ハイパフォーマンス
コンピューティング 71－7
（１９９８．５．１５）

High Performnce Routing over Massively Parallel Routers

太田　昌孝、Manolo SOLA（東京工業大学）
藤川　賢治（京都大学）、児島　彰（広島市立大学）
福盛　秀雄、村岡　洋一（早稲田大学）

近年ギガビットルータという声も聞こえるが、すでに市販されている超並列計算機をルータとして利用することを考えると、数十万ポートの入出力ネットワークインタフェースをもち、内部のスループットが数十Ｔｂｐｓの超並列ルータは、簡単に構成できる。

ところが、超高速ルータのボトルネックは、スループットではなく、超高速インターフェースでのルーティングテーブルの検索にある。この部分を並列化するため、入力インタフェースにおいて、パケットのヘッダ情報に基づいたハッシュ関数によりパケットを複数のプロセッサにふりわけ、各プロセッサであらためてルーティングテーブルを検索する。同種のパケットは同じプロセッサで処理されるため、キャッシュの効きもよくなる。

High Performance Routing over Massively Parallel Routers

Masataka Ohta, Manolo SOLA (Tokyo Institute of Technology)
Kenji Fujikawa (Kyoto University), Akira Kojima (Horishima City University)
HIdeo Fukumori, Youichi Muraoka (Waseda University)

People are recently talking about Giga-bit routers. But, it is easy to construct a massively parallel router with hundreds of thousands of network interfaces and tens of Tera bps of throughput based on a massively parallel computers available today in the commercial market.

However, the performance bottleneck of the high speed routers is not at the throughput but at the routing table lookup rate at the high speed interfaces. To make the lookup parallel, incoming packets are distributed to multiple processors based on hash values of the destination information in packet headers. Routing table is accessed in each processor in parallel, then. As the packets to the same destination are processed on the same processor, locality of cache to access the routing table is improved.

１．はじめに

インターネットがこのまま発展すると、近い将来電話局からは電話交換機がなくなり、すべてルータに置き換えられる。このとき、ルータには、電話交換機程度、１０万ポートくらいのインターフェースが欲しい。また、画像伝送などの応用を考えると、各家庭への速度は 100Mbps 程度は欲しい。100Mbps のポートを１０万個もつために必要な交換能力のスループットは、1Tbps となる。

末端ルータが 100Mbps のポートを１０万個程度もち利用者の多くが大域的な通信を行うとすると、末端ルータと幹線ルータとの間は 1Tbps のインタフェースが必要であり、幹線ルータ間には 10Tbps 程度のスループットの交換能力が必要である。

また、インターネットでは一対多、多対多の通信のためのマルチキャストや、品質保証を伴った通信（フロー）のためのＱｏＳルーティングが、ルータの基本的機能として必要となる。これまではこれらの機能は類似したマルチキャスト要求やフローを一つにまとめることにより効率的に処理できると考えられていた。しかし、マルチキャストにおいてはまとめることは不可能であることが証明されたし［１］、ＱｏＳルーティングにおいても、おそらく不可能であろうと思われる。

つまり、ルータには、これらの高級な要求を満たすため、ルータ上のマルチキャストやフローの数に比例したある程度の計算能力が必要である。

そこで、超並列計算機の技術を使って超並列ルータを構成することを考える。

すでに市販されている SR2201 では、4096 プロセッサを持ち、計算能力としては十分なものがある。また、各プロセッサは、相互に 300MB ／秒(=2.4Gbps)で衝突なしに同時に通信でき、合計 10Tbps 程度のスループットの交換能力をもつ。

将来の超高速ルータに対する要求は、既に市販されている超並列計算機を超並列ルータとして利用するだけで、一応満たすことができる。超並列ルータのスループットは、超並列計算機の相互結合網の線の数に比例して、いくらでも増やすことができる。計算能力も要素計算機の数(P)に比例して増やすことができ、このとき相互結合網の複雑性は線形ではないものの O(PlogP)程度であり、特に問題にはならない。

ところが、超並列ルータ網の幹線ルータに必要な超高速インターフェースでは、個々のパケットの処理という別の問題が生じ、性能上昇のボトルネックとなる。

例えば 10Tbps の速度のルータで、平均パケット長を 500 バイトとすると、必要なパケット処理能力は、2.5Gpps である。つまり、0.4ns に一つのパケットが処理できないといけない。

ルータは、個々のパケットの処理に際して、ルーティングテーブルを検索し、その送出先を決める。しかし、テーブルの検索はメモリー読み出しを伴う操作であり、あまり高速には行えない。なお、高速インターネットルータは難しいとして高速 ATM スイッチに期待をかける向きもあるかと思うが、ATM を利用しても入力インターフェースで大きなルーティングテーブルの検索（と後述する書き込み）が必要であることは同じで、逆に ATM ではセル単位でテーブルの検索が必要となり、事情が１０倍程度悪化するだけである。

さらに、品質保証を行う場合にはテーブルの書き込みも必要になる。品質保証のためには、あらかじめ必要な品質を申告して（シグナリング［２］）一部のデータを優先的に扱う必要があるが、誰もが品質保証を要求したのでは、優先的扱いの意味がなくなる。そこで、品質保証のためには、申告された品質要求に比例して課金するなどして本当に品質保証が必要なデータのみ品質保証するための仕

組みが必要となる。このとき、ルータが実際に流れたデータの量を監視して申告された以上の量が流れてきていないかなどの統計をとらないと、課金などの処置が正しく行えない。そこで、パケットの到着の際には個別のフローに対して統計をとる必要があり、テーブルの書き込みが必要となるわけである。

また、ルーティングテーブルの大きさは、現在のインターネットプロトコルである IPv4 では数万エントリ必要であり、ルーティングテーブルの検査においてキャッシュはあまり効かない。次世代のインターネットプロトコルとして開発されている IPv6 では、アドレス割り当てを工夫することによって、似た行き先のルーティング情報を階層的にまとめ、ルーティングテーブルの大きさを抑えることができると期待されている。しかしながら、これが可能なのは、品質保証を考えない（ベストエフォート）一対一（ユニキャスト）通信においてのみである。

マルチキャストや品質保証を考えると、ルーティングテーブル検索のワーキングセットは、マルチキャストやフローの数に比例した大きなものとなり、素直に実現したのではキャッシュは効かない。

ところが、キャッシュなしで大きなメモリを読み書きする場合の遅れは数百 ns 程度であり、0.4ns という目標に比べて千倍程度遅い。そこで、なんらかの手段によりルーティングテーブルの検索自体を並列化する必要がある。

2．インターフェースごとの並列ルーティング

シングルプロセッサルータでは、入力パケットをプロセッサにあつめ、そのプロセッサ並列ルーティングのための一つの手法は、ルータの各インターフェースに専用のプロセッサをそれぞれ張り付けることである。各プロセッサは並列にルーティングテーブルの検索を行えばよい。個々のインターフェースから入ってくるパケットは独立に処理できるため、この並列化はいちおう有効である。しかし、幹線ルータや、末端ルータの幹線側インターフェースのようにインターフェースの速度自体が高速である場合、この方式ではそのための付加を並列化して分散することはできない。

また、ルーティングテーブルは各プロセッサでそれぞれ持たなければならず、キャッシュの効きも期待できない。

3．ハッシュによる並列ルーティング

本節では、品質保証を伴わないベストエフォート通信のルーティングテーブル検索の並列化について論じる。

前節での考察により、ルーティングの高速化のためには、ルーティングテーブルの検索以前に、並列化をはからなければならないことがわかる。

そこで、入力インターフェースにおいてはテーブル検索を行わずにパケットの分散だけを行い、分散した後であらためてルーティングテーブルを検索すればよい。

このような方式では、パケットをインターフェースから個別プロセッサに分散し、個別プロセッサからさらに出力インターフェースに分散するという二段構えのデータ転送が必要になり、一見非効率的である。

しかし、データ転送量はたかだか２倍であり、またインターフェースからプロセッサまでの配線は一本道であり一般の超並列計算機のプロセッサ相互接続網のような複雑な構成は必要ないので、実はそれほどの手間ではない。

残る問題は、パケットをどのようにプロセッサに分散するかである。

まず考えられるのは、パケットをラウンドロビン、あるいは、ランダムに各プロセッサにわりあてることである。

この時、ランダムなばらまきはプロセッサの負荷にばらつきが生じ、場合によっては処理しきれないパケットが落とされる可能性がある。実はラウンドロビン方式を用いても事情は同じで、たまたま出力が特定のポートに集中した場合、そこで出力待ちが生じ、バッファが必要となり、バッファあふれにともなうパケット落ちの可能性がある。しかし、インターネットにおいてベストエフォートサービスでの高負荷にともなうパケット落ちは当然のことであり、全く問題にならない。

ラウンドロビンあるいはランダムなパケット配布の問題の一つは、各プロセッサがそれぞれ大きなルーティングテーブルを持たなければいけないことである。

そこで、同種のパケットを同じプロセッサで処理するため、インターフェースからプロセッサへの分散を、パケットの行き先を決定する部分に対するハッシュにより行うことを考える。

パケットの行き先は、ベストエフォート通信では行き先アドレスだけで決まる。そこで、同じ行き先アドレスを持つパケットは、常に同じプロセッサで処理されるようにすれば、各プロセッサはルーティングテーブルのうち自分が担当する部分だけを持てばよい。

とはいえ、ユニキャスト通信では、ルーティングテーブルはある程度階層化されているがハッシュをほどこすと階層化の構造はまったく反映されなくなるため、この方式でも各プロセッサの持つユニキャストルーティングテーブルの大きさはあまり小さくならない。ユニキャストルーティングテーブルの本格的な縮小は、IPv6 の登場とその階層的アドレス割り当てを待つしかなさそうである。

しかし、もともと階層化によるとりまとめが不可能なマルチキャストルーティングテーブルに対しては、各プロセッサはそのテーブルのうち自分にかかわる部分だけを持てばいいので、各プロセッサのルーティングテーブルの大きさはプロセッサの数に比例して小さくなり、キャッシュなどのメカニズムも有効に作動する。

4．品質保証を考慮した並列ルーティング

前節のルーティングテーブルを検索前にパケットを分配してしまうという考え方は、品質保証を考慮したルーティングの並列化にも有効であるが、プロセッサの割り当てに工夫が必要である。

ランダムあるいはラウンドロビン方式でパケットを分配すると、同じフローに属するパケットが多数のプロセッサで処理されることになる。ベストエフォートの場合は、テーブルはほとんど更新されず通常は読み出すだけですむが、品質保証を行う場合は、同じフローに所属するパケットの量を監視するためのテーブルへの書き込みが各プロセッサからパケット単位で生じ、結局共有ルーティングテーブルの更新が性能のボトルネックとなってしまう。

いっぽう、パケットの行き先アドレスをハッシュした値に基づいてパケットをプロセッサに割り振れば、同じフローの属するパケットは常に同じプロセッサで処理されるため、品質保証の監視のためのテーブルの書き込みが必要となっても、それは各プロセッサでのローカルな処理であり、共有テーブルの更新に伴うボトルネックは生じない。

ただ、品質保証を行う場合、ベストエフォートと異なり勝手にパケットを落としていいわけでないので、プロセッサ間の負荷分散が重要となる。ハッシュの結果がたまたま同じプロセッサとなる行き先へ品質保証要求が殺到すると、そのプロセッサの処理能力を越えてしまい、品質保証が行えない。

この問題の解決策としては、品質保証が必要なフローに対しては、あらかじめフローのプロセッサへの割り当ての負荷が分散するようにしておけばよい。幸い、品質保証を行う際にはあらかじめ必要な品質を申告するわけであり、その際にはいろいろな処理をする時間の余裕がある。

そこで、プロセッサの負荷を均等になるようにフローを割り当てるプロセッサをきめればよい。しかしながら、行き先アドレスに基づくハッシュを使ったのでは、フローがどのプロセッサにわりあてられるかを制御できない。そこで、品質保証が必要なフローに限っては、ラベルスイッチングの手法［３］を導入し、前段のルータとあらかじめ打ち合わせておいて、フローについては特殊なリンク層のラベルをつけておくことにする。

そして、フローに属するパケットについては、行き先アドレスではなく、ラベルの値に基づき処理するプロセッサを決めればよい。ラベルの値は、通常のベストエフォートパケットとの区別が容易につくように割り当てておく必要がある。

５．採算性

10Tbps の相互結合網をもつ超並列計算機はそこそこに高価なものである。しかし、帯域が一人 100Mbps とすると、１０万人で共有することができる。そこで、現在の電話料金や CATV に払っている程度の月額一万円を各加入者が負担すれば、年間１２０億円、３年で償却するとして３６０億円であり、末端、幹線を含めて平均１０台の超並列ルータが必要だとしても一台あたり３６億円かけることができる。

まして、当面の応用がテレビ放送程度の画質の画像伝送でよければ、帯域は 6Mbps もあればいいので、１６０万人で共有でき、月額千円でもルーター台６０億円かけることができ、採算性の問題はないであろう。

６．終わりに

インターネットの利用者の増加、インターネット上での画像配信、ADSL やケーブルモデムによるインターネットアクセス網の高速化などにより、今後はますます高速なインターネットバックボーンが必要となる。そこで、高速ＡＴＭスイッチに期待する動きもあるが、ルーティングテーブルの検索がボトルネックとなる以上、ＡＴＭは問題を悪化させるだけにすぎない。しかし、市販されている程度の性能の超並列計算機をもとに、パケット単位でハッシュとラベルスイッチングの技法を活用すれば、ベストエフォート通信、品質保証通信ともに高いパフォーマンスで行える超並列ルータが構成できることが示された。

参考文献

［１］M. Sola, M. Ohta, T. Maeno, "Scalability of Internet Multicast Protocols", http://web.jet.es/sola/inet98.html, to be presented in INET'98.

［２］R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Funcional Specification", RFC 2205, Sept. 1997.

［３］M. Ohta, "Conventional IP over ATM", Expired Internet Draft, ftp://ftp.jain.ad.jp/pub/ids/draft-ohta-ip-over-atm-00.txt, March 1994.