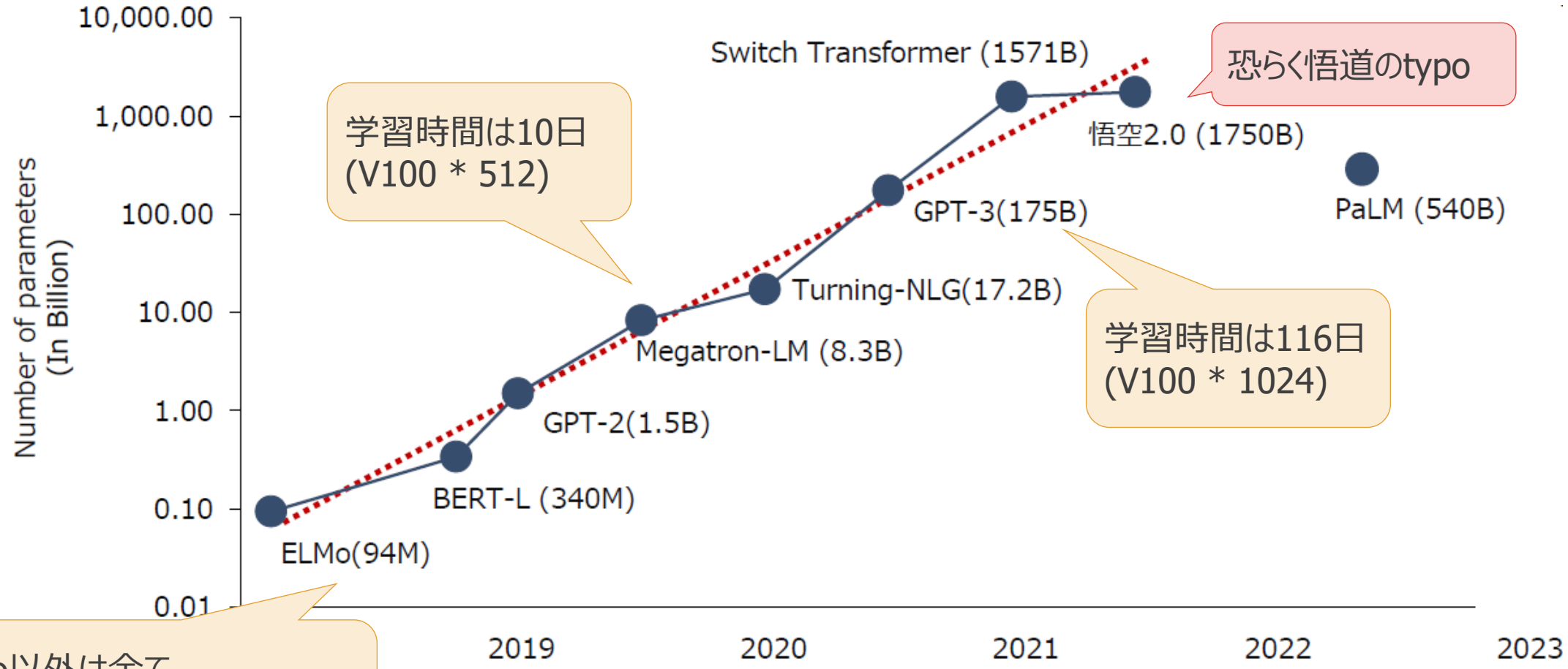
2. コンピュート基盤
- AI特化のHW
- 超大規模に耐えられる設計



3. 学習データ
- 十分な日本語データ
- ドメイン特化のデータ

※今日は主に2についてお話しします

昨今のモデルサイズのトレンドと学習時間



ELMo以外は全て、
Transformer系の言語モデル

https://note.com/api/v2/attachments/download/a29a2e6b5b35b75baf42a8025d68c175

大規模モデルの学習にかかるコスト

Optimal LLM Training Cost				
Model	Size (# Parameters)	Tokens	GPU	Optimal Training Compute Cost
MosaicML GPT-30B	30 Billion	610 Billion	A100	$ 325,855
Google LaMDA	137 Billion	168 Billion	A100	$ 368,846
Yandex YaLM	100 Billion	300 Billion	A100	$ 480,769
Tsinghua University Zhipu.AI GLM	130 Billion	400 Billion	A100	$ 833,333
Open AI GPT-3	175 Billion	300 Billion	A100	$ 841,346
AI21 Jurassic	178 Billion	300 Billion	A100	$ 855,769
Bloom	176 Billion	366 Billion	A100	$ 1,033,756
DeepMind Gopher	280 Billion	300 Billion	A100	$ 1,346,154
DeepMind Chinchilla	70 Billion	1,400 Billion	A100	$ 1,745,014
MosaicML GPT-70B	70 Billion	1,400 Billion	A100	$ 1,745,014
Nvidia Microsoft MT-NLG	530 Billion	270 Billion	A100	$ 2,293,269
Google PaLM	540 Billion	780 Billion	A100	$ 6,750,000

https://www.semianalysis.com/p/the-ai-brick-wall-a-practical-limit

解

決

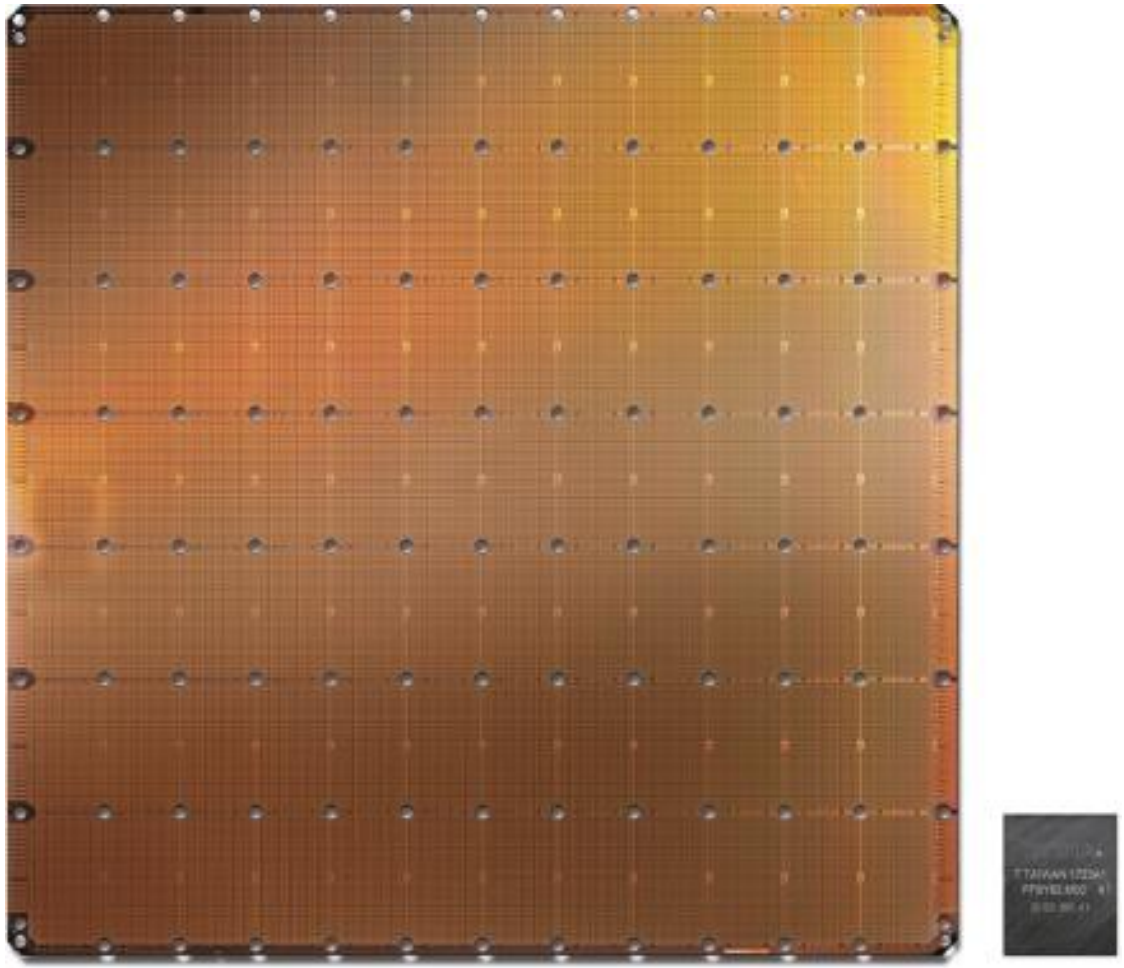
Cerebras CS-2 のご紹介

Cerebras CS-2 とは？

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

Wafer Scale Engine



46,225㎟ の面積
最大のGPUの**56倍**

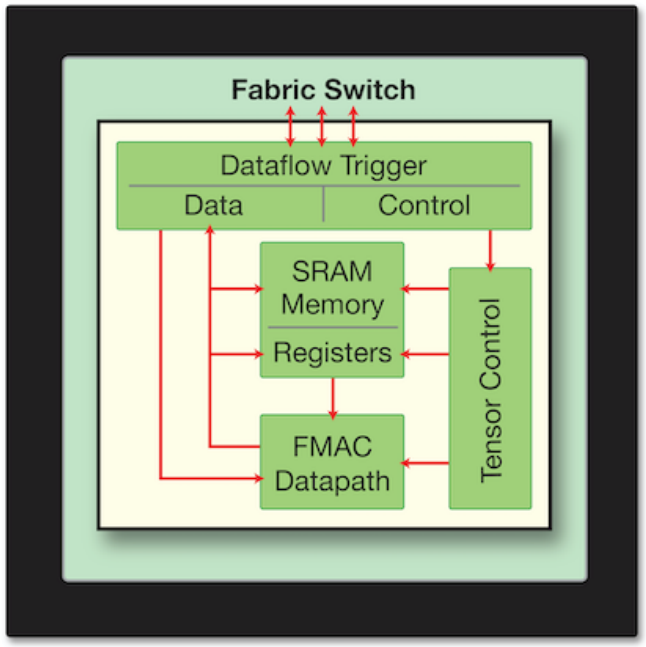
85万コア
最大のGPUの**123倍**

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**

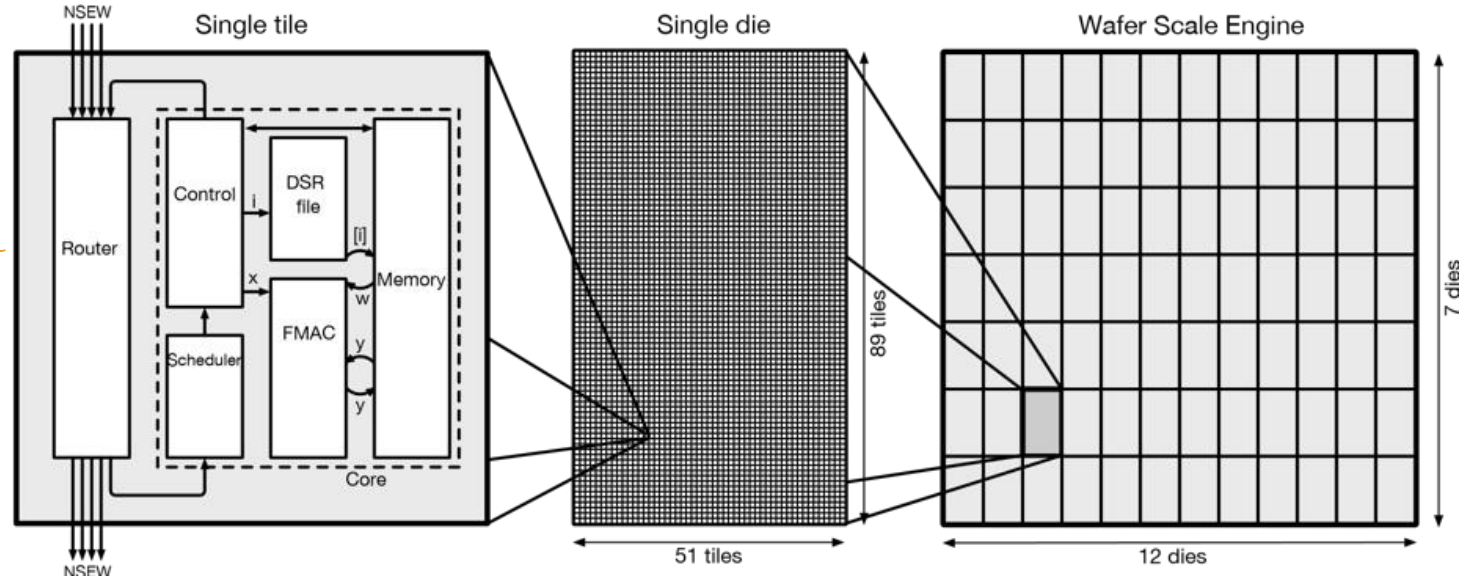
220 Pb/s コア間接続
最大のGPUの**45,000倍**以上

WSEの内部アーキテクチャ概要



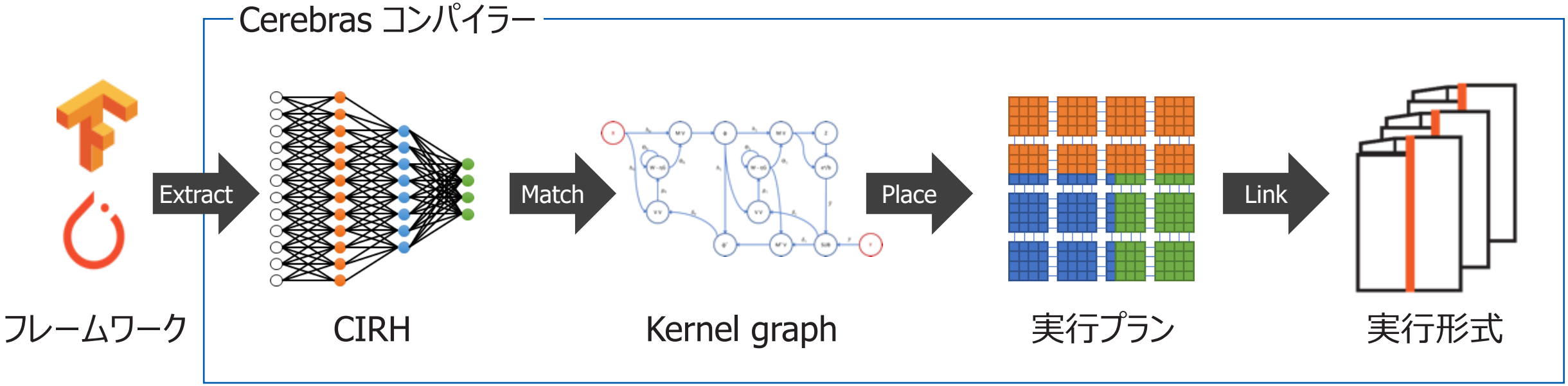
- ✓ データフローアーキテクチャー
- ✓ 全てのコアが2Dファブリックで相互接続
- ✓ 4方向隣接コアとは1クロックサイクルで双方向通信
- ✓ ゼロをスキップすることで疎行列計算も高速化
- ✓ オンチップメモリを使った高速アクセス
- ✓ 各コアに高速(1クロックサイクルでアクセス可能)SRAMを搭載

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



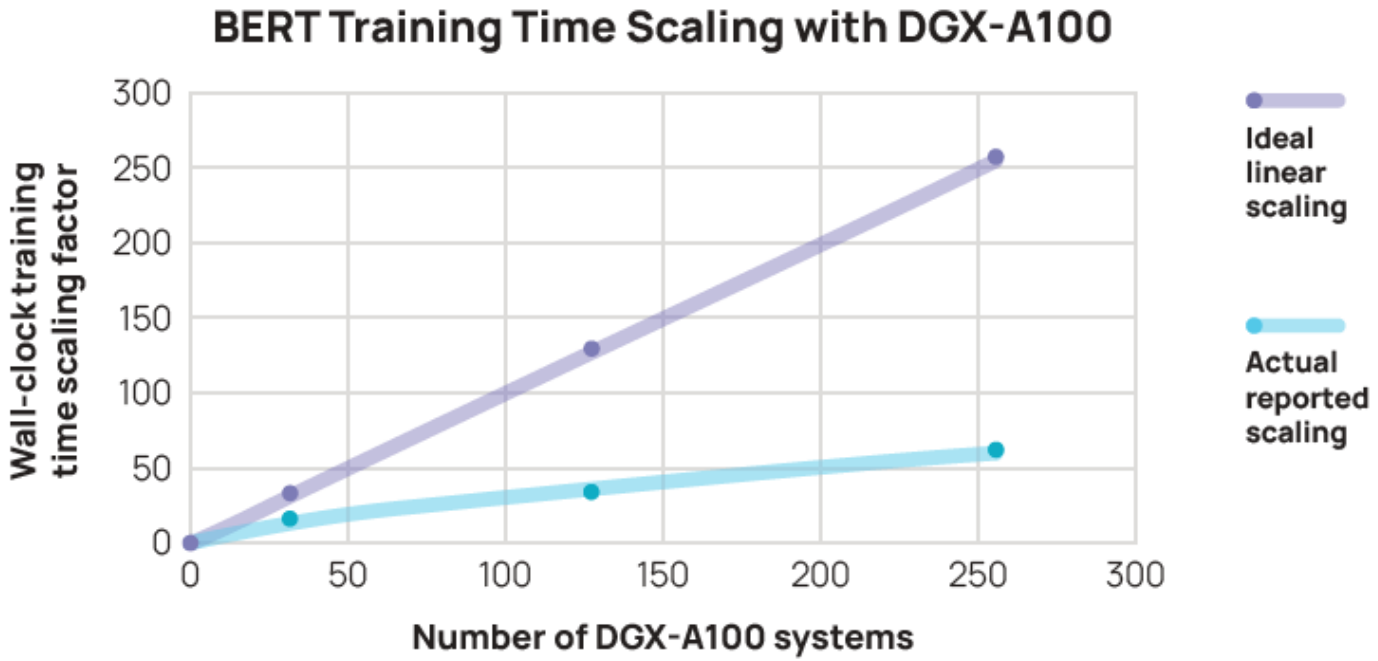
Cerebras ソフトウェアプラットフォーム (Csoft)

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

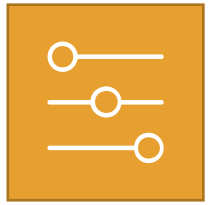


Why Cerebras ?

- Cerebrasの特徴、それは **1チップ** であるということ
- 1チップであるということは、チップ間通信のオーバーヘッドを考える必要がありません
- 以下はMLPerf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
- 数百枚と言ったオーダーでは、期待する性能の数分の1程度しか性能が出ていません



巨大モデルに対してのテクニック：2つの実行モード



Pipelined

Stationary weights,
streaming activations

- モデル全体が同時にCS-2ファブリックにマッピング
- アクティベーションとデルタスパース性を活用する本質的な機能
- 非常に低いレイテンシー

Weight Streaming

Stationary activations,
streaming weights

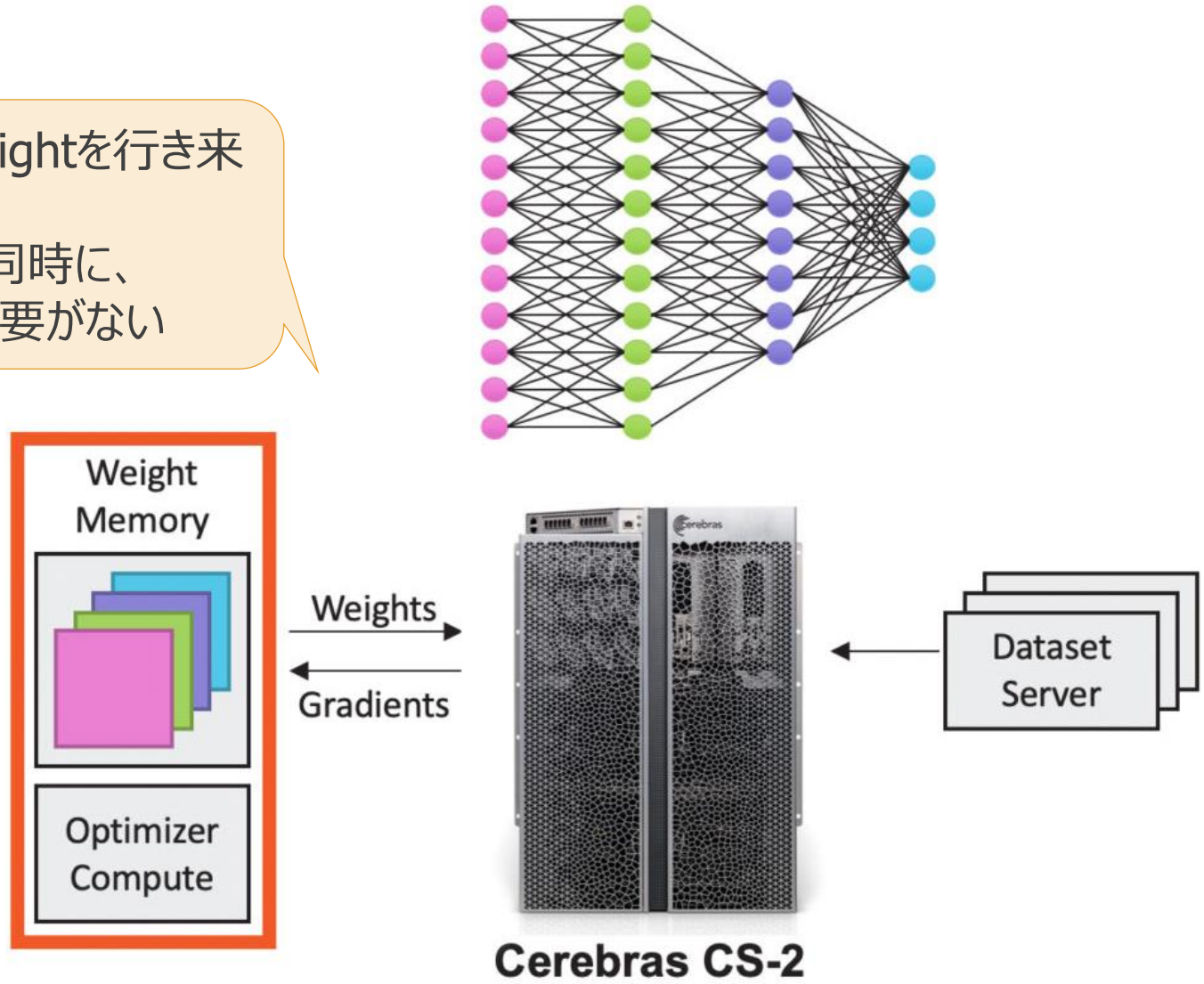
- ネットワークをCS-2ファブリックに一層ずつローディング
- 動的な重みのスパース性を活用する本質的な機能

どちらを利用するかはモデルのサイズによって選択:

通常サイズモデル vs. 超大規模モデル

Weight Streamingのイメージ

- Layer (層) ごとにWeightを行き来させる
- そのため全Weightを同時に、メモリ上に展開する必要がない



Cerebrasが提供する効果的な学習

オープンなモデル

- 様々な生成モデルが公開されているため選択肢は多いです
- ここでは商用で使える学習済みモデルについてお話しします

AI専用の基盤

- AIにはAIのために作られた専用の基盤が必要です
- Cerebras CS-2は他のアクセラレータと比べて一線を画す存在です

様々な提供形態

- オンプレミスのみならず、クラウドサービス (ファインチューニング、スクラッチ) といった形態で利用できます
- 弊社エンジニアによるサポートもありますので安心してご利用ください

フリーで使える学習済みモデル：Cerebras-GPT

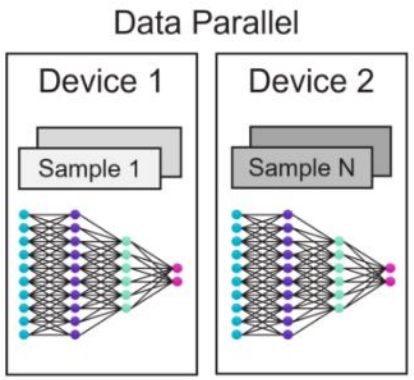
- Cerebrasは110M～13Bのパラメータを持つ7つのGPT StyleモデルをOSS化しました
- これらはチンチラ式で学習させているため、十分な精度を出すことができます
- Hugging Faceで公開中 (https://huggingface.co/cerebras)

Model	Model architecture	Training data	Model weights	Checkpoints	Compute-optimal training	License
OpenAI GPT-4	Closed	Closed	No	No	Unknown	Not available
Deepmind Chinchilla	Open	Closed	No	No	Yes	Not available
Meta OPT	Open	Open	Researchers Only	Yes	No	Non-commercial
Pythia	Open	Open	Open	Yes	No	Apache 2.0
Cerebras-GPT	Open	Open	Open	Yes	Yes	Apache 2.0

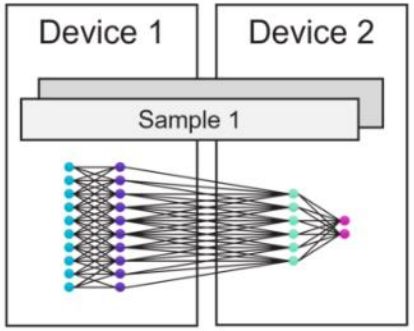
Cerebras-GPTと他のモデルとの学習方法の比較

Model	Hardware	Scaling technique
Facebook OPT	992 x Nvidia A100	Fully-sharded data parallel + Megatron tensor parallel
Eleuther GPT-NeoX	96 x Nvidia A100	ZeRo data parallel + Megatron tensor parallel + Pipeline parallel
Cerebras-GPT	16 x Cerebras CS-2	Data parallel

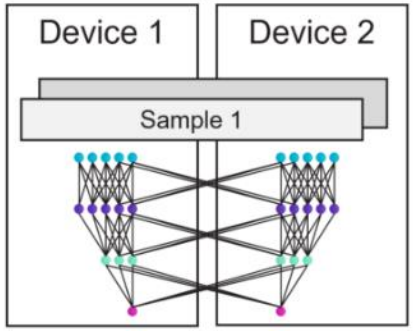
ここで本当に言いたいことは、
CS-2を使うとData Parallelで
簡単に実装できますということ！



Multiple samples at a time
Parameter memory limits



Multiple layers at a time
Communication overhead
N^2 activation memory



Multiple splits at a time
Communication overhead
Complex partitioning

Blog書いています

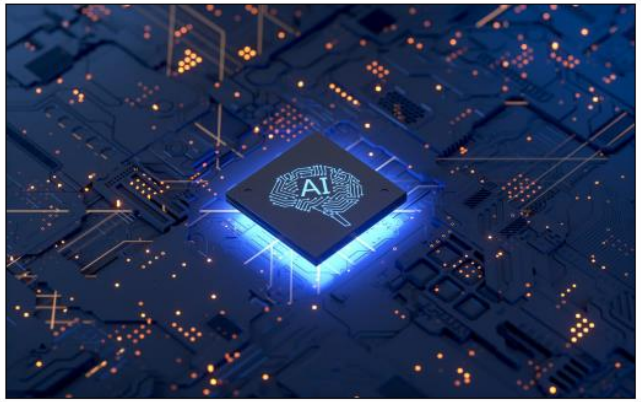
＼Cerebrasに関するブログはこちら／

- Cerebras-GPTを早速使ってみました
- 残念なことに英語だと優秀なのですが、日本語だと支離滅裂な回答をしてきます
- 理由はズバリ、学習データのせいです
- この辺りの詳細はブログにまとめていますので、是非以下のブログをご覧ください
- 実は・・・このブログではNVIDIA DGX A100を使っています

Cerebras-GPTがリリースされました

みなさん、こんにちは。Cerebrasプリセールスエンジニアのnakadaです。
今回は、2023年3月にCerebrasからリリースされたCerebras-GPTについて解説します。

CerebrasGPTでファインチューニング



こんにちは。Cerebrasプリセールスエンジニアのnakadaです。先日、当ブログでCerebras-GPTについてお知らせしました。今回は、そのCerebras-GPTでファインチューニングを実施しましたので結果を共有いたします。

https://cn.teldevice.co.jp/blog/p40369/

ファインチューニングの所要時間例

大規模モデルではファインチューニングにおいても、非常に時間がかかりました
3GBのデータでさえ、**1通り学習するのに1ヶ月以上**かかりました

- 学習済みモデルが公開されているとはいえ、それをそのまま使うことは少ないです
  (通常はドメイン特化のデータを用いてファインチューニングします)

項目	値	備考
モデル	Cerebras-GPT 13B	モデルだけで51.6GB
ハードウェア	NVIDIA DGX A100	A100 (40GB) * 8枚
学習データ	日本語Wikipediaダンプ	3GB、75万サンプル
アーキテクチャ	Data Parallel、DeepSpeed Zero3	
設定	エポック 2、バッチサイズ 8	OOMを考慮し、1GPUあたり1バッチ
性能	45秒/イテレーション	約94000イテレーションで学習が一巡する

Cerebras CS-2のご利用方法

Cerebras AI Model Studio (クラウドサービス)

- スクラッチトレーニング
- ファインチューニング
- 国内CSP様との協業も進行中

オンプレミス

- CS-2はたった1台 (1ラック) で高速なパフォーマンスを提供します
- さらにはCS-2を16台並べた **Andromeda** と呼ばれるクラスタ構成もあります
- 国内に設置することでセキュリティリスクの低減も可能です

Cerebras AI Model Studio のリリース

https://www.cerebras.net/press-release/cerebras-systems-and-cirrascale-cloud-services-introduce-cerebras-ai-model-studio-to-train-gpt-class-models-with-8x-faster-time-to-accuracy-at-half-the-price-of-traditional-cloud-providers

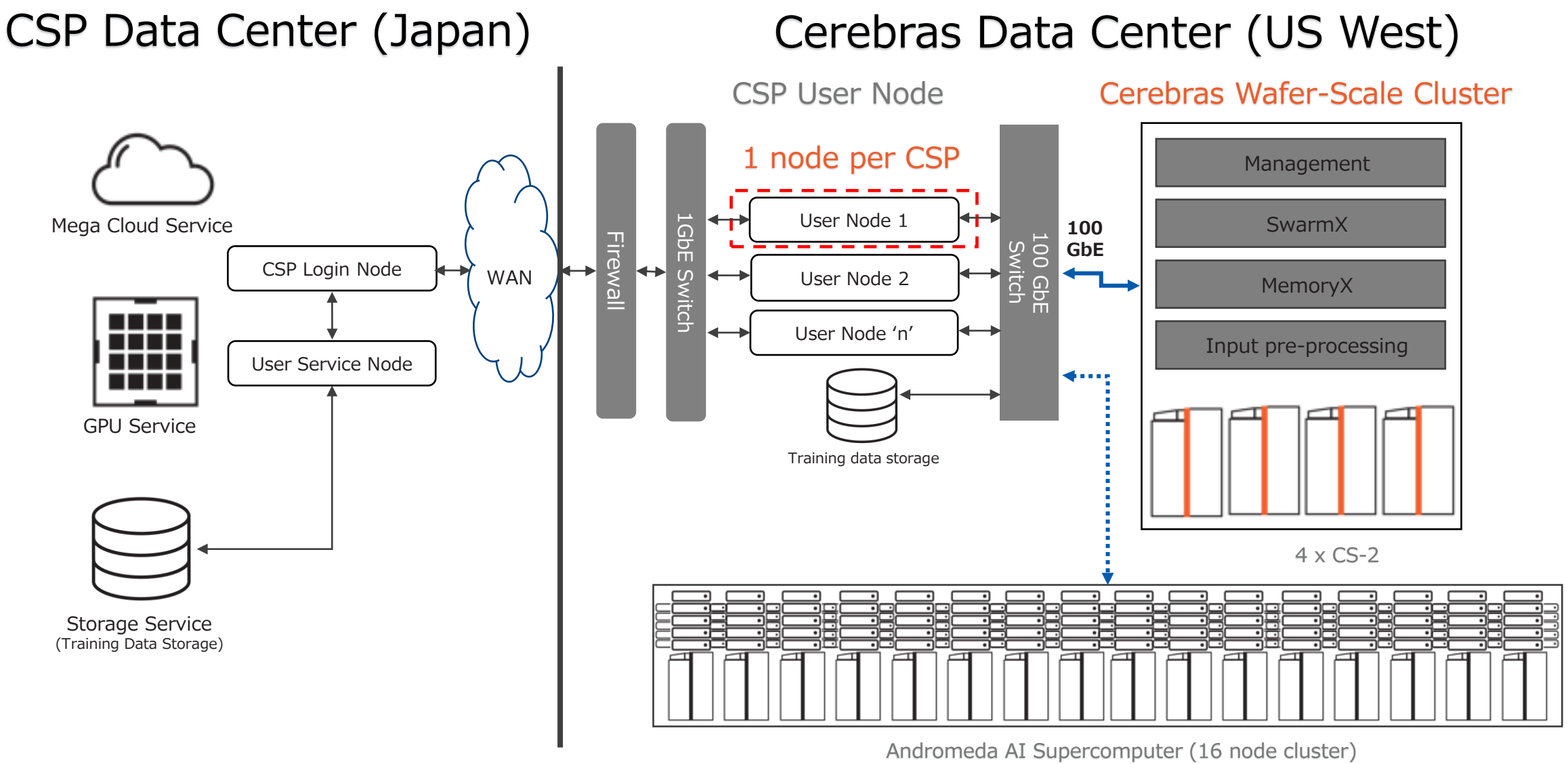
Cerebras AI Model Studio について

- AI Model Studio は何日使ったか、ではなくどのモデルを学習したか、で価格が決まります
- この表は自社のデータセットで**ゼロから学習させたい**ユーザーがターゲット
- パブリッククラウド利用比較で半額、8倍の学習パフォーマンス
- これとは別にファインチューニング用の価格表もあります (ここでは割愛)

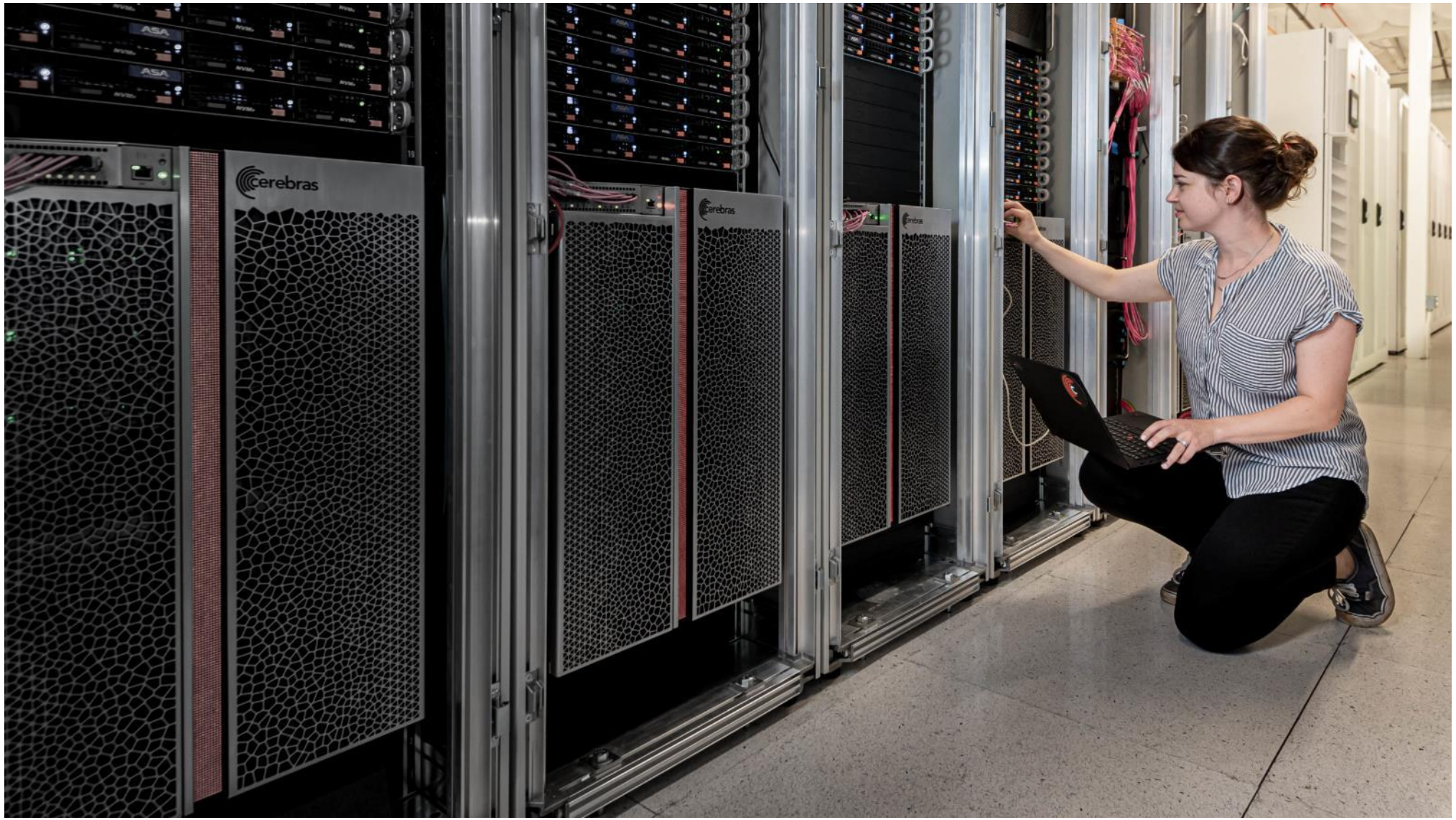
モデル	パラメータ数	学習に必要なトークン数	想定日数	価格
GPT-3 XL	13 億	260 億	0.4 日	¥ 412,000
GPT-J	60 億	1,200 億	8 日	¥ 7,425,000
GPT-3 13B	130 億	2,600 億	39 日	¥ 24,750,000
GPT 70B	700 億	14,000 億	個別見積り	個別見積り
GPT 175B	1,750 億	35,000 億	個別見積り	個別見積り

＊ 学習用の価格表から抜粋

国内CSP経由の接続を近日サポート予定

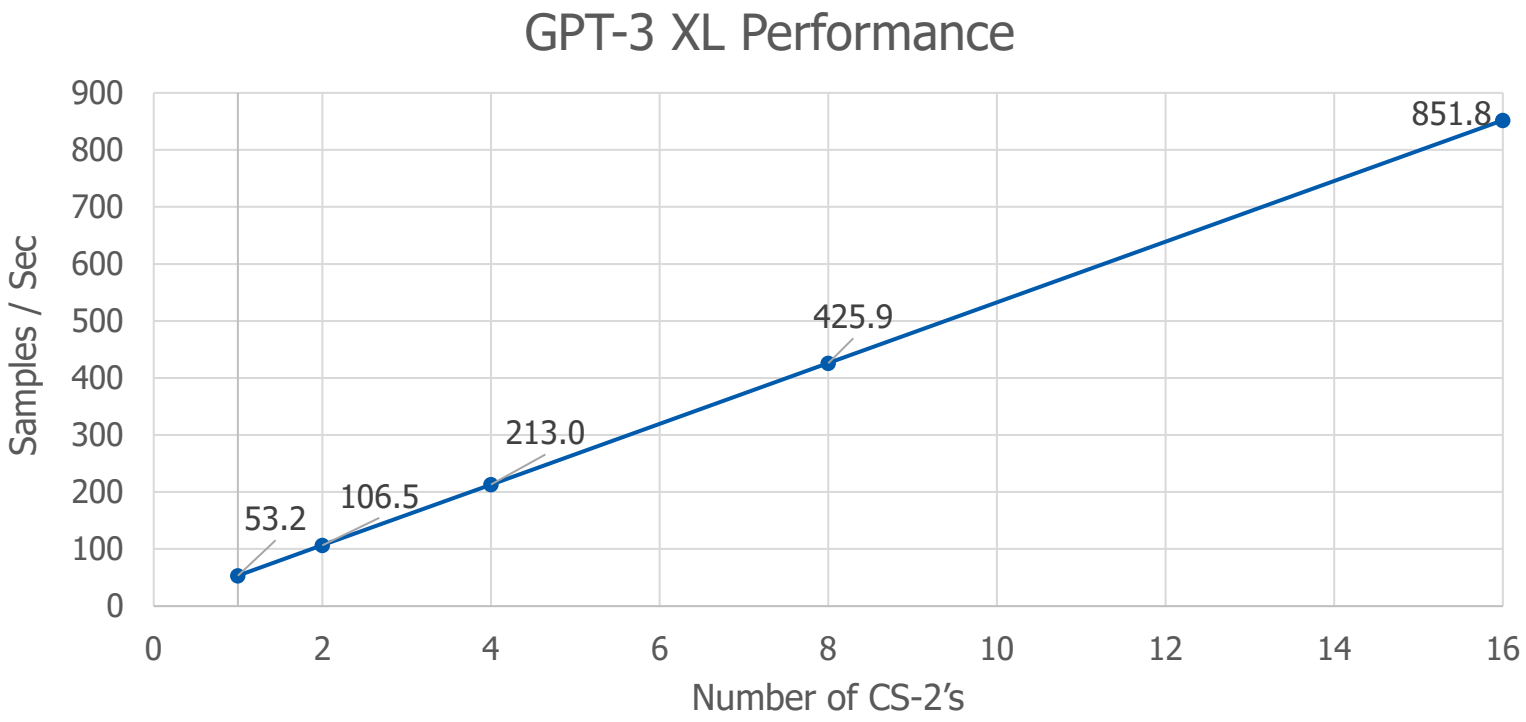


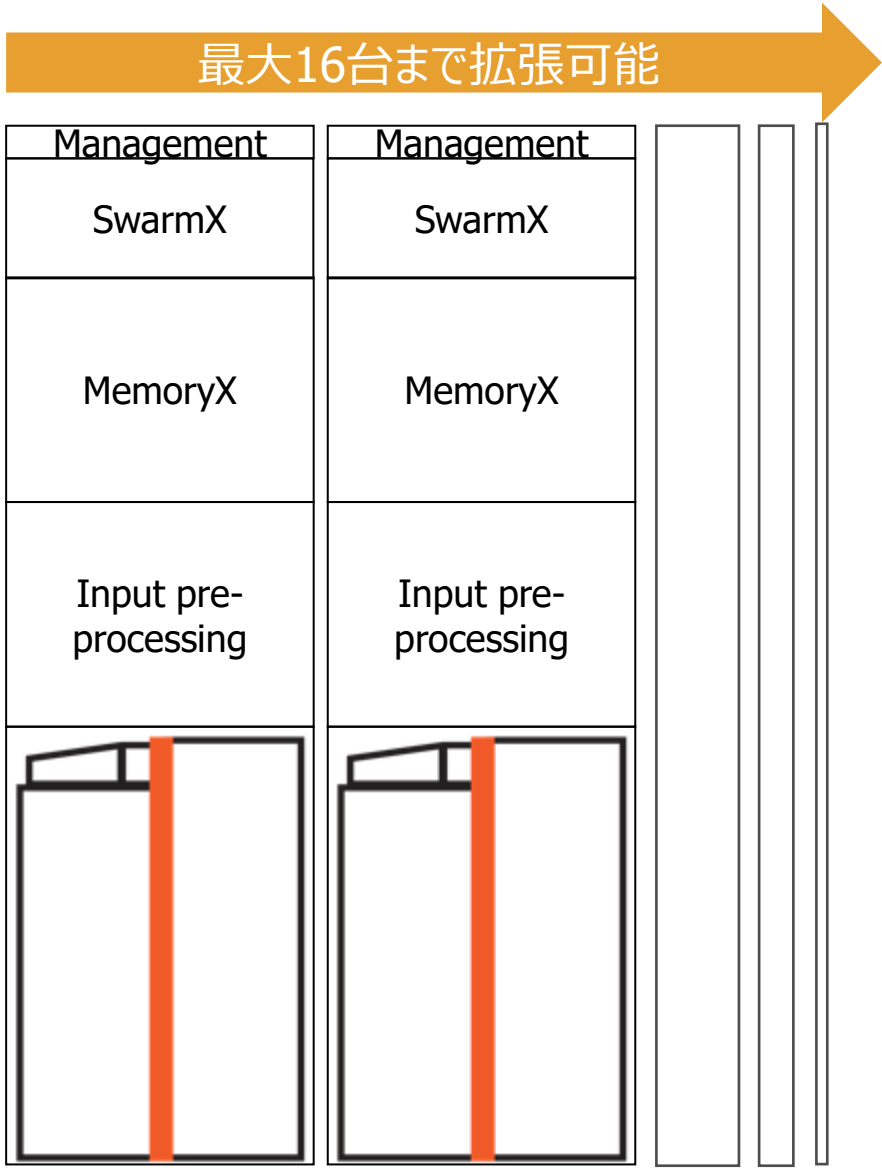
機器のイメージ：1台～最大16台まで拡張可能



オンプレで持ちたいお客様向け

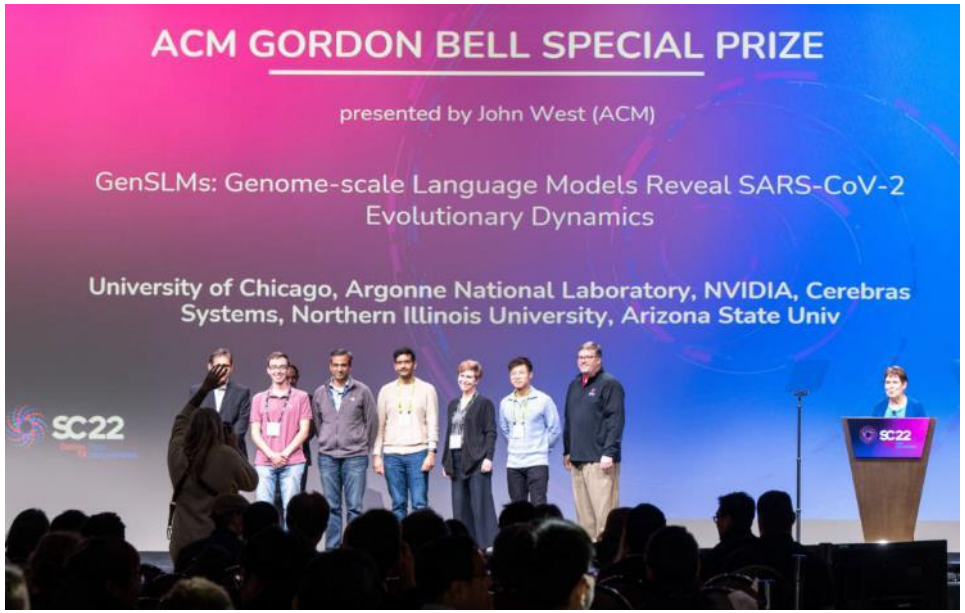
Cerebras CS-2は拡張すれば**リニアに性能向上**
(ChatGPTクラスのモデルも学習可能)

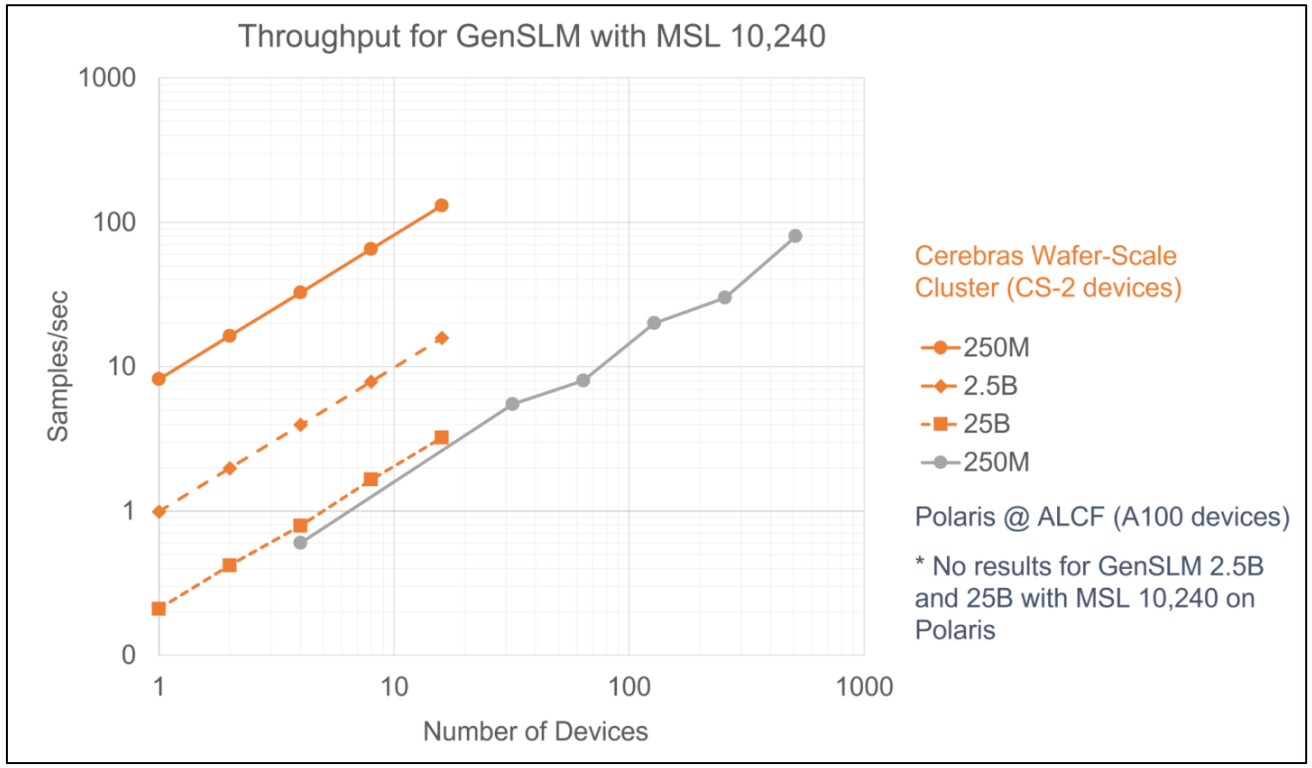




参考：ゴードンベル特別賞について

- SC22にて米国アルゴンヌ国立研究所、NVIDIA、Cerebras等の共同論文が受賞
- Max Sequence Length が10,240に達するデータの学習に成功





＊ Cerebrasウェハースケールクラスタ上でのMSL 10,240による
GPT-Jの25Bパラメータまでのリニアスループットスケーリング

小さくスタートしたいお客様向け

NVIDIA最新世代H100搭載：TED GPUサーバーパッケージ（SuperMicro社製）

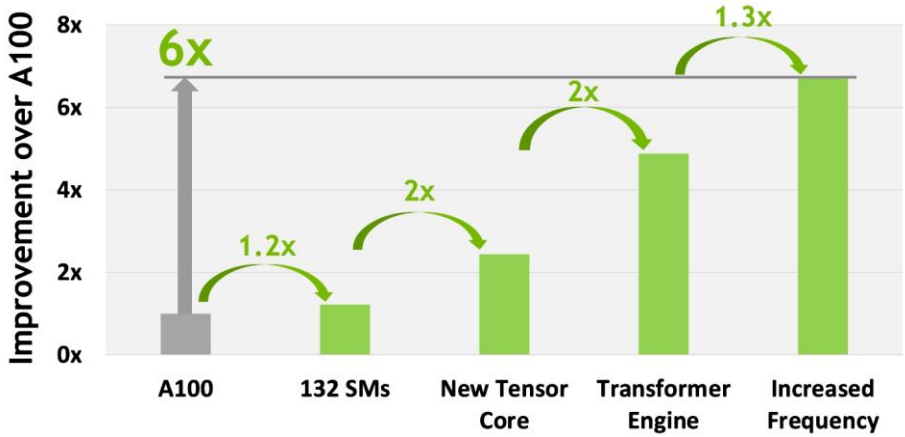
特徴

- NVIDIA最新世代H100を搭載
- 大規模化するデータセット環境下でも高速な分析を実現
- 注目のChatGPTを始めとする言語解析分野で高パフォーマンスを提供

筐体イメージ

こんな人におすすめ

- NVIDIAの最新世代GPUでベンチマーク検証したい
- 大量の言語解析を行う必要がある



今だけ！

パッケージご購入いただいた方限定でNGC活用ユーザートレーニングを特価にて提供いたします（詳細は担当営業にお問い合わせください）

必要に応じて、DGX/DGX SuperPODまで拡張可能

東京エレクトロンデバイスのAI関連サービスご紹介

TED AI Lab

- AIアクセラレータの販売
  - Cerebras CS-2
  - NVIDIA製品(H100 GPU, DGX H100, DGX Superpod)
- TAIP (TED AI Infrastructure Package)
  - AIアクセラレータ+ストレージ+ネットワーク機器をパッケージとして販売・構築・保守
- TAILES (TED AI Lab Engineering Service)
  - TED AI Labを使って評価をしていただくための技術サポートを提供
  - スタートアップセッション⇒評価⇒Q&A対応⇒クロージングセッション

まとめ：東京エレクトロンデバイスからのご提案

Cerebras-GPT with NVIDIA GPUs

- サーバーパッケージやDGX、SuperPODなどモデル規模に合わせて最適ハードウェアをサポート込みで提供
- 初期セットアップやNVIDIAのコンテナライブラリ「NGC」簡易トレーニングも実施可能

Cerebras GPT with AI Model Studio

- 国内CSP経由で利用可能（学習は米国インフラ）
- Cerebras/TEDエンジニアが学習をサポートする「Fine tuningオプション」も準備済み

Cerebras Wafer Scale Cluster

- DGX SuperPODクラスの学習基盤をワンラックでご提供

東京エレクトロンデバイス：取り扱い製品一覧

Connect Beyond

Security

テレワーク/クラウドアクセス関連ソリューション

SDP	SWG	CASB
SSE/SASE		
netskope		

エンドポイント	シークレット管理	HSM
Active EDR		
SentinelOne	HashiCorp Vault	ENTRUST THALES

社内/トラストネットワーク関連ソリューション

Wi-Fi	VPN	Firewall	WAF
Cognitive Wi-Fi ARISTA Extreme networks	f5	paloalto NETWORKS FORTINET	f5 CLOUDFLARE

DNS/DHCP	Cloud Network Platform
Infoblox NEXT LEVEL NETWORKING	CDN、IDS、FW、WAF、SandBox、Proxy、VPN CLOUDFLARE

脆弱性対策

セキュリティ検証	脆弱性管理
Automated Security Validation PENTERA	脆弱性自動診断 Balbix

セキュリティ運用

SIEM	SOAR
LogRhythm	splunk>phantom

Infrastructure

クラウド管理/実行

CSPM	IaC
netskope	HashiCorp Terraform

ネットワークソリューション

IP Clos	L2/L3スイッチ		ADC
JUNIPER NETWORKS	DCNW+NDR ARISTA NVIDIA	キャンパス Extreme networks	f5 citrix

AI/DLソリューション

GPU	GPU System	Accelerator
NVIDIA	AIRI AI-READY INFRASTRUCTURE	cerebras

仮想マシンインフラ

HCI	3Tier
NUTANIX	PURESTORAGE

ファイルストレージソリューション

Scale Out		Scale Up
Power Scale DELL Technologies	PURESTORAGE	Unity XT DELL Technologies

バックアップソリューション

Cloud Backup	
rubrik	VEEAM

その他


Q&A

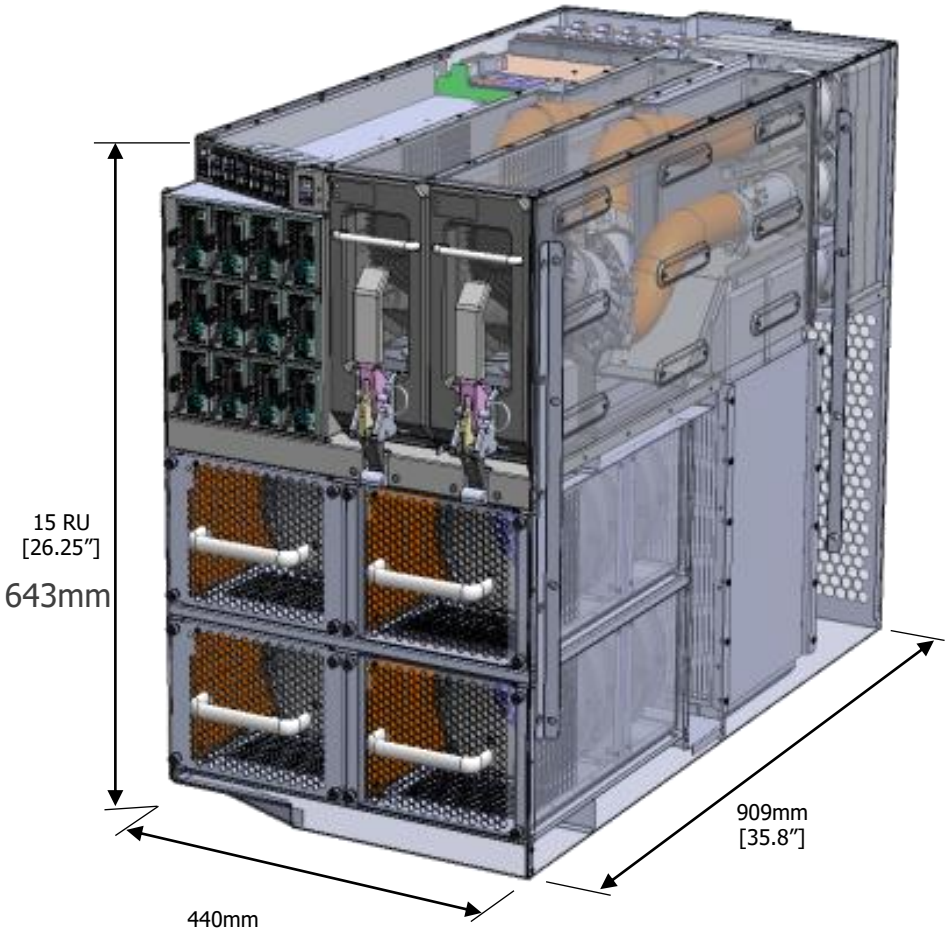
onnect Beyond

backup

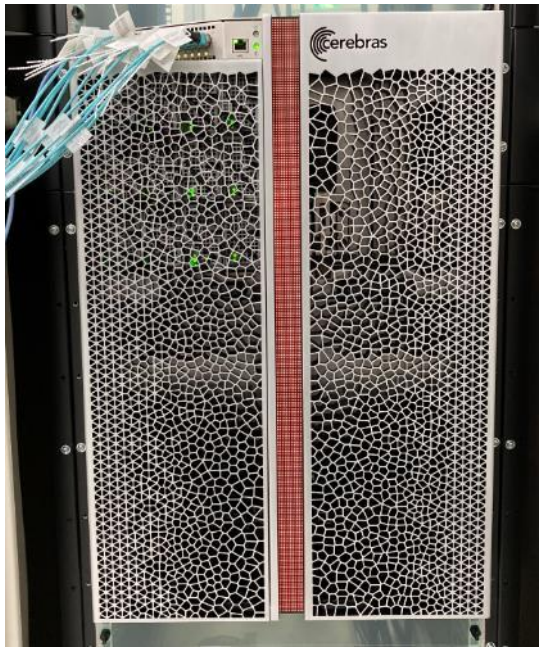
Cerebras Systems 会社概要

創業	2016年4月
米国本社	**サニーベール カリフォルニア USA**
エンジニアリングオフィス	サン・ディエゴ カリフォルニア USA
カナダテクノロジーセンター	オンタリオ カナダ
日本法人	**セレブラス・システムズ合同会社**
総調達資金	7.2億ドル超
従業員数	480 (90%以上がエンジニア)
主要製品	CS-2
テクノロジー	Wafer Scale Engine 2 (WSE2)
取得済・出願中特許	80
日本販売代理店	**東京エレクトロン デバイス株式会社**
代表事例顧客	アルゴンヌ国立研究所、国立科学財団 ピッツバーグ・スーパー・コンピューティングセンター ローレンスリバモア国立研究所、エジンバラ大学パラレルコンピューティングセンター 国立エネルギー研究所、グラクソ・スミス・クライン、アストラゼネカ、トタルエナジーズ

CS-2製品外観とハードウェア仕様(左図は空冷仕様品)

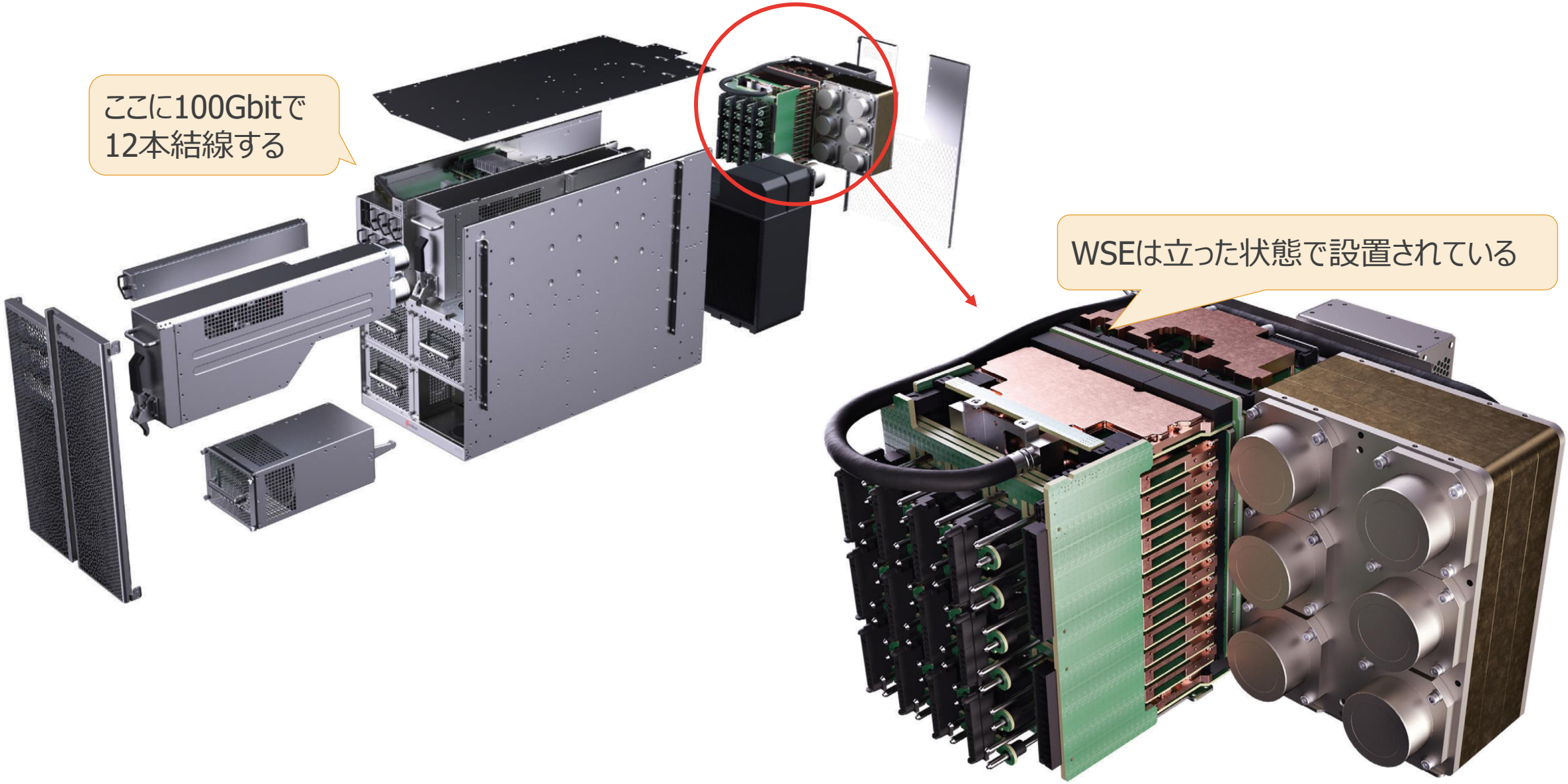


サイズ	15 Rack Units
重量	254 kg (≒17 kg/RU)
最大消費電力	23kW
Process	TSMC 7nm
冷却方式	水冷または空冷
システムIO	1.2 Tb/s (12 x 100 Gig Ethernet)



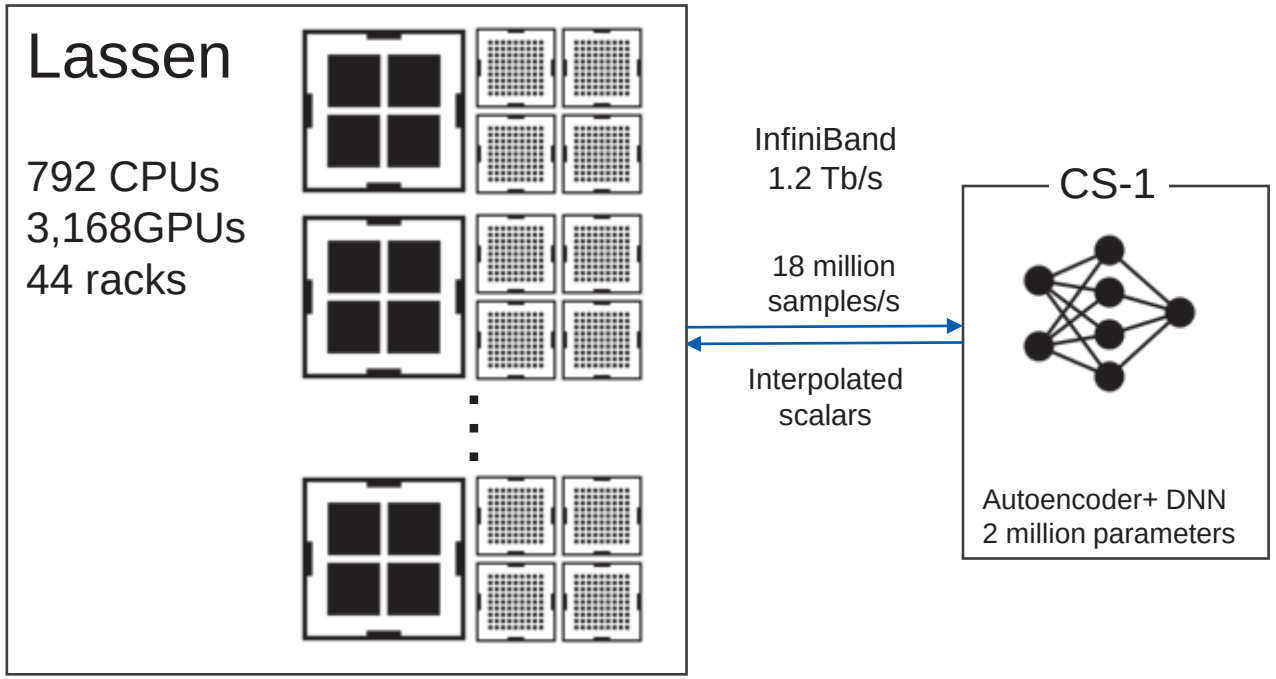
当社AIラボに設置されている実物(正面と背面)

筐体3Dイメージ



ローレンス・リバモア国立研究所: HPCを加速させるAIサロゲートモデル

AI + HPCワークロードのコンバージド化に向けた
ヘテロジニアスシステムレベルの最適化

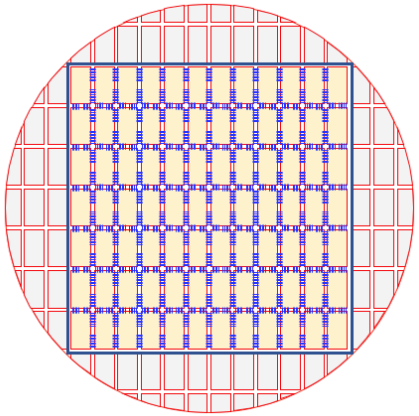


20時間以内に稼働可能
CS-1 (Lassen GPUの64倍の性能) は全く新しい実験への扉を開く
＊ ここでは旧世代バージョンのCS-1を使用しています

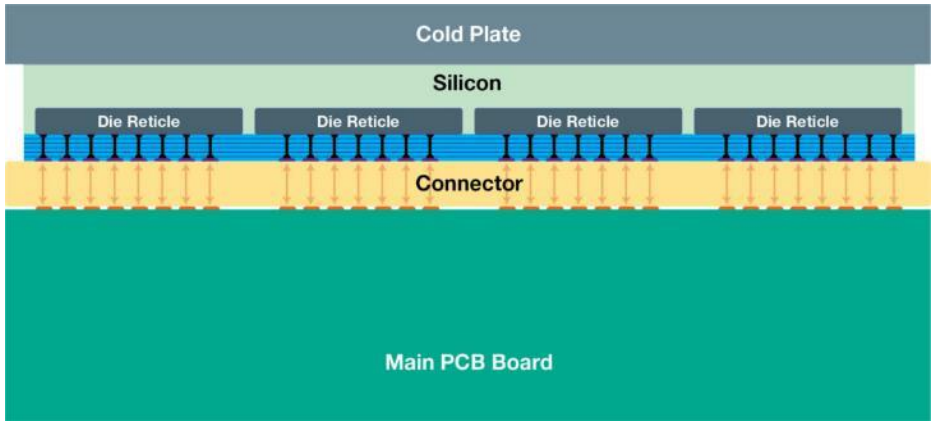
WSEを実現する上での課題

- 大きなチップを作ったは良いけど、大きな課題がありました

1. スクラブラインを超えた接続
2. 歩留まり
3. 熱膨張
4. 電力と冷却



数十万本のオンシリコン
ワイヤーでダイを接続



熱膨張に対応するために
様々な部品を独自に開発

- 詳細をと知りたい方はCerebras Systems社のホームページをご覧ください

https://cerebras.net/blog/wafer-scale-processors-the-time-has-come/

WSEを実現する上での課題 ～解決編～

- 各課題は以下のようにして解決されました

1. Reticle間の接続

   TSMCのInFO_SOW*プロセスによって回路面の上部に配線層を形成

   従来型Flip-Chip MCMに比べ配線密度2倍, PDN*インピーダンス0.03倍

2. 歩留まりへの対処

   コア間の接続経路を冗長化して不良コアをスキップ

3. 熱膨張率の差異への対処

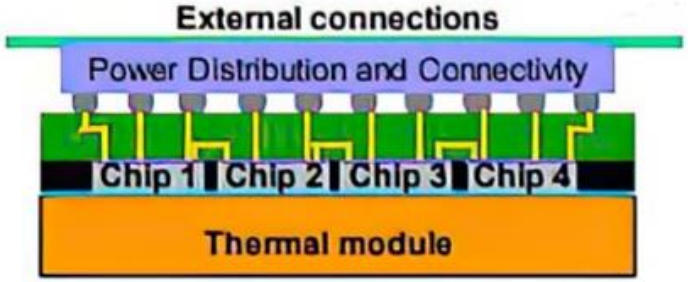
   特殊な2D膜状コネクターを開発し熱膨張の差を吸収

4. 電力供給と熱設計

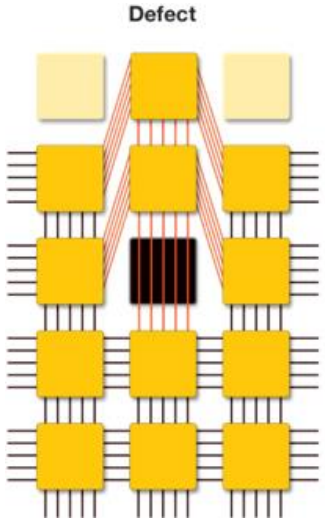
   回路面から電力を供給

   シリコン基板の裏側にヒートシンクを設置し水冷で冷却

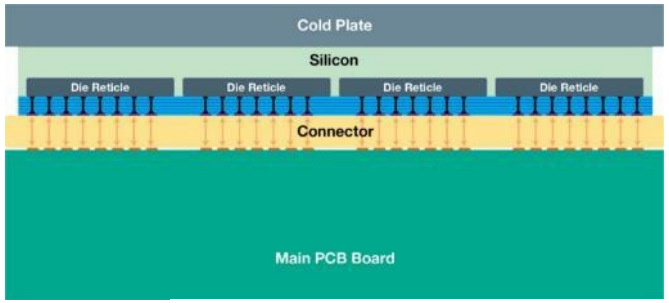
①InFO_SOW実装技術



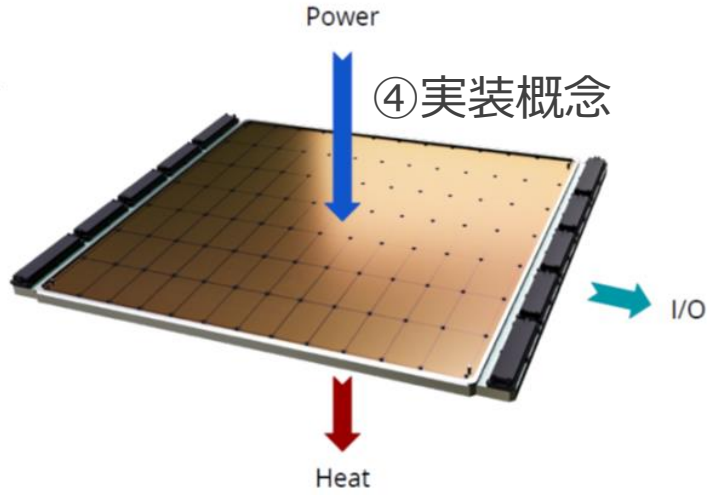
②配線冗長化



③特殊2D膜状コネクター



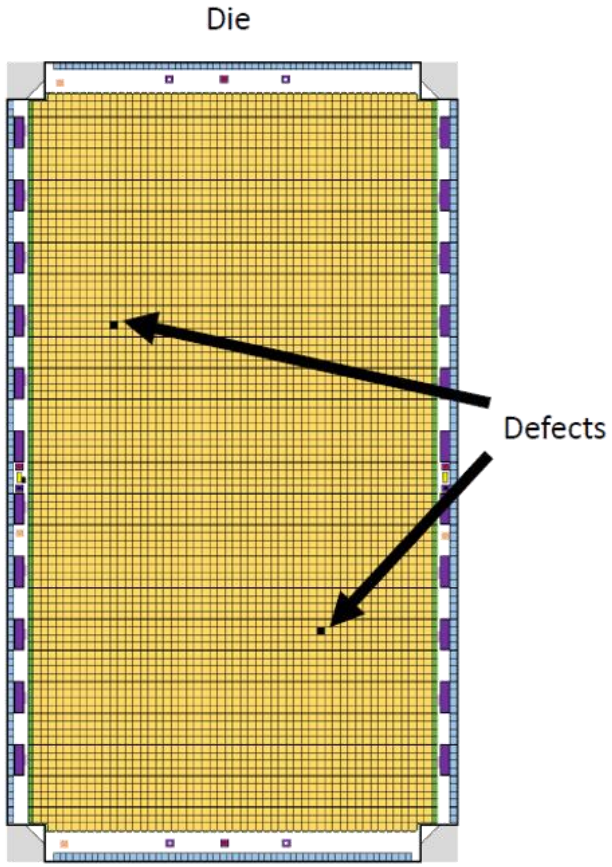
④実装概念

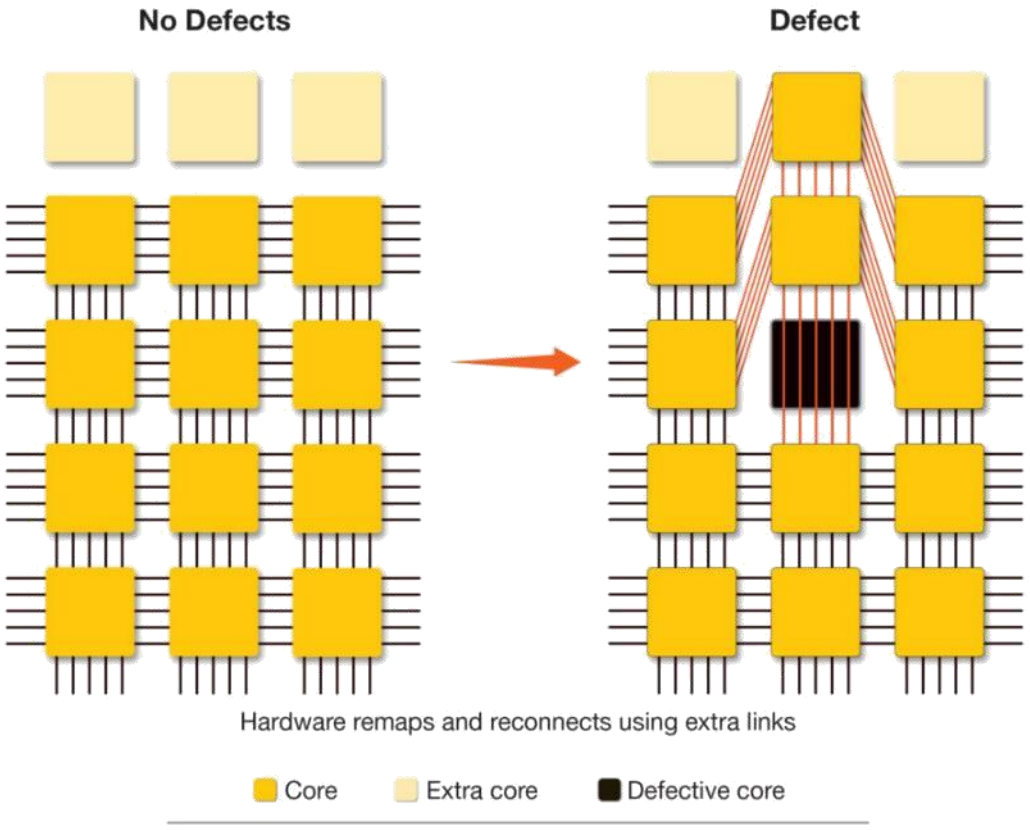


*)InFO_SOW: Integrated Fan-Out System on Wafer
*)PDN: Power Distribution Network

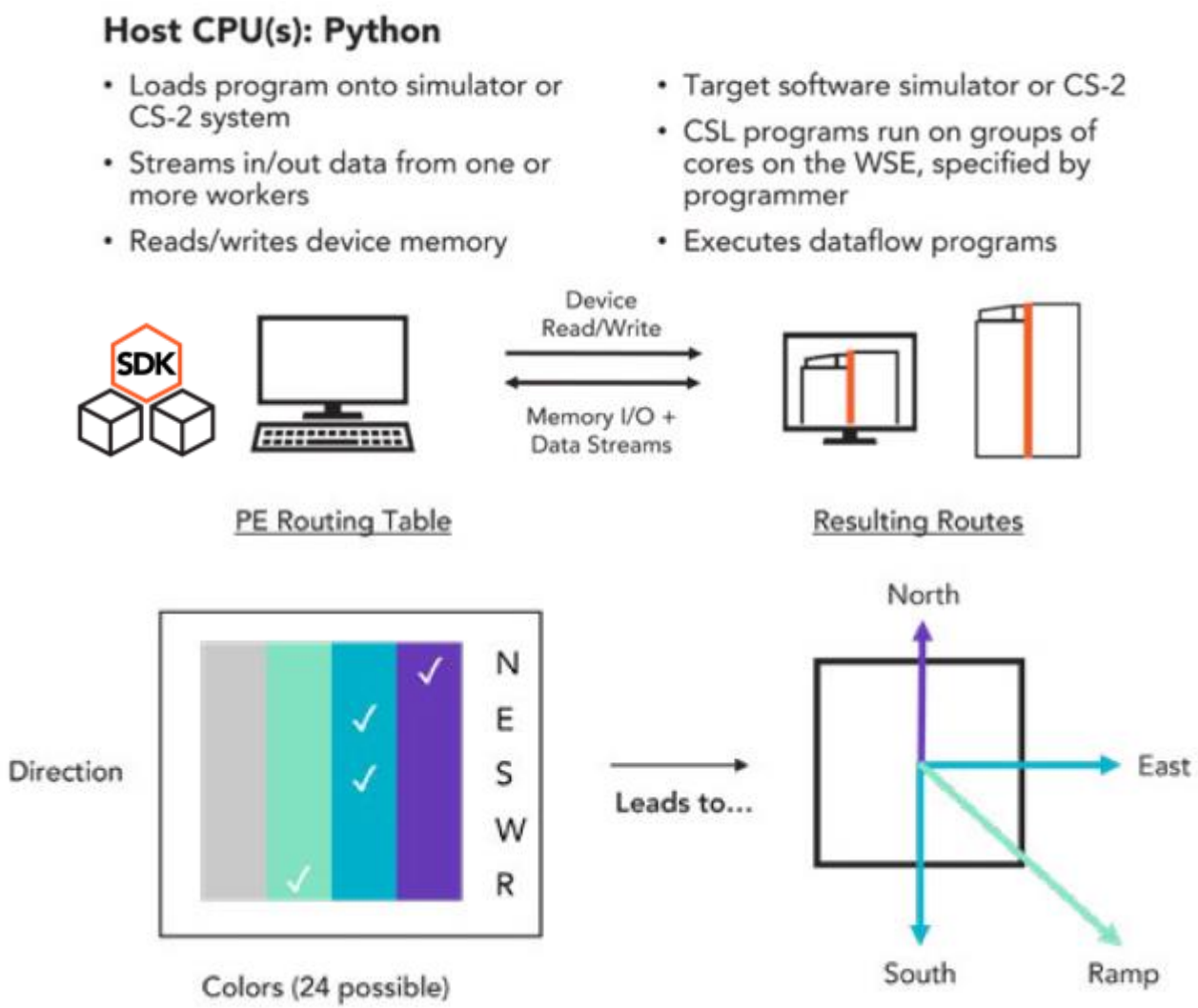
歩留まり改善に向けた取り組みの補足

- どんなに頑張ってもいくつかのコアは欠陥となってしまいます
- そのため予備のコアを予め組み込んでおいて欠陥を避けて通るような仕組みを作成しています





Cerebras SDK



- CS-2をHPC用途で使用されるユーザー向けにSDKが提供されています
  - 2023年1月現在(β版 0.6.0)
- SDKではC言語をベースにしたCSL (Cerebras Software Language)でローレベルのデータフロープログラミングを行うことができます
- SDKはx86上で動作する下記機能を含みます
  - ドキュメント (チュートリアル、スタートガイド)
  - コンパイラ、デバッガ、シミュレータ
  - サンプルコード
- CSLでプログラミングされたコードはコンパイラで実行コードにコンパイルされ、x86上のシミュレータあるいはCS-2実機で実行することができます

Cerebras-GPT パラメータ情報

Model	Parameters	Layers	d_model	Heads	d_head	d_ffn	LR	BS (seq)	BS (tokens)
Cerebras-GPT	111M	10	768	12	64	3072	6.0E-04	120	246K
Cerebras-GPT	256M	14	1088	17	64	4352	6.0E-04	264	541K
Cerebras-GPT	590M	18	1536	12	128	6144	2.0E-04	264	541K
Cerebras-GPT	1.3B	24	2048	16	128	8192	2.0E-04	528	1.08M
Cerebras-GPT	2.7B	32	2560	20	128	10240	2.0E-04	528	1.08M
Cerebras-GPT	6.7B	32	4096	32	128	16384	1.2E-04	1040	2.13M
Cerebras-GPT	13B	40	5120	40	128	20480	1.2E-04	720 → 1080	1.47M → 2.21M

Model Details

- Developed by: Cerebras Systems
- License: Apache 2.0
- Model type: Transformer-based Language Model
- Architecture: GPT-3 style architecture
- Data set: The Pile
- Tokenizer: Byte Pair Encoding
- Vocabulary Size: 50257
- Sequence Length: 2048
- Optimizer: AdamW, (β1, β2) = (0.9, 0.95), adam_eps = 1e−8 (1e−9 for larger models)
- Positional Encoding: Learned
- Language: English
- Learn more: Dense Scaling Laws Paper for training procedure, config files, and details on how to use.