

Comet による分散共有メモリの提案とその評価

水野 裕識、陣崎 明

新情報処理開発機構 並列分散システム富士通研究室
〒211-8588 川崎市中原区上小田中 4-1-1
E-mail: {mizuno,zinzin}@flab.fujitsu.co.jp

「超高速インターネット」の実現には、IP を高速に中継するルータが必要不可欠である。ルータの機能は、経路情報の管理とそれに基づくパケットの中継に大別される。現在インターネットのバックボーンでは5万経路を超える経路を管理する必要があり、ルータでは経路管理にCPUが追い付かなくなってパケットの中継処理が滞るといった大きな問題がある。この問題の解決に向けて、本稿では分散共有メモリによって構成した並列分散システムを用いて経路情報を共有する方式を提案する。その結果として1) 新経路の更新は各ルータで行えば、全体の更新処理量を抑えられること、2) 更新経路の情報は高速ネットワークにより速やかに共有可能になる2点について報告する。

インターネット、ルータ、分散共有メモリ、ルーティングプロトコル

Proposal of a distributed shared memory on “Comet” and the evaluation

Hironori Mizuno , Akira Jinzaki

Parallel and Distributed Systems Fujitsu Laboratory, RWCP
1-1,Kamikodanaka 4-Chome, Nakahara-ku, Kawasaki, 211-8588 Japan
E-mail: {mizuno,zinzin}@flab.fujitsu.co.jp

For the achievement of “Super-high-speed internet”, routers are required to relay IP at the high speed. The functions of router are divided roughly into the management of routing information and the relay of the packets based on it. It is necessary to treat the updates which exceeds 50,00 routes in the backbone of the internet now. This causes a big problem that the router which is overloaded with much routing information cannot relay the packets. To solve this problem, we propose a parallelly decentralized system that the update processing is decentralized and the bottleneck is not caused by a routing reference necessary for the relay. Consequently, we report on the two examination results, 1) if a new route information is updated at each router, the amount of the entire update processing can be suppressed, and 2) information on the update route can be promptly shared through a high-speed network.

internet,router, DSM, routing protocol

1. はじめに

次世代「超高速インターネット」の実現に向けて、我々は超高速ネットワーク(LAN:Gigabit Ethernet, WAN: ATM)を用いた並列分散システムの研究および応用システムとして超高速ルータ‘Comet’[1]を構築している。

ルータの処理は、経路情報の管理とそれに基づくパケットの中継に大別される。ルータはルーティングプロトコルを用いて数十秒ごとに相互に経路変更情報を交換することで、動的に経路情報を管理する。

現在インターネットバックボーンの相互接続点で観測[7]される経路数は50000(国内経路3000, AS数25)程度にまで増大している。経路交換を行うサーバは毎分数百経路以上の更新を行う必要があり、経路数に比例した更新処理がルータに対する負荷となっている。

またパケットの中継について、今後音声や動画がインターネットに流れるようになると、中継するルータのトラフィック量も爆発的に増加すると想定される。このようなトラフィックをインターネットに通過させるには、ルータに利用可能帯域、伝送遅延、ジッタのようなデータの品質を保証する必要がある。また特定経路に特定のパケットを通過させるポリシールーティングを実現するには、経路を選別する仕組みが必要になることから、経路情報は今後一段と複雑化する。このように各ルータでの経路情報の交換、それに伴う更新処理の負荷は高まる一方だと考えられる。

こうした問題のなかでルーティング情報の共有化とその負荷の軽減[11]に関して、我々は分散共有メモリによって構築する並列分散システムを用いて、経路情報を共有する方式を検討していることを述べる。

各ルータはATMにより接続された広域分散共有メモリに経路情報を置いて、経路が更新されるとその結果を共有メモリに反映する方式をとる。

これまで分散共有メモリは高速LANで実現されるが、インターネットの経路情報の管理が、数十秒単位で行われることを想定すると、遅延が大きい広域ネットワークを用いた分散共有メモリでも十分に耐えら

れる。

従来の分散モデルでは経路情報が送られてくると、更新処理を行うためCPUの負荷は大きくなる一方であった。我々のアプローチの場合、全体で1つの経路表を分散共有メモリで実現し、新規経路の書き換えを行う必要のあるルータのみが経路表を更新する方式を採用する。経路情報の更新処理そのものが並列に分散化されて処理のオーバーヘッドが抑えられる利点を持つ。

2章は、インターネットの経路制御の問題点を説明する。3章は分散共有メモリの実現方式を提案し、4章でインターネットで行われてきた経路情報のやりとりと比較して、本方式を評価する。5章では本方式の見通しについてまとめる。

2. インターネットの経路制御

インターネットの経路情報はルーティングプロトコルによって緩やかに伝搬する。分散共有メモリを用いて共有する場合の必要条件を述べる。

2.1. 経路の交換について

現在インターネットを流れるIPパケットは“hop-by-hop”の中継によって、最終的に受信側に届けられる。

IPが正しく中継されるには、次に転送すべきルータのアドレスが得られなければならない。図1に示すようにある組織の境界にあるルータは、外の隣接組織と経路を交換することで、そのアドレスを獲得する。これらを一覧にしたルーティングテーブルは、実際にIPパケットが転送される際のフォワーディングテーブルに反映される。

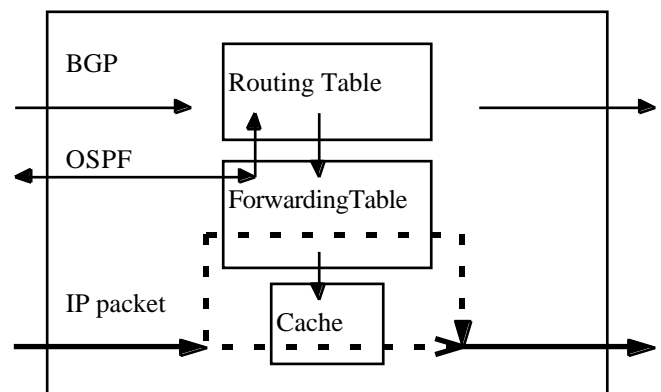


図1 インターネットルータの構成図

組織は自律システム(AS)と呼ばれ、AS 単位の経路交換には、BGP プロトコル[10]を、AS の内側にあるルータ間は OSPF[9]を用いるのが一般的である。

しかしながら、経路情報の増加とともにトポロジーの変化やマルチホームによる到達性の急激な変化[6]、さらにルータの設定ミスや内在するソフトバグなどと相互に関係して、経路制御の不安定さを生じるようになった。この不安定さには、経路更新が頻繁に繰り返される”Route Flap[5,8]”や、ネットワークの利用状況による週間単位での周期性も報告されている。米国インターネットの接続拠点の1つである MAE-East には、60 を超すサービスプロバイダーのトラフィックが流れる。このように大規模接続拠点には他の接続拠点と経路情報の交換だけを行う計算機がある。1996年1月から C.Laboitz[4]らがこの計算機を流れる BGP を監視したところ、1日辺り平均 125 経路の更新があること、また秒あたり 100 を超える集中的な経路の更新を観測している。

BGP は経路の差分のみ転送するので経路情報の転送量は大きくないが、経路を更新する頻度が多くなると CPU の負荷は大きくなる。特に分散モデルはローカルに経路表を持つことが前提なので、この経路表を更新する処理の負荷が問題となってきた。一方、変更が必要なルータがローカルの経路表を更新し、その結果のみを伝搬すれば更新処理そのものが分散化されるので全体の処理量は削減される。

2.2. 分散共有メモリ方式の適用について

分散共有メモリを広域環境で実現するには、高速化、高帯域化とは異なるパラメータとして距離による遅延が大きく関与する。そのため物理的に離れた計算機間に生じる遅延時間を考慮[2]しなければならない。

高速 LAN で実現される分散共有メモリ方式に比べて、広域網の分散共有メモリは比較的ルーズに接続される形態と見なせる。一方、広域網における分散共有メモリでインターネットの経路を仮想的に共有すると、共有すべき経路情報は BGP に比べて早く更新がなされる。インターネットの経路情報が、数十秒単位で交換されることから、従来の経路情報の交換から見ると経

路を一層タイトに共有する例と見なせる。

一般的に分散共有メモリ方式では、分散した各ノードは必要なページのコピーを持ち、読み込みにページフォルトが発生すると、ページのリクエストを発行して、ページの内容を自身ノードのメモリにコピーする。これはオンデマンド方式と呼ばれる。またページの書き込みは、一貫性を保つために他のノードのコピーを無効(Invalid)にしてから、書き込みオーナーとなってデータの書き込みを行う。

しかしながら、ページフォルトやデータを無効化するには、広域環境を前提にすると比較的大きな遅延を伴うことから一般的な分散共有メモリの方式をそのまま利用することはできない。

一般的な分散共有メモリ方式[3]を利用して通信を行なおうとすると、ローカルメモリに存在しないデータは毎回獲得しにいかなければならない。経路の参照に失敗すると全てのルータに経路情報の問い合わせ要求を行ない、広域網を経由して送られる経路情報を待たなければならなくなる。仮に何千何万のセッションを張るたびに、トポロジーの状態を獲得するため各経路を参照し、経路の収集を行なう方式は現実的ではない。

共有メモリ方式の並列処理を行う場合の同期と排他制御の必要性について述べる。

経路情報の共有に同期が必要かどうかについては、仮に書き込みが行われる前の経路を読み出し、そのルータにパケットが向かったとしても、パケットは到着保証がされないだけで、この動作は今のインターネットでも起こりえるため同期機構は重要でないと考えられる。

また一般に排他領域の実行は、スピンロックなどのセマフォ機構によってページ単位に行なわれる。ハードウェアでは Testset のように1クロックでセマフォを獲得する機構で実現し、データにアクセスした後にセマフォを解放する。経路情報の変更は短時間に同時に起こらないとすれば、排他制御も重要ではないと考えられる。

3. 広域網上の分散共有メモリ方式について

経路情報の共有化を実現するとローカルに経路の更新、管理を行うようになって、更新処理の負荷が抑制されることを説明する。

3.1. 分散共有メモリの適用方針

要求に応じて経路トポロジを集める要求を発行し、経路情報を収集する手法は、広域網の距離に比例する伝送遅延を考えると必ずしも適切ではない。一般的に自律システム AS がある組織を反映した管理形態であることから、ひとたび AS を登録すると登録データは頻繁に変更されることは少ない。この性質から経路情報は新規に登録する時やデータの変更時に共有メモリに書き込みを行なうと良い。このため隣接ノードから転送される経路の更新処理は 2.1 で述べたように全体で抑えられる。

BGPでは差分経路を update メッセージによって隣接ルータに伝搬する方式を行う。変更が生じたルータのみが共有メモリに書き込みを行い、各ルータに経路を放送することで、各ルータのメモリにその経路情報を反映することができる。こうして各ルータは全体とずれのない経路表をグローバルメモリとしてローカルに持つことができる。トポロジを表わす経路情報であれば、仮に全経路を集めても数 M から数十 M バイトあれば足りるので実現性は高い。

次にトポロジ以外の経路情報として例えばインタフェースの帯域や負荷の状態などの管理情報あるいは安全性の高いセキュアな経路を確保するためには、オンデマンドで経路情報を獲得する必要があることを述べる。

これらの一連の情報すべてを全ノードから放送すると、帯域が無駄に消費されてしまうので現実的でない。画像のようなある一定の帯域を定常的に占有するトラフィックをインターネットに流すことを想定する。送信元から受信までの通過経路はグローバルメモリに格納されている。多くの経路の中から到達可能な経路を得る。具体的には全経路トポロジが格納されているグローバルメモリに対して探索を行い、経路の候補を複数得る。そして、候補のなかから中継路にあるルータ

に対して帯域や負荷の情報を問い合わせる。そして最終的に複数経路から最適な経路を決める。

このような例を実現するためには、要求に応じて適切な情報を問い合わせることでそれらの情報を高速に獲得するオンデマンドの機構も必要になることがわかる。

3.2. 経路情報の共有に適した分散共有メモリ

3.1 で説明した経路情報を共有する分散共有メモリ方式について説明する。その特徴としてトポロジの経路情報は各ノードに共有メモリのコピーをグローバルメモリ (GlobalRT) に格納される点にある。図 2 に示すように、グローバルメモリは他のルータのグローバルメモリの書き換えが行われると一貫性を保つように広域ネットワーク経由で書き換わる。経路情報の読み出しは、他のグローバルメモリと基本的に一貫性を持っている自身のグローバルメモリから行う。経路情報の書き込みも同様にグローバルメモリに対して行なうが、書き込まれた経路情報は広域ネットワークに放送される。放送された経路情報を受けとる各ルータでは、一貫性を保つようにグローバルメモリに格納する。

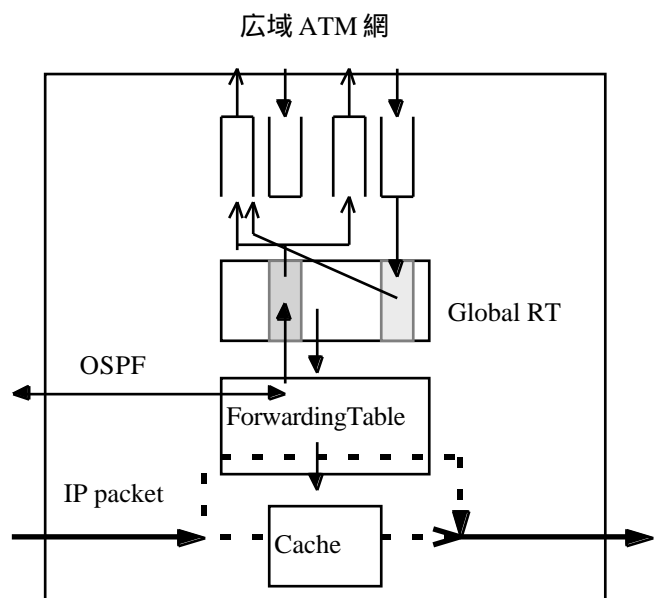


図 2 広域網分散共有メモリルータの構成

3.3. Comet による 1 つの実現手法

3.1 で説明したグローバルメモリによる経路情報を Comet を利用して仮想的に共有する 1 手法について説

明する。

グローバルメモリ上のデータを扱う単位はページといったあるメモリ領域（例、経路情報の組）ではなく、送信量を少なくするためにも1経路情報とする。経路情報の一部として宛先アドレスと次に向かう行き先アドレスの組を扱うとすると、Ipv4の場合8バイトである。

ATMで接続された2拠点間は、IP over ATMで決められたLLC/SNAPヘッダーを持つIPパケットを転送する。信頼性の低い経路に経路を転送することも考慮して、BGPプロトコルではTCP/IPを用いた。NTTのMegalinkサービスの場合、物理層でのセル廃棄率は 10^{-9} であることと、距離に応じた転送遅延、往復では2倍の遅延が生じることを考えると、転送経路情報について毎回ackを返すTCP/IPは必要ないと考えられる。そこでackを待つことなしに転送を可能にするため、経路情報はUDP/IPで転送する。

広域網に物理的に分散したルータはATMで接続されると仮定する。ルータの接続規模が少ない場合にはVCIによってフルメッシュの接続が可能である。現在広域接続する場合にはATMを選択することになるだろう。ATMセルはそのヘッダーにVPI(8ビット)とVCI(16ビット)と呼ぶ2つの識別子を持つ。Megalinkサービスの場合VPIは固定で、VCIによってパスを区別する。初期段階では接続数が少ないので、フルメッシュに相互接続する。

4. 評価検討

更新に必要な処理コストの削減と経路情報を転送に必要な時間を見積り、現在の経路共有の方法と比較する。

4.1. BGPとの比較

BGPによる分散モデルと経路を共有するモデルの違いは、経路情報を分散して更新するかどうかにある。BGPのような分散モデルでは各ルータは完全なる経路情報を管理しながら転送される経路の更新を随時行なう。更新する頻度が増加するにつれて、更新される負荷は全体で増大する。

これに対しn台のルータで更新処理を分担する本方式

では1台あたり $1/n$ の処理量になるため、全体では $O(n)$ の速度向上を見込める。

更新処理を分散すると各ルータは一貫した経路情報を共有しなければならない。BGPはAS-PATH属性を送りあうので、各ASのルータは部分経路から全体トポロジーを構築することができる。共有メモリ方式でも、全体経路のコピーをローカルに持つので、全体トポロジーの構築は可能である。

BGPは分散を前提にしたプロトコルで"Hop-by-hop"に経路を伝播する。更新される経路は様々な拠点を經由して伝播する可能性を持つので、更新間隔をある程度大きくするよう設計されている。また、隣接ノードがupしていることを知らせるために、"keep alive"メッセージは30秒に一度の間隔で送信される。このメッセージが180秒間来ない場合に、隣接ノードは停止した見なし経路表の隣接ノード情報を外す、あるいは経路はフラッシュされる。

今回提案した共有メモリ方式は経路はupしたという状態がイベント駆動方式に伝播するので、一定間隔という方式に基づいて情報情報を伝えるのではない。そのため、経路がupしたという情報が遠隔地のルータに転送され利用可能として登録されるまでの時間は、従来と比べると非常に短い時間で行われる。経路情報が複雑化すれば、up状態だけでなく隣接ノードの様々な情報もオンデマンドに収集する必要がある。

4.2. 広域での検討

実際に広域で経路情報を共有することを想定して、複数サイトを相互接続した。具体的には、広域実験が可能な環境を持つWIDE ProjectのATMバックボーンを利用して、富士通研究所から慶應大学SFCを經由の奈良先端大学の間でPVCを張った。

富士通研究所と慶應大学はその距離27kmを135Mbpsの帯域で接続し、慶應大学と奈良先端大学間は600km、45Mbpsの帯域で接続する。1経路の変更で転送されるデータ量を8バイトとすれば、富士通研で更新された1経路が奈良先端大学に転送されるまでの遅延時間は3.3[msec]になる。国内の全経路を3000とすると、24kバイトになるが45Mbpsの帯域で計算すれば

4.3[msec]かかり、その転送時間は7.6[msec]になる。

定常的に50経路の更新が各拠点で行われると仮定すると、各拠点のルータでは50経路を転送しつつ、100経路を受けて、合計150経路を同時に更新管理する必要がある。更新の負荷がある場合でも、奈良で更新された最終経路は最初の経路を送出してから数十ミリ以内の遅延で富士通研に到着し更新できる。

更新情報の放送について、富士通研で行われた経路の更新は、慶應大学、奈良先端大学に別のパスで送る方式を説明した。この場合富士通研からは2度転送しなければならない。仮に慶應大学のルータが受信経路情報を奈良先端大学に転送すれば、富士通研からは1度だけの転送で済む。接続ルータの増加に伴う経路の転送順序については今後検討していく。

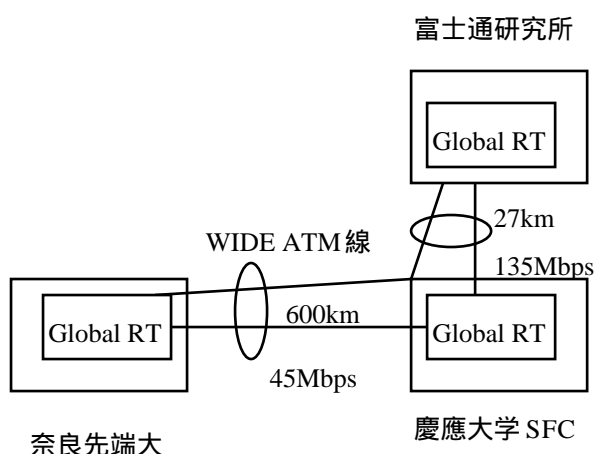


図3 WIDE 広域実験網を利用した仮想共有

5. まとめ

本稿ではインターネットの経路情報を広域で共有する方式を提案し、これまで1ルータで行ってきた経路の更新処理を分散して行なう方式について検討した。我々のアプローチに従えば経路情報の更新処理は各ルータで計算されるので、それぞれのルータに新経路が到達するたびに更新処理を行う必要がなくなって、その結果全体の更新処理量が低く抑えられることを確認した。

また、共有メモリ方式では更新した経路情報を広域に配置されたルータに伝搬する必要があるが、ATMの

ような高速ネットワークとその上を一方向に転送するプロトコルを併用することにより、BGPのような間欠的に伝搬する方法と比較しても遅延は低く抑えられることを確認した。さらに、今後経路情報が複雑化してもオンデマンドに経路情報を速やかに獲得できるためソースルーティングの実現に向けた経路選択が可能になる見通しを得た。

今後は方式の詳細検討を行いながら並列分散システム Comet の上に本方式を実装し、実際に動作させる予定である。

参考文献

- [1] 陣崎、中村、村井、"並列ネットワークサーバ Comet のアーキテクチャとその応用", SWoPP'98, CPSY-5, 1998年8月
- [2] 小口、相田、斎藤、"広域環境における分散共有メモリの検討", マルチメディア通信と分散処理 64-24, March, 1994
- [3] B.Nitzberg, V. Lo, "Distributed Shared Memory: A Survey of Issues and Algorithms", Computer, pp.52-60, 1991
- [4] C. Labovitz, G.Robert Malan, and F. Jahanian, "Internet Routing Instability", CSE-TR-332-97,
- [5] "BGP Route Flap Damping", <http://engr.ans.net/route-damp/>
- [6] 上水、"マルチホーム環境のための経路制御について-経路の不安定さに関する一考察", マルチメディア通信と分散処理 71-3, July, 1995
- [7] "Routing Arbiter web pages", <http://www.ra.net>.
- [8] "Internet Performance, Measurement & Analysis(IPMA)", <http://www.merit.edu/ipma/>
- [9] J. Moy, "OSPF Version 2", RFC 2178, Cascade Communications Corp, July 1997
- [10] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4(BGP-4)", RFC 1771, T.J. Watson Research Center, IBM Corp., cisco Systems, March 1995
- [11] X. Xiao, L. Ni, "Parallel Routing Table Computation for Scalable IP Routers", Proceeding of the IEEE International Workshop on Communication Architecture, and Applications for Network-based Parallel Computing, pp.144-158, Las Vegas, Feb 1998

