



最適な通信プロトコル選び

Mark Mahowald

Copyright 2009, 29West, Inc.

2009年5月1日

要約

この5年間、メッセージング市場は数々の革新の恩恵を受けてきました。10ギガビットイーサネット、Infiniband、カーネルバイパス、高速かつ安価なマルチコアサーバといった新技術が実装されるとともに、さまざまな組み合わせの検討が必要となっています。さらに、メッセージングアプリケーションの必要条件を考えなければなりません。配信を行うのはLANなのかWANなのか、目標とするメッセージレート、望ましいメッセージング方法（配信はレシーバ側主導型なのかソース側主導型なのか）、ファンアウト（同一データが必要なレシーバの数）、配信モデル（キューイング、パーシステンス、またはストリーミング）などがその例です。通信プロトコルの選択はパフォーマンスに多大な影響を与える可能性があります。本ホワイトペーパーでは、選択するプロトコルのトレードオフについての検討してみたいと思います。

メッセージングアーキテクチャ：検討材料

金槌を持つ者にとってはすべてが釘

メッセージング製品の中には、一つの通信プロトコルのみ（例えばTCPのみ、あるいはUDPマルチキャストのみ）をサポートするものがあります。その場合は、特定のプロトコルの制限にすべてがひきずられることとなります。理想には、開発しようとしているアプリケーションがどのようなものであっても最適な通信プロトコルが使用できるよう、多種多様なプロトコルをサポート可能なメッセージング製品を開発者が利用できることです。例えば、1台のマシン上でレイテンシーを最小限に抑えて通信する場合には共有メモリによる通信、ローレイテンシー・高ファンアウトの通信にはUDPマルチキャスト、ユニキャストアドレッシングが必要なアプリケーションにはTCPまたはUDPユニキャストといった具合です。また、単一のメッセージングAPIによりこれらすべての通信プロトコルを活用できること、基本的なすべての動きを100%モニタリング可能であること、望ましいネットワーク技術・トポロジーをいかなるものでも使用できることなども有益です。

レイテンシーはあらゆる個所で重要な要素

レイテンシーがメッセージングソリューションのどこで発生しているのかを知りたい時、必要なのはメッセージのデータパスを端から端まで追いかけることです。一般的なデータパスと全体的なレイテンシーは次の通りです。

- 1) 送信側から送り出されケーブルにたどり着くまでのレイテンシー。一般的にはカーネルやネットワークのスタックを含む、この段階でのレイテンシーはカーネルバイパスや DMA 技術により軽減することが可能です。
- 2) デーモン、サーバ、メッセージング用ハードウェアといったメッセージングの中間処理層によるレイテンシー。これらの多くは、処理途中でのデータコピーを実施します。
- 3) 受信側ネットワークスタックにおけるレイテンシー。送信側と同様、カーネルやスタックのバイパス技術による低減が可能です。

レイテンシーを最小限に抑えるためには、これらのレイテンシーをできるかぎり取り除くことが必要です。

最適なツール

このセクションでは、各通信プロトコルの特徴と、どのような場面で長所が発揮されるのかを細かく見ていきます。

共有メモリ (IPC) 通信

共有メモリ通信は、ネットワークレイテンシーやすべての中間処理ステップが取り除かれており、全体的なレイテンシーの点で他を圧倒しています。29West の IPC トランSPORTを使用すれば、開発者は同一マシン上のアプリケーション間でのレイテンシーを3マイクロ秒にまで低減でき、また、同じデータを同じ API を使い TCP または UDP マルチキャストを介して LAN または WAN ベースのレシーバに送ることができます。

古き良き TCP

TCP はメッセージングの通信プロトコルとしての利点を数多く備えていますが、レイテンシー、スケーラビリティ、公平性の点においては難点がいくつかあります。メッセージングに使用する場合にキーとなる長所・短所は以下の通りです。

- どこにでも届く： 幅広い方式の LAN/WAN やファイアウォールをシームレスに通すことが可能です。
- レシーバ主導型： TCP はレシーバ側でデータの流れると速度を落とすため、レシーバ側でオーバーランが発生しません。（ソース側での速度低下、レイテンシー増加などの問題が起こりえますが、レシーバ側でのオーバーランは発生しません。）この特

徴を持つため、TCP は実効スループットのテストに最適です。送信側のアウトプットを最大にしてどれほどのスループットが可能か調査でき、自己制御という特徴を持っています。29West が TCP での配信に基づいてスループット数を引用するのはこのためです。このようなテストは簡単に実施でき、TCP と一般的な低価格帯ハードウェアを使用したテストで 1 秒当たりのメッセージ数は 240 万を超えました。

- 「礼儀正しい」動作: ネットワークが輻輳状態になった時には、TCP は他のトラフィックが流れるよう一時的に送信を譲歩します（レイテンシーは当然増加します）。ネットワーク利用の「公平性」が目標であれば、この特徴は素晴らしいものです。しかし、トレーディングの環境では多くの場合、重要度の低いアプリケーションに対するファイル転送を優先させ、取引価格の送信が遅れることは望ましいことではありません。
- ファンアウト 1: TCP は、トラフィックのロスや順番通りのメッセージ配信を自己修復するポイント to ポイントの通信プロトコルです。一対一の通信に限定されているため、一対多のデータストリームはソース側ソフトウェアでコピーし何度も送信する、もしくは何らかの中間処理アプリケーションでこれを行う必要があります。どちらの場合もネットワークの外でコピーが行われるため、メッセージの複数回にわたる送付によりネットワークの負荷が高まり、コピーが行われる個所でレイテンシーが増加します。
- 公平性: UDP マルチキャストとは異なり、TCP のマルチプルデリバリーでは複数回の送信を順番通りに行わなければなりません。コピーされたデータを最後に受け取るレシーバは、他のレシーバに比べ当然不利な立場に立たされます。あるレシーバは最初に、別のレシーバは最後にメッセージを受け取ります。レシーバ間の公平性がビジネスの観点から求められる場合、一般的に TCP は最適な選択とは言えません。

信頼性のあるユニキャスト UDP 通信

TCP があらゆる OS に組み込まれているのに、敢えてユニキャスト UDP ベースの通信プロトコルを検討するのはなぜでしょうか。理由としては、「公平性」指向を回避するためや、より細かい制御を可能にするため、などが挙げられます。輻輳状態のネットワークリンクで、TCP トラフィックを飛び越せればいいのと思う場面は少なくありません。ポイントツーポイントの UDP プロトコルを使えば、これが可能となります。29West では、このプロトコルの使用は、ゲートウェイを使用するアプリケーションで WAN に対して競合トラフィックよりも優先順位を高くしたい場合に効果的であることを実際に見てきました。

ユニキャスト UDP のもう一つの利点は制御が可能になることです。送信順序や送信レートなどをより細かく制御したいのであれば、精緻に設計された UDP ユニキャスト通信プロトコルが最適と言えます。

信頼性のあるマルチキャスト UDP 通信

リライアブルマルチキャストは、一つのメッセージのコピーを多数のレシーバに送信する必要のあるハイパフォーマンス・メッセージングに最適な標準として長年にわたり使われ

ています。金融マーケットでは 1985 年から使用されており、高性能であるがゆえに適切な実装にはそれなりの手間が必要であるものの、レイテンシーおよびスループットの面での優位性は非常に優れています。マルチキャストの主要な長所および短所を以下に挙げます。

- **公平性：** すべてのレシーバが公平に受信でき、ネットワークやアプリケーションの負荷が最小であり、ローレイテンシーを実現可能な、真の意味での一対多データ配信。精密に設計されリライアブルマルチキャストシステムでは、スイッチでのデータコピーが回線速度で自動的に行われるため、すべてのレシーバがデータを同時に受信します。
- **ネットワーク・トラフィック：** ソース側や中間処理層が同一メッセージを何度も（場合によっては数百回）送信しないため、ネットワークの負荷が最小限となり、これによりネットワークの入力トラフィックが劇的に減少します。このため、帯域不足を理由としたネットワークインフラ更新の必要性が減り、実装コストの削減に役立ちます。
- **ネットワークでのコピー：** マルチキャストにつきもののネットワークでのコピーは、レイテンシーの原因となる中間処理層（デーモンやサーバなど）がないことを意味し、これによりスケーラビリティが大きく拡大するとともにレイテンシーが大幅に低くなります。

すべてのレシーバが高速で、また、全員が送信されるメッセージをすべて受け取りたい場合には、UDP マルチキャスト通信以外の選択肢は無いといっても過言ではありません。しかし、すべてのメッセージの受信を望むレシーバが全員ではない場合や、速度に対応できないレシーバが存在する場合はどうでしょうか。このようなケースでは、トレードオフを考える必要があります。

- **送信側主導型：** UDP マルチキャストは本来、ソース側の速度に合わせます。レシーバのペースが遅い場合にソース側の速度を落とす、プロトコルレベルでのバックプレッシャ機能（「フロー制御」）はありません。
- **信頼性：** UDP マルチキャストは組み込まれた信頼性を持たないため、ロスのないデータ配信が求められる場合には、何らかのリカバリプロトコルを準備しなければなりません。

この二点は、UDP マルチキャストの適切な活用に関し、リライアブルマルチキャスト通信プロトコルが欠かせない主な理由です。

スピードが非常に速い送信者と非常に遅いレシーバが存在するケースすべてに適した唯一のソリューションというものはありません。万能なメッセージング製品はさまざまなツールを提供してくれますが、すべてのケースで設計について意思決定を行わなければなりません。すべてのレシーバを最も遅いレシーバのペースに合わせるのか、あるいは遅いレシーバにデータをロスさせ、リセット後、ストリームに再参加してもらうのか、という問題です。マルチキャストのような高性能ツールでは、ロスメッセージのリカバリや遅いレシーバに対処する方法を確保しておく必要があります。

TCPによる送信でも同様の問題が発生しますが、TCPの場合には、開発者による制御の範囲や設計面での妥協の自由度が著しく低下することにご注意ください。TCPは、古いデータの破棄による対処を望むか望まないかとは無関係に、データ配信の速度低下（レイテンシー）を強行します。

マルチキャスト通信プロトコルの設計に関するもう一つの問題は、ロストメッセージのリカバリです。リライアブルマルチキャストプロトコルの設計が不十分な場合、再送信や否定応答メッセージ（NAK）によるストーム発生という問題が起こりえます。この問題は多くの場合、大量のパケットロスや関連したNAKのトラフィックにより劣化する傾向を持つ、従来のデーモンベース設計と関連しています。29Westが2003年に他社に先駆けて発表した製品のような最新のアプリケーション・to・アプリケーション設計では、詳細なモニタリング情報、送信データおよび再送信データの双方のレートの制限、NAKタイマーのインターバルにおける柔軟性、リカバリについての取り決めが提供されています。これにより、スパイクが引き起こすオーバーランや、度重なる再送信要求などで送信者側リソースを独占してしまうレシーバの存在によるネットワークの性能低下を防ぐことができます。

この点について豊富にある詳細情報は、本ホワイトペーパーの目的とは異なる範ちゅうに属するためここでは省略します。これまでの120件を超える世界中での29West製品の本番環境での使用ケースを見ますと、弊社製品は本ペーパーでご紹介した通信のすべてに対応しておりますが、大部分のお客様は安定性および効率という理由から、弊社のリライアブルマルチキャストをネットワークの重要個所に実装しています。

正しい選択とは何か

一つの通信プロトコルや配信モデルに頼りすぎでは、全体としてのニーズに応えることはできません。この問いに対する魅力的な答えは、既存インフラの有効に活用し、最高のパフォーマンス、安定性、制御の提供によって全体的なTCOを低く抑えることだと29Westは考えています。

正しい選択はアプリケーション開発者の手に委ねるのがベストであると我々は信じています。お客様のニーズとネットワーク設計に最適な通信を選択いただけるよう、29Westでは高性能かつフレキシブルなAPIですべての通信プロトコルをシームレスに提供しています。

弊社にとって、企業の強みとなる全社規模のメッセージングとは以下のことを意味します。

- 中間処理層をすべて取り除くことによりレイテンシーを最小限に抑える成熟した製品
- パワフルなマルチ通信、マルチプラットフォーム対応のAPI（C, .NET, Java）

- マルチパラダイムな配信モデル——ストリーミング（IPC および LAN/WAN）、パーシステンス、キューイング、TCP ファンアウトおよびキャッシュ
- ネットワークのあらゆる技術と形態を活用する能力
- 設計時に意図したシステムパフォーマンスが実現されていることを確認するための、ネットワークおよびアプリケーションレベルでの詳細なモニタリング

29West は、企業の強みとなる全社規模のメッセージングを提供します。お客様からのご質問にはどのようなものでもお答えいたしますので、ぜひお問い合わせください。

29West のメッセージング製品についての詳細は <http://www.29West.com/> をご覧いただくか、japan@29West.com 宛てに E メールでご連絡ください。弊社はシカゴ、ニューヨーク、ロンドン、東京にオフィスを構えており、お客様のニーズをうかがい、そのニーズに弊社がどのようにお応えできるのかをご説明できる機会を歓迎いたします。

29WEST 日本支社

〒102-0083 東京都千代田区麹町 3-5-2 BUREX 麹町

Tel: 03-6268-9803 Fax: 03-6268-9804

E-mail: japan@29West.com