Preferred Networks

MN-Coreシリーズの展望

2024/12/19

Preferred Networks, Inc
AI Computing Div.

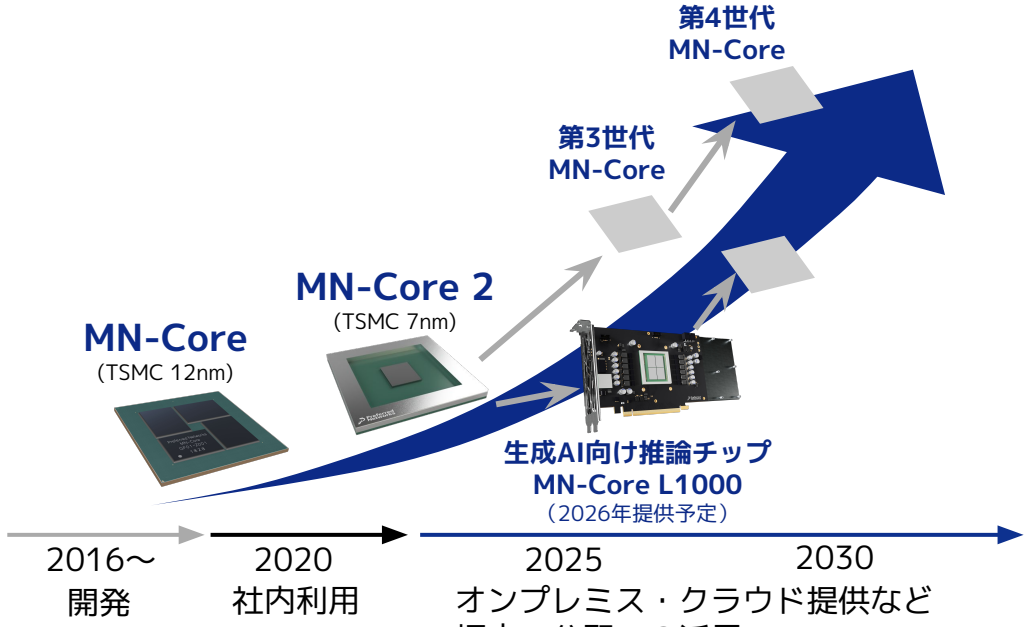
AIチップ

MN-Core™シリーズの進化と今後の展開

PFNは2016年に半導体開発を開始しました。

●MN-Core： 2020年にスーパーコンピュータMN-3に搭載。社内の研究・開発に利用。
2023年にはMN-Coreによる計算力を試験的に外部ユーザへ提供開始。

●MN-Core 2：2023年に試験運用開始。
2024年からMN-Core 2 の販売、クラウドサービスPFCP™の提供開始。



詳細: https://projects.preferred.jp/mn-core/

Preferred Networks

ご提供形態①

- Cloud-based AI Computing Service
  Preferred Computing Platform™ (PFCP™)

MN-Server 2 specifications
Accelerators: 8 MN-Core 2 chips (FP64 96TFlops, FP32 392TFlops, TF32 784TFlops, TF16 3.1PFlops)
CPU: Intel® Xeon® Platinum 8480 + (2.0GHz) 2 processors, total of 112 cores
Double-precision (FP64) performance: 8,960 GFlops
Memory: 1,024GiB
Storage: 960GB of System SSD and 15.3TB of operation SSD
Inter-node network: 100Gbps Ethernet x4

Operating system and software
Provides container image compatible with MN-Core 2 that runs on a Kubernetes pod

Charges (excl. tax)
Monthly charge: 10,000 yen
Monthly exclusive use of MN-Server 2: 1.7 million yen/server

Preferred Networks

ご提供形態②:オンプレミス

- MN-Core 2 Devkit
  - 本体価格:200万(税抜)
  - 本体価格＋初期費用セット:250万(税抜)
  - カタログ
  - 納期:25年3月前(台数限定)

- MN-Server 2
  - 本体価格:2000万(税抜)
  - 設置、運用保守などは別途ご相談
  - カタログ
  - 納期:受注生産、台数によるので要相談

- 基本的には、弊社販売パートナーの株式会社アルゴグラフィックス様からの販売になります。

Preferred Networks

MN-Coreの特徴(ビジネスサイドビュー)

- 強み
  - 共通
    - 電力性能比/省電力
    - PFNのAI人材
    - (国産/サポートが日本語)
  - MN-Core 3世代目
    - 価格性能比
    - FP64対応
  - MN-Core L1000
    - 推論性能、従来の約10倍
    - 小規模(Box型)に買える = 相対的に購入価格を抑えられる
- 弱み
  - エコシステム
    - 少ないWeb情報 => ドキュメント、PFNのサポートでカバー
- MN-Coreを採用するメリット
  - オルタナティブを持つ = 市場原理が働く

MN-Core L1000 開発背景

- MN-Coreアーキテクチャと3D Stack DRAM を活かした推論に特化したボード
- 技術的背景
  - **バンド幅に優れたチップ** は、生成AI/LLMの広まりによって**需要が大きく伸びる** ことが期待
  - **バンド幅** を確保する技術として、**3D Stack DRAM**が期待されているが、熱やnon uniformなメモリアクセスに対する対処が困難
    → MN Coreアーキテクチャは、**3D Stack DRAMとの相性がよい** 構造
  - MN-Coreをベースとした開発によって、競合他社よりも素早く商品を市場投入可能
- ポイント
  - **2026年前半**に実機にて**競合よりも大幅に優れた Transformer推論が可能なことを示す**
    ことがマーケット参入の鍵。

Preferred Networks

MN-Core L1000搭載製品企画

今後変更の可能性あり

提供価値：セキュアな超高速推論の超手軽なデプロイ

機能概要

- MN-Core L1000を**1枚 or 2枚搭載**する
- ThunderBolt等でサーバーと接続する（CPUレス）
- ネットワークに接続するとAPIサーバーが立ち上がる
  （最低限の性能CPU付き、NAS的な方式）

顧客像

- LLMのオンプレ推論を行う事業者
- オフィス、研究室、病院等に設置

イメージ図

Preferred Networks

提供ソフトウェア

コンセプト：簡単な設定でLLMサービスが利用できる

構築：
モデルの簡単なデプロイ

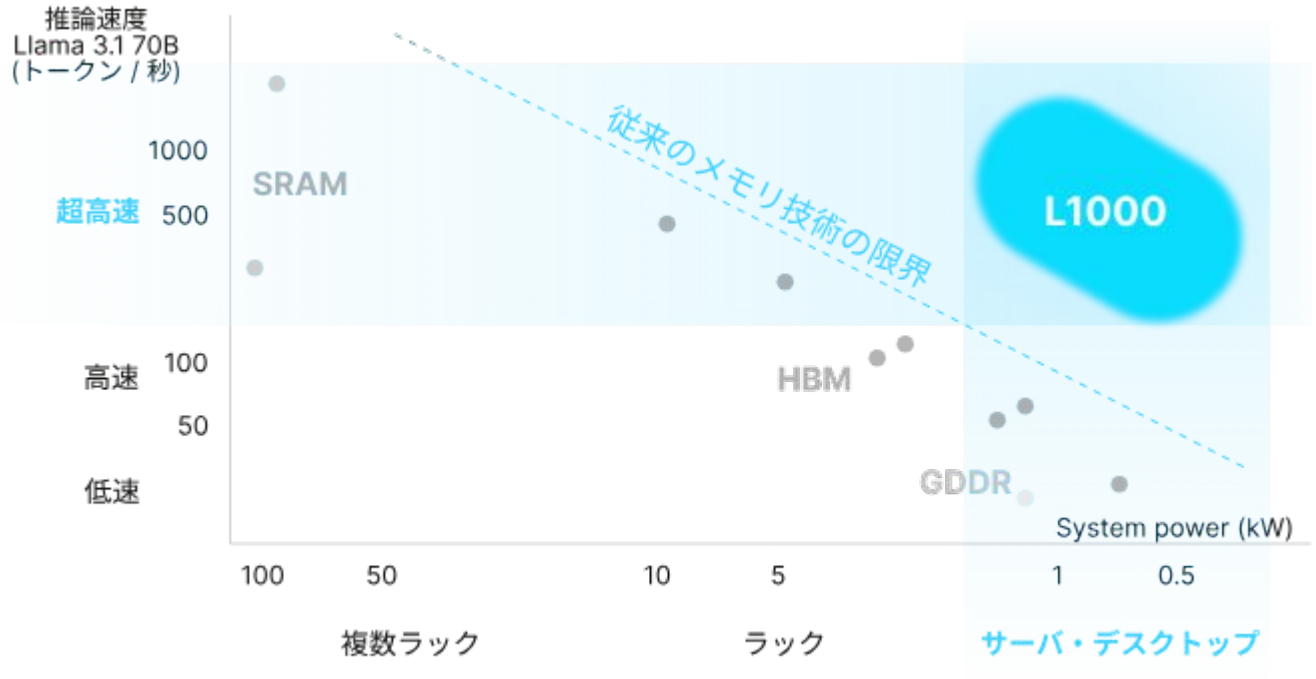
- GUIを用意
- パラメータファイルを用意するだけで
  1クリックでモデルをデプロイ可能
- 主要オープンモデル、ファインチューンモデル
  対応（Llama, Mixtral, Plamo）
- ネットワークを接続すればそれのみで
  構築完了

運用：
Chat UI、OpenAI APIとの互換API提供

- Open AI API, Azure APIに対応したAPIサーバーを用意
- ChatUI等のアプリケーションをデフォルトで用意
- lang chain、RAGサービスとのインターフェース整備
- その他デフォルトアプリケーションも
  検討

Preferred Networks

L1000が目指すターゲット



Preferred Networks

ターゲット層の位置付け
LLM のユースケースには大きく分けて二つ

大規模、スループット重視

顧客像

- apiベンダー（OpenAIなど）、秒間クエリ数が1000を超えるような大規模サービス運営者 (character.aiなど)

求めるスペック

- 単位コスト当たりのスループット

QPSの観点から可能なバッチサイズ：1000超

MN-Core L1000のターゲット

小規模、Latency重視

顧客像：一つのインスタンスに対するQPSがあまり多くない

- 独自モデルをデプロイしている人
- オンプレapiをデプロイしている人

求めるスペック

- 実用に耐える Latencyと精度のLLMのTCO

QPSの観点から可能なバッチサイズ：1~10程度

エコシステム

- 大変重要視しています
- でも、まだまだこれからです
- MN-Coreに興味がある方を中心に情報提供する場を作りつつ、コミュニティを形成していきたいと思っています
- 今後も弊社の動向をウォッチしていただけると嬉しいです。
  - X: @PreferredNetJP
  - 今後開催する講習会

Preferred Networks

Making the real world computable