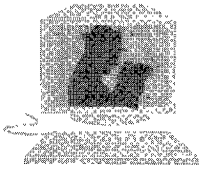
スーパールーティング

太田 昌孝
東京工業大学 総合情報処理センター
mohta@necom830.hpcl.titech.ac.jp



スーパーコンピューティング

本誌7月号のインタラクティブ・エッセイは「これでいいのか？日本のスパコン」というタイトルだったが，修士論文のテーマが汎用並列計算機であった筆者には感慨深いものがあった．

今どきいくらベクター処理に工夫をこらしたところで個々のプロセッサは家庭用のゲーム機に毛が生えた程度の能力しか持たないわけで，スパコンは超並列計算機にならざるを得ない．超並列といえるほど並列度が上がると，メモリ共有型ではメモリアクセスレイテンシが非常に大きくなるので，要素プロセッサはプロセッサとメモリを含むことになる．

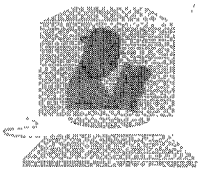
並列計算機で重要なのは，要素プロセッサの総合結合網である．並列計算の場合，解くべき問題やアルゴリズムによっては相互結合網にそれほどの能力が要求されない場合もあり，インターネットなどを利用したクラスタコンピューティングで十分である．ある種の計算にキャッシュが効くのと同じことである．

しかし，問題やアルゴリズムによっては真に大域的な通信が大量に必要な場合がある．自明な例では，ランダムな数列の並列ソートは，いかにアルゴリズムを工夫しても全データの大域的な通信は避けようがない．

こういう場合が扱えてこそ，スパコンである．

相互結合網をK×Kの要素スイッチから構成する場合，N個の要素プロセッサを結合するには$\log_k N$段のスイッチを経由する必要がある．各プロセッサがスイッチの能力と同じ速度で相互結合網と通信するとすると，スイッチの数は最低でも$N/K*\log_k N$個必要となる．この最低値を達成するトポロジーが存在し，多少のバリエーションはあるが等価なもので，ハイパークロスバー，バンヤン，シャッフルエクスチェンジなどと呼ばれている（図-1）．逆にこのような相互結合網を利用すれば，並列計算可能な問題は，すべてそこそこの効率で計算でき，汎用並列計算機のでき上がりである．

重要なことは，本格的な相互結合網を備えた超並列計算機では，そのコストのほとんどが相互結合網に費やされることである．そこで相互結合網の低コスト化はきわめて重要な課題である．



スーパールーティング

という話も昔は面白かったが，今どきスパコンのような特殊な機器は市場規模が小さすぎてつまらない．一方，インターネットに要求される速度は今後もいくらでも増大し，それに見合った超高性能のルータが多数（今の電話局の数程度？）必要となる．10万加入者（大きめの電話局に相当する）がそれぞれ10Mbpsしか利用しないとしても交換能力は1Tbps必要だし，幹線で100万加入者が1Gbpsずつ利用すると1Pbpsの交換能力が必要となる．

現状では並列処理を行わないルータの能力は10Gbps程度でしかないので，1Tbpsを実現するだけでも100台の要素ルータによる並列処理が必要となる．今後の半導体技術などの進歩は，ルーティングの速度だけでなく端末機器の速度も向上させ帯域への要求も増大するため，並列化の必要性は消えない．1Tbpsの要素ルータを1000台並べて1Pbpsの超並列スーパールータとするなどといったことになる．

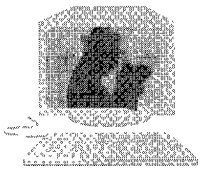
超並列ルータは，入力されたパケットを適当な出力に高速に転送することが仕事である．そのためには容量の大きな相互結合網を持つ必要がある．また，どの入力からのパケットがどの出力に出やすいといったことをあらかじめ仮定したトポロジーを工夫して相互結合網を節約するわけにもいかない．

そこで，スーパールータにはまともなスパコンの相互結合網と同程度の相互結合網が必要である．もちろん，スーパールータのコストのほとんどは相互結合網にかかることになる．

相互結合網では途中のスイッチで同時に同じ行き先のパケットが複数入力された場合にはバッファが必要となる．バッファによる遅延を嫌ってバッファを減らすために複雑な処理を考える人は多いが，本コラム3月号で紹介した方式を利用すれば，各スイッチでバッファしても問題となるほどの遅延は発生しない．

なお，スーパールータの場合，スパコンと違って，相互結合網での少々のビット誤りや欠けは問題にならないので相互結合網への要求はかなり緩和される．スーパールータ内の通信も含めて，インターネットのプロトコルはパケット落ちに対応できるのが当然であり，誤りを含むパケットは捨てればよい．

逆に，スパコンでも，相互結合網で誤りが起きても全計算を中断するのでなく，うまくソフトウェアで再送を行うようなアーキテクチャ上の工夫が必要であろう．そうでないと，相互結合網のコストが必要以上に跳ね上がる．クラスタコンピューティングではTCPなどを利用してこれを行っていることもあって，低コストになるわけである．



負荷分散

10Gbpsの要素ルータを100台並べて1Tbpsのスーパールータを構成したとしよう．ここで2地点間の通信速度が500Gbps必要だとすると，50台の要素ルータを並列に利用する必要がある．このときいかにうまくパケットを各要素ルータに分散するかが，効率的な並列化に最も重要である．

もちろん，パケットの分散のために複雑な処理をしていたのでは，効率は低下してしまうので，並列処理屋の腕のみせどころとなる．

効率的な分散のためには，たとえば，パケットの行き先アドレスなどをハッシュして担当要素ルータを決めることができる．これでだいたい負荷は分散するし，同じTCPに属するパケットは同じ要素ルータで処理されパケットの順序が逆転しない（逆転するとTCPの速度は低下する）のでベストエフォート通信に向いた方式といえる[1]．

しかし，QoS保証通信では，だいたいの負荷分散ではたまたま負荷が集中した場合QoS保証ができなくなるので，厳密な分散が必要である．幸いなことにQoS保証通信は事前に帯域などの資源予約のためのシグナリングが行われるため，シグナリングの際に厳密な負荷の分散を行えばよい[1]．

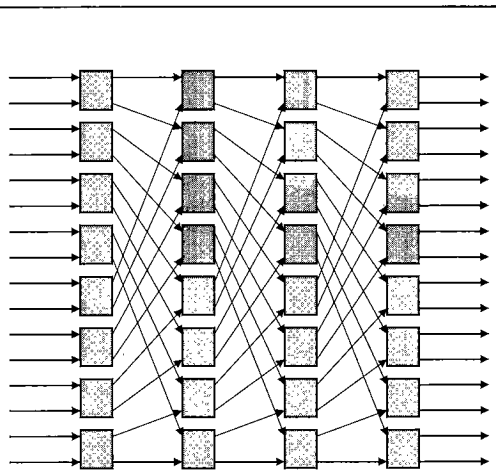
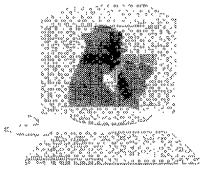


図-1　効率的な相互結合網のトポロジー



光相互結合

相互結合網にコストがかかるのは，それが膨大な配線量を意味するからである．電線1対で300Mbps伝送できるとして，10Gbpsの伝送には33対の電線が必要となる．10Gbpsの要素ルータを100台ならべた1Tbpsのスーパールータで相互結合網が10×10のスイッチ2段からなるとすると，10Gbpsの伝送路が300必要であり，総配線数は10,000対である．

電線の場合300Mbpsというのはほぼ限界であり，半導体などの処理速度が向上して通信速度が向上するとますます配線数が増大する．

一方，相互結合網を光ファイバで構成することを考えると，ファイバに10Gbpsを通せば300本の伝送路で同じスーパールータが構成できる．ファイバ上の伝送速度は半導体などの処理速度に比例して向上すると考えてよいので，将来的にも問題は起きない．

スパコンで培った相互結合網の技術を生かし，光相互結合方式のスーパールータを作成，その一部をスパコンにフィードバックするのが正しい技術の発展の方向ではないだろうか．

参考文献
1) Ohta, M., Sola, M., Fujikawa, K., Kojima, A., Fukumori, H. and Muraoka, Y.:Hash Parallel and Label Parallel Routing for High Performance Multicast Router with Fine Grain Qos Control, Proceedings of Internet Workshop '99, pp.13-16 (Feb. 1999).

(平成12年8月11日受付)