オペレーティング・システム 56−13
(1992. 8. 19)

実時間用相互接続網の評価

戸田賢二* 西田健次 内堀義信 島田俊夫

電子技術総合研究所†

概要

並列実時間処理システムの相互結合網としてどのような構成と制御機構が適するかを評価する．まず，クロスバと多段網について幾つかの制御方式について最悪通信遅延を検討し，優先度制御を行なうには多段網の構成が適していることを示す．多段網の制御方式としては，優先度フィールドが小さいときは優先度別バッファ方式が，ある程度大きい場合は優先度先送り方式が適している．さらに，優先度先送り方式で古いパケット優先としたときの性能，仮想チャネルの技法の導入方法及びその効果，優先度キューの実現コストを示す．

Evalutaion of Real-Time Interconnection Networks

Kenji TODA Kenji NISHIDA Yoshinobu UCHIBORI
and Toshio SHIMADA

Electrotechnical Laboratory

Abstract

Network topology and control schemes are evaluated in terms of interconnection networks for real-time parallel processing. The worst case delay of a crossbar and a multistage network (Omega) is discussed. The discussion draws the advantage of multistage network with a proposed priority forwarding scheme for wide priority field. Its performance on oldest packet first policy is shown. Virtual channel flow control technique and implementation of a priority queue are also discussed.

*Email: toda@etl.go.jp

†〒305 つくば市梅園1-1-4、 Tel: 0298-58-5875 Fax: 0298-58-5882

1 はじめに

並列処理により実時間処理性能が飛躍的に向上することが期待されている．しかし，このためには演算要素間の相互結合網が実時間性をもつことが不可欠である．すなわち，スループットや平均通信遅延という通常の能力に加え，通信遅延に対する予測可能性が実現されなければならない．

本稿では，まずクロスバと多段網において様々な制御方式での最悪通信遅延を検討する．我々が提案している優先度先送り方式については，古いパケットを優先し通信遅延を一定化しようとした場合のシミュレーション結果を示す．次に，仮想チャネルの技法の導入コスト及びその効果について述べる．また，優先度キューの動作速度と実現コストを示し，優先度先送り方式の実現可能性を明かにする．

2 クロスバと多段網での最悪転送遅延の検討

2.1 クロスバ

クロスバは，N^2（N は網のサイズ）のハードウェア量が必要であるが，出力ポートが一致しない限り転送経路上でのパケット衝突が発生しない非ブロッキング網であるので，この制約の下では通信遅延が一定であることが保証されている．しかし，パケットの転送パターンに制約のない状況では，通信遅延は出力ポートの衝突に影響される．以下に各々の衝突時の調停方式における最悪転送遅延を述べる．

- ランダム，固定優先度
  意図したパケットが転送される保証がないため最悪ケースの転送遅延は∞となる．
- ラウンドロビン
  調停に要する時間が転送サイクルに加算され，最悪ケースで N 回の転送サイクルが必要である（全ての入力ポートから同一の出力ポートへのパケットが投入された場合）．
- パケットの持つ優先度により制御を行なう
  優先度判定に $\log N$ 段の優先度比較が必要である．先に述べたラウンドロビン回路及びこの優先度判定回路は出力ポート毎に必要となる．

ラウンドロビン方式での最悪転送遅延は N となるが，これが発生するのは全てのパケットが同一の出力ポートを目指した時であり，通信パターンの制御による改善は他のブロッキング網と比べ容易である．ただし，優先度制御を行なうための時間的，回路的コストは大きい．

2.2 多段網

多段網は，クロス等の非ブロッキング網，ベネス，ガンマ，冗長段を持つオメガ等の入力ポートから出力ポートへの経路が複数ある冗長性のある網，オメガ等の単一経路網等に分類できる [ICN]．多段網は，ハードウェアコストと通信性能の比に優れた構成を取り易く，例えばオメガ網の場合の全体の段数は $\log N$ であり，ハードウェア量は $N \log N$ である．従って，大規模な網を構成し易い反面，出力ポートが異っていても経路が交差する通信パターンでは，パケット同士の衝突が発生する．この調停は，多段網の場合は各段のルータで制御を行なえばよい．各調停方式毎の最悪遅延は次のようになる．

ここでの議論では，ルータの各入力ポートにバッファを用意し，衝突が発生した場合又は次段のルータのバッファが一杯で出力できない場合，転送されてきたパケットをバッファに格納するという制御を仮定する．なお，優先度制御を行なうときは優先度キューをバッファとして用いると仮定する．n，N は各々，ルータサイズ，網のサイズである．

- ランダム，固定優先度
  最悪通信遅延が∞となる．
- ラウンドロビン
  全ての入力ポートから同一の出力ポートへのパケット転送があった場合，$\frac{n(N-1)}{n-1} + N$ の通信遅延となる [PF1].
- パケットの優先度のみで各ルータ上で調停を行なう
  転送したいパケットの経路上により低い優先度のパケットが存在し，これが別のパケットとの調停に負け，結果として目的のパケットの転送がブロックされる優先度逆転の現象を招いてしまう．このため，最高優先度のパケットであっても最悪転送時間は∞となる．
- 低優先度パケットの消去
  次段のバッファが一杯であっても，そのバッファ中の最低優先度パケットより高い優先度のパケットの転送は受け付ける．この場合，転送されてきたパケットをバッファに格納するために，当該最低優先度のパケットを消去する．転送するパケットが，その経路上のパケットと比べ一番優先度が高ければ，パケットの通過時間は，網の段数＋1の転送時間となることが保証できる．

  バッファが一杯の場合，そこでの最低優先度の情報を前段に知らせる信号線が必要である．また，最高優先度以外のパケットは消去される可能性があるため，そのチェックと再送のコストは大きなものになる．

- パケットスワップ
  次段のバッファが一杯であっても，そのバッファ中の最低優先度パケットより高い優先度のパケットの転送は受け付ける．この場合，転送されてきたパケットをバッファに格納するために，送ろうとするパケットと送り先の最低優先度のパケットをスワップする．パケットを後方へ戻すための時間又は専用のデータ線が必要になる．専用線を用意した場合でも，スワップのためのパケットの書き戻しと同時には前段からの新たなパケットの入力を受け付けることは困難であり，スループットの低下を招く．

- 優先度別バッファの使用
  優先度毎にバッファを用意し，そこには対応する優先度のパケットしか格納しない．衝突時は優先度による調停を行なう [Kun89][Dal90][Raj91]．高い優先度パケットを低い優先度パケットがブロックしないため，優先度逆転を発生しない．優先度毎のバッファはFIFOキューでよいが，入力ポート毎に優先度の数のバッファが必要である．また，バッファの状態を前段へ知らせるため信号線が優先度毎に必要である（優先度毎のバッファの状態は，エンコーディング可能で，状態変化だけを伝達するようにすれば，一杯になったバッファの優先度と受け付け可能になったバッファの優先度を知らせるために $2\times\log_2 P$（ただし，P は優先度フィールド長）の信号線があれば良い（異なる出力ポートを選択した場合，複数の優先度バッファからの同時出力が可能な設計にすれば，その数だけ追加の信号線が必要になる）．

- 優先度先送り
  低い優先度のパケットの後ろでより高い優先度のパケットがブロックされていることをその原因となっているパケット（群）に知らせることにより，優先度逆転を解消する方法であり，いわゆる優先度継承 [Tok89] を多段網に応用したものである．優先度先送りには，複数の種類が可能であるが，ここでは最も単純な基本優先度先送り（先送りする情報は，優先度のみ，先送りの方向はその段のパケットの転送方向，先送りされた優先度はそのパケットの転送によって消去される）について説明する．

  次段のバッファが一杯でパケットがブロックされたら，そのパケットの優先度情報を次段に先送りする．次段では送られてきた優先度を自分のバッファ中の最高の優先度と比較し，送られてきた方が高ければ，その優先度を保持し，調停に用いる．その次ぎの段（バッファ中の最高優先度パケットの出力方向）のパケットのバッファが一杯であれば，さらにそれを，先送りする．以上の操作により，優先度逆転が解消される．

  優先度の先送りはパケットブロック時に行なわれ，パケット転送用のデータ線を共用できるため，ルータ間の信号線を増設する必要はない．ルータ内の追加回路は，先送りされた優先度優先度を保持するバッファが余分に必要な程度である．本方式による最高優先度パケットの最悪通信遅延は s^2+2s+1（s は網の段数でオメガ網では $s=\log_n N$）となる [PF1]．

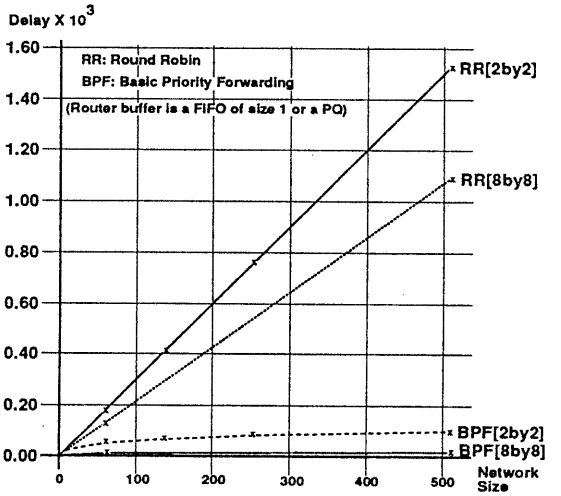

図 1: 最悪遅延の理論値

図1は，2×2と8×8のルータの場合についての理論値をグラフ化したものである．基本優先度先送り方式では，最悪遅延が $\log^2 N$ オーダーであり，大規模な網にも適用可能である．

3 通信遅延を一定化する場合の優先度先送り方式の評価

パケットに固定した優先度を与えた場合についてはすでに [PF1][PF2] で評価しており，優先度先送り方式は，様々な負荷条件の下でも優先度が高くなるに従い小さな通信遅延を安定して与えることを示した．ここでは，パケットの優先度として網での滞在時間を与え，滞在時間の長いものほど高い優先度を持つ，すなわち通信遅延を一定化する，場合についての評価を行なう．シミュレーションの仮定は優先度以外は上記の文献と同一である．

一様乱数を入力した時の2×2ルータのオメガ網のスループットを図2に示す．図中，“p2”，“p8”は各々，入力ポートのバッファが優先度キューであり，サイズが2，8であることを示す．RRよりBPFのスループットが高いのは，BPFによりパケットの詰まりが解消する効果があるためである．

図3と4は，8段のオメガ網（256×256）における転送遅延のヒストグラムであり，各々パケット発生確率0.3と0.4の場合である．発生確率0．3ではBPFの最大遅延はRRの1／2程度であるが，発生確率0．4では，BPFの最大遅延が0.3の時と比較し1.4倍程度であるのに対し，RRでは14倍となっている．これは，BPFとRRのスループットの差が主な要因である．

図5は，バッファサイズを8にしたときのものであり，同一パケットを16回続けて生成するという条件を入れた場合である．BPFはRRの4割程度の最大遅延を与えている．

以上の評価から，優先度先送り方式とラウンドロビン方式でそのスループットが同程度になるように網の実現ができるなら，優先度先送り方式の転送サイクルがラウンドロビンの2倍程度までは，通信遅延の保証の観点から十分許容できることがわかる．

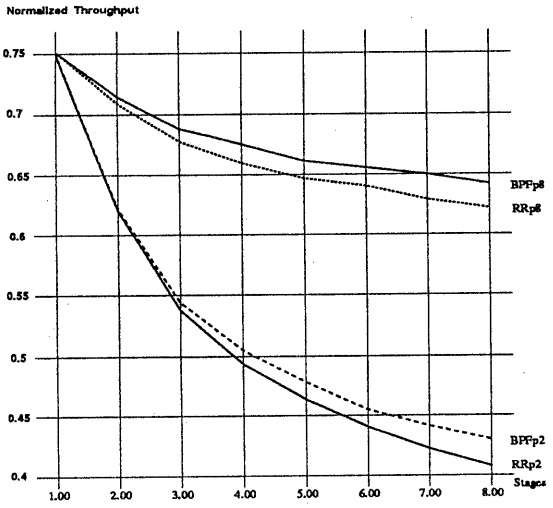

図2: オメガ網のスループット（パケット古いもの優先）

4 仮想チャネルと優先度制御方式

ルータの入力ポートバッファの通常の実現では，バッファの先頭のパケットがその出力先が一杯であるため待たされた場合，次にそれとは異なる出力先のパケットが到着してかつその出力先が一杯でなかったとしても，そのパケットは先に待たされているパケットにブロックされ出力が行なえずそのバッファで待たされることになる．これを解決するために，入力ポートバッファを出力先毎に分割したりする手法により特定方向への経路が詰まっている状況でも，別の方向への経路を確保する，チャネル仮想化が提案されている[Dal]．この仮想チャネルによりブロッキングが減少し経路の有効利用が行なえるためスループットの大幅な改善が行なえる．

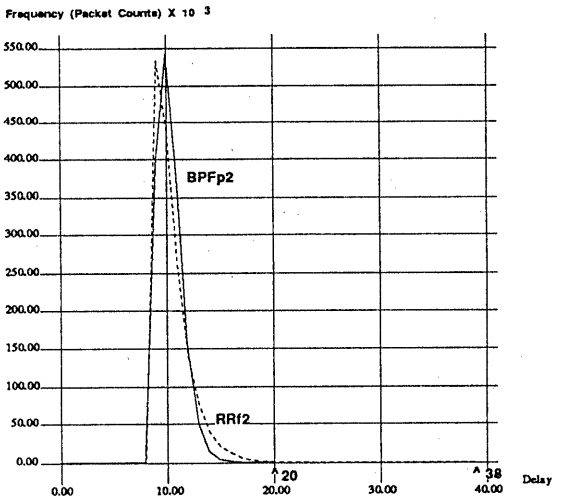

図3: 転送遅延ヒストグラム（8段オメガ網，バッファサイズ2，パケット発生確率0．3）

このための実現コストとしては，バッファ量の増大（n×nルータの場合，ルータ全体でn^2個必要）及び前段へバッファの状態を知らせるための信号線の増加（各バッファが独立して動作する，すなわち同時出力可能ならn本／ポート，同時には1パケットしか出力しないなら一杯になったバッファと一杯でなくなったバッファの2つのアドレスをエンコーディングして$2\log_2 n$本／ポート必要）がある．ただし，バッファの状態の信号線については，主力先毎に状態を知らせず，どれかの出力バッファが一杯になったら，そのポート全ての入力を禁止する簡略化も（スループットの低下を招くが）可能である．

優先度別バッファ方式と仮想チャネルの併用は，バッファ数が膨大に必要になり困難である．ただし，優先度別バッファ方式自体，優先度毎にチャネルを仮想化したことになるため，優先度毎に経路が異っている場合は，仮想チャネルの効果が現れる．

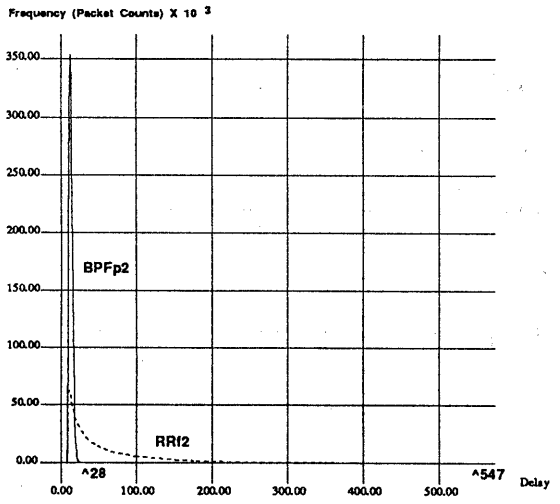

図 4: 転送遅延ヒストグラム（8段オメガ網，バッファサイズ2，パケット発生確率0．4）

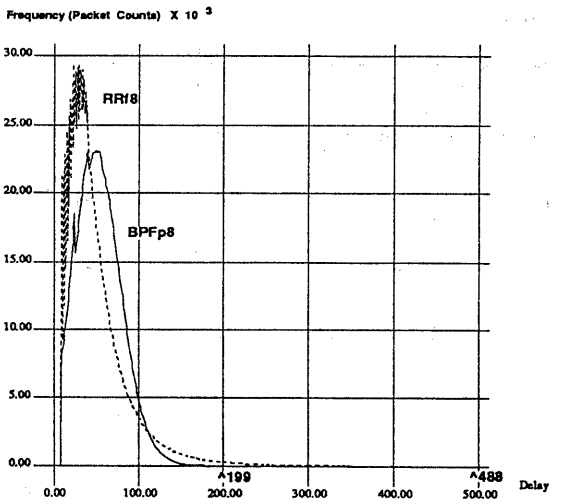

図 5: 転送遅延ヒストグラム（8段オメガ網，バッファサイズ8，パケット発生確率0．4，グループサイズ16）

優先度先送り方式は仮想チャネルと両立可能である．ただし，当該方式では入力ポートバッファとして優先度キューを使用するため必要とされるゲート数はFIFOの場合と比較して大きくなる．優先度キューの構成を工夫して，行き先の異なるパケットを途中から取り出せる機構を付加する実現等も考慮の対象となろう．

仮想チャネルを実現した優先度先送り方式でのルータは2×2ルータによるオメガ網のスループットを図6に示す．ここでの仮定は，入力パケットの行き先及び優先度は，ネットワークサイズまでの一様乱数．ネットワーク全体は同期動作し，パケットは1サイクル毎に転送され，入力ポートごとにサイズ8の優先度キューを持つ．最大優先度のパケットがブロックされたら優先度先送りを行ない（出力できれば，他方向への出力は行なわない），かつ，もう一方の出力先を持つできるだけ高い優先度のパケットを出力しようとする．そちらの出力先もブロックされていれば，その方向にも優先度先送りを行なう，というものである．BPF（通常の優先度先送り方式）と比較してVBPF（仮想チャネルBPF）は，どのサイズの網でも約0.2のスループット向上を得ている．図7は，パケット発生確率が0.8のときの各優先度毎の最大通信遅延である．この様な高いトラッフィクの下でも優先度に従った通信遅延を示している．

5 優先度キューの動作速度及び実現コスト

優先度キューのコストと動作速度は，優先度先送り方式の性能を決める最大の要因である．我々は，双方向シフトレジスタを用いた優先度キューの実現[PQ]を考案し，実際にCADによるデザインを行なった[VLPQ]．その結果，優先度フィールド32ビット，キューサイズ8の場合，1．5ミクロンプロセスで，クリティカルパスの遅延がティピカル20nS（ワーストケース36nS）で，全体のゲート数は約7Kであった．1.0ミクロンプロセスでは，ティピカル14nS（ワーストケース24nS）になると推測される．この実現ではキューの各要素毎に32ビットコンパレータを実装している．従って，キューサイズや優先度フィールドがこれほど必要でない場合は全体のゲート数の大幅な縮小が可能である．また，優先度フィールドの縮小は動作速度を改善する．

6 考察及び今後の予定

優先度別バッファ方式は，優先度の数が少ない場合（優先パケットと非優先パケットの区別しかない場合等）に有効であるが，優先度の数が多くなると実現が困難となる．

優先度先送り方式の動作速度を，優先度制御を行なわずラウンドロビン方式でFIFOを用いる実現と比較すると，先送りされた優先度とバッファ中の最大優先度の比較等は優先度キューの動作と同時に行なえるため前者のオーバヘッドは殆んど優先度判定にかかる時間だけである．

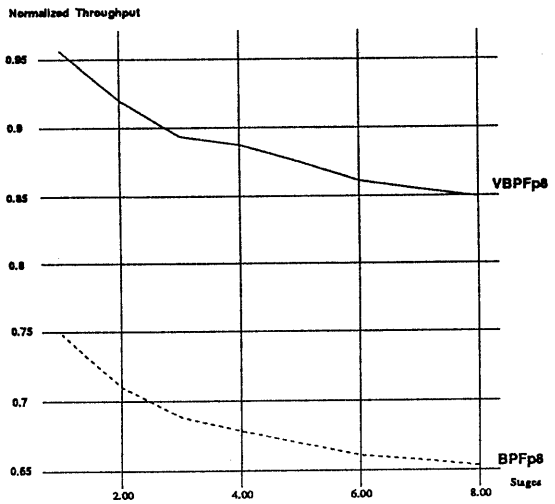

図 6: チャネルの仮想化を行なった優先度先送り方式のスループット

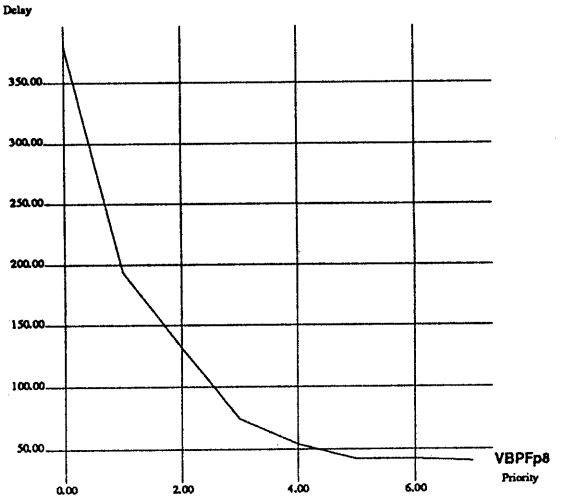

図 7: 仮想チャネル優先度先送り方式による優先度毎の最大転送遅延（パケット発生確率 0. 8）

今後，優先度先送り方式と仮想チャネルの併用について，シミュレーションやハードウェアコストの見積り等を通じ，キューのサイズ等のパラメータや仮想チャネルの実現方法等を詰めて行きたい．また，網の構成についての検討も行ないたい．

謝辞

現在の研究の機会を与えて下さった、柏木寛 電子技術総合研究所所長並びに弓場敏嗣 情報アーキテクチャ部長に感謝致します。また、有益な意見を下さった情報アーキテクチャ部の皆様に感謝致します。

なお，本研究は，科学技術庁の平成 4 年度科学技術振興調整費による「センサフュージョンの基盤的技術の開発に関する研究」の一環として行ったものである．

参考文献

[PF1] 戸田，西田，坂井，島田: 多段ネットワークにおける優先度制御方式の提案及びその評価，情報処理学会 計算機アーキテクチャ研究会（SWoPP'91），ARC89-22, 159/167 (1991)

[PF2] 戸田，西田，坂井，平木，島田: 優先度先送り方式を用いたオメガネットワークの性能評価，電子情報通信学会 コンピュータシステム研究会, CPSY91-53, 9/14 (1991)

[Kur89] Kurisaki, Lance, and Lang, Tomas, "Multistage Networks Including Traffic with Real-Time Constraints", ICPP, pp. I-19-I-22, 1989.

[Raj91] Rajkumar, Ragunathan, "Priority Inversion and Interprocessor Networks in Real-Time Systems", Workshop on Architectural Aspects of Real-time Systems, pp. 60-65, December 1991.

[Dal90] Dally, William J., "Virtual-Channel Flow Control", Int. Simp. on Comp. Architecture 1990, pp.60-68, 1990.

[Tok89] Tokuda, H., Mercer, C.W., Ishikawa, Y., and Marchok, T.E., "Priority Inversions in Real-Time Communication", Proc. of Real-Time System Symp., IEEE Computer Society TC Real-Time Systems, pp. 348-358, December 1989.

[PQ] 西田，戸田，坂井，平木，島田，「実時間用並列処理計算機 CODA-r のアーキテクチャ」，SWoPP 大沼' 91 資料（情報処理学会 ARC），1991 年 7 月.

[VLPQ] Nick Michell, Kenji Toda, Kenji Nishida, "A VLSI Priority Queue", ETL TR-91-27, Aug. 1991.

[ICN] Chuan-lin Wu and Tse-yun Feng, "Tutorial: Interconnection Networks for parallel and distributed processing", IEEE Computer Society Press, 1984.