2P-2

バッファ付き入出力サブシステムにおける負荷均衡制御のための漸近近似手法

山本　彰*　　坪井俊明**　　北嶋弘行*　　難波龍雄***　　土井　隆***
(*日立製作所システム開発研究所，　**日立マイクロコンピュータエンジニアリング，
***日立製作所小田原工場)

1．はじめに

半導体メモリの大容量化、低価格化に伴い、計算機システムの入出力系、特に、入出力制御装置内に半導体メモリより構成されるバッファを組み込むことにより、コスト・パフォーマンスの向上を実現するアーキテクチャの開発が行われている[1]。以上のようなアーキテクチャにおいて、入出力装置を複数の制御装置に接続する場合、１台の入出力装置のデータが複数の制御装置のバッファに分散することをさけるため、それぞれの入出力装置の制御権を１台の制御装置に与え、入出力装置と制御装置間のデータ転送は制御権下にある制御装置とのみ実行する方式をとることがある[2]。この時、ある制御装置の制御権下にある入出力装置のみに負荷が偏った場合、適切な入出力装置を選択して、制御権の移行を行うことにより、制御装置間の入出力負荷を均衡させる必要が生ずる。

制御権を移行する入出力装置の選択方式としては、制御権移行後の負荷状況を解析モデル等により予測計算し、この結果最も負荷が均衡するという結果が得られた装置を移行対象として、選択するのが妥当であると考えられる。通常、この計算は、入出力処理の稼働中に実行するため、極めて高い高速性が要求される。さらに、この選択処理は制御装置内で実行するため、制御装置内で収集可能なデータで予測計算を実行しなければならない。本講演では、制御装置内で収集可能なデータにより、制御権を移行する入出力装置の選択を高速に実行可能とする漸近近似手法について述べる。ただし、本講演では、簡単のため、チャネルと入出力装置のデータ転送速度が等しい構成を取扱いの対象とする。

2．モデル化対象システム

図１にモデル化対象システムを示す。モデル化対象システムは、制御装置内のバッファを中間バッファとして、主記憶と制御装置内のバッファとの間のデータ転送処理、および、CPU内でのデータ処理を実行する上位系と制御装置内のバッファと入出力装置のデータ転送を実行する下位系が並列に動作していることになる。この場合、制御装置内のバッファが上位系と下位系の同期をとることになる。例えば、入出力装置に対し読み取り動作を実行している場合、上位系は制御装置内のバッファにデータがなくなると、待ち状態に入ることになる。この待ち時間をバッファデータ待ち時間と呼ぶ。

各入出力装置はそれぞれ２台の制御装置に接続されている。各制御装置は、制御権を有している入出力装置とのみ、バッファと入出力装置の間のデータ転送処理を実行する。制御装置が、チャネルからある入出力装置に対する入出力要求を受け付けた時、この入出力装置の制御権を有している場合、自装置内のバッファを用いて、上位システムとの間でデータ転送処理を実行する。制御権を有していない場合、相手の制御装置内のバッファと上位システムのデータ転送処理を実行する。このため、ある制御装置がある入出力装置の要求を受け付けた時、もう一方の制御装置がこの入出力装置の制御権を有している方のバッファと上位システムとの間でデータ転送中の場合、待ち状態に入ることになる。以下、この待ち時間をバッファ転送路待ち時間と呼ぶ。

負荷均衡制御とは、各制御装置内のバッファに対して生ずるバッファ転送路待ち時間、バッファデータ待ち時間を観測し、この差が一定値を超えたとき、両者のバッファ転送路待ち時間、バッファデータ待ち時間を均衡させる入出力装置を選択し、制御装置間で入出力装置の制御権を移行させるものである。

3．漸近近似手法

漸近近似手法の基本的な仮定[3]を以下に示す。仮定１で示すように発生する待ちを漸近的に発生する待ちと呼ぶ。

仮定１：ＩＦ（待ち時間を０と仮定して算出したその資源の利用率≦１）ＴＨＥＮ　待ち時間＝０
　　ＥＬＳＥ　待ち時間＝その資源の利用率を１とするだけの待ち時間

さらに、中間バッファを有する系の漸近近似手法における仮定[3]を以下に示す。

仮定２：ＩＦ（バッファがない時の上位系のスループット＜バッファがない時の下位系のスループット）
　　ＴＨＥＮ　システム全体のスループット＝上位系のスループット
　　ＥＬＳＥ　システム全体のスループット＝下位系のスループット

Asymptotic Model for Load Balancing of I/O Subsystems with Buffer Memory

以上より、バッファがないとして算出したスループットが小さい方の系には、バッファデータ待ち時間は発生せず、大きい方の系のみにこの待ち時間が発生することになる。チャネルの転送速度と入出力装置のデータ転送速度が等しい時には、CPUのデータ処理が存在する分だけ、上位系の処理時間が長くなり、上位系のスループットが下位系に比較して大きくなることはない。従って、仮定2より、以下では、上位系の解析のみを行う。さらに、上位系には、バッファデータ待ち時間は発生しないことになるため、負荷均衡の制御としてはバッファ転送路待ち時間のみを考慮すればよいことになる。

図2に上位系のモデルを示す。モデル内には各入出力装置に対応して処理要求が1つ存在する。処理要求はCPU上でデータ処理等を実行している時には、各処理要求対応に存在する入出力発行間隔サーバ上に滞在し、入出力処理を開始する時、まずチャネル・キューに入り、どちらか一方のチャネルが空くとチャネル・キューから出る。この時、もう一方のチャネルがこの入出力装置の制御権を有している制御装置内のバッファとのデータ転送処理を実行している時には、転送路キューにはいる。転送路が空くとチャネル側転送サーバをつかみ、制御装置内のバッファと主記憶の間の転送処理に入ることになる。通常の解析モデルでは、入出力発行間隔時間、チャネル側転送時間を入力データとし、チャネル待ち時間、バッファ転送路待ち時間を計算結果として算出する。しかし、負荷均衡制御は制御装置内で実行するため、モデルの入力とするデータは制御装置内で計測可能でなければならない。従って、ここでは、図2に示したように、入出力発行間隔時間＋チャネル待ち時間を制御装置入出力発行間隔時間として計測し、この時間とチャネル側転送時間を入力データとして、バッファ転送路待ち時間を算出する方式をとる。以下、バッファ転送路待ち時間の算出にあたり設けた仮定を示す。

仮定3：入出力処理を実行する際発生する待ち時間のうち、バッファ転送路待ち時間のみが漸近的に発生するものとする。

仮定3が成立するとそれぞれの制御装置のバッファ転送待ち時間はよく知られた以下の式[3]により算出できる。各記号の定義は表1に示す。ここで、MAX(a, b)はaとbのうち大きい方の値をとる関数である。

$$W_A = MAX\{0, (N_A-1)S_A - X_A)\}, \quad W_B = MAX\{0, (N_B-1)S_B - X_B)\} \qquad (1)$$

4．制御権移行入出力装置選択方式

制御権を移行する入出力装置を選択する際の処理は、(1)式を基本にしてそれぞれの入出力装置の制御権を移行した後の両制御装置のバッファに発生するバッファ転送路待ち時間を算出する。この結果、それぞれのバッファ転送路待ち時間の差が最も小さかった装置を制御権移行対象として選択するものである。

5．おわりに

バッファ付き入出力サブシステムにおいて制御装置間の負荷を均衡させるための漸近近似手法についての報告を行った。

参考文献　1）山本他：カートリッジ型MTにおける先読み・纏め書きスジューリング方式の提案と解析，情報処理学会第35回全国大会　(1987)　2）山本他：カートリッジ型MTにおける予測制御型負荷均衡制御方式，情報処理学会第36回全国大会 (1988)　3）西垣，山本：資源割当て優先度のある多重プログラミング・システムのボトルネック解析，情報処理学会論文誌，Vol.23,No.5,pp.562-569,(1982)

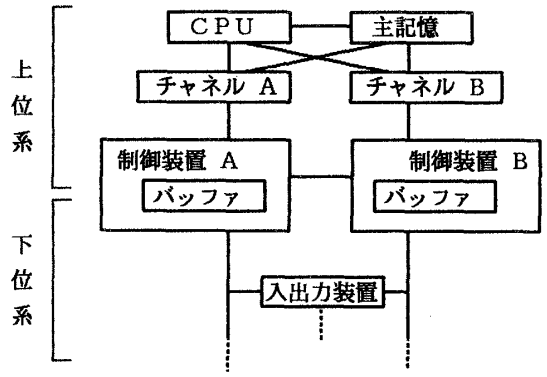

図1．モデル化対象システム構成

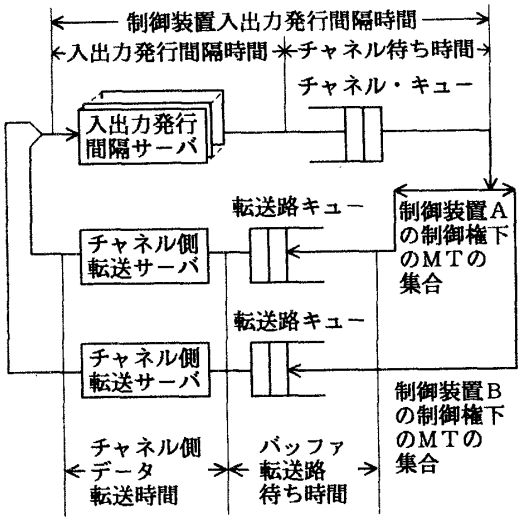

図2．上位系のモデル

表1　記号の定義

N_A, N_B	制御装置A，Bの制御権下にあるMTの台数
X_A, X_B	制御装置A，Bの制御権下にあるMTの制御装置入出力発行間隔時間
S_A, S_B	制御装置A，Bの制御権下にあるMTのチャネル・データ転送時間
W_A, W_B	制御装置A，Bの制御権下にあるMTのバッファ転送路待ち時間