コンピュータの中のグラフ：大規模並列計算機のネットワーク設計

Graphs in Computers: Designing Interconnection Networks for HPC Systems

藤原　一毅[*1]
Ikki Fujiwara

[*1] 国立情報学研究所
National Institute of Informatics

1. はじめに

スーパーコンピュータに代表される大規模並列計算機(HPC) の内部には、数十コアのプロセッサを持つ計算ノードが数百台～数万台あり、それらがネットワークで結ばれている。アプリケーションは数万～数百万のプロセスが互いに通信しながら計算を進めるため、ネットワークの性能が全体の性能を大きく左右する。なおかつ、ビッグデータ解析をはじめとする最近の並列アプリケーションは小さなパケットを頻繁にやりとりする傾向があり、ネットワークの遅延が全体を律速する状況が生じている。次世代のスーパーコンピュータが 100 倍の性能向上を達成しようとするならば、プロセッサの性能を 100 倍にするだけでは足りず、それらを結合するネットワークにも革新的な設計が求められる。本稿では、次世代スーパーコンピュータのネットワーク設計に向けてどのようなアプローチが有望なのか、そこに離散構造処理の知見をどのように応用するべきかについて、計算機研究者の立場から概説する。

2. ネットワークの構成とモデル化

スーパーコンピュータというと、体育館くらいの部屋に冷蔵庫のようなラックが整然と並んだ姿を思い浮かべる。スパコンの典型的な設計では、ラック内に 100～1,000 台程度の計算ノードが収められ、各計算ノードが 10 個前後のネットワークルータ（ポート）を内蔵している。このルータ同士を互いに接続して作られるネットワークを直接網という。例として、理研の京コンピュータ（6 次元トーラス）や IBM の BlueGene/Q シリーズ（5 次元トーラス）がある。直接網は計算ノードを頂点とする正則無向グラフとしてモデル化できる（図 1 左）。

より汎用品的な設計では、ラック内に 1～数台のネットワークスイッチと 40～120 台程度の計算ノードが収められる。各計算ノードは 1 個のネットワークポートを介してスイッチに接続され、スイッチ同士はさらに上位のスイッチを介して結合される。このようなネットワークを間接網という。例として、東工大の TSUBAME 2.0 や中国国防科技大の天河二号（いずれもファットツリー）がある。間接網はスイッチおよび計算ノードを頂点とする無向グラフとしてモデル化できるが、平均距離等の計算でスイッチを表す頂点を含めないことに留意する（図 1 右）。

スーパーコンピュータのネットワークは従来、巨大なデータを速く流す広帯域志向の設計が主流であり、直接網では 2～3 次元のメッシュやトーラス、間接網ではファットツリーや Myrinet Clos[1] などのトポロジがよく使われてきた。

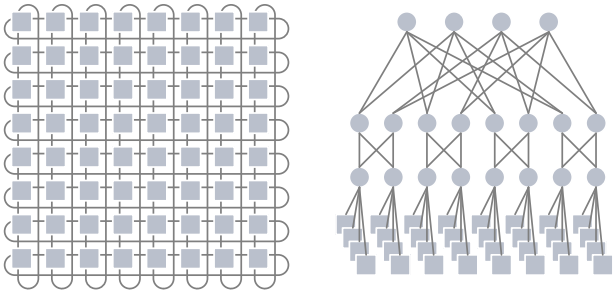

図 1　直接網（左：2 次元トーラス）と間接網（右：ファットツリー）。四角はノード、円はスイッチを表す

将来のスパコン用ネットワークは対照的に、小さなデータを早く届けるために低遅延志向の設計が求められる。

送信ノードから受信ノードまでの通信遅延は、輻輳にともなう待ち時間を別にすれば、(1)送信ノードがネットワークにパケットを注入する時間、(2)パケットがケーブルを伝搬する時間、(3)パケットが経由ノード（スイッチ）を通過する時間、(4)受信ノードがネットワークからパケットを取り出す時間、の総和としてモデル化できる。このうち(1)と(4)は定数である。(2)は光ファイバ中の光速の逆数＝5.0 ナノ秒／メートルである。(3)は製品によって異なるが、現行製品で 200 ナノ秒／ホップ程度である。スパコンは大きくても数十メートル四方なので(2)は(3)に比べて十分小さく、通信遅延を支配するのは経路上の経由ノード数、すなわちグラフ上の 2 頂点間の最短パスのホップ数であると言える。したがって、低遅延なネットワークを設計するには、直径・平均距離が小さいトポロジを採用することが第一義的に重要である。

3. ランダムトポロジ

与えられた頂点数と次数をもつ無向グラフの中で、直径・平均距離が小さいのはランダムグラフであることが経験的に知られている。この性質に着目し、スパコンのネットワークにランダムトポロジを採用することで低遅延化を図ろうとする研究が近年注目されている[2-4]。従来スパコンで使われてきた 3 次元トーラスと、同じ 6 次のランダムトポロジの直径・平均距離を比較した（図 2）。ランダムトポロジの採用により、ノード数が大きくなればなるほど、直径・平均距離を劇的に削減できることがわかる。ランダムトポロジは拡張性や耐故障性の面でも優れた性質を持っている反面、安直に実装すると大量のケーブルが山のように積み重なる事態になりかねない。筆者らはランダムトポロジのケーブル問題を緩和する手法を提案している[5]。

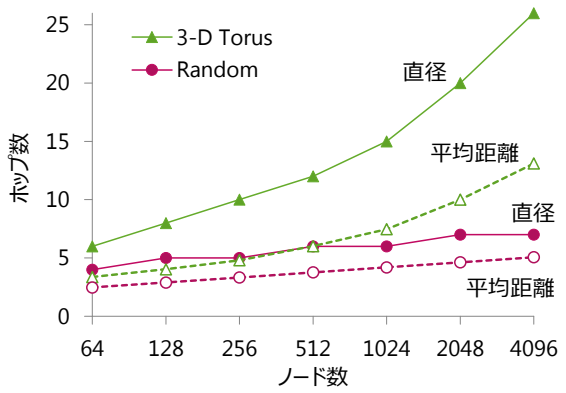

図 2　トポロジ別の直径と平均距離（次数 6）

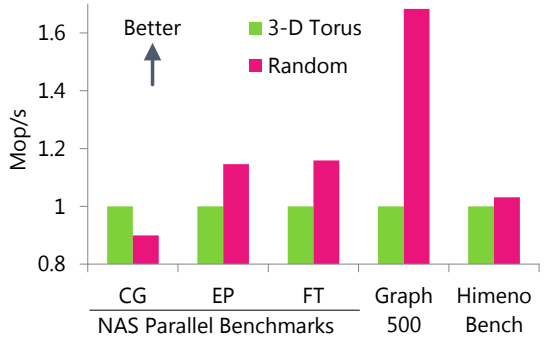

図 3　トポロジによるアプリケーション性能の差
（64 ノード、次数 6、トーラスの性能で正規化）

ランダムトポロジは低遅延ネットワーク設計の切り札として有望なアプローチだが、エンジニアの立場からは「でたらめな設計が最善だと言われても納得しがたい」という意見が聞かれる。頂点数と次数が与えられたなら、直径・平均距離が最小のトポロジは決定論的に定まるはずだという直観である。この疑問に対する我々の答えはまだ見つかっていないが、次に述べる order/degree 問題を追究する中で答えが見つかるかもしれない。

4. Order/Degree 問題

与えられた次数・直径を満たす中で頂点数がなるべく大きいグラフを見つける問題は degree/diameter 問題 (DDP) と呼ばれ、これまでに多くのグラフ理論家が取り組んできた成果が蓄積されている[6]。しかし、スパコンのノード数は予算や電力などの外部要因で決まるため、DDP の解をそのままトポロジとして採用できる可能性は低い。計算機ネットワーク設計者はあくまで、与えられた頂点数・次数を満たす中で、直径がなるべく小さい無向グラフを見つけたい。この問題は order/degree 問題 (ODP) と名付けられているが[7]、筆者の知る限り、これまで誰も真剣に取り組んでこなかった（語弊があればお詫びする）。

そこで、筆者らは ODP をオープンサイエンスの俎上に載せようと考え、小直径グラフ探索コンペ “Graph Golf” を開催している[8]。参加者は与えられた頂点数・次数を満たすグラフを作成・投稿し、その直径・平均距離の小ささを競う。優れた貢献者は国際会議 CANDAR’15 で表彰する予定である。“Graph Golf” の究極の目標は、すべての頂点数・次数の組合せに対し最小の直径・平均距離を持つグラフのカタログを作り、計算機ネットワーク設計者に提供することである。コンペは誰でも参加可能なので、読者もぜひ挑戦していただきたい。

5. トポロジとアプリケーション性能の関係

並列アプリケーションの性能は、トポロジの直径・平均距離だけでなく、通信パターンや輻輳による待ち時間などネットワーク上のさまざまな要因に左右される。計算機ネットワーク設計者は、それらの要因を考慮したシミュレーションによって、設計したネットワーク上でのアプリケーション性能を予測する。図 3 は、筆者らが改良に参加したフローレベル・ネットワークシミュレータである SimGrid を用いて、64 ノードの 3 次元トーラス（平均距離 3.37）とランダムトポロジ（平均距離 2.47）上における 5 種類の並列アプリケーションの性能をシミュレートした結果である。アプリケーションごとに特有の通信パターンがあり、最適なトポロジもそれぞれ異なるが、Graph500 のように通信パターンが不定なアプリケーションでは特に、平均距離の小さいランダムトポロジが性能向上に寄与することがわかる。

6. おわりに

大規模並列計算機の設計にあたっては、本稿で述べたネットワークトポロジ最適化のほかにも、ラック配置最適化（二次割り当て問題）や動的ネットワーク再構成の最適化（ジョブスケジューリング問題）など、さまざまな離散構造処理の問題が顔を出す。次世代スパコンの設計には従来手法の延長ではない革新的なアイディアが求められており、計算機研究者と数理研究者との協同によるブレイクスルーが不可欠であると筆者は考えている。

参考文献

[1] Myricom, “Guide to Myrinet-2000 Switches and Switch Networks,” 2001, http://www.myricom.com/scs/myrinet/m3switch/guide/

[2] J.-Y. Shin, B. Wong, and E. G. Sirer, “Small-world datacenters,” in Proc. 2nd ACM Symposium on Cloud Computing, 2011, pp. 1–13.

[3] M. Koibuchi, H. Matsutani, H. Amano, D. F. Hsu, and H. Casanova, “A case for random shortcut topologies for HPC interconnects,” in Proc. 39th International Symposium on Computer Architecture (ISCA), 2012, pp. 177–188.

[4] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey, “Jellyfish: networking data centers randomly,” in Proc. 9th USENIX conference on Networked Systems Design and Implementation (NSDI), 2012, p. 17.

[5] M. Koibuchi, I. Fujiwara, H. Matsutani, and H. Casanova, “Layout-conscious random topologies for HPC off-chip interconnects,” in Proc. 19th International Symposium on High Performance Computer Architecture (HPCA), 2013, pp. 484–495.

[6] “The Degree/Diameter Problem,” *Combinatorics Wiki*, http://combinatoricswiki.org/wiki/The_Degree/Diameter_Problem

[7] M. Miller, and J. Širán, “Moore graphs and beyond: A survey of the degree/diameter problem,” *Electronic Journal of Combinatorics*, Dynamic Survey #DS14, 2013.

[8] “Graph Golf: The Order/Degree Problem Competition”, http://research.nii.ac.jp/graphgolf/