

はじめに

今日のネットワークでは、予測可能なパフォーマンスが求められ、デバイスを静的に割り当てていた従来のネットワーク以上に、アプリケーションの変化への柔軟な対応が必要です。かつての企業ネットワークが特定のアプリケーションに合わせた設計だったのに対し、データセンターの新しいクラウドは、膨大な数のアプリケーションに対応できる設計になっています。遅延の発生や伝送時の高レイテンシーが当然になっている現在のオーバーサブスクリプトなネットワークに比べると、当然ながら、発想は根本的に異なります。

アプリケーションに対応した予測可能なネットワーク・パフォーマンス

黎明期の Web（転送量 256KB）、電子メール（1MB）、ファイル転送（10MB）を前提としたかつてのクライアント/サーバ型の設計とは違い、データセンターの新しいクラウド・ネットワークでは、特定のパフォーマンスを確定的に得られるようにすることが必須です。現在のアプリケーションには、超低レイテンシーが求められる高頻度アルゴリズム・トレーディングや地震探査解析のように、具体的で定義が明確なものもあります。あるいは、ストレージと仮想マシン・イメージの大量移動や、Web 2.0 アプリケーション向けの大規模データ解析もあります。このようなデータセンターで求められるのは、ノンブロッキングで予測可能なパフォーマンスです。スイッチング・アーキテクチャで鍵となる要素の 1 つが、物理マシンと仮想マシンにまたがったアプリケーションのスケール・アウトに対応してパフォーマンスが均一化されているかどうかです。ノンブロッキング帯域幅と予測可能なレイテンシーが全ノードに等しく備わっていないてはなりません。最近のマルチコア・プロセッサでも、ネットワーク帯域幅は重要です。こうしたことから、一貫性のあるパフォーマンスとバランスのよさ（テラビット級のスケラビリティ、予測可能な低レイテンシー、ノンブロッキング・スループット、複数の 1/10GbE の処理と将来の 40/100GbE にも対応した高速接続）は、クラウド・ネットワーキング・アーキテクチャには欠かせません。

データセンターのスイッチング・アーキテクチャの基盤

クラウド向けのデータセンターのスイッチングでは、2 種類のスイッチング・アーキテクチャが利用されています。1 つはカットスルー・スイッチングです。ハイ・パフォーマンス・コンピューティング（HPC）クラスタのアプリケーションで超低レイテンシーを実現します。もう 1 つはストア・アンド・フォワード・スイッチングです。大容量メモリと仮想出力キュー（VOQ）のしくみによって、Web ベースのデータセンター・アプリケーションで必要となるバッファリングを実現します（図 1 を参照）。Arista 7000 ファミリーは、リーフ（葉）とスパイン（幹）の 2 層で構成される低レイテンシーの HPC ネットワーク設計に最適なスイッチです。Arista 7048 は、非対称な 1/10GbE 接続を利用する高密度な次世代データセンターに適合し、ストレージや Web ベース・アプリケーションを支えます。

HPC アルゴリズム・トレーディングでの低レイテンシー設計

- トポロジの一貫性
- 低レイテンシー
- カットスルー・アーキテクチャ
- 現時点で 10 ギガビット
- 将来の 40 および 100 ギガビットにも対応

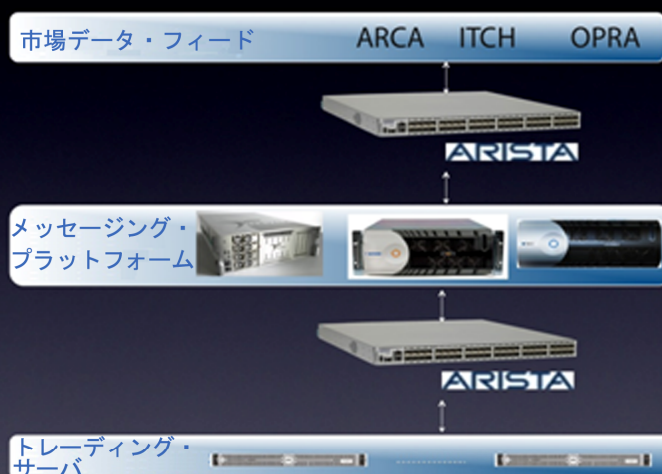


図 1a : 低レイテンシー・ネットワークでの HPC アプリケーション

大規模データセンター

- 演算速度およびストレージ・リンク速度が非対称
- 1GbE ノードと 10GbE ノードが混在
- 高い可用性とスケーラビリティ
 - 電力の冗長性
 - MLAG による冗長トポロジ
 - 信頼性のあるEOSスイッチ
- Arista 7500 シリーズのスパイン（幹）スイッチ
 - 最大 10 テラビットの容量
 - 低レイテンシー
 - ディープ・バッファ
 - VOQ アーキテクチャ

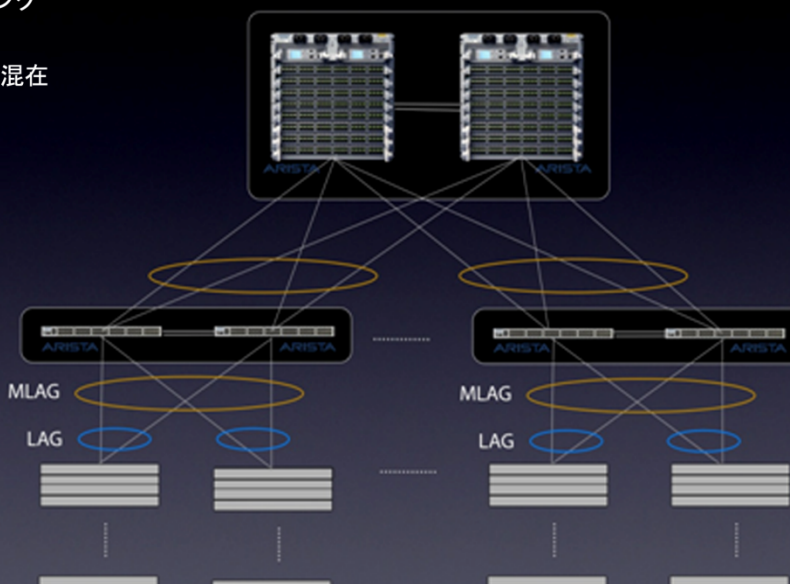


図 1b : 非対称な大規模データセンター設計

低レイテンシーのハイ・パフォーマンス・コンピューティング（HPC）クラスタ

リアルタイムの市場データ・フィードを利用して高頻度トレーディングを行う金融サービスなどの最新アプリケーションでは、カットスルーと共有メモリによるスイッチング技術で超低レイテンシーを実現する必要があります。そのレイテンシーは、マイクロ秒単位から、時には数百ナノ秒というレベルです。この方式による 10GbE アーキテクチャのメリットは、ポート・レベルでのバッファリングがほとんどなく、レイテンシーを最小化できることです。これによって、データ・フィード・ハンドラーやクライアント、アルゴリズム・トレーディング・アプリケーションに対して、ネットワーク経由で情報をほぼ瞬間的に伝送できます。カットスルー・スイッチングは、定型的で対称なトラフィック・パターンを持つリーフ・サーバに最適なアーキテクチャです。たとえば、HPC、地震探査解析、高頻度トレーディング・アプリケーションなどに向いています。その前提は、ネットワークの負荷が 50%未満で輻輳がなく、低レイテンシーの重要性が高いことです。カットスルー・スイッチングは数マイクロ秒の節約になり、特にラージ・フレームおよびジャンボ・フレームの packets で効果があります。迅速性が求められる環境では、この規模の低レイテンシーの実現は数百万ドルの節約につながることもあります。まさに「時は金なり」です。Arista 7100 シリーズは、受信した packets をメモリにバッファリングすることなく転送するため、超低レイテンシーの実現には理想的です。また、高速なマルチキャストにも対応し、キューイング遅延やシリアル化遅延を抑えます（図 2 を参照）。

超低レイテンシーのカットスルー・スイッチング

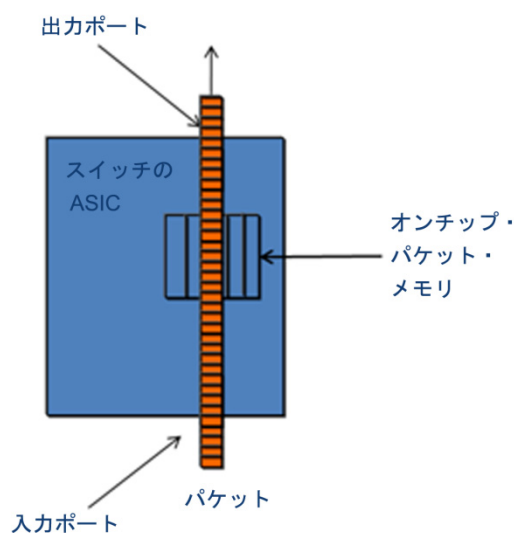


図 2 : HPC クラスタに適した低レイテンシーのカットスルー・スイッチング

新時代のデータセンターにふさわしい確定的なパフォーマンス

スパインの利用、リーフでの各種リソースへのシームレスなアクセス（ストレージ・リソース、演算処理リソース、アプリケーション・リソース）、データセンターのバックボーンなど、負荷の高いネットワークでは、スイッチング・ネットワークのスケール・アウトでも予測可能で均一なパフォーマンスを実現することが重要な要件の 1 つです。典型的な例としては、MapReduce のクラスタ、分散検索、データベース・クエリ・システムなど、大きなデータ・ブロックを必要とするアプリケーションが挙げられます。パフォーマンスの均一性は欠かせない要件です。レイテンシーは、多少遅めの 3~6 マイクロ秒でも問題ありません。それでも、レイテンシーが 20~100 マイクロ秒という従来のスイッチに比べれば、桁違いのパフォーマンスだからです。このような用途では、バッファを 1 ポートあたり数メガバイトの規模に増やして適切に設計したストア・アンド・フォワード型のスイッチング・アーキテクチャが最適です。現在のストア・アンド・フォワード・スイッチングのアーキテクチャでは、仮想出力キュー（VOQ）を利用して、Any to Any のトラフィックへの対応を強化しています。スイッチ・ファブリックの輻輳や「行頭ブロッキング」など、従来のスイッチをたびたび悩ませてきた問題は、VOQ で抑制できます。バッファリングの拡大と VOQ によって、アプリケーションの柔軟性は高まり、ネットワークの全体的動作が向上します。多数のデバイスから同じサーバへの集中でトラフィックが突然急増したときでも、大容量バッファによって輻輳は緩和されます。そんな急増が生じる例としては、ストライプ化されたストレージ・サーバ群からアプリケーション・サーバがデータを受け取り、すべての応答が同時に発生したときなどがあります。この場合、スイッチには、ストレージ・データを損失なしで一貫して保持しておける十分なバッファが必要です。大容量バッファリングは、10Gb ネットワークから 1Gb ネットワークへの非対称伝送の際にも、リンク速度の違いを吸収するために重要です。

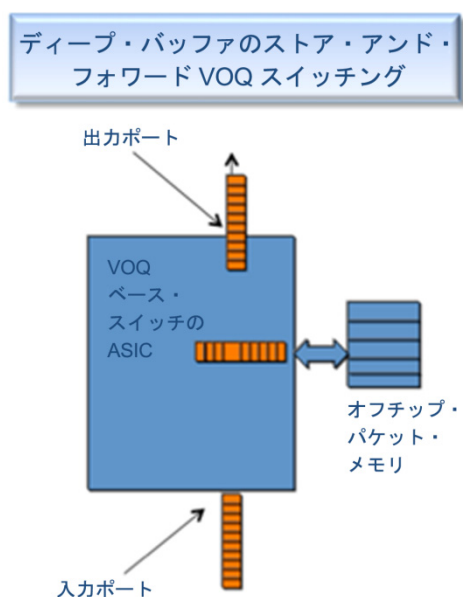


図 3 : 大規模データセンターでは大容量バッファと VOQ によるストア・アンド・フォワード・スイッチングで非対称のトラフィックへの対応が必要

クラウド・ネットワーキングの事例

クラウド・ネットワーキングは、Arista のリーフ・スイッチとスパイン・スイッチによる 2 層構成で設計できます（図 4 を参照）。ソーシャル・ネットワーキング・アプリケーション大手としておなじみの Facebook の事例を考えてみましょう。Facebook では、サーバ 3 万台でクラウド・ネットワークを構築し、memcache のクラスタあたりサーバ 800 台を配置して、発生する要求は 5 千万～1 億件、その間にアクセスするメモリは 28 テラバイトだそうです。5 ミリ秒のアクセス時間を要する従来のデータベース・アクセス手法ではなく、memcache アーキテクチャを使用することで、アクセス時間を 0.5 ミリ秒に抑えています。再伝送の遅延を抑えることと、持続的接続を増やすことも、パフォーマンス向上につながります。この環境では、アクセスを保証する大容量バッファがポイントです。大容量バッファ、先進の輻輳制御機構、VOQ アーキテクチャを搭載した Arista 7048 と 7508 シリーズのスイッチは、ストレージ、検索、データベース・クエリ、Web のトラフィックが大量に発生するアプリケーションに最適です。

次に、低レイテンシーの高頻度トレーディング（HFT）アプリケーションの事例について考えてみましょう。HFT とは、プログラムを使用して、タイミング、価格、注文数といったリアルタイムの条件に基づいた金融取引を自動的に実行するもので、ヘッジ・ファンド、年金基金、ミューチュアル・ファンドなどの機関投資家が広く利用しています。実行中のアプリケーションは、受け取った入力や情報に反応し、1 マイクロ秒足らずのうちに取引を処理します。まさに、目にも留まらぬ早業です。関連証券や市場取引についての国際間でのリアルタイムな情報交換には、金融処理のプロトコルが広く利用されています。今後は、マルチスレッド化するアプリケーションが増え、コンピューティング・クラスタ・ノード間での低レイテンシー接続がクラウドとスイッチング・アーキテクチャにとって要件となることが見込まれます。Arista 7000 ファミリのスイッチは、超低レイテンシーの HPC 設計に最適です。

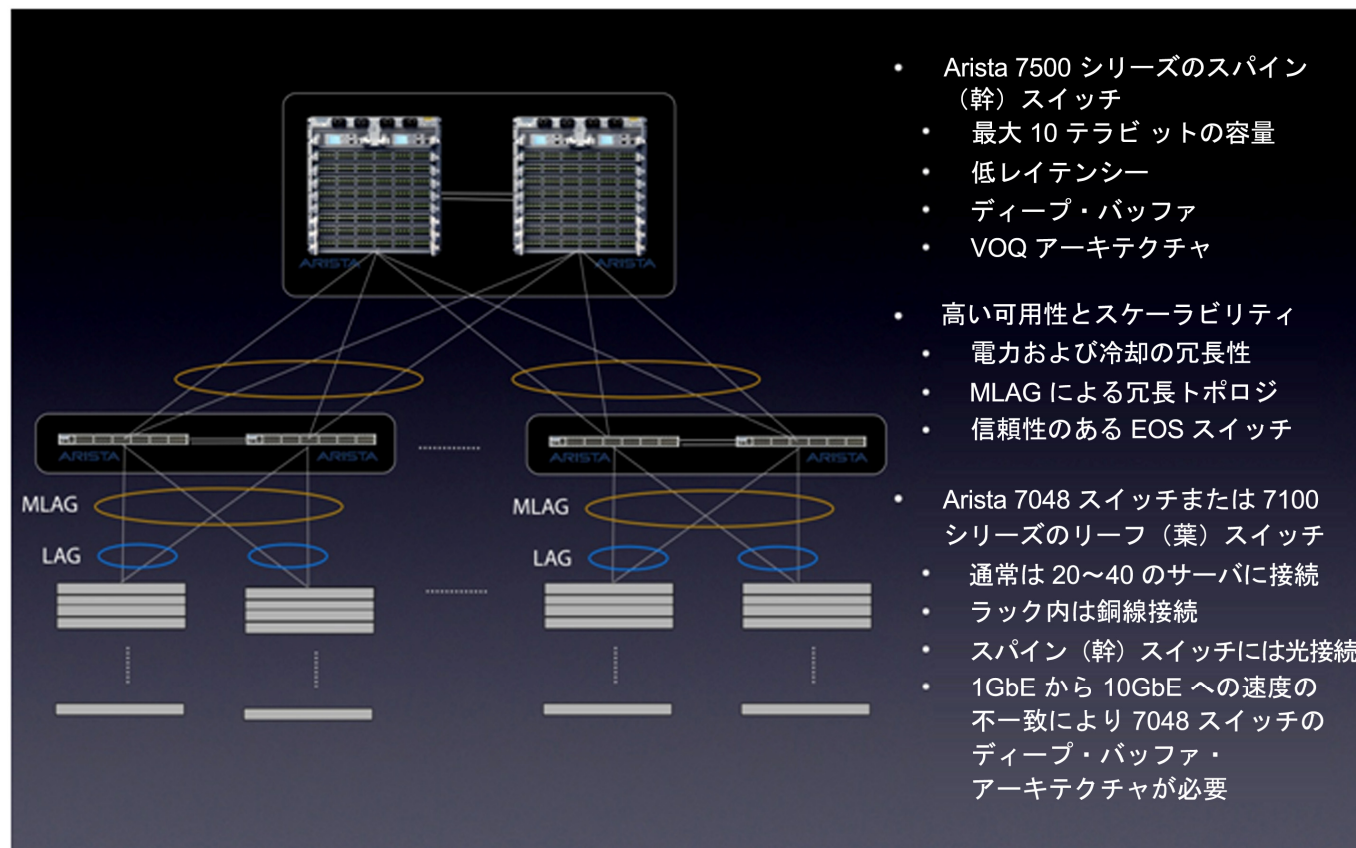


図 4 : Arista 7000 ファミリを基盤としたクラウド・ネットワーキング設計

まとめ

クラウド・アプリケーションの中には、データセンター向けの Arista の新しいスイッチング・アーキテクチャを生かせる「キラール・アプリ」が増えています。たとえば、次のようなアプリケーションです。

- 金融分野の高頻度トレーディング・アプリケーション
- ハイ・パフォーマンス・コンピューティング（HPC）
- クラスタを利用した演算処理アプリケーション
- ビデオ・オンデマンド
- ネットワーク・ストレージへのアクセス
- Web 解析、MapReduce、データベース、検索クエリ
- 仮想化

90 年代後半に設計されたネットワークは、静的なアプリケーションと電子メールへの対応が主な課題でした。しかし今日では、動的なアプリケーションやトラフィック・パターンが増え、アプリケーションのリアルタイム・アクセスに対応した新しいスイッチング手法が必要となっています。今後のクラウド・ネットワーキングでは、パフォーマンスの確実性、低レイテンシー、Any to Any の通信パターンといった面が整備されます。Arista のスイッチング・アーキテクチャと、豊富なラインナップの Arista 7000 ファミリは、クラウド・ネットワーキング・ソリューションの構築に最適です。