ホワイトペーパー
FUJITSU AI Zinrai ディープラーニング システム
FUJITSU Storage ETERNUS NR1000A Series
増大する学習データに対応するディープラーニングシステム

2018 年 12 月
第 1.0 版

本書では、FUJITSU AI Zinrai ディープラーニング システムと FUJITSU Storage ETERNUS NR1000A Series で実現するディープラーニングについて説明します。
ImageNet データセットを使用して TensorFlow ベンチマークを実行し、学習性能、GPU 負荷、ストレージ I/O 負荷の結果を説明しています。

目次

1. はじめに

人工知能(AI)はさまざまな分野で活用されるようになっています。特にディープラーニングはより発展が望まれており、ディープラーニングで扱われるデータは増加傾向にあります。スマートフォン、タブレットのようなエッジデバイスやセンサーに代表される IoT デバイスから送信されるデータは、これからも増大していくことが予測されます。

最適なディープラーニングの学習モデルを作り上げるには、多数の学習ジョブを実行していかなければなりません。さらに共有環境ではより多数のユーザが利用することになり、それに伴いアクセスするデータ量は増大していきます。

ディープラーニングの社会的な利用を模索していた頃、試験的に作られるディープラーニングではパソコンのローカルディスクに教師データを格納していました。しかし、上記のような業務システムに組み込まれるディープラーニングシステムや複数ユーザで共同利用するディープラーニング基盤では、大量のデータを格納できるストレージが必要です。また、障害に備えて、データの可用性、バックアップも必要です。

そして、GPU の飛躍的な性能向上により、サーバで並列に処理できるディープラーニングの学習ジョブ数も増加しています。そのような多数の学習ジョブに対してデータを安定的に供給し続けられる基盤システムが必要であり、サーバの処理能力とストレージの処理能力、その間のデータ転送能力のバランスが良いことが望まれます。

FUJITSU Storage ETERNUS NR1000A Series は、オールフラッシュアーキテクチャを採用した高速処理、コンポーネント冗長化、RAID 方式による高可用性、SnapShot 機能による高速なバックアップリカバリー、SnapMirror 機能による遠隔地ミラーリングを兼ね備えた NAS ストレージであり、ディープラーニングの学習データの格納場所に最適です。

FUJITSU AI Zinrai ディープラーニング システムは、高性能なハードウェアとディープラーニングで必要となるソフトウェアを組み合わせた垂直統合型のディープラーニング基盤であり、簡単システム運用を実現する運用機能や使いやすい Web GUI によりディープラーニングを実行するコンピューティングノードとして最適です。

本書では、これらの最先端のストレージとコンピューティングノードを組み合わせ、学習データを ETERNUS NR1000 A300 に配置し、Zinrai ディープラーニング システムでディープラーニングの学習を行った場合の学習性能、GPU 負荷、ストレージ I/O 負荷、および拡張性について考察します。

2. コンピューティング – Zinrai ディープラーニング システム

2.1 即時利用が可能な Zinrai ディープラーニング システム

Zinrai ディープラーニング システムは、GPU (Graphics Processing Unit) を含むサーバとソフトウェアを一括で提供するディープラーニング基盤です。

ディープラーニングを活用した分析のためのシステムであり、NVIDIA® Tesla® V100 をサポートした、ディープラーニング実行環境が統合された環境を提供します。将来的には富士通が独自開発したディープラーニング専用ユニット(DLU)もサポート予定です。ハードウェア・ソフトウェアをディープラーニングに最適な組み合わせで検証・構築しているため、お客様はインストールや設定を行うことなく、スピーディーに利用開始することが可能です。

2.2 学習規模に合わせた段階的拡張

Zinrai ディープラーニング システムに搭載する GPU はお客様要件に応じて、1～8 枚から選択可能です。導入後の利用者・学習量の増加にあわせて段階的な拡張が可能です。

2.3 簡単システム運用の実現

Zinrai ディープラーニング システムはマルチユーザ環境でも簡単にシステム運用することが可能です。複数のユーザが利用する環境においても、各ユーザは運用管理者に割当てられた GPU を占有でき、常に安定した性能を維持可能です。また、ユーザは割当てられたユーザ環境にのみアクセス可能なため、ユーザ間のセキュリティを確保できます。

2.4 豊富なフレームワークの提供

Zinrai ディープラーニング システムは、GPU を共有するバッチ型学習環境と GPU を専有する対話型学習環境を提供します。用途によって使い分け、組み合わせた利用が可能です。

バッチ型学習環境は、フレームワーク Caffe が利用可能です。ウェブブラウザから学習ジョブを投入できます。投入された学習ジョブは、投入順(FIFO[先入れ・先出し])に学習が実行されます。また学習の進捗状況を確認できるダッシュボードを提供します。

対話型学習環境は、広く利用されているフレームワークである Caffe, Chainer, TensorFlow, MXNet に加えて、富士通が開発した時系列として連続したデータのディープラーニング技術を用いて、畳み込みニューラルネットワークを使用した学習と分類を行う Time Series Data Analysis が利用可能です。また、それ以外のフレームワークもお好みに応じてインストールし、ご利用いただくことも可能です。

3. ストレージ － ETERNUS NR1000A

3.1 広範な接続性・運用性とパフォーマンスを兼ね備えた、万能ストレージ

ETERNUS NR1000A series とは、もともとファイルサーバ専用機や仮想基盤向けのストレージとして根強い人気を博していた NR1000F series のオールフラッシュ向け改良版です。多彩なアクセスプロトコル・頻出アプリケーションへの機能サポートといった特長はそのままに、SSD 搭載に特化したストレージとして一層のパワーアップを果たしました。その卓越した性能は、低いレイテンシのもと大量の I/O を処理する必要があるディープラーニングの基盤としても、お客様にとって最適な選択肢のひとつとなります。

3.2 SSD の性能を最大限に引き出す “ONTAP9” システム

フラッシュ自体が高速なのはもちろんのこと、その特性をより活かす機構として本シリーズに独特の ”WAFL” ファイルシステムにより、Random Write 処理の高速化と SSD の長寿命化が図られております。さらに、最新 OS である ONTAP9 では SSD に対する読み取りに特化した最適化処理を実装しており、レイテンシの低減・IOPS 性能の向上を実現しております。

3.3 スモールスタート・スケールアウトが可能な拡張型アーキテクチャ

テスト導入から本番用へ、そして順次拡張といった一連のアクションと相性が抜群のスケールアウト型となっており、システムを稼働させたまま容量・性能を自由自在に拡張することで、ハードウェアの乗換え・データ移行といったシステム規模の拡大に伴うお悩みを払拭します。
今回ご紹介する “NR1000 A300” では最小2ノード(1 HA-Pair)・11.5TB(物理容量)から構成できますが、最大で24ノード(12 HA-Pair)・140PB (物理容量) までの拡張余地を備えており、最大性能は4,200,000IOPS, 110GB/s にまでも達します。

3.4 データの堅牢な保護と高度な利活用を提供

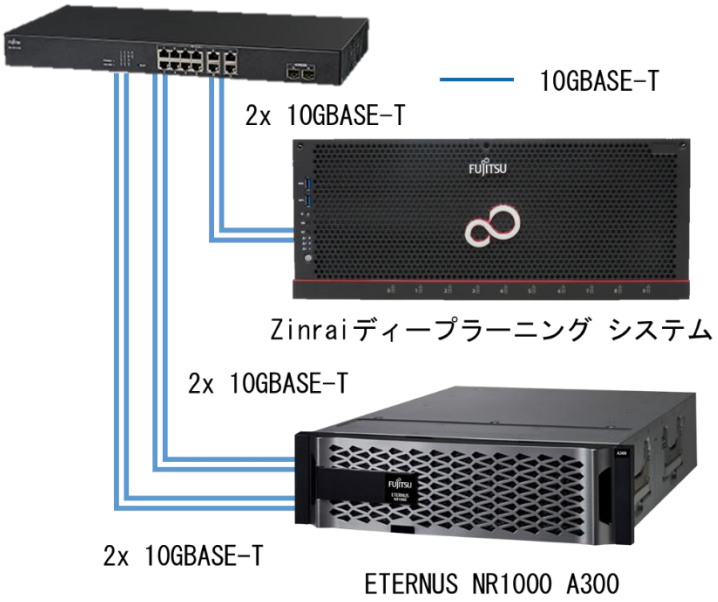
もともと高信頼ファイルサーバ・仮想基盤として有力視されていたストレージシステムであり、ストレージコントローラーをはじめとした各コンポーネントの徹底した冗長性確保・ファイルシステムの安定性・データ保護といった基本的要件は豊富な運用実績に裏打ちされており、データを守り運用を止めないといった可用性・信頼性の観点は万全といえます。
これに加えて、ディープラーニングシステムにおいては大量データの配置・転送・利活用といった視点も重視されます。本機ではデータのバックアップやサイト間転送として有用なSnapMirror、アクセス頻度の落ちたコールドデータのクラウド移行を自動化するFabricPoolといった仕組みを内包。「速いだけ」にとどまらず、データのライフサイクルに寄り添ったきめ細かいデータマネジメントを実現します。

4. 検証構成

ディープラーニング学習時間を短縮するには複数の GPU を最大限活用するための低レイテンシかつ高 I/O スループットを維持できるシステム設計が必要不可欠です。この要件を満たすためには高速、広帯域幅、低レイテンシのイーサネット・ファブリックをサポートするストレージシステムが必要となります。複数の Zinrai ディープラーニング システムに絶えずデータを供給することで、各 GPU のパフォーマンスを最大化します。

図 1 は、検証に使用した 1 対 1 構成の Zinrai ディープラーニング システム × ETERNUS NR アーキテクチャです。 1 台の 10GBASE-T 対応レイヤー 2 スイッチを介して 1 つの ETERNUS NR1000 A300 高可用性 HA-Pair からデータ フィードされる 1 台の Zinrai ディープラーニング システムで構成されます。Zinrai ディープラーニング システムは、スイッチに 2 本の全二重化された 10GBASE-T で接続されています。ETERNUS NR1000 A300 はコントローラーごとに全二重化された 2 本の 10GBASE-T を介してスイッチに接続されています。

図 1） 1 対 1 構成の測定環境



5. パフォーマンステスト

1 対 1 構成の設定（1 台の Zinrai ディープラーニング システムと 1 台の ETERNUS NR1000 A300 HA-Pair）において、ETERNUS NR1000 A300 の FlexVolume に格納された ImageNet データセットを使用して、TensorFlow ベンチマークを実行しました。このテストでは、ファイルシステムとして NFSv3 を選びました。
環境設定：

- OS：Ubuntu 16.04 LTS
- NVIDIA ドライバ：NVIDIA-Linux-x86_64-390.30.run
- Docker：docker-ce_18.03.1~ce-0~ubuntu_amd64.deb
- nvidia-docker：nvidia-docker2_2.0.3+docker18.03.1-1_all.deb
- Dockerfile：tensorflow/tensorflow:1.11.0-gpu
- フレームワーク：TensorFlow 1.11.0-gpu
- ベンチマーク：TensorFlow ベンチマーク
  https://www.tensorflow.org/performance/benchmarks
  [commit：4c7b09ad87bbfc4b1f89650bcee40b3fc5e7dfed]
- データセット：ImageNet データセット
  http://www.image-net.org/

最初にストレージ I/O に影響を受けないようにプログラム上で合成したデータを用いてベンチマークを実行し、ストレージに影響されない GPU 性能を計測しました。CPU 性能などのストレージ以外のボトルネックが発生せず、GPU 使用率がほぼ 100%になることを確認しました。次にストレージ上のデータを用いて計測しました。

計測方法として、下記のポイントを採用しました。

- 各モデルを学習する GPU 性能を、1 秒あたりに処理された画像の数として測定する。
- 学習させるモデルは計算量に関わるネットワークの複雑さを考慮し、一般的なモデルから ResNet-50 を採用する。
- 学習に使用する GPU 数、バッチサイズを変えて、複数回測定する。
- ストレージから十分な速度でデータを供給できることを確認するため、学習中に GPU 使用率がほぼ 100%になるよう負荷を掛ける。

図 2 は、GPU の数が 1 枚、2 枚、4 枚、6 枚、8 枚(*1)の場合にモデルで測定した学習パフォーマンスの結果をまとめたものです。GPU の数を増やすとリニアに学習速度が速くなることが分かります。これはコンピューティングノードをスケールアウトすることで、優れた学習パフォーマンスを得られることを示唆しています。
(*1) 8 枚は予測値

図 2) ImageNet のデータを使用した学習速度

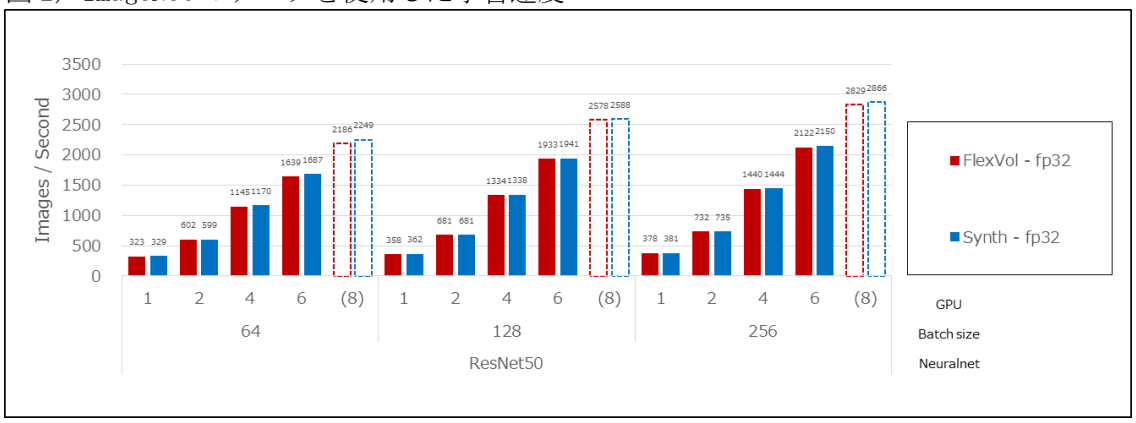
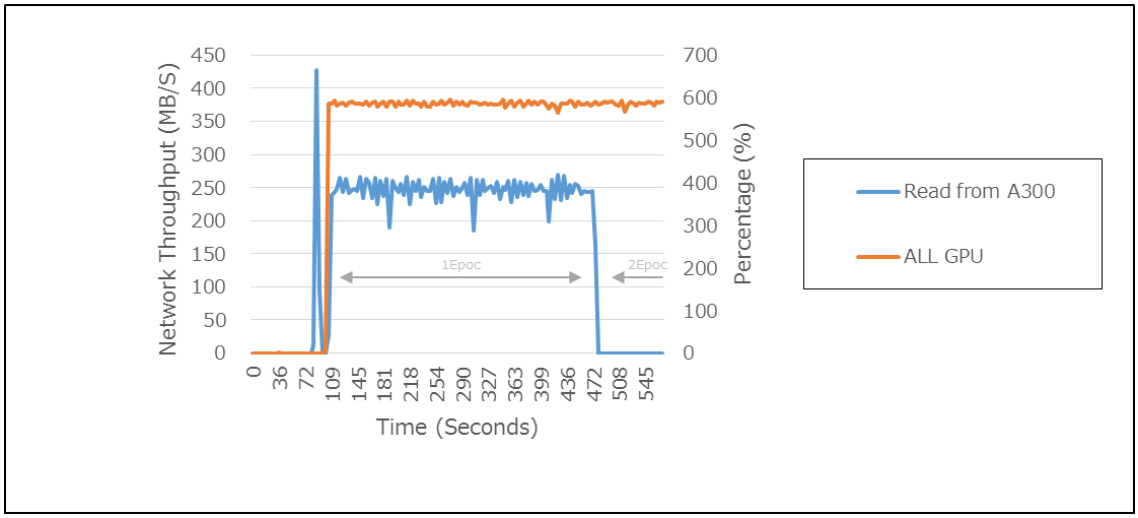


図 3 は、GPU 6 枚で ResNet-50 モデルを学習しているときの GPU 使用率を示しています。青色の線は 6 枚すべての GPU 使用率の合計を、オレンジ色の線は A300 からの読み取りスループットをそれぞれ示しています。高い GPU 使用率と 1 秒あたり約 2100 個の画像学習速度を維持するため、A300 からの読み取りスループットはおよそ 250MB/s に達しています。データセットを Zinrai ディープラーニング システムのメモリに読み込むのにおよそ 472 秒かかっています。472 秒の前後で GPU 使用率と学習速度は変化ありません。この学習速度では、ストレージ I/O がボトルネックにならず、GPU に十分なデータを供給できたことを示しています。

図 3) ResNet-50(バッチサイズ 256)で 1 秒あたり画像約 2100 個の速度での GPU 利用率と A300 からの読み込みスループット



6. 拡張性

スケールアウトとは、ストレージ環境の拡大に合わせて、共有ストレージインフラ上のリソースプールに、ストレージ容量やコンピューティングノードをシームレスに追加することです。ホスト接続とクライアント接続に加えてデータストアも、リソースプールのうちの任意の場所にシームレスに移動できます。したがって、既存のワークロードは利用可能なリソース上で均等に分散でき、新しいワークロードは容易に導入できます。テクノロジの更新(ドライブ シェルフやストレージ コントローラーの追加や交換)も、環境をオンラインに保ったままで、データのフィードを続けながら実行できます。

富士通は、Zinrai ディープラーニング システムのコンピューティング能力と ETERNUS NR1000 A300 のハイパフォーマンスアーキテクチャとを組み合わせることで、ディープラーニングワークフローを数時間のうちに導入し、必要に応じてシームレスにスケールアウトすることを可能にする、魅力的なソリューションを実現しました。

ディープラーニングに着手する場合は、まず 1 対 1 構成から始めて、ワークロードの増大に合わせてスケールアウトしていくのがよいでしょう。Zinrai ディープラーニング システムを 2 台、3 台とスケールアウトした構成(1 対 3 構成)においても優れた学習パフォーマンスを発揮すると予想されることが、今回の検証結果から確認されています。表 1 に、Zinrai ディープラーニング システムと ETERNUS NR1000 A300 の幅広い構成で実現できる容量とパフォーマンスの拡張を示します。

表 1) A300 を使用したスケールアウト シナリオでの容量とパフォーマンスの指標

ストレージの台数	サーバの台数	スループット	最大物理容量	実効容量(*2)
HA-Pair×1	3	9.7GB/s	5,875.2TB	20,870.0TB

(*2) ストレージ容量削減比率 5:1 に基づく実効容量

表 1 の情報は、ETERNUS NR1000 A300 と ONTAP 9.4 のパフォーマンス指標に基づいています。各 A300 は 9.7GB/s のスループットを実現し、3 台の Zinrai ディープラーニング システムからのトラフィックを処理できます。今回の検証構成では、1 対 3 構成で 10GBASE-T 対応レイヤー2 スイッチのポート数が上限に達しますが、ETERNUS NR1000 A300 のスループットには余裕があるため、ポート数の多いスイッチに交換することによって、さらに多くの Zinrai ディープラーニング システムを接続することが可能です。また、より大量のストレージ容量が必要な際には、ETERNUS NR1000 A700 といった上位のストレージシステムを使用することもできます。

Zinrai ディープラーニング システムと ETERNUS NR1000A Series のラックあたりの搭載数は、使用中のラックの電力と冷却の仕様によって異なります。システムの最終的な配置は、数値流体力学的な分析、エアフロー管理、データセンター設計によって決まります。

7. まとめ

Zinrai ディープラーニング システムと ETERNUS NR1000 A300 を組み合わせてディープラーニングの学習をした場合、I/O 性能がボトルネックになることなく、GPU の性能をフルに使い切れる環境を実現できたことを検証しました。今回検証したディープラーニング環境は、データやワークロードに応じてシームレスにスケールアウトすることが可能です。

ますます活用が期待されるディープラーニングにおいて、学習データの重要性は高まっています。大量のデータの通信、蓄積、処理を行えるディープラーニングシステムが必要となりますが、Zinrai ディープラーニング システムと ETERNUS NR1000A Series の組み合わせはその解を与えるものになります。

将来、Zinrai ディープラーニング システムは現行の電力性能比 10 倍を実現する DLU を搭載することが予定されています。コンピューティングノードのさらなる性能向上に対して、ETERNUS NR1000A Series ならば、さらに高性能なモデルが用意されており、適切に学習データの提供を行うことができるでしょう。

今後も富士通は、NAS ストレージ業界のリーディングカンパニーであるネットアップ社と協力しまして、ディープラーニングに最適なソリューションを提供してまいります。

付録. コンポーネントリスト

表 2 は、本書で説明した検証構成のコンポーネントを示したものです。

表 2) コンポーネントリスト

サーバ	
製品名	Zinrai ディープラーニング システム
装置数	1
GPU	NVIDIA Tesla V100
GPU 数	6 (最大 8)
CPU	Intel Xeon E5-2690v4
CPU 数	2
メモリ容量	512GB (最大 512GB)
内蔵ストレージ	4TB 7200rpm SATA HDD x 6
ストレージ	
製品名	ETERNUS NR1000 A300
HA-Pair 数	1 (最大 12)
ドライブ種別	960GB SSD
ドライブ搭載数	24 (最大 384)
スイッチ	
製品名	SH-E514TR1
装置数	1
基本インターフェース	100/1000/10GBASE-T: 12 10GBASE-SR/LR/CR: 2

改版履歴

改版年月	版数	改版内容
2018 年 12 月	1.0	初版

富士通株式会社

〒105-7123
東京都港区東新橋1-5-2
汐留シティセンター