

はじめに：アプリケーションのレスポンス・タイムはクラウド環境の成否の鍵

仮想化と柔軟性を特徴とする次世代のクラウド・アーキテクチャに移行するデータセンターが増える中、耐障害性を備えた高性能なクラウド・ネットワーキングは今や必須となっており、常時稼働のアプリケーションをいつでも利用できるようにすることが求められています。レスポンス・タイムに対する要件は、この 10 年間で非常に厳しくなりました（図 1 を参照）。従来とは要件が 2~3 桁異なる場合もあるほどです。レスポンス・タイムはさまざまな業種でビジネスの成否にかかわる最重要要件となっています。たとえば、金融機関、教育/研究機関、Web 2.0 企業はもとより、リッチメディア企業（ビデオ編集、デジタル・アニメーション、放送関連企業など）、石油/ガス精製企業、医療関連企業など、その範囲はますます広がっています。

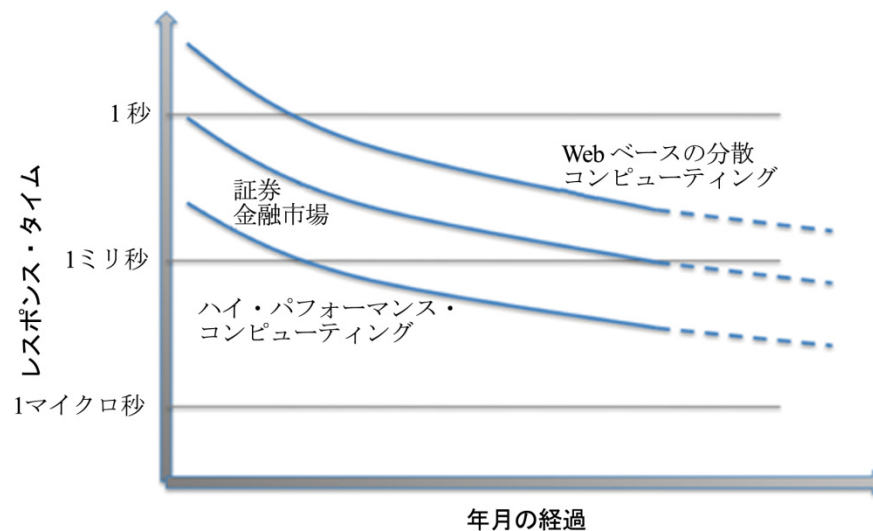


図 1：アプリケーションのレスポンス・タイムの推移

コマ何秒という単位のレスポンス・タイムが求められる現在では、ネットワークと演算処理の両方のレイテンシーを同時に短縮することが不可欠です。そのためには、アーキテクチャ面からの手法によって、多種多様なクラウド・アプリケーションのレスポンス・タイムを着実に向上することが必要となっています。以前は、クラウド環境にニッチ製品や独自技術を導入する事例も多くありました。しかし今では、広く利用されている標準ベースの技術を使って、低レイテンシーのインフラストラクチャを構築し、汎用的なクラウド環境を実現することがきわめて重要です。また、レイテンシーの向上をインフラ・レベルで実現するということは、3~5 年先のレイテンシーの要件に対応できる将来性も欠かせません。

低レイテンシー・クラウド・ネットワークの設計

厳しいレイテンシー要件があるワークロードで見られる特徴の 1 つが、分散コンピューティングのアーキテクチャを利用していることです。ワークロードのトランザクションごとに、処理ノード間で多数のやり取りが発生します（図 2 を参照）。場合によっては、数千のマシンやデータ・ストアが関与することもあるほどです。具体的な事例には、次のようなものがあります。

- アルゴリズム・トレーディングでは、あるマシンが送信した市場データのフィードを別のマシンが受信し、計算処理により取引の判断を下して、さらに別のマシンに対して取引の指示を行います。
- ハイ・パフォーマンス・コンピューティング（HPC）では、数千台のコモディティ・サーバをクラスタリングして"スーパー・コンピュータ"として機能させ、Lustre などのクラスタ・ファイル・システムに接続して複雑な数式処理やシミュレーション処理を実行します。
- Web ベースのクラウド・アプリケーションでは、動的にクラスタリングした数千台のサーバを利用して、複数のステージによる処理メカニズムを実装し、ペタバイト級のデータを処理しています。たとえば、Web 2.0 企業やインターネット検索企業のアプリケーションや、成長著しい Hadoop などのオープンソース・プロジェクトがこれにあたります。Hadoop は、分散コンピューティング（MapReduce）、分散ファイル・システム、分散データベースを実現するためのフレームワークで、スケーラビリティと信頼性を備えています。

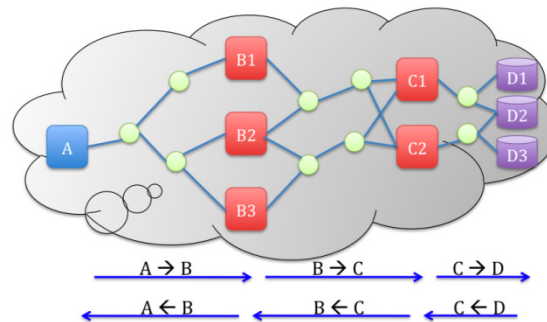


図 2：ネットワーク接続したマシンが単一のワークロードで行うやり取り：トランザクション

システム全体のレスポンス・タイムを向上するには、演算処理とネットワークの両方のレイテンシーをできるだけ小さくする必要があります。一般的なコンピューティング処理では、コモディティ・サーバを利用することが多く、レイテンシー最適化の自由度には限りがあります。基本的には、より高密度かつ高速なマルチコア CPU と、より大容量かつ高速なメモリ/キャッシュを搭載し、より高速な周辺機器接続が可能なサーバを選択します。レイテンシー短縮のもう 1 つの面が、クラウド・ネットワークです。こちらは、ファブリック全体の最適化によって、伝送でのレイテンシーを小さくする必要があります。具体的には、次の 4 つをキー・ポイントとして、ネットワーク・レイテンシーを最適化することが必要です。

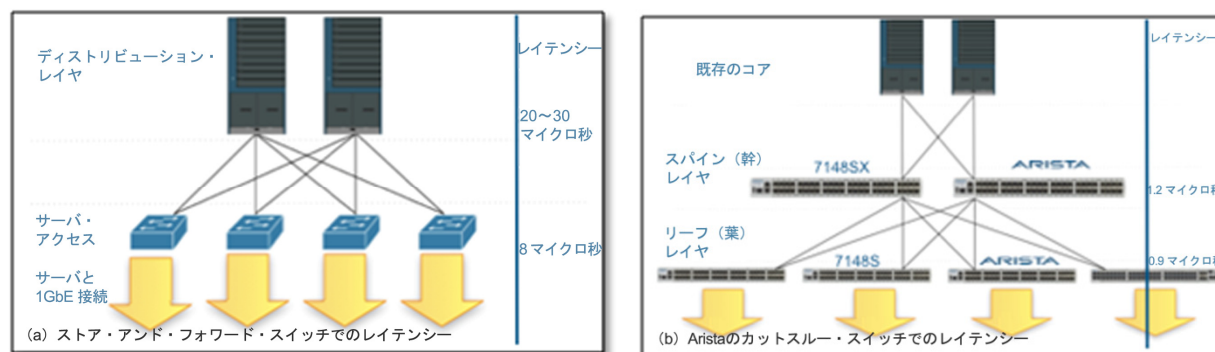
- 各ネットワーク・ノードのレイテンシーを小さくする。
- ステージ間で経由するネットワーク・ノードの数を減らす。
- ネットワーク輻輳をなくす。
- トランスポート・プロトコルのレイテンシーを小さくする。

Arista のメリット :

Arista Networks では、アーキテクチャ型的手法を利用して、低レイテンシーのクラウド・ネットワークを実現しています。ワイヤレートでノンブロッキングの超低レイテンシー 1Gb/10GbE スイッチと、Extensible Operating System (EOS) により構築した、耐障害性のあるクラウド・ネットワークを通じて、データ、マルチメディア、ストレージ、演算処理のトラフィックを伝送できます。Arista 7xxx シリーズのスイッチ・ファブリックを利用したクラウド・ネットワークには、次のようなメリットがあります。

- 前述の 4 つのキー・ポイントすべてでネットワーク・レイテンシーを小さくできます（詳細はこの後の説明を参照）。
- 現在のみならず、将来のレイテンシーの要件にも対応できる万全のネットワーク・インフラを構築できます。
- 豊富な利用実績があり、運用面もよく知られている標準イーサネットを採用できます。

ネットワーク・ノードのレイテンシー：金融処理や HPC など、超低レイテンシーが求められるワークロードでは、マイクロ秒単位でレイテンシーを切り詰めることが重要です。最適化された 1/10Gb イーサネット・スイッチを導入すると、1500 バイトのイーサネット・フレームの場合で、レイテンシーを 10 マイクロ秒以上短縮できます（9 キロバイトのジャンボ・フレームの場合はさらなる短縮が可能）。さらに、カットスルー・スイッチ・アーキテクチャなら、従来のストア・アンド・フォワードのアーキテクチャに比べて、数十マイクロ秒を追加で短縮できます。カットスルー・アーキテクチャを採用した Arista の高密度 10Gb スイッチでは、スイッチングのレイテンシーは 0.6~1.2 マイクロ秒の範囲となります（図 3 を参照）。



(a) 従来型のストア・アンド・フォワード設計	(b) Arista のカットスルー設計
ラック内のレイテンシー : 8 マイクロ秒	ラック内のレイテンシー : 0.6 マイクロ秒
ラック間のレイテンシー : 36 マイクロ秒	ラック間のレイテンシー : 2.4 マイクロ秒

図 3 : 従来のストア・アンド・フォワード・スイッチよりもレイテンシーに優れた
Arista のカットスルー・スイッチ

ネットワーク・ホップ数：データセンターのアーキテクチャは、サーバ・アクセス、ディストリビューション、コア・スイッチの各レイヤで構成される 3 レイヤ・アーキテクチャが一般的です。この場合、サーバ・ラック間でやり取りするパケットは、一度コア・レイヤを経由する必要があります。一方、リーフ（葉）スイッチとスパイン（幹）スイッチで構成される Arista の 2 レイヤ・トポロジ（図 4 を参照）では、パケットが経由するホップ数が少なくて済むため、レイテンシーを大幅に短縮できます。

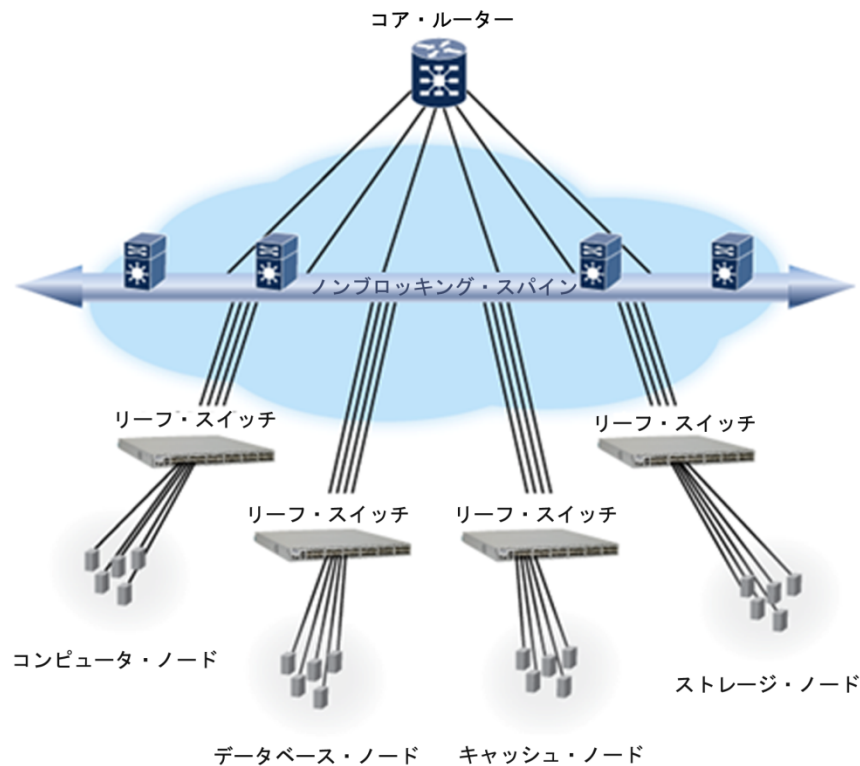


図 4：従来型の 3 レイヤ（アクセス、ディストリビューション、コア）トポロジに比べてレイテンシーを短縮できる Arista の 2 レイヤ（リーフとスパイン）トポロジ

ネットワーク輻輳：多くのアプリケーションでは、IP/イーサネット・パケットの伝送に TCP プロトコルを使用しています。信頼性とスループットを確保するために、TCP のステート・マシンでは、ウィンドウ制御により動的なネットワーク輻輳に対応しています。輻輳時にはウィンドウが小さくなるため、スループットが低下し、伝送全体のレイテンシーが増加します。従来型のデータセンター・アーキテクチャはオーバーサブスクリライブのため、輻輳とパケット損失につながりやすく、TCP のレイテンシーが大きくなっていました。Arista 7100 シリーズのスイッチはワイヤレート・スイッチ（10Gb イーサネットの各ポートがラインレート）です。このため、1:1 の確保された帯域でオーバーサブスクリライブのないクラウド・ネットワーク・アーキテクチャを実現できます（図 4 を参照）。輻輳のない大規模ファブリックの構築によって、TCP 通信時のウィンドウ・サイズが大きくなり、高スループットと低レイテンシーが可能となります。Web、ファイル/ストレージ、マルチメディア、データベースなど、さまざまなアプリケーションのワークロードによりマシン間で発生する膨大な数の TCP 通信にとって、TCP のスループットとレイテンシーの向上は多大なメリットをもたらします。

トランスポート・プロトコルのレイテンシー：一部の低レイテンシー環境（HPC、Lustre ファイル・システムの利用アプリケーション、アルゴリズム・トレーディングなど）では、TCP を完全に回避して、メッセージ・パッシング・インターフェイス（MPI）やリモート・ダイレクト・メモリ・アクセス（RDMA）によるマシン間の直接通信が望ましい場合があります。これまでは、超低レイテンシー、高スループット、ロスレスなどの特徴から、Infiniband などの独自の接続技術が利用されてきました。しかしユーザー企業からは、標準ベースのイーサネットを使って、同様の機能や優れた価格性能比を実現するよう求める声も上がっています。パブリック・クラウドまたはプライベート・クラウド全体を単一のイーサネット・ファブリックで管理できれば経済的だからです。ラインレートとカットスルーによる低レイテンシー・パフォーマンスが可能な Arista 7100 シリーズの 10Gb イーサネット・スイッチなら、単一の低レイテンシー・クラウド・ネットワークを実現できます。また、Arista のアーキテクチャは、レイヤ 2/3 マルチパスと輻輳制御に関する IEEE および IETF の現在および将来のイーサネット標準をサポートしており、輻輳のない大規模ドメインの展開が可能です。低レイテンシー、ワイヤレート、輻輳なしという Arista のクラウド・ネットワークを利用した大規模レイヤ 2 ドメインなら、アプリケーションのワークロードをクラウド内のどの場所にもすばやく移行できます。

低レイテンシー・アプリケーションの例：証券金融市場

超低レイテンシー設計の最も顕著な例が、アルゴリズム・トレーディングや市場データ・フィードです。この世界では、「トレーディング・アプリケーションでの 1 ミリ秒のアドバンテージは、大手証券会社にとって年1億ドルの価値がある¹」とされています。ミリ秒単位での時間短縮がきわめて重要であることから、証券会社では、トレーディング・システムを立会場に配置し、市場データ・フィードに直接アクセスすることで、距離に起因するレイテンシーの障壁をなくしています。最近の設計では、トレーディング・トランザクションで 1 ミリ秒のレスポンス・タイムを目指し、さらに数年以内にミリ秒未満の単位への突入も視野に入れています。しかし、速度を追求した結果、Infiniband などを利用した複雑な相互接続を導入してしまうケースもよくあります。その場合、クラウド規模での展開や管理は厄介です。

Arista では、Solace Systems などのパートナーとの協力のもと、ラインレート・イーサネット・スイッチ・ファブリックを利用して、31 マイクロ秒というレイテンシーを実現した包括的なアルゴリズム・トレーディング・ソリューションを提供しています（図 5（b）を参照）。

¹ 「Business at Light Speed」 Richard Martin、InformationWeek、2007 年 4 月
(<http://www.informationweek.com/news/infrastructure/showArticle.jhtml?articleID=199200297>)

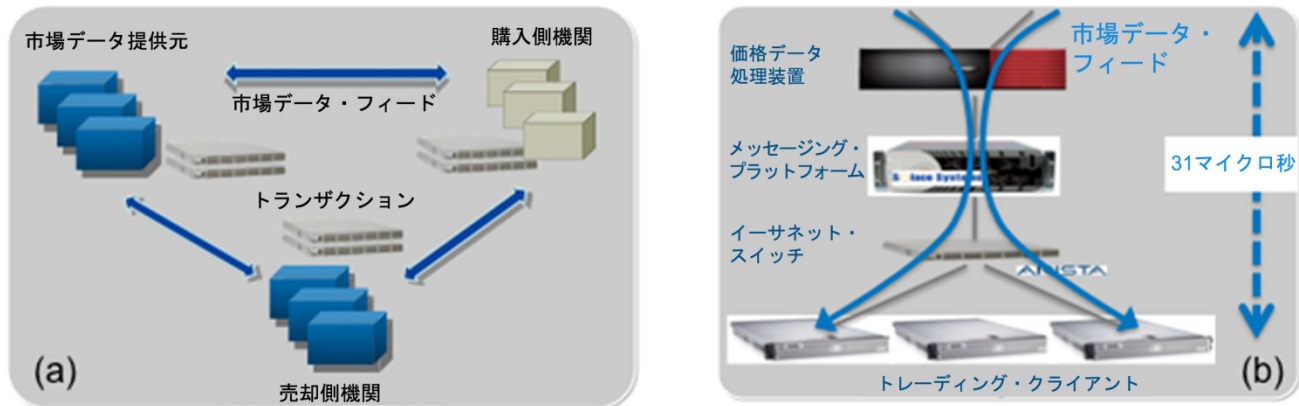


図5 (a) : 市場データ・フィードとアルゴリズム・トレーディング
 5 (b) : Arista と Solace Systems が提供する超低レイテンシー・
 アルゴリズム・トレーディング・ソリューション

まとめ

スケーラビリティと効率性を備えたクラウド・ネットワークへの移行が進む中、10Gb イーサネットのインターフェイスを採用して、共通のイーサネット・クラウドでデータ、ストレージ、演算情報を同時転送する事例が増えています。レイテンシーを最適化したアーキテクチャを基盤として超高速なレスポンス・タイムを実現している分野は、証券金融市場、ハイ・パフォーマンス・コンピューティング、Web ベース・コンピューティングなど、多岐にわたります。ユーザーにサービスを継続的に利用してもらえるかどうか、HPC クラスタのスケーラビリティを確保できるかどうか、市場が乱高下しているときに迅速なアルゴリズム・トレーディングで利益を上げられるかどうかは、レスポンス・タイムの短縮が決め手です。Arista Networks の 7xxx シリーズなら、クラウド・コンピューティングのレイテンシーとスケーラビリティに関するこのような現実的問題をスムーズに解決できます。