Vol.2016-HPC-155 No.33
2016/8/10

パケット交換網の万能性と一般化 Benes 網による実現

太田昌孝[†1]

概要：回線交換網では全順列を実現可能な Rearrangeably Non-blocking のような性質が望ましいものとされるが，パケット交換網では Non-blocking は当然である．パケット交換網に望まれる性質は，単に全順列に留まらず全入出力帯域を使いきる全てのトラフィックパターンをボトルネックなしに許容することで，これを万能性と呼ぶこととする．万能相互結合網は，二つのバタフライ網や Benes 網やその一般化により実現可能である．

キーワード：バタフライ，Benes，万能相互結合網，高性能ネットワーキング

Universality of Packet Switching Networks and its Implementation by Generalized Benes Network

MASATAKA OHTA[†1]

Abstract: While properties such as "rearrangeably non-blocking", which means all the permutations are achieved, is desired for circuit switching networks, packet switching networks are, by nature, non-blocking. A property desired for packet switching networks is to allow all the traffic pattern to use up all the input/output bandwidth, not merely all the permutations, without bottlenecks, which we call "universality". Universal interconnection networks may be implemented with two butterfly networks, Benes networks or their generalization.

Keywords: Butterfly, Benes, Universal Interconnection Network, High Performance Networking.

1. はじめに

電気制御され光ファイバ遅延線をバッファに使う光パケットスイッチについては既に報告したが[1, 2], この度 64K ノードを 16Tbps の多波長光パケット（40GBaud DP-QPSK 変調の 100 波長を利用）で結合する場合（素のバンド幅 1.024Ebps）についてその消費電力を評価したところ，各出力ポートに 10 本の光ファイバ遅延線からなるバッファを備えた 4 ポート光スイッチ 8 段のバタフライ網で 1.49pJ/bit, 同 15 段の Benes 網で 5.3pJ/bit（光パケットの生成や受け取りのための EO/OE は除く, 実効バイセクションバンド幅はどちらも 0.53Ebps）という結果が得られた[3]．また，単純化した光パケットスイッチのモデルで消費電力を最小にするポート数は 3 であるという結果も得られた[4].

相互結合網のトポロジーとしては，多ポートスイッチを前提とした Flattened Butterfly [5]や，EO/OE の数を抑えた Dragonfly [6]等があるが，各段での EO/OE が不要で小ポートスイッチが最適な光パケットスイッチを前提とした場合には単なるバタフライや Benes トポロジーで十分である．そもそも 16Tbps のような高速では極めて短距離でも光伝送は必須で[6]の前提も成り立たない．また，[5, 6]では，ボトルネックを避けるために混雑状況を計測し適切な経路を選ぶことが前提とされているが，短期的な通信が多数発生し混雑状況の変動速度が速い場合にはそのような経路の選択は困難となる．

そこで，バタフライや Benes トポロジーを前提に，理想的なパケット交換による相互結合網のあり方を探ってみた．

回線交換網ではブロッキング特性が問題となり全順列を実現可能な Rearrangeably Non-blocking やある順列から別の順列に変更部分だけの回線切断で移行できる Strictly Non-blocking のような性質が望ましいものとされるが，パケット交換網では Non-blocking は当然である．バタフライ網は，全体を接続するための段数が最小で，入出力帯域に等しいバイセクションバンド幅を持ち，均一な全対全通信もバイセクションバンド幅で行えるが，ある種の順列のようなトラフィックパターンでは内部にボトルネックが生じ良い性能が出ない．

そこで望まれるのが，全入出力帯域を使いきる全てのトラフィックパターンをボトルネックなしに許容するような網で，このような網を万能網，この性質を万能性と呼ぶこととする．同じ入出力対間に複数の経路があってもかまわない．

この時，万能相互結合網は，バタフライ網を 2 段連結した網や Benes 網やその一般化により実現可能であることや，そのための経路の選択方法が簡単であることがわかったので以下に報告する．

†1 東京工業大学
Tokyo Institute of Technology

2. バタフライ網と Benes 網

*N=n*m* としたとき，*N* 対 *N* のバタフライ網は，単に *N* 対 *N* のクロスバ網であるか，*n* 個の *m* 対 *m* バタフライ網と *m* 個の *n* 対 *n* バタフライ網から図１のように構成される．

また，*N* 対 *N* の Benes 網は，*N* 対 *N* のバタフライ網とその左右反転したものを接続し，接続場所の両側のクロスバ網 は入出力直結なので 1 段にまとめたものである(図２)．

ポート数 *k* のクロスバ網のみから構成される *N* 対 *N* のバタフライ網の段数は $\log_k N$ となるが，クロスバ網を 1 段経由することで接続できる相手が *k* 倍に増えることを考えるとこの段数は *N* 対 *N* の網として最小である．Graph Golf [7] ではわざわざ双方向のグラフで半径を最適化するものを探求しているが，分散管理されたインターネットでは隣接するルータ間で監視パケットを相互に送りあってリンクの健全性をチェックできるメリットは大きいが，データセンターやスーパーコンピュータは管理主体が一つであり，内部のネットワークも集中的な管理をしてよいので，双方向性にこだわらずに最適なトポロジーを追求すべきであり，その答がバタフライ網であることは自明である．

[7]ではバンド幅が問われていないのも問題だが，バタフライ網が均一な全対全のトラフィックに対して全入出力帯域に等しい帯域を提供できることは，図１からも自明であろう．

しかし，バタフライ網では図３のようなトラフィックパターンでは一部のリンクに負荷が集中してボトルネックとなってしまう．

3. 万能性の実現

万能性では同じ入出力対間に複数の経路があってもかまわないことを積極的に利用すると，万能性をもった網は，二つのバタフライ網を縦列接続することで実現できる．

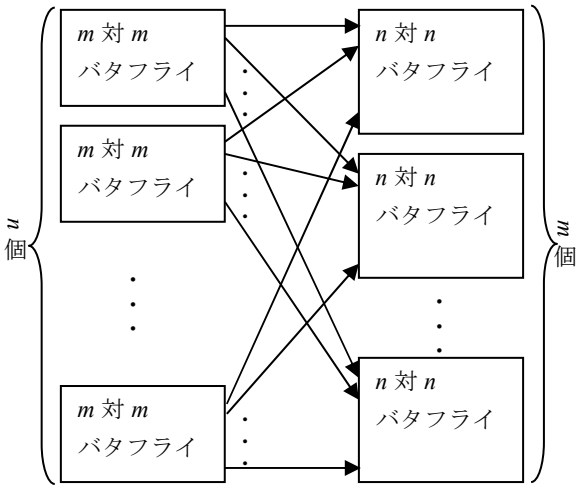

図１　*N* 対 *N* バタフライ網の構成（*N=n*m*）

すなわち，前段のバタフライ網では入力されたパケットを全ての出力に均等に出力し，後段のバタフライ網ではパケットを本来の目的地に出力する（図４）．この時，前段も後段もトラフィックパターンは均一な全対全になるため，どこにもボトルネックは発生しない．

図２から，Benes 網は二つのバタフライ網を接続したものと交換能力は等価であり，やはり万能性が実現できることは自明である．

図２では前段と後段の *m* 対 *m* バタフライの内部構造は左右対称であるとしているが，万能性の実現にはこの制約は不要である．そこで，接続部分のクロスバ網のポート数だけ合わせて二つのバタフライ網を接続した網を一般化 Benes 網と呼ぶこととする．もちろん，一般化 Benes 網は万能性を持つ．

4. ボトルネック回避のための経路選択

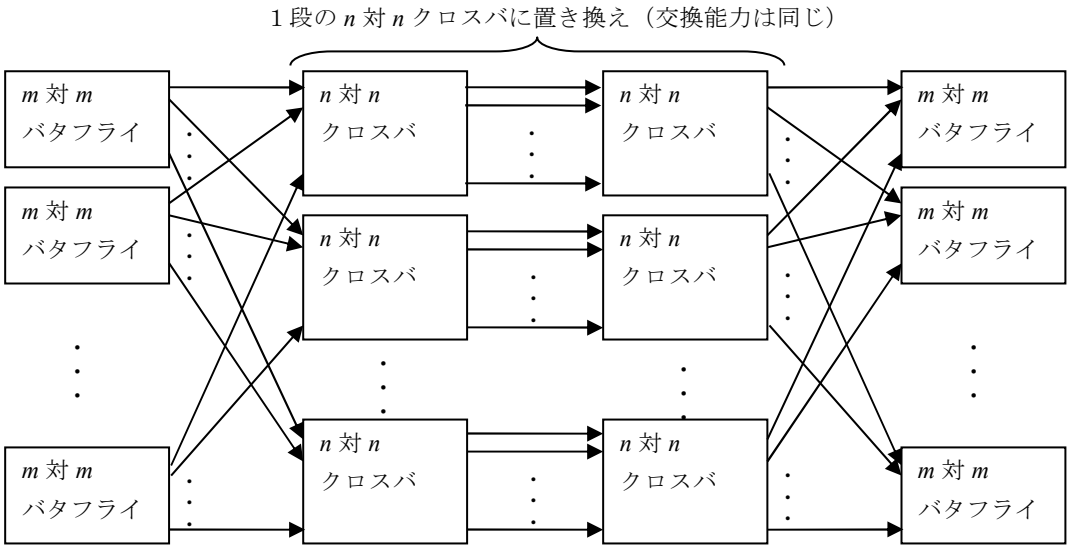

図２　*N* 対 *N* バタフライ網とその左右を反転した網からの *N* 対 *N* Benes 網の構成（*N=n*m*）

前節では前段のバタフライ網での経路は「入力されたパケットを全ての出力に均等に出力」すればいいだけであり，この実現方法としては，入出力対間ごとにラウンドロビンで行ってもよいが，別の入出力対間のパケットが次にどういう経路を辿るかまで考慮するわけではないので，完全にランダムに行っても結果は同じであろう．

すなわち，リンク上のトラフィック分布はポワソンと考えてよく，運が悪いとどこかのリンクに一時的にトラフィックが集中してパケットが落ちることもあるが，低確率でパケットが落ちること自体は全く問題にならないどころか，TCP による伝送帯域制御には必須であり，落ちたパケットは再送すればよいだけである．また，図４のように同一地点間で複数の経路を利用すると遅延が経路ごとに異なりパケット順序の逆転が起き TCP がパケット落ちと誤認することがある．しかし，これが起きるのはパケットが長い遅延線を通る場合，つまり混雑度が大きい場合である．これは，バッファが完全にいっぱいになる前にもある程度混雑度が大きければ低確率でパケットを落とす RED [8]機能に他ならず TCP の安定性のためにはむしろ望ましい．

ただし，ネットワーク中の遅延が小さくないと，TCP の帯域調整は高速に動作しない．このために必要なのは，パケットスイッチのバッファをあまり大きくしないことであり，[3]のバッファが 10 種類の光ファイバ遅延線からなるのは，このためには実は良い事である．

インターネットについてよく知らないと，パケットを落とすことが悪いことであるかのように思い，過剰なバッファを用意したり，バックプレッシャーによりパケット落ちを完全に防ごうとしたりしがちであるが，これは遅延の増大を招き TCP の能力を発揮できなくさせるだけである．昔のインターネットではこのようなことは有り得なかったが，最近では CATV 系の ISP がこのようなことをやりがちであり，混雑時には秒単位の遅延が引き起こされることすらあり Buffer Bloat 問題として知られている．データセンターやスーパーコンピュータ内の網でも事情は全く同じである．

Buffer Bloat が起きるような状況でも，網に余計な機能を追加したり端末で複雑な制御を行ったりすることで，時間をかけて全体で大量のデータを転送する時には通信遅延の影響を見えなくすることも場合によっては可能であろうが，本末転倒であるし，他のトラフィックが混在した場合（スーパーコンピュータの計算ノード群を２つ以上に分割して別々のジョブを走らせる場合等）にはそちらへの影響は避けがたい．

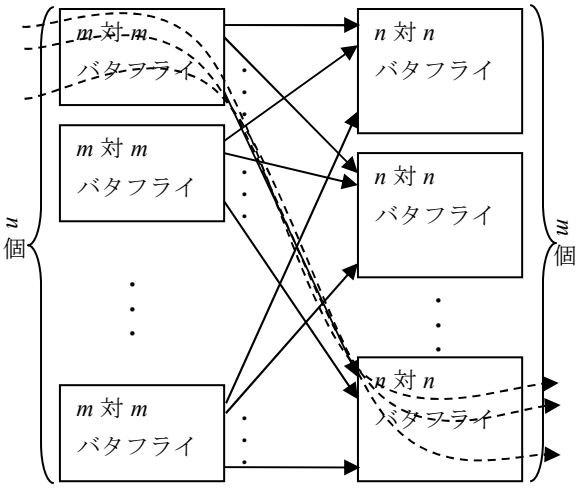

図３　バタフライ網でのボトルネックの発生

5. おわりに

全入出力帯域を使いきる全てのトラフィックパターンをボトルネックなしに許容できる網の性質を万能性と定義し，万能性が，バタフライ網を 2 段連結した網や Benes 網や Benes 網から左右対称性の要求を除いた一般化 Benes 網により実現可能であることを示し，また，万能性を実現するための経路制御は単に前段のバタフライ網（や，（一般化）Benes 網でそれに相当する部分）でランダムな経路を使えばよいことを示した．

[3]で提案された光パケットスイッチを用いれば，このための網を 4 ポート光スイッチ 15 段の Benes 網で 5.3pJ/bit のエネルギー消費で実現できることとなる．

未解決の課題として「回線交換網として Rearrangeably Non-blocking なトポロジーはパケット交換網として万能であるか」逆に「パケット交換網として万能なトポロジーは回線交換網として Rearrangeably Non-blocking であるか」がある．一見自明であるかのように思えるが，入出力対間を 1 つの回線で接続しなければならない回線交換網と，複数の経路を使ってよいパケット交換網との違いがあるため，よくわからない．後者の特別な場合として「一般化 Benes トポロジーは Rearrangeably Non-blocking であるか」という問題も考えられる．

参考文献

[1] 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”，SWoPP，2006.
[2] 太田昌孝，“超並列計算機向け多段光超高速内部接続”，SWoPP，2007.

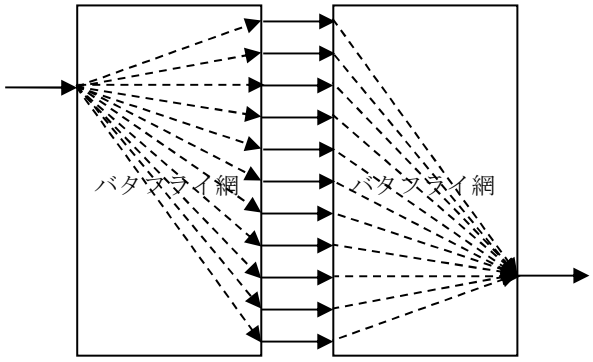

図４　バタフライ網２段での万能性の実現

[3] M. Ohta , “Optical Switching of Many Wavelength Packets: A Conservative Approach for an Energy Efficient Exascale Interconnection Network”, IEEE High Performance Switching and Routing (HPSR), 2016
[4] M. Ohta, “Optimal Radix for High Speed Optical Packet Switching”, IEEE HPSR 2016 Workshop, 2016
[5] J. Kim, W. J. Dally, and D. Abts, “Flattened Butterfly : A Cost-Efficient Topology for High-Radix Networks”, Proc. of the International Symposium on Computer Architecture (ISCA), 2007.
[6] J. Kim, W. J. Dally, S. Scott, D. Abts, “Technology-Driven, Highly-Scalable Dragonfly Topology”, Proc. of the International Symposium on Computer Architecture (ISCA), 2008.
[7] “Graph Golf”, http://research.nii.ac.jp/graphgolf/.
[8] S. Floyd, V. Jacobson, “Random Early Detection gateways for Congestion Avoidance”, IEEE/ACM Transactions on Networking, 1993