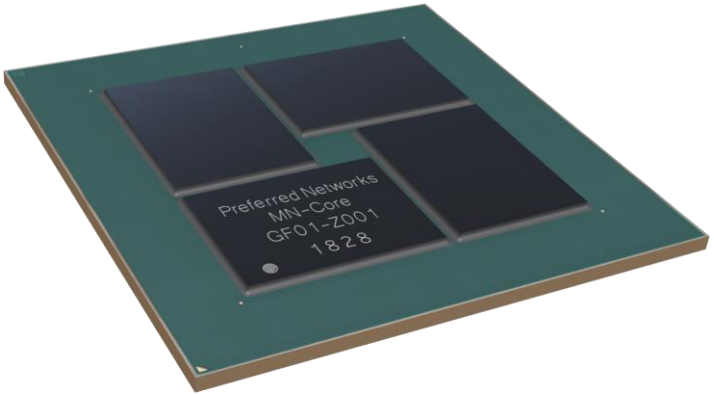
2020/12/15

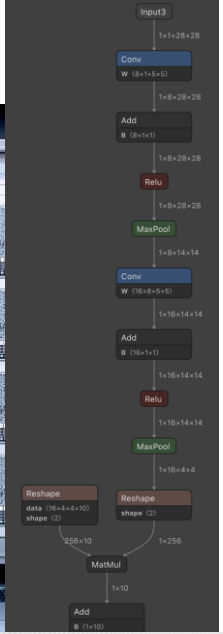
PFNにおけるDeep Leaning アクセラレータの開発について

平木　敬
Preferred Networks

Preferred Networks

本資料はPFNの金子 紘也さん作成の資料
をベースにしています





世界で最も電力性能の
高いスパコンと
認定されました
@2020/06

Name	CPU [Core]	Memory [GiB]	GPU	MN-Core	Links	Notes
2	90% (42.5/47.0)	75% (280.67/374.07)	-	0% (0/4)	Issue bmcURL	(alpha)
3	99% (46.5/47.0)	79% (296.67/374.07)	-	100% (4/4)	Issue bmcURL	(alpha)
4	97% (45.5/47.0)	95% (354.67/374.07)	-	25% (1/4)	Issue bmcURL	(alpha)
5	101% (47.5/47.0)	99% (370.67/374.07)	-	100% (4/4)	Issue bmcURL	(alpha)


The GREEN 500 CERTIFICATE

MN-3 - MN-Core Server, Xeon 8260M 24C 2.4GHz, MN-Core, RoCEv2/MN-Core DirectConnect

Preferred Networks, Japan

is ranked

No. 1 in the Green500

among the World's TOP500 Supercomputers

with 21.11 GFlops/Watt Linpack Power-Efficiency

on the Green500 List published at ISC 2020 Digital Conference, June 22nd, 2020

Congratulations from the Green500 Editors

Wu-chun Feng
Virginia Tech

Kirk Cameron
Virginia Tech

自己紹介 – 平木　敬(Kei Hiraki)

- 2019/4- 株式会社Preferred Networks, Senior Researcher
  - MN-Coreの開発/実装, 社内教育の講師

- MN-Core は私が開発に関係した7個目のコンピュータシステム
- 石川先生、佐藤先生には長年お世話になっています

目次

- PFNにとっての計算能力の位置付け
- 代表的なDeep Learningの高速化手法
- なぜ今プロセッサ開発なのか？
- MN-Coreの概要と開発チームの働き方
- 最近の成果

PFNを支える技術と事業内容

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

Manufacturing | Transportation | Bio & Healthcare

Personal Robot | Visual Inspection | Entertainment

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）

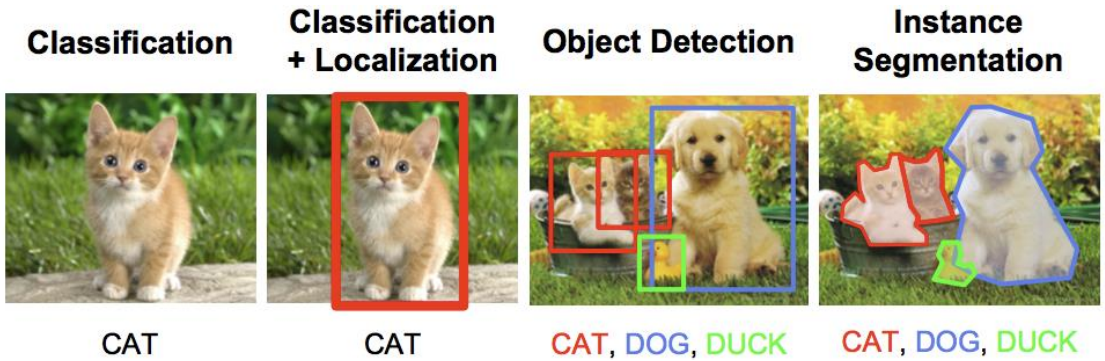
- Anomaly Detection（異常検知）
- Optimization（最適化）
- Time series data（時系列データ）

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

Infrastructure （インフラ技術）

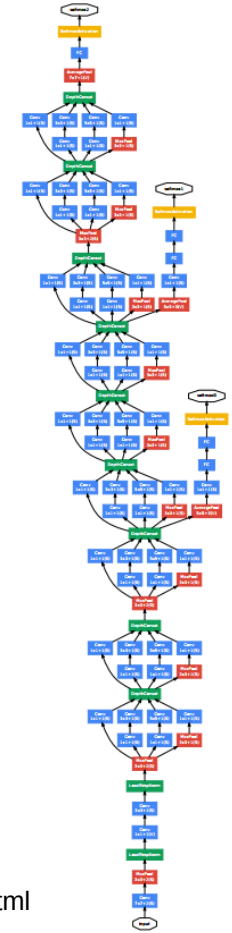
ディープラーニング（深層学習）とは

- 層が深く、幅も広いニューラルネットワーク
  を利用した機械学習手法の一手法
- 画像認識、音声認識、強化学習、自然言語処理
  などで劇的な精度向上を果たし、その多くが既に実用化されている



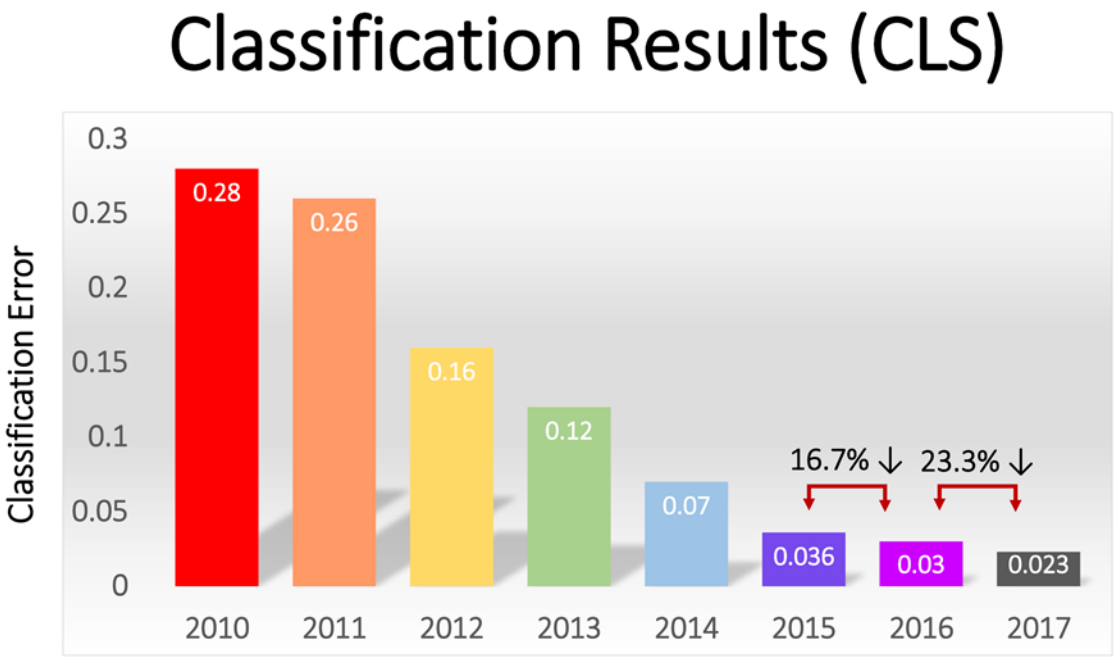
The graph was excerpted from https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/object_localization_and_detection.html

*http://memkite.com/deep-learning-bibliography/



2014年の一般画像認識コンテストで優勝した
22層からなるのGoogLeNetの例 [Google 2014]

Image Classification の進歩



The graph was excerpted from Eunbyung Park (2017).
Overview of ILSVRC 2017

2012: AlexNet
2014: GoogLeNet
2016: ResNet

既に人の認識率を超えつつある

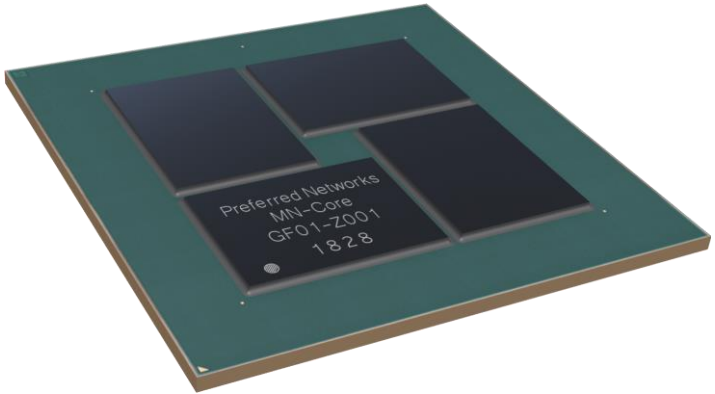
目次

- PFNにとっての計算能力の位置付け
- 代表的なDeep Learningの高速化手法
- なぜ今プロセッサ開発なのか？
- MN-Coreの概要と開発チームの働き方
- 最近の成果

Deep Learning の高速化

- Scale-out（分散深層学習）
  - たくさんのプロセッサをつなげて高速化
    - Keywords: データ並列、モデル並列 , 計算と通信のオーバラップ

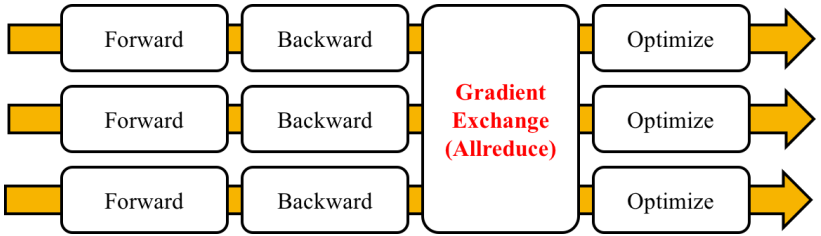
- Scale-up（専用アクセラレータ）
  - はやいプロセッサを使って高速化
    - Keywords: Inference/Training専用アクセラレータ

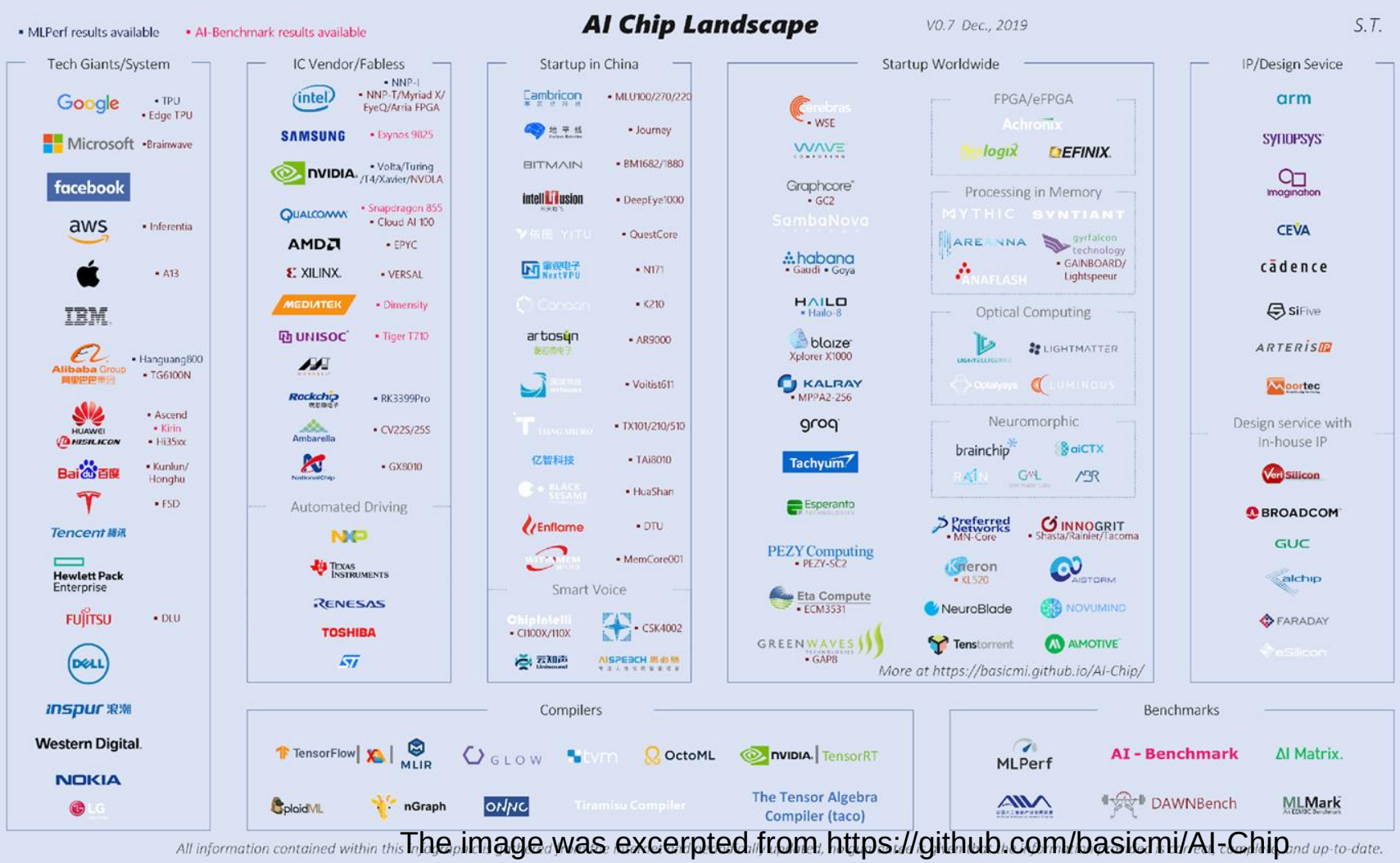


同期型データ並列による学習の高速化

- 同期型データ並列によって、2年弱で100倍以上高速化している
  - 元々の論文では8枚のGPUで数週間要していたものは、今や2.2min
  - バッチサイズを増やせる問題については、GPU台数に対してほぼリニアに性能向上が達成できる程度にノウハウがたまりつつある
- 常に適用できる万能な手法ではない
  - バッチサイズを増やしても精度や学習の安定性に問題が出ないモデルのにのみ適用可能

Company	Processor	Date	Training time
PFN	TITAN X *128	17/1	4h
Facebook	P100 *256	17/6	1h
PFN	P100 *1024	17/11	15min
SONY	V100 *2176	18/11	3.7min
Google	TPUv3 *1024	18/11	2.2min





The image was excerpted from https://github.com/basicmi/AI-Chip

目次

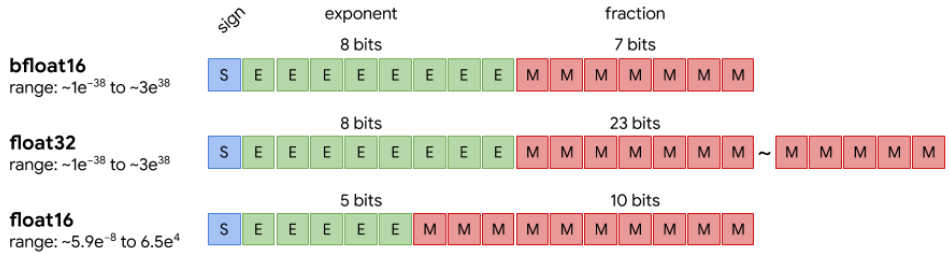
- PFNにとっての計算能力の位置付け
- 代表的なDeep Learningの高速化手法
- なぜ今プロセッサ開発なのか？
- MN-Coreの概要と開発チームの働き方
- 最近の成果

なぜ今DL専用アクセラレータ開発なのか？

- 「大きな計算能力が必要」という需要の観点以外にも、ハードウェア開発を加速させる背景がある
  - Deep Learningの応用範囲が拡大していること
  - 分散学習では高速化が難しいワークロードが一部存在すること
  - Deep Learningにおいて必要とされる演算精度がこれまでの科学技術計算と異なること
  - 演算手順が計算グラフによって宣言的に定義されること

演算精度について(Training)

- Trainingでは混合精度 (fp16乗算+fp32加算)の活用が注目されている
  - NVIDIA Volta: Tensor Core (4x4の混合積和演算, fp16 matmul, fp32 accumulate)
  - Google TPU: BFLOAT16 (brain float)
    - fp16よりもdynamic rangeが広い
    - 勾配のunderflow対策



- Deep learningに最適な数値表現とはなにか？という問題には答えは出ていない
  - [fp16でもCNN, RNN, GANなどのTrainingがある程度うまくいくという報告]
  - 使う観点ではcuDNNなどが対応を始めているが、正しく利用するためにはノウハウが必要

The figure was excerpted from https://cloud.google.com/tpu/docs/bfloat16

演算精度について(Inference)

- Inferenceでは、よりAggressiveな最適化が可能
  - Int8, Int4, binary
- 学習済みモデルをターゲットアーキテクチャに対して最適化する
  - モデルの量子化 (Quantization)
    - モデルのN-bit整数化
  - モデルの剪定 (Pruning)
    - 構築済みモデルのSparse化 (主にSpMV Acceleratorとの組み合わせ)
  - 小さいモデルへの蒸留 (distillation)
    - 小さなモデルに教師モデルの分布を学習させる
- Emerging deviceを利用したものも様々提案がある(が、現時点ではまだMNISTなどのToy Problemが解ける程度という印象)

なぜ今DL専用アクセラレータ開発なのか？

- 「大きな計算能力が必要」という需要の観点以外にも、ハードウェア開発を加速させる背景がある
  - Deep Learningの応用範囲が拡大していること
  - 分散学習では高速化が難しいワークロードが一部存在すること
  - Deep Learningにおいて必要とされる演算精度がこれまでの科学技術計算と異なること
  - 演算手順が計算グラフによって宣言的に定義されること

なぜ今**PFN**がDL専用アクセラレータ開発なのか？

- 「大きな計算能力が必要」という需要の観点以外にも、ハードウェア開発を加速させる背景がある
  - Deep Learningの応用範囲が拡大していること
  - 分散学習では高速化が難しいワークロードが一部存在すること

    -> 実際に社内にワークロードが存在し、幅広いタスクの高速化が図れる
  - Deep Learning において必要とされる演算精度がこれまでの科学技術計算と異なること

    -> 専用アクセラレータによって同じ計算をより低消費電力/低コストに実行可能
  - 演算手順が計算グラフによって宣言的に定義されること

    -> シンプルなHW + それを活かす高度なコンパイラという構成によるアーキテクチャレベルの性能向上の余地がある

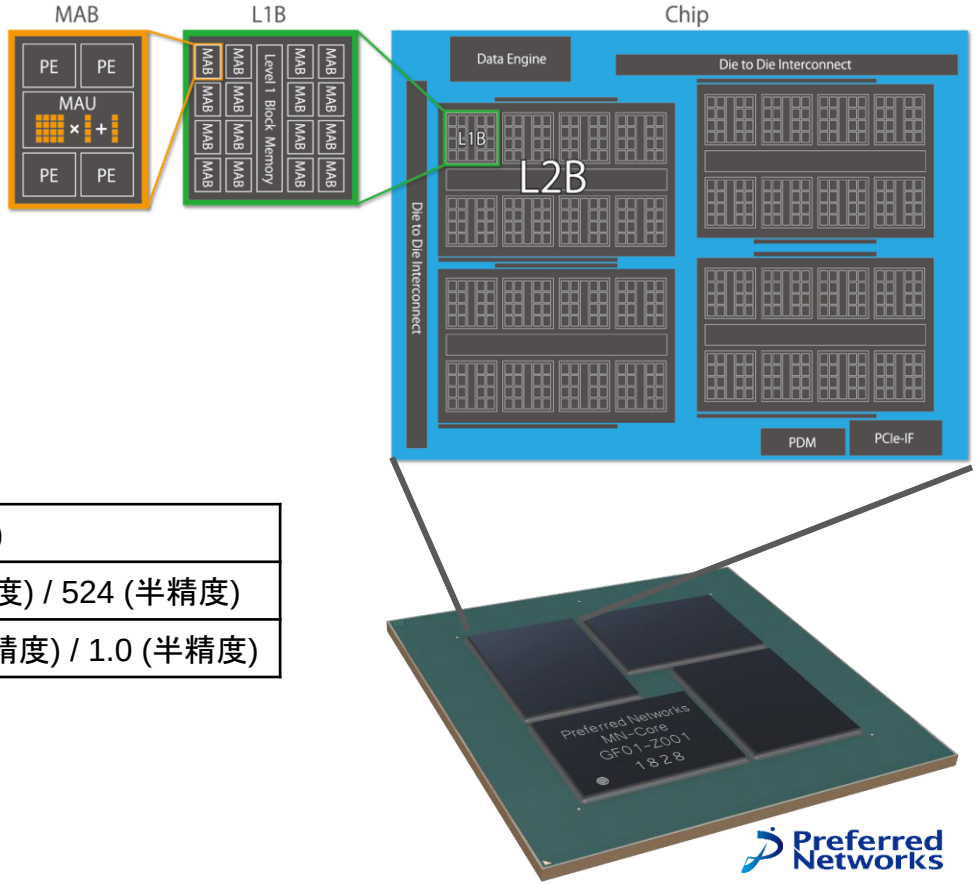
<u>ソフトウェアに強い企業だからこそハードウェアを作っている</u>

目次

- PFNにとっての計算能力の位置付け
- 代表的なDeep Learningの高速化手法
- なぜ今プロセッサ開発なのか？
- MN-Coreの概要と開発チームの働き方
- 最近の成果

MN-Coreの概要

- 超大規模なSIMD型プロセッサ
  - Single Instruction, Multiple Data (単一命令・複数データ)
  - 各階層のメモリは陽にプログラムが制御(scratchpad)
- 各階層間では分配、結合、放送、縮約といった複数の転送モードをサポート



消費電力 (W、設計値)	500 (4die, 1package合計)
ピーク性能 (TFLOPS、設計値)	32.8 (倍精度) / 131 (単精度) / 524 (半精度)
電力性能 (TFLOPS / W、設計値)	0.066 (倍精度) / 0.26 (単精度) / 1.0 (半精度)

MN-Coreを搭載したクラスタ(MN-3)

- 1rack辺り 4MN-Core Server
- 1MN-Core Server辺り4MN-Core Board
  - サーバ間はRoCEv2で接続
- 1MN-Core Board辺り1MN-Core Chip
  - MN-Core Board間はMNCore DirectConnectで相互接続

(表) MN-Core Board

チップ	1 MN-Core チップ
インターフェース	PCI Express Gen3 x16
メモリサイズ	32 GB
消費電力	600 W (予測値)

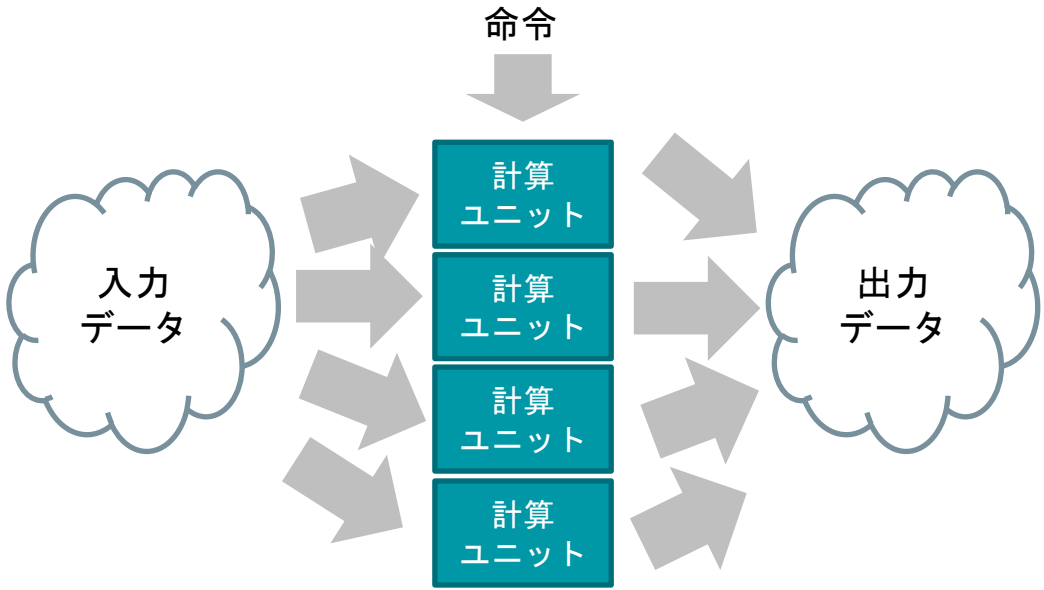
(表) MN-Core Server

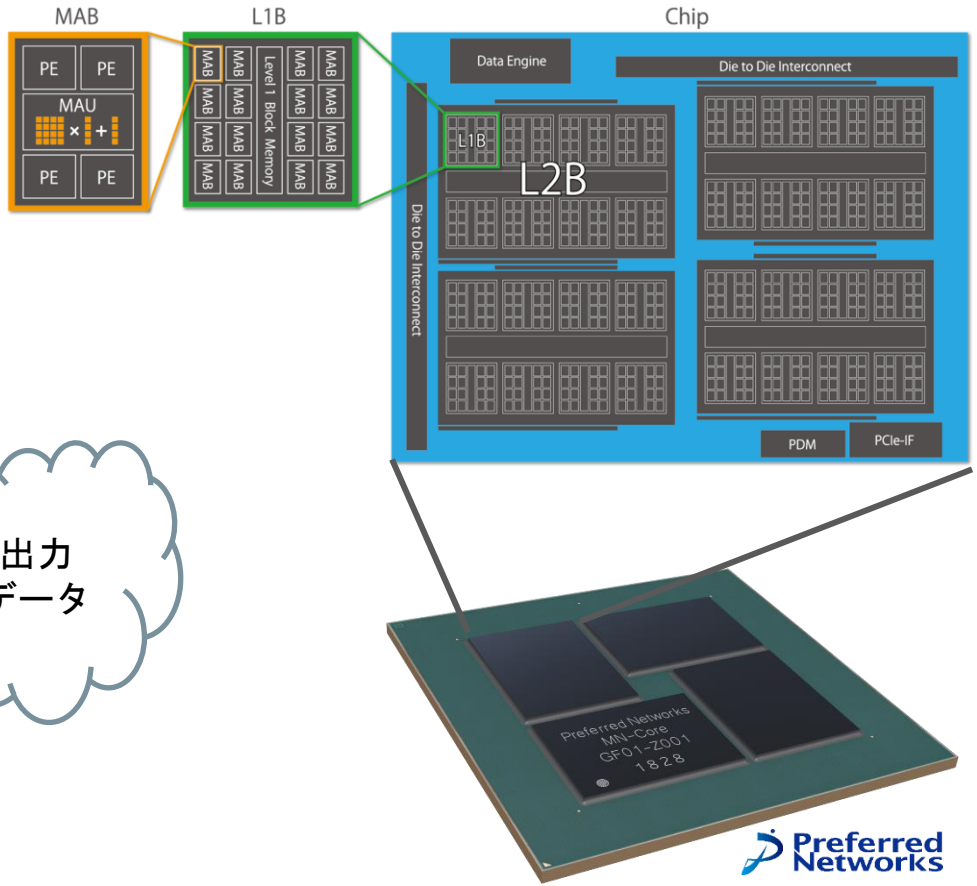
MN-Core 搭載数	4 MN-Core Boards
CPU	Dual socket up to TDP 200W
メモリ	DDR4 up to 2666MHz / Up to 3TB ECC 3DS LRDIMM, 1TB ECC RDIMM
ストレージ	Up to 24 SAS/SATA drive bays / 8x 2.5" SAS/SATA supported natively, 2x 2.5" NVMe supported natively
電源ユニット	4 2000W (2+2 Redundant) Titanium Level
サイズ	H311mm, W437mm, D737mm (7U Rack-mountable)

Preferred Networks

SIMDプロセッサのプログラミング (超概略版)

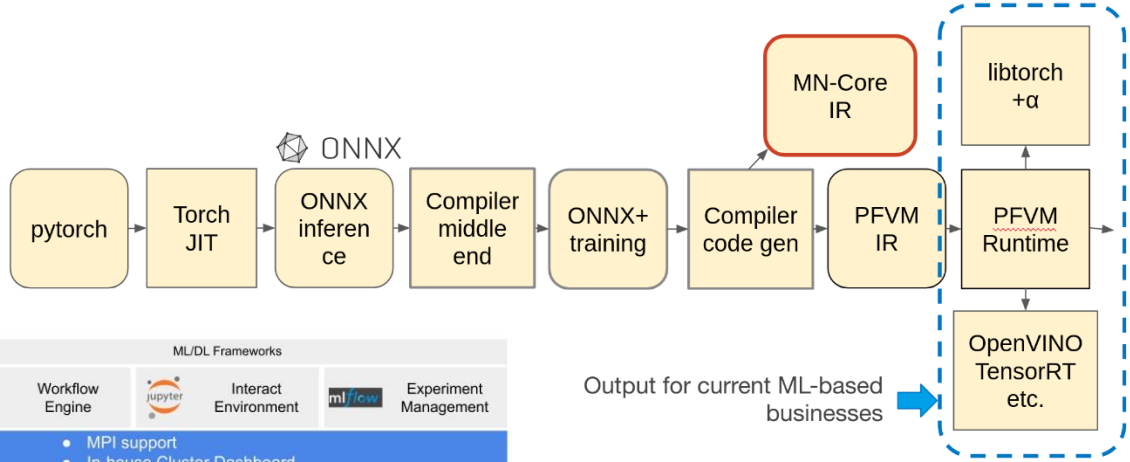
- 計算ユニットが休むことのないようにデータを流す
- 可能な限りデータの移動を減らすような演算スケジュールを組む





MN-Coreのためのソフトウェア(開発中)

- MN-Coreを動作させるための様々なruntime software
- 計算グラフから最適なMN-Core向けプログラムを自動生成するためのコンパイラ
- ユーザコードの改変を最小限に抑えるためのユーザ向けI/F
- クラスタの一部として動作させるためのk8sやMiddlewareとの統合

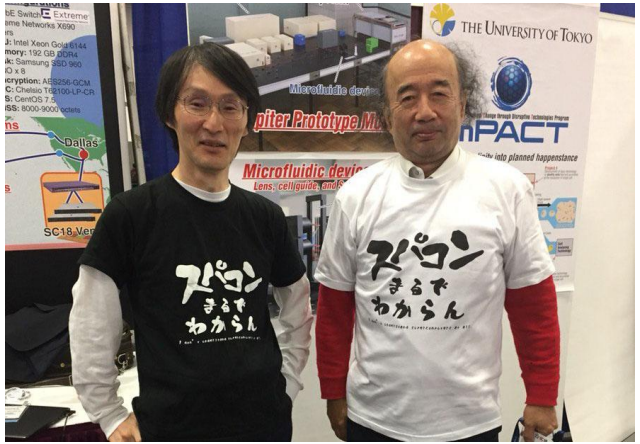


Name	CPU [Core]	Memory [GiB]	GPU	MN-Core	Links	Notes
2	90% (42.5/47.0)	75% (280.67/374.07)	-	0% (0/4)	Issue bmcURL	(alpha)
3	99% (46.5/47.0)	79% (296.67/374.07)	-	100% (4/4)	Issue bmcURL	(alpha)
4	97% (45.5/47.0)	95% (354.67/374.07)	-	25% (1/4)	Issue bmcURL	(alpha)
5	101% (47.5/47.0)	99% (370.67/374.07)	-	100% (4/4)	Issue bmcURL	(alpha)

MN-Core開発チーム

- ハードウェア/ソフトウェアエンジニアの垣根なくフレキシブルに働いている
  - そもそも肩書として区別が会社全体として存在していない
- HDLをソフトウェアエンジニアが読みつつ・・・などはよくある

神戸大学 牧野淳一郎教授と共同で
アーキテクチャ検討を実施



開発メンバー(一部)

MN-Coreの開発

- ハードウェア開発
  - アーキテクチャ検討
  - 詳細仕様検討
  - 論理設計
  - 物理設計
  - 検証

プロセッサ tapeoutまで

- ソフトウェア開発
  - 機能/精度検証用ソフトウェア実装
  - 記述性確認用アプリケーションの作成
  - ランタイム/コンパイラの先行開発

- システムとしての実装
  - ボード/サーバの設計
  - 工事スケジュール
  - システムの初期検証

(一例): 私が実作業していたところ

プロセッサES入手後

- ソフトウェア開発
  - コンパイラの開発
  - GREEN500 Benchmark開発
  - モデル検討/社内での実利用

- システムとしての実装
  - 社内k8sとの統合
  - Monitoring

- 様々な仕事があるが、ハードウェア/ソフトウェアエンジニアで明確な区切りがあるわけではない
  - 時期によってそれぞれ自身の得意分野を活かしながら業務が移り変わっている
    - ex. 既存のプロセッサ上で高速なカーネルを書くのが得意な人は、チップができるまでは自分が使うチップの仕様検討をしていたりする

MN-Coreチームの開発スタイル

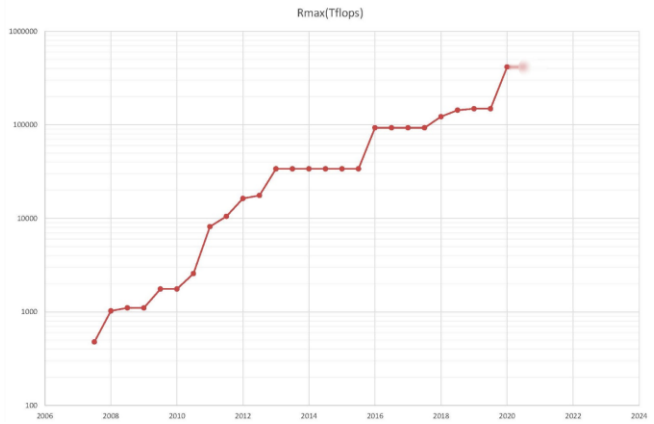
- HW/SWも一つのリポジトリで開発を行っている(monorepo)
  - 社内の開発リポジトリのActivity: 5000+PRs, 20KCommitsぐらい
    - 最近ではコンパイラ周辺の開発が加速しているのでもっと増えているはず
- CIを大切にする文化がある
  - PRの影響範囲ごとに自動的にRTL Simが流れたり、実機を用いたコンパイラのtest-runが走る
  - Weeklyで検証の状況やコンパイラの性能などに関する自動CIが回る
  - かなり大規模なコードベースで、大人数で作業をしているためかなり大事
- 基本的に仕事はslack上, 週1程度で分野ごとに定例
  - コロナで完全にリモート勤務になりましたが少なくともSW系の人にはあまり影響なし
  - HW系の人も、実際の計測などを除くとリモートでもできることは多い
    - RTL Simのような待ち時間が長い作業の場合、リモートのほうがより都合が良かったという話も一部あったりする

目次

- PFNにとっての計算能力の位置付け
- 代表的なDeep Learningの高速化手法
- なぜ今プロセッサ開発なのか？
- MN-Coreの概要と開発チームの働き方
- 最近の成果

TOP500とGreen500

- TOP500は、世界中にあるスパコンのプログラム実行時の最高性能による順位リストである
    - 連立一次方程式をLU分解を用いて解くWorkload
- TOP500.orgに申告されたシステム性能をNo.1からNo.500までリストにしたもの
- 1993年に始まり、以後毎年6月と11月に更新されます
- Green500は、TOP500に申告されたシステム性能をNo.1からNo.500までリストにしたものの性能を、消費電力で除算して、電力あたり性能を求め、順位リストにしたもの



GREEN500 Challenge

- 2020/06: 21.11GFlops/W (#1 in the world, 1.621PFlops) at zone0+1
  - #2 is a NVIDIA DGX SuperPOD (21.108GF/W)
- 2020/11: 26.04GFlops/W (#2 in the world, 1.652PFlops) at zone0
  - #1 is a NVIDIA DGX SuperPOD (26.195GF/W)
- 2020/06 -> 2020/11で電力効率20%以上UP, 40->32ノードでRmax同等


The GREEN 500 CERTIFICATE

MN-3 - MN-Core Server, Xeon 8260M 24C 2.4GHz, MN-Core, RoCEv2/MN-Core DirectConnect

Preferred Networks, Japan

is ranked

No. 1 in the Green500

among the World's TOP500 Supercomputers

with 21.11 GFlops/Watt Linpack Power-Efficiency

on the Green500 List published at ISC 2020 Digital Conference, June 22nd, 2020

Congratulations from the Green500 Editors

Wu-chun Feng
Virginia Tech

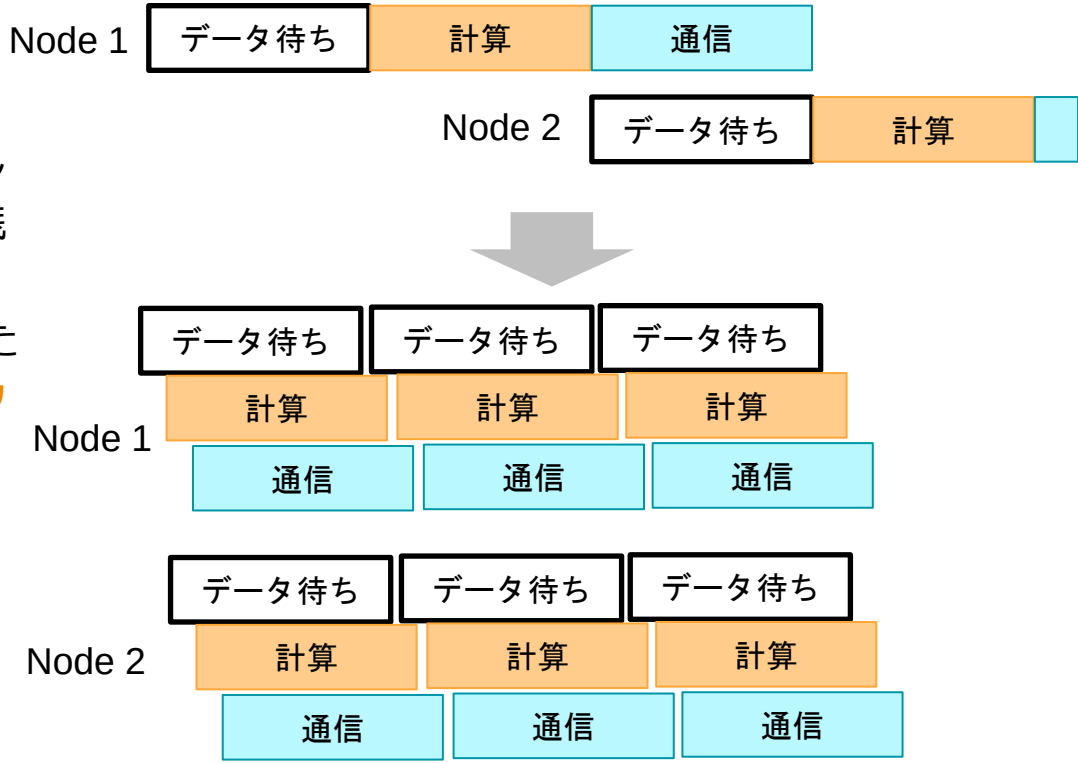
Kirk Cameron
Virginia Tech

開発のポイント

- 前回 (2020年6月のランキング) では、まずは動作させることを優先し、性能的な最適化余地がたくさん残っていた
- 今回は、MN-Coreの特性を吟味した上で設定を最適化した他、ソフトウェアの最適化を行った
  - 内部のルーチンの分割の最適化
  - 実行スケジュール最適化

  等々

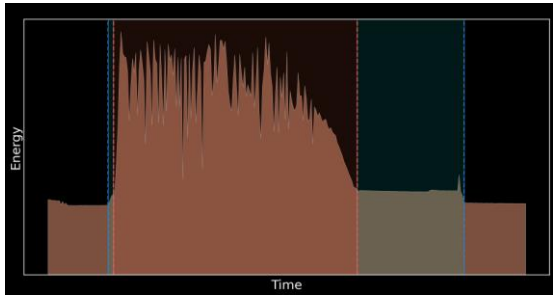
工夫の例 (イメージ図)



Top/Green500 Challengeで得られたもの

- システム全体(アクセラレータ/ホスト/インタコネクト)が安定かつ高性能に、一定の期間動き続ける必要があり、システム自体のBringupとして非常に有効だった
- HPLは基本的な計算カーネルが行列積、かつ低レイテンシな通信が性能を出すために必須である。これは分散深層学習のWorkloadが要求する条件でもあり、今回得られた最適化のための知見が実際にDeep Learning向けコンパイラのコード生成にも応用可能
- システム全体を正確(level3)に計測する試みを通じてクラスタの系全体への理解が深まった
- やり切ればきちんと世界一が獲れるシステムが作れるという経験

	Level 1	Level 2	Level 3
電力計測器の精度	最低定格精度5%以上	最低定格精度2%以上	Revernue Grade(ANSI C12.20) or, SPEC power認定 or, 最低定格精度 1%以上
電力サンプリング周波数	1 Hz以上	1 Hz以上	AC: 5KHz以上 DC: 120Hz以上
電力計測項目	瞬時電力値	瞬時電力値	総エネルギー、電圧、電流
必要な計測範囲	計算実行期間のみ	計算実行時間を含む全処理時間	計算実行時間を含む全処理時間
電力計測期間の例外規定	厳密な時間管理を必要としない	厳密な時間管理を必要としない	厳密な時間管理が必要
計算ノードの計測	一部分の計測による類推が許容	一部分の計測による類推が許容	全ノード計測が必須
その他システムの計測 (インターコネクトやストレージシステム等)	1/10以上の計測による類推あるいはカタログ定格による類推	1/8以上の計測による類推あるいはカタログ定格による類推	全ノードの計測が必要
設置環境情報の計測	オプション	オプション	オプション



Level3計測部門では
No1と認定@2020/11

さいごに

- 計算機インフラはDeep Learningをコアとする研究開発の競争力の源泉の一つ
  - Deep Learningが多量の計算能力を必要とするだけでなく、Deep Learningというワークロード自体の特性を生かした新しいアーキテクチャが様々開発されている
- PFNでは計算機インフラの研究開発にも多くの投資を行っている
  - 複数のコンピュータを並べてScale-out
  - 高速なアクセラレータ(MN-Core)を開発し単一ノードの性能をScale-up
- PFN独自設計のDeep Learning向けアクセラレータ「MN-Core」が実際に稼働しつつあり、電力性能比の世界ランキングで１位@2020/06を獲得した
- Deep Learning向けコンパイラの開発もかなり力を入れて進めている

計算機を１から作るのは楽しい

Q&A