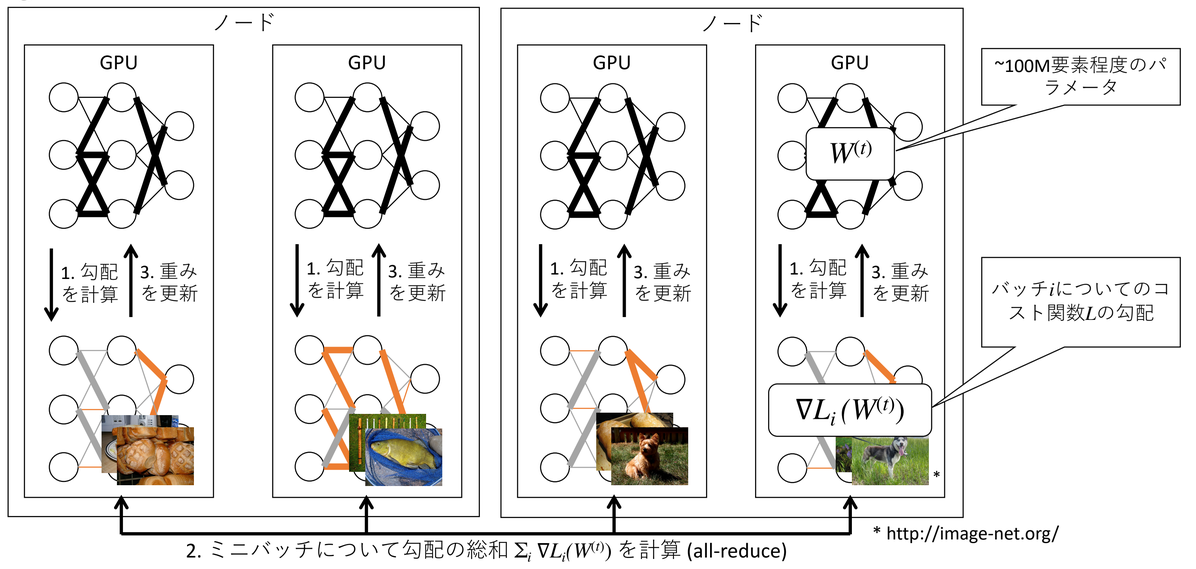
# 大規模並列環境における低精度型を用いたディープラーニングの学習精度の検証

大山洋介[1*], 野村哲弘[1], 佐藤育郎[2], 松岡聡[3 1]
[1]東京工業大学　[2]デンソーアイティーラボラトリ　[3]理研計算科学研究センター　*oyama.y.aa@m.titech.ac.jp

## 研究背景

### データ並列学習

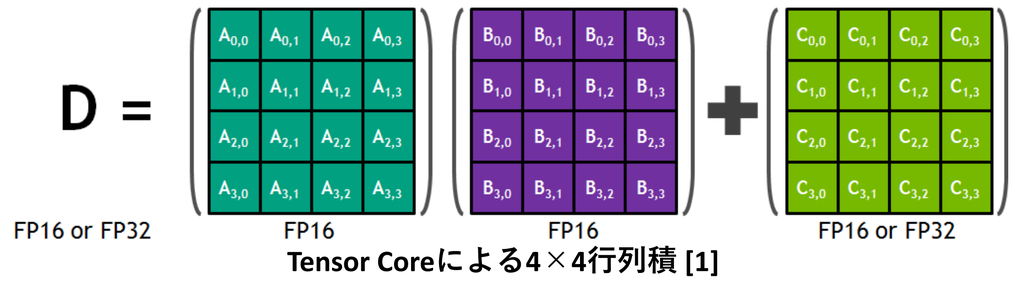
- 深層学習 (DL) ではSGD (確率的勾配降下法) による並列学習が一般的
  - 1反復で用いる複数のデータサンプル (ミニバッチ) についての計算を並列化する
  - GPU同士でコスト関数の勾配の総和計算 (all-reduce) を行う
- GPUの高速化にともないGPU間・ノード間通信の高速化が重要となる




**DNNのデータ並列学習**

### 低精度型の利用

- DLでは単精度浮動小数点数 (32 bit) よりも精度の低い数値計算が適用　可能であるといわれている
  - NVIDIA Volta GPUではTensor Core (混合精度演算ユニット) を搭載
  - 一部のDLフレームワークでは低精度な通信アルゴリズムを採用
    - 1 bit SGD (Microsoft CNTK), 16 bit all-reduce (ChainerMN)
- 使用される計算精度の決定は実験的な知見によるところが大きい



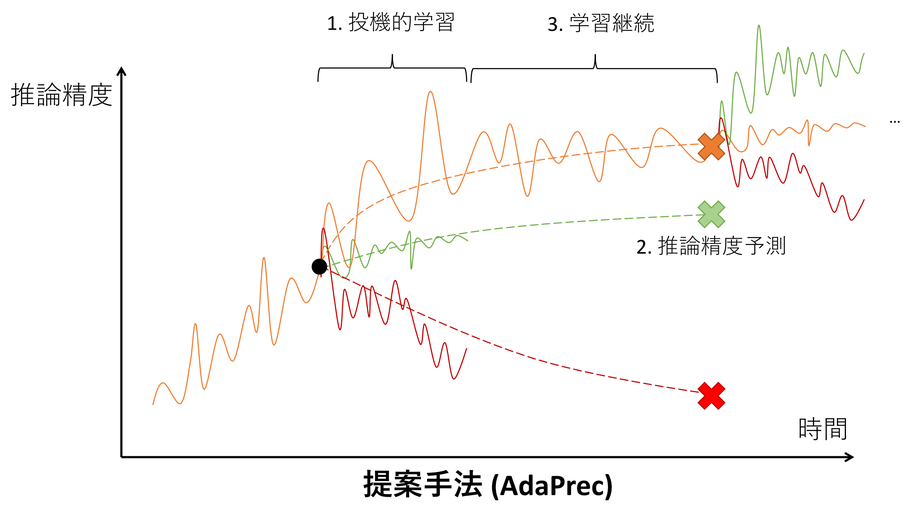

**Tensor Coreによる4×4行列積 [1]**

**問題提議**
- ディープラーニングにおいて最適な通信精度は？
- 自動で通信速度・精度を最適化するには？

## 提案手法

### AdaPrec

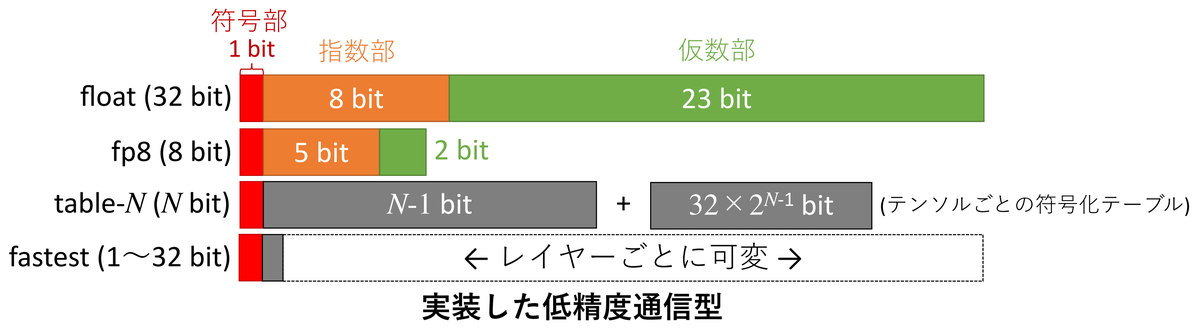
- 適合的 (ADAptive) に通信精度 (PRECision) を変更する手法
1. 複数の異なる通信精度で一定ステップを学習
2. 次回の1. 開始時の推論精度 (Top-N accuracy) を学習曲線モデル [2]により予測
3. 単位時間あたりの推論精度の向上が最大の通信精度を用いてさらに一定ステップ学習




**提案手法 (AdaPrec)**

### 低精度All-reduce実装

- 複数の異なるMPI 加算all-reduceアルゴリズムを実装
- **単精度 (float)**: 通常の`MPI_FLOAT`
- **8 bit (fp8)** [3]: 独自定義の8 bit浮動小数点型により通信値を符号化
- **出現頻度に基づく符号化 (table-$N$)**: $99k/2^{N-1}$ パーセンタイル ($k$=1,..$2^{N-1}$) を動的に計算し丸める
- **レイヤーごとの通信速度に基づく通信 (fastest)**
  - ベンチマークの結果，全結合層ではtable-1，畳み込み層ではfp8，バイアスについてはfloatを採用




**実装した低精度通信型**

## 評価

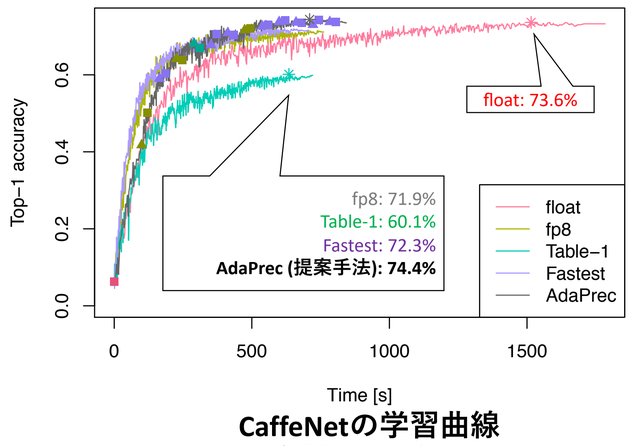
**評価環境 (Reedbush-H)**

| | |
|---|---|
| ノード数 | 120 |
| CPU | Intel Xeon E-2695v4 × 2 |
| - メモリ容量 | 256 GiB (DDR4) |
| GPU | NVIDIA Tesla P100 × 2 |
| - 演算性能 (単精度) | 10.6 TFlop/s |
| - 演算性能 (半精度) | 21.2 TFlop/s |
| - メモリ容量 | 16 GiB (HBM2) |
| - メモリ帯域幅 | 732 GiB/s |
| インターコネクト | 4xFDR Infiniband × 2 |
| - 帯域幅 | 14 GiB/s |
| DLフレームワーク | Caffe 1.0 (MPI対応版 [3]) |
| CUDA | 8.0 |
| MPI | OpenMPI 2.1.1 |

**学習設定**

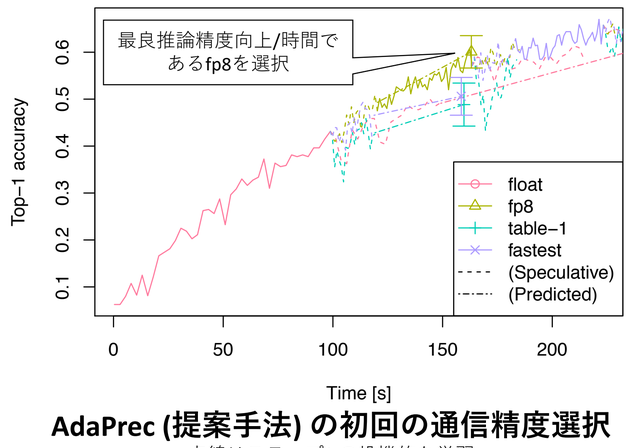
| | |
|---|---|
| モデル | CaffeNet |
| - パラメータ数 | 61.0 M |
| ミニバッチサイズ | 256 |
| データセット | ILSVRC2012データセット中の16クラス |
| Optimizer | Momentum SGD |
| Learning rate | 0.01 (1-(Epoch数)/100)$^2$ |
| Momentum | 0.9 |
| Epoch数 | 100 |

- CaffeNet (AlexNetに類似したCNN) を東京大学情報基盤センター Reedbush-Hの4 GPUで学習
  - 一定Epochの学習において、単一の低通信精度 (fp8, table-1) で学習する場合と同等の速度かつ単精度 (float) を上回る推論精度を達成
  - 通信速度に優れるが推論精度に悪影響を及ぼす通信手法 (table-1) を大部分のケースで除外することに成功




**CaffeNetの学習曲線**

＊は各通信手法の最良Accuracy
▲，■はそれぞれAdaPrec (提案手法) のステップ1.，3. 開始時点




**AdaPrec (提案手法) の初回の通信精度選択**

点線はステップ1.の投機的な学習
一点鎖線の先端は予測されたステップ3. 後の精度と±標準偏差

[1] L. Durant, O. Giroux, M. Harris and N. Stam, "Inside Volta: The World's Most Advanced Data Center GPU," May 2017, https://devblogs.nvidia.com/inside-volta/.
[2] A. Klein, S. Falkner, J. T. Springenberg, and F. Hutter, "Learning Curve Prediction with Bayesian Neural Networks," in International Conference on Learning Representations (ICLR) 2017 Conference Track, 2017.
[3] 大山洋介, 野村哲弘, 佐藤育郎, 松岡聡, "ディープラーニングのデータ並列学習における少精度浮動小数点数を用いた通信量の削減," 情報処理学会研究報告, Vol. 2017-HPC-158, 2017.