ホワイトペーパー

自律走行ソフトウェア企業導入事例

PURE STORAGE と NVIDIA による深層学習のためのリファレンスアーキテクチャ

PURESTORAGE®

はじめに

人工知能（AI）は、さまざまな分野においてイノベーションを促進しています。機械学習の活用が期待されている領域の一つが、完全に自律した自動運転車です。このアイデアは、つい数年前までは未来のものと考えられていました。しかし、高度なトレーニングアルゴリズムの出現と、コンピューティングやビッグデータ分野の大きな進歩を背景に、自動車業界では、まさにサイエンスフィクションのような技術の実現を視野に入れるようになっています。トレーニングアルゴリズムやコンピューティングシステムには、大量のデータを迅速に供給しなければなりません。そのためには、最新の分析技術のために開発されたストレージソリューションが必要です。

NVIDIA®と Pure Storage®は、企業が深層学習と AI をプライベートクラウド環境で活用できるようにすることを目指し、協力体制で臨んでいます。このドキュメントでは、完全自律走行ソリューションの 5 年以内の実現を目標に、NVIDIA DGX-1™と Pure FlashBlade™を基盤とする AI インフラストラクチャを導入した、ある匿名のお客様の事例を取り上げます。

自動車業界では、自律走行の実現に向けた競争が始まっており、早期の市場展開を目指した開発が急務となっています。ミュンヘンを拠点とする自動車関連のコンサルティング企業、Beryll Strategy Advisors 社の Jan Burgard 氏は、次のように述べています。

「自動運転車を他社に先駆けて提供することはきわめて重要です。なぜなら、このテクノロジーが自動車業界のコアビジネスを覆す可能性があるからです。他社に先駆けることで、このテクノロジーの脅威を正しく理解し、将来の方向性を決めるうえで、有利な立場に立つことができます。」

自律走行の実現には、多くの優秀な技術者、膨大な計算能力、大規模かつ最新のデータプラットフォームが必要です。そして、それぞれに要求される規模は今後も増大する一方であると予想されます。このドキュメントで取り上げるお客様は、ミッションを支えるインフラストラクチャを構築する手段として複数の選択肢を検討し、最終的に NVIDIA DGX-1 と Pure FlashBlade™を選択しました。この 2 つのソリューションを導入するに至った技術的な要因を中心に説明します。

「自動運転車は、社会に重大な影響をもたらすイノベーションです。自動車事故の 94%はヒューマンエラーが原因だといわれています。自動運転車の実用化により、事故を大幅に減らすことができるでしょう。」

— Audi 社

お客様の要件

このお客様のシステムでは、開発のさまざまなフェーズにおいて、地理的に分散したチームにより機械学習が利用されています。NVIDIA DGX-1 と Pure FlashBlade がこの環境向けに選ばれたのは、密度の高いフォームファクターでコンピューティングとストレージの両方を提供することができるためです。これらのソリューションでは、従来のソリューションでは数十台のラックを要するところを数 U の機器に集約することができます。このお客様のシステムでは、データセンターの設置面積を最適化し、電力と冷却のコストを最小限に抑えることが不可欠であったため、この密度の高さは重要でした。

プライマリデータセットには、平均サイズが 1～2.5 MB の高画質の JPG 画像が数千万枚あります。これらの画像は、モデルの大規模なトレーニングと、環境内での前処理や探索のジョブを実行するために使用されます。詳細は、「アーキテクチャの概要」セクションを参照してください。

> 完全なシミュレーションテストの実施中は、各 GPU がプライマリデータセットのファイルを無作為かつ継続的に高速で読み取ります。この規模のデータセットから学習用の画像を無作為に選択すると、キャッシュが役に立たなくなるため、ストレージアプライアンスからデータを直接取り込む必要がありました。FlashBlade の導入を決定するまで、多数のディスクを搭載したホワイトボックスサーバーや、他社のスケールアウトストレージなど、お客様はさまざまなストレージソリューションを試しましたが、GPU のスピードに対応できるものはありませんでした。

この AI スーパーコンピューターのもう一つの重要な要件はシンプルであることです。データサイエンティストは、インフラストラクチャの管理ではなく、モデルのチューニングとトレーニングに時間をかけたいと考えています。DGX-1 は、深層学習向けに事前に最適化と構成を行ったフレームワークとツールを備えているため、深層学習のワークフローのために環境をチューニングする必要がほとんどありません。

同様に、FlashBlade は、かつてない容易さで管理することが可能で、環境内の任意のトレーニングノードにデータを均等に供給できます。その結果、データサイエンティストは、データの管理に時間を割く、つまりトレーニングのたびにデータをコピーする必要がなくなります。また、パフォーマンスを最大化するためにハードウェアノードをチューニングする必要もなくなります。データサイエンティストの業務が大幅に効率化されました。

アーキテクチャの概要

このアーキテクチャは、主に次の 2 点に重点を置いて設計されました。一つ目は、反復型であるというプロジェクトの性質に応じて、動的で俊敏性に優れたインフラストラクチャを構築することです。二つ目は、今後、AI ベースのソリューションや機能が成熟しても、コンピューティングとストレージへの投資が無駄にならないようにすることです。今日では、固定的なソリューション向けに専用のインフラストラクチャを構築すべきではありません。このような例は、多数のディスクから構成される Hadoop 型のアプローチでよく見られますが、今日求められているインフラストラクチャはそれとは異なり、データを研究するチームが開発、テストしている内容に応じて、規模を柔軟に変化させられるものです。

図 1 は、深層学習用インフラストラクチャとしてお客様が最初に展開した環境の構成を示しています。NVIDIA DGX-1 ノードのクラスタリングの詳細は、こちらを参照してください。

この自律走行ソフトウェア企業のデータサイエンティストは、クラスタリングされた DGX-1 インスタンスだけを使用するわけではありません。完全なトレーニングシミュレーションに先立って、反復型の開発作業のために、CPU のみを搭載したコモディティノードや、NVIDIA Tesla P100 などの単一（または複数）の GPU を搭載したサーバーを使用します。

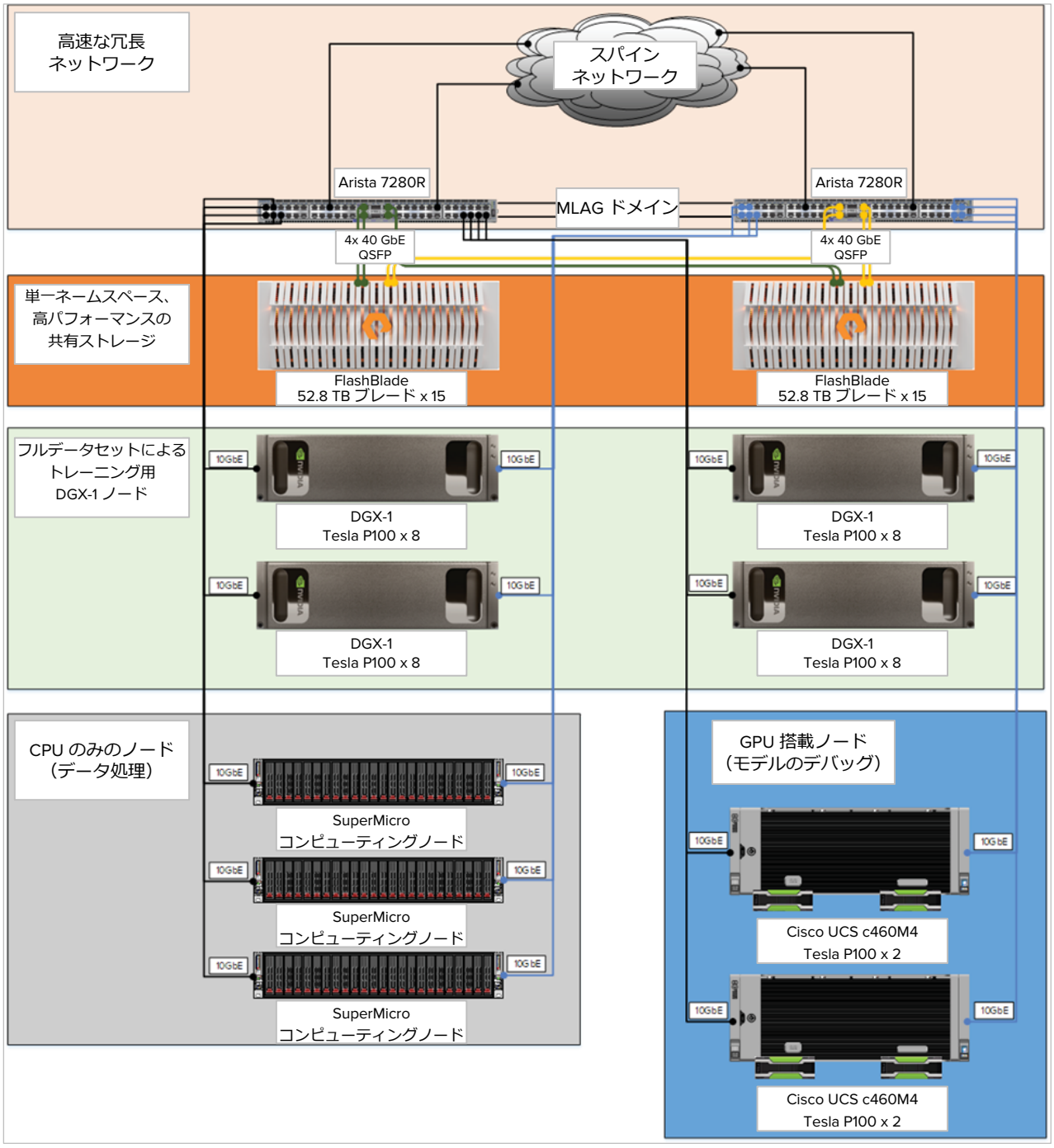

図 1：自律走行ソフトウェア企業におけるアーキテクチャ

図 1 の各部の簡単な説明を次に示します。

ネットワーク： 図の最上部には、冗長化された 10/40/100 Gb 対応の高速イーサネットネットワークスイッチのペアがあります。帯域幅は、将来的な拡張のため、十分な余裕を持たせてあります。Arista 社のスイッチを MLAG ドメインを介してペアリングしています。コンピューティングとストレージのすべての要素が、ロードバランシングのため、また単一のポートまたはスイッチで問題が発生した場合に備えて冗長性を確保するため、各スイッチに接続しています。AI のトレーニングに使用する複数のコンピューティングノードはすべて、FlashBlade に格納されているトレーニングデータセットにアクセスできます。

FlashBlade： FlashBlade は、これまでのストレージシステムとは異なります。ソフトウェアからハードウェアにおよぶ大規模な並列処理アーキテクチャにより、AI データのパイプライン全体で、どのような非構造化データのワークロードでも、かつてない優れたパフォーマンスを発揮します。

前述のように、この自律走行ソフトウェア企業のデータサイエンティストは、クラスタリングされた DGX-1 インスタンスだけを使用するのではなく、完全なトレーニングシミュレーションに先立って、複数のタイプのノードを利用して反復型の開発を行います。データの取り込みから、CPU のみのサーバーを使用したデータ処理、軽量 GPU サーバーを使用したモデルのデバッグ、そして DGX-1 サーバーを使用したトレーニングの実稼動まで、FlashBlade が高パフォーマンスのデータハブとなります。データサイエンティストが、利用可能なノード間でデータセットを移動する必要がないため、同じデータのコピーが複数作成されることはありません。

> さらに特筆すべきは、トレーニングの実行中に単一の SSD やコンポーネントの障害が発生しても、研究チームがデータやそれまでのトレーニング結果を失うことがないという点です。また、FlashBlade では、トレーニングデータセットのデータが増大した場合、ブレードを追加することで、帯域幅パフォーマンスを線形的に向上させることができます。このため、データセットやトレーニングアルゴリズムの規模や複雑さが増すなかで、将来のストレージ容量とパフォーマンスのニーズの正確な予測が可能となります。

NVIDIA DGX-1 クラスター： このシステムの目的の一つは、トレーニングの実行中、コンピューティングシステムに十分なトレーニングデータを常に供給し続けることです。膨大なデータセットの取り込みとトレーニングの実行において、FlashBlade が燃料だとすれば、DGX-1 はエンジンです。FlashBlade と同様に、既存のクラスターに DGX-1 ノードを追加すると、全体的な処理能力が定量的に向上します。このため、最初にニーズに基づいて適切な規模を決めてから、いずれかのコンポーネントが飽和状態に近づいたときに、ノードを追加して処理能力を向上させることができます。この事例のお客様は現在、DGX-1 システムを 4 ノード使用しています。コンピューティング容量を拡大する必要が生じても、FlashBlade のパフォーマンスが十分であるため、シームレスに拡張できます。

GPU を搭載するノードと CPU のみのノード： エンジニアリングプロジェクトでは通常、全体的な成果を達成するために、システム内の一部のコンポーネントを使用する時間が長くなります。データサイエンティストの作業もこの点では変わらず、DGX-1 クラスターに常時アクセスする必要はありません。作業は主に次の 3 つの要素に分類されます。

- トレーニングデータの照合、クリーニング、フィルタリング、処理、モデルトレーニングに適した形式への変換
- トレーニングデータの一部を使用したモデルのテストとデバッグ
- トレーニングデータの全体を使用したモデルのトレーニング

一般的に、最初の 2 つの作業は、CPU または GPU を搭載したコモディティサーバーで行います。FlashBlade をこれらのノードに共通のデータハブにすることで、3 つの作業の間でのデータ移動を最小限に抑え、生産性をさらに向上できます。

ベンチマークテストの環境

Pure Storage は、業界標準のベンチマークに基づく結果を得るために、多くの労力を投じて、社内で DGX-1 と FlashBlade の組み合わせのベンチマークテストを実施しました。導入を検討中のお客様には、DGX-1 と FlashBlade の最適な組み合わせを確認するために、可能な限り自社のデータとモデルでテストを行うことをお薦めします。下記に示すのはベンチマークテスト結果の一部です。詳細は、Pure Storage の担当者までお問い合わせください。

Pure Storage 社内のテストでは、P100 GPU 数 8、CUDA コア数 28,672、ホスト CPU 数 2（合計 80 コア）、システムメモリ 512 GB の DGX-1 サーバー1 台（Pascal アーキテクチャベースのシステム）を使用しました。DGX-1 の V100 ベースのバージョンではより優れた結果が得られると予想されます。トレーニングデータは 15 ブレードの FlashBlade に格納しました。テストはすべて NVIDIA 提供のコンテナ（nvcr.io/nvidia/tensorflow v17.07）を使用して実行しました。

トレーニングの入力には、ImageNet 2012 データセットを使用しました。これは深層学習の研究でもっとも一般的に使用されているデータセットの一つです。入力データでは、150 KB の JPG 画像が、より大きなファイル（それぞれ 135 MB 以内）にまとめられていました。ファイルシステムのキャッシュ（fsc）オプションは無効にしました。TensorFlow のパフォーマンス最適化のための一般的なベストプラクティスに従い、1 秒間に処理される画像数が一定になるまで各モデルをトレーニングしました。

DGX-1 の 10 Gbps リンクを両方活用するには、複数の TCP 接続を使用して入力データにアクセスする必要がありました。FlashBlade ファイルシステムを 2 つの異なるデータ VIP にマウントし、トレーニング中はファイル読み取りアクセスを両方のマウントポイントで多重化しました。

ベンチマークの結果

TensorFlow を使用して 4 種類のモデルのベンチマークテストを行った結果を次のグラフに示します。モデルごとに 2 つの結果があります。合成データによるテストは、GPU のスループットの評価が目的です。データをシステムの RAM で無作為に生成して GPU に供給しました。この構成では、FlashBlade に I/O を送信せず、実際の ImageNet データセットを使用しません。合成データによる結果は、実現可能な最高のトレーニングパフォーマンスを表します。これを、FlashBlade に格納されている実際の ImageNet データセットを使用した結果と比較しました。

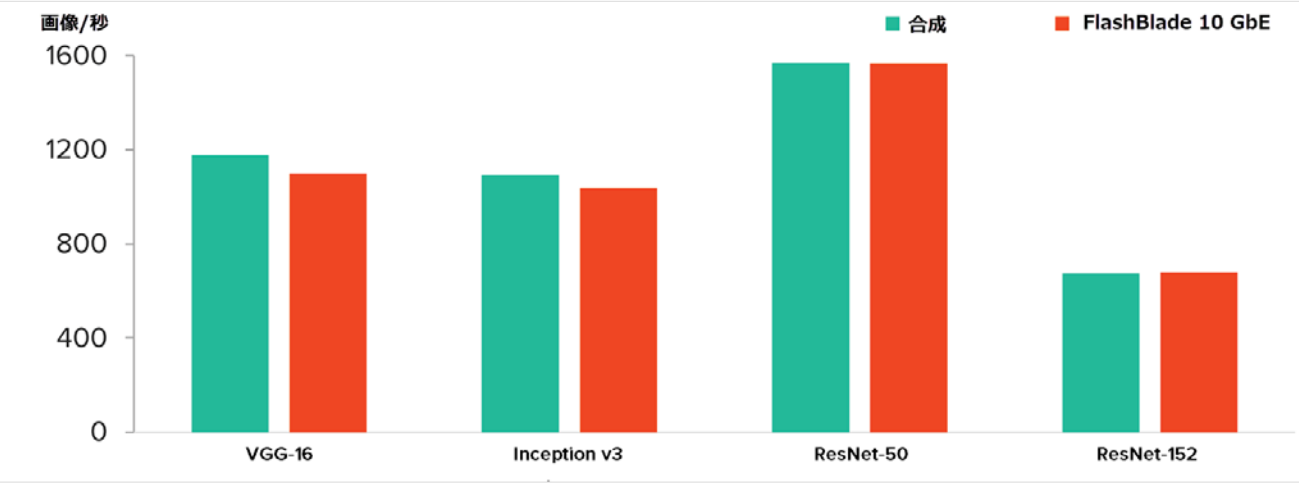

図 2：ベンチマークの結果

この結果から、FlashBlade は、深層学習のトレーニングについて、合成データをシステムの RAM から直接読み取る場合と同等のパフォーマンスを示すことがわかります。FlashBlade の構成では、ネットワークファブリック、NIC、PCIe 接続を通じてデータを外部ストレージから読み取っているということを考えると、これは驚くべき結果であるとも言えるでしょう。この結果から、FlashBlade は深層学習のトレーニングのために最高のパフォーマンスを提供できることがわかります。

ImageNet は広く使用されているベンチマークツールであるため、さまざまなフレームワークやソリューションを比較できます。しかし、お客様の現実的な要件が反映されないこともあります。たとえば、この事例のお客様は、ImageNet データセットの 150 KB よりも大幅に大きいファイルを使用しています。したがって、幅広いファイルサイズに対するパフォーマンスを理解することが重要です。共有ストレージによって達成可能なパフォーマンスは、入力ファイルのサイズから影響を受けます。

- 小さなファイル（50 KB）の場合、7 ブレードの FlashBlade システムで、ランダムデータの読み取りが約 5 GB/s であり、15 ブレードシステムでは 10 GB/s を超えることがあります。
- 大きなファイル（1 MB 以上）の場合、読み取りパフォーマンスは 1 ブレードあたり 1 GB/s をわずかに超え、たとえば 15 ブレードのシステムでは 15 GB/s になります。

GPU の処理能力は将来的に向上すると考えられます。たとえば、V100 は、トレーニング速度が P100 の 2 倍以上になる見込みです。このため、求められるストレージシステムのパフォーマンスも高くなります。一方で、GPU の処理能力が高まると、階層が多いニューラルネットワークを利用できるようになるため、入力バッチごとの計算量が増大します。ストレージシステムに求められるパフォーマンスの向上は、モデルの階層が多くなることで、ある程度は相殺されると考えられます。

まとめ

AI と深層学習の分野に関して一般的な共通認識があるとすれば、それは変化と進歩によって可能性の幅が広がり続け、新しい用途への適用が促進され、自動運転車などの分野で成熟が進むということでしょう。進化を続ける深層学習をサポートするには、AI データのパイプラインの基盤となるインフラストラクチャが、柔軟性に優れ、線形的な拡張性を持つことが必要です。さらに、管理が容易でなければなりません。この事例のお客様は、さまざまな選択肢を検討した結果、NVIDIA と Pure Storage の共同ソリューションを選択しました。最高水準のコンピューティングとストレージのプラットフォームを組み合わせたこのソリューションは、自律走行車というイノベーションを強力に支援しています。

ps_wp7p_dgx1-flashblade-anonymous_ltr_01