■6 群（コンピュータ －基礎理論とハードウェア）－ 5 編（コンピュータアーキテクチャ（II）先進的）

9 章 インタコネクション技術

（執筆者：天野英晴）［2010 年 4 月 受領］

■概要■

並列計算機の構成要素間を接続するネットワークを相互結合網あるいはインタコネクションと呼ぶ．このネットワークは，他のコンピュータネットワークとは異なり，高速，高信頼性，軽量なパケット転送技術，メッシュ，トーラス，Fat Tree，ハイパーキューブなどの接続トポロジーが発達した．現在，このインタコネクション技術は，クラスタや，大規模マルチプロセッサ用のネットワークだけではなく，チップ内のネットワークすなわち NoC（Network-on-Chip）として利用され，活発な研究分野となっている．

ここでは，インタコネクション技術を直接網，間接網，パケットスイッチ技術に分けて概観する．

【本章の構成】

9-1 節はインタコネクション技術の概観であり，全体的紹介を行う．9-2 節は構成要素間をリンクで直接接続した直接網（分散網）について，その接続トポロジーやそれぞれの性質を概観する．9-3 節は構成要素間をスイッチを用いて接続する間接網（集中網）について，その接続トポロジー，性質等を概観する．

最後に 9-4 で，相互結合網ならではのワームホールルーティング，仮想チャネル，適応型ルーティングなどのパケット転送技術をまとめる．

■6群 － 5編 － 9章

9-1 相互結合網の概観

（執筆者：吉永　努）［2008年10月 受領］

相互結合網とは，複数のデバイスを相互に結合し，デバイス間での通信を実現するためのネットワークのことをいう(**図 9･1**)．相互結合の対象となるデバイスには，コンピュータ，コンピュータの構成要素となるプロセッサやメモリのほか，電話機など様々な機器が含まれる．ただし，狭義には高性能な並列計算機の通信サブシステムを相互結合網と呼ぶことがある．

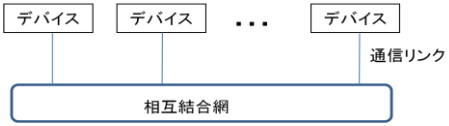


図 9･1　相互結合網の概観

以下に，相互結合網のトポロジーによる分類，データ交換方式，性能指標について述べる．

9-1-1 ネットワークトポロジーよる分類

相互結合網の形状のことをネットワークトポロジーと呼ぶ．相互結合網は，そのネットワークトポロジーによって**図 9･2** に示す四つに分類できる．共有媒体網は，初期の Ethernet（10 BASE-2 や 10 BASE-5）LAN などに用いられた方式で，1 本のバスを複数のデバイスが共有するネットワークを指す．直接網は，相互結合される個々のデバイスがルータを内蔵し，デバイス同士が通信リンクを介して直接接続されるネットワークである．これに対して，間接網では相互結合されるデバイスがルータを内蔵せず，デバイス同士は間接的にスイッチと通信リンクを介して接続される．直接網と間接網については，それぞれ 9-2 節と 9-3 節で説明する．ハイブリッド網は，上記三つの網に含まれる方式を組み合わせて実現されるもので，例えば直接網であるハイパーキューブの各次元をクロスバスイッチで接続するハイパクロスバ網などが提案されている．

■共有媒体網
　•共有バス
■直接網（9.2参照）
　•メッシュ，トーラス，ハイパーキューブなど
■間接網（9.3参照）
　•クロスバ，Clos，バタフライなど
■ハイブリッド網
　•ハイパクロスバ，ハイパメッシュなど

図 9･2　ネットワークトポロジーによる分類

9-1-2 データ交換方式

相互結合網における通信データの交換方式は，回線交換とパケット交換に大別することができる．回線交換はコネクション型通信とも呼ばれ，コネクションの確立，データ通信，コネクションの解放の三つのフェーズで動作する．コネクションの確立は，データの送信元デ

バイスから宛先デバイスまでのネットワーク中の通信経路を設定する．その後，データ通信を行うが，データ通信中は回線が占有されルーティングは必要ない．そのため大規模なデータ通信に適するが，他の通信と回線上のネットワーク資源を共有できない，小規模なデータ通信に対してはコネクション確立時間がオーバヘッドとなる，などのデメリットも存在する．回線交換の例として，アナログ電話の通信方式があげられる．

パケット交換は，通信データをパケットと呼ぶ単位に分割して通信を行う方式である．個々のパケットが宛先デバイスのアドレス情報を有し，通信経路上のルータで中継処理が行われる．データ通信に先立つコネクション確立を行わなくてもよいため，コネクションレス型通信を実現することができる．また，一つの回線を複数のパケットが時分割利用することができる．パケット交換の例として，インターネットプロトコルの通信方式があげられる．パケット交換技術については，9-4 節で詳しく述べる．

9-1-3 相互結合網の性能指標

相互結合網の代表的な性能指標として，通信遅延とバンド幅（スループット）が用いられる．通信遅延は，データの送信元デバイスが通信を開始してから宛先デバイスが受信を完了するまでの時間を表す．**図 9･3** に一般的な相互結合網に与える通信負荷を変化させたときの通信遅延の特性を示す．通信遅延は通信距離に依存するため，相互結合網の平均遅延で示す場合が多い．無負荷遅延（zero-load latency）は，ネットワーク中で通信衝突が発生しないときの遅延の下限値を表す．遅延は通信負荷が高くなるにつれて通信衝突が増えることによって上昇し，通信負荷 T_s において無限大となる．この T_s をネットワーク飽和スループットという．

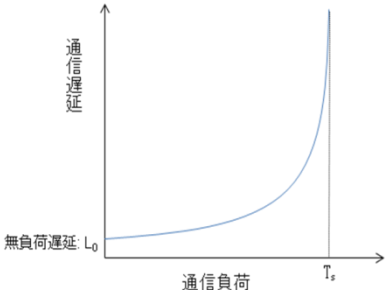


図 9・3　通信遅延

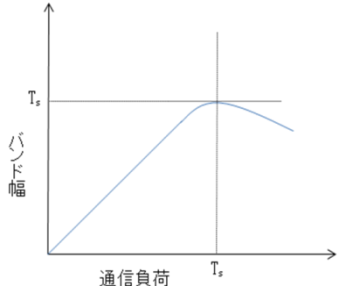


図 9・4　バンド幅

図 9･4 に一般的な相互結合網に与える通信負荷を変化させたときのバンド幅の特性を示す．バンド幅は，単位 bits/s などで表す単位時間当たりのデータ転送能力を表す．バンド幅には，通信リンクのバンド幅，受信デバイスで計測する受信バンド幅，ネットワークを最少の通信リンク数で 2 等分したとき，その分割面を通る通信リンク群の合計バンド幅を示す二分バンド幅などがある．

図 9･4 は横軸が送信デバイス当たりの送信負荷，縦軸が受信デバイス当たりの受信バンド幅を表す．T_s 未満の通信負荷ではほぼそれに比例したバンド幅が得られるが，通信負荷が T_s を超えるとネットワークが過負荷な状態となりバンド幅が損なわれることがある．

■参考文献

1) 富田眞治, “並列コンピュータ工学,” 昭晃堂, 1996.
2) 天野英晴, “並列コンピュータ,” 昭晃堂, 1996.
3) 情報処理学会編, “情報処理ハンドブック,” オーム社, pp.407-415, 1989.
4) J. Duato, S. Yalamanchili, and L. Ni, “Interconnection Networks,” IEEE Computer Society Press, 1997.
5) W. Dally and B. Towles, “Principles and Practices of Interconnection Networks,” Morgan Kaufmann Publishers, 2004.
6) T. M. Pinkston and J. Duato, “Interconnection Networks,” Appendix E in “Computer Architecture: A Quantitative Approach,” 4th Edition, Morgan Kaufmann Publishers, 2006.

■6群 － 5編 － 9章

9-2 直接網

（執筆者：鯉渕道紘）［2009年12月 受領］

直接網は，個々のデバイス同士を直接リンクで結合するネットワークのことである．デバイスは通常，ルータを内蔵することで複数のリンクを結合することができるため，直接網では様々なトポロジー，ルーティングを採用することができる．

直接網は，以下の3点により特徴付けられる．

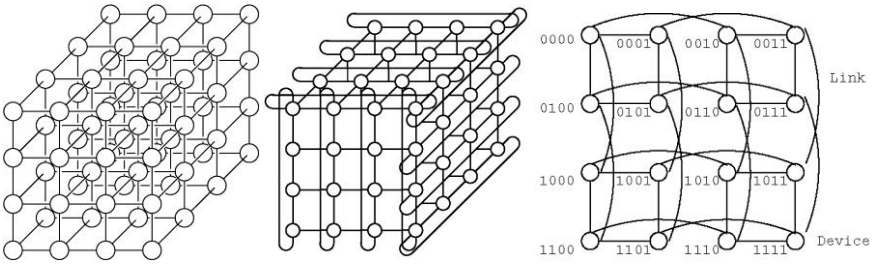
- **トポロジー**：理想的なトポロジーは各デバイスがすべてのデバイスと直接つながっている完全結合であるが，実装面の点から実現することが困難である．そこで，性能面と実装面のトレードオフの議論から多数のトポロジーが提案されている．
- **ルーティング**：通信データは，パケットに分割され転送され，多くの場合，複数のデバイスを経由して目的地に到達することになる．ルーティングが，その中継するデバイス群を決定する．
- **スイッチング**：パケットがデバイス間を転送される場合に，スイッチングによって経路をパケットに割り当てる（9-4節で詳細を説明する）．

9-2-1 トポロジー

トポロジーは，隣接デバイスに接続しているリンク数である次数（degree），デバイス間の最大距離である直径（diameter），対称性（symmetry）などで特徴付けられる．

図9･5に代表的なトポロジーであるメッシュ（mesh），トーラス（torus），ハイパーキューブ（hypercube）の例を示す．メッシュはネットワークの端にあるノードとそうでないノードとでは，隣接するノード数が異なる．

メッシュにおいてすべてのノードに対して隣接するノード数が等しくなるように，端のノード間を結んだトポロジーをトーラスと呼ぶ．トーラスは各次元のノードをリングで結合した構造をもつ．ハイパーキューブはノードを n 桁の2進数で表現し，それぞれの桁が1ビット異なるもの同士をリンクで結ぶ．図9･5は16ノードの例なので，例えばノード0101は1101，0001，0111，0100の四つのノードとの間にリンクをもつ．次数は2進数で表したときの桁数に等しいので，全体のノード番号を N とすると $n = \log_2 N$ となる．



(a) 3次元メッシュ　(b) 3次元トーラス　(c) ハイパーキューブ

図9･5　トポロジーの例

9-2-2 ルーティング

ルーティングは，パケットを目的地まで配送する経路を定める．ただし，ルーティングを定めるアルゴリズム（ルーティングアルゴリズム）が，無制限にパケットに対して非最短経路がとることを許すと，パケットが永久に宛先に到達できない状況であるライブロックを引き起こす．更に，ルーティングアルゴリズムの設計は，デッドロックを避けることが要求される．

デッドロックとは，ネットワークを通過中のパケットが，起こる可能性がない事象を待ち続けることにより，転送することが不可能となる状態のことをいう．

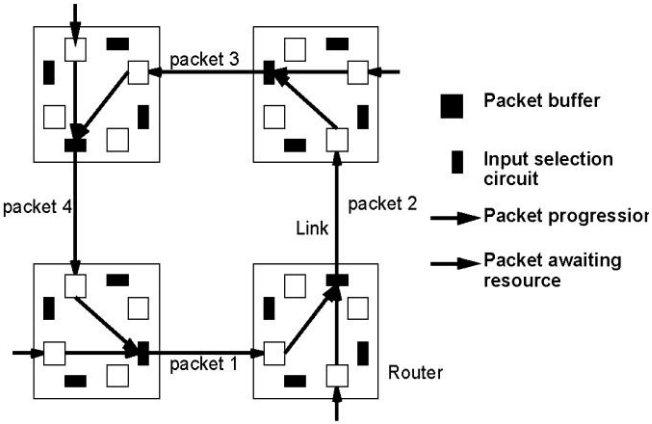


図 9・6　デッドロックの例

デッドロックが生じるのは, スイッチのチャネルバッファ間に循環依存があるためである．**図 9･6** にデッドロックの例を示す．図 9･6 では，四つのパケットがそれぞれ行き先のパケットバッファが空くのを待っているが，互いにバッファを占有し合い，動きが取れなくなっている．デッドロックは，ネットワーク内でパケットの紛失がほとんど生じない相互結合網で生じる現象であり，古典的なイーサネットのようにスイッチのバッファオーバフローを避ける制御をしないネットワークでは発生しない．前者をロスレス（lossless）ネットワーク，後者をロシー (lossy) ネットワークと呼ぶ. デッドロックはネットワークトラフィックが多く，また，ネットワーク資源が少ないほど，その発生確率が高くなる．

ライブロックは送信元から宛先までの最短経路だけを利用する，あるいは，非最短ホップ数を制限するルーティングを利用することで防ぐことができる一方，デッドロックの対応策はやや複雑で，デッドロック回復方式とデッドロックフリー方式の二つがある．

デッドロック回復方式はデッドロックが発生した場合，abort-and-retry（パケット破棄，再送）やパケット破棄せずにネットワーク資源を割り当てし直す方法により，パケット転送を保証する方法である．ただし，デッドロックが頻繁に発生すると性能の低下が大きくなるという問題がある．一方，デッドロックフリー方式はデッドロックが発生しない経路群を設定することでパケット間のデッドロックの発生を防ぐ方法である．

以上より，ライブロック，デッドロックの問題に対処しつつ，トポロジーの規則性をでき

る限り利用して経路を分散させ，最短経路をなるべく取れるルーティングアルゴリズムが一般的に並列計算機などの直接網で採用されている．

■6群 － 5編 － 9章

9-3 間接網

（執筆者：横田隆史）［2008年10月 受領］

9-3-1 間接網の定義

デバイス間の通信を，それらとは独立に存在するスイッチにより行う方式において，スイッチ及びスイッチ間のリンクの集合を指す．演算ノード間，演算ノード・メモリモジュール間，演算ノード・入出力間（ファイルなど）を接続するものとして多く用いられる．

通常，各デバイスは一対の双方向通信リンクにより結合網と接続される．個々のデバイスにおけるピーク通信性能はこのリンクの物理的転送速度により規定されるが，結合網全体の通信性能はトポロジーや通信方式により大きく変わる．

間接網におけるデバイス間の通信は，本質的に1以上のスイッチを経由する．間接網における距離は，経由スイッチの数により測られる．

間接網では多くの場合，入力から出力に至る単方向性の接続を基本に考える．双方向性の通信は，単方向通信の重畳として考えることが多い．

9-3-2 間接網の代表的なトポロジー

(1) クロスバ（crossbar）

入力ポート数 N, 出力ポート数 M に対して $N \times M$ のスイッチマトリクスを構成することで得られる接続方式である（**図 9･7**）．複数の入力が同時に同一の出力を目指さない限り，どのような入力･出力の組合せに対しても距離1で通信が可能になる．このため，低レイテンシ，高スループットが強く要求されるシステムにおいて多く用いられる．しかし必要なハードウェア資源が接続ポート数の2乗に比例することよるスケール性の欠如が欠点とされる．

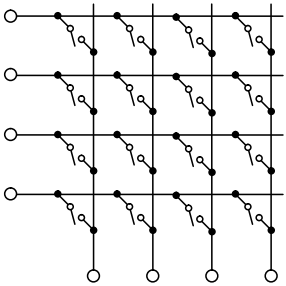


図 9･7　クロスバスイッチ

(2) 多段網（Multistage Interconnection Network：MIN）

多段網は，比較的小規模のスイッチモジュールを一列に並べた段（stage）を複数設け，段間を適切な方法により接続することで構成される結合網である．各スイッチモジュールはクロスバスイッチ及びパケット保持用のバッファ（必要に応じて）で構成され，その入出力ポート数を基数（radix）と呼ぶ．

直接網など他の形式の結合網にもいえることだが，結合網では，いかなる入力・出力の組

に対しても情報を伝搬できる到達性を保証する必要がある．これに加えて間接網では，それぞれ独立な複数の入力ポートからの情報が，重複しない複数の出力ポートに同時に接続される状況を考えねばならない．ここで，要求されたすべての入力・出力の組が同時に実現可能か否かにより，ブロッキング網／ノンブロッキング網に分類される．

代表的なノンブロッキング網として，中間段の構成により多重経路を可能にする Clos 網があげられる（**図 9･8**）．ブロッキング網では多くの場合多重経路を許さず，入力と出力の組合せにより経路が唯一に決まる．多くの場合，同一基数のスイッチモジュールを用い，段間の接続方法による構成が定式化される．代表的なものとして，パーフェクトシャッフル（perfect shuffle）接続を用いるオメガ（Omega）網やデルタ網（Delta），キューブ（cube）接続を用いる間接キューブ網（indirect cube）などがある．また，これらをまとめて交換網（permutation network）と呼ぶこともある．

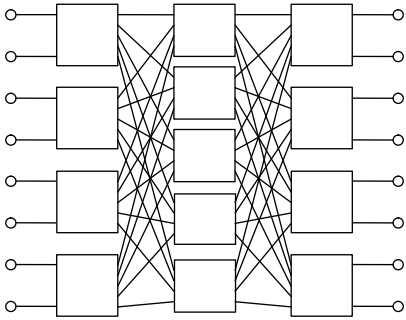


図 9・8　Clos 網の例

図 9･9 に多段網の一般化表現を，8 ノード Omega 網を例に示す．この例では基数 2 のスイッチを 3 段（G_0～G_2）用い，これらを段間接続（C_0～C_3）により接続している．

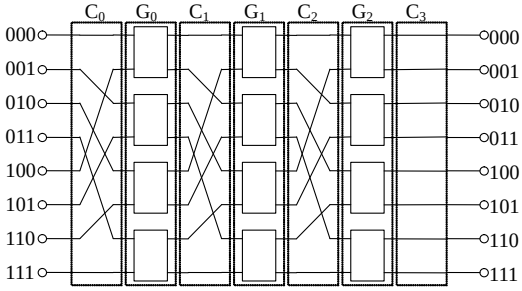


図 9・9　間接網の一般化表現（Omega 網の例）

かつてはスイッチデバイスの実装や配線上の問題から基数 2 程度のスイッチモジュールが用いられるケースが多かったが，最近では高速シリアル通信技術の発展や半導体デバイスの微細化により高基数（high radix）のものも多く出現している．

(3) 不等距離間接網

代表的なものとして tree，fat tree，ハイパクロスバ網があげられる．tree 網は *n* 進木構造の葉（leaf）部にデバイスを，ノード部にスイッチモジュールを配したものである．fat tree 網は，葉から根（root）に向かうに従って転送路の容量を増すことで根付近のボトルネックの解消を図ったものである．多段網の通信を単方向から双方向に拡張し，途中の段での通信路の折り返しを許すことにより不等間隔間接網（双方向多段網）が得られる．fat tree 網は，こうした双方向多段網により構成される事例が多い．

(4) その他，不定形網

柔軟なスイッチモジュール間接続を許容することで，不定形の網構成をとることも可能である．現実には使用するスイッチモジュール（多くの場合ファブリック（fabric）と呼ばれる）により Mytinet，Infiniband，Fibre Channel などとして商用化されている．

9-3-3 通信方式

(1) データ交換方式

データ交換方式に関して間接網のトポロジーに依存する要素はほとんどなく，9-1-2 節に示した回線交換，パケット交換（9-4 節）が用いられる．

(2) 制御方式

入力・出力ポート間の接続を結合網全体で一斉に切り替えるか（同期制御），あるいは入力ポートからの要求に応じて随時切り替えるか（非同期制御）の制御方法がある．システム全体が SIMD 動作する並列計算機においては，同期制御方式の親和性がよいといえる．MIMD 動作の並列計算機でも結合網の制御の煩雑さを避ける目的により同期制御を採用する例もある．

9-3-4 その他

(1) 性能向上手法

間接網での性能向上手法として，トポロジーの工夫，ルーティング手法，アプリケーションへの適応があげられる．

トポロジー上の工夫では，クロスバスイッチを最大性能の指標とし，できるだけ少ない資源で最大性能に近づける試みがなされている．入出力ポート間の経路の多重度を増す方法が主に使われる．

入出力ポート間に多重経路が存在するトポロジーでは，適応ルーティングにより性能向上を図ることができる．代表的な例として Thinking Machines 社 CM-5 での fat tree 網がある．

共有メモリ型の並列計算機では，同期用の共有変数を介しての同期操作などの際に特定のデバイスにアクセスが集中する場合がある．これにより結合網内においてアクセスが集中した箇所を基点として樹状に輻輳が拡大する樹状飽和（tree saturation）をまねき，網全体の性能を低下させる．このため，同期操作で多く用いられるアトミックな操作に対応するパケットをスイッチモジュールの中でひとつにまとめるコンバイニング（combining）が提案されている．また，マルチキャストや集合通信（collective communication）に適した制御方法も議論

されている．

(2) 耐故障手法

耐故障性は，基本的には多重経路を提供することにより実現される．例えば Clos 網や fat tree 網では中間段のスイッチにおいて多重経路が存在する点において耐故障性があるが，一方，デバイスと直接接続する段では冗長性がなく耐故障性に欠ける．スイッチ間のリンクの増加や，段の追加などにより耐故障性を増す手法が提案されている．

■6群 － 5編 － 9章

9-4 パケットスイッチ技術

（執筆者：松谷宏紀）［2008年10月 受領］

9-1-2項で述べたとおり，通信データの交換方式は回線交換とパケット交換に大別される．パケット交換では，通信データはパケットと呼ばれる単位に分割され，一つの回線上を複数のパケットが時分割で利用することができる．ここではパケット交換について詳しく述べる．

9-4-1 パケットの構造

一般的なパケットの構造を**図 9･10** に示す．図のように，パケットの先頭にはパケットを識別するヘッダが付与され，その後ろにデータ本体が格納されるボディ部が続く．ヘッダには，宛先デバイスのアドレス，送信元デバイスのアドレス，パケット長などが格納される．

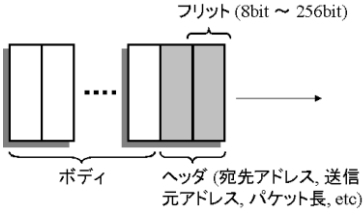


図 9・10　パケットの構造

通常，デバイス間のリンクは8 bitから大きいもので256 bit程度のデータ幅をもつ．このようなリンクにおいて，1 サイクルで転送されるデータの単位をフリットと呼ぶ．図のように，パケットは複数のフリットから構成され，リンク上をサイクルごとに1フリットずつ転送される．

9-4-2 パケットの転送方式

パケット交換では，パケットは通信経路上のノード（ここではスイッチとデバイスの組をノードと呼ぶ）で中継されながら宛先デバイスまで配送される．パケットを中継するノードは一定量のパケットバッファを有し，パケットの受信（バッファリング），及び，転送を繰り返す．パケットの中継処理は，受信及び転送処理の粒度とタイミングに応じて，以下の4種類の転送方式に分類される．

(1) ストアアンドフォワード（store-and-forward）方式

各ノードはパケット全体を格納できる大きさのバッファを有し，**図 9･11** に示すように，パケット全体を受信してからそのパケットを次のノードに転送する．

各ノードにおいて，パケット全体を受信するまで待ってから次のノードに転送するという処理を繰り返すため，ストアアンドフォワード方式は，以下に示すワームホール方式やバーチャルカットスルー方式に比べて通信遅延が大きいという欠点がある．それでも，制御が容易という理由から，インターネットプロトコルではストアアンドフォワード方式が広く用いられている．

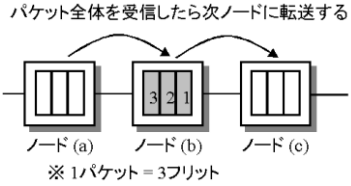


図 9・11　ストアアンドフォワード方式

(2)　ワームホール（wormhole）方式

各ノードはヘッダフリットを格納できる大きさのバッファを有し，**図 9･12** に示すように，フリットを受信するたびにそのフリットを次のノードに転送する．通信経路上のノードのバッファが空いている限り，パケットが複数のノードにまたがって移動する様子はさながらイモムシ（ワーム）の動きに似ている．

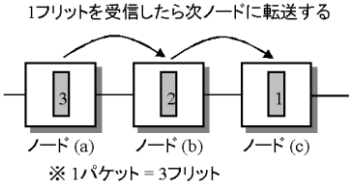


図 9・12　ワームホール方式

各ノードは，パケット全体の受信を待つことなく，受信したフリットを即座に次ノードに転送できるため，ストアアンドフォワード方式に比べて通信遅延は小さい．また，必要なバッファ量も少なく実装コストが小さい．しかし，パケット同士の衝突が起きると，複数ノードのバッファを占有したままの状態で待ちが発生する（このような状況をヘッドオブラインブロッキングと呼ぶ）ため，通信性能が悪化してしまう．ただし，この問題は仮想チャネルによって軽減できる．

(3)　バーチャルカットスルー（virtual cut-through）方式

各ノードはパケット全体を格納できる大きさのバッファをもつが，ワームホール方式と同様に，フリットを受信する度にそのフリットを次のノードに転送できる．

各ノードがパケット全体を格納できるバッファをもつため，ヘッドオブラインブロッキングが起きると，すべての後続のフリットは，先頭フリットがいるノードのバッファに吸い込まれるようにして格納される．そのため，ワームホール方式のように経路上のノードのバッファを占有したまま待ちが発生することはない．しかも，通信遅延はワームホール方式と同様に小さいため，並列計算機ではバーチャルカットスルー方式がよく用いられている．

(4)　サーキットスイッチング（circuit switching）方式

まず，セットアップフリットを宛先まで送ることで通信経路を確保する．そのうえで，ボディ部をノンブロッキングで転送する．この方式はどちらかというと回線交換方式に近い．

9-4-3 仮想チャネル（virtual channel）

ワームホール方式などのパケット交換では，ノードはリンクごとに最低 1 セットのバッファをもっていれば通信を行うことができる（リンクとバッファの組を物理チャネルと呼ぶ）．

一方，単一のリンクに対し，バッファを複数セットもたせることで仮想的に複数の物理チャネルがあるかのように見せることができる．このようにして多重化されたチャネルを仮想チャネルと呼ぶ．同一物理チャネル内の複数の仮想チャネルは，単一の物理リンクを時分割で利用する．

(1) ヘッドオブラインブロッキングの回避

仮想チャネルをもたせることで，ワームホール方式の欠点であるヘッドオブラインブロッキングを軽減できる．**図 9･13** 中にはパケット A，B，X が図示されている．パケット B の目的地はノード(d)であるにもかかわらず，パケット X が移動するまでパケット B は移動することができない．

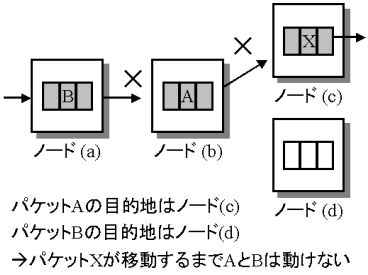


図 9・13　ヘッドオブラインブロッキング

一方，**図 3･14** ではノード(b)に 2 本の仮想チャネルを割り当てている．これによって，パケット B はノード(b)の 2 本目の仮想チャネル（バッファ）を用い，パケット A を追い越して宛先にいち早く到達することができる．

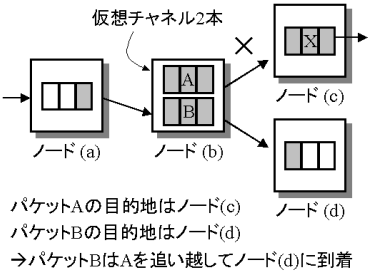


図 9・14　仮想チャネルによるブロッキング回避

(2) デッドロックの回避

9-2-2 項で述べたようにノードのバッファ間に循環依存があると，パケット転送のデッドロ

ックが発生し永久にパケットを転送できなくなる．仮想チャネルによって一つの物理チャネルに複数セットのバッファをもたせることができるので，循環依存が生じないようにバッファを使い分ければデッドロックを回避できる．このような手法はトーラスでデッドロックを回避する場合など様々な用途に応用されている．

9-4-4 ルータの構造

図 9･15 に *n* 本のチャネルをもつルータの構造を示す．入力パケットは入力バッファにてバッファリングされ，宛先アドレスより出力チャネルを決める（経路計算）．その後，クロスバのアービトレーションを行い，クロスバを経由してパケットを出力チャネルに転送する．

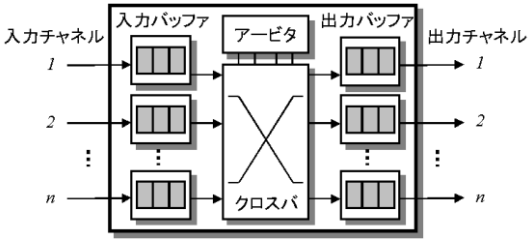


図 9・15　ルータの概略図