Preferred Networks

プライベートスーパーコンピュータの悩みと愉しみ

2023/08/30

株式会社Preferred Networks
計算基盤担当VP 土井裕介

自己紹介

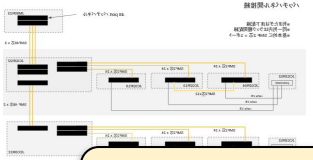
- 2000-2016 株式会社東芝
  - 分散KVS、RFID、データ表現、無線IoT等
- 2016-現在 株式会社Preferred Networks
  - Computer Networks x Deep Learning
  - → 社内IT整備 (チーム立ち上げて脱退)
  - → クラスタ整備 (チーム立ち上げ)
  - → 計算基盤担当 (2019/9より)

現在の担当領域: クラスタMiddleware (kubernetes等) クラスタ調達設計 (物理基盤含)、独自開発アクセラレータ (MN-Core)

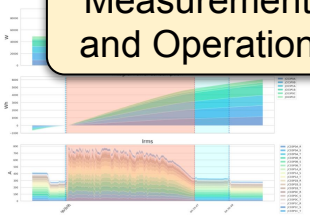
Preferred Networks

計算基盤いろいろやっています

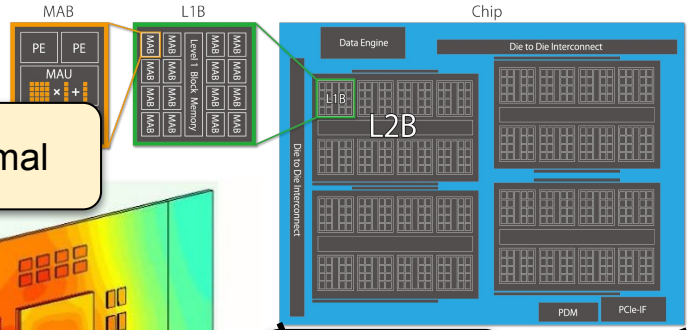
DC and Cluster

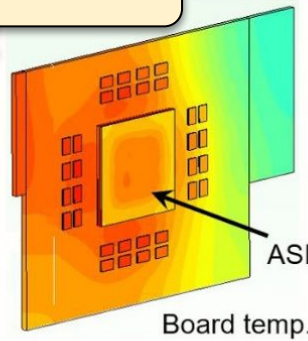


Measurement and Operation

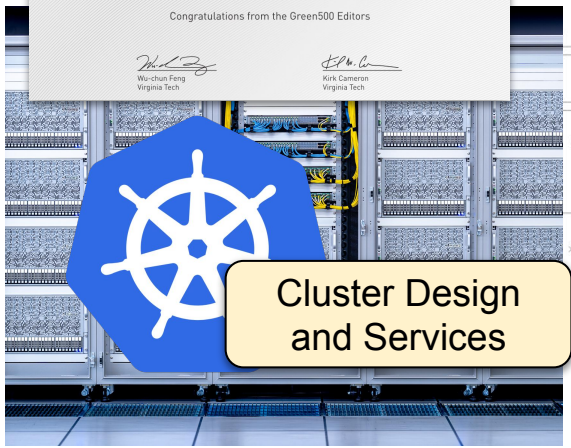


Thermal

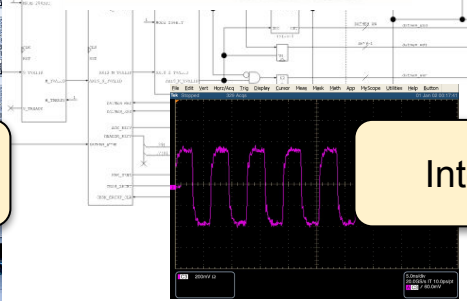




ASIC Design



Cluster Design and Services



Interconnect

Preferred Networks

Preferred Networks 会社概要

Preferred Networks

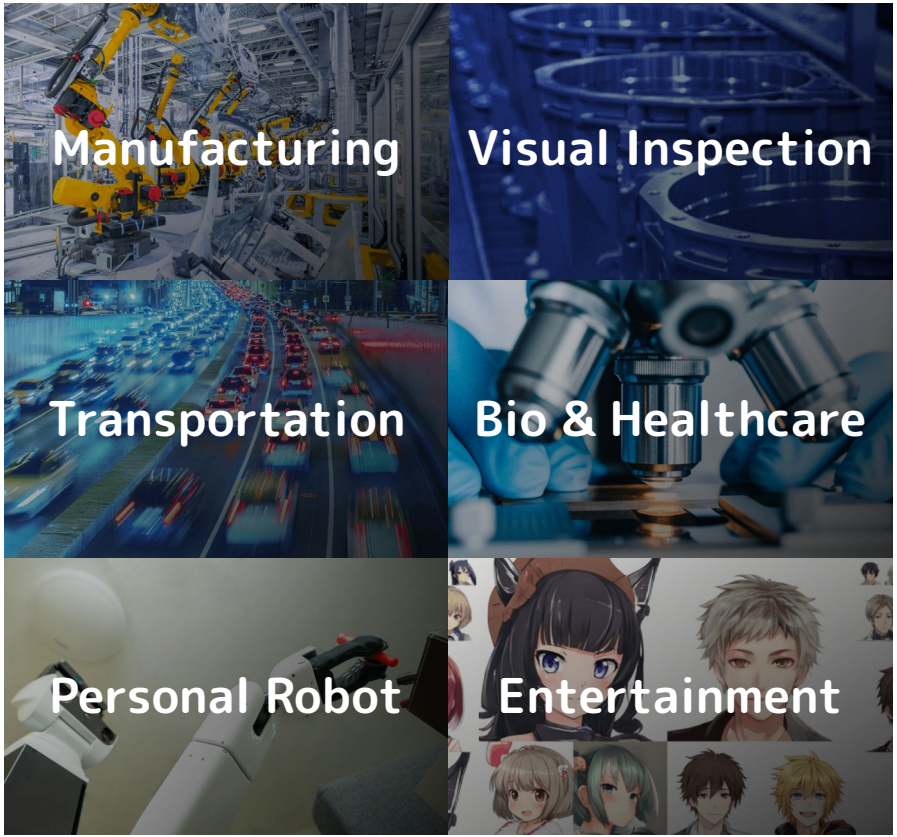
現実世界を計算可能にする

自分たちの手で革新的かつ本質的な技術を開発し、未知なる領域にチャレンジしていく。

私たちはソフトウェアとハードウェアを高度に融合し、自動車やロボットなどのデバイスをより賢く進化させます。常に変化する環境や状況に柔軟に対処できる賢いデバイスができれば、物理世界をリアルタイムにセンシングし、現実世界そのものが計算可能になります。

技術を使って、自分たちが見たことが無い、まだ知らない世界を知りたい。すでにわかっている領域で勝負するのではなく、技術の力で想像を超えた世界に挑戦していきます。

会社情報



設立	2014年3月26日
経営陣	代表取締役 最高経営責任者　西川 徹 代表取締役 最高研究責任者　岡野原 大輔
所在地	本社 東京都千代田区大手町1-6-1大手町ビル 米国子会社 Preferred Networks America, Inc. 330 Primrose Rd., Suite 300, Burlingame, CA 94010
従業員数	約300名

グループ企業の提供サービス: Matlantis（PFCC）

グループ企業の提供サービス: Kachaka (Preferred Robotics)

Preferred Networks

豊富な計算資源と高度な技術を基盤に複数の事業を創出

Business

Manufacturing	Transportation	Bio & Healthcare
Personal Robot	Visual Inspection	Entertainment

PFN Technology

Computer Vision（コンピュータビジョン）

Object Detection（物体検出）	Pose（ポーズ推定）
Navigation（ナビゲーション）	Scene（シーン解析）
Visual Inspection（外観検査）	Image Segmentation

Data Analytics（データ解析）

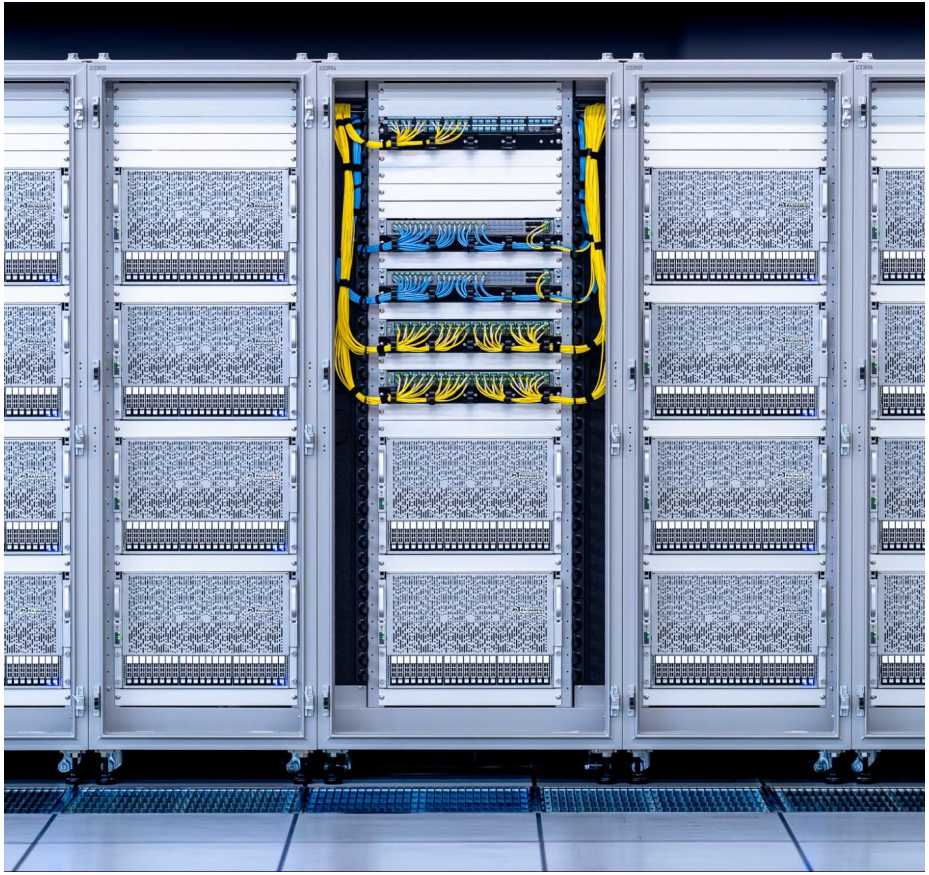
- Anomaly Detection（異常検知）
- Optimization（最適化）
- Time series data（時系列データ）

Machine Learning and Deep Learning（機械学習と深層学習）

Infrastructure （インフラ技術）

自社計算基盤の紹介

自社計算基盤：MN-2, MN-3



深層学習をはじめとするPreferred Networks(PFN)の中核技術は膨大な計算を要求します

PFNでは、多量の計算を効率的に実行するために独自の計算機クラスターを複数運用しており、現在はMN-2、MN-3が稼働しています

MN-3は、PFNが自社開発した深層学習用プロセッサーMN-Coreを初めて用いた計算機クラスターです

- **MN-2：1500GPU規模クラスタ**
- **MN-3：160 MN-Core規模クラスタ（写真）**

MN-3のサーバ1台あたりのスペック（MN-3は、以下サーバが48台で構成）

MN-Core	MN-Core Board x 4
CPU	Intel Xeon 8260M 2way （48物理core）
Memory	384GB DDR4
Storage Class Memory	3TB Intel Optane DC Persistent Memory
Network	MN-Core DirectConnect（112Gbps） x 2 Mellanox ConnectX-6（100GbE） x 2 On board（10GbE） x 2

PFNのオンプレKubernetesクラスタ

拠点名: MN-J

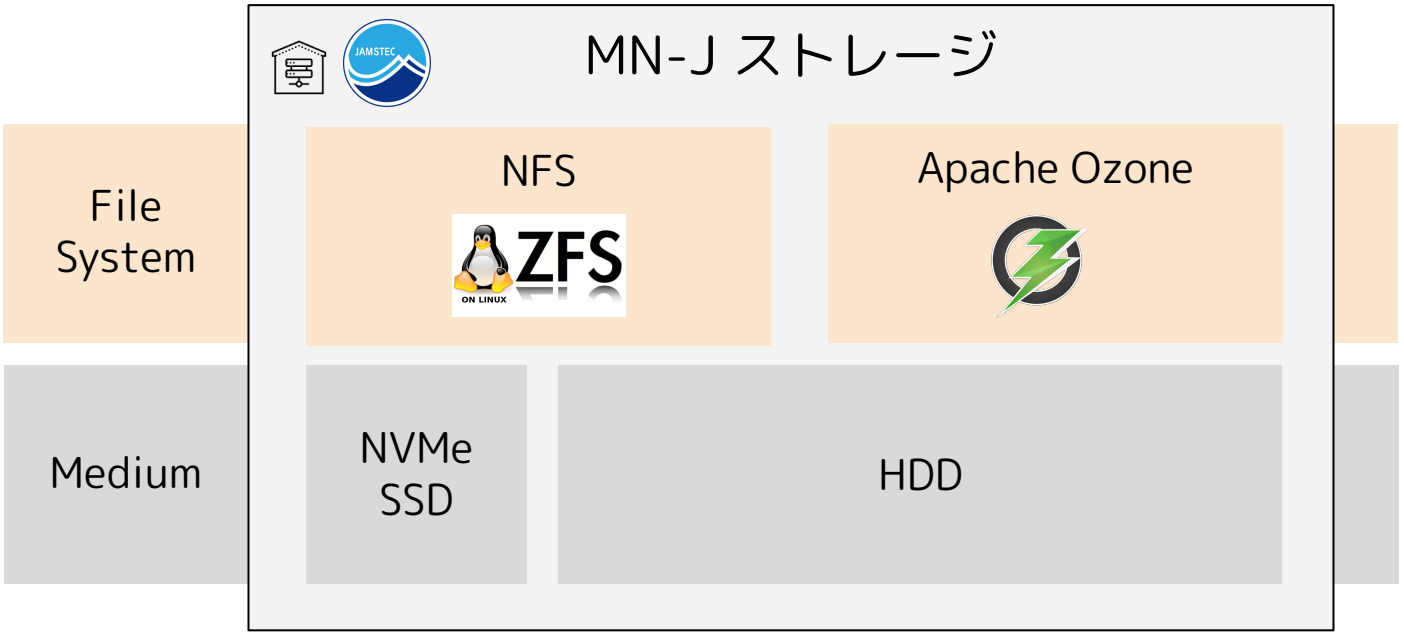
MN-2a	MN-3	MN-2b (A100)	MN-2b (A30)
128 nodes (1024 GPUs)	**48 nodes (192 MN-Cores)**	**42 nodes (168 GPUs)**	**42 nodes (252 GPUs)**
36 CPU Cores	48 CPU Cores	128 CPU Cores	80 CPU Cores
DDR4 384GB	DDR4 384GB	DDR4 1024 GB	DDR4 512 GB
V100 (16 / 32 G) SXM2 x 8	MN-Core x 4	A100 (80G) SXM4 x 4	A30 (24G)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

Preferred Networks

PFNのオンプレストレージクラスタ

拡大中!

トータル **約 10 PB** (論理容量)



Icon pack by Icons8 - https://icons8.com

Preferred Networks

何故「プライベート」(2016)

Preferred Networks

なぜわざわざ自前の計算機を用意するのか?

- クラウドでいいじゃないか → 当時のクラウドGPUは...

- ABCIつかえば → 当時はなかった
(今はお世話になりまくりです!)

Preferred Networks

なぜか? ⇒ それが必要だったから

- 計算機が足りない
- 計算機間の通信路が足りない
- ドライバとかいろいろいじくりまわしたい

Preferred Networks

プライベート「スーパーコンピュータ」

- 作っているモノは、いくつかの意味でスパコンに近い
  - GPUどっさり
  - **“Interconnect” の存在**

Preferred Networks

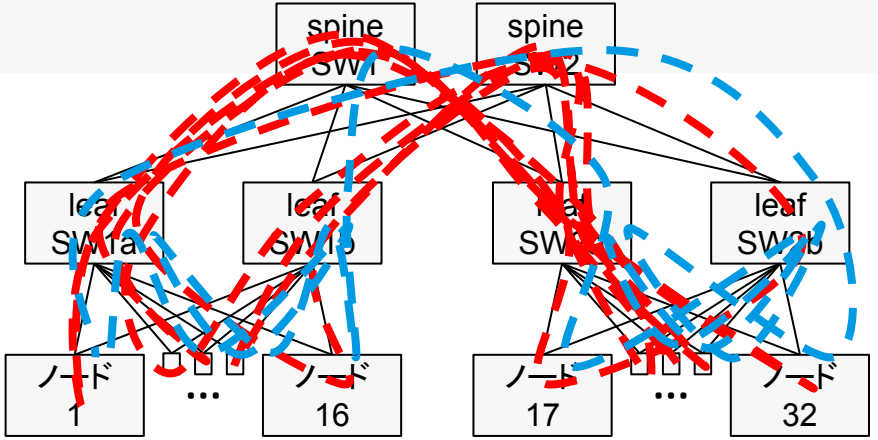
トポロジのマッチング

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

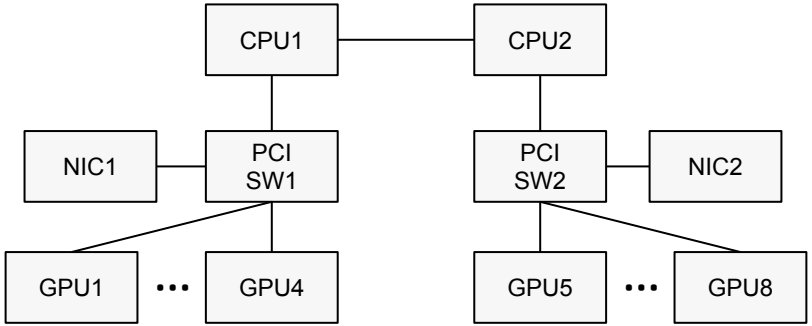
leafの中ではfull bisectionで2リンク分の
速度が(理論上)出る→ tree allreduce
leafをまたぐとleaf-spineの帯域で律速
→ ring allreduce

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: leaf-spine 2:4:32

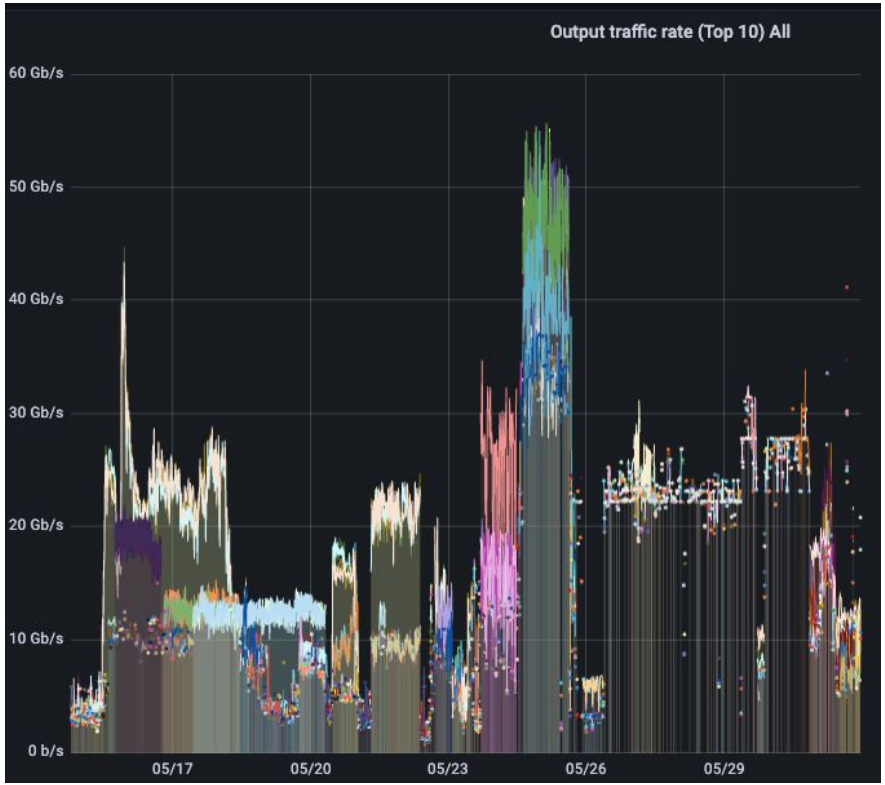


例: 古典的GPUサーバ CPU:GPU = 2:8

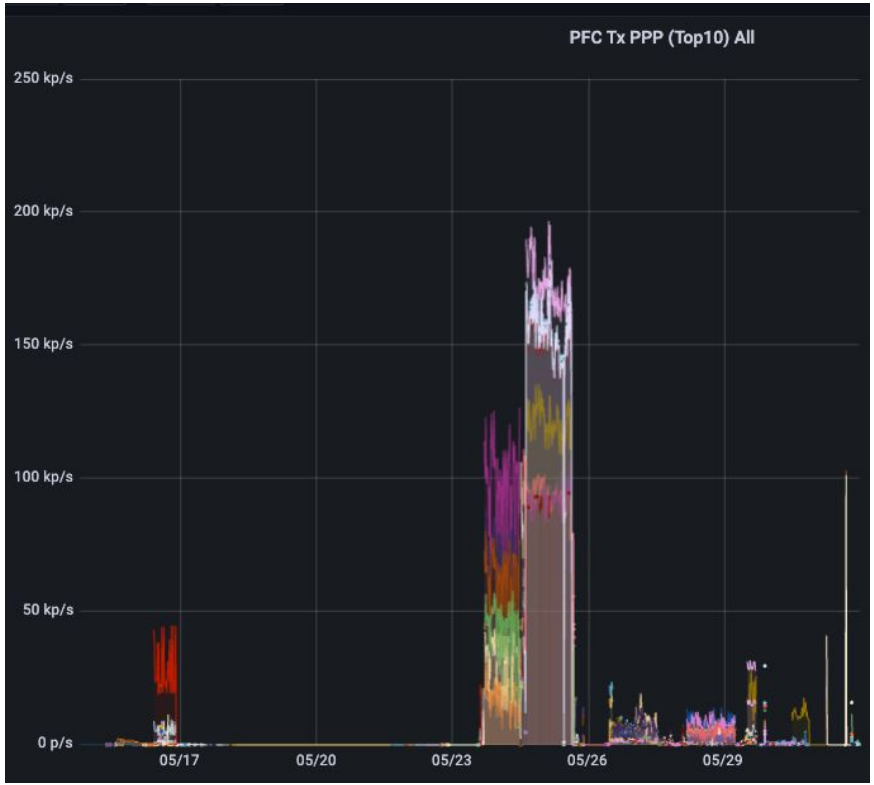


Preferred Networks

実例



インターフェイスごとのトラフィック（100GbE）

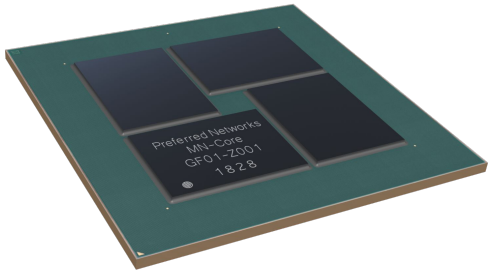


PFC count

Preferred Networks

プライベート「スーパーコンピュータ」

- 作っているモノは、いくつかの意味でスパコンに近い
  - GPUどっさり
  - “Interconnect” の存在
  - **「謎の半導体」**



Preferred Networks

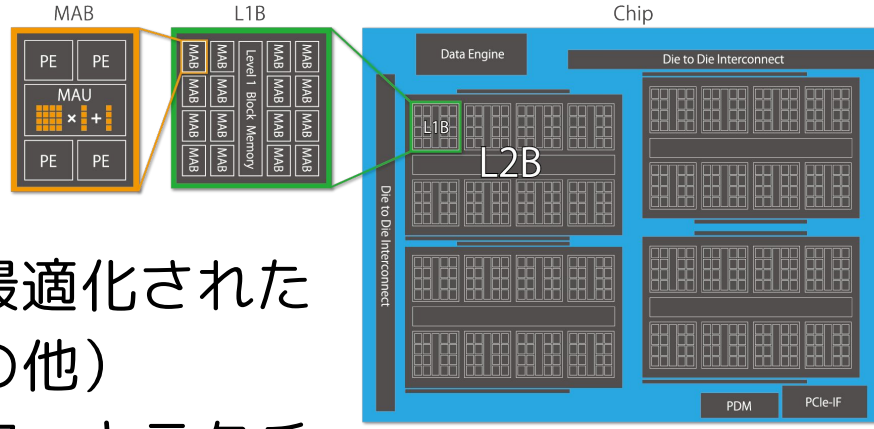
何故「プライベート」
⇒ MN-Coreをモノにするために

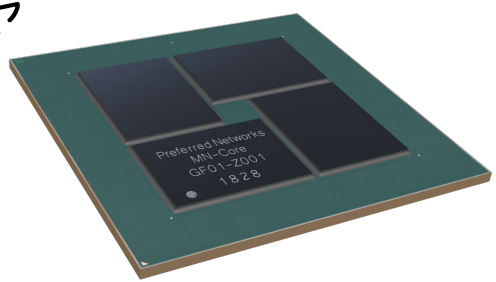
Preferred Networks

MN-Core：自社開発の機械学習チップ

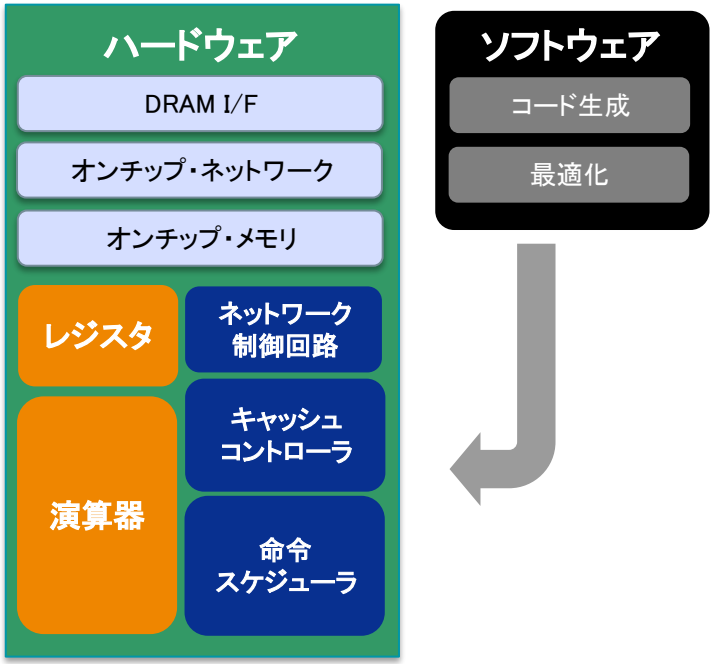
「自社の計算需要に最適なチップを内製」
（神戸大学と共同研究）

- 計算機の性能が深層学習の
  性能の最も重要な差異化要因
- チップ内階層構造と機械学習に最適化された
  ネットワーク（放送、縮約、その他）
- ソフトウェア最適化を優先したアーキテクチャ
  - ソフトウェア側でハードウェアの動作を
    詳細に制御できる構造

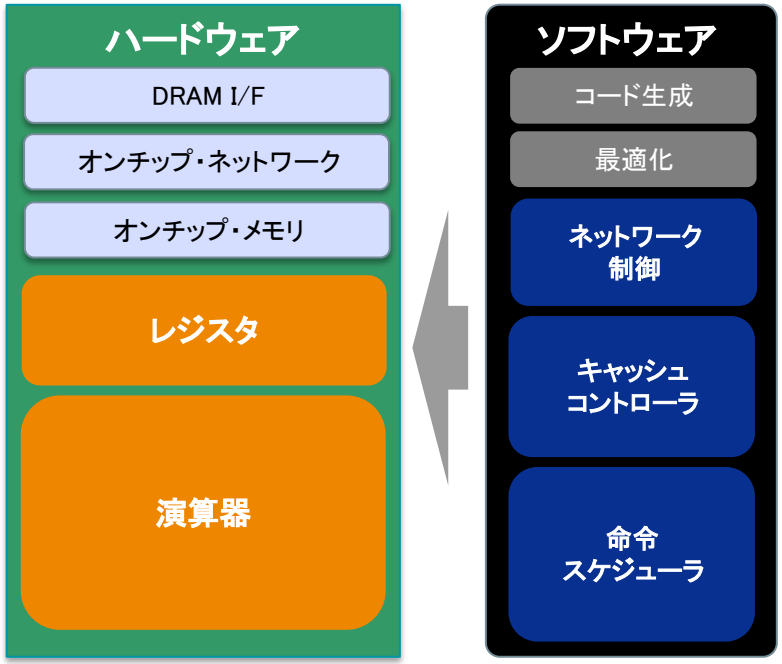




ハードウェアの演算器数を最大化

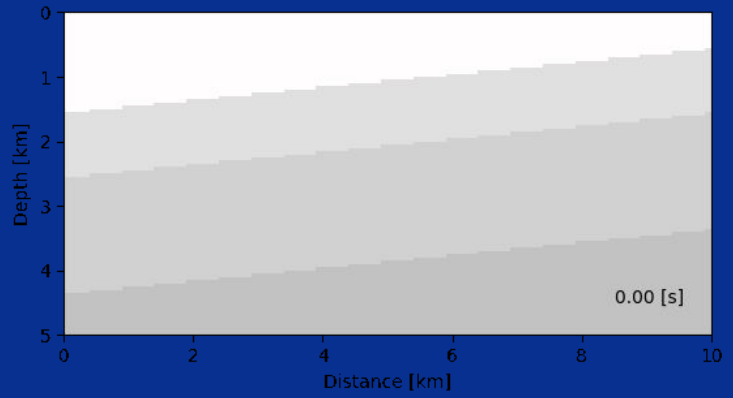


汎用プロセッサー



MN-Core

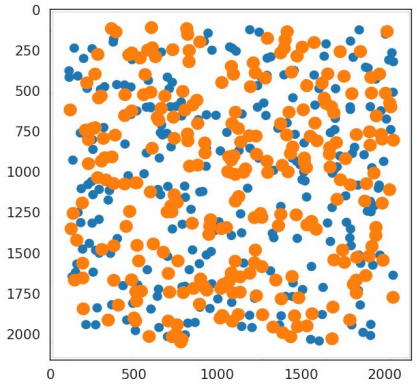
CGによる物体認識の学習



地震波による地層解析

テキスト条件付き画像生成

例: “Christmas tree in a datacenter”



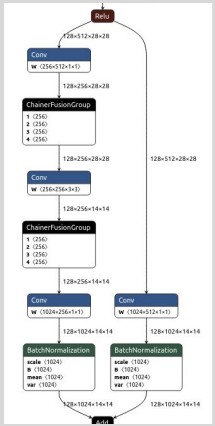
分子動力学

MN-Core コンパイラスタック (内製)

PyTorch

ONNX

ONNX model



L3IR
DNN op level, SIMD parallelism strategy, ……

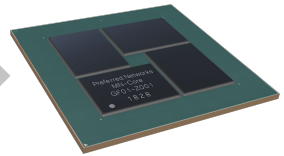
- MNGraph
- Global Layout Planner
- Re-computation Scheduler

L2IR
ndarray op level, optimized DNN op impl, ……

- PEVector
- MNTensor
- Reshape Impl.
- Generic Conv Impl.

L1IR
MN-core op level, memory allocation, scheduling, optimization, ……

- Scheduling Graph
- Layer Impl.
- Instruction Merger



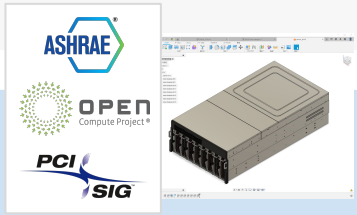
Preferred Networks

「今」と「これから」

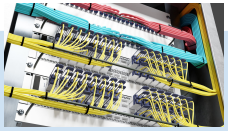
Preferred Networks

物理基盤構築運用のためのもろもろ

技術
- サーバ
- ネットワーク
- ファシリティ



計算機新技術検討、
機器選定・筐体設計

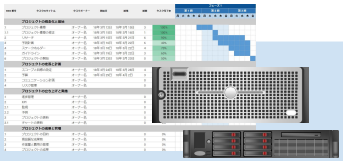


NW新技術検討、
論理・物理設計

電力/冷却設備検討、
DC/設備設計

など

計画
- 長期計画
- 年次計画
- 設備企画



ロードマップ作成

予算計画、
設備計画

基盤設計、
工事計画

など

管理
- 調達
- 設備工事
- 保守運用

資材/機材調達、
ロジ/Kitting手配

施工管理、
チーム間連携

保守手配・部材管理

など

Preferred Networks

「これから」の悩みと愉しみ

MN-Core™ 2 ベースのクラスタの立ち上げ + さらに先へ
上から下まで全部やってて(大変|愉しい)

↑関連blog (本日公開)

カジュアル面談お申し込みはこちら↑

Preferred Networks