Preferred Networks

MN-Coreにより実現する高効率計算機システム

2023/10/26

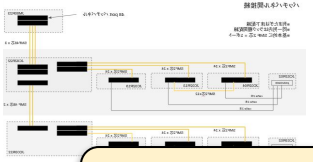
株式会社Preferred Networks
計算基盤担当VP 土井裕介

Preferred Networksの主な研究/事業化領域

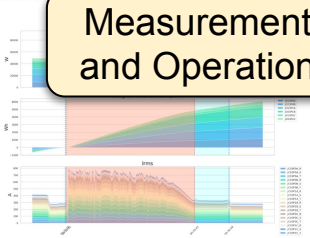
PFNは、深層学習(AI)などのソフトウェア技術と、それを支える計算インフラなどのハードウェア技術を融合し、各産業領域で最先端技術の実用化・事業化に取り組んでいます。

「計算基盤担当」とは?

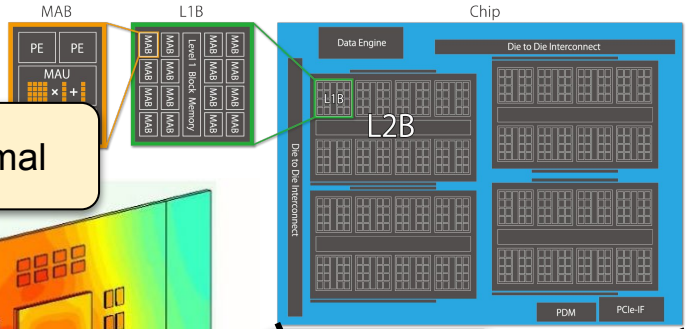
DC and Cluster

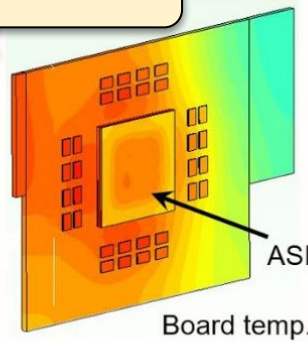


Measurement and Operation

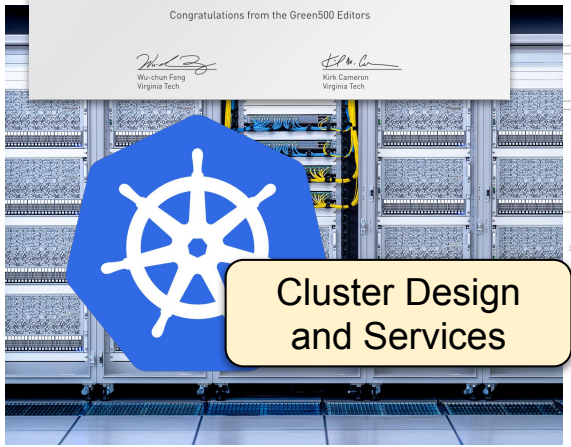


Thermal

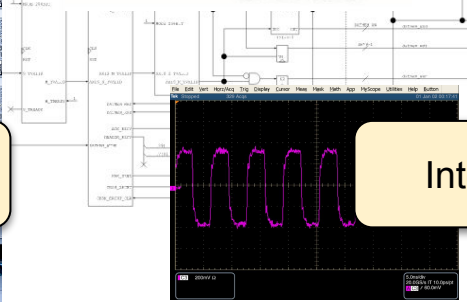




ASIC Design



Cluster Design and Services

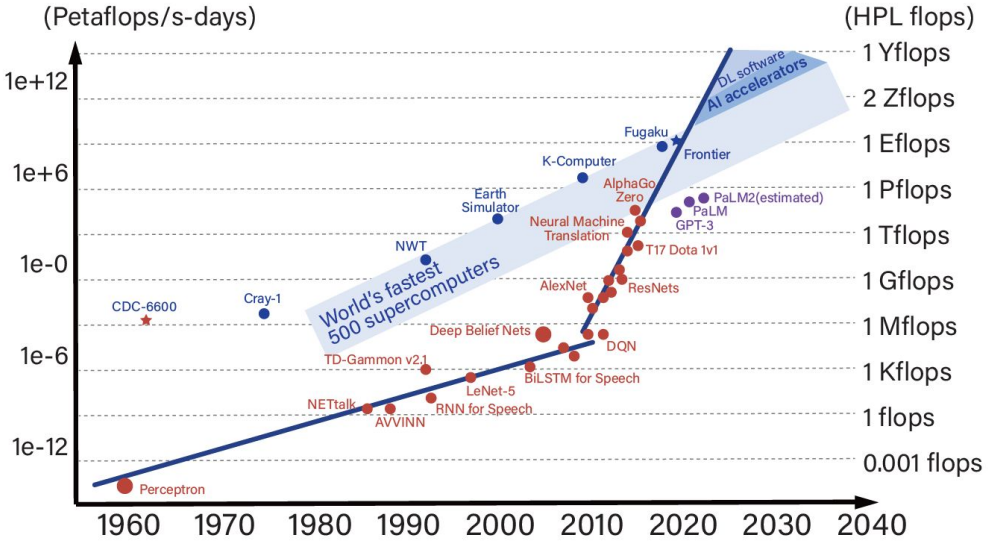


Interconnect

Preferred Networks

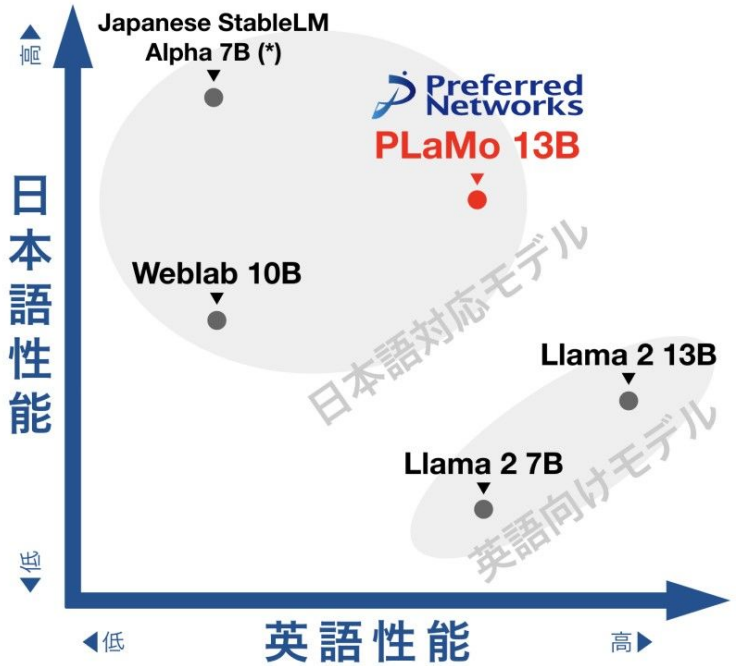
生成AIの進化と爆発的に増加する計算量需要

- ChatGPTに代表される生成AIの登場によって**AIの社会実装が急速に加速**
- 生成AIモデルの特徴はその膨大なパラメータ数であり、**必要とされる計算量は指数関数的に向上**
  - 計算量需要の増加は計算機の高速化の速度(ムーアの法則)を超えつつある
- MFM(Multimodal Foundation model)が主流になると、**今後さらに計算量は増加していく**
  - 画像、音声処理



Source: OpenAI (openai.com/research/ai-and-compute) modified by Preferred Networks

PLaMo 13: PFNが開発する大規模言語モデル



詳細：https://tech.preferred.jp/ja/blog/llm-plamo/

今年の９月下旬にPLaMo 13Bというモデルを PFNが開発しOSSで公開

- モデルサイズ：
  13B パラメータ
- 学習トークン数：
  1.4Tトークン

日本語と英語の両方で高い性能を達成

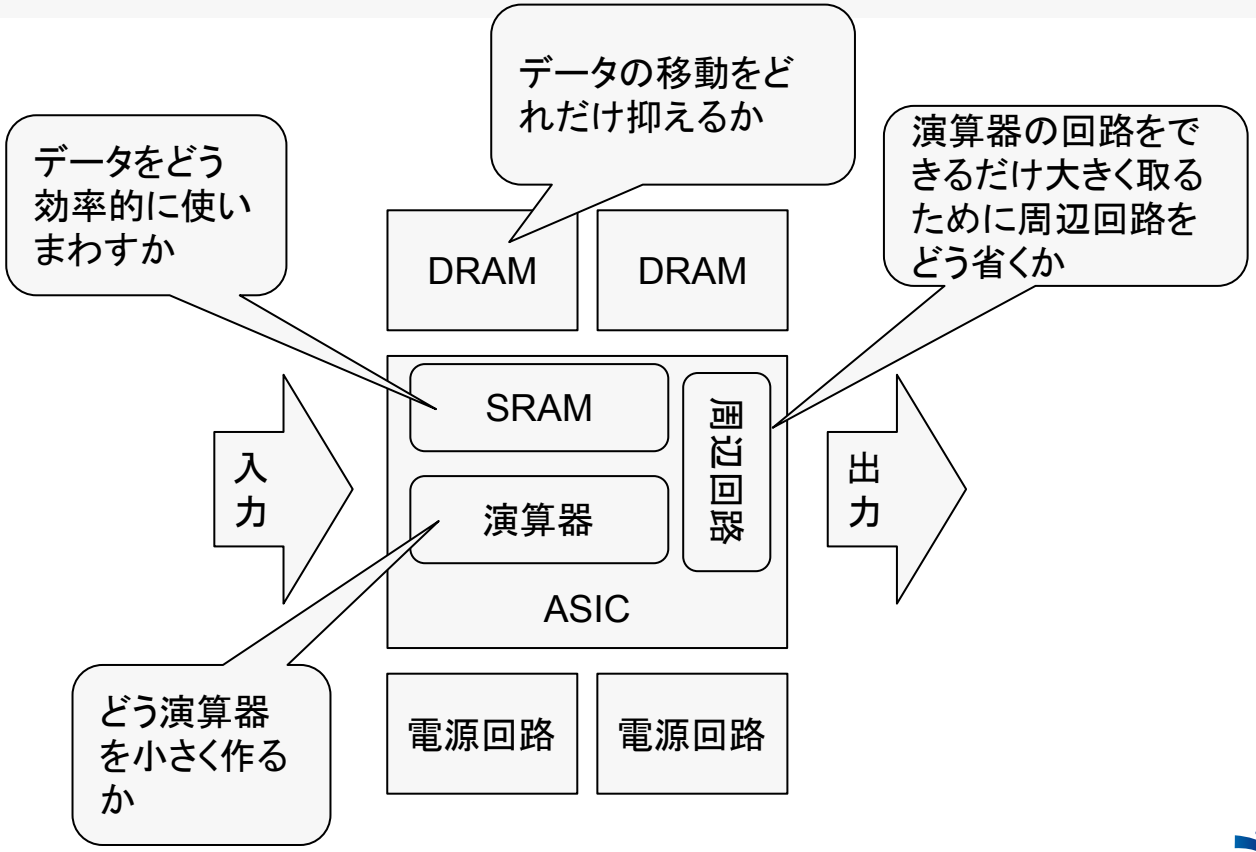
Preferred Networks

計算量への要求 ⇒ 効率の追求

- とはいえ現実的な設備の規模には限度がある
  - コスト、電力供給、GHG排出量
  - 1%の削減がおおきな差になる規模
- レイヤをまたがる効率の追求
  - ASIC、サーバ、ファシリティ、ワークロードで個々に追求していた**効率を、連動して追求**する必要がある

いろいろな効率

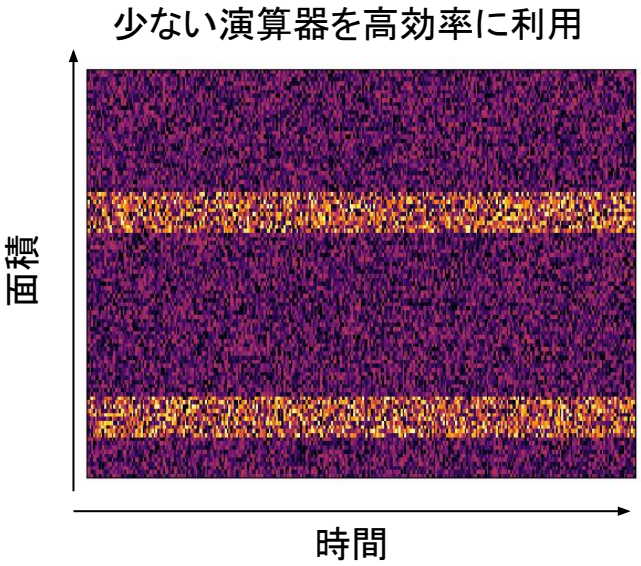
シリコン周辺の効率と課題 (例)

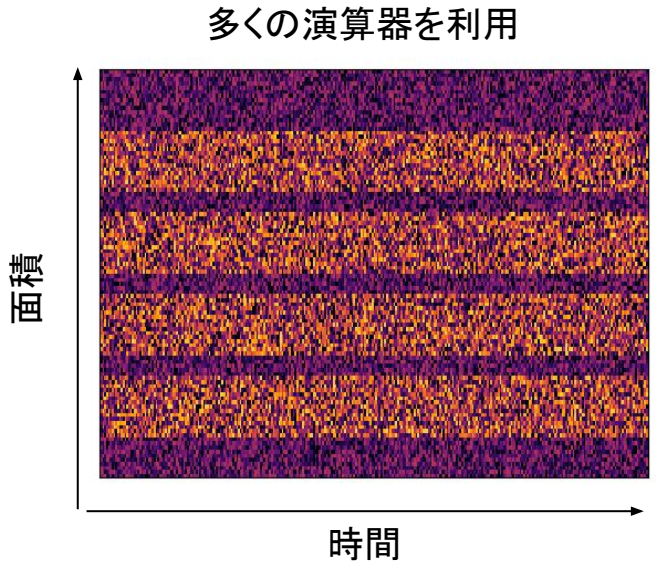


Preferred Networks

シリコンの効率 (時間方向×空間方向)

計算効率: 定義としては時間方向の効率のみ
性能向上には面積方向の効率も重要 → 高電力化





※説明のためのイメージです。何かのシミュレーションではありません

Preferred Networks

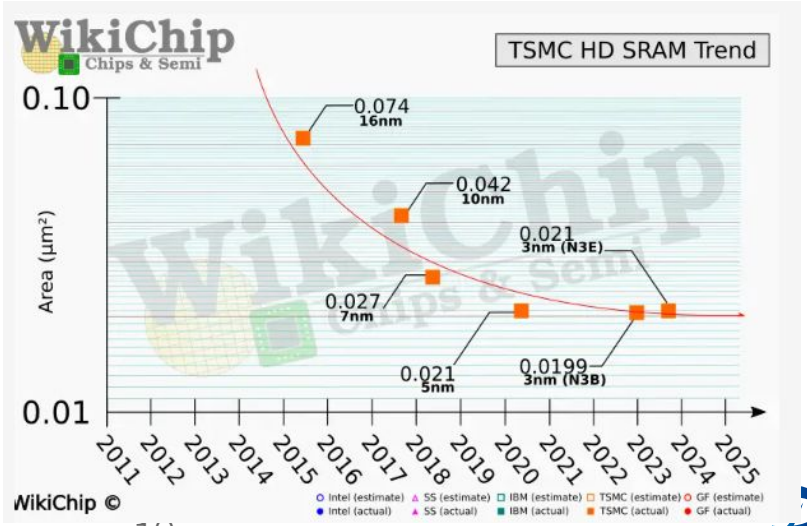
メモリとデータ移動のエネルギーとコスト

高効率化:
データをなるべく
動かさない (SRAMの活用)
だが
End of SRAM Scaling
N5→N3で縮まらない

どうする?

Task (256b ops)	Energy (Without sequencing overhead)
Two 2-operand Double Precision Floating Point Operations	15 pJ (10nm logic)
Small L1 Cache SRAM Read	30 pJ (10nm logic)
10mm move on logic die	180 pJ (10nm logic)
Low-Power discrete DRAM off-chip read	1500pJ (same timeframe as 10nm logic, DRAM portion only, add ~600 for logic side)

p.28 of https://passlab.github.io/mchpc/mchpc2019/presentations/MCHPC_Pawlowski_keynote.pdf



https://fuse.wikichip.org/news/7343/iedm-2022-did-we-just-witness-the-death-of-sram/

Preferred Networks

演算器周辺の話: 超大規模SIMDの場合

例: 倍精度4096並列SIMD計算機 @ 1GHz = 32TB/s
(4096 x 8bytes x 1GHz)

効率向上のためにはchipの中でデータをできるだけ使いまわす必要あり (HBM3の100GB/sでも遅すぎる)

- 中間データの再利用 (cache, scratchpad, register forwarding, etc)
- 再計算: DRAMからfetchするより値を再生成したほうが早い場合も (!!)

高速化のための再計算

https://tech.preferred.jp/ja/blog/mncore-compiler-optimization-with-recompute/

2023.09.26 Engineering

再計算を用いたMN-Core向けコンパイラの最適化

Area Chip Machine Learning / Deep Learning

Tag # MN-Core # コンパイラ

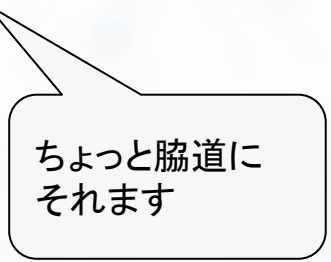
Shinichiro Hamaji
Engineer

私がPFNに入ってから知った、もっとも好きな技術トピックの一つである、MN-Core™向け再計算のご紹介をします。再計算(recomputation、rematerializationやcheckpointingなどのキーワードで呼ばれることもあります)は、その名の通り同じ計算を複数回することで、GPUメモリを節約するために再計算を利用するテクニックは広く知られています。PFNでも、再計算を使ったメモリ節約アルゴリズムに取り組み、実際の事業でフル活用しています。

MN-Core向けの再計算は、消費メモリ削減でなく、高速化を主目的としています。再計算で計算する量が増えるにも関わらず、高速化が達成できるというのが、私がとても面白いと思う点です。カラクリを紹介していきます。

Preferred Networks

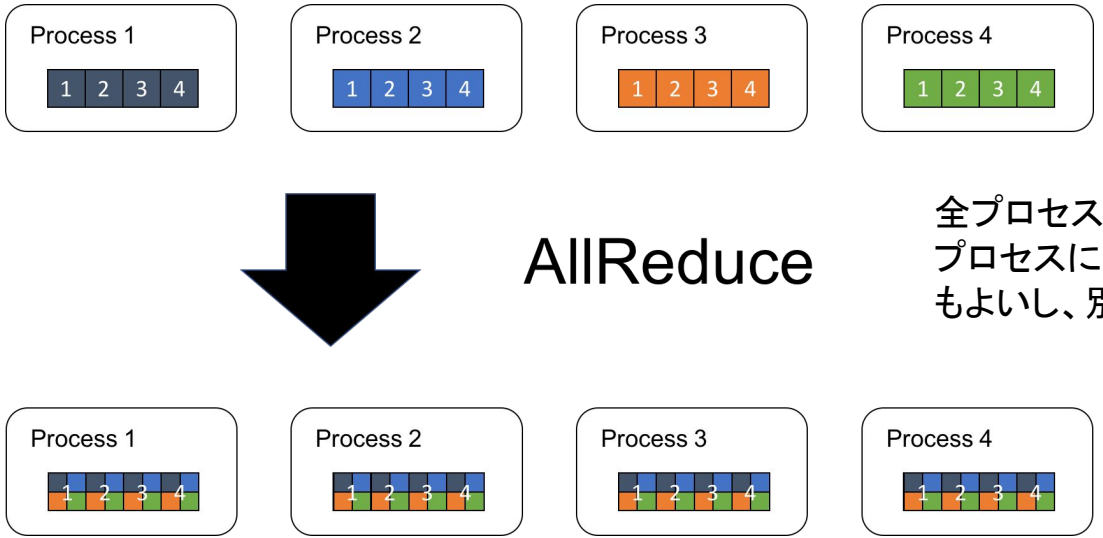
効率の話 (深層学習と分散計算)



Preferred Networks

分散計算

- 効率が下がっても分散計算が必要
  - 実速度とモデルサイズの2つの問題
- 分散計算の事例: AllReduce



全プロセスのローカルデータを集約 (reduce)して全プロセスに配布する (プロセスは同一ノードであってもよいし、別のノードにあってもよい)

Preferred Networks

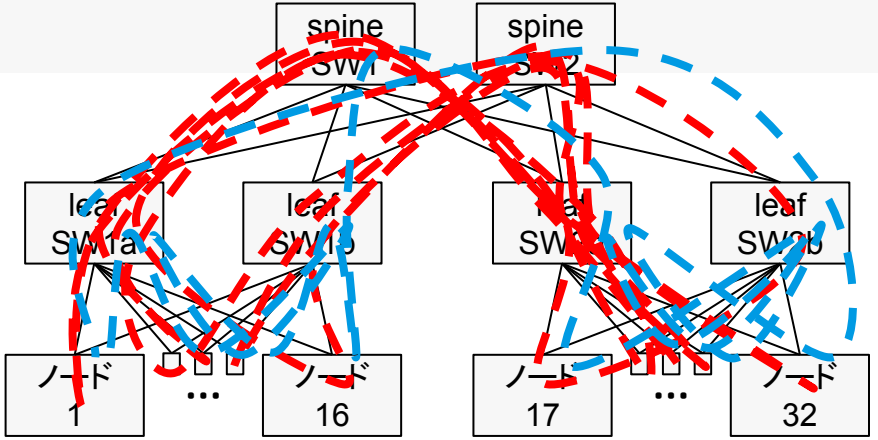
トポロジのマッチング

分散計算のオーバーレイネットワークと
物理的なネットワークが存在

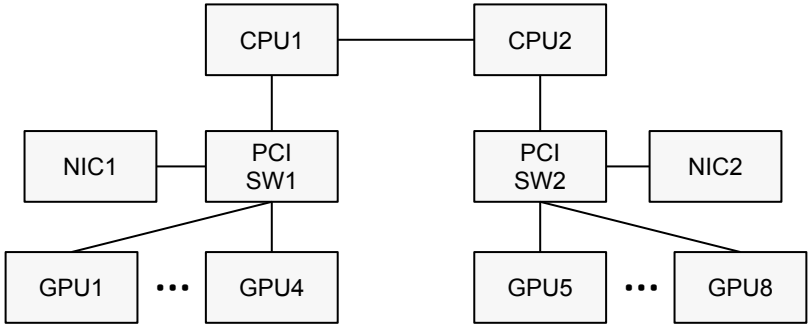
leafの中ではfull bisectionで2リンク分の
速度が(理論上)出る→ tree allreduce
leafをまたぐとleaf-spineの帯域で律速
→ ring allreduce

問題: topologyとマッチしていないと?
例1: ring(ノード1, 16, 2, 17, …)
例2: ノード1のGPU1→ノード2のGPU5
(各NICから1portの場合)

例: よくあるleaf-spine 2:4:32

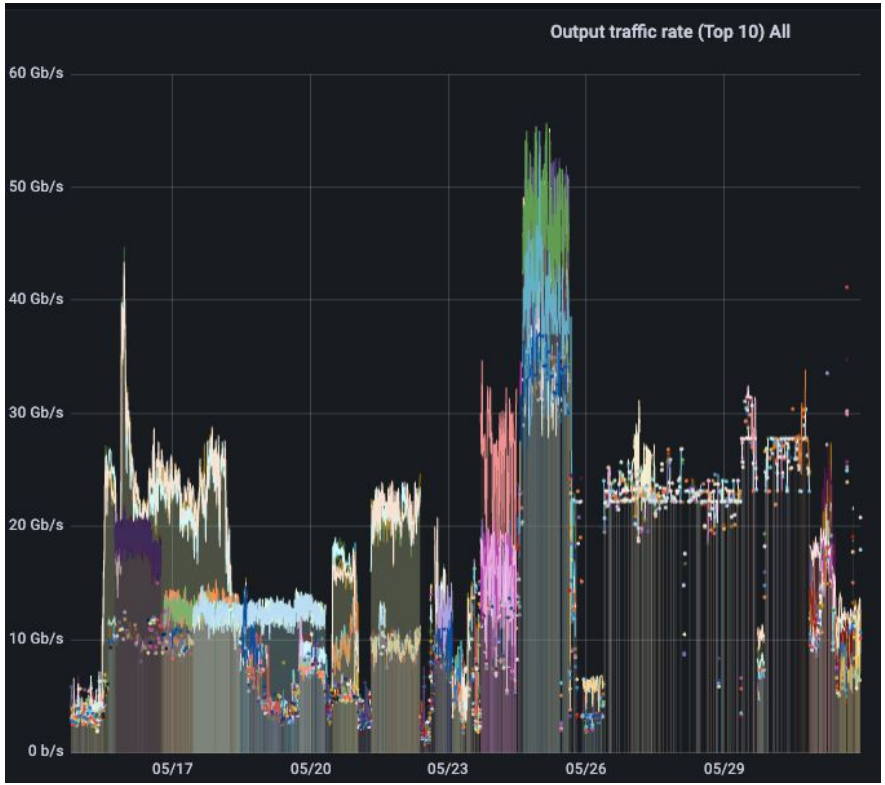


例: 古典的GPUサーバ CPU:GPU = 2:8

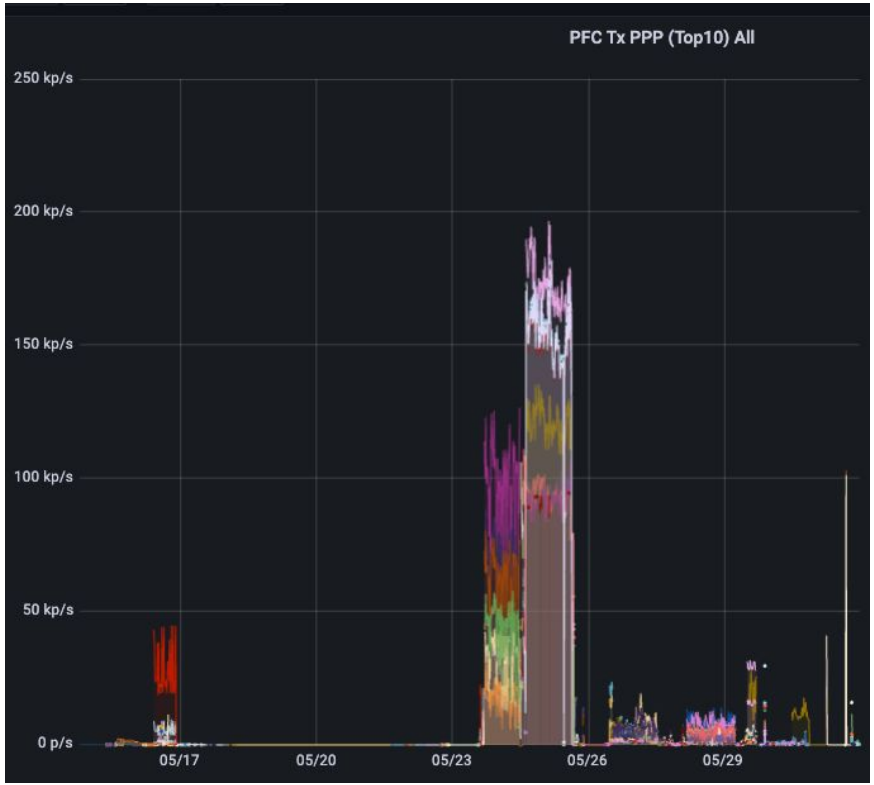


Preferred Networks

実例



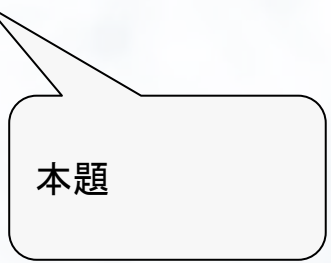
インターフェイスごとのトラフィック（100GbE）



PFC count

Preferred Networks

高効率を実現するアーキテクチャ



Preferred Networks

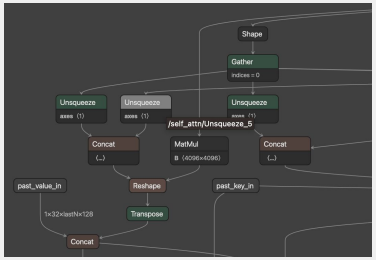
高効率x低消費電力 な AI半導体をつくるには

AIワークロードはこれまでの汎用計算と異なる特性を持つ
HWアーキテクチャの革新と、それを使いこなすSWの全体最適が重要

```
for(int y = 0;y < ih;y++){
    for(int x = 0;x < iw;x++){
        auto acc = 0;
        for(int yk = 0;yk < kh;yk++){
            for(int xk = 0;xk < kw;xk++){
                acc += image[y+yk][x+xk] *
                    filter[yk][xk];
            }
        }
        image[y][x] = acc / (kw * kh);
    }
}
```

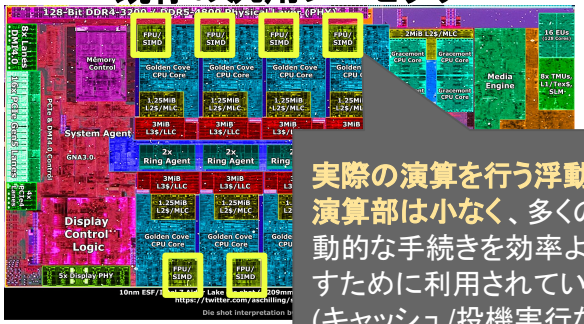
ワークロードごとにオーダーメイドで定義される、条件分岐や繰り返しで定義される**動的な手続き**

AIに必要とされる計算



AIモデルとして標準化された、計算の依存関係が明にグラフとして表現された**静的な手続き**

既存の汎用プロセッサー



実際の演算を行う浮動小数点演算部は小なく、多くの面積は動的な手続きを効率よく動かすために利用されている
(キャッシュ/投機実行など)

静的な手続きを前提とした
アーキテクチャ革新の可能性

Preferred Networks

MN-Core独自の設計思想

“Fully-Deterministic Architecture”

ソフトウェアから**透過的/確定的**に動作 ➡ **ハードウェアを最大限活用**

2つの特徴

1: 高いシリコン利用効率を狙ったチップ設計

時間方向および空間方向の両方での資源利用効率を意識した設計

＋

2: ソフトウェア最適化を前提としたアーキテクチャ

ハードウェア資源をソフトウェアで細粒度制御できるMN-Coreアーキテクチャ

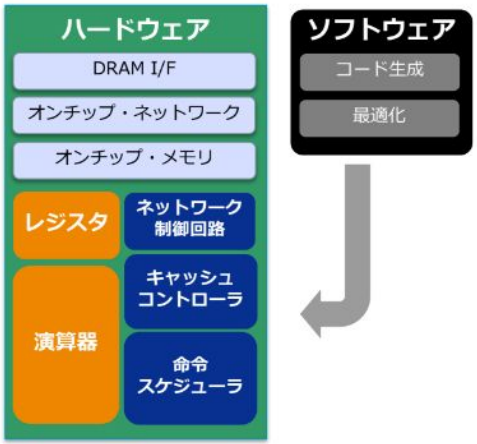
完全に事前スケジューリング可能な深層学習/AIワークロードを前提とした計算機アーキテクチャの進化

Preferred Networks

MN-Core独自の設計思想

一般的な計算機

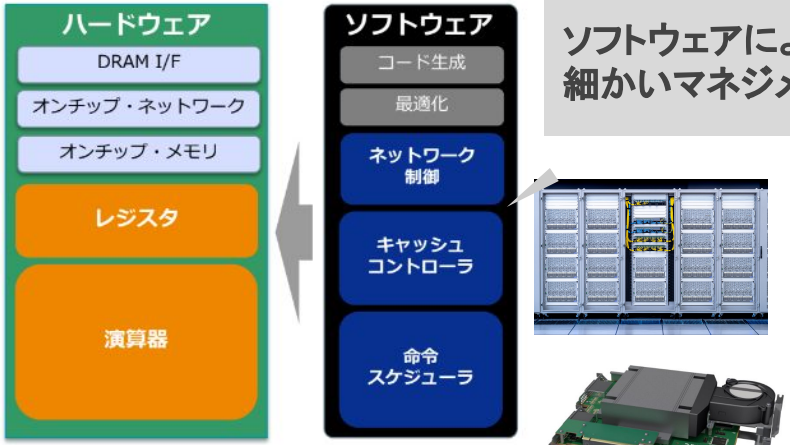
ハードウェアの内部の
挙動はブラックボックス



汎用プロセッサー

“Fully-Deterministic Architecture”

ハードウェアの内部の
挙動を透過的/確定的に制御



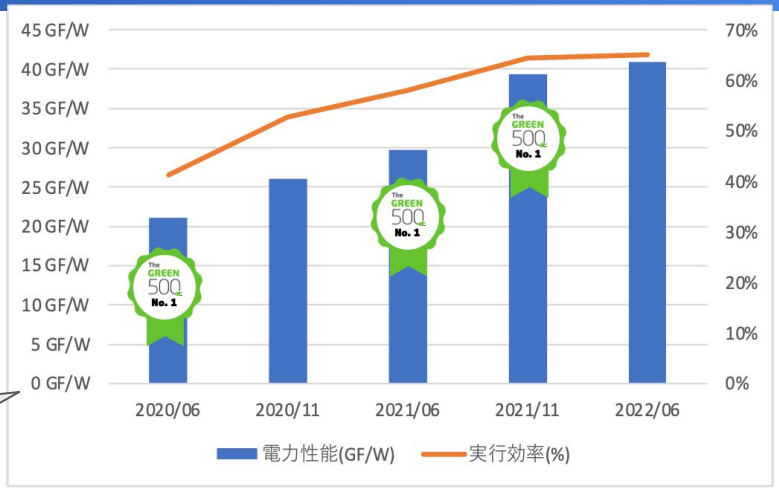
**ソフトウェアによる資源の
細かいマネジメントが可能**

MN-Core

ソフトウェアに強い企業だからこそ作れるハードウェア

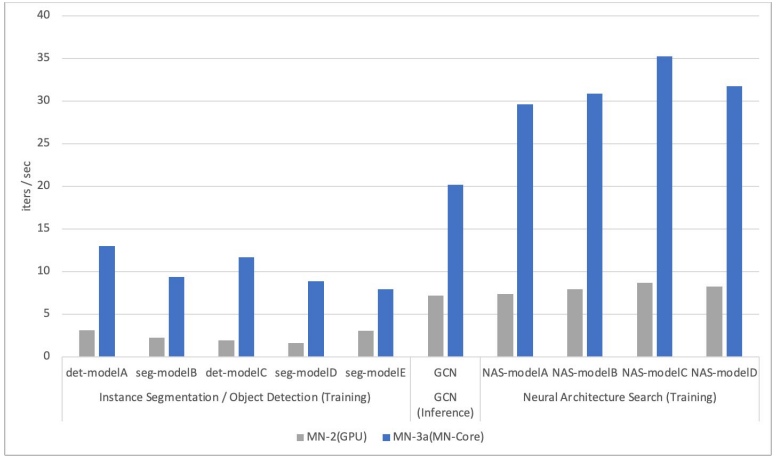
MN-Core™シリーズの実績 - HPC/AIベンチマーク

MN-Core™シリーズ初代を搭載した初の
スパコン(MN-3a) を自社で運用し、
様々なベンチマークにおいて好成績を確認



スパコンの電力効率ランキングにおいて複数回の首位を獲得（他部門は富岳などのナショナルフラッグシップスパコンが独占）

画像認識やグラフ処理(GCN)などの
AIワークロードにおいてGPUに対して
平均的に3～6倍以上の高速化を実現

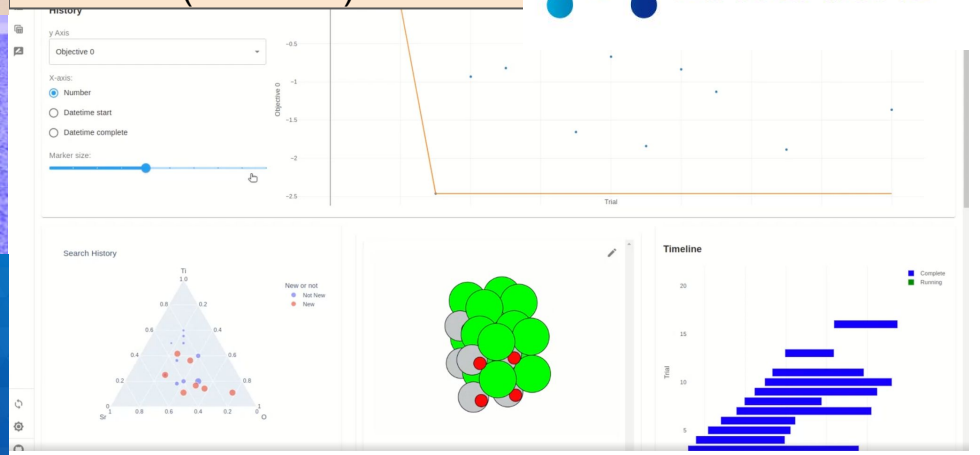


MN-Core™シリーズの実績 - AIアプリケーション

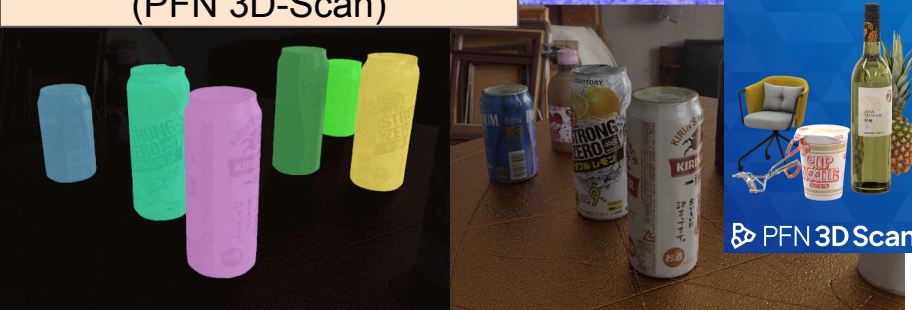
製品レベルのAIワークロードにおいて日常的に利活用されている

ロボティクス向け画像認識
(Kachaka)

AIベース材料探索
(Matlantis)



3Dモデル生成/復元
(PFN 3D-Scan)



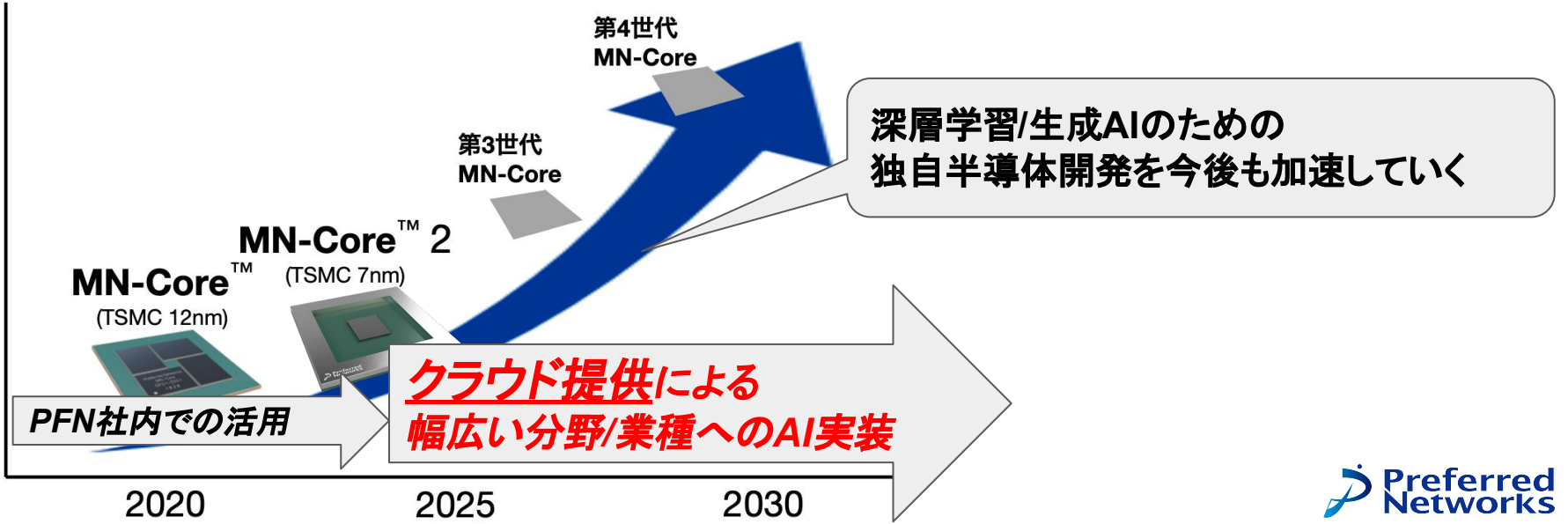
Preferred Networks

MN-Core™シリーズの進化と今後の展開

2020年~ MN-Coreをクラスタ計算機MN-3として運用開始（2016年~半導体開発を開始）

2023年 **MN-Coreを用いた計算力の外部ユーザへの試験的なクラウド提供を開始**
第2世代MN-Core (MN-Core2)の試験運用を開始

2024年 MN-Core 2を用いた大規模クラスタ計算機の構築、クラウド提供（予定）



Preferred Networks

まとめ

まとめ

- LLMを含めた深層学習には効率の良い計算機が不可欠
- 効率はいろいろなレイヤで実現する必要がある
  - アーキテクチャ ASIC設計 冷却 通信...
- MN-Coreのアーキテクチャはシリコン効率を重視
  - ソフトウェアでがんばれば効率化可能なシンプルで透過なハードウェア
- 既にたくさん使っています!