SambaNova SYSTEMS

MPLS 2023

生成AIを加速する データフローアーキテクチャ

SambaNova Systems
Head of Marketing for APAC
林 憲一
Kenichi.Hayashi@SambaNova.ai

自己紹介：林 憲一（1967年5月東京生まれ）

SambaNova Systemsアジア太平洋地域マーケティング責任者

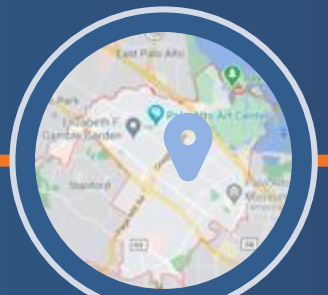
- 信州大学社会基盤研究所特任教授
- 略歴
  - 1991年　東京大学工学部計数工学科卒
  - 1991年 ～ 1998年 富士通研究所及び富士通でスパコンの研究・開発
  - 1998年 ～ 2004年 米Sun Microsystemsでエンジニアからマーケティング
  - 2004年 ～ 2005年 米エンジニアス・ソフトウェアでマーケティング部長
  - 2005年 ～ 2006年 子育て休業
  - 2006年 ～ 2010年 マイクロソフトでWindows Server製品マーケティング
  - 2010年 ～ 2019年 NVIDIAでエンタープライズマーケティング本部長
  - 2019年 ～ 2023年 日本ディープラーニング協会
  - 2020年 ～ 信州大学社会基盤研究所特任教授
  - 2022年 ～ SambaNova Systemsでアジア太平洋地域マーケティング責任者

SambaNova SYSTEMS

SambaNova Systemsについて

2017
創業

パロアルト
オースティン、
ロンドン、東京



ML/AI
ソフトウェア定義型
ハードウェア

500+
HW/SWAIエンジニア

スタンフォード大学
EE/CS教授
Kunle Olukotun

CEO
Rodrigo Liang

スタンフォード大学
コンピュータ科学教授
Chris Ré

SoftBank Investment Advisers
BLACKROCK
WALDEN International
G/
intel Capital
TEMASEK
GIC
REDLINE Capital Management
ATLANTIC BRIDGE
Celesta
Micron
SAMSUNG CATALYST FUND
SK telecom

シリーズDまでに11億ドル以上を調達

SambaNova SYSTEMS

AI新時代：ファンデーション（基盤）モデル

GPU（ノイマン型）が
最も苦手とする計算
⇓
非ノイマン型計算機
の必要性

基盤モデル：大規模スパース演算
2020年代

深層学習：小規模密行列の並列演算
2010年代

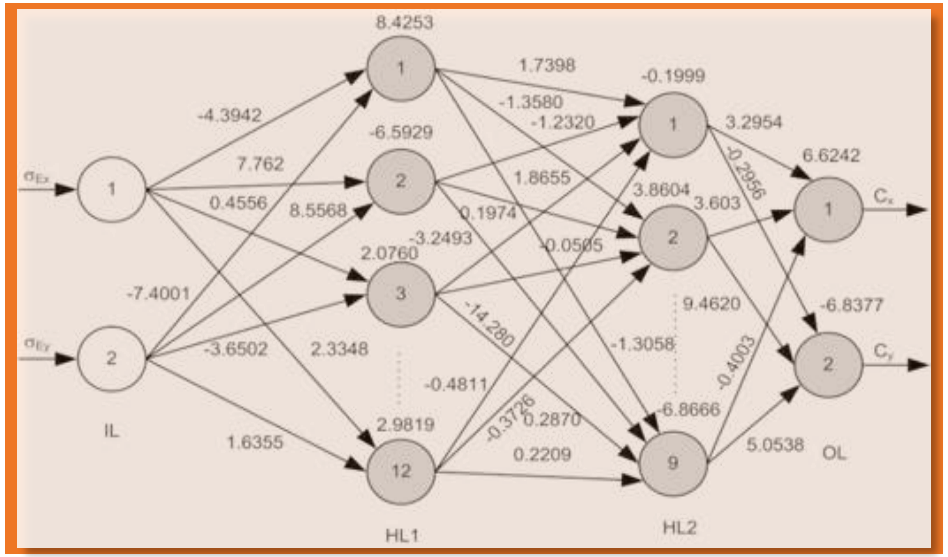
機械学習
2000年代

GPU（ノイマン型）
が最も得意とする計算
⇓
GPU=AIという誤解

SambaNova SYSTEMS

現代AIはデータフローの問題

```
#include <iostream>
using namespace std;

int _tmain (int argc, _TCHAR* argv[])
{

     int iVal1 = 0, iVal2 = 0, iVal3 = 0;

      printf("Enter three numbers:");
      scanf("%d %d %d", &iVal1, &iVal2, &iVal3);

     if (iVal1 >= iVal2)
      {
          if(iVal1 >= iVal3)
            printf("Largest number = %.2d", iVal1);
          else
            printf("Largest number = %.2d", iVal3);
      }
      else
      {
          if(iVal2 >= iVal3)
            printf("Largest number = %.2d", iVal2);
          else
            printf("Largest number = %.2d", iVal3);
      }

 getchar ();
 return 0;
}
```



ソフトウェア 1.0

- コードで書かれている (C++, ...)
- ドメインの専門知識が必要

現代AIはデータフロー（ソフトウェア2.0）

- コードではなくデータがモデルを鍛える
- ニューラルネットワークの重みで記述

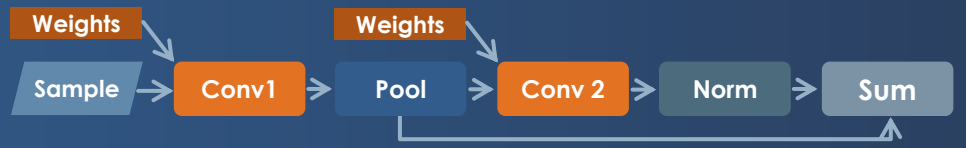
Andrej Karpathy. Scaled ML 2018 talk

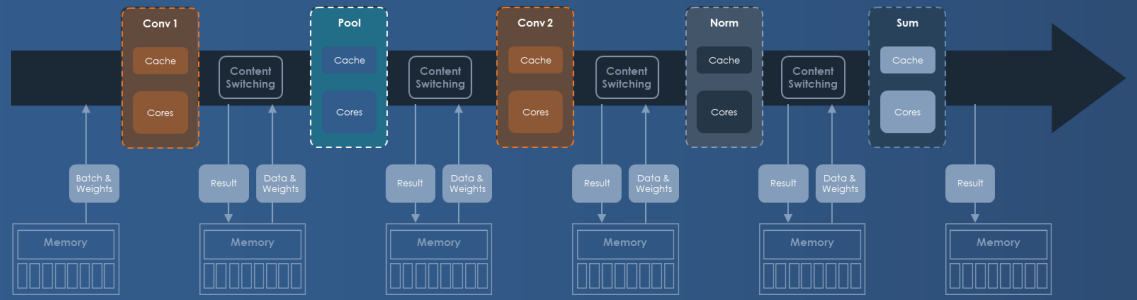
SambaNova SYSTEMS

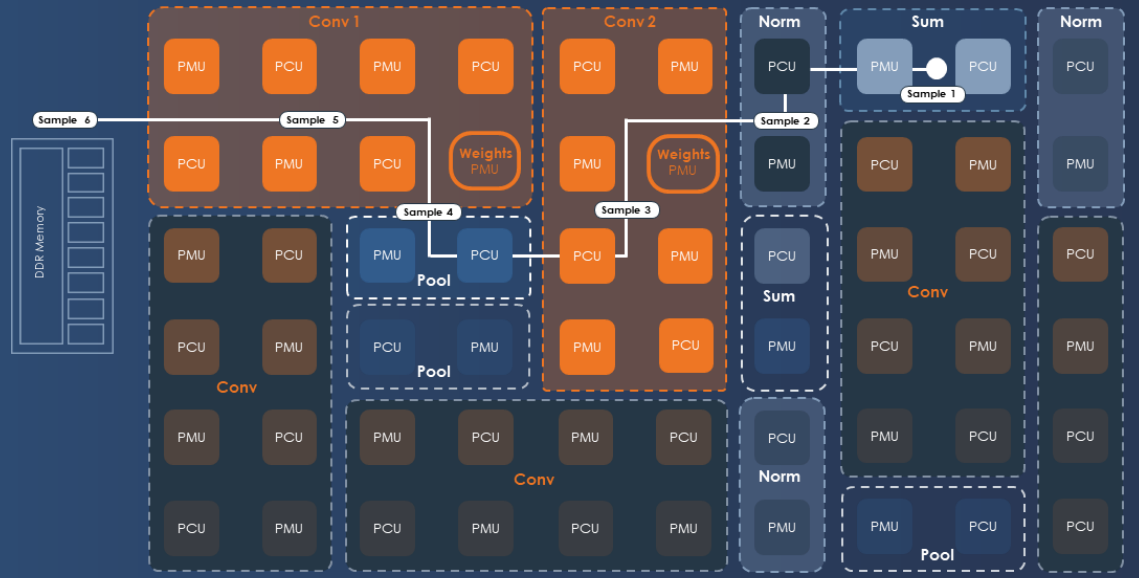
従来手法とデータフローの違い

データフローアーキテクチャは局所性と並列性を活用

畳み込みグラフ







計算カーネル毎にメモリアクセスが必要なため、高速メモリを演算器の近くに配置する必要 → **高速メモリは容量が小さい**

外部メモリアクセスが最小化できるので、高速メモリが不要 → **メモリ大容量化可能**
データが移動しなければ演算が発生しない
→ **スパース性にも強い**

SambaNova SYSTEMS

SambaNovaはフルスタックAIプラットフォーム

SambaNovaはAIスタックの全てのレイヤーでイノベーションを実現

アズ・ア・サービス	事前学習済みのSOTAモデル
システム	統合システムにより、すぐに使える最適化
ソフトウェア	データフローコンパイラ
半導体	再構成可能なデータフローアーキテクチャ

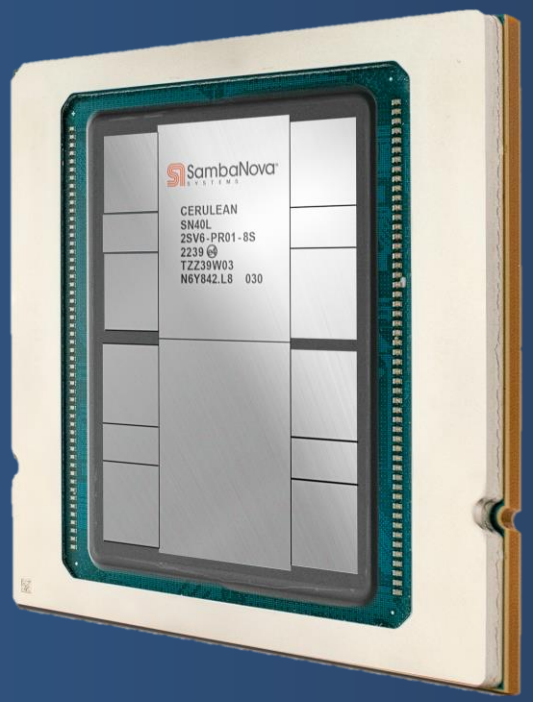
SambaNova SYSTEMS

SN40L 言語モデルに最適化された新しいSambaNova RDU

再構成可能なデータフローユニット

SN40L RDU

- TSMC 5nmテクノロジー
- 1,026億トランジスタ
- 1,040個のRDUコア
- 638TFLOPS (bf16)

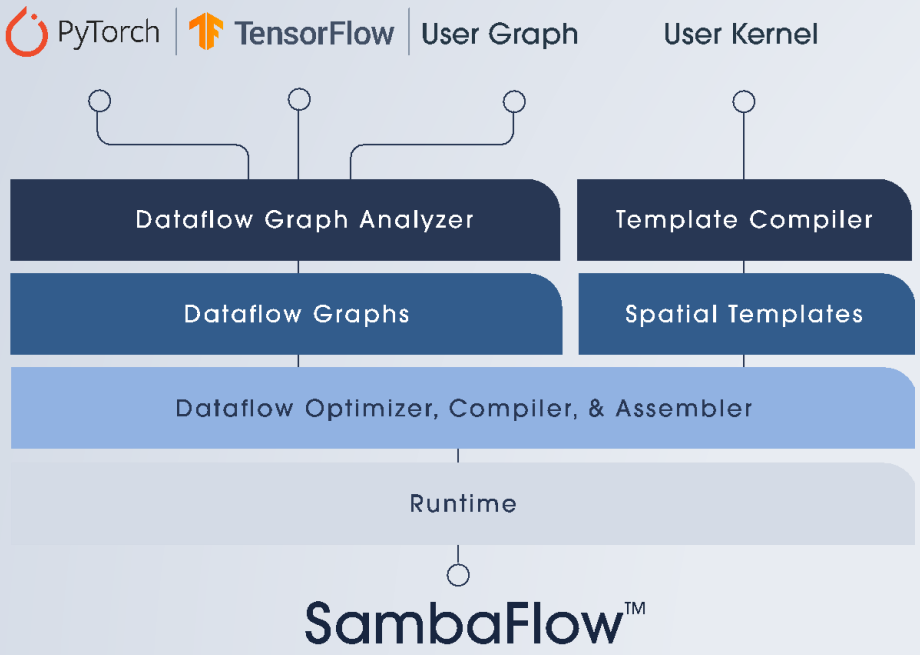


- 3層データフローメモリ
- 520MBオンチップメモリ
- 64GB広帯域メモリ
- 1.5TB大容量メモリ

生成AI学習および推論

SambaNova SYSTEMS

SambaFlowソフトウェア



- CUDAなどの専用命令は不要で、PyTorchなどのプログラムをそのまま利用可能
- SambaFlowにより空間的データフローアーキテクチャに最適な実行形式を生成

アズ・ア・サービス
事前学習済み基盤モデル

システム
DataScale®

ソフトウェア
SambaFlow™

半導体
RDU

SambaNova SYSTEMS

SambaNova DataScale

米アルゴンヌ国立研究所に導入されたシステム

DataScale

- ラックに最適化された統合システム
- 1ユニットは10Uサイズ
- 1ユニットにCPUを2基、
  RDUを8基（12TBメモリ）搭載

アズ・ア・サービス
事前学習済み基盤モデル

システム
DataScale®

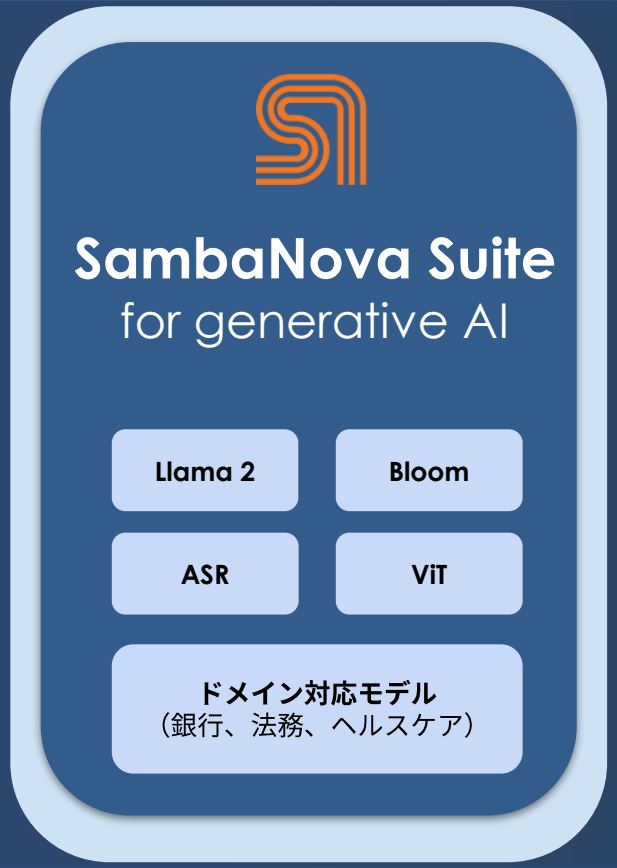
ソフトウェア
SambaFlow™

半導体
RDU

SambaNova® SYSTEMS

SambaNova Suite

企業向けに最適化され、オンプレミスまたはクラウドで展開可能な生成AIプラットフォーム



- 事前学習済み基盤モデル
  - ＋ GPT、Bloom、Llama-2
  - ＋ Composition of Expert (CoE)
- お客様環境での展開
  - ＋ オンプレミスまたはクラウド
- お客様のデータによるファインチューニングにより、安全で精度の高い生成AIを実現
- サブスクリプション価格
- SambaNovaによる完全な管理

アズ・ア・サービス
事前学習済み基盤モデル

システム
DataScale®

ソフトウェア
SambaFlow™

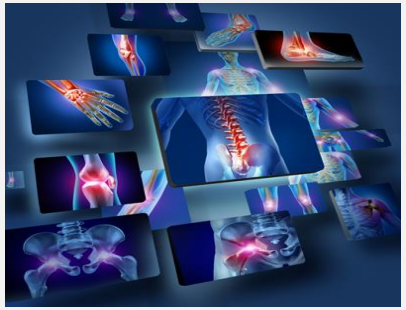
半導体
RDU

SambaNova® SYSTEMS

幅広い業界をカバーする4つの注力エリア

データフローアーキテクチャによる革新的な変化

超高精細コンピュータビジョン



生成AI

レコメンデーション

科学のためのAI

SambaNova SYSTEMS

真の解像度を持つコンピュータビジョン

従来のソリューションの精度限界を打ち破る

従来手法

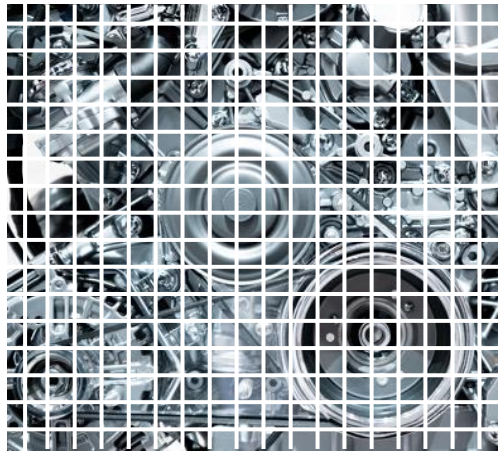
アーキテクチャ上の必要から回避策を使って学習

1. ダウンサンプリング

- 広い視野でも大きな特徴がぼやける
- 解像度が低く、ディテールが失われる

2. パッチング

- 高解像度でも狭い視野
- 大きな特徴を検出できない
- 境界部分の情報が失われる



SambaNova SYSTEMS

そのままの自然な形状で学習

元の高解像度画像の真の解像度でモデルを学習

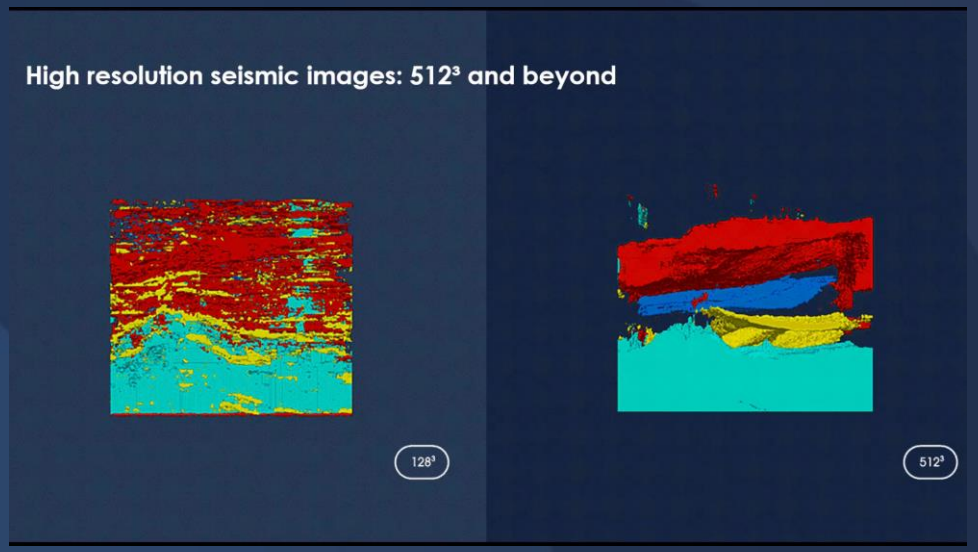
- 広い視野と高解像度を1つの解析パイプラインで実現
- 妥協のない精度

SOTA(State-of-the-art)の精度

SambaNova SYSTEMS

高精細3次元コンピュータビジョン

- SambaNovaの大容量メモリにより、512 x 512 x 512を超える、より高解像度の3次元セグメンテーションが可能になります。
- 512 x 512 x 512以上の高解像度3次元画像は、画像分析をより短時間で正確に行うことを可能にし、結果としてさらなる対象物の発見を可能にします。



「現在利用可能な従来からのソリューションは、石油・ガス産業における数百立方キロの地質データの地震解析のような複雑なAIワークロードに対応できるようには設計されていません。SambaNovaは、地質調査や探査に活用される最大の3Dネットワークを処理できる包括的なAIソリューションの提供を実現します。」

マーシャル・チョイ
SambaNova Systems
製品担当上級副社長

SambaNova SYSTEMS

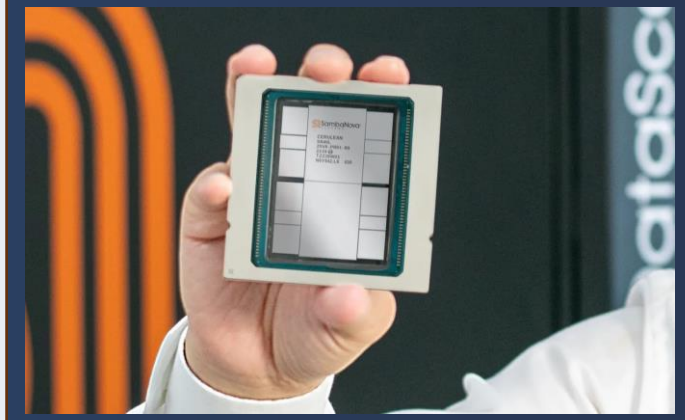
NVIDIA一色のAI半導体に一石　理研が選んだSambaNova

米サンバノバシステムズという人工知能（AI）関連の製品やサービスを手掛ける企業がある。まだ日本での知名度は低いが、AI用のハードといえば米エヌビディア一色のような現状に一石を投じる可能性がある。

例えば2023年3月にはスーパーコンピューター「富岳（ふがく）」を管理・運用している理化学研究所計算科学研究センターが「SambaNova DataScale（サンバノバデータスケール）」を採用すると発表した。これはサンバノバが提供するラック型のハードウエアノードで、企業のデータセンターなどに設置することを想定した製品だ。サンバノバの特徴はハードウエアだけでなくソフトウエアやソリューションまでをまとめてパッケージ化して提供している点だ。例えば大規模言語モデル（LLM）は特にプログラムを作らなくても実行できるような構成にしている。サンバノバデータスケールのようなオンプレミス（自社保有）向けの製品だけでなく、サブスクリプション（定額課金）モデルで契約できるクラウドサービスも提供している。もっとも基本的にハードウエアを利用者が占有する形態だそうなので、個人や中小企業がおいそれと手を出せるような代物ではなさそうだが。

サンバノバ製品の「心臓」が「RDU（リコンフィギュラブル・データフロー・ユニット）」と呼ぶAI半導体だ。AI半導体というと、米グーグルの「Tensor（テンサー）」チップのようなNPU（ニューラルネットワーク・プロセッシング・ユニット）や、エヌビディアのGPU（画像処理半導体）を思い浮かべる方が多いだろう。GPUは、1つのデータを1つの命令で処理するスカラー型の積和演算器を大量に備えている。AIにおける演算の大半が積和演算で占められているから、GPUを使うと高速化できる。TensorチップのようなAI向けのプロセッサーは詳細を明らかにしていないことが多いが、複数のデータを一括処理するベクトル演算をハードウエアで処理できるようにしたものと考えてよさそうだ。ハードウエアにより高速化・並列化しているものの、GPUやNPUはノイマン型のコンピューターであることに変わりはない。命令とデータはいずれも、メモリーに置かれているので、大量のデータを取り扱うとメモリーアクセスがボトルネックになる。これは「フォン・ノイマン・ボトルネック」などとも呼ばれる。例えばLLMをローカルで動かそうとすると、大量のメモリーを積んだGPUが必要になる。米メタが公開した「Llama 2」の比較的小規模な7B（70億パラメーター）モデルでも、メモリーは12ギガ（ギガは10億）～16ギガバイト程度必要だ。

一方でサンバノバはデータフロー型の仕組みを採用している。データフロー型とは、複数の処理ブロックをデータが流れていき、それに応じて処理を進めるアーキテクチャーだ。最近「AI」という単語はデフォルトで深層学習（ディープラーニング）を指すことが多い。深層学習の処理は、ある階層の演算結果が次の階層の入力値となり、そうした階層が多数重なって実行される。各層は多数のノードから成り、ノードの間を横断してデータをやり取りする動きはほとんどない。このため並列性が高く、GPUのような比較的単純な計算を多数同時に実行するプロセッサーが得意としているわけだ。しかしGPUの演算器が実行した結果はどこに蓄えられるかというと、メモリーである。階層内や、階層最後の段階で実施した演算は一度メモリーに蓄えられ、次の計算の際にメモリーから読み出される。エヌビディアのGPUであれば、GDDR5やGDDR6といった規格のメモリーが用いられる。これに対しサンバノバのプロセッサーRDUでは、深層学習における各階層をハードウエアコンポーネントとして定義し、その流れ通りに配置して計算させる。こうすることによって必要なデータがコンポーネント内でとどまり、メモリーアクセスがボトルネックとならないというのがサンバノバの説明だ。つまりメモリー階層で考えたときに多少大きなデータであっても、プロセッサーに近いところにとどまってデータを処理ができるというわけだ。（日経新聞2023年10月19日）



ご参考

SambaNova SYSTEMS

日本語ウェブサイトとX(Twitter)

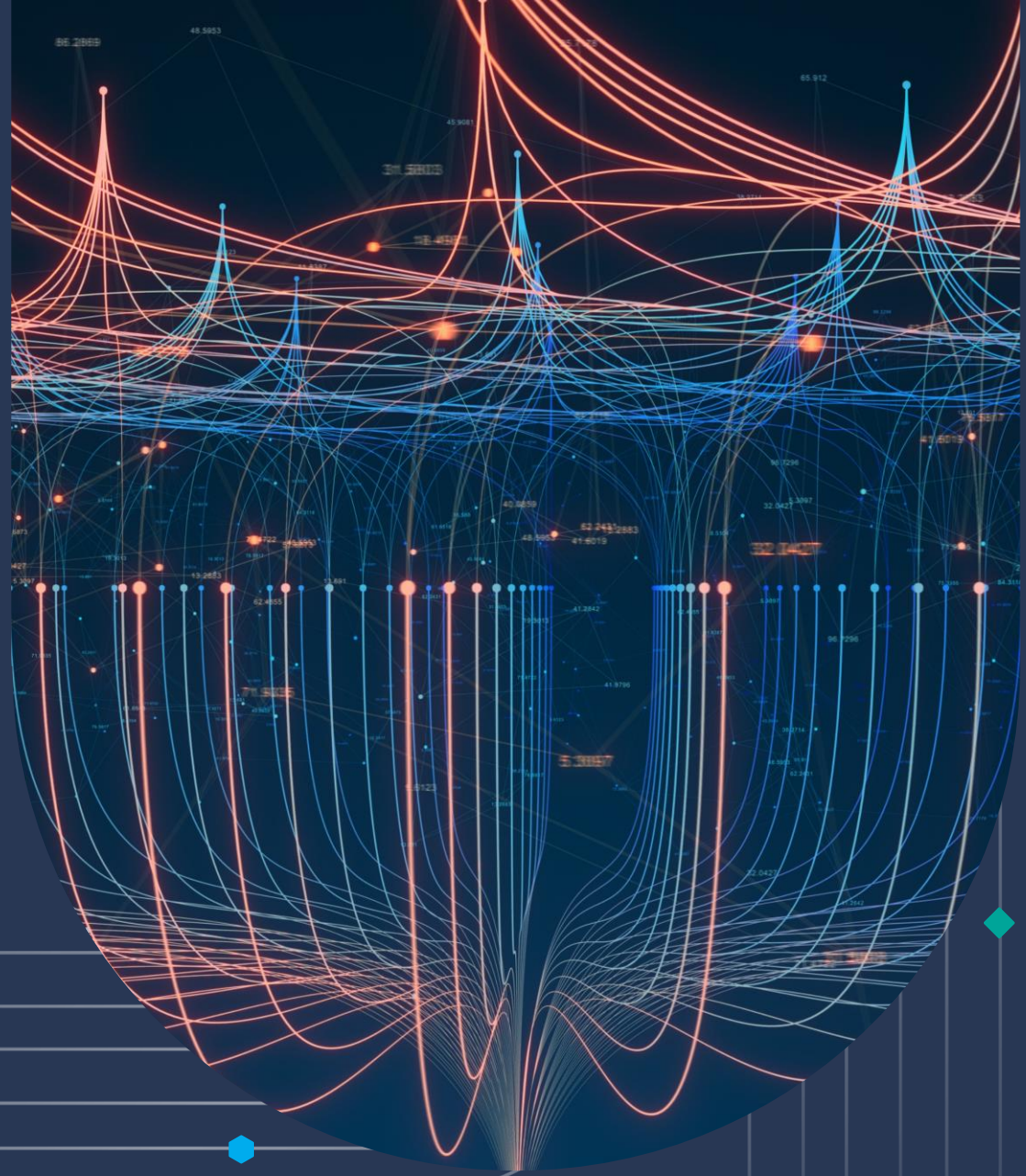
https://sambanova.ai/jp

@SambaNovaAI_jp

SambaNova SYSTEMS

Thank you!

@SambaNovaAI_jp | SambaNova.ai/jp



SambaNova SYSTEMS