FUJITSU
shaping tomorrow with you

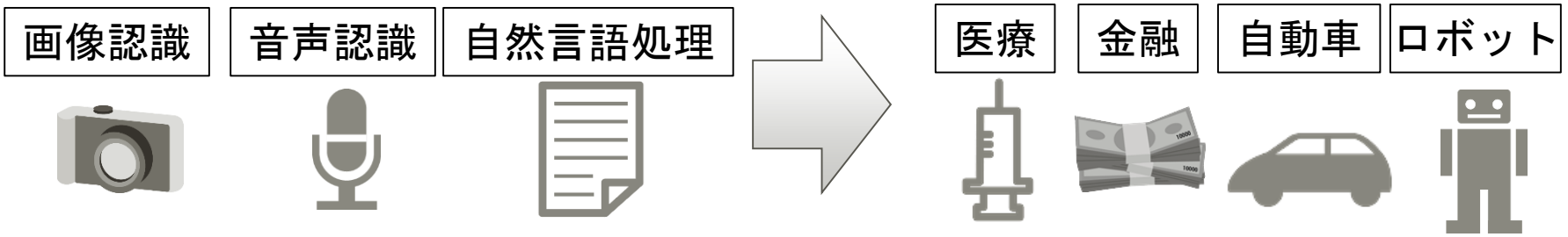
大規模ニューラルネットの高速Deep Learningの実現に向けて

株式会社富士通研究所
白幡　晃一
k.shirahata@jp.fujitsu.com

開発の背景

FUJITSU

■ Deep Learningは様々なアプリケーションへ適用拡大



■ ニューラルネットの大規模化

ニューラルネットの層数は増えるほど認識等の精度向上のため、大規模化が進む

年	ネット名	層数	エラー率
2012	AlexNet	8	16.4%
2014	VGGNet	19	7.3%
-	人間	-	5.1%
2015	ResNet	152	3.6%
2016	Ensemble 2	～152	3.0%

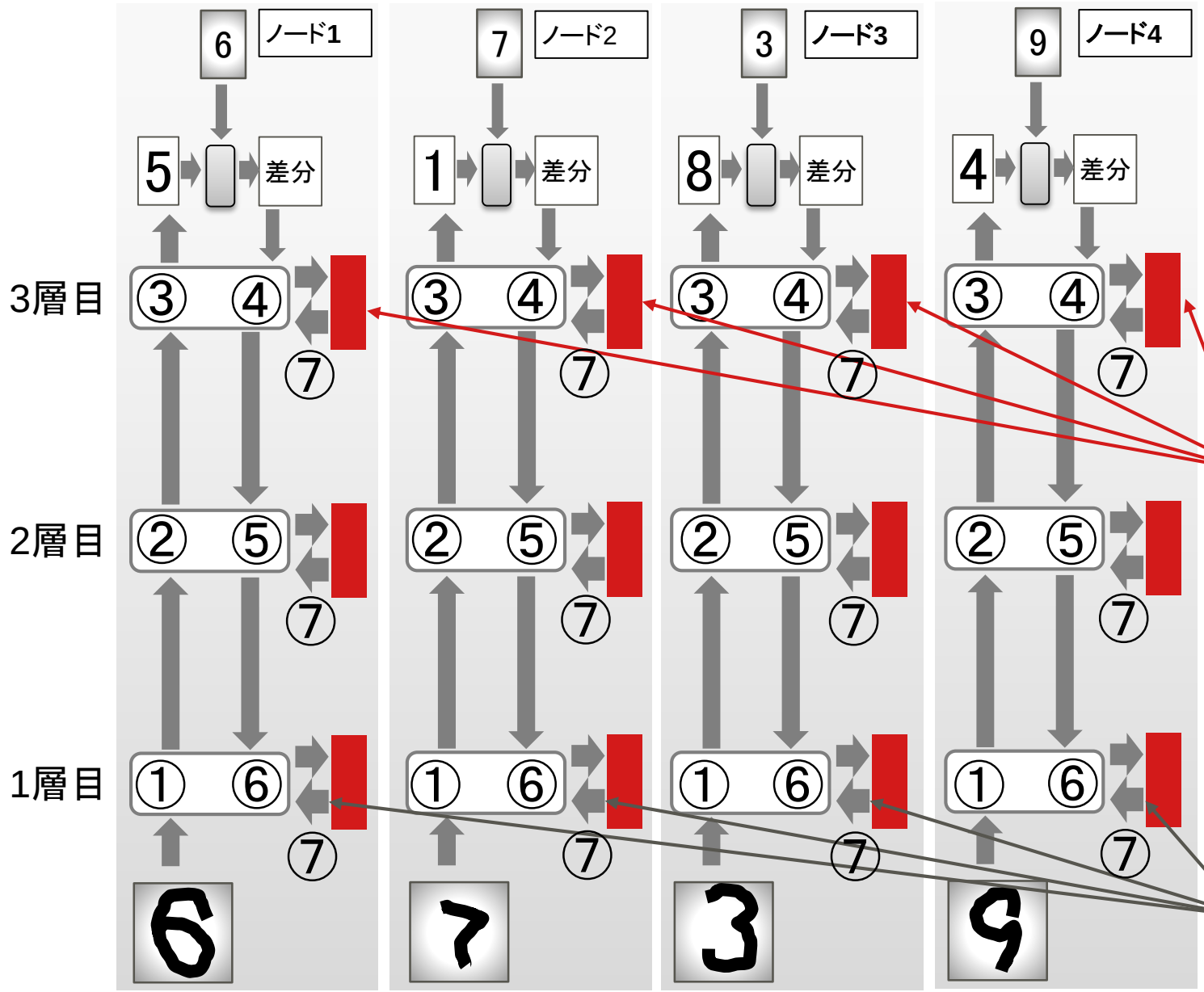
学習時間が増大 → 複数GPUの並列動作による高速化が必要

メモリ使用量が増大 → GPUメモリ使用の効率化が必要

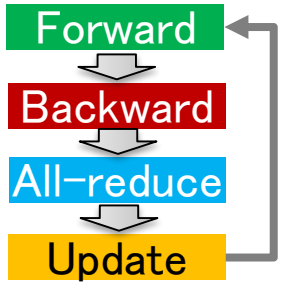
FUJITSU

Deep Learning処理の大規模並列化による高速化技術

複数のコンピュータで並列化(データ並列)



4ノードで
並列学習した例



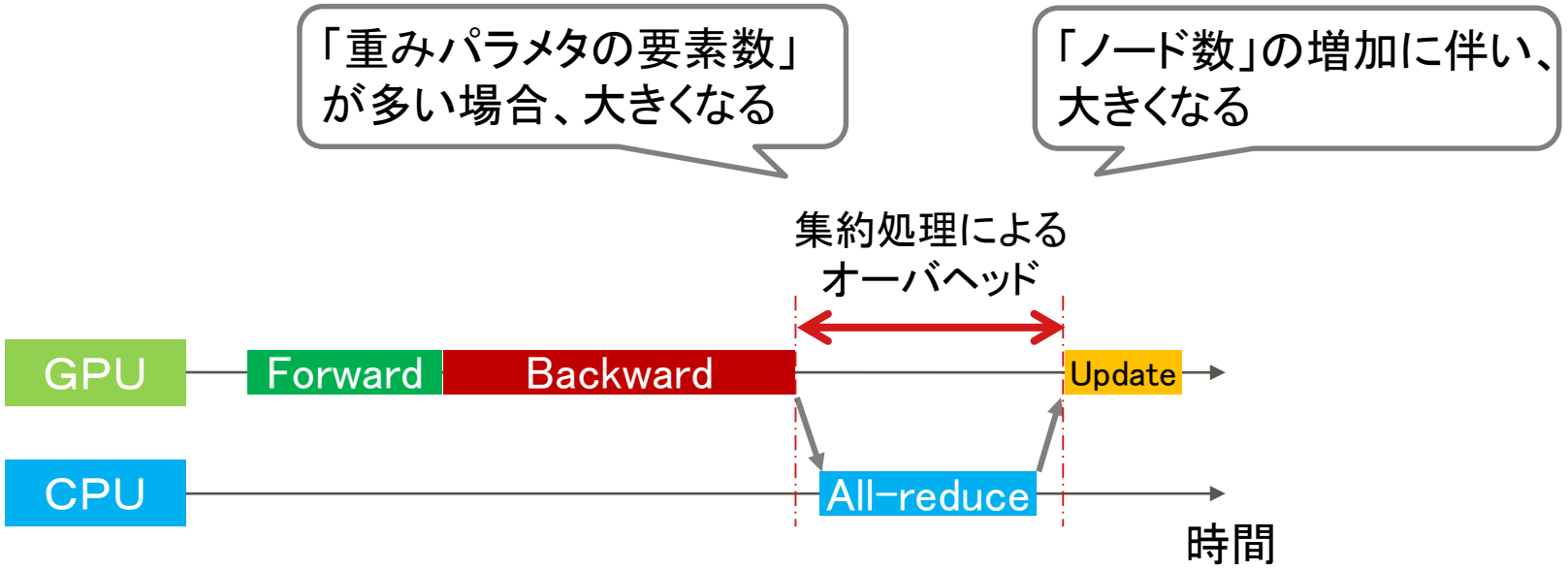
MPIの
All-reduce
処理で集約し、
更新用データ
を計算

⑦を終えたら、
次の①を開始

複数ノードで行う学習処理の課題

FUJITSU

- All-reduce処理が加わる事でGPUが動作しない時間が発生



基本的なアイディア

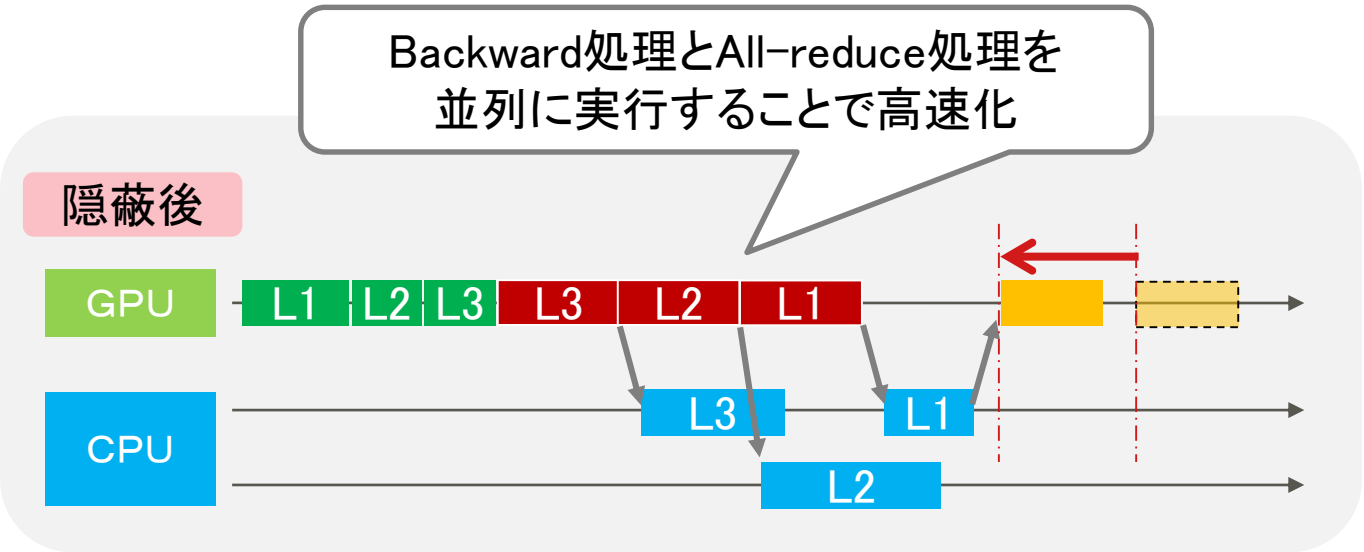
- 集約処理時間を他のGPU処理時間に隠蔽
- 集約処理時間を短縮

Backward処理時間への隠蔽

方法

各層のBackward処理が終わるごとに
層単位でAll-reduce処理を開始する

各層のForward処理
各層のBackward処理
All-reduce処理
Update処理



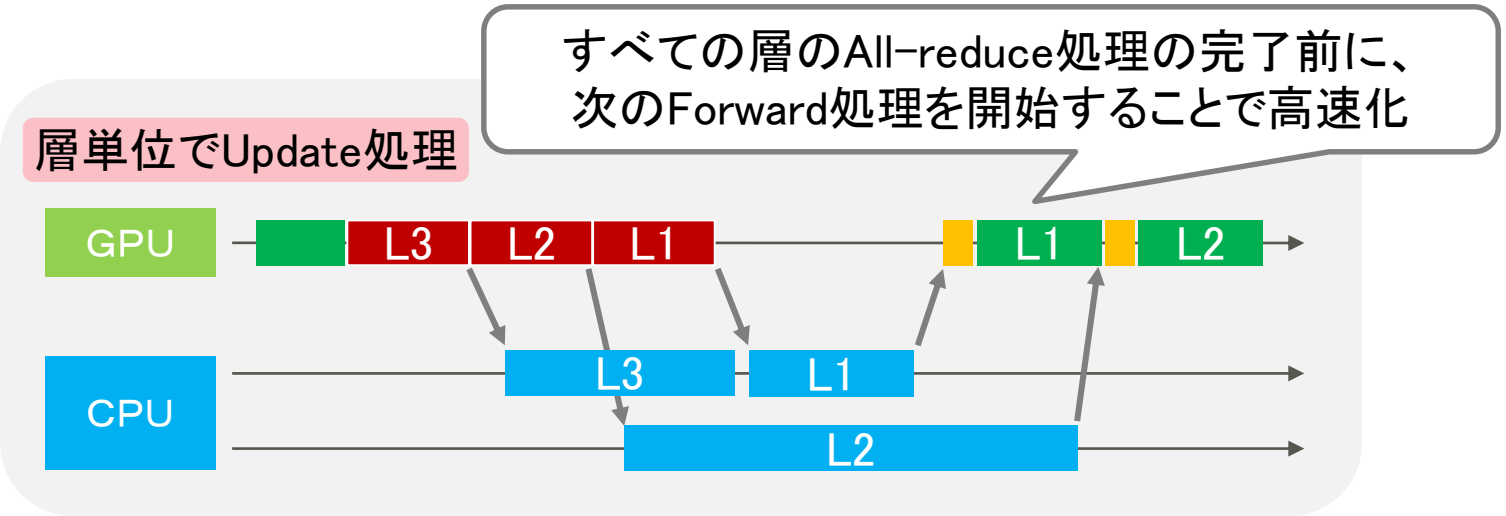
Forward処理時間への隠蔽

FUJITSU

方法

・Update処理を**分割**
・Forward処理の開始を**層ごと**に判定

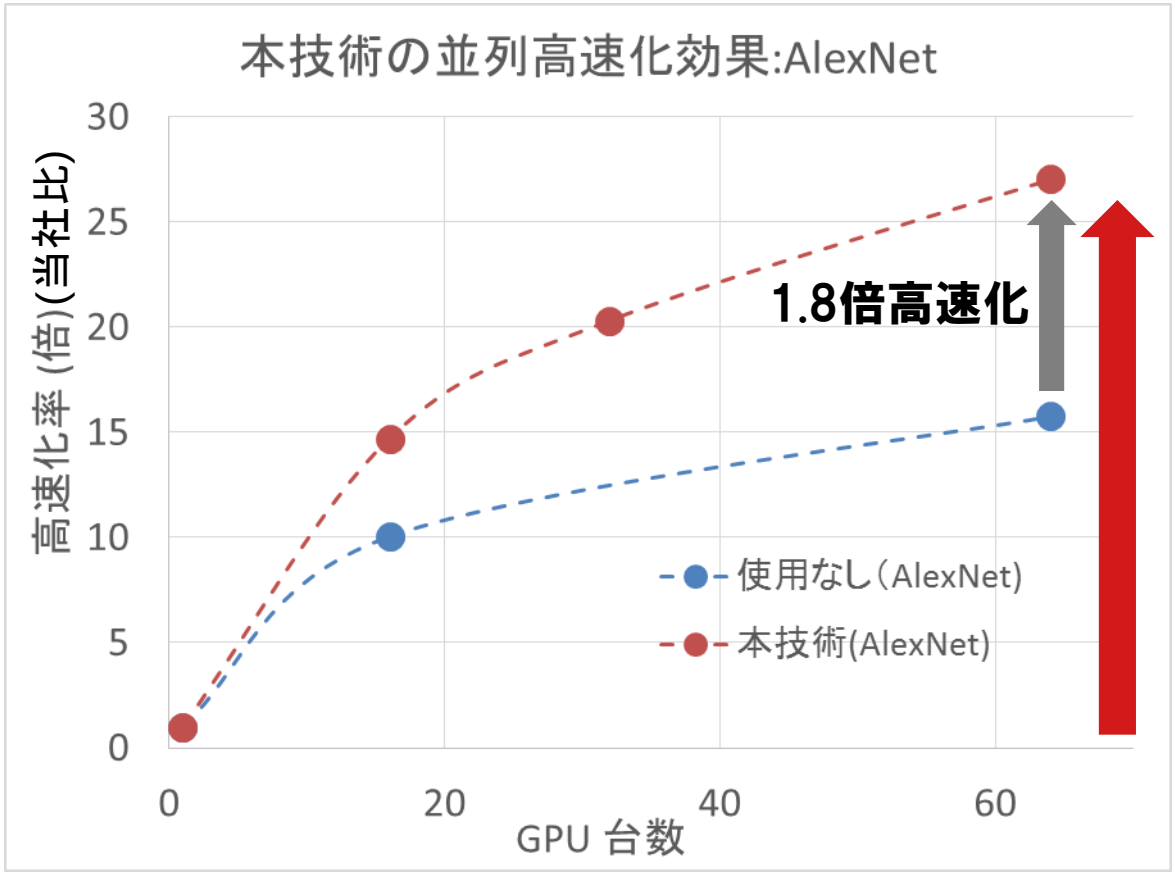
各層のForward処理
各層のBackward処理
All-reduce処理
Update処理



評価: Distributed Caffe

FUJITSU

複数 GPU 使用時の 1 GPU 使用時に対する学習時間の高速化率



評価環境

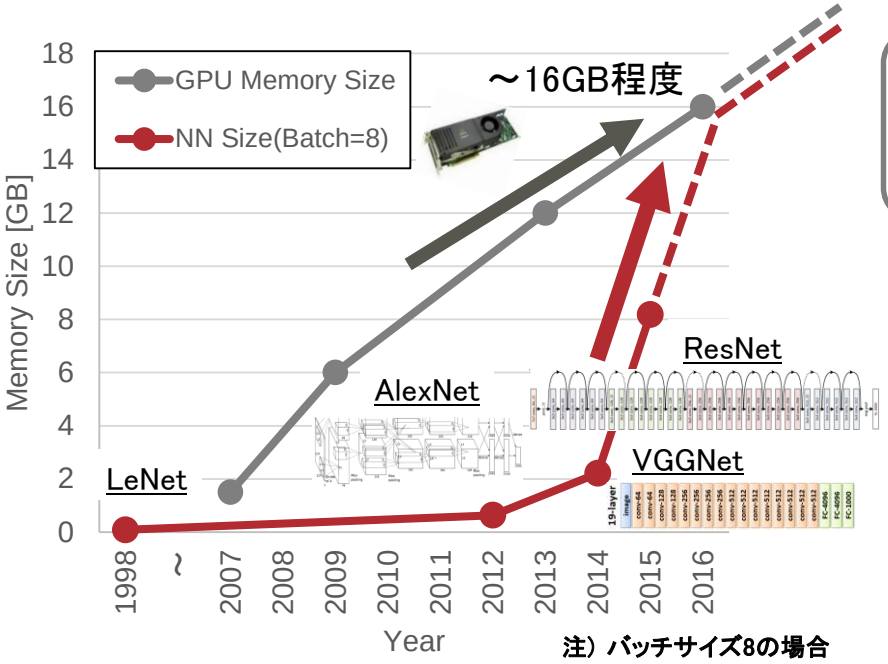
・Tesla K20X GPU を 64 GPU (1ノードあたり1GPU) まで使用

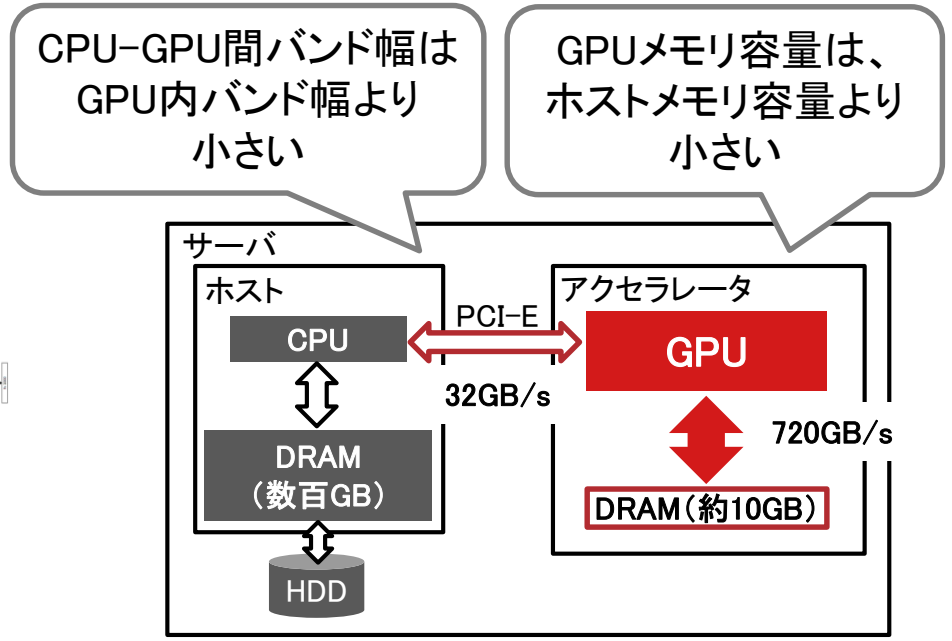
FUJITSU

Deep Learningのニューラルネットを拡大するメモリ効率化技術

GPUを用いたDeep Learningの現状

- ニューラルネットの規模は年々急速に拡大
- GPUの高速な演算性能を活用するには、一連の演算に使用するデータを可能な限りGPUの内部メモリに格納する必要がある



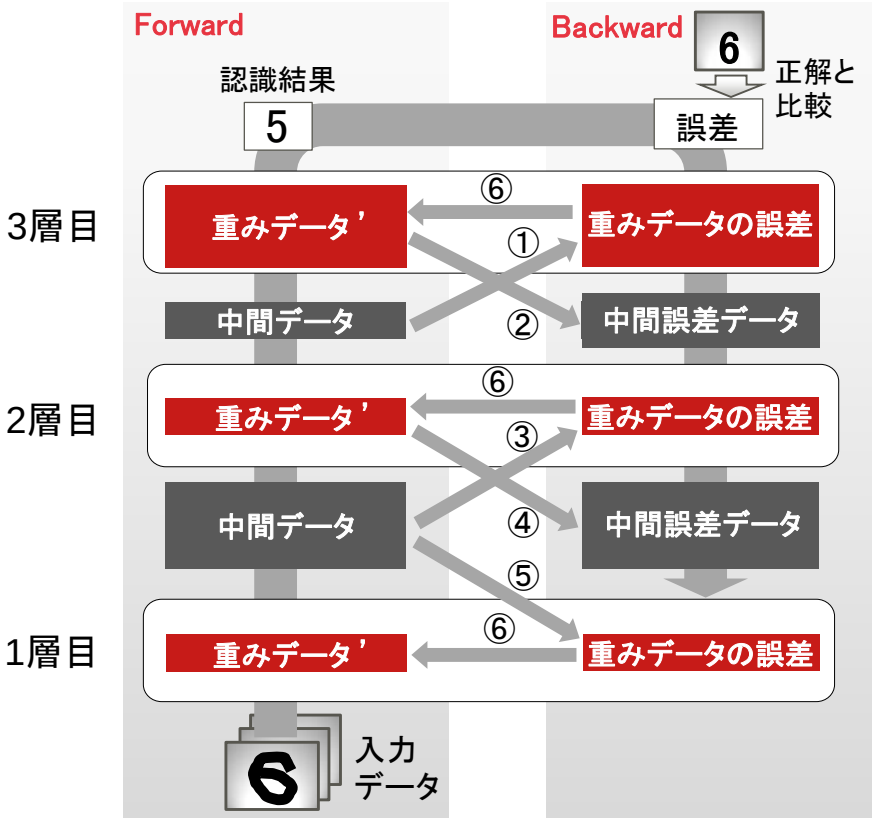


GPUが搭載しているメモリ量が小さいため、高速に学習できるニューラルネットの規模が制限されるという課題がある

Deep Learningのメモリ使用量の問題

FUJITSU

■ 学習時にGPUのメモリ使用量が増加する



①と②、③と④はそれぞれ独立に演算可能

重みを更新するための重みデータの誤差と、誤差を前層に伝えるための中間誤差データを計算

重みデータ、中間データに加えて誤差データを確保

基本的なアイディア

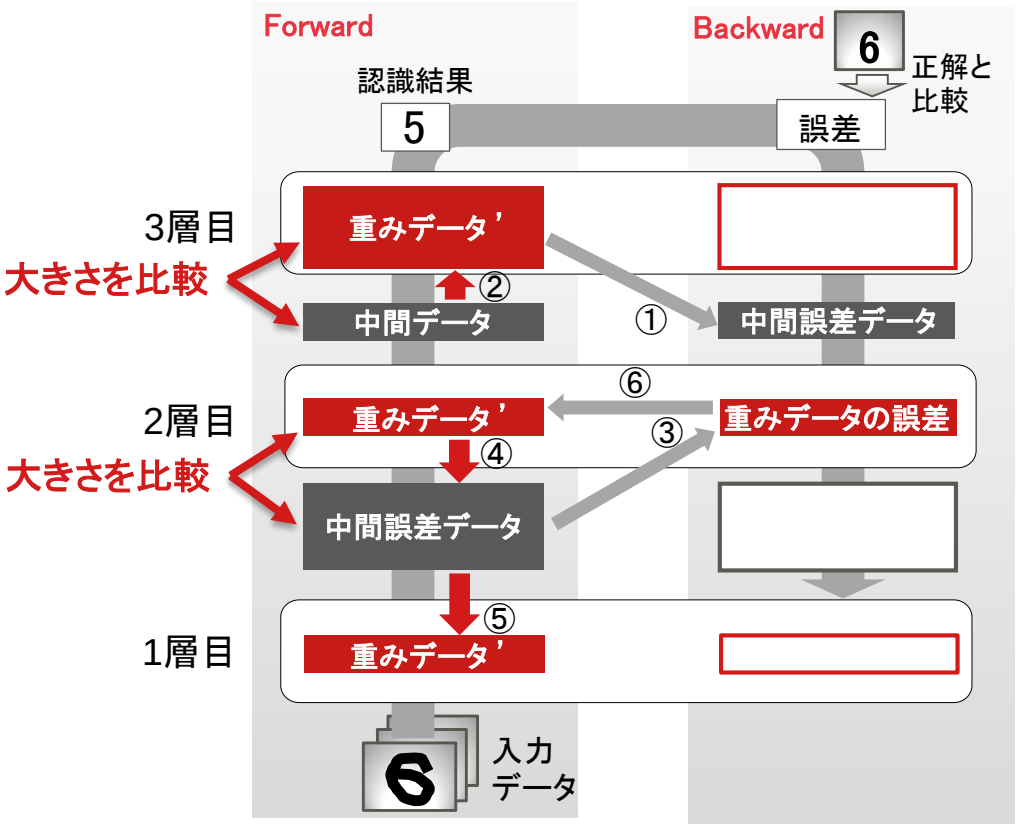
- **2種類の誤差データ計算の独立性に着目**
- **メモリ領域の再利用によりメモリ使用量を削減**

Deep Learningのメモリ効率化技術

FUJITSU

方法

ニューラルネットの構造を解析し、より大きなメモリ領域を再利用するように演算順序とメモリ配置をスケジューリング

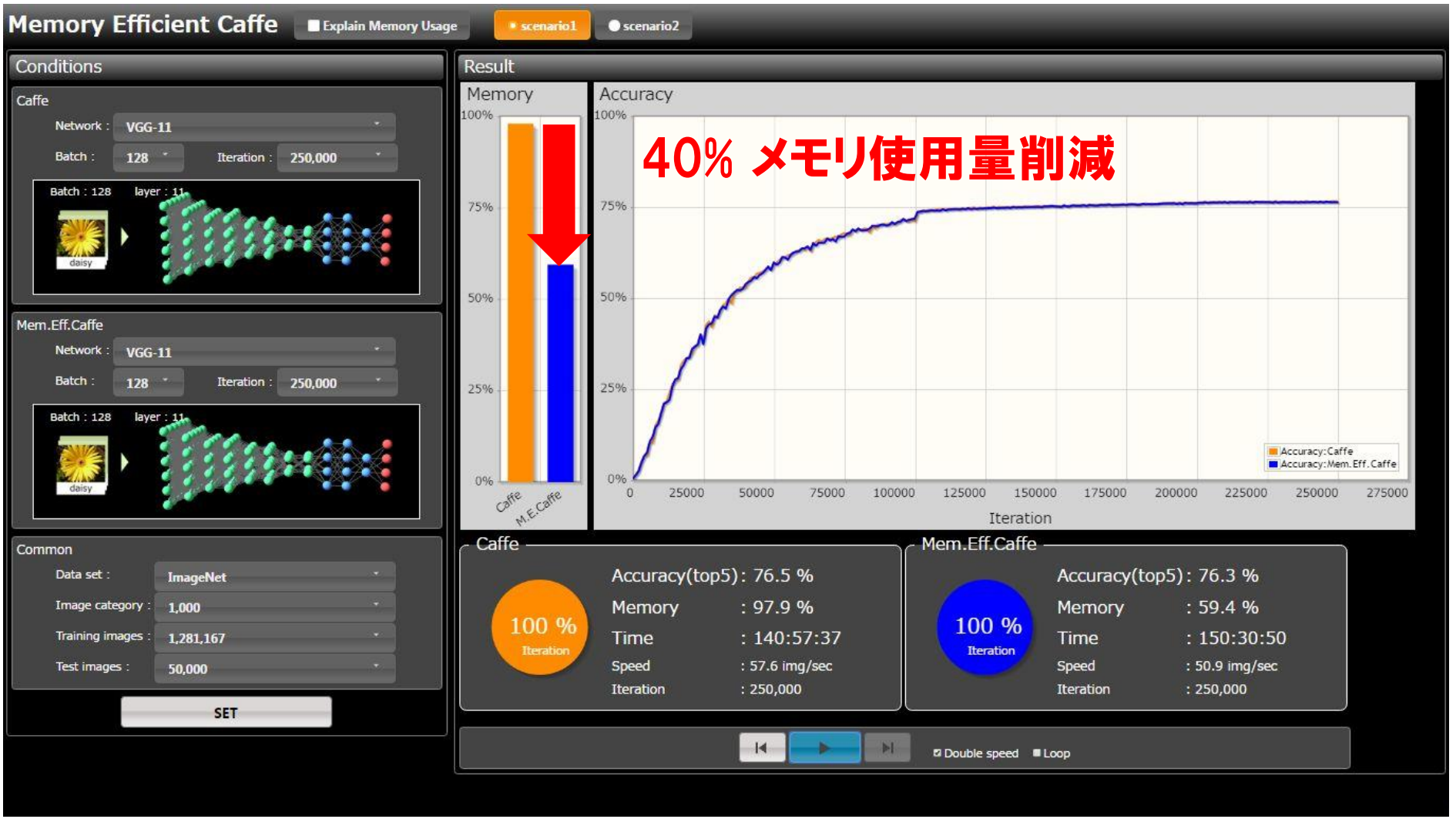


重みデータ＞中間データの層(Fully-connected層など)では、重みデータ領域を上書きしてメモリ使用量を削減

中間データ＞重みデータの層(Convolution層など)では、中間データ領域を上書きしてメモリ使用量を削減

評価: メモリ効率化Caffe

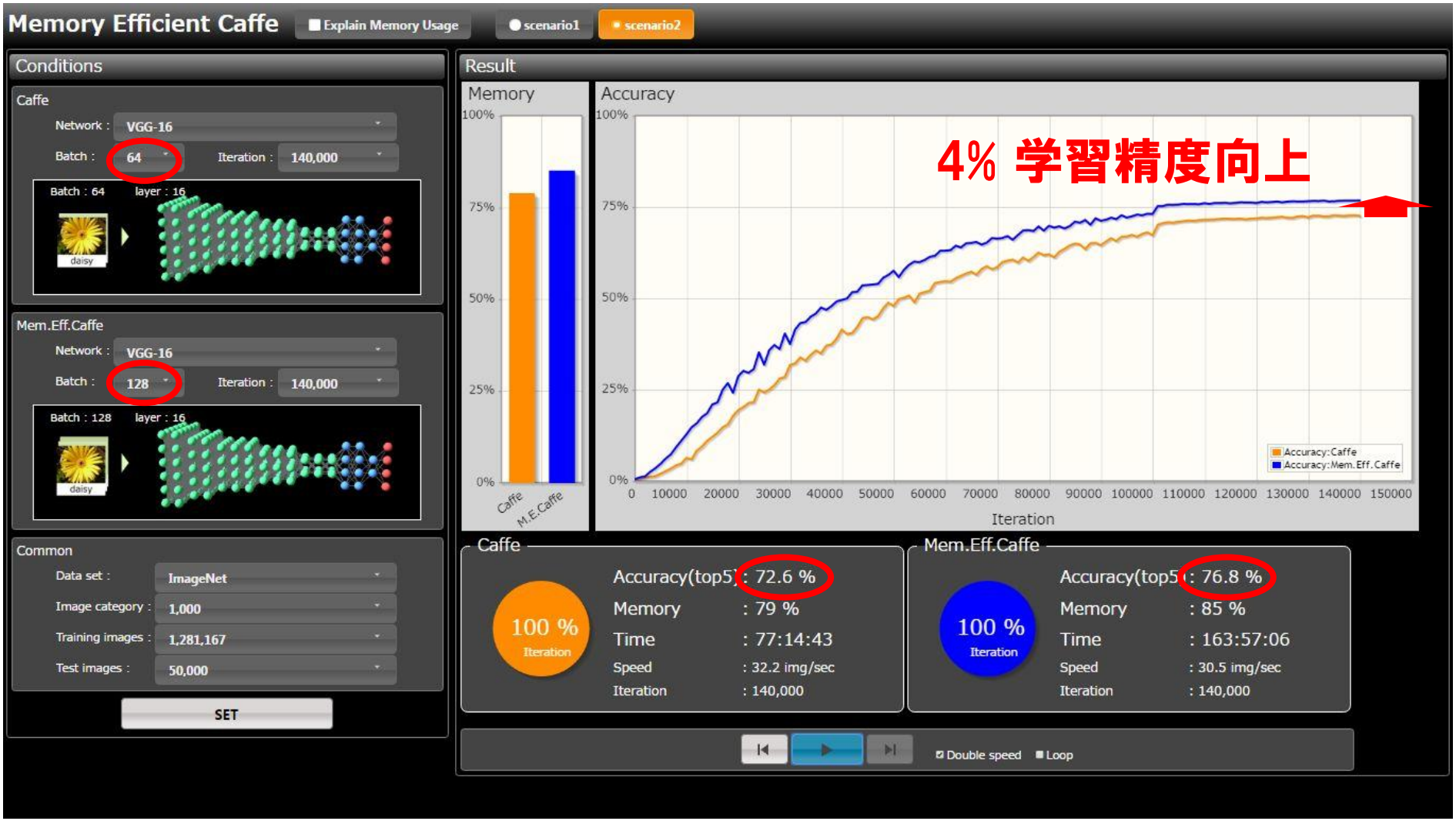
- 同じニューラルネット(VGGNet)を使用し、同条件で比較



評価: メモリ効率化Caffe

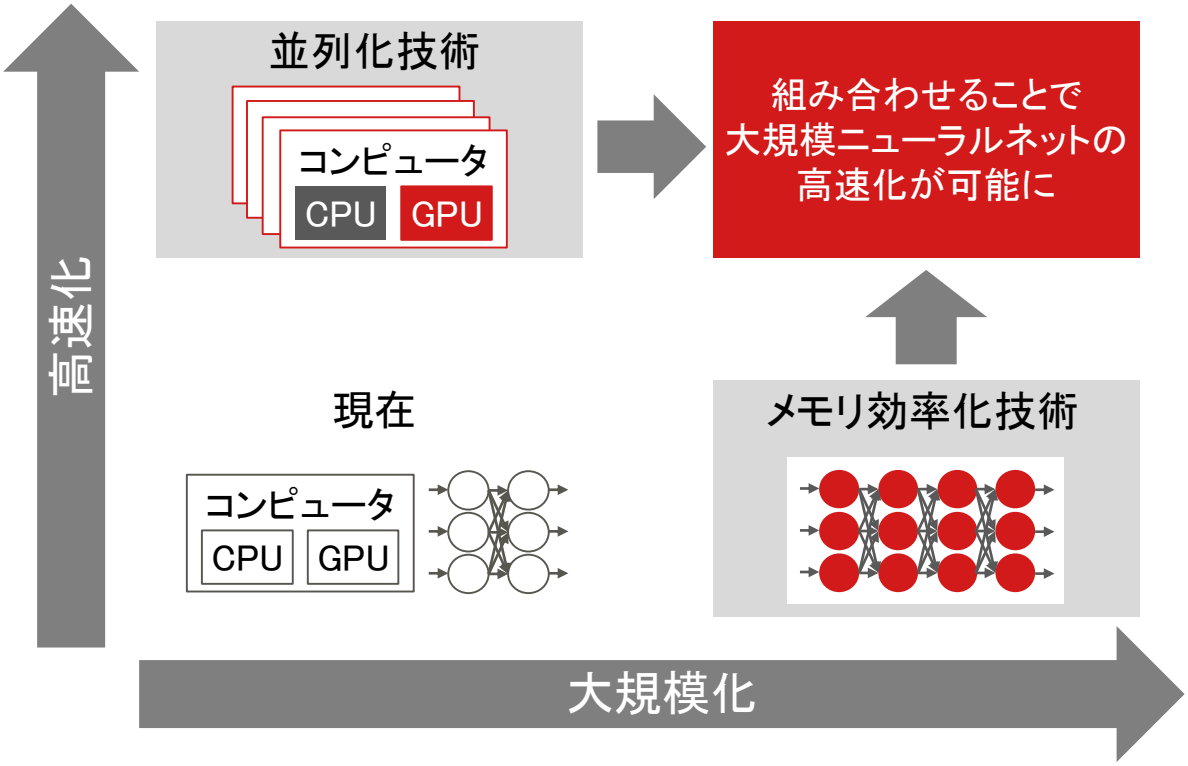
FUJITSU

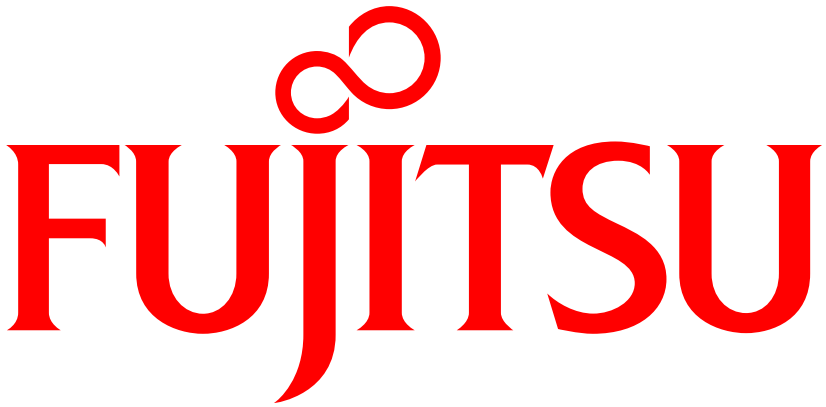
■ 削減したメモリ領域を活用してミニバッチサイズを2倍に拡大



今後

- 今回紹介した技術は富士通株式会社のAI技術「Human Centric AI Zinrai（ジンライ）」の一つとして、2017年4月から順次実用化を予定
- 「Deep Learning処理の高速化技術」と「メモリ効率化技術」を組み合わせて、技術の改善を行う





shaping tomorrow with you