Vol. 27 No. 1 情報処理学会論文誌 Jan. 1986

ショートノート

高速光トークンリングを用いたプロセッサ間結合システムの性能評価†

小柳津 育郎†† 魚住 栄市††
森 繁仁†† 星子 隆幸††

チャネル間を高速のトークンリングを介して接続するプロセッサ間結合システムについて，フロー制御を考慮した性能評価の一方法を提案する．

1. まえがき

近年，計算機による処理量の増大，システムの段階的成長や高信頼化等に対する要求を疎結合マルチプロセッサによる負荷分散で解決し，さらにはソフトウェア開発効率の向上，システム処理能力向上をプロセッサの機能分散化により解決していく傾向が顕著である．このような多数台のプロセッサを相互接続した複合システムへの適用を目的として，高速の光トークンリング方式を導入したプロセッサ間結合システムの検討を進めている[1]．トークンリングそのもののトラヒック特性については，橋田[2]，Bux[3]らにより明らかにされている．しかしながら，現実のシステムでは個々のプロセッサは有限の処理速度をもつチャネルとノードアダプタを介してリングに接続され，かつ送信側と受信側の各階層ごとにデータが正しく届いたかどうかを確認(フロー制御)しながら情報転送が実行される．このため実質的な情報転送量はネットワーク自体の転送能力に比べて低下する．

本論文では，上述の背景のもとに，高速の光リングを用いたプロセッサ間結合装置を具体的なシステムに適用する際のモデル化と性能評価の一方法を述べ，適用領域を考察する．

2. プロセッサ間結合装置の機能概要

本装置は，**図1**に示すように光リングのノードアダプタ (PCI) を介してプロセッサ間をチャネル結合する．1台の PCI は1台のプロセッサを接続し，リングには最大127台の PCI が接続できる．PCI は，複数のサブチャネルをもち，自側チャネルとの転送時だけチャネルを保留し，その他のときにはチャネルを解放するブロックマルチプレクス機能を具備しており，上位プロセッサに対して論理的に独立した複数の標準チャネル結合インタフェースを提供する[1]．

3. システムのモデル化

3.1 システム構成

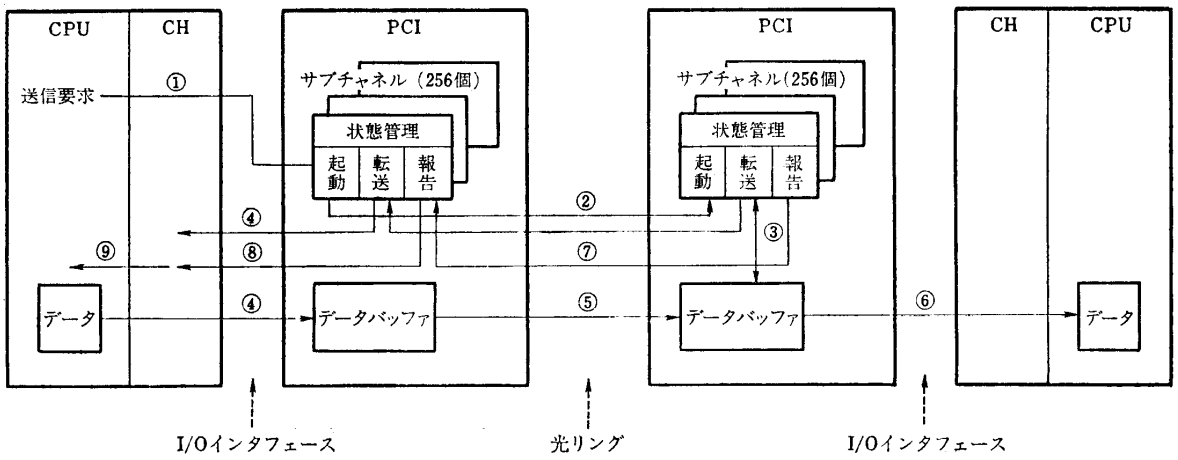
複数の通信処理プロセッサ (FEP) と情報処理プロセッサ (HOST) が光リングを介して接続されるシステムを対象にモデル化する．このシステムでは，端末からトランザクションの入力電文が投入されると，FEP においてビット組立，文字組立，電文組立等の通信処理を行った後，光リングを通して HOST に電文を転送する．HOST 上で情報処理が行われた後，出力電文は光リング，FEP を経由して端末に送信される．通常，1台の HOST は複数の FEP を経由してきたトランザクションの処理を分担するが，ここでは，簡単にするため，**図2**のように1組の FEP-HOST で系が構成され，N系統の FEP-HOST が光リングを共用する構成とする．

3.2 HOST-FEP 間電文転送モデル

1回のトランザクションは入力電文と出力電文からなる．**図3**のように，入/出力とも1電文当たり，フロー制御のための2回と電文そのものと合計3回のデータ転送が必要になる．

† Performance Evaluation for Processor-to-Processor Communications System with a High-Speed Optical Token Ring by Ikuro Oyaizu, Eiichi Uozumi, Shigehito Mori and Takayuki Hoshiko (NTT Electrical Communication Laboratories).

†† NTT 電気通信研究所

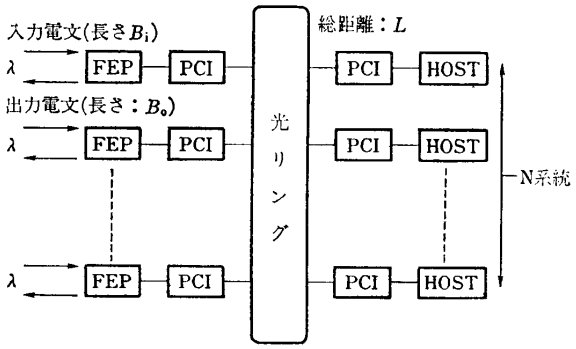


［動作説明］

① 送信 CPU は，チャネルを介して PCI を起動する．PCI はサブチャネルに起動状態を表示後，チャネルを解放する．
② 起動 PCI はトークンを獲得し起動の制御フレームを生成して，相手 PCI を起動する．
③ 受信側 PCI はデータバッファを獲得すると転送状態になり，トークンを獲得し転送要求の制御フレームを起動側 PCI に送る．
④ 送信側 PCI はチャネルに転送状態を報告して，CPU からのデータをデータバッファに読込みデータフレームを組立てる．
⑤ つづいて，トークンを獲得しデータフレームを送出し，受信側 PCI のバッファに書き込む．
⑥ 受信側 PCI は，受信終了をチャネルに報告して CPU にデータを送る．データを送り終るとチャネルを解放する．
⑦ 受信側 PCI はトークンを獲得して送信側 PCI に終了の制御フレームを送る．
⑧ 送信側 PCI はチャネルと再結合して終了報告を行う．終了報告後チャネルを解放する．
⑨ 送信チャネルは CPU に割込みを起こし，データ転送の終了を報告する．

図 1 プロセッサ間結合装置の概要

Fig. 1 The outline of the processor-to-processor communications system operation.



(装置名の略称)
HOST：情報処理プロセッサ
FEP：通信処理プロセッサ
PCI：プロセッサ間結合装置
(記号の意味)
λ：系当たりのトランザクションの到着率
N：光リングを共用する系の数
L：光リングの総延長距離 (km)
B_i：入力電文のデータの長さ（バイト）
B_o：出力電文のデータの長さ（バイト）

図 2 システム構成モデル

Fig. 2 The system configuration.

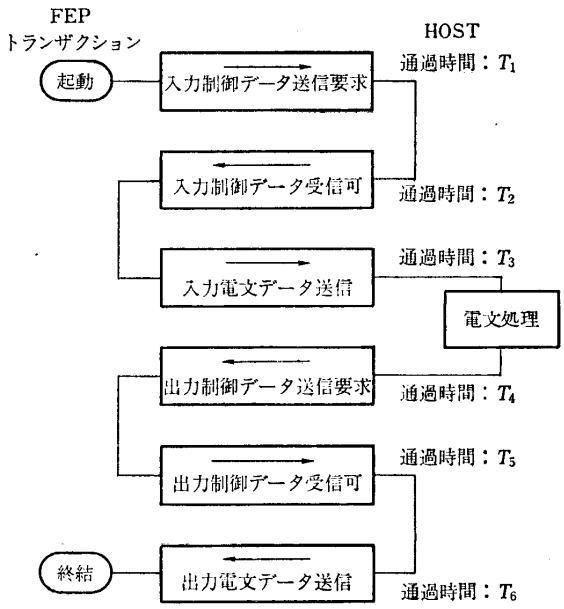


図 3 電文転送モデル

Fig. 3 The model for data transmissions between FEP and HOST originated with a transaction.

3.3 プロセッサ間結合装置データ転送モデル

2章の PCI/光リングの機能概要から，図3の箱枠の中の1回のデータ転送動作は，**図4**のようにモデル化できる．光リングには，転送データそのものと，両 PCI 間の同期に要する3回の制御情報の計4回のフレーム転送の負荷がかかる．

4. トランザクション通過時間の算出

4.1 算出方法

PCI のブロックマルチプレクス動作のために，かつN個の系が光リングを共用するため，バッファ競合やチャネル競合による PCI 処理待ちと光リングのトークン待ちが生ずる．

トランザクションの発生に伴う全データ転送の PCI/光リングの通過時間 (T_{all}) は，各待ち行列の直列接続型を用いて各待ちの総和で近似する．次の式で与えられる．

$$T_{all}=\sum_{i=1}^{6}(T_i)$$
$$=\sum_{i=1}^{6}[4\times W_1+W_2+W_3+3\times W_4+X_a+3\times X_b+X_{ci}+X_{di}+X_{ei}+X_f].$$

ここで，

T_i：図3の各データ転送に要する通過時間
W_1：トークン待ち時間
W_2：受信バッファ空き待ち時間
W_3：転送待ち（チャネル待ち）時間
W_4：起動・報告待ち時間
X_a：起動処理時間
X_b：制御フレーム伝搬時間
X_{ci}：図3の i 番目のデータ送出処理時間
X_{di}：図3の i 番目のデータフレーム伝搬時間
X_{ei}：図3の i 番目のデータ受信処理時間
X_f：終了処理時間

待ちモデルとして，トークン待ち W_1 は橋田の制限式多重待ちモデル[2]を，PCI の処理待ち W_2～W_4 には M/M/1 モデルを適用する．また，算出において，端末から系へのトランザクションの到着を到着率 λ の指数分布に従うと仮定し，各待ちモデルでの到着率は，端末からの1回のトランザクション投入により各待ちモデルを通過するサービス回数を λ に乗じたものを使用する．各時間の具体的算出式は付録に示した．

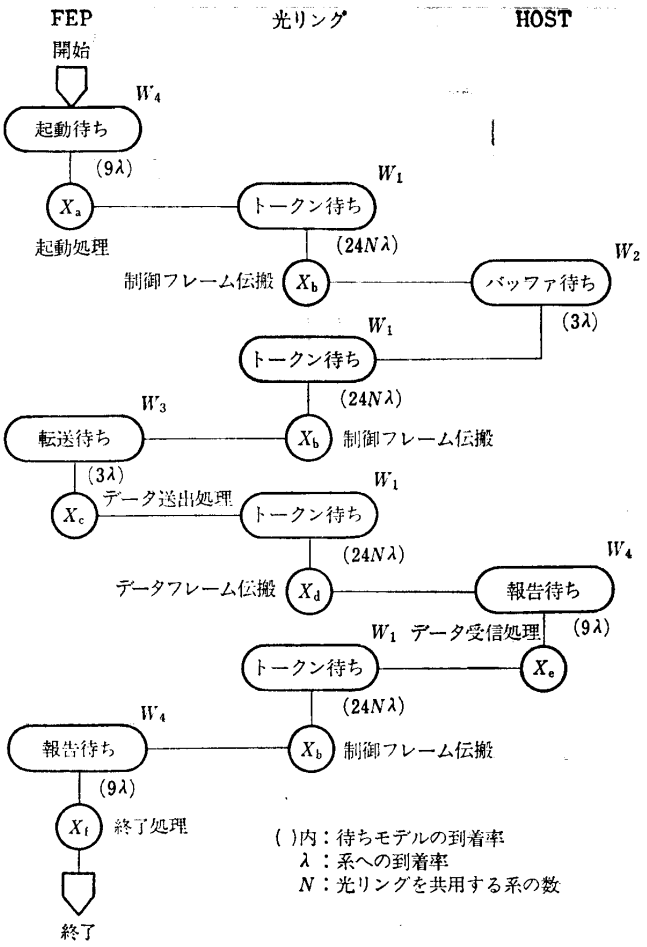


図4 データ転送モデル (FEP→HOST)
(FEP←HOST は対称形である)
Fig. 4 The model for a data transmission.

4.2 数値例と算出結果

典型的な数値例として，

光リングの転送レート：100 M ビット/秒
I/O インタフェース転送速度：2 M バイト/秒
入力電文長：100 バイト

における通過時間算出結果の例を**図5**に示す．通過時間に占める待ち時間は，系統数が小さい領域では PCI のバッファ待ちが主要因であるが，系統数が 10 を超える付近からトークン待ちが主要因となる．前者のケースでは PCI の装置使用率が，後者のケースでは光リングの使用率が，それぞれ 0.4 を超えるあたりから待ち時間が急激に増加する．これらの結果から，PCI および光リングの使用率上限 0.4 を条件に，出力電文長とリング総延長をパラメータとして PCI/光リングの適用領域を整理すると**図6**のようにまとめることができる．

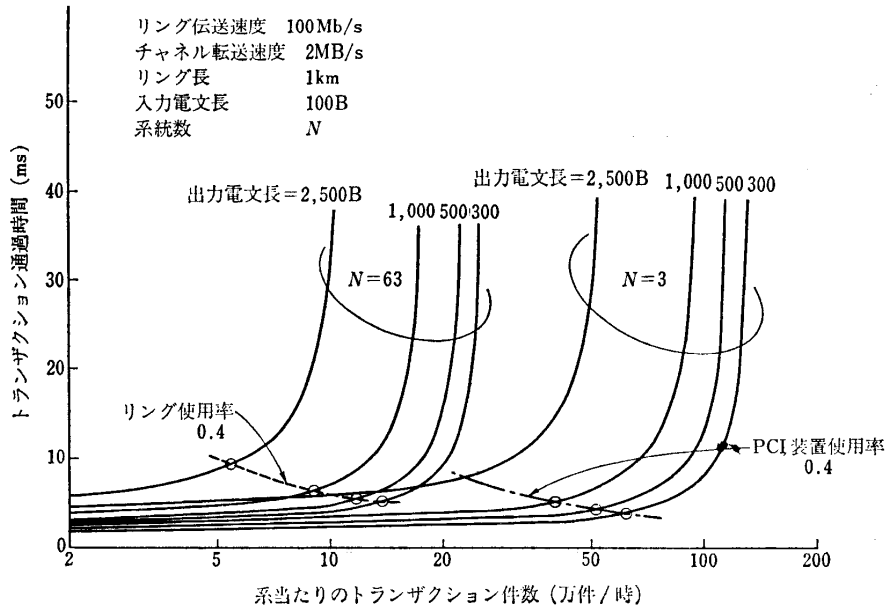


図 5 出力電文長と通過時間
Fig. 5 The total transaction elapse time.

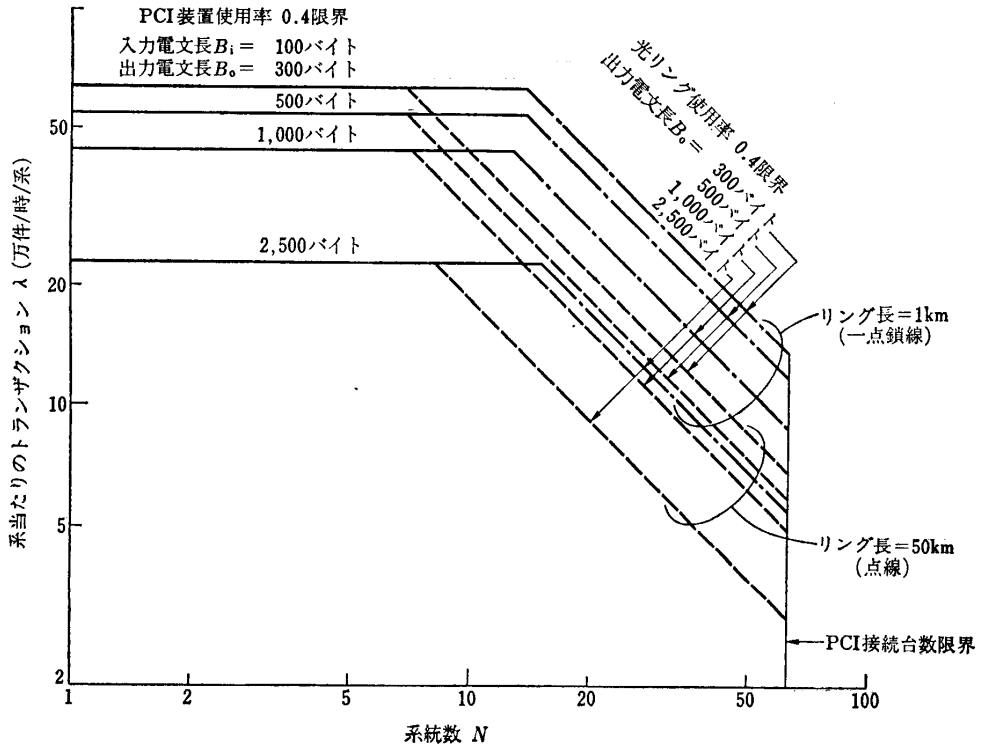


図 6 プロセッサ間結合システムの適用領域
Fig. 6 The throughput limitation of the processor-to-processor communications.

付表 データ通過時間算出式
Appendix The elapse time of a data transmission.

項番	項目	記号	算出式	数値例
1	起動処理時間	X_a	T_p	50 μs
2	制御フレーム伝搬時間	X_b	$T_s+T_L/2$	
3	データ送出処理時間	X_{ci}	$T_{cpi}+T_p$	
4	データフレーム伝搬時間	X_{di}	$T_{di}+T_L$	
5	データ受信処理時間	X_{ei}	$T_{cpi}+T_p$	
6	終了処理時間	X_f	T_p	50 μs
4	トークン待ち時間	W_1	$\overline{F}/2F+\alpha\cdot\overline{F_1}/2(1-\alpha\cdot F_1)$; $F_1=F+H$ $F=T_L+(n-1)\rho\cdot H$ $\rho=\alpha\cdot T_L/(1-\alpha\cdot n\cdot H)$ $\alpha=(\lambda\cdot n/2)\cdot P$ $P=4\times6$ ただし，P：フレーム転送回数/トランザクション	
5	バッファ待ち時間	W_2	$L_1\cdot T_B/(1-L_1)$; $L_1=\alpha\cdot T_B/4$	
6	転送待ち時間	W_3	$L_2\cdot T_C/(1-L_2)$; $L_2=\alpha\cdot T_C/4$	
7	起動/報告待ち時間	W_4	$L_3\cdot T_C$	
8	受信バッファ保留時間	T_B	$W_1+X_b+W_3+T_C-(T_L-T_M)+W_4+X_e$	
9	送信動作によるチャネル占有時間	T_C	$X_c+W_1+X_d$	
10	平均フレーム送出時間	H	$(B_i+B_o+4B_c+(3\times6+6)A)/V_L/P$; B_i：入力電文データ長 B_o：出力電文データ長 B_c：制御データ長 A；制御フレーム長	B_c：90 バイト A：43 バイト
11	リング一周時間	T_L	$5L+D\cdot n/V_L$; D：遅延バイト数/ノード V_L：リング転送速度（バイト/秒）	$D=5$ バイト $V_L=12.5$M バイト/秒
12	制御フレーム送出時間	T_s	A/V_L	3.4 μs
13	データフレーム送出時間	T_{di}	$(B_i+A)/V_L$	
14	データの I/O インタフェース転送時間	T_{cpi}	B_i/V_I ; V_I：I/O インタフェース転送速度（バイト/秒）	$V_I=2$M バイト/秒

λ：件/秒/系，n：ノード数 ($2\times N$ N：系統数)，L：リング長 (km)
$B_i=B_c$ ($i=1,2,4,5$)　$B_i=B_i$ ($i=4$)　$B_i=B_o$ ($i=6$)

5. むすび

この報告では，チャネル接続の高速の光リングを用いたプロセッサ間接続装置をシステムに適用する場合について，フロー制御を考慮したモデルによる性能評価の一方法を提案した．今後，実システムにおいて本手法の評価を行う予定である．

謝辞 本検討に当たりご指導いただいた NTT 電気通信研究所トラヒック研究室室員の各位に感謝いたします．

参考文献

1) 小柳津，魚住，星子：高速の光ファイバループを用いたプロセッサ間結合方式に関する検討，情処分散処理システム研究会資料，23-5 (1984).
2) 橋田：ポーリング制御における待合せモデルの解析，昭和 43 年度信学会全国大会，p. 129 (1968).
3) Bux, W.: Local-area Subnetworks: A Performance Comparison, *IEEE Trans. Commun.*, Vol. COM-29, pp. 1465-1473 (Oct. 1981).

(昭和 60 年 6 月 24 日受付)
(昭和 60 年 9 月 19 日採録)