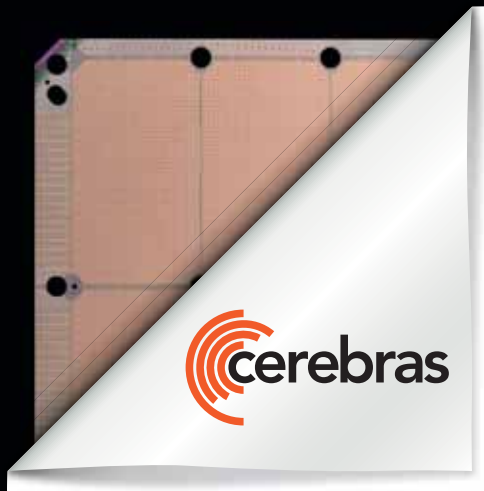


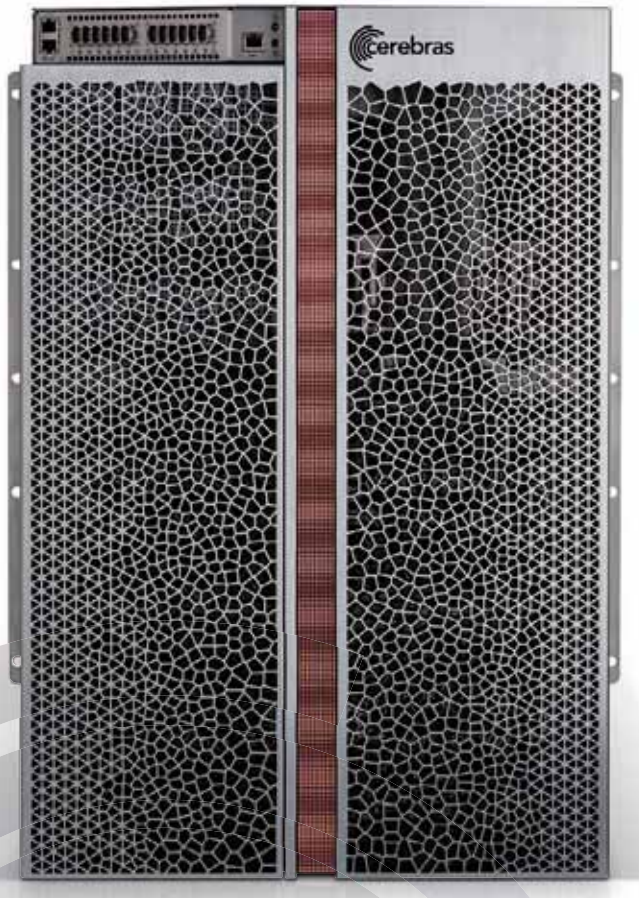
Cerebras CS-2で高速AI処理を実現

Cerebras CS-2の能力

85万コアを実装したウェハースケールエンジンチップ(WSE)

- 40GBの“オンチップ”メモリー(SRAM)
- 20PB/sのメモリーバンド帯域幅
- コア間220Pb/sのファブリック帯域幅

1 チップにAIモデルがロードでき、
超高速AI処理を実現



一般的なGPUでの高速処理の実現方法

学習時間短縮のため、
GPUの台数を増やすこと(スケールアウト)で解決したいが…

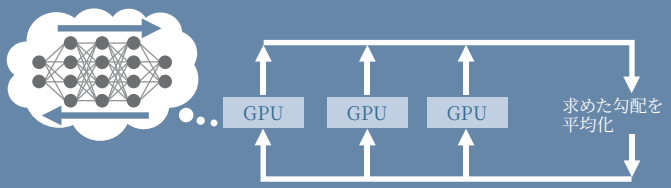
» 台数増加に伴いハードウェア・ソフトウェア設定も増加

» スループットは向上するが
Loss収束までの時間が短縮するとは限らない。

大規模AIモデル学習にCerebras CS-2を
利用することで、複数GPUの分散処理を考慮することなく、
AIモデル学習の収束が高速化可能に

なぜ、Loss収束速度が上がらないのか？

複数GPUでAIモデルをデータパラレルで学習させる際、ミニバッチ単位でデータを学習しLoss値を減らすためにGPU毎で勾配を求め、GPU毎の勾配計算結果を一度平均化する必要があります。



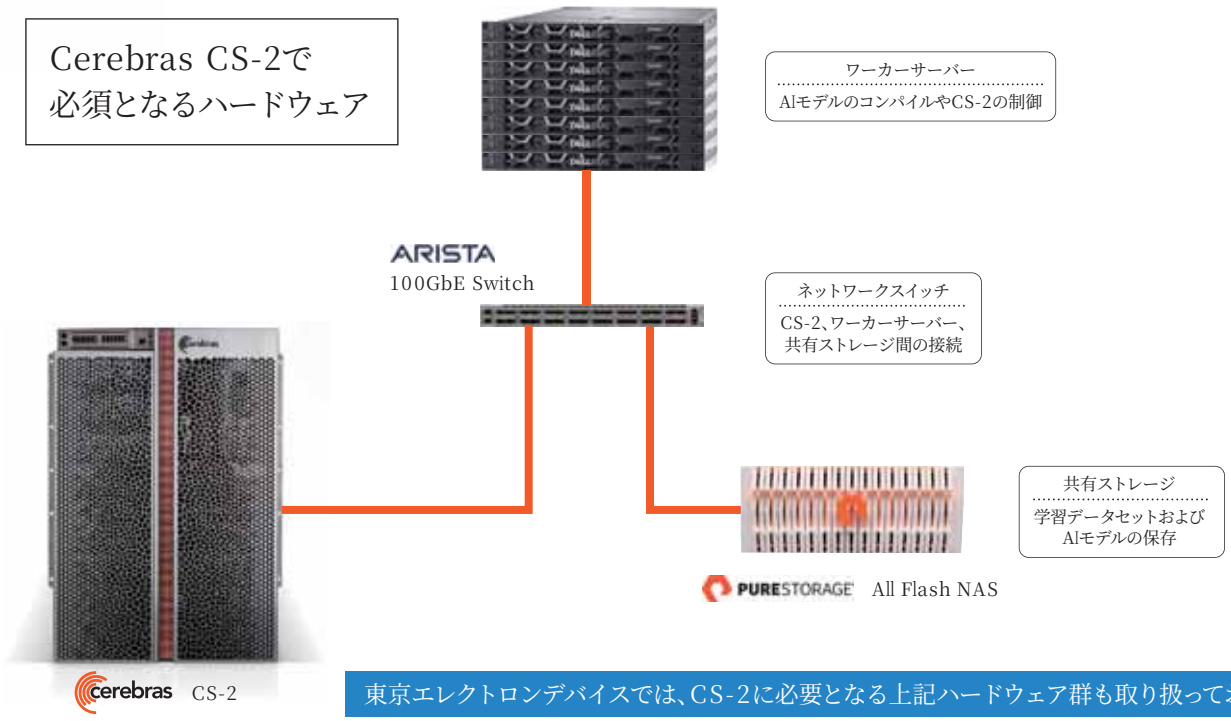
この処理により、GPU個数、データセット内容(データの偏り)によっては、Lossの収束時間に影響が出てしまう可能性がありますが、CS-2は1チップで学習可能なため、複雑な分散処理に必要な設定等を考慮する必要が無く、AIモデルのLossをより早く収束することが可能になります。

実際の結果は

ホワイトペーパー
金融サービスアプリケーションの
NLPモデルトレーニングの高速化と高精度化
をご覧ください。

Cerebras CS-2で高速AI処理を実現

Cerebras CS-2で必須となるハードウェア



東京エレクトロンデバイスでは、CS-2に必要となる上記ハードウェア群も取り扱っております

クラウド利用 VS オンプレミス

AI開発のゴールは、AIモデルの高精度実現です。

精度実現には
グリッドサーチ手法などによるハイパーパラメータの調整が必要

膨大なAI学習時間が費やされ
AI学習中にコンピュータリソースを利用し続ける必要があります。

開発費削減のためのクラウド利用のはずが思ったより削減できていないという声もあります。
オンプレ環境では利用時間に縛られることなく、自由にAI開発が可能になります。

東京エレクトロン デバイス株式会社
CN BU https://cn.teldevice.co.jp

新宿:〒163-1034
東京都新宿区西新宿3-7-1 新宿パークタワーS34階
Tel.03-5908-1990 Fax.03-5908-1991

大阪:〒540-6033
大阪府大阪市中央区城見1-2-27 クリスタルタワー33階
Tel.06-4792-1908 Fax.06-6945-8581

名古屋:〒451-0045
愛知県名古屋市西区名駅2-27-8 名古屋プライムセントラルタワー8階
Tel.052-562-0826 Fax. 052-561-5382

つくば:〒305-0033
茨城県つくば市東新井15-4 関友つくばビル7階
Tel.029-848-6030 Fax.029-848-6035

お問い合わせは、Webサイトの右記フォームよりお願いします。 https://cn.teldevice.co.jp/product/cerebras/form.html