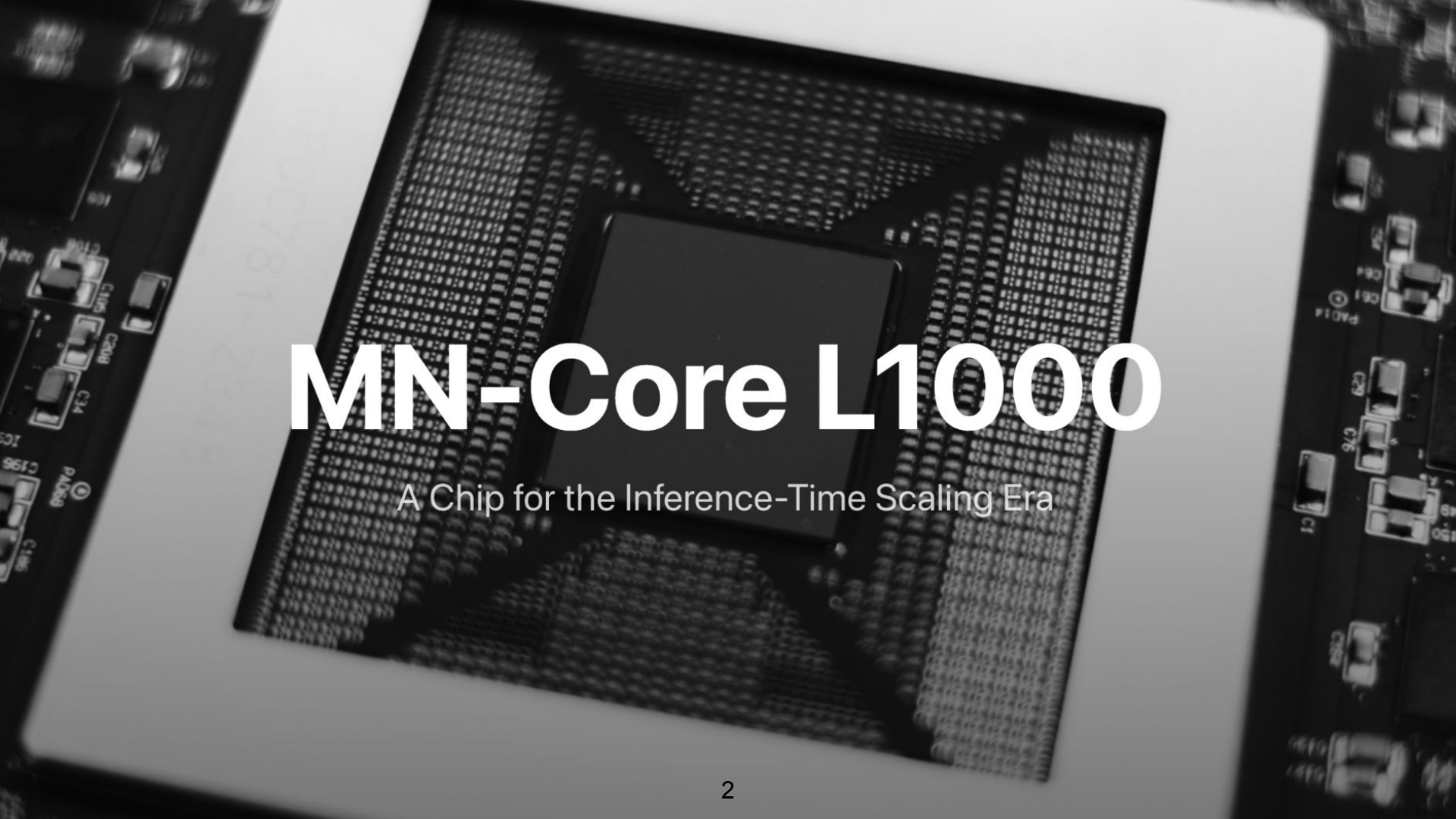
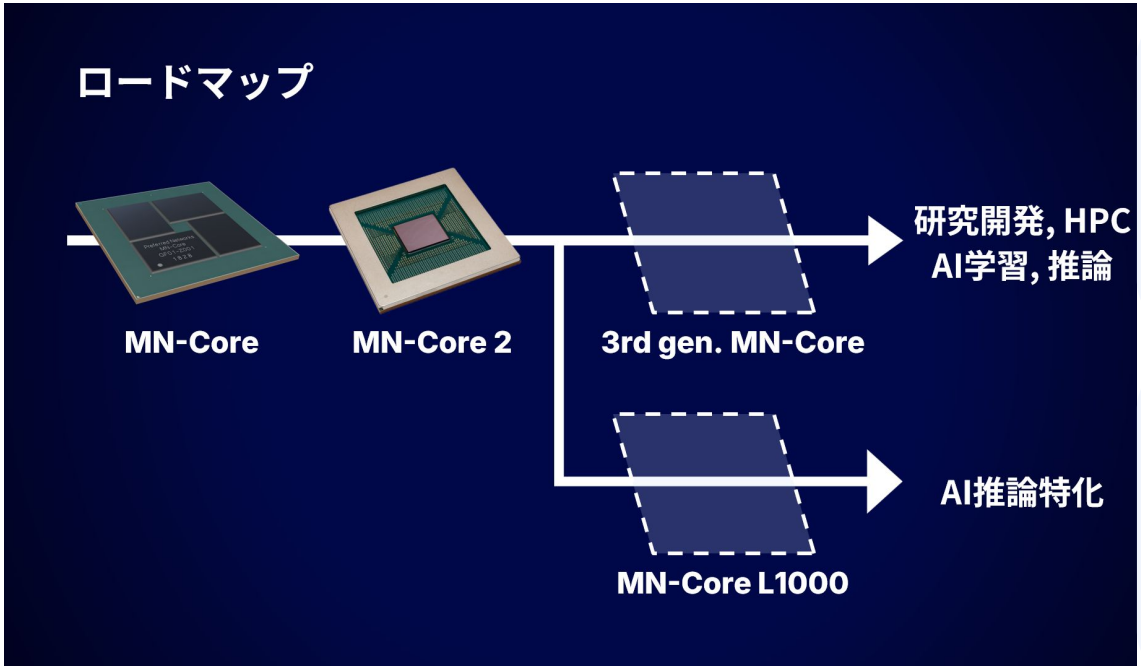
Preferred Networks

Memory bandwidth is All you need.

株式会社 Preferred Networks


MN-Core L1000
A Chip for the Inference-Time Scaling Era

Q.なぜ推論特化チップを？



MN-Core (第三世代)
高度な電力制御による
ピーク性能の改善
(> 10x peak performance)

MN-Core L1000
3D積層DRAM採用による
高メモリバンド幅
(> 10x inference performance)

Preferred Networks

A.ビジネスチャンスがありそうだったから

- 計算資源そのものが競争力の源泉となりつつある
- AI開発を効率化するAI半導体の**市場規模は2027年までに現在の2倍以上**の1,194億ドル（1.7兆円）に達すると予想されている

Gartner Forecasts Worldwide AI Chips Revenue to Reach $53 Billion in 2023

By 2027, AI Chips Revenue Is Forecast to More Than Double

Semiconductors designed to execute artificial intelligence (AI) workloads will represent a $53.4 billion revenue opportunity for the semiconductor industry in 2023, an increase of 20.9% from 2022, according to the latest forecast from Gartner, Inc.

“The developments in generative AI and the increasing use of a wide range AI-based applications in data centers, edge infrastructure and endpoint devices require the deployment of high performance graphics processing units (GPUs) and optimized semiconductor devices. This is driving the production and deployment of AI chips,” said Alan Priestley, VP Analyst at Gartner.

AI semiconductor revenue will continue to experience double-digit growth through the forecast period, increasing 25.6% in 2024 to $67.1 billion (see Table 1). By 2027, AI chips revenue is expected to be more than double the size of the market in 2023, reaching $119.4 billion.

Table 1. AI Semiconductors Revenue Forecast, Worldwide, 2022-2024 (Millions of U.S. Dollars)

	2022	2023	2024
Revenue ($M)	44,220	53,445	67,148

Source: Gartner (August 2023)

Transformerの寿命は？

Transformerアーキテクチャの持続性は高い

- **歴史的観点**：2018年に開発されて以降、様々な研究者が改良に取り組んでいるが、根本的なブレイクスルーはなかなか起きていない
- **ビジネス的観点**：大規模生成モデルは一つのモデルを作るのに多大なコストが必要なため、挑戦的なニューラルネットアーキテクチャを採用しづらい
- **技術的観点**：Transformerアーキテクチャに関する研究が世界各国で進んでいるため、異なるアーキテクチャを採用すると、総合的な研究力で遅れる事となる

Transformerアーキテクチャの推論は高いバンド幅を要求

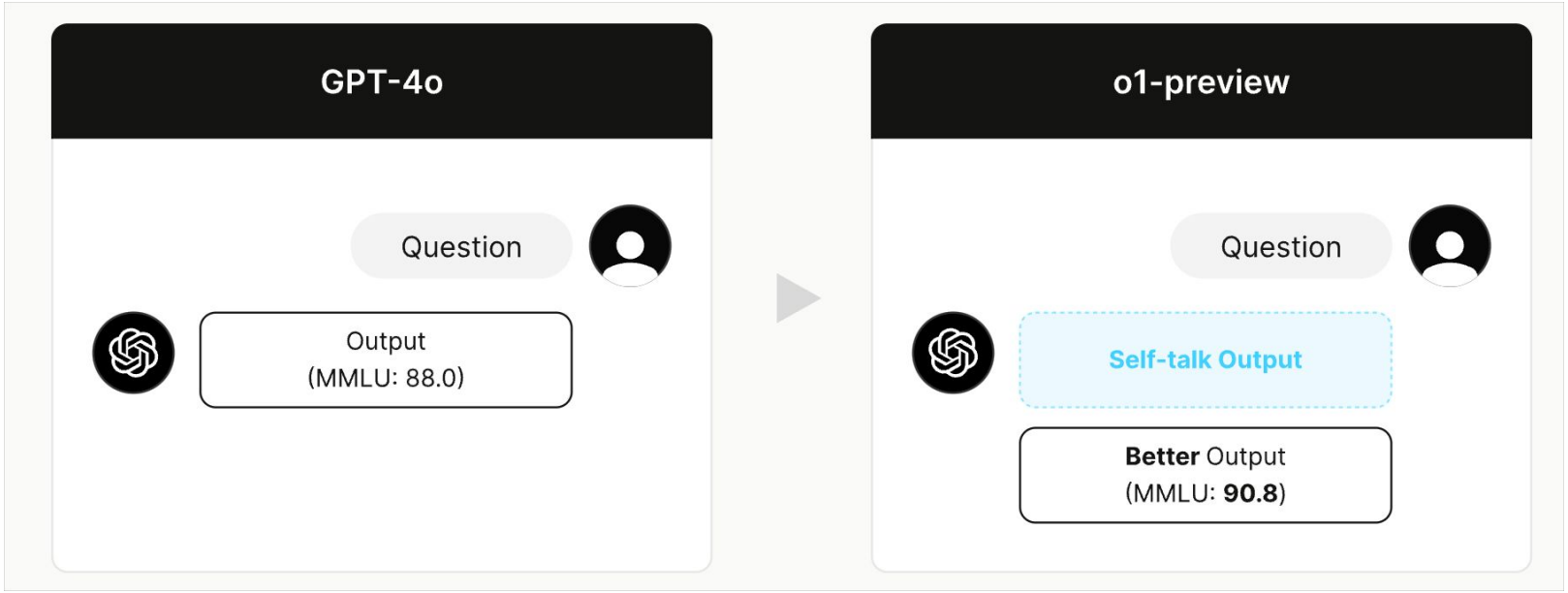
- 1トークン毎に出力する構造が維持される限り、高いバンド幅が必要
    - Mamba等のState Space Modelであってもこの傾向は変化せず

最近のLLM: 推論時スケーリング則：長く、深く考える

ユーザからは隠された、自身のみが見える「隠れトークン」を生成

人間がメモを取りながら考えるように、内部での思考、振り返りを行う

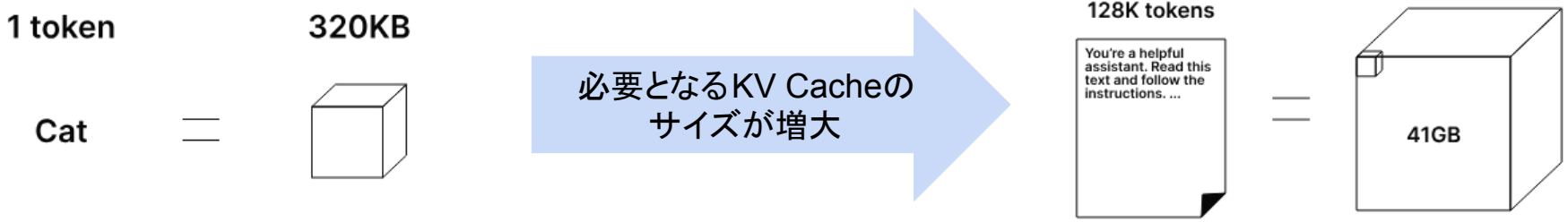
=> 性能が向上した一方推論コストも大幅に上昇



Preferred Networks

Memory Bandwidth is All You Need.

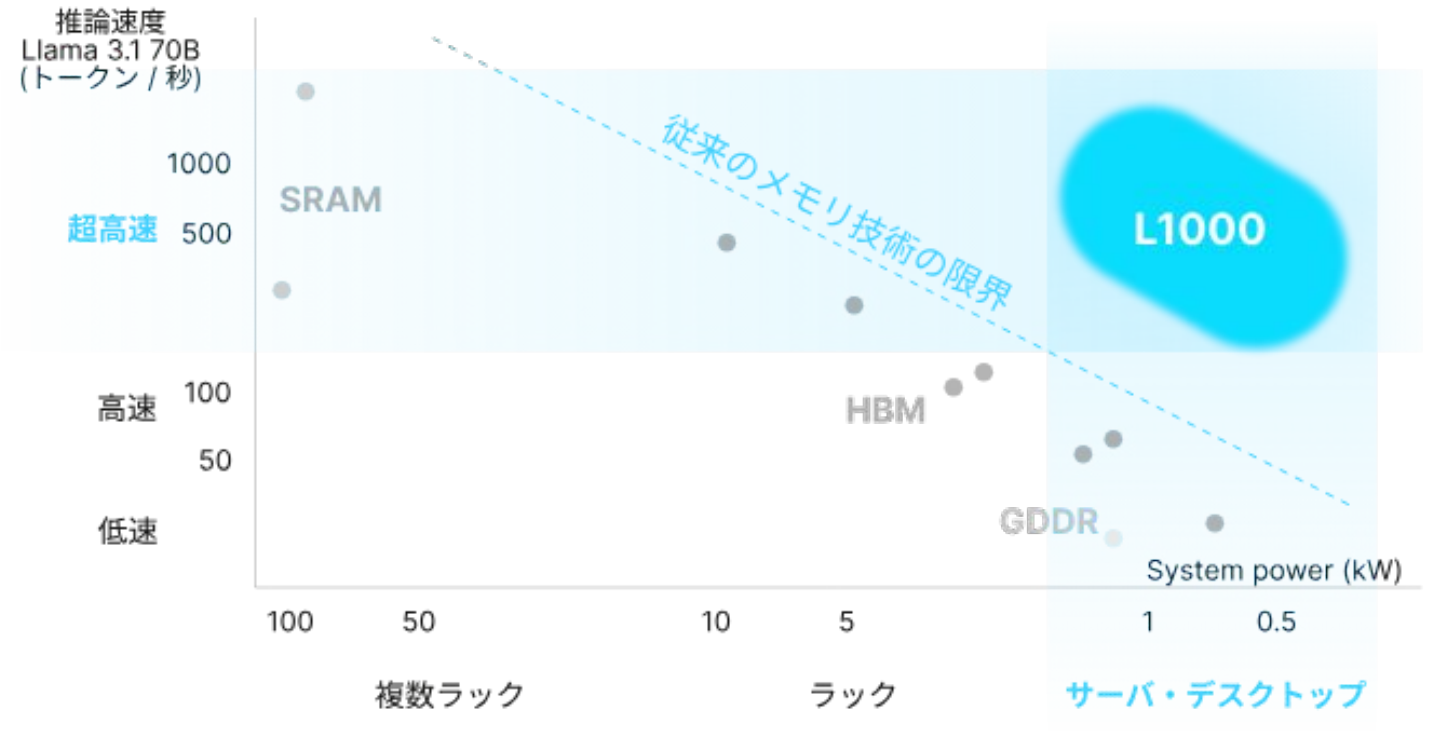
推論時スケーリングにより 高度な推論に必要とされる Token数は増加
メモリ帯域幅が 推論処理の 社会実装において重要



	MMLU Score	Output Length	Price / 1M Tokens	Output Cost
GPT-4o	88.7	2.8k	$10	$0.028
o1-preview	**92.3**	15.7k (**5.6x**)	$60 (**6x**)	$0.942 (**33.6x**)

Average result of 30 example prompts. Cost is only calculated for output tokens.
Source: OpenAI, Artificial Analysis

Preferred Networks



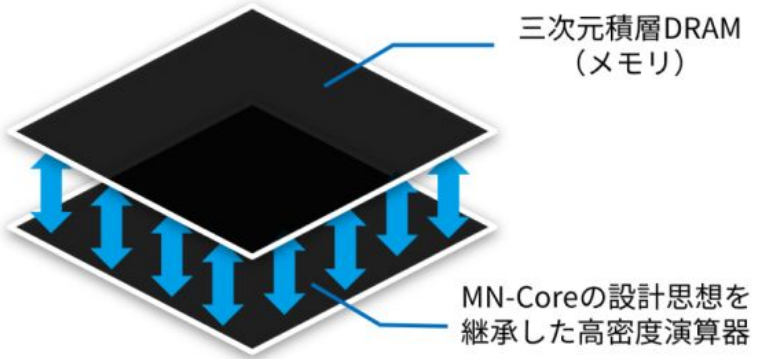
Preferred Networks

3次元積層 DRAM

DRAM DieとLogic Dieを垂直に積層
圧倒的に高いメモリバンド幅と大容量を同時に実現

実用的な容量

積層DRAMウエハ
がHBM級の容量を実現



超高速アクセス

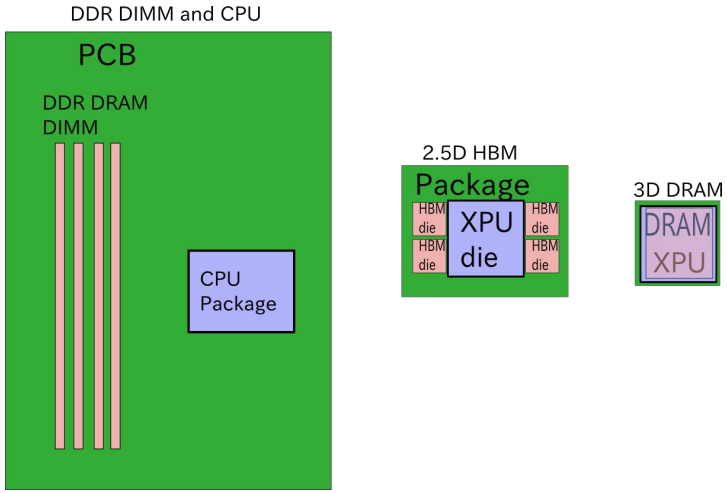
平面での接続
によりHBMを超える帯域を実現

ウェハ間の距離の削減
によりSRAM級の通信速度を実現

Preferred Networks

既存技術との違い

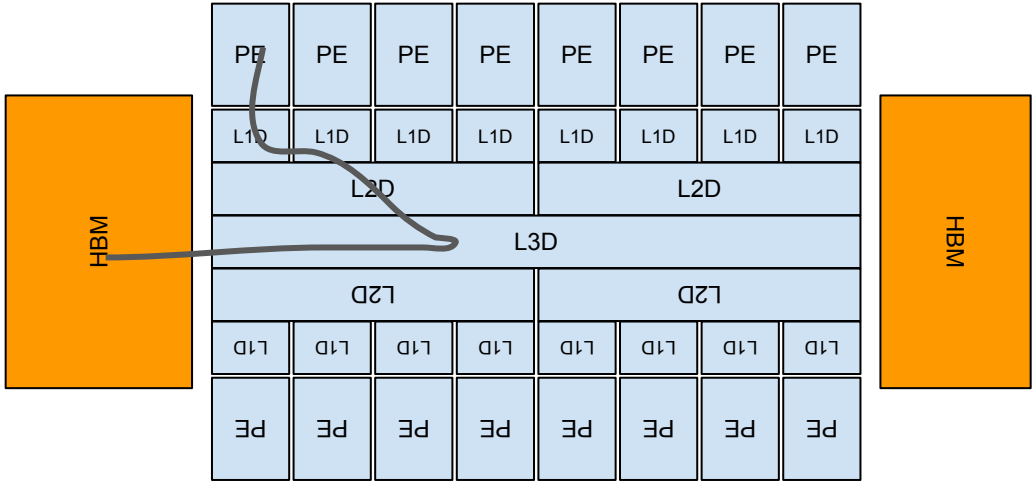
- DRAMとの接続方法の違いと表現することが可能
  - DIMM
    - CPU Package外にDRAM IFがある
    - 非常に遠い配線を引き回してパッケージに入れている
  - HBM
    - CPUパッケージ内にDRAM IFがある
    - Logic dieの下にSiPを使用して、
      至近距離での接続
      - 線でIFを生やすことが可能
  - 3D-Stacked DRAM
    - CPUパッケージ内にDRAM IFがある
    - Logic dieの上にDRAM dieがある
      - 面でIFを生やすことが可能

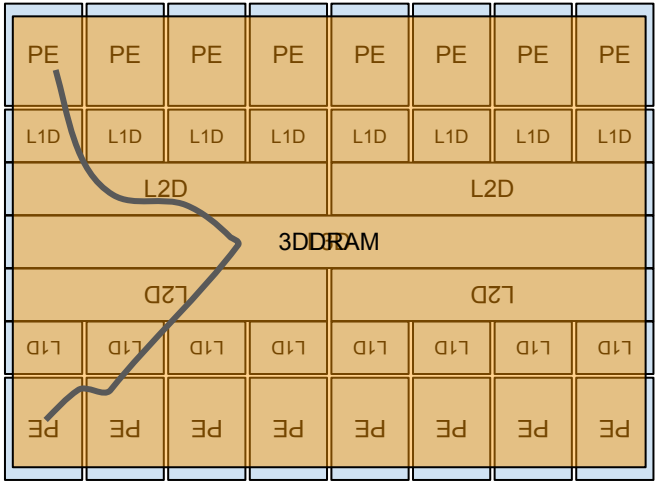


Preferred Networks

3次元積層DRAMをMN-Coreが採用する理由

- 3次元積層DRAMの利点を最大限活用するためには？
  - 共有メモリを採用したアーキテクチャでは最大限の活用が難しい





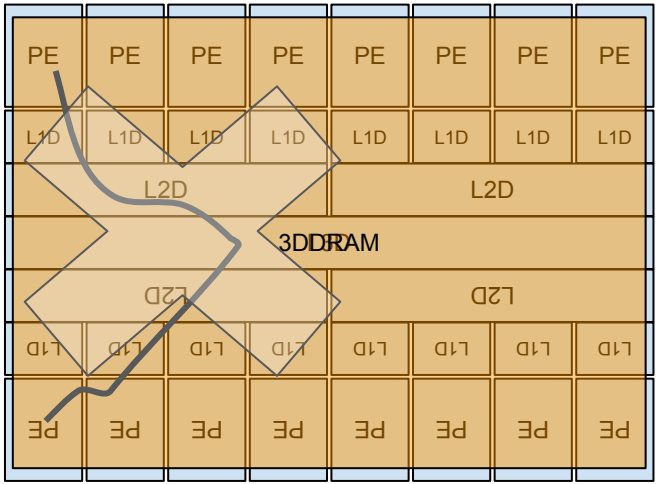
HBMとキャッシュを採用したアーキテクチャで 3D-Stacked DRAMを採用しても、”物理的なデータの移動距離 ” は大きく変わらない
つまり、3D-Stacked DRAMの利点を活かすアーキテクチャではない

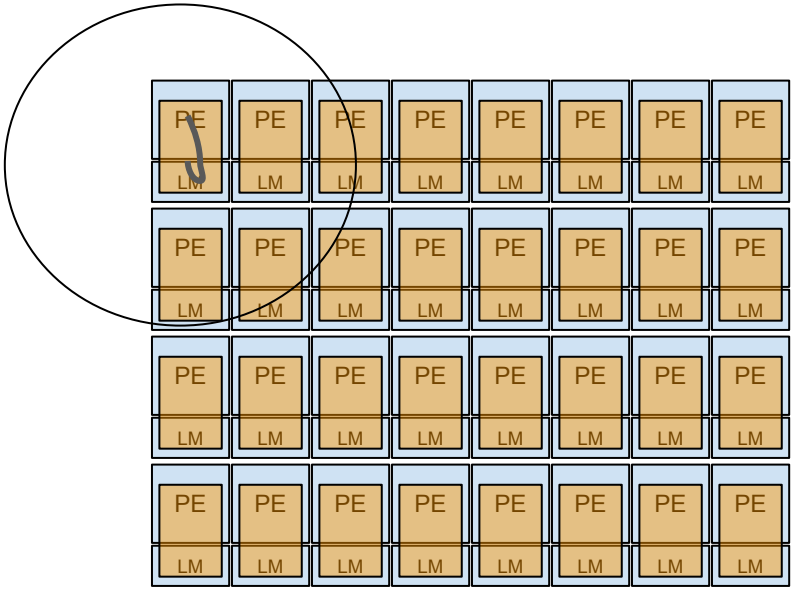
新しい
アーキテクチャが必要

Preferred Networks

3次元積層DRAMをMN-Coreが採用する理由

- 3D-Stacked DRAMを最大限活かすためには、従来型のShared Memory アーキテクチャから、Distributed-memory アーキテクチャへと転換する必要がある
  - メモリ空間が明示的に独立しているMN-Coreアーキテクチャと非常に親和性がある

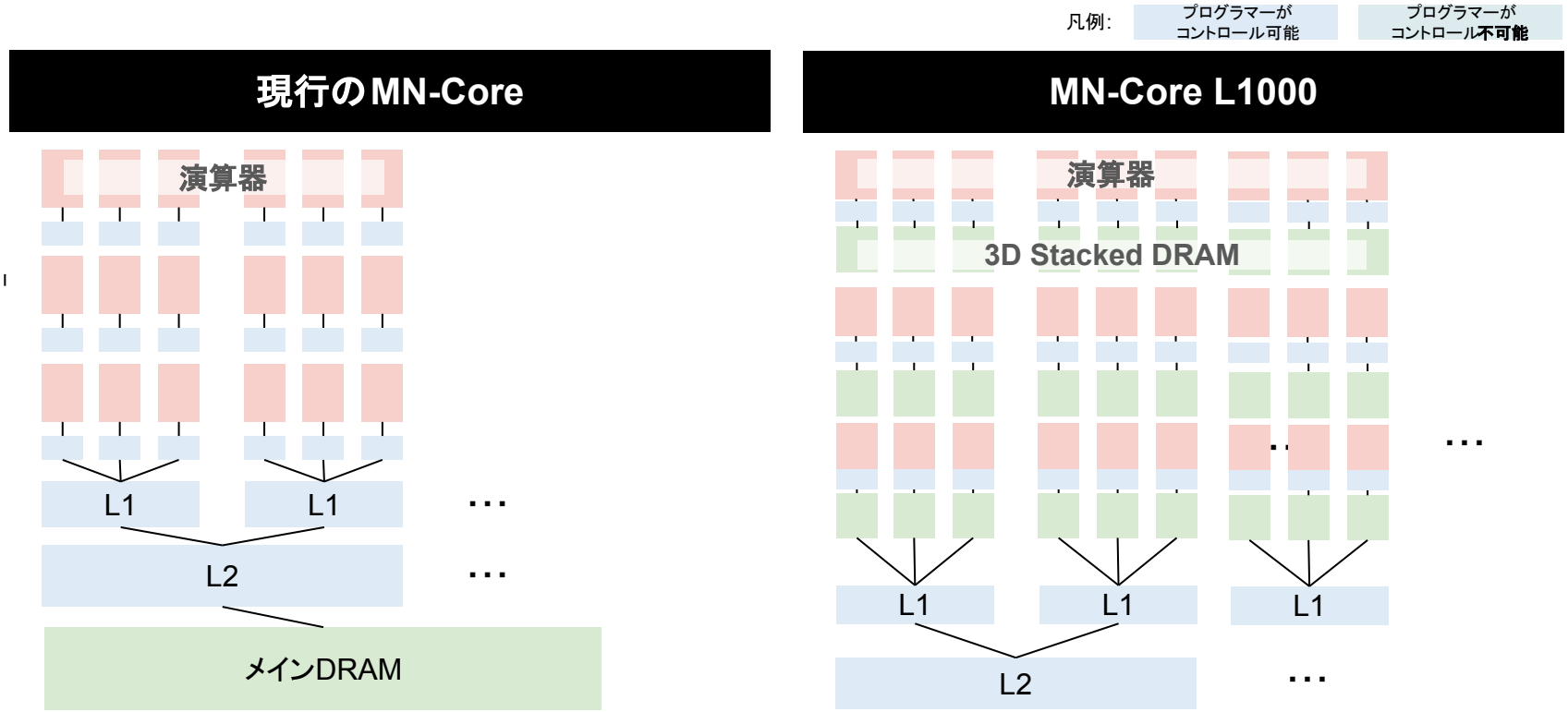




Preferred Networks

MN-Core L1000と 3D-Stacked DRAMの親和性

- MN-Core アーキテクチャで3D-Stacked DRAMを採用するとどんなふうになるか



Preferred Networks

まとめ

- LLM推論向けプロセッサ MN-Core L1000について解説しました。
- 3D-Stacked DRAMの技術的概要と、既存技術との違いについて解説しました。
- MN-Core シリーズと3D-Stacked DRAMの親和性について解説しました。

- Memory bandwidth is all you need!
- ご期待ください！