Vol.2016-HPC-157 No.13
2016/12/22

光ファイバリンクによる相互結合網のトポロジー

太田昌孝[†1]

概要：バタフライや Benes トポロジーは配線コストが膨大になると誤解されがちだが，三次元配置を利用すると三次元トーラスと配線量はそれほどの違いはない．バタフライや Benes ではリンク長は長くなるが光ファイバは損失が無視できるので問題ではない．光ファイバの大容量を生かせば，エクサスケールの相互結合網を無理なく実現できる．

キーワード：Exascale 相互結合網，バタフライ，Benes，三次元トーラス，Scalably Wired Clustered バタフライ

Topology of Interconnection Network with Optical Fiber Links

MASATAKA OHTA[†1]

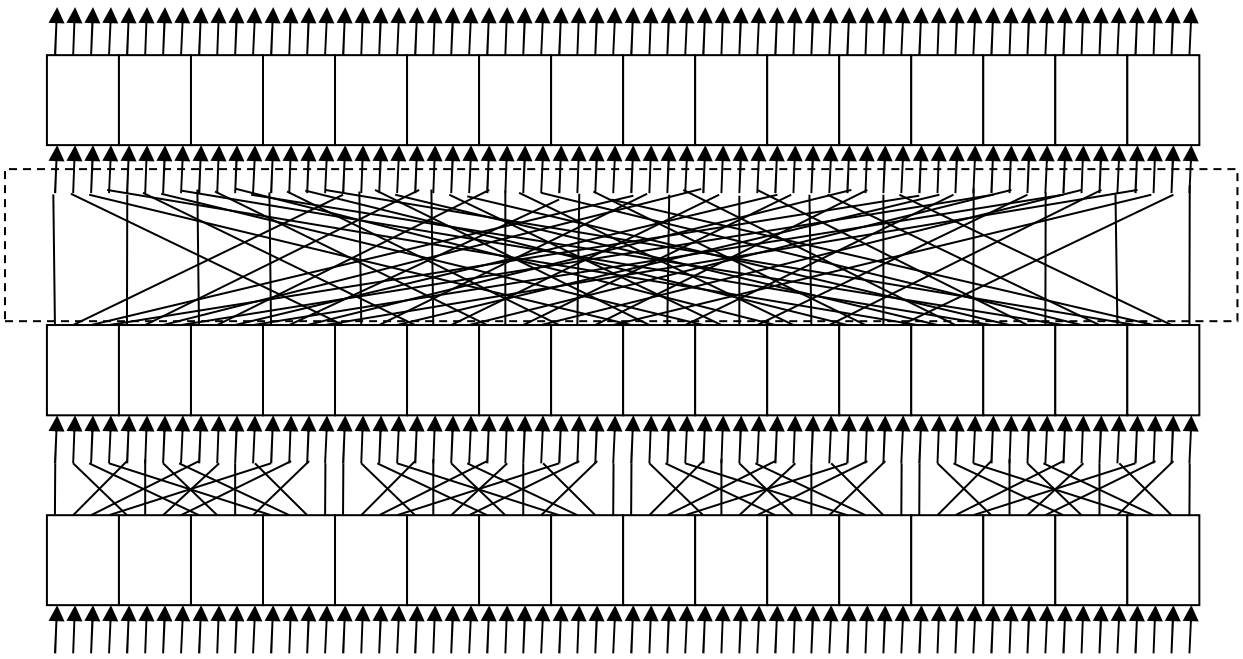
Abstract: Though it is often misunderstood that wiring cost of butterfly/Benes topologies is a lot, the cost is not so different from that of 3D torus, if 3D positioning is used. Long links required for butterfly/Benes is not a problem for optical fiber links with negligible loss. If large capacity of optical fiber is fully utilized, Exascale interconnection can naturally be realized.

Keywords: Exascale Interconnection Network, Butterfly, Benes, 3D Torus, Scalably Wired Clustered Butterfly

1. はじめに

多数（*N* 台）の計算ノードや要素スイッチ間を光ファイバリンクで結合する相互結合網のトポロジーについて、バタフライや Benes と三次元トーラスを使った場合について，配線コスト（総配線長），スイッチコスト，バイセクションバンド幅の観点で比較する．バタフライや Benes は長距離配線があり大変と思われがちで，実際に同軸ケーブルリンクでは相互結合網内でも距離による減衰が無視できず長距離では太いケーブルが必要になる．しかし，光ファイバリンクでは長距離リンクでも同じケーブル（コード）が使える．また，光ファイバの通信容量を生かせば同軸ケーブルより遥かに細いコードで高バンド幅の通信が実現できるが，多波長光パケットによる光パケット交換網1ではこのような通信が可能である．

以下、リンクの容量を 1，計算ノードの間隔を各次元で



†1 東京工業大学
Tokyo Institute of Technology

1, 配線の距離はマンハッタン距離，ラディックス k の要素スイッチのコスト（内部の複雑性や消費電力）を特に断らない限り k^2 としてう評価する．またオーダーが最も高い項以外は無視する．Benes 網の配線量やスイッチコストはバタフライ網の約倍なので，以後はバタフライ網のみ評価する．

図 1 は 64 台の計算ノードを 1 列に並べラディックス 4 の要素スイッチを用いバタフライトポロジーで結合する相互結合網の図であるが，配線の量は見てのとおり多く，配線コスト（横方向）は $N^2/2 \cdot k/(k\text{-}1)$となる（以後は $k/(k\text{-}1)$ の要素は無視する，つまり図の点線部分だけ考える）ため N が大きくなると実用的な配線パターンとは言いがたい．しかしながら全部の計算ノードを 1 列に並べてそれらを 3 次元トーラストポロジーで結合した場合の配線コストも $O(N^{3/5})$となり同じく実用的ではない．つまり図 1 の問題は計算ノードの配置に三次元性を生かしていないことにある．

計算ノードを三次元空間で各次元に $N^{1/3}$ 台の立方体状に配置すると三次元トーラスが三次元空間内で $12N$ の配線コストで実現できる．一方， N ポートのバタフライ網は $N^{1/3}$ ポートのバタフライ網を $N^{2/3}$ 台並べたものを 3 段重ねて構成することができ，三次元空間で効率的な配置が可能である．三次元トーラスと同様の計算ノードの配置で，この時の配線量は $3N^{4/3}/2$ となり，三次元トーラスとの差は $N^{1/3}/8$ とそれほど大きくないどころか N が小さい場合にはむしろバタフライのほうが有利となる．

実際 64 台の計算ノードを縦横高さ方向に 4 台ずつ並べた場合，三次元トーラスでは配線量は 576，スイッチコストはラディックス 7 の要素スイッチが 64 台必要なため 3136，バイセクションバンド幅は 32 となる．一方，バタフライでは 16 台のラディックス 4 の要素スイッチを 3 組用意し，各組をそれぞれ縦横高さ方向の結合に使うと，配線量は 384，スイッチコストは 756，バイセクションバンド幅は 64 となり，バタフライのほうが三次元トーラスより有利となる．

同軸ケーブルの時代には三次元トーラスではリンク長が最大 2 なのは利点と言えたが，光ファイバリンクでは関係ない．逆に，$O(N)$のバイセクションバンド幅を実現するためには，計算ノードを縦横高さが $O(N^{1/3})$の二つの塊に分けたとき塊間に $O(N)$本のリンクが存在する必要があり，リンク長が $O(N^{1/3})$になったり配線コストが $O(N^{4/3})$になったりするのは必然であり，この意味で三次元空間を利用したバタフライは最適である．

逆に，N が大きくなった時に配線コストを三次元トーラス同様の $O(N)$に抑えるためには，計算ノード $O(N^{1/3})$台を適当に結合して 1 つのクラスターとし $O(N^{2/3})$個のクラスター間をバタフライ網で結合すれば良い．これを SWC (Scalably Wired Clustered)バタフライと呼ぶことにする．クラスター内の配線量はスケーラビリティには影響しないものとする．

N が大きな場合として，2 節では[1]のように 2^{16} の計算ノードを 16Tbps のリンクで結合した場合，3 節では 2^{18} 個の計算ノード 2^6 個のクラスターとした SWC バタフライ網について，三次元トーラスと比較しつつ考察する．

2. 多波長光パケットによる超高速リンクの場合

[1]では 40Gbaud の DP-QPSK (Dual Polarization Quadrature Phase Shift Keying)変調された光を 100 波長利用した 16Tbps の多波長光パケットを使ったラディックス 4 の光パケット交換機により 2^{16} の計算ノードをバタフライや Benes トポロジーで結合しその消費電力をそれぞれ 1.49pJ/bit と 5.3pJ/bit と見積もった． 多波長パケットの考えにより 1 本の光ファイバで 16Tbps のパケットが運べるわけである．バタフライの場合の光パケット交換機の段数 8 段，台数は 131072 台である．計算ノードはラック内（垂直方向）に 16 台収納し，4096 本のラックを 64*64 の方形に配置することを仮定している．

同様の配置で自然な三次元トーラスを構成した場合，配線コストは 786432，スイッチコストは 3211264，バイセクションバンド幅は 4096 となる．ただしスイッチの消費電力については，別に考察する必要がある．

まず個々のスイッチの消費電力はラディックス(7)の二乗に比例するが信号強度にも比例する．SNR を一定とすると信号強度は雑音強度に比例し，雑音強度は光増幅器の段数つまりスイッチの段数に比例する．最大段数は 72 段とバタフライの 9 倍，スイッチの数は 2^{16} とバタフライの 1/2 なので，消費電力は$(7/4)^2*9/2=13.8$ 倍の 21pJ/bit となってしまう．

一方バタフライの場合，配線コストは 4718592 と三次元トーラスの 6 倍だが，スイッチコストは 2097152 と三次元トーラスと同程度（消費電力コストは上記のとおり 1 割未満），バイセクションバンド幅は 65536 と三次元トーラスの 16 倍となる．

バタフライの配線コストが実際にどのくらい大変かを評価しておこう．ラック内は 8 個の単位光スイッチ 2 段のバタフライ，ラック間は 48 個の単位光スイッチ 3 段のバタフライをラックあたり縦横 16 個ずつで結合することになる．

大変なのはラック間の長距離結合なので，以下ではラック間の横方向の結合だけ考えると，ラックあたりの単位光スイッチは 12 個，長距離結合なのは図 1 の点線部分で 64 個のリンクを含むので，64 本の光ファイバコードを組み合わせて図 1 の点線部分に相当するケーブルを作っておくこととする．コードの直径を 2mm とするとこれを 64 本収容するケーブルの直径は 2cm もあれば十分である．ラック間にはこのケーブルが 16 本渡されることとなるが，ラックの大きさからすると無視できるケーブル量である．

結線の規則性も，図 2 のように各ラックに 4 ポート光パケット交換機 3 段を 4 組配置し，4 ラックおきに 16 個のラックを長距離結合するケーブルを，対応する光パケットスイッチのポートに接続すればいいだけなので，極めて規則的であり誤接続の可能性もまずない．

現在のイーサネットでよくある構成では 4 本の光ファイバで 100Gbps しか運ばず，この 640 倍のケーブルが必要になるのでかなり大変であろう．

3. SWC バタフライの場合

超高速リンクを用いない等の理由で長距離リンクのケーブル量が問題になる場合に SWC バタフライの効果を評価しよう．

2^{18} 個の計算ノードを縦横高さ 2^6=64 個並べて三次元トーラスで結合する場合，配線コストは 3145728，スイッチコストは 12845056, バイセクションバンド幅は 4096 となる．

SWC バタフライでの結合では，2^{18} 個の計算ノード 2^6 個で 1 クラスターとし, 2^{12} のクラスターを結合すると同じバイセクションバンド幅となるが，クラスター内接続を考えない場合の配線コストは 393216，スイッチコストは（ラディックス 16 として）196608 となり三次元トーラスと比べて無視できるコストしかかからない．クラスター内接続を考えるとそうもいかないが，クラスター内接続をラディックス 8 の単位スイッチ 2 段（両端）と 9 の単位スイッチ 1 段（中央）の 3 段 Benes で構成する（72 の入出力のうち 64 は計算ノードと，1 はクラスター間との接続で使うとしてあまった 7 はクラスター間の三次元トーラスにでも使えばよい）とすると，総配線コストは 393216+2126440=3014656 と三次元トーラスより少し少なく，総スイッチコストは 196608+7372800=7569408 と三次元トーラスの 6 割たらずとなる．

つまり, SWC バタフライは三次元トーラスに比べて同程度の配線コストで消費電力を 6 割程度にできることとなる．また，クラスター間結合網を複数用意すれば総コストのわずかな増加でバイセクションバンド幅を増強することもできる．

4. おわりに

多数の計算ノードを結ぶ相互結合網のトポロジーとして，長距離リンクで特に減衰がなく超高速（16Tbps）リンクも実現可能な光ファイバの特性に基づいて，バタフライや Benes トポロジーと三次元トーラスの配線コスト，スイッチコスト，バイセクションバンド幅を比較した．

超高速リンクでは，ラック間に 2cm 程度のケーブルを縦横それぞれ 16 本はわせ規則的に結線するだけで Exascale 相互結合網が実現できる．光パケット交換を行う場合は消費電力が段数に比例するため，三次元トーラスではバタフライや Benes に比べて消費電力が非常に多くなる．

また，リンク速度があまり速くない場合，バイセクションバンド幅が三次元トーラス程度でよければ，いくつかの計算ノードをクラスター化してクラスター間をバタフライ網で結合する SWC (Scalably Wired Clustered)バタフライトポロジーで三次元トーラスより低コストで結合できる．

参考文献

[1] 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