2007－HPC－111 (31)
2007／8／2

超並列計算機向け多段光超高速内部接続

太田 昌孝†

†東京工業大学大学院情報理工学研究科 〒152-8552 東京都目黒区大岡山 2-12-1

E-mail: † mohta@necom830.hpcl.titech.ac.jp

概要 超並列計算機の相互接続網を回線交換網ではなくパケット網とし、光パケットスイッチの使用について見当した。光パケット網は回線交換網と異なり交換パターンへの制約はないが、光バッファが必要となる。実用的な光バッファは、光パケット多重の考えにより 1Tbps でエンコードした光パケットでは光ファイバ遅延線により容易に実現できる。現在市販の部品を組み合わせて実現した 1Tbps*8 ポート光パケットスイッチは 600W程度の消費電力で実現可能であり、それを 5 段、2560 台使用した 4096 ポートスイッチの速度は、名目 4 Pbps、実効 1.8Pbps、消費電力は 1.54MW 程度となった。

Multi-Stage Optical Ultra High Speed Interconnection for Massively Parallel Computers

Masataka OHTA†

†Graduate School of Information Science and Engineering, Tokyo Institute of Technology

2-12-1 O-okayama, Meguro-ku, Tokyo, 152-8552 Japan.

E-mail: †mohta@necom830.hpcl.titech.ac.jp

Abstract Packet, not circuit, switching for interconnection networks of massively parallel computers is discussed using optical packet switches. Unlike circuit switched networks, optical packet switched networks do not have restriction on exchange pattern but require optical buffers. Practical optical buffers can be constructed easily with optical fiber delay lines with optical packet multiplexing to encode packets at 1Tbps. A 1Tbps*8 port optical switch built from commercially available parts consumes 600W, 2560 of which can be combined in five stages to form a 4096 port switch with net speed of about 4Pbps, effective speed of about 1.8Pbps and power consumption of about 1.54MW.

1. はじめに

通信方式としての回線交換に対するパケット交換の優位性は明らかであり、通信網ではすでに回線交換による SONET/SDH 等はほぼ消えパケット交換によるインターネットが全盛である。しかし、超並列計算機の内部接続ではあいかわらず回線交換が幅をきかせている。

そこで、テラビット級の速度と低消費電力が望める光パケットスイッチをもとに、ペタビット級相互結合パケット網を構成し、性能や消費電力を評価した。

2. 回線交換と超並列計算機における問題点

回線交換の大きな欠点は、コネクションオリエンテッドであることで、接続トポロジの変更に多大な手間がかかる。また、一般に、通信帯域は回線により決まり固定的であり、その大域が常に保証はされるが、通信量が変動すると帯域に無駄が生じる。

また、回線数が増えるとスイッチが複雑化するため、回線の数もあまり多くできない。昨今の超並列計算機ではノード数が極めて多いため、ノードあたり数回線しか用意できず(回線数をノード数の定数倍にしておかないと、回線数は爆発的にふえるため)、個々のノードは他の少数のノードとしか通信できず、多次元での領域分割を行うと個々のノードが同時に利用できる回線数よりそのノードが通信すべき相手ノードの数のほうが多くなり、途中で回線の構成変更が必要になる。回線交換において全ての回線間の組み合わせを実現する(Non-blocking)ための技術としては、クロスバスイッチ、Clos スイッチ、Benes スイッチ等の技術があるが、それぞれ一長一短がある。回線数を N とすると Benes スイッチの回路規模は O(NlogN)であるが、回線間の任意の組み合わせを実現する場合、既存の回線の経路を切断して

変更する必要があり(Rearrangable non-blocking)、その際、経路が変更された回線を流れているデータは失われるため、回線の切り替えを全ノード一斉に行う場合以外には、超並列計算機での実用性に乏しい。Clos スイッチは、3段のクロスバスイッチにより、既存の回線の経路を変更することなく全ての回線間の組み合わせを O(N^1.5)で実現できる（Non-Rearrangable non-blocking）が、定数項がかなり大きいし、そもそもオーダがあまり小さくない。クロスバスイッチは単刀直入に Non-Rearrangable non-blocking を実現できるが、規模が O(N^2)で、回線数が多いと実用性はない。光が平面上や空間中で交差できることを利用しMEMSにより作成した光クロスバの規模は一応 O(N)だが、精度の限界からあまり N は大きくできない。

しかもそもそも、Non-blocking という性質は、電話網のように各電話機は同時に複数の相手とは通話できない場合には大きな意味を持つが、超並列計算機では、そうでもない。個々のノードが自分に割り当てられた回線数以上の数の他のノードと同時に通信することはできず、ノード間の通信パターンはいずれにせよ制約されるからだ。

そこで、どうせノード間の通信パターンが制約されるなら、いっそ回線交換のパターンを制約して O(N)の規模でそこそこの交換パターンを実現しようという Fat Tree のような考え方が生まれる。しかし、ハードウェアの規模の削減により、交換パターンだけでなくバイセクションバンド幅も制約される。

交換パターンの制約は、ノードの CPU をルータとして使用することで一応回避はできるが、高い性能を持った高価な計算資源を単なるパケット交換に使用し、またその度に光と電気の変換を行うのは FLOPS 的にも消費電力的にも資源の浪費である。

そこで、パケット交換の出番である。交換パターンの制約は、回線交換に内在するものであるが、同時に複数のノードが他の複数のノードと通信するのが当然のパケット交換では、このような事態は生じない。

3. パケット交換と超並列計算機

パケット交換（各パケットのヘッダにパケットの行き先が直接書いてあるデータグラム型パケット交換）の大きな利点は、コネクションレスであることで、事前の通信路の設定なしに各機器は他の任意の機器と通信できる。これが、通信網において回線交換がほぼ消滅しインターネットが主流となった大きな理由である。また、通信単位は同じ通信路を流れるパケット間で共用されるため、通信路の帯域を効率的に使用できる。

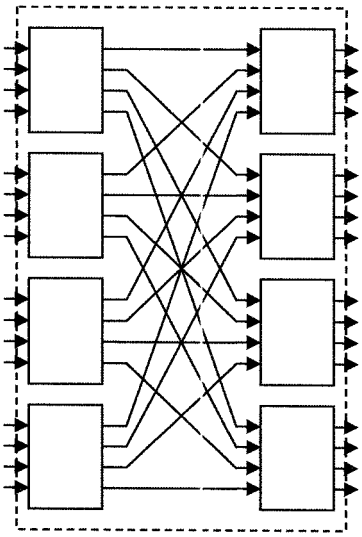

図 1　4 ポートスイッチ 8 個による
16 ポートスイッチの構成

超並列計算機においても、相互結合網にパケット交換を使用すれば、トポロジによりバイセクションバンド幅は変わるものの、とにかく全体が繋がってさえいれば任意のノード間でパケットを常に交換でき、交換パターンには何の制約もない。

K ポートのパケットスイッチにより N 個のノードを相互接続する場合、あるノードから他のノードまでに経由するパケットスイッチの平均段数は少なくとも $O(\log_K N)$であり、バイセクションバンド幅が N に比例する相互結合網のパケット中継の計算量は $O(N\log_K N)$となる。

図 1 のようにして K ポートの要素パケットスイッチから K^2 ポートのパケットスイッチを再帰的に構成することができる。このままでも全入力ポートから全出力ポートに同程度の量の通信を行う場合のバイセクションバンド幅は広いが、交換パターンによっては局所的に負荷が集中する可能性がある。横方向の段数を増やし経路の冗長性を確保し、初段のほうでパケットの行き先を複数の経路にランダムに振れば負荷の局所的な集中は統計的に回避することができ、また、故障したスイッチを通る経路を回避することで耐故障性を向上させることができる。Benes スイッチと同様に、バイセクションバンド幅が N に比例する相互結合網を O(NlogN)のハードウェア量で構成することができる。このときパケットスイッチでは Non-blocking 問題は発生しないが、かわりにバッファが必要となる。

パケット交換網の固有の問題としては、交換パ

ターンの制約を取り除いたことに必然的に由来する、バッファが必要であることと、パケット落ちが避けられないことがある。交換パターンに制約のないパケットスイッチには同時に同じ出力へのパケットが入力されることがあるが、その場合には、片方のパケットを待たせるためにバッファに蓄えておく必要がある。同じ出力へのパケットが複数の入力から入力されつづけた場合にはバッファに蓄えるべきパケット数が増えてついにはバッファの容量を超え、パケットを破棄せざるを得なくなる。

インターネットでは、この問題には、端末が必要に応じてパケット落ちを検出し、再送したり負荷の集中によるパケット落ちを緩和するために通信速度を低下させることで、対処している。

負荷集中によるパケット落ちがなくとも、大規模な通信網や大規模な超並列計算機による長時間計算ではエラーによるパケット落ちは避けられず、どのみち端末による再送は必須であり、超並列計算機でも同様に対処すればよい。

なお、図1でバッファは、個々の要素パケットスイッチ内部にある。

4. 光パケット交換と超並列計算機

ペタビット級の超並列計算機にはペタビット級の相互結合網が望まれる。ペタビット級の相互結合パケット網を構成する要素パケットスイッチも、速くて低消費電力であればあるほどよいため、光の広帯域性を生かした光パケット交換に期待が集まったが、これまで実用化は困難であった。

4.1 光バッファ

光パケットスイッチ実用化の最大の困難は、光パケットバッファにある。光パケットをいちいち電気に変換して半導体メモリに収納したのでは、光の広帯域性は失われ交換を電気で行うのと同じである。固定時間のメモリでよければ光ファイバ遅延線をメモリとすることはできるが、実用的なパケットバッファを実現するために必要な遅延線の数や長さが膨大であると思われてきた。ところが、筆者が[1]において提案し[2]で紹介した光パケット多重という考え方と、[3, 4]で提案した光ファイバ遅延線によるバッファ構成と制御方式により、この困難は解消した。

光パケット多重とは、パケット多重以外の多重化は伝送では使っても交換では使わず、伝送で利用可能な全帯域で個々のパケットをエンコードするというもので、光パケットが占める時間は短くなり、短い光ファイバでバッファできる。例えば1TbpS(10Gbps*100波長)が利用可能であれば、1500Bのパケットは12nsとなり、わずか2.5mの光ファイバ（屈折率 1.46）の遅延量に相当する。ヘッダ情報は少数の波長に収め、光パケットスイッチではそれらの波長のみを電気信号に変換して処理し制御に使い、残りのほとんどの波長は一括して光スイッチで出力先を切り替えることで、光の広帯域性を生かした光パケットスイッチとなる。2:2光スイッチとしては、100psで動作し消費電力 125mW のものが既に市販されている[5]。[3]では、光パケットの順序入れ替えを許すことで光バッファに必要な遅延線の数を減らせることを示し、[4]では順序入れ替えを許す制御回路をパイプライン化により高速できることを示した。

具体的には、1Tbpsで、平均パケット長500B(4ns)、最長パケット長1500B、最小パケット間隔2ns（つまり、100%負荷時666Gbps）で負荷率65%(433Gbps、ポワソン)の場合、15本の遅延線(最長 833m）でパケット落ち率 0.0017%を達成できる。なお、個々のTCPのトラフィックはポワソンとはとてもいえず効率よい中継のためには多量のバッファが必要だが、インターネット幹線や超並列計算機内部では多数のTCP等が平均化されるため、全体のトラフィックはポワソンと見て差し支えない [6]。

4.2 相互結合網の規模と消費電力

要素光パケットスイッチのポート数を 8 とすると、図 1 の方法で、要素光パケットスイッチ16 個で 64 ポートスイッチを構成でき、さらに2048 個で 4096 ポートスイッチを構成できる。各ポートの能力が 433Gbps とすると、速度は名目 4 Pbps、実効 1.8Pbps となる。負荷集中の防止と障害回避のためには 1 段追加につき要素パケットスイッチが 512 台増加し、1 段追加した場合（合計 5 段）は合計 2560 台となる。

8 入力で 15 本の遅延線を使用した光バッファは 240 個の 2:2 光スイッチで構成できるため、8 ポートの要素光パケットスイッチの光スイッチの消費電力は 240W となり、光スイッチ駆動回路等の電力も含めて 1 台で 600W とすると、相互結合網全体では 1.54MW となる。

なお、Fat Tree での構成を考えてみると、8 ポート要素光パケットスイッチ 192 台で 512 ポート(3 段)のパケットスイッチが得られるので、これをエッジスイッチとして 10 台並べエッジ側に10*410 ポート使い、コア側にでる 102*10 ポートを 2 台のコアスイッチで交換すると合計要素光パケットスイッチ数は 2304 台（追加段数なし）とむしろ増え段数も 9 段になり、何よりバイセクションバンド幅が減るため、何のメリットもない。

4.3 相互結合網の遅延と TCP/UDP

超並列計算機内部でこのような光要素スイッチを図1のように接続することで、パケットは相互結合網内部をほぼ光のままで通過する。相互結合網内で要素スイッチを5段通過し、各段の最大遅延が光ファイバ 1000m（最長遅延線長に配線長を加えた）分とすると、全体の遅延は最悪 24 μs となる。TCP の1本の理論速度は

MSS/RTT/sqrt(パケット落ち率)

程度なので[7]、MSS=1440B、RTT=48 μs、パケット落ち率 0.0017%としても、56Gbps となる。ただ、各光ファイバ上の速度が 433Gbps なのでこれほど速い TCP ではその本数も少なくなりトラフィック変動が平均化されないが、TCP の出力をほぼ等間隔にする Paced TCP 等の技術により、光ファイバ上のパケットのポワソン性を維持できる。また、光バッファではパケット順序入れ替えがおきるため、パケット順序入れ替えにあまり過敏に反応しない TCP を使用すべきである。

総和演算などのために少量のデータを迅速かつ確実に送りたい場合は、一応 ACK や再送は行うとしても、UDP により短い間隔で ACK を待たずに同じパケットを複数回（データは少量なので、問題はない）送れば、ほぼ確実にどれかのパケットは短い遅延線を通り目的地に到着するので、ACK を待つまでもなく目的地で計算は進行する。

なお[3]では、光ファイバ遅延線の数を節約しつつパケット落ち率を減らすため、光ファイバ遅延線長を等比数列的に長くしているが、光ファイバ遅延線数を例えば 31 本に増やせば、同じ負荷率でパケット落ち率を大幅に減らしつつ最長遅延線長をさらに短くできる。

4.4 パケット優先度と帯域保証（回線交換）

光パケットのヘッダに優先度ビットを設け光バッファの制御で考慮することで、総和演算などのためのパケットが落ちる確率や遅延時間を前節で述べた以上に減らすことができる。

また、一部のパケットを回線交換的に一定の帯域で低い遅延とパケット落ち率で運びたい場合も、パケット優先度を利用できる。この時はそのパケットが通る経路を全体で管理し、個々の要素ルータの入出力で高優先度のパケットの占める割合が一定以下となるようにする必要がある。

4.5 さらなる高速化

超並列計算機内部の配線長は短いので、その減衰や分散はほぼ無視でき、光ファイバ1本に例えば 10Tbps(40Gbps*250 波長)通すことにも特に困難はない。その場合光バッファに必要な光ファイバ長や光バッファによる遅延は、さらに短くなる。

5. おわりに

光パケットスイッチによる超並列計算機の相互接続パケット網の構築可能性を論じた。回線スイッチと異なりパケットスイッチには交換パターンに制約はないかわり、光バッファが必要である。光ファイバ1本に1 Tbps(10Gbps*100 波長)を通し市販の部品により構築した 4096 ポートスイッチの速度（≒バイセクションバンド幅）は名目 4 Pbps、実効 1.8Pbps、消費電力は 1.54MW 程度となり、十分実用的であるといえる。

今後の課題は、光ファイバ1本に通す光の速度を増加させた場合と、超並列計算機向けに多少低速でもより低消費電力の光スイッチを使用した場合の評価である。

文　献

[1] 太田 昌孝、「全光データパスルータの構成要素」、信学技報 PN、2005 年 8 月.

[2] 太田 昌孝、「光パケット多重ルータによるテラビット級広域分散計算」、情報処理学会研究報告、HPC、V. 2006、N.87、2006 年 7 月.

[3] 太田 昌孝、「光パケット多重ルータのファイバー遅延線による光バッファ構成 」、信学技報 PN、2006 年 3 月.

[4] 太田 昌孝、「パケット順序入れ替えを許す実時間光バッファの制御方式」、信学技報 PN、2007 年 3 月.

[5] "Ultra-High-Speed (sub-nanoseconds) 1x2, 2x2 Optical Switches/Modulators", 2006 Brief Switch Brochure, EOSpace, http://www.eospace.com/pdf/EOSPACE-custom OpticalSwitch2006.pdf, 2006.

[6] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, T. Roughgarden, "Routers with Very Small Buffers", ACM SIGCOMM Computer Communication Review, Vol 35, No. 2, July 2005.

[7] S. Floyd, "Connections with Multiple Congested Gateways in Packet-Switched Networks Part1: One-way Traffic", ACM SIGCOMM Computer Communications Review, Vol. 21, No. 5, October 1991.