計算機セミナーレポート
相互結合網 (Interconection Network)

前川 勉 (富田研)

提出日 :1998年7月13日(月)

1 はじめに

並列コンピュータの重要な構成要素である相互結合網について、その分類、評価項目を示した後、静的網、動的網について述べる。

2 相互結合網の分類と評価項目

以下の説明において、ノードとは相互結合の対象となる演算器やプロセッサを意味し、リンクとはデータ転送路のことをいう。

2.1 分類項目

1. 制御方式
    - 集中制御方式
      中央に設置された一つの制御装置によってすべてのスイッチ類の動作が制御される。—SIMD方式に適合した制御方式。
    - 分散制御方式
      スイッチ自体に制御装置が内蔵されており、各スイッチは自律的に動作する。—MIMD方式に適合した制御方式。
2. 交換方式
    - パケット交換方式 (packet switching)
      メッセージに宛名を付けて結合網に送り出すので、送受信ノード間に物理的リンクを確立する必要がない。 短いメッセージの通信や通信相手が頻繁にかわるような場合に有効。
    - 回線交換方式 (circuit switching)
      送受信ノードに物理的リンクを確立して、メッセージ交換がおこなわれる。リンクの確立に時間を要すが、大量データの高速転送に適している。
3. トポロジ相互結合網の幾何学的形状は大きく不規則網と規則網に分類される。規則網では、さらに静的網と動的網に分類される。

2.2 評価項目

1. 距離 (distance)
   静的網ではノード間の最小中継ノード数+1、動的網では、送受信ノード間でのスイッチ段数で定義。特に静的網においてノード間距離の最大値を直径とよぶ。
2. 次数 (degree)
   静的網のノード一つ当りに直接結合されるノード数。
3. 総スイッチ/総リンク数
   静的網でのリンク総数、動的網でのスイッチ総数。

4. 拡張性 (scalability)
   規模の小さなシステムから、大きなシステムまで順次拡張できることが望ましい。

5. ３次元実装の容易性
   大規模システムでは３次元空間での実装が大きな問題となる。

6. 耐故障性 (fault tolerance)
   ノードやリンクに故障が生じた場合に代替装置が必要。

7. 多様な網の埋め込み能力 (embedability)
   並列処理の応用プログラムはその論理構造に適合した特定のトポロジ上で最も高速に実行できる。したがって、静的網が多様な網を包含できれば、より多様な応用に効率よく対応できる。

8. ルーティングの容易性
   送信元と受信先が与えられたとき、相互結合網内のルータやスイッチを高速に制御し、メッセージを受信先にとどける必要がある。

2.3 基本通信パターン

以下、送受信ノードの数をともにNとし、送信ノード番号X、受信ノード番号Yの2進表示をそれぞれ$(a_n, ..., a_2, a_1), (b_n, ..., b_2, b_1)$とする。

1. エクスチェンジ置換
   $y = E_i(x) = (a_n, ..., \bar{a}_i, ..., a_1)$

2. シャフル置換
   $y = \sigma(x) = (a_{n-1}, a_{n-2}, ..., a_1, a_n)$

3. バタフライ置換
   $y = \beta(x) = (a_1, a_{n-1}, ..., a_2, a_n,)$

4. ビット逆転置換
   $y = \rho(x) = (a_1, a_2, ..., a_n)$

5. シフト置換
   $y = \alpha(x) = |x + 1|_{2^n}$

6. 恒等置換
   $y = i(x) = (a_n, ..., a_2, a_1)$

7. 逆置換
   置換 πに対してその逆置換をπ^{-1}を、
   $\pi\pi^{-1} = \pi^{-1}\pi = i$
   で定義する。

3 静的網

以下、ノード数をN，リンクは双方向転送が可能であるとする。

3.1 完全網

直径は1、次数は$N - 1$, リンク総数は全体で$N(N - 1)/2$。大規模システムの実現は困難。

3.2 スター網

中央のノードに他のすべてのノードが結合している。中央ノードの次数は$N - 1$で他のノードの次数はすべて1。

3.3 リング網系

- 単純リング網
  シフトレジスタで構成できる。ベルトコンベヤ上にスロットが回転しており、空スロットにデータを乗せて転送する。直径は$N/2$、次数は2である。

- 有限リング網
  リング網の各ノードから、$2^m(m = 1, 2, \ldots)$だけ離れたノードにバイパスを張ったもの。

- 階層リング
  リングを階層的に結合したもの。

3.4 トーラス網系

1. 2次元メッシュ網
   メッシュ網では隣接する四つのノードと通信できるが、最左と最右ノード、最上と最下ノードは結合されない。直径は$2\sqrt{N}$、次数は4である。

2. 2次元トーラス網
   2次元メッシュ網の最左端と最右端、最上端と最下端同士のノードを結合したもの、直径は$\sqrt{N}$、次数は4である。

3. 3次元トーラス網
   直径は$\sqrt[3]{N}$、次数は6。

4. ピラッミド網
   二次元メッシュを階層的に構成する。最下位層がhとし、層iで2^i× 2^iメッシュとすると、
   $N = 1 + 2^2 + \ldots + 2^{2i} + \ldots + 2^{2h}$
   $= (4^{h+1} - 1)/3$
   次数は7、直径は$2h$で$\log 3N - 2$である。

5. 再帰トーラス網(RDT,Recursive Diagonal Torus)
   下位のトーラス網から上位のトーラス網を再帰的に構成する。(JUNP-1で採用)

3.5 トリー網系

1. 単純トリー網
   次数は3、直径は$2\log N$。ルートノード付近の負荷が重く、故障に弱いが、VLSIチップに効率よく埋め込むことができる。

2. Xトリー
   木の各レベル内に部分的な鎖状結合を施し、バイパスによって高速化と耐故障性の向上を図ったもの。

3. ファットトリー (fat tree)
   トリーのルートノードに行くにしたがって、リンクの転送能力を増加させる方式。

3.6 ハイパーキューブ網系

1. ハイパキューブ網
   超立方体、2進n-キューブともよぶ。2進n-キューブは$N(= 2^n)$個のノードからなる。次数、直径はともに$\log N$である。

2. ベースmn-キューブ
   2進n-キューブを拡張した網である。$N = m^n$である。次数、直径ともに$\log_m N$である。

3. 環立方体網(CCC,Cube Connected Cycle)
   2進n-キューブの各ノードをn個のノードからなるリングで置き換えたもので。全ノード数Nは、$n2^n$となっている。次数は3、直径は$3n/2$である。

4. CCT cube網
   2進m-キューブをサブネットとして、それらを完全網で結合したシステムである。サブネット内では、ハミング距離がmであるノード間にバイパスが張られ、サブネット内での直径を$m/2$としている。上位の完全網へは、サブネットの各ノードから一つづつリンクが張られる。全ノード数Nは2^{2m}、次数は$m + 2$、直径は$m + 1$である。

3.7 シャフル網系

以下のものがある。

1. シャフルエクスチェンジ網
2. de Bruijn網
3. 循環オメガ網 (circular omega)
4. Ster Graph網

4 動的網

動的網では、送受信ノード間には、スイッチ群が設定されており、各スイッチを制御することによって、所定のノード間の通信が可能となる。以下、送受信の各ノード数を $N(=2^n)$ とし、送受信ノード間の対応関係を通信パターンとよぶ。

4.1 クロスバ網

クロスバ網は閉塞 (block) なしで任意の時刻に任意の送受信ノードを高速に結合できる。通信パターンは $N!$個、送受信ノード間の通信に必要な時間 (通信時間) はスイッチ段数で 1 である。スイッチの個数は N^2となり、Nが大きいほど実現が困難。

4.2 多段結合網

1. 非閉塞網
   通信パターンは $N!$である。また、任意の送受信ノード間に新な通信要求や変更があった際に、すでに接続中の他の送受信ノード間通信のためのスイッチ設定を変更しないで常に要求が満たされる。

   - クロスバ網は一段の非閉塞網になっている。
   - 3 ステージ網は $m \geq 2n-1$ のとき非閉塞網となる。

2. 再構成型非閉塞網
   再構成型非閉塞網では、任意の送受信ノード間の結合要求 (総個数 $N!$) を満たす。しかし、新しい接続要求や変更が起こった場合、接続中の他の送受信ノード間のスイッチ設定を変更する必要が生じる。

   - Benes 網
     N 送受信ノードを有する Benes 網 (以下、NBense 網と略記) は送信側および受信側の交換スイッチと二つ $N/2$Bense 網を結合して得られる。交換スイッチは 2 入力 2 出力で直通と交換の二つの動作モードが設定可能である。
   - バイトニックソータを用いた結合網
     バイトニックソータで、受信ノード番号に基づいてソーティングすることで、所定の送受信ノード間の通信が実現できる。
   - 3 ステージ Clos 網
     $2n-1 > m \geq n$ のとき再構成型非閉塞網となる。

3. 閉塞網
   閉塞網では、可能な通信パターン総数 $N!$のうちごく一部の結合を実現するものである。$N/2$ 個の 2 入力 2 出力交換スイッチを $\log N$ 段連結した閉塞網で可能な送受信結合パターン数は、$(\sqrt{N})^N = 2^{(N \log N)/2}$であり、全結合パターン数 $N!$と比べるとごくわずかである。(例 :$N=8$ のとき $8! = 40320$ に対して、$2^{12} = 4096$ である。)

   - オメガ網
     オメガ網はシャフルエクスチェンジを $\log N$段連結してえられる。オメガ網は、$\Omega = (\sigma E)^n$ と表せる。
   - 間接 2 進 n-キューブ
     先に述べた n-キューブ網を多段結合網で実現したもの。間接 2 進 n-キューブは、
     $C = E\beta_2 E\beta_3, \ldots, E\beta_n E\sigma^{-1}$で表わされる。
   - バンヤン (Banyan) 網
     バイヤン網は、
     $Y = E\beta_2 E\beta_3, \ldots, E\beta_n E$

と表され、間接2進キューブの最終段の逆シャフル置換がないものである。

- ベースライン (baseline) 網
  ベースライン網は、
  $B = E{\sigma_n}^{-1} E{\sigma_{n-1}}^{-1}, \ldots, E{\sigma_2}^{-1} E$
  で表され、Benes 網の前半部分に相当する構成となっている。

4.3 単一結合網

多段結合網の1ステージ分のスイッチからなる。送受信ノード間で通信を行う際には単一段結合網を繰り返し利用することになる。静的網でノードがスイッチで構成されているものも単一段結合網とみなせる。したがって静的網を単一段結合網、動的網を多段結合網とよぶ場合もある。

参考文献

[1] 富田 眞治 , 並列コンピュータ工学 , 昭晃堂 (1996)

[2] 富田 眞治 , コンピュータアーキテクチャ1 , 丸善 (1994)