Preferred Networks

# Storage Systems in Preferred Networks

SNIA-J 最新技術動向講演会
2024/10/22

Kota Uenishi

# Preferred Networks 会社概要

## ミッション： 現実世界を計算可能にする

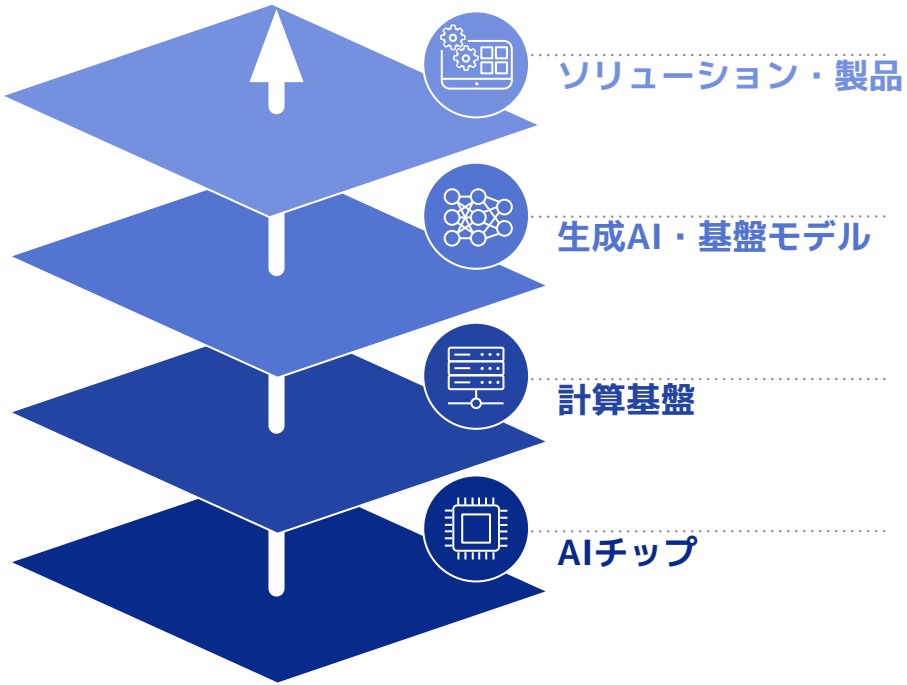
| | |
|---|---|
| **設立** | 2014年3月26日 |
| **本社** | 東京都千代田区 |
| **代表取締役** | 西川徹（最高経営責任者）、岡野原大輔（最高研究責任者） |
| **従業員数** | 約300名（2023年2月） |
| **事業内容** | 深層学習やロボティクスなどの先端技術を応用したソフトウェア・ハードウェア・ネットワーク技術の研究・開発・販売 |
| **主要子会社** | 株式会社Preferred Computational Chemistry（2021年6月）<br>株式会社Preferred Robotics（2021年11月）<br>株式会社Preferred Elements（2023年11月） |
| **出資企業** | トヨタ自動車株式会社　ファナック株式会社<br>日本電信電話株式会社　ENEOS ホールディングス株式会社<br>中外製薬株式会社　株式会社博報堂DYホールディングス<br>株式会社日立製作所　三井物産株式会社　みずほ銀行株式会社<br>東京エレクトロン株式会社 |



Preferred Networks

# PFNの事業: AI技術のバリューチェーンを垂直統合

**PFNは、チップ、計算基盤、生成AI・基盤モデル、ソリューション・製品まで、AI技術のバリューチェーンを垂直統合し、ソフトウェアとハードウェアを高度に融合することで、競争力の高い技術の開発および産業応用を進めています。**




様々な産業・消費者向けのソリューション・製品群

MATLANTIS　Preferred Networks Visual Inspection　PFN 3D Scan　rypko　kachaka

PLaMo 13B
大規模言語モデル

PLaMo
マルチモーダル基盤モデル
(2024年リリース予定)

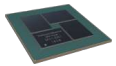
Preferred Potential (PFP)
物質の電子状態・
エネルギー計算モデル


GPUクラスタ


MN-3
(MN-Core™
クラスタ)

MN-4
(MN-Core™2
クラスタ

MN-Core™ 2による
計算能力のクラウド提供
(2024年開始予定)


MN-Core™


MN-Core™ 2

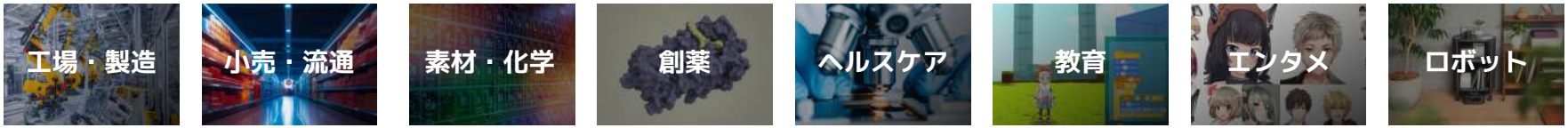
次世代機



Preferred Networks

# PFNの事業: AI技術の水平展開

**PFNは、AI技術のバリューチェーンを垂直統合し、産業、コンシューマー、社会に向けて様々な領域でソリューション・製品を水平展開しています。**




**産業**
生産性向上・品質改善
属人化回避・人手不足解消

**社会**
安心・安全な社会
高度な教育・医療

**コンシューマー**
人間の能力の拡張
新しい創作表現・娯楽体験

**生成AI・基盤モデル**

**計算基盤**

**AIチップ**



Preferred Networks

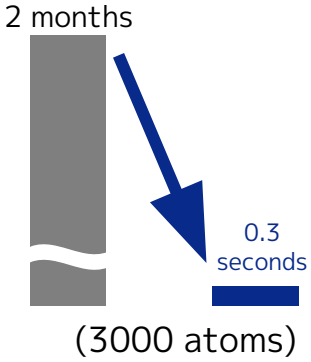
# 材料探索

MATLANTIS

**2021年7月　汎用原子レベルシミュレータMatlantisの提供開始**

https://matlantis.com/

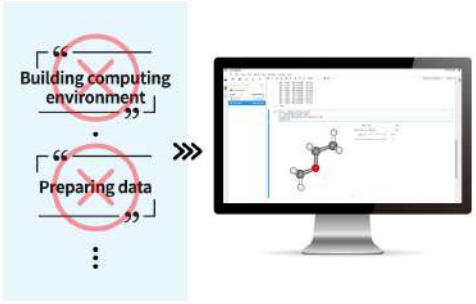
- 原子スケールのシミュレーションを10万倍から数千万倍高速化
- **72の元素**をサポート。未知の分子や結晶などに対してもシミュレーションできる汎用性
- クラウドサービスで、ハードウェアやソフトウェアのインストール不要
- JupyterLabでプログラミングによる柔軟なカスタマイズが可能




**従来のDFTと比較し**
**10万〜1000万倍の速度**


**72元素のあらゆる**
**組み合わせに対応**


**触媒や電池開発等、**
**多くのケースに適応可能**




**クラウドサービス**

※Matlantisは、PFNとENEOSの共同出資会社であるPreferred Computational Chemistry（PFCC）が販売しています



Preferred Networks

材料探索を高速化する汎用原子レベルシミュレーター

# MATLANTIS

## 数字で見るMatlantis™

### シミュレーション速度

**20M 倍**

3,000原子で従来のDFT
計算と比較した場合

### 最大対応原子数

**20k 原子**

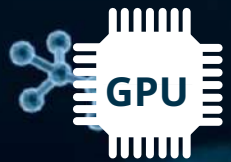
Professional Planで
Pt bulkの場合

### 対応元素数


**72 元素**

地殻上に存在する
元素の質量の99.9969%

### 1 GPU換算の計算量




**1,650年**

3,300万以上の構造からなる訓練
データの生成にかかった計算量を
1台のGPUで処理した場合の年数

### クライアント数

**80+ 組織**

企業および学術機関
などの団体

### 累計原子数

**2.6T 原子**

2023年1月～3月の3か月間で
シミュレーションした原子の総数

2024年1月現在
Matlantisのニューラルネットワークポテンシャル「PFP」は、PFNのスーパーコンピュータおよび
国立研究開発法人産業技術総合研究所のAI橋渡しクラウド（ABCI） を用いて開発されました。

# エンターテインメント
# PFN 3D Scan

2022年 日経優秀製品・サービス賞「最優秀賞」受賞

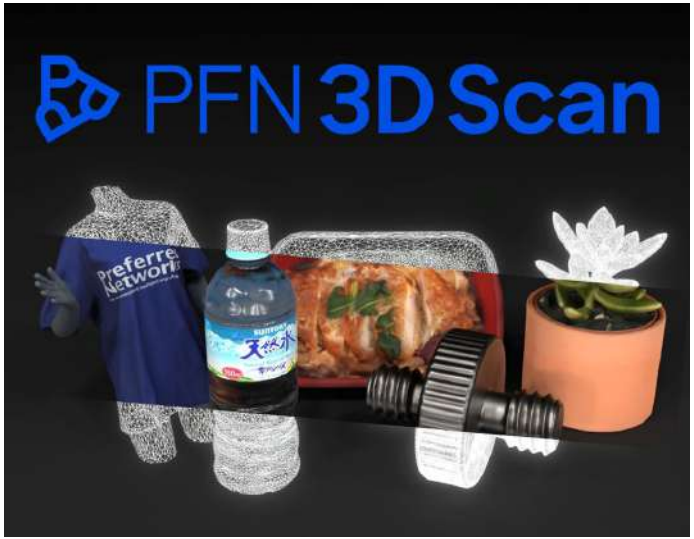
## 高品質・多様な物品に対応する3Dスキャンサービス

https://pfn3d.com/

- 2022年6月リリース
- 従来の3Dスキャンでは3Dモデル化が難しかった、

  透明・金属・黒色の物品などにも対応可能
- すでに3万点近い物品のスキャンを実施

https://youtu.be/RX4j6wxWev8

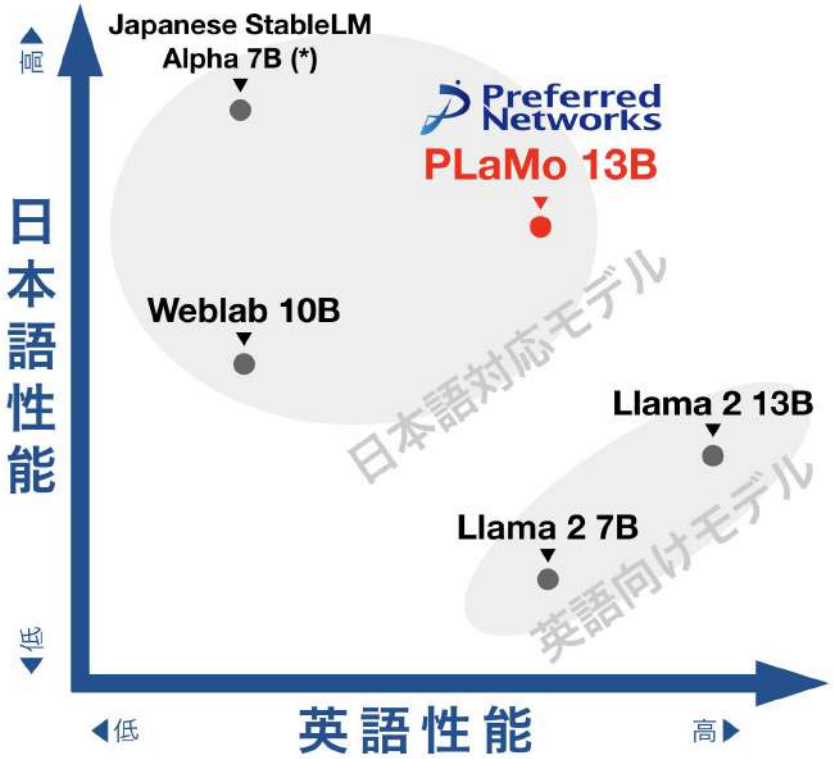





Preferred Networks

# PLaMo-13B PFNのマルチモーダル基盤モデル

研究/商用利用可能な130億パラメータの大規模言語モデルを去年9月に公開



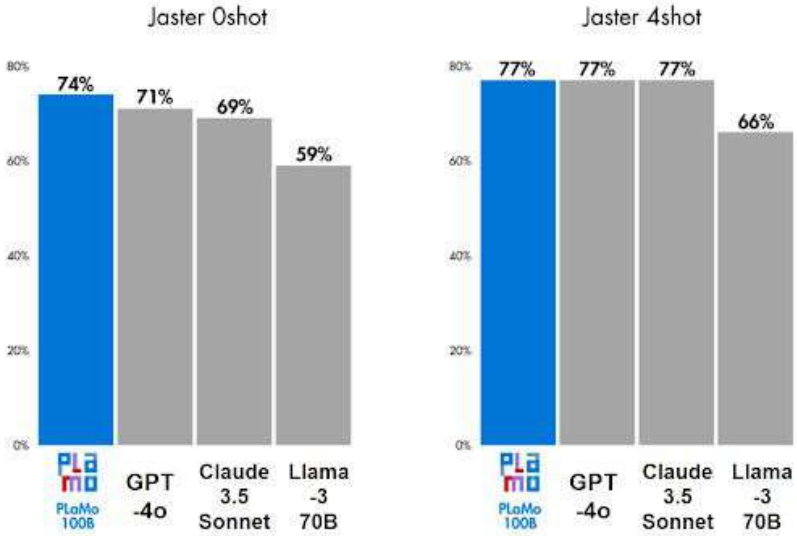

- 2023年9月公開当時、日英2言語をあわせた能力で世界トップレベル
- 経産省ABCIのスーパーコンピュータ A100 480GPUで1ヶ月弱利用。日本語、英語の1.4兆トークン（数兆文字）を使って学習
- 学習時の実行効率は41%で世界の他の学習基盤と比べても高い（効率的に学習資源を使える）

- 開発実績を元に基盤モデル開発・提供を行うPreferred Elementsを2023年11月に設立

Preferred Elements

Preferred Networks

# PLaMo-100B-Pretrained の公開




- 研究利用、商用利用可能
- 1000億パラメータ
- 10/15 公開
  - https://huggingface.co/pfnet/plamo-100b
- 100% 子会社Preferred ElementsがGENIACプロジェクトの支援のもとフルスクラッチで開発
- 法人向け PLaMo Prime および軽量版 PLaMo Lite を開発中

Preferred Elements

Preferred Networks

# PFN の計算基盤

MN-2a, MN-2b

The GREEN 500 世界1位!! (ISC20, ISC21, SC21)

MN-3

MN-J

**MN-2a**
V100 x1024

**MN-2b**
A100 x168, A30 x252

**MN-3**
MN-Core x192

100G IP-Clos Network

Login Node

NFS
(home, archive)

**Object Storage**
SATA HDD 14 TB x3000

パブリック
クラウド

**北海道DC**
H100 x512

ストレージサーバ

Preferred Networks

# MN-J の計算機

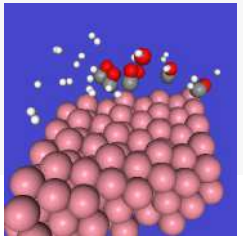
| MN-2a | MN-3 | MN-2b (A100) | MN-2b (A30) |
|---|---|---|---|
| **128 nodes** (1024 GPUs) | **48 nodes** (192 MN-Cores) | **42 nodes** (168 GPUs) | **42 nodes** (252 GPUs) |
| 36 CPU Cores | 48 CPU Cores | 128 CPU Cores | 80 CPU Cores |
| DDR4 384 GB | DDR4 384 GB | DDR4 1024 GB | DDR4 512 GB |
| V100 (16 / 32 G) SXM2 x 8 | MN-Core x 4 | A100 (80 G) SXM4 x 4 | A30 (24 G) PCIe x 6 |
| 100 GbE x 4 RoCEv2 with SR-IOV | 100 GbE x 2 MN-Core DirectConnect | 100 GbE x 2 RoCEv2 with SR-IOV | 100 GbE x 2 RoCEv2 with SR-IOV |





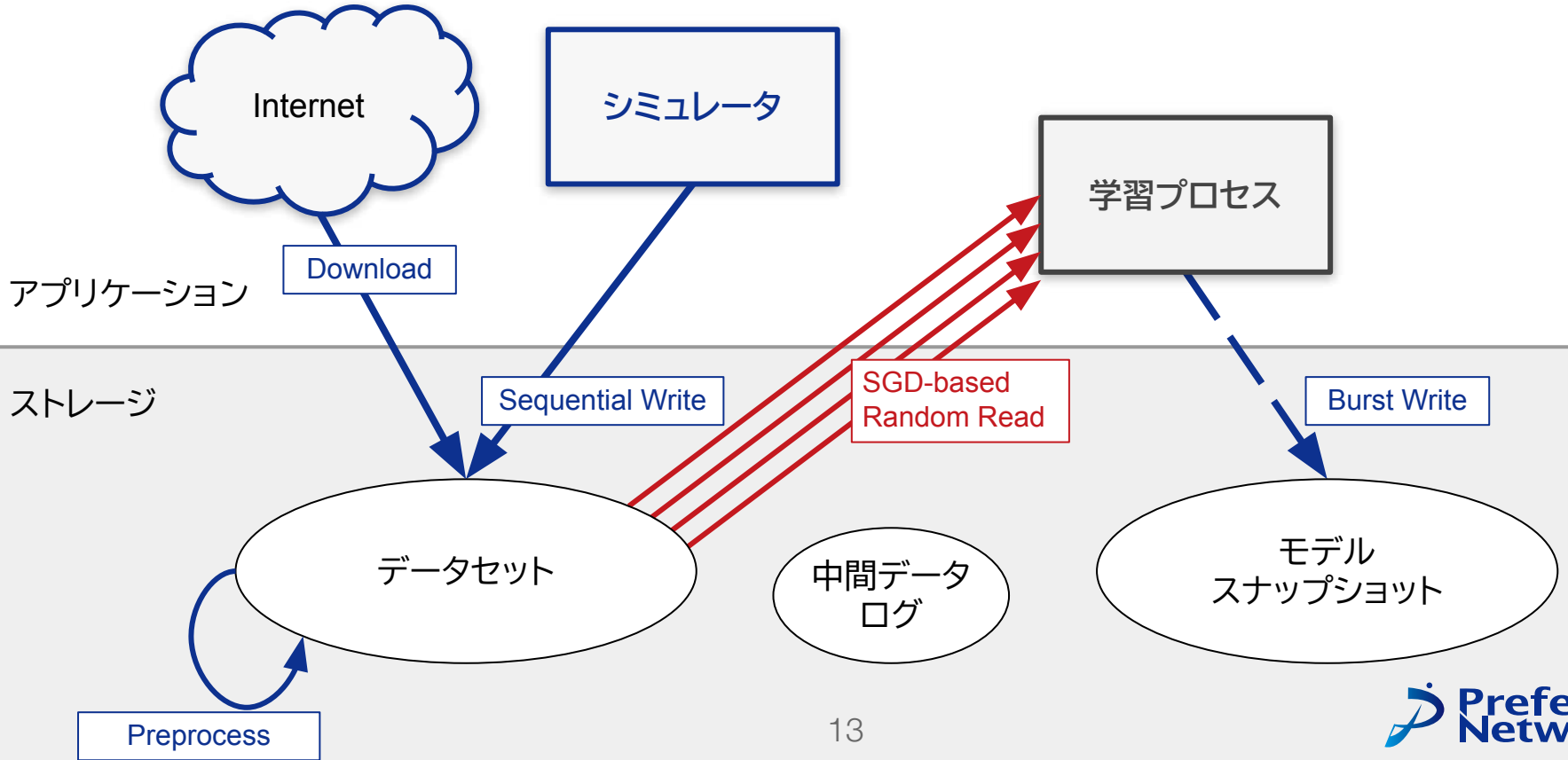
Preferred Networks

# MN-J のワークロード


- Matlantis
  - DFTシミュレータにより大量の学習データを生成
  - 再現性のために中間データを多めに保管していた時期があった
  - >100TB
- Kachaka
  - 家庭内などエッジで動作する軽量・頑健な画像認識モデルを学習するために、家庭を模した 3D モデルからシミュレートした画像を大量にレンダリングして学習データとした
- PLaMo
  - テキストの学習データ自体はそこまで大規模ではないが、学習スナップショットの保存が課題。10B で 360 GB とか、これを数時間に 1 回保存する（スループットだけでなく安定的な Write が課題）
  - パラメータが増えるとプロポーショナルに大きくなる

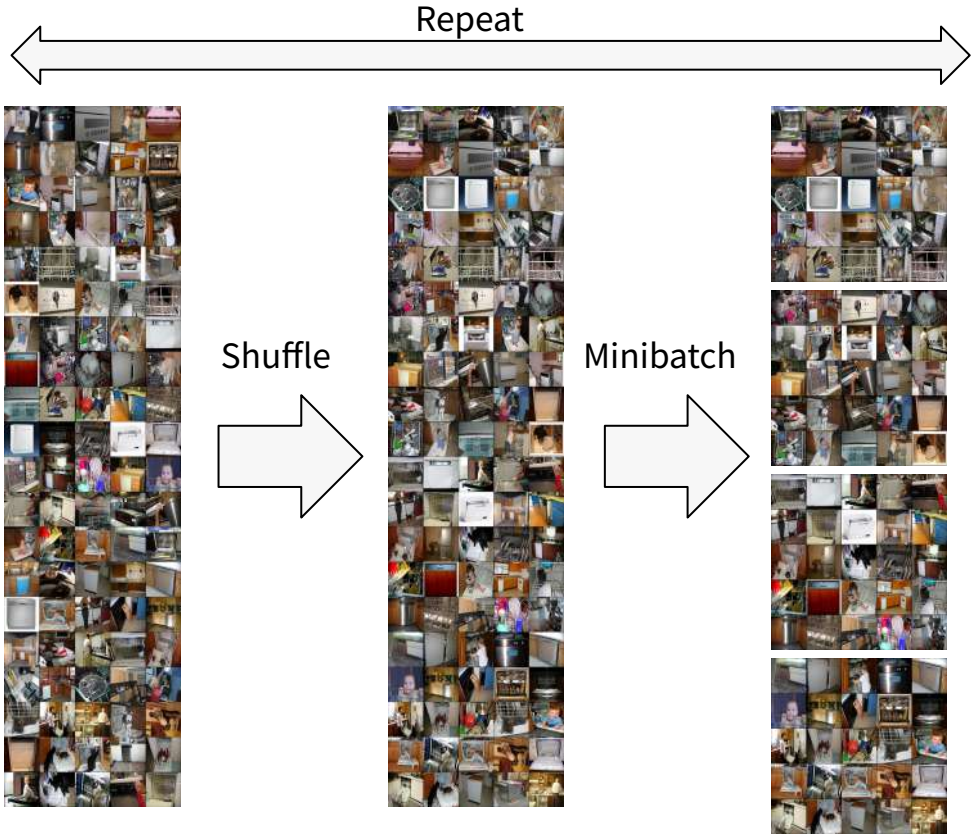
# ストレージから見た深層学習のワークロード（1/2）






Preferred Networks

# SGD-based Random Read




MNIST
- 60k images
- Typical batch size: 10~100

CIFAR10
- 60k images
- Typical batch size: 64

ImageNet-1k
- 1580k images
- Typical batch size: 32

Open Images Dataset (OID)
- 8M images
- Typical batch size: ???

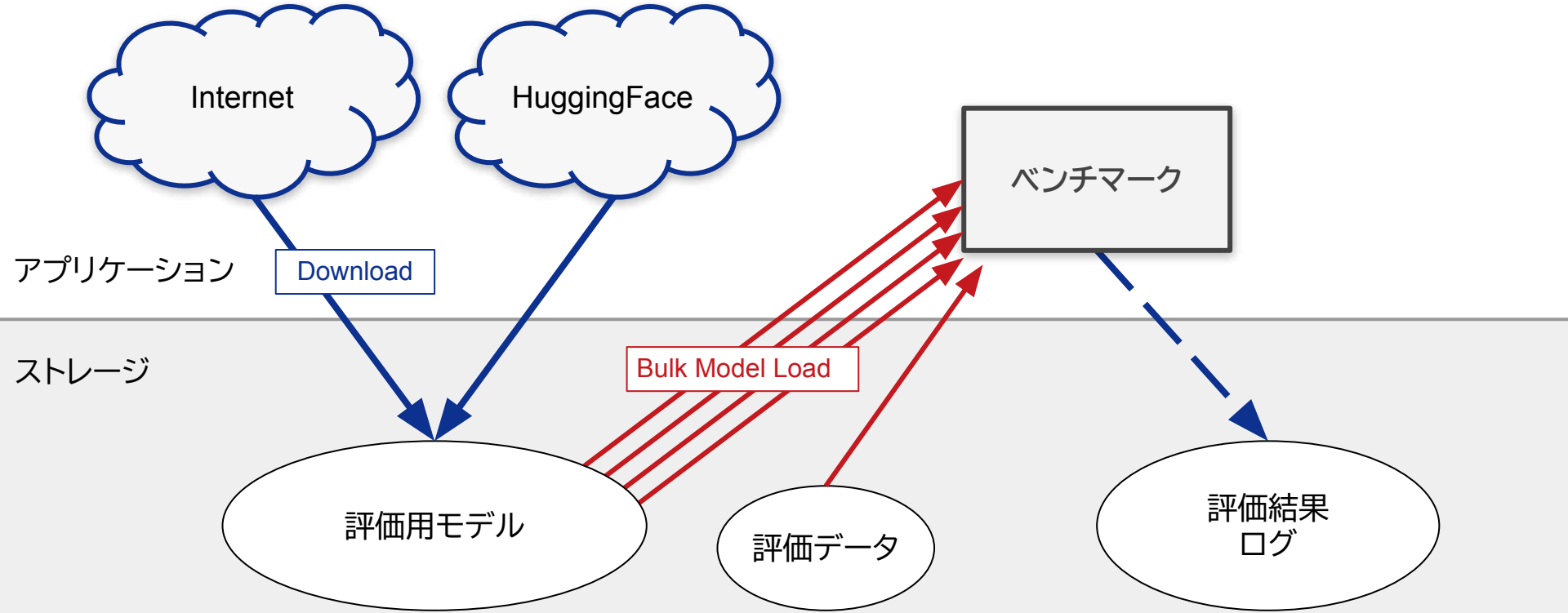


Preferred Networks

# Burst Write

- 故障に備えて学習の途中状態をスナップショットとして保存する
- サイズ
  - CNN-based 画像モデルだと <1GB
  - LLM の場合はパラメータ数にもよるが一般的にはパラメータ数 10B で 120 GB, 100B で 1.2TB
  - 1param あたり 12B
- 速度
  - スナップショットの保存間隔をストレージや計算の MTBF 以内に収めないといけない
  - 一般的には数十分〜数時間以内
  - Kubernetes によるプリエンプションもありうる

# ストレージから見た深層学習のワークロード（2/2）






Preferred Networks

# モデル評価のワークロード

- PLaMo-100B-Pretrained を始めとする自社モデルの事後学習評価
  - 事後学習のタスクごとに多様なパラメータで何度もモデルをストレージからロードして利用
- HF上にある多種多様なモデルを評価比較
  - 数百GBのLLMを100種類ダウンロードして1000件のタスクを評価する
-

# AI インフラに必要なその他の要件

- データセットの保管
  - 画像や動画、テキストベースのデータセット
  - シミュレーション結果
  - 2024/05 時点で総量は 10 PB 程度
- 学習時の高速なデータ読み出し
- 推論時の高速なモデルロード
- システムの水平拡張性
  - 事業が続く限りデータは増え続ける
  - 異なる世代のハードウェアが同一システム下で同居できること
- 単一の名前空間
  - アプリケーションを移植しなくても利用継続できること

# 要件に対する実装

**SGD-based Random Read / Bulk Model Load**

- 低レイテンシ、高い帯域
- 同じデータセットをシャッフルして繰り返し読む
- 大きなモデルを高速ロード

**Dataset Generation by Simulation**

- 増え続けるデータに対応する
- 永続性、バックアップ

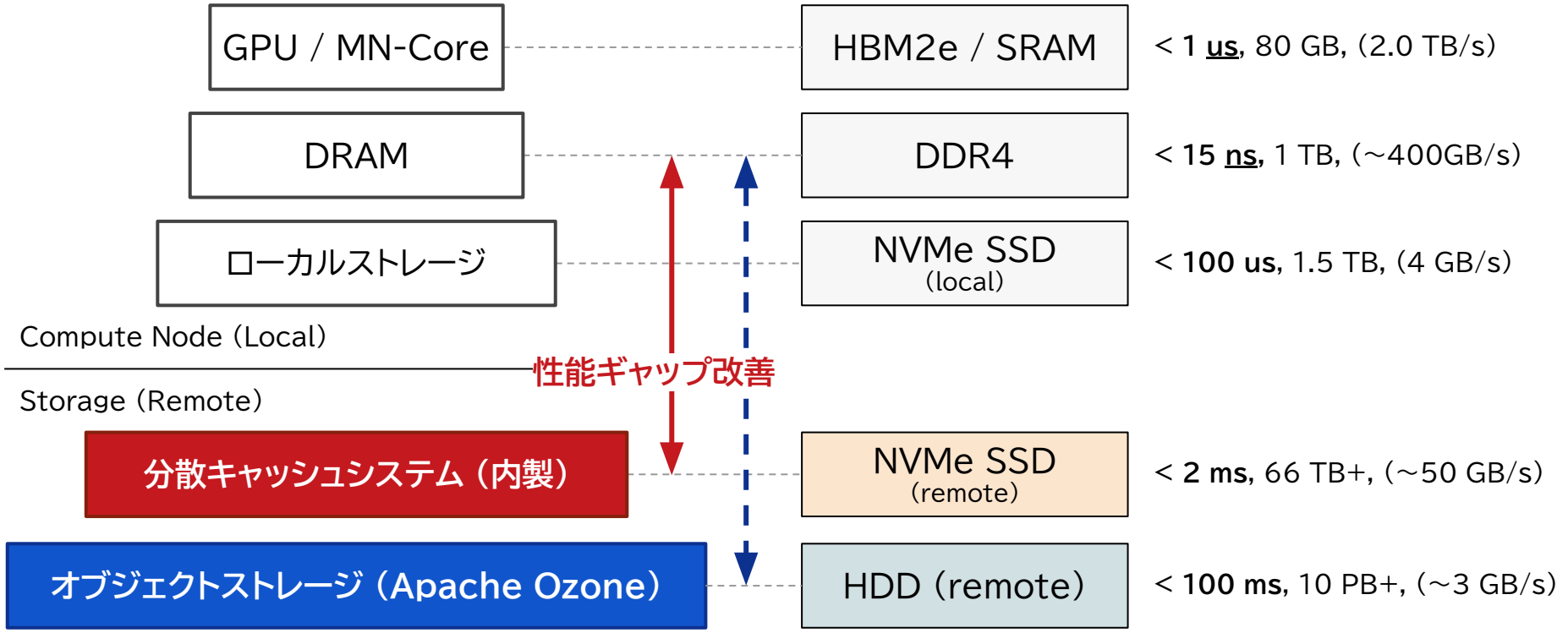
**分散キャッシュ**
NVMe SSD x 10^2

**オブジェクトストレージ**
Apache Ozone
HDD x 10^3



Preferred Networks

# I/O 階層



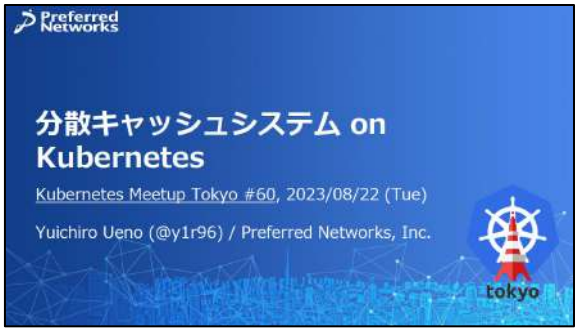



Preferred Networks

# 分散キャッシュシステム（1/2）

- キャッシュシステムの要件
  - SSD の性能を引き出せる
  - データノードを追加して簡単に容量を拡張できる
  - 複数の利用者間で資源を平等に共有できる
- 特徴
  - sendfile(2) によるゼロコピー実装
  - Consistent Hashing によるデータ分散、Envoy による L7 Routing
  - LRU によるデータの入れ替え、クォータ機能
  - Non-POSIX Stateless API (HTTP)

詳細は下記のスライド・ブログを参照




https://tech.preferred.jp/ja/blog/distributed-cache-for-deep-learning/
https://speakerdeck.com/pfn/k8s-tokyo-60-distributed-cache-system



Preferred Networks

# 分散キャッシュシステム（2/2）

- 分散キャッシュシステムはローカルファイルシステム（i.e. JBoF）を想定した設計
  - RAID なし、NVMe-oF なし、永続性なし
  - 高速（OS からみて ~10us）で大容量な一時領域
  - まずは簡単に HTTP で動くものをつくった
  - NVDIMM が狙っていた領域？
- Discussion
  - NVMe-oF をつかわなかった
    - 運用性
    - Consistent Hashing で負荷分散をしたかったので N:M になる
  - 計算機との RDMA（GPUDirect Storage etc.）はしていない
    - Kubernetes のエコシステム（Envoy）に実装されていない
    - Consistent Hashing で負荷分散をしたかった
    - クライアントは User-land ライブラリにせざるをえなかった
  - FUSE という選択肢
    - コンテナ上の CAP_SYS_ADMIN 問題

c.f. pfnet-research/meta-fuse-csi-plugin

# オブジェクトストレージ

- オープンソースソフトウェア選定の検討事項
  - 問題発生時にコードを読める、直せる、データを復旧できる
  - ユーザ空間で動作する
  - 手弁当な小体制でも運用できること

- これまでの運用実績
  - HDFS（～ 2021）→ Apache Ozone（2021 ～）→ ？？？

- Cloudera Manager による省力運用
- 安定性

- Small File Problem への対処
- S3 API サポート



Preferred Networks

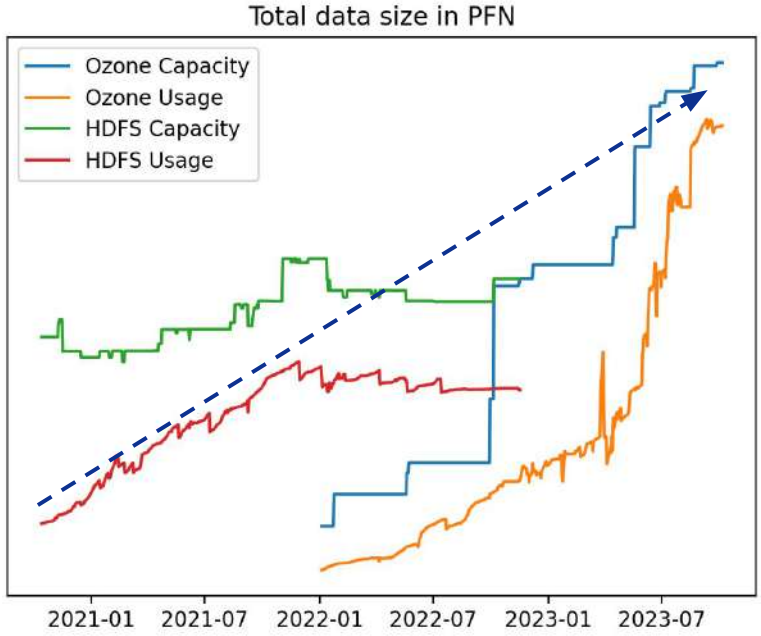
# オブジェクトストレージの構成

現在は 3 系統の Apache Ozone クラスタを自社運用中

| | クラスタ構成 #1 | クラスタ構成 #2 | クラスタ構成 #3 |
|---|---|---|---|
| 総物理容量 | 18.1 PB | 12.1 PB | 9.0 PB |
| データノード台数 | 36 | 88 | 24 |
| デバイス | SATA HDD | SATA HDD | SATA HDD |
| インターコネクト | 100 GbE x2 | 100 GbE x2 | 100 GbE x2 |
| ソフトウェア | Apache Ozone 1.3.x | Apache Ozone 1.3.x | Apache Ozone 1.4.0 |
| 運用開始 | 2021 | 2022 | 2024 |
| 構成 | | OM/SCM HA | OM/SCM HA |

# 社内の容量需要の増加

- ストレージの容量需要は一定ラインで継続
- 毎日データノードの HDD が数本埋まる（10 ～ 20TB/day）

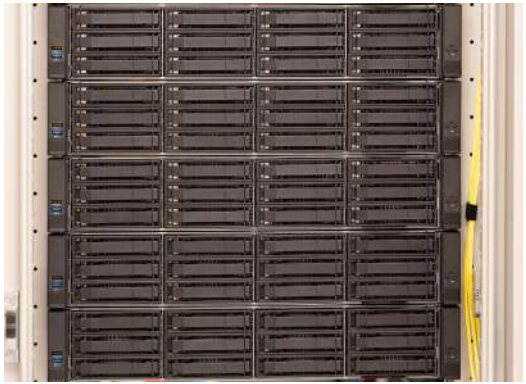




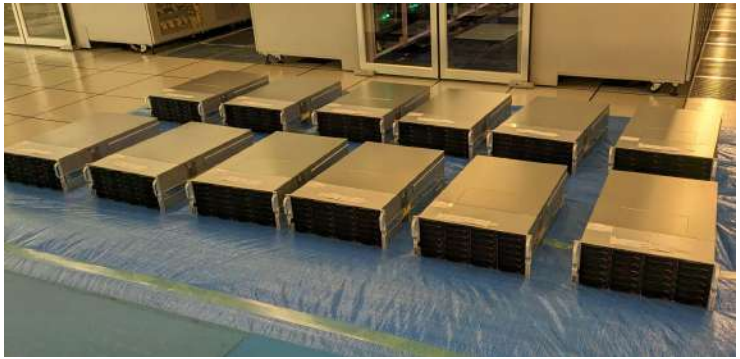

Preferred Networks

# HDD 増設の歴史





2022



2023

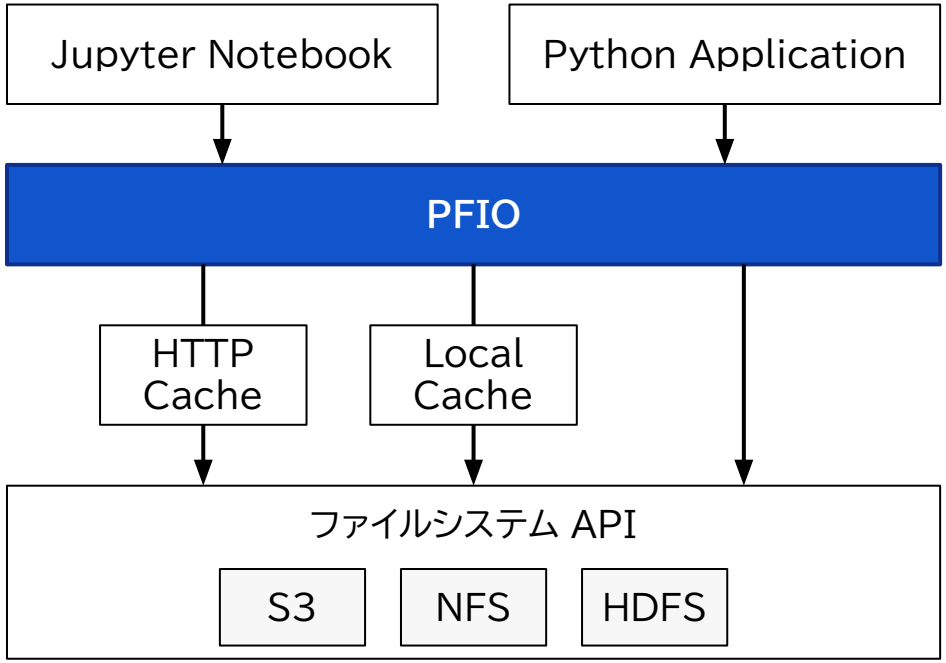
増設はほぼ年 1 回
サーバー数十台、HDD 数百台



Preferred Networks

# インタフェースの抽象化

https://github.com/pfnet/pfio
https://pfio.readthedocs.io/en/latest/index.html




```
[ozone]
scheme = s3
endpoint = ...
```

```python
import pfio

backend_fs = pfio.v2.from_url('ozone://bucket/');
with pfio.v2.HTTPCachedFS(url = 'http://cache/',
                          fs = backend_fs) as fs:
    with fs.open('bin-1m.dat', 'rb') as fp:
        buf = fp.read()
        print(len(buf))
```

図: 分散キャッシュと S3 API の抽象化

```
% python3 ./example.py
1048576
0.12042808532714844
```

→

```
% python3 ./example.py
1048576
0.017198801040649414
```

図: キャッシュによるレスポンスタイム改善



Preferred Networks

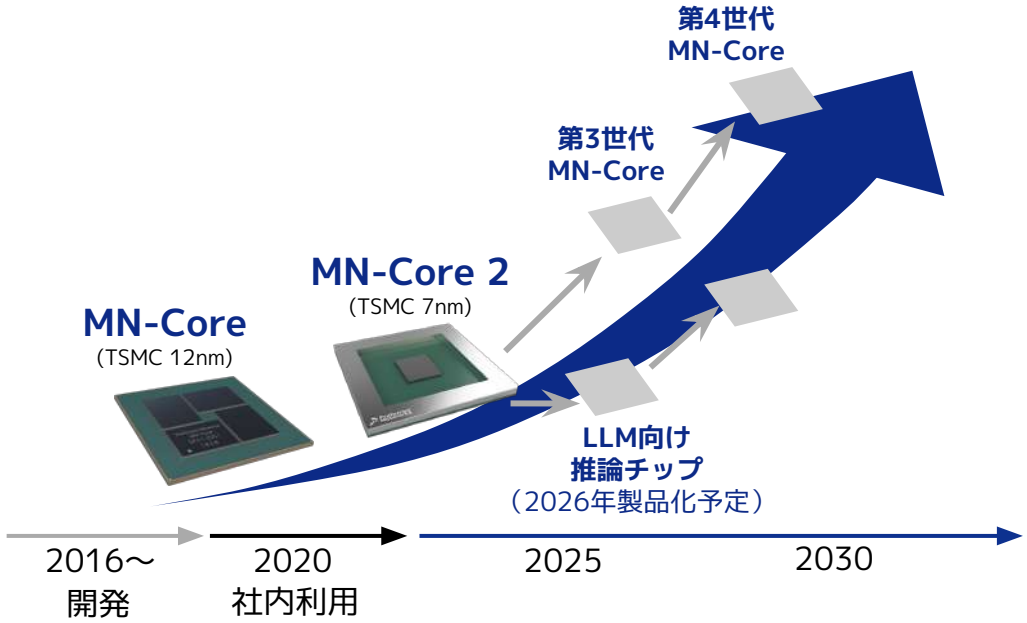
# 今後の展開（会社）



Preferred Networks

# MN-Core™シリーズの進化と今後の展開

●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 の販売を開始。クラウドでの提供も予定。




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



Preferred Networks

# MN-Core クラウドサービス化

PFN、深層学習用プロセッサ

preferred.jp/ja/news/pr20231016/

Preferred Networks | Products | Projects | Research & Development | News | Company | Careers | Contact

JA EN

Home > News > PFN、深層学習用プロセッサーMN-Coreを Matlantisの深層学習モデル（PFP）の計算基盤として実装

News

News Release #CEATEC #Matlantis #MN-Core

PFN、深層学習用プロセッサーMN-Coreを Matlantisの深層学習モデル（PFP）の計算基盤として実装

実際のワークロードで安定して3倍程度の高速化を確認

2023.10.16

株式会社Preferred Networks（本社：東京都千代田区、代表取締役 最高経営責任者：西川徹、以下、PFN）は、神戸大学と共同開発した深層学習を高速化するプロセッサーMN-Core™を、汎用原子レベルシミュレータMatlantis™の深層学習モデル（PFP[*1]）の計算基盤として実装し、Matlantisの販売をおこなう株式会社Preferred Computational Chemistry[*2]（以下、PFCC）を通じてENEOS株式会社（以下、ENEOS）に提供を開始しました。

PFNは、MN-Coreを搭載したスーパーコンピュータMN-3を2020年に稼働させ、膨大な計算力を必要とする深層学習の計算速度の高速化を進めてきました。PFNが事業として取組む深層学習の実用的なワークロードにおいてMN-Coreの性能評価を行った結果、GPUを搭載したPFNのスーパーコンピュータMN-2と比較して、三次元モデル換元で10倍以上、画像認識モデル自動最適化で7倍以上の高速化を確認しています。これらの性能評価に加え、専用ソフトウェアスタックの開発も進捗したことから、これまで社内の研究開発利用に制限していたMN-Coreを、Matlantisのコア技術である深層学習モデルの計算基盤として実装しました。MN-Coreを計算基盤とするMatlantisは、2023年8月からクラウド経由でENEOSに提供され、分子動力学計算を使った密度・粘度・熱伝導・比熱などの物性データ生成などの実際のワークロードで安定して3倍程度高速化されるなどの有用性が確認されています。

Matlantisは、材料探索の多様なワークロードに柔軟に対応するため、複数種のGPUなどを組み合わせたシステムを構築しています。MN-Coreを Matlantisの計算基盤として使用することにより、PFNはハードウェアから深層学習アーキテクチャまで高度にチューニングを行い、材料探索シミュレーションに最適化した計算資源を安定的に提供していきます。また、MN-CoreによるMatlantisの高速化で、これまで以上に現実世界に即した

山田てるみ
@telmin_orca

これがMN-Coreね！！

Translate post



12:52 PM · Oct 20, 2023 · **2,304** Views

5 25 1

https://www.preferred.jp/ja/news/pr20231016/

# MN-Core2 クラウドサービス構想

Preferred Networks

News Release #MN-3 #MN-Core #MN-Core 2

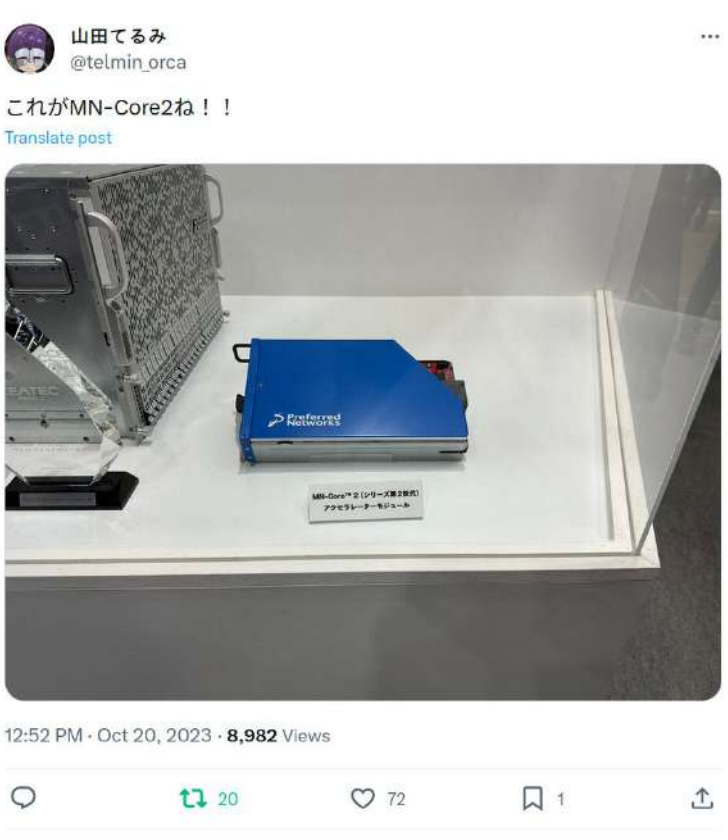
## PFN、深層学習を高速化するプロセッサーMN-Core 2の開発および、MN-Coreシリーズのクラウドサービス構想を発表

**2023年春からPFNのパートナー企業向けにMN-3の計算資源を提供し、順次設備とユーザーを拡大予定**

2022.12.14

株式会社Preferred Networks（本社：東京都千代田区、代表取締役 最高経営責任者：西川徹、プリファードネットワークス、以下、PFN）は、深層学習を高速化するディープラーニング・プロセッサー MN-Core™ 2（エムエヌ・コア・ツー）を、東京ビッグサイトで開催されている SEMICON Japan 2022のキーノート講演において本日発表しました。





誰もが MN-Core™ シリーズを利用できる AI クラウドサービス

# Preferred Computing Platform

Preferred Networks

Preferred Computing Platform（以下、PFCP）は株式会社 Preferred Networks（以下、PFN）が構築運用する深層学習・AI ワークロード向けのクラウドサービスです。PFNが開発する独自アクセラレータであるMN-Core™ シリーズを唯一利用でき、最先端の性能と効率性を備えています。

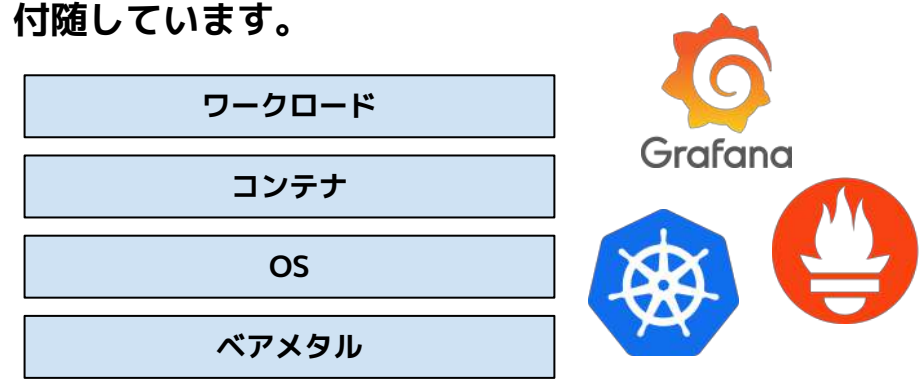
## 強力な計算ノード

MN-Core ボードを4基搭載した MN-Core サーバを複数専有して利用できます。すべてのノードは深層学習に最適化された高速なネットワークで相互に接続されています。次世代の MN-Core 2 や NVIDIA GPU を搭載したノードも順次提供予定です。

| MN-Core 2 サーバの構成 | |
|---|---|
| アクセラレータ | MN-Core 2 ボード × 8 |
| CPU | Intel Xeon SPR 8480+ 56C x 2 |
| Memory | DDR5 64GB x 16 (2TB) |
| Interconnect | NVIDIA ConnectX-6 (100GbE) × 2 |

## フルマネージドサービス

深層学習・AI ワークロード向けに拡張された Kubernetes クラスタをマルチテナントで利用できます。大規模分散学習から推論サーバの高可用な運用まで幅広く行なえます。ワークロードの状況を観測するためのマネージドなモニタリングサービスも付随しています。



# 今後の展望（ストレージ）

# 今後の展望

- SGD-based Random Readワークロード
  - 分散キャッシュの改善
    - 最新世代NVMe SSDの導入
    - RDMA等の導入
- Simulation Data
  - 自社SDSの開発(今はどこも自社開発している)
  - 高密度HDDサーバーの導入
  - HDDはよりアーカイブストレージ化していくが、NANDの単価次第
- LLM Snapshot
  - Transparent HSM-like なシステムの検討

# 今後の展望: PCIe の高速化

- OSまではソフトウェアが追いついているが、その上のインフラソフトウェアが Gen5 の高速化にすら追随できていない
  - e.g. HDD でいうところの RAID（狭義）等の技術が追いついておらず、運用しやすい形でファイルシステムAPIを提供する手段が限られている
  - e.g. Gen5 NVMe が安定動作するサーバー機材が限られている（?）
  - e.g. HDD を大量に並べても IO スループットが出ないケースがある
- 400 Gbps の NIC ベンダーは実質 1 社独占であり、SSD だけが高速でも使いにくい状態（PFNは200Gbpsをスキップしました）
  - 他社が 400 を発表する前に 800 Gbps NIC が発表された
- ストレージ API としては以下を継続利用
  - Relaxed POSIX（NFS, etc）
  - S3（HTTP）

# Preferred Networks

Making the real world computable