FUJITSU

shaping tomorrow with you

富士通のPCクラスタ研究開発への取り組み

2017.12.14

(株)富士通研究所
中島 耕太

はじめに

FUJITSU

- PCクラスタの技術トレンドの変化
  - ムーア則の鈍化に伴う性能向上の鈍化
  - AIの台頭

- ムーア則の鈍化に伴うプロセッサの動向
  - 性能限界を攻めるために電力限界を睨んだ細かい性能制御
  - 電力制約を見越した性能見積もり技術が必要に

- AIの台頭
  - AI (DNN)がHPCの主要なワークロードに
  - 新しいワークロードに合わせたHPCシステムの設計が必要

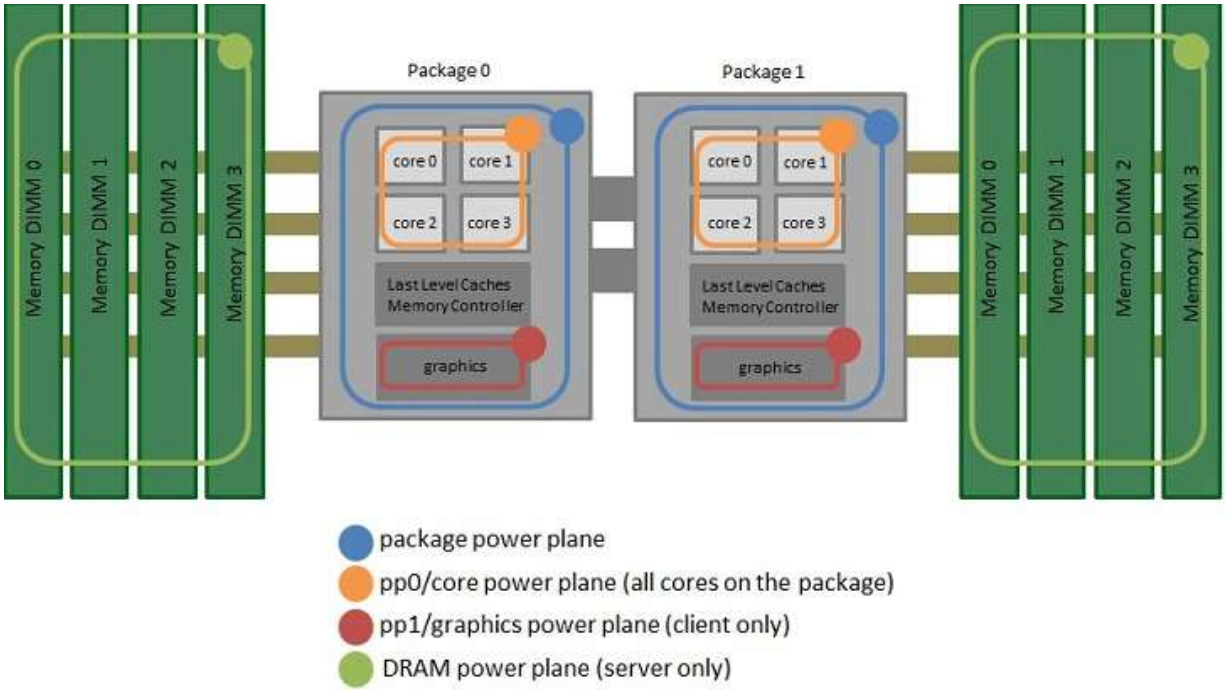
電力制約下での性能チューニング技術と
AIワークロード向け分散FS最適化技術についてご紹介

FUJITSU

電力制約下での性能チューニング

RAPL (Running Average Power Limit)

- 消費電力上限を制限する機能
- Sandy Bridge以降のIntel CPUに搭載
  - CPUパッケージ、CPUコア、メモリの消費電力上限を制約
  - Knights Landingにも搭載

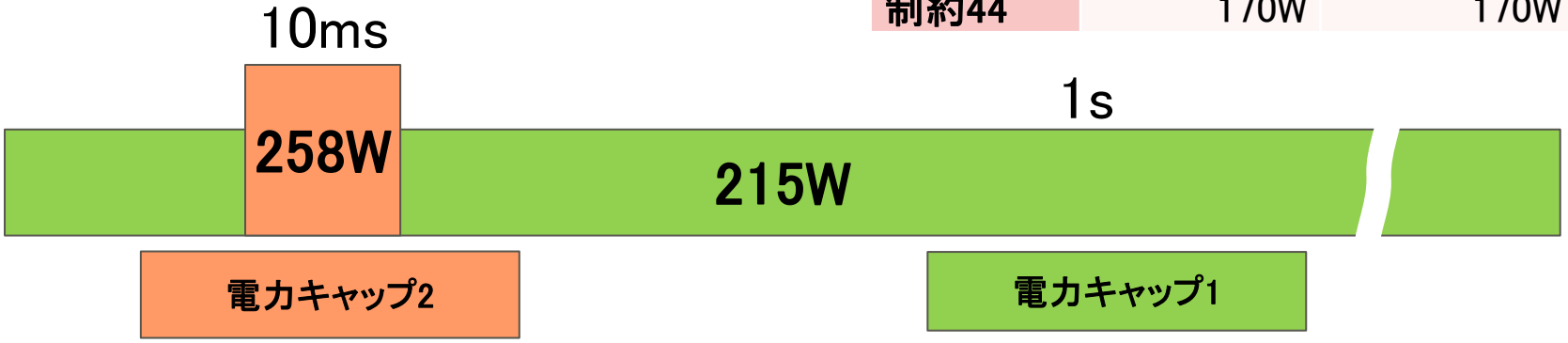


電力制約がどのように性能影響するかが知りたい

RAPLによる電力制約

- 10msと1sの2種類の制約
- 制約の組み合わせ

	キャップ1	キャップ2
制約11	215W	258W
制約12	215W	215W
制約21	200W	258W
制約22	200W	215W
制約23	200W	200W
制約31	185W	258W
制約32	185W	215W
制約33	185W	185W
制約41	170W	258W
制約42	170W	215W
制約43	170W	190W
制約44	170W	170W

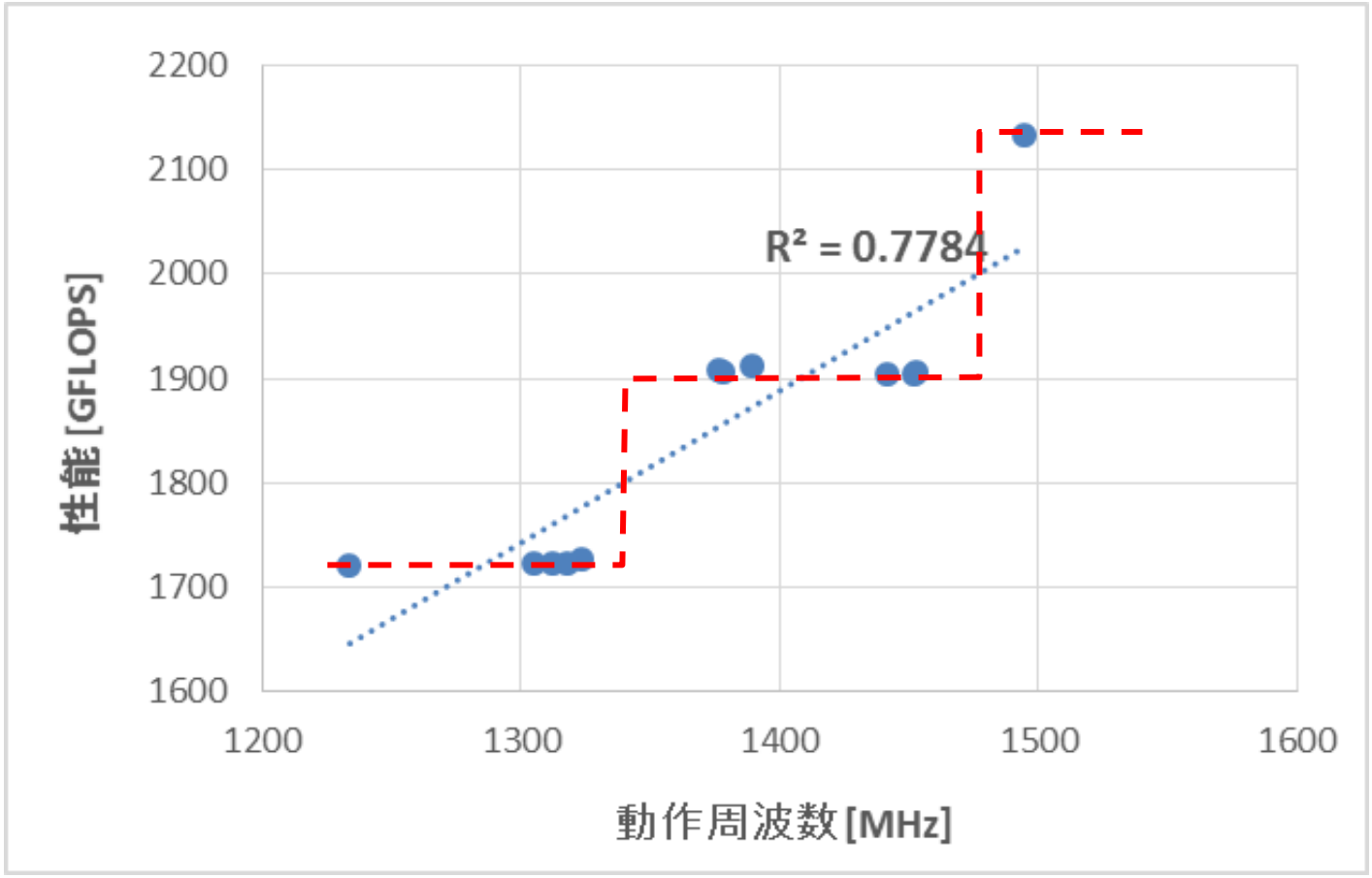


行列積性能の電力制約による変化を調査

KNL上の動作周波数と行列積性能の関係

FUJITSU

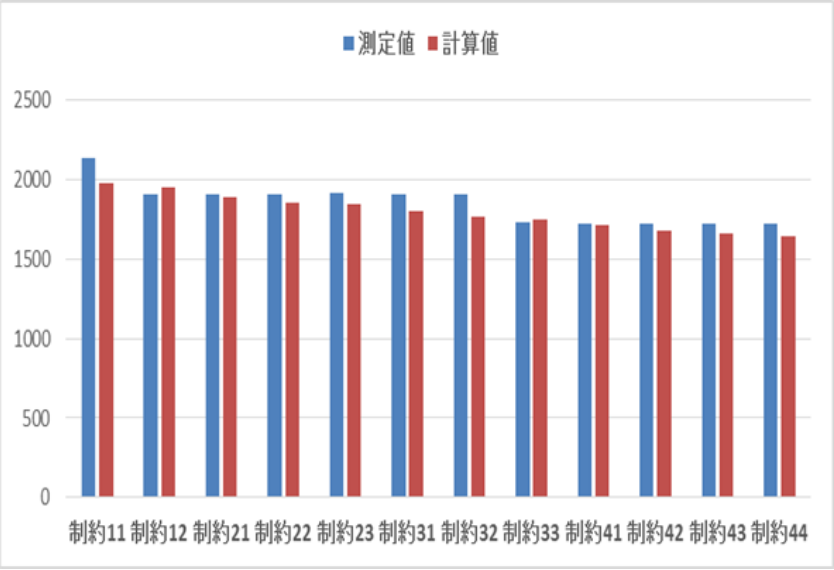
- 動作周波数を計測し、制約と行列積性能との関係をマッピング

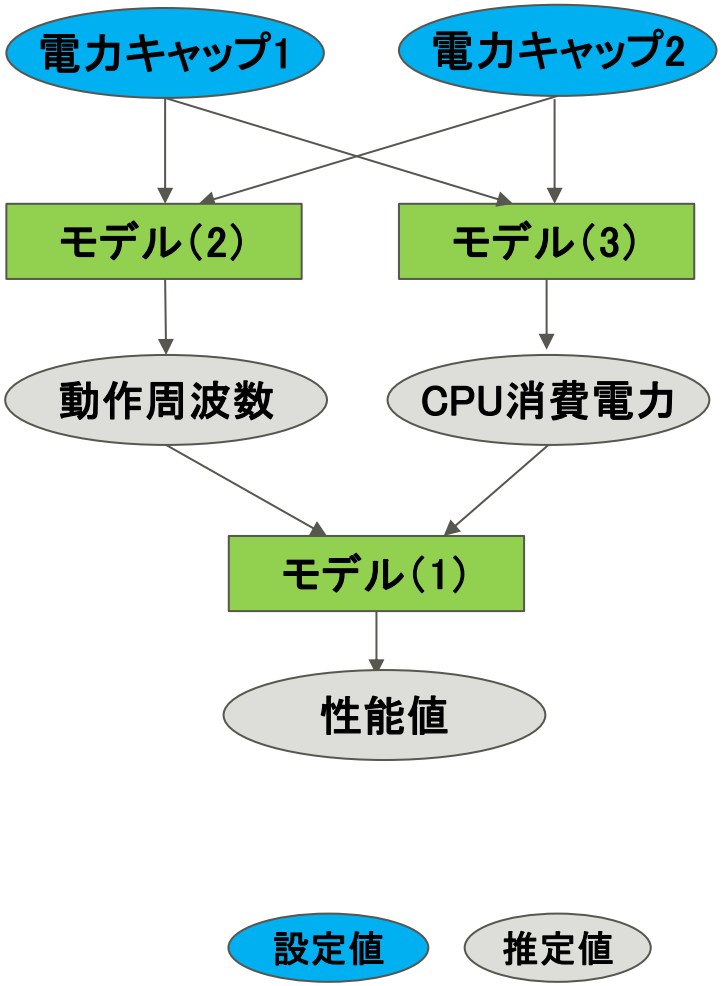


動作周波数だけでは性能をモデル化できない

性能モデルによる性能推定

- 電力制約の情報を加えて性能を予測
- 性能値の推定
  - 電力キャップ値から動作周波数(2)とCPU消費電力(3)を算出
  - 算出した動作周波数とCPU消費電力から性能値(1)を算出
- 誤差
  - 算出値と測定値の誤差：3%





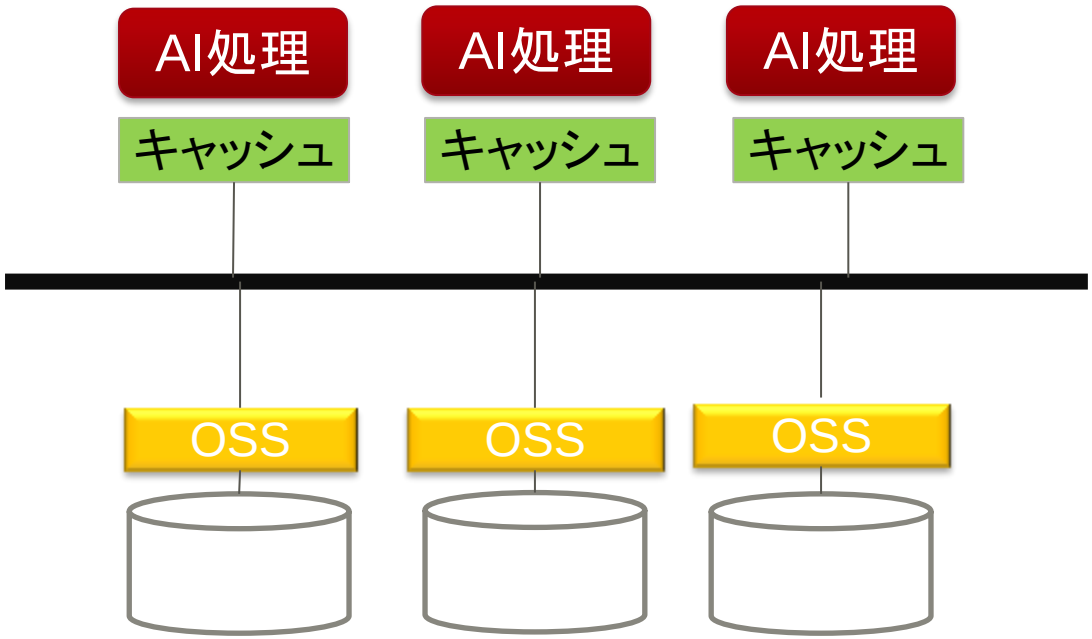
FUJITSU

AIワークロード向け
共有ファイルシステム

並列化したDeep Learningフレームワーク

FUJITSU

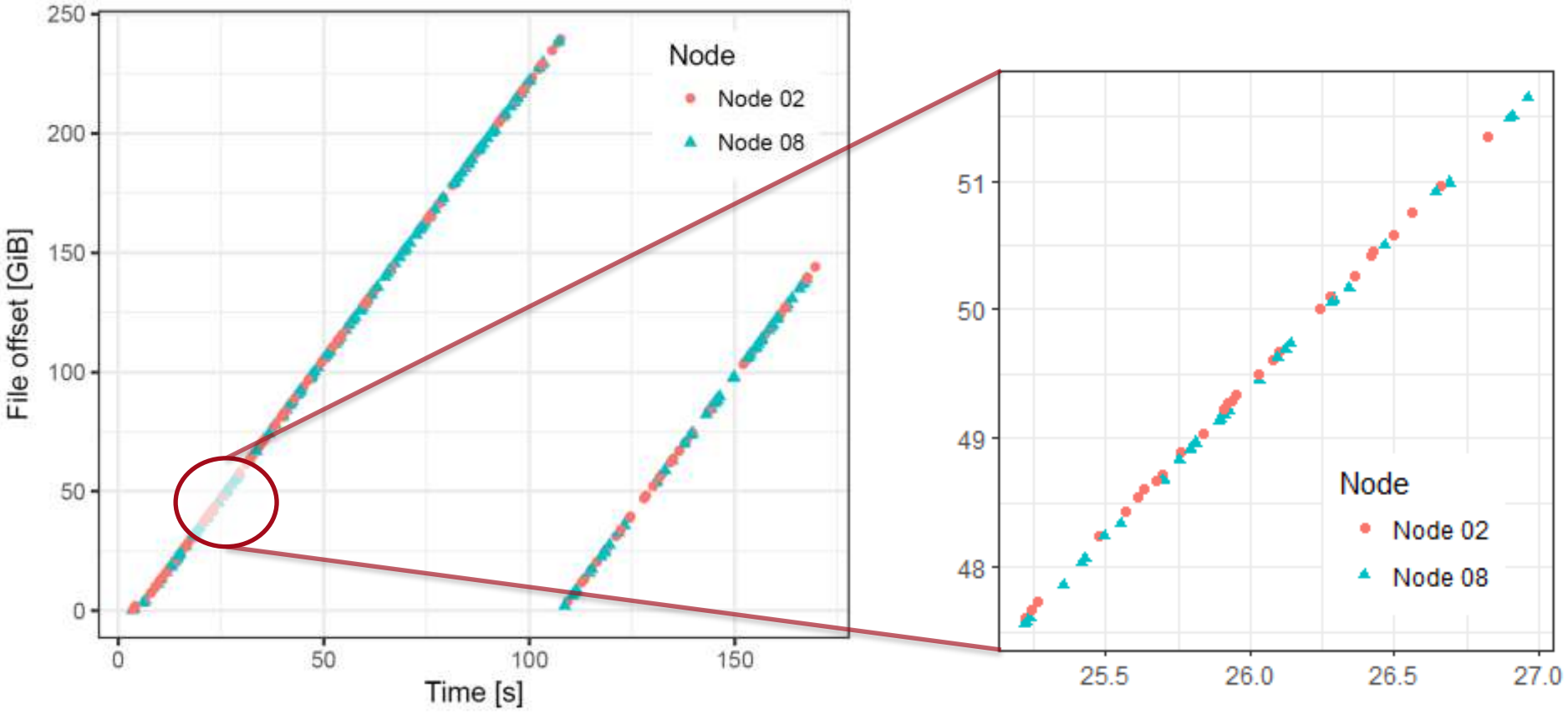
- Distributed Caffe: 富士通研が開発したMPI化したCaffe実装
- 分散並列DL処理が学習データにアクセス
- 学習データを分散FS(FEFS)上に配置



データアクセスパターンを分析

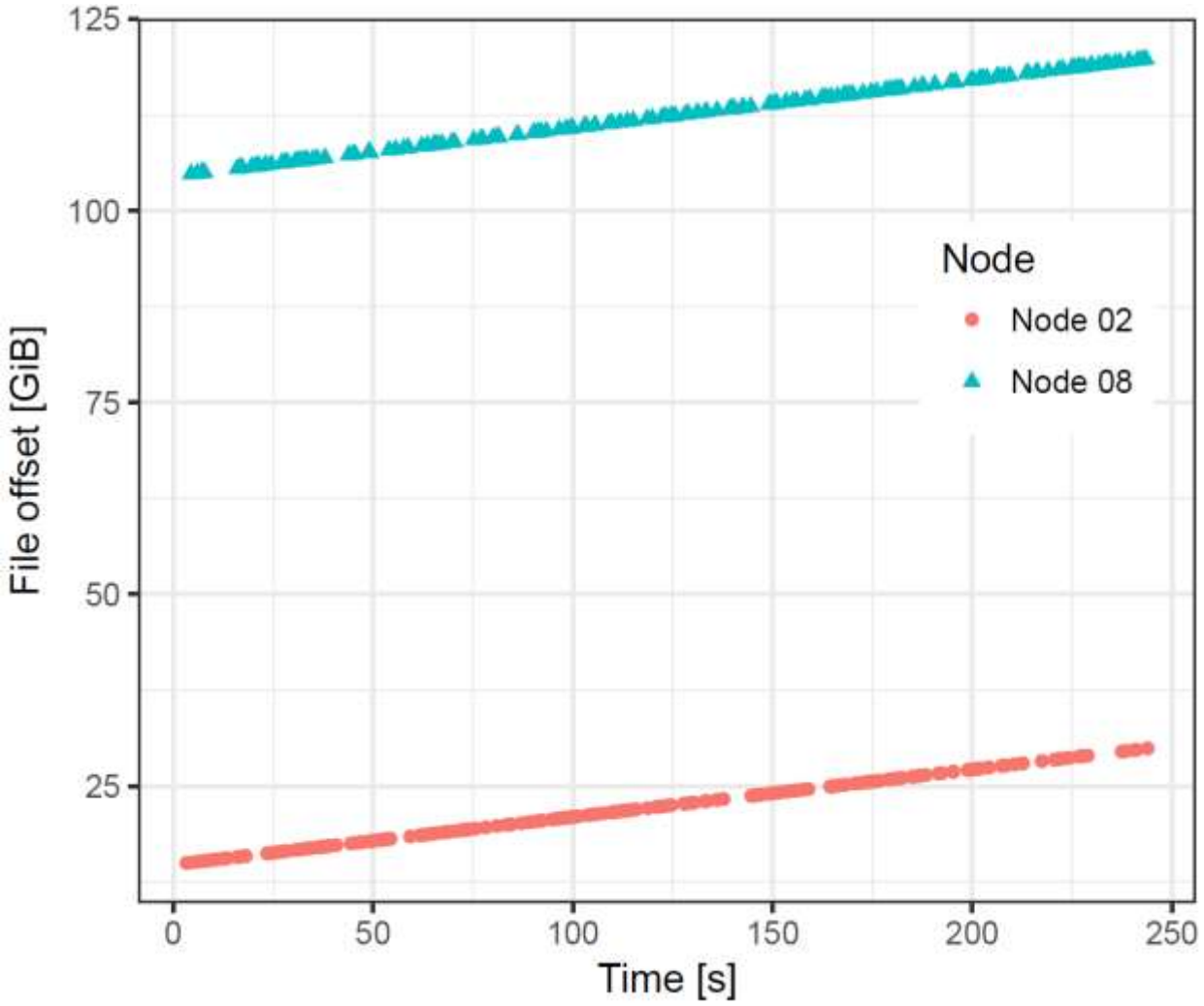
Deep Learningワークロードでの分析

- 実際に学習ワークロードを走行させた際のアクセスを分析
- 全体としてはほぼアドレス順にアクセスされる
- ノード毎にアクセスが混じる

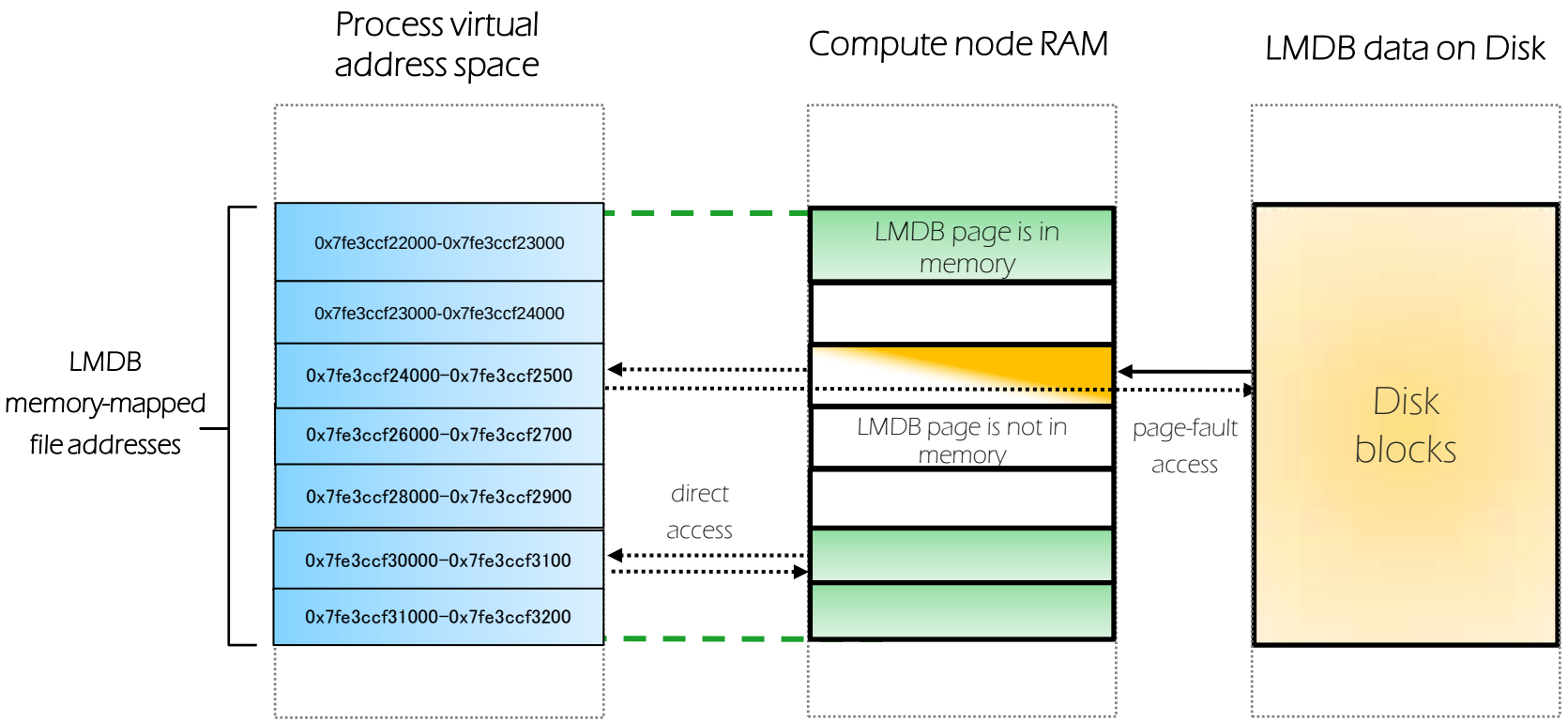


データの並び替えによるアクセス整理

- データを並び替えると、各ノードからのアクセスはほぼシーケンシャルアクセスに



Caffeでの学習データへのアクセス

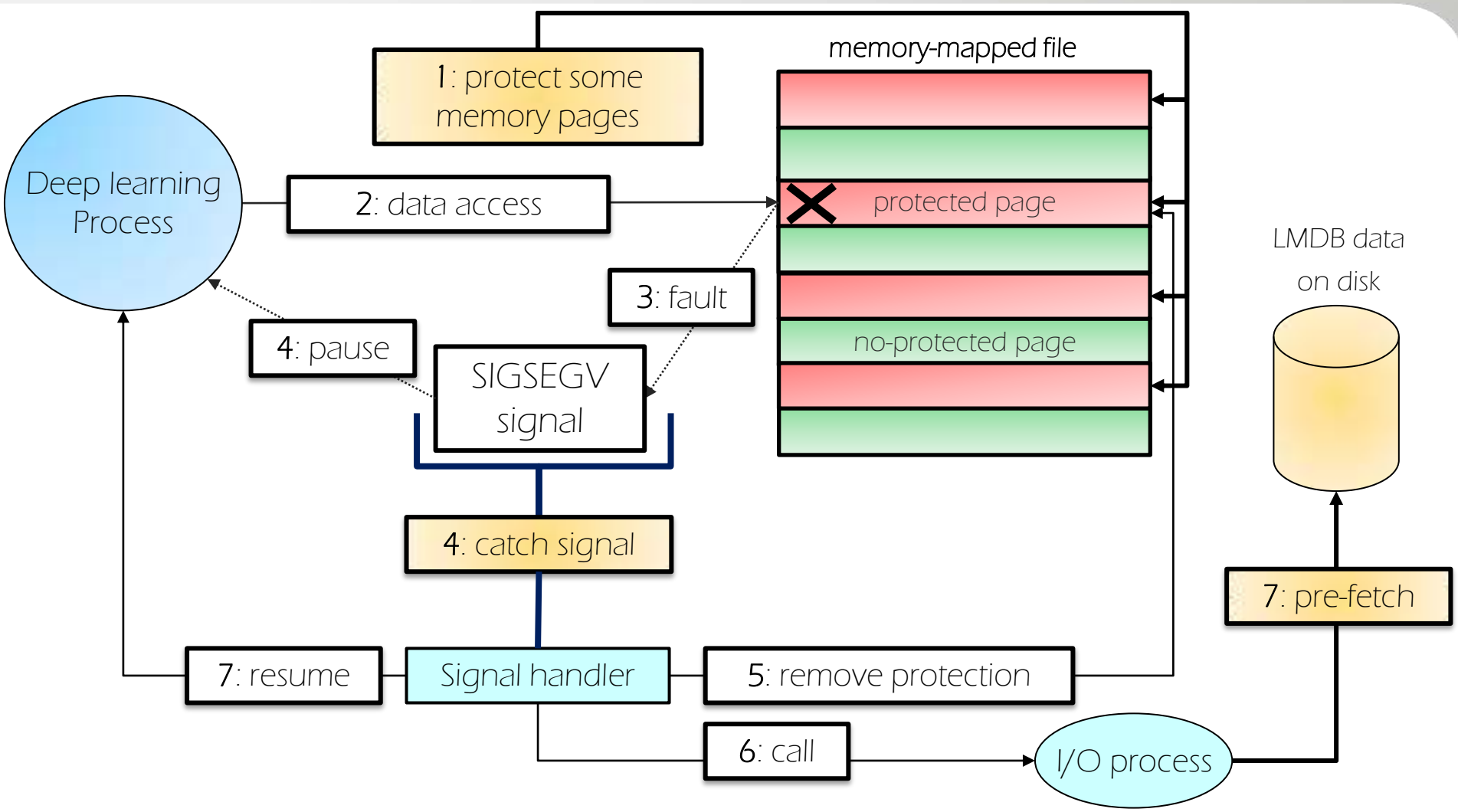


- 学習データはLMDB上に格納
- LMDB領域にはmmapによりアクセス
- Mmapされた領域のアクセス位置の特定が課題

データの原本はリモートにあるためmmapされた領域の先読みが重要

Mmapされた領域のアクセス位置の検知

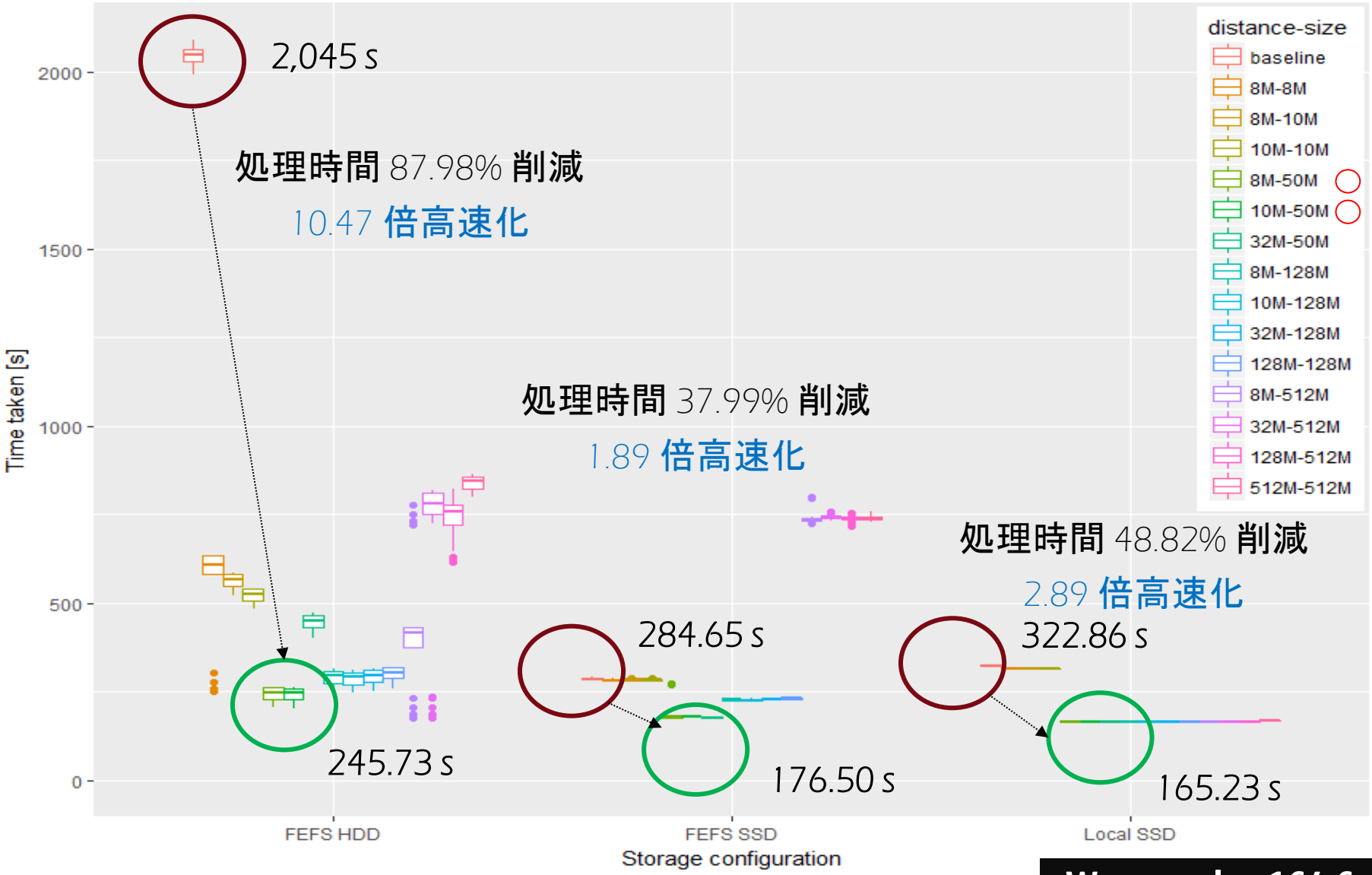
FUJITSU



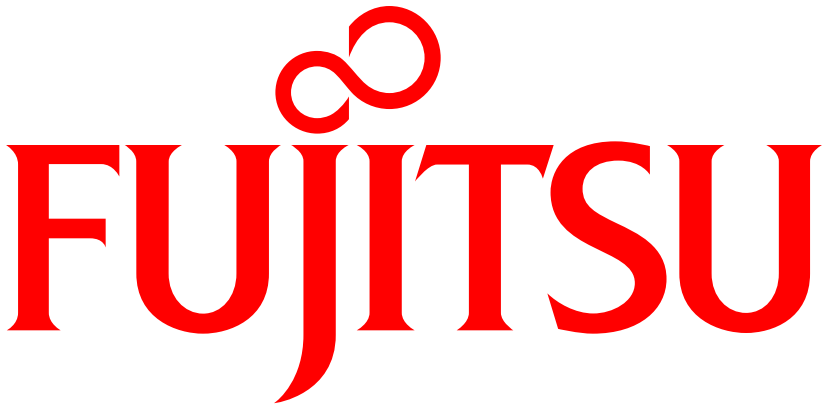
- LMDBをマップした領域にmprotectをかけ、シグナルを生成
- シグナルハンドラでアクセス位置を特定

先読みによる性能改善

FUJITSU



Warm cache: 164.6s



shaping tomorrow with you