計算機アーキテクチャ 113−17
（1995. 8. 24）

超並列計算機 JUMP-1 における ディスク入出力サブシステムの シミュレーションによる評価

大谷 智，中條 拓伯，金田 悠紀夫
神戸大学工学部情報知能工学科

e-mail:satoshi@timpani.seg.kobe-u.ac.jp

JUMP-1 は，要素プロセッサ，2次キャッシュメモリ，高速な同期／通信を行う MBP より構成されるクラスタを RDT と呼ばれる相互結合網で接続した分散共有メモリ型超並列計算機である．クラスタと入出力サブシステムは STAFF-Link と呼ばれる高速なシリアルリンクにより接続される．入出力用のバッファを共有入出力バッファとして本体の共有メモリ空間にマッピングすることにより，メモリアクセスとして，入出力アクセスを行う．本稿では，ディスク入出力サブシステムの基本的な性能をイベント駆動型のシミュレーションにより評価し，Video On Demand をアプリケーションとして実行した場合の性能の予測を行なった．

Performance Evaluation of a Disk Input/Output Subsystem of the Massively Parallel Computer:JUMP-1 by using simulation

Satoshi OTANI,
Hironori NAKAJO and Yukio KANEDA
Department of Computer and Systems Engineering , Faculty of Engineering, Kobe University

JUMP-1 is a distributed shared-memory massively parallel computer which consists of many clusters via network called RDT. Each cluster consists of processing elements, secondary cache memory and MBP which controls fast synchronization and communication among clusters. Disk I/O and image display subsystems are connected to clusters via fast serial links called STAFF-Link. Since shared I/O buffer is mapped onto global shared-memory space, each I/O access can be dealt as a memory access. In this paper, we evaluate fundamental performance of disk I/O and estimate execution performance of Video On Demand by an event driven simulation method.

1 はじめに

近年，プロセッサ能力が飛躍的に向上し，大容量メモリの使用が容易になってきた．それに伴い，計算機システムへの要求もより複雑になり，これに応えるには複数のプロセッサでの並列，分散，協調による高性能化が必要となっている．その一つの有力な候補として超並列計算機が挙げられ，様々な大学，研究機関で開発，研究が進められ商用機も登場している．

文部省科学技術研究補助金・重点領域研究においても分散共有メモリ型超並列計算機のプロトタイプ JUMP-1 の開発がなされている[1]. JUMP-1 は，クラスタ間を RDT (Recursive DiagonalTorus)ネットワーク [2]と呼ばれる階層トーラス状の相互結合網を介して結合されたマシンである．またクラスタ内では，要素プロセッサ(PE)の他に MBP(Memory Based Processor)と呼ばれる非局所処理に特化したプロセッサがあり，効率の良い分散共有メモリシステムを実現する．

このような分散共有型メモリでは，メモリ性能と同様にそれを支える入出力性能も必要と考えられる．特に多数のプロセッサで構成される超並列計算機では，扱うデータ量も多く効率の良いデータ供給がなければ，そのシステムのプロセッサ能力を有効利用することもできない．

現状では，多くのスーパコンピュータの入出力システムはある特定のノードに専用の高速入出力バス(HiPPI 等)を設置し，そのバスに様々な入出力装置を接続する形態が一般的である．しかし，多数のプロセッサから構成される超並列計算機に対して専用入出力バスを接続すると接続するノードやその近傍でボトルネックが生じる．

そこで JUMP-1 のディスク入出力サブシステムでは，STAFF-Link(Serial Transparent Asynchronous First-in First-out Link)と呼ばれる高速なシリアルリンク[3]を用い，各クラスタの MBP と接続した入出力装置を分散・設置した．入出力相互接続網とプロセッサの相互接続網とは，独立に構成した形態をとる．そして各々の入出力ユニット上には，入出力用バッファ(共有入出力バッファ)が存在し，そのバッファは JUMP-1 のグローバルアドレス空間にマッピングされる．この入出力バッファを用いることにより，全クラスタで入出力機器が共有でき，入出力機器へのアクセスをメモリアクセスとして扱うことができる．このユニットを複数もつことにより入出力装置へのアクセスの分散化を図っている．またこのユニットは，SPARC station 5(SS5)上に構築されディスク入出力に関する低レベルの管理を行なっている．このことより本ディスク入出力サブシステムは以下の特徴を持つ．

- 共有入出力バッファを用いたアクセス
- 入出力装置の分散により大容量のデータのシステムの構築が可能

本稿では，これらの特徴を待ち行列型シミュレーションにより検証し，本システムの性能評価を行う．

第 2 章では JUMP-1 ディスク入出力サブシステムについて述べ，第 3 章でシミュレーションモデルとアクセスパターンに対する本システムの特性について述べる．

2 JUMP-1 入出力サブシステム

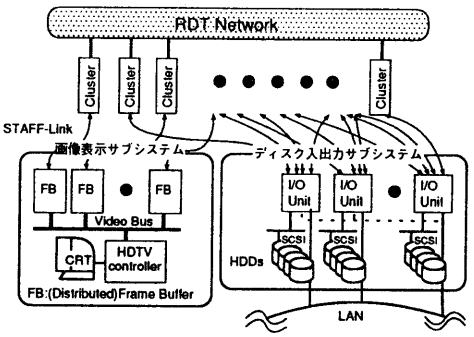

図 1: JUMP-1 システムの全体構成

図 1に JUMP-1 システム全体構成を示す．入出力サブシステムは，ディスク入出力サブシステムと画像表示サブシステムに分けられる．
この入出力サブシステムの特徴を以下に示す．

STAFF-Link の使用

STAFF-Link は，入出力機器に対して分散アクセスを図るために，多数の入出力ユニットを複数のクラスタに接続する形態をとる．そのため入出力ユニットの設置場所を考慮に入れる必要がある．各クラスタと入出力ユニット間には，STAFF-Link と呼ばれる高速シリアルリンクを用い，これによりケーブル長の制限を少なくし，

設置場所に自由度を与えた．

共有入出力バッファを用いたアクセス

各入出力機器は，それぞれ共有入出力バッファと呼ばれる入出力用のバッファを持つ．そしてこのバッファをJUMP-1のグローバルメモリ空間にマッピングさせる．これにより全てのディスクアクセスは，2.2で後述するJDD，IDDといったソフトウェアの支援により，共有入出力バッファのアドレス，データサイズ，データ等の内容を持つパケットのやり取りで行なわれ，memory mapped I/Oを実現する．

2.1 JUMP-1ディスク入出力ユニットの構成

図2にJUMP-1ディスク入出力ユニットの物理的構成を示す．このユニットは，SPARC station 5(SS5)上に構築され，ディスクの入出力に関する低レベルの管理をディスク入出力ユニット側で処理することができる．よってクラスタ側では，デバイスの構成要素，形態については考慮に入れる必要はない．以下にディスク入出力ユニットの構成要素を示す．

- SPARC chip
  クラスタからの要求に対してディスク入出力に関する各種の処置を行う．SS5のプロセッサであるmicroSPARC-IIになる．
- ディスク装置
  SCSIバスを介してSS5に接続されているディスクである．
- 共有入出力バッファ
  JUMP-1クラスタ側で管理される入出力用のバッファメモリである．JUMP-1のグローバルアドレス空間にマッピングされ，全てのクラスタから直接アクセスすることが可能である．SS5の内部メモリを用いて実装される．
- DMAC(Direct Memory Access Controller)
  STAFF-Linkと共有入出力バッファ間のデータのやり取り，クラスタへの割り込みパケットを転送する．

2.2 ディスク入出力サブシステムのアクセス方式

JUMP-1のディスク入出力サブシステムは，クラスタ側のOSから論理的に単一のブロックデバイスとして扱われ，各ディスク入出力ユニットのディスクブロックには，システムで一意のブロック番号が付けられている．ユーザのアプリケーション上のディスクへの読み書きは，クラスタ側のOSからJUMP-1側のOS上のデバイスドライバ(JDD)とディスク入出力ユニット制御用OS上のデバイスドライバ(IDD)のサポートにより共有入出力バッファへのメモリアクセスとしてJUMP-1のクラスタメモリへのアクセスと等価に実行される[4]．以下にディスク入出力におけるJUMP-1クラスタ上のファイルシステム及びメモリオブジェクトマネージャ，JDD，IDDの役割を示す．

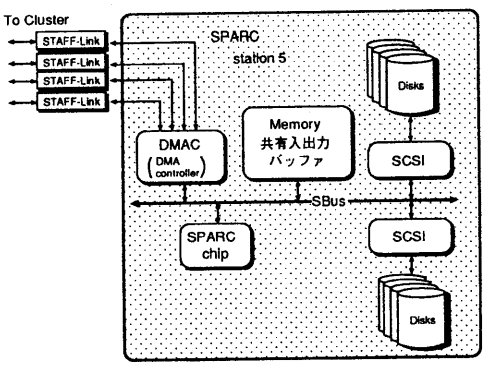

図2: ディスク入出力ユニットの物理的構成

ファイルシステム及びメモリオブジェクトマネージャ：各入出力ユニットの空きブロック情報，共有入出力バッファの利用状況といった，ディスク入出力に関する情報の管理を行なう．

JDD: JUMP-1クラスタ上のOSカーネルからのディスク入出力要求をRDTのメモリアクセスパケットに変換し，変換したパケットをディスク入出力ユニットに対して転送する．

IDD: JDDより送られたパケットの解釈，ディスク装置や共有入出力バッファへのアクセスを行なったり，アクセスの完了を知らせる割り込みパケットの送信を行う．

以下にクラスタのOSから送られる要求パケットの種類(3種類)を示す．

- **共有入出力バッファへのリード要求**
  共有入出力バッファのアドレス，データサ

イズがパラメータとして渡される．基本的には，RDTにおけるメモリアクセスパケットと同じである．

- **ディスクリード要求**
  ブロック番号，共有入出力バッファのアドレスをパラメータとして渡される．
- **ディスクライト要求**
  ブロック番号，共有入出力バッファのアドレス及びデータをパラメータとして渡される．

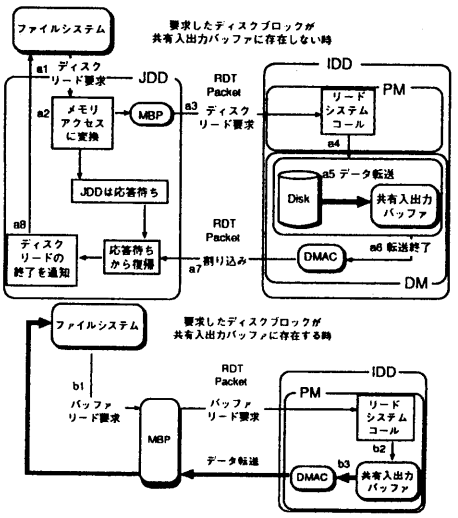

図 3: リードアクセスの手順

2.2.1 リードアクセス

リードアクセスの手順は図3の通り進められる．要求するデータが共有入出力バッファに存在しない時，OSのカーネルは，ディスクリード要求を発行する(a1)．JDDは，このディスクリード要求をメモリアクセスに変更し(a2)，そしてMBPによりRDTパケットとしてデータのある入出力ユニットに送る(a3)．送られてきたRDTパケットをIDD側では，パケットマネージャスレッド(PM)により解釈し，そして解釈された要求をディスクアクセス待ちキューに入れる(a4)．ここでPMは次に来るパケットの処理に移る．一方ディスクマネージャスレッド(DM)は待ちキューより要求を取り出し，ディスクブロックのデータを共有入出力バッファのアドレスに転送する(a5)．転送終了後DMは，割り込みパケットを要求元のクラスタのOSカーネルに送る(a6)．DMは，次の要求の処理に移る．以上でディスクリード要求は終了する．改めてOSのカーネルからバッファリード要求を発行する(b1)．目的のデータが格納されている共有入出力バッファへの読み出しをPMが行い(b2)，共有入出力バッファからクラスタにデータ転送を行う(b3)．この処理(b1-b3)は，要求するデータが共有入出力バッファに存在する時にも同様に行われる．

2.2.2 ライトアクセス

ライトアクセスは以下の手順をとる．(図4) ファイルシステムによりディスクライト要求が発行される(c1)．これをJDDは，メモリアクセスに変換し，ディスク入出力ユニット側に送る(c2)．IDDでは，PMが送られてきたパケットの解釈を行い共有入出力バッファにデータを書き込む(c3)．そしてディスクアクセス待ちキューにデータを除いた要求を送り(c4)，PMは次の要求パケットの処理に移る．DMは待ちキューより要求を取り出し，共有入出力バッファから指定したディスクブロックにデータを転送する(c5)．転送終了後(c6)，割り込みパケットを要求元のJDDに送り(c7)，JDDからクラスタに書き込み終了を通知する(c8)．

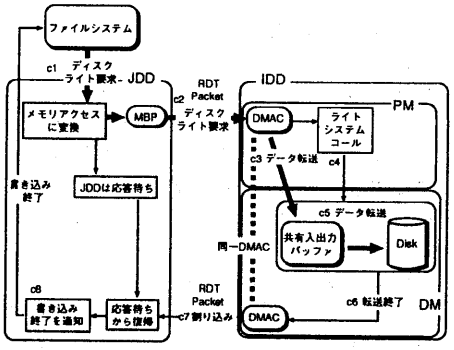

図 4: ライトアクセスの手順

3 シミュレーションによる性能評価

本ディスク入出力サブシステムの共有入出力バッファを用いたアクセス方式，分散入出力システムにより大容量のデータが扱える有効性を検証するため，待ち行列シミュレーション[5]により評価を行った．以下に，本ディスク入出力サブシステムのアクセスでの待ち行列モデルを示し，シミュレーションを行った結果を示す．

3.1 シミュレーションモデル

2章で述べたディスク入出力サブシステムに対するアクセス方式にもとづいたディスク入出力ユニット毎の待ち行列モデルをリードアクセス，ライトアクセスについてそれぞれ図5，図6に示す．また以下に今回のシミュレーションで用いたパラメータを示す．

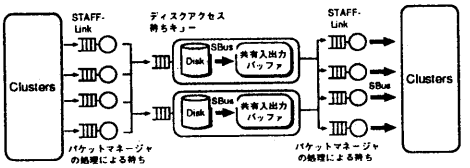

図5: 基本リードモデル

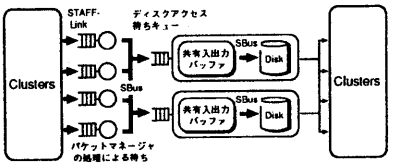

図6: 基本ライトモデル

モデルの全体構成

- ディスク入出力サブシステム
  クラスタ：256台，
  ディスク入出力ユニット：64台
- ディスク入出力ユニット
  SBus：1本，SCSI：2本，DISK：2台
  共有入出力バッファの容量：4MB
- 最大転送速度
  DMAC：8.3MB/sec
  SBus：20.0MB/sec
  SCSI(FAST SCSI2)：10.0MB/sec
  STAFF-Link：140Mbps(17.5MB/sec)
- 各処理時間
  IDDパケットの解釈時間：0.8ms
  MBPの処理時間：0.1ms

表．1:ディスクのパラメータ

ディスク容量	22.8GB
セクタの大きさ	512bytes
セクタ/トラック	16
シリンダー数	6000
ヘッド数	50
平均シーク時間	12.0 ms
回転速度	5400rpm

3.1.1 シミュレーション1

I/Oユニット1台に対する特性を検証するため以下の条件でシミュレーションを行なった．

- ディスクのブロックに対してランダムにアクセスを行う．

アクセスサイズ

- バッファへのリード：4KB
- ディスクへのリード
  (a)トラック単位(80KB)
  (b)ブロック単位(4KB)
- ディスクへのライト：
  トラック単位(80KB)

アクセス頻度

- リードアクセス：10 ～ 500Kbyte/s
- ライトアクセス：10 ～ 300Kbyte/s

3.1.2 シミュレーション2

- 実際のアプリケーション例(VOD)
  シーケンシャルデータに対する性能を評価するため，VOD(Video On Demand)のアクセスパターンを用いてシミュレーションを行う．
  ディスク入出力サブシステムのディスクに図7の通りにMPEG-2の規格を満たすストリームデータを配置する．このシステムに複数のカスタマ(クラスタ)がストリームを同時に要求する．ただしこの要求はビデオ再生のみの要求とする．
  ディスク入出力ユニットは，要求されたデー

タをクラスタに転送する．データは，クラスタ側でデコードされ画像として表示される．クラスタ側には十分大きいバッファがあることとし，ディスク入出力サブシステムは，一定制限時間内にデータをクラスタに転送しなければならないとする．

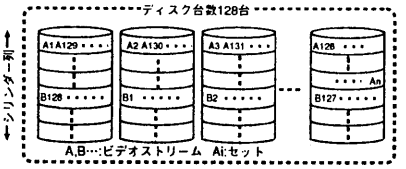

図 7: ディスクブロック配置図

ブロックの配置方法
ストリームデータ(シーケンシャルデータ)をトラック単位毎に分割し，その分割されたデータを 1 セットとする．このセットを全てのシリンダーに対し図 7の通り格納する．またそれぞれのストリームの先頭ブロックは，それぞれ分散して配置する．ここでのアクセスは，トラック単位でディスクにアクセスを行う．
以下に VOD におけるパラメータを示す．

- 1 ストリームの大きさ：5.3GB
  (ビデオの再生時間 90 分)
- MPEG-2 に要求される転送速度：8Mbps

3.2 シミュレーション結果

3.2.1 シミュレーション 1

図 8は，リードアクセスでのアクセス頻度に対する応答時間を示したものである．そして図 9～ 11は，それぞれシステムの構成要素の時間利用率を示している．また図 12にライトアクセスでのアクセス頻度に対する応答時間を示す．本システムはトラック単位でのアクセスにおいてリード・ライト共に 220KB/s 程度のアクセス頻度に対応でき，ブロック単位でのリードのアクセスにおいて 400KB/s 程度のアクセス頻度に対応できる．

応答時間の劣化は，ディスクの稼働率の増加により発生する．そして図 8，10が示す通り，バッファ，ディスク間の転送時間による性能の劣化も見られた．しかし，STAFF-Link は図 11の示す通り今の性能で十分対応できる．

3.2.2 シミュレーション 2

VOD アプリケーションを用いた時のカスタマに対するユニット一台の 1 トラックの応答時間を図 13に示し，またその時の各ファシリティの時間利用率を図 14に示す．図 13が示す通り応答時間の急激な変化がおきない 200 人程度では，カスタマに安定した供給が行なえる．このアプリケーションに対するアクセスでは，データのディスクからバッファへの転送よりバッファからクラスタに転送する時間が増大する．この時間の増大によりディスク・バッファ間の転送とバッファ・クラスタ間の転送が SBus の競合を発生させ，パケットマネージャのパケット解釈にかかる時間の増大が応答時間の劣化を引き起こしている．

3.3 考察

ランダムアクセスにおいてトラック単位での転送は，約半分のアクセス頻度しか対応できない．しかしトラック単位での転送は，一度に 20 ブロック分を転送していることを考えるとデータの転送時間による性能の劣化でありディスクシーク時間の隠蔽はできていると考えられる．そしてランダムアクセスにおいて図 9が示すようにディスク･バッファ間の転送とバッファ･クラスタ間の転送を SBus 1 本で対応しているためにバッファがボトルネックとなっている．VOD アプリケーションでのアクセス(シーケンシャルアクセス)では，トラック単位の転送が有効に用いられており，一台の入出力システムで 200 人のカスタマに同時に別々のビデオを提供でき，転送速度に換算すると 1 入出力ユニットあたり約 3.0MB/s，システム全体では約 192MB/s の転送能力があると言え大容量のデータシステムにも対応できると考えられる．図 14の Disk に対する STAFF-Link の時間利用率の増加によりトラック･バッファ間のトラフィックが応答時間に影響することがわかり，ディスク・バッファ間の転送及びディスクのシーク時間の隠蔽が行われていることがわかった．

これらよりこのディスク入出力サブシステムでは，共有入出力バッファでのクラスタ，ディ

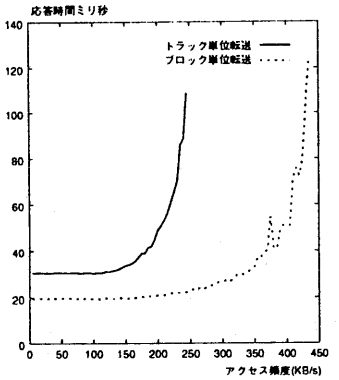

図 8: リードアクセスの応答時間

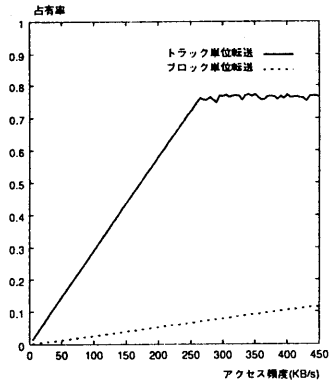

図 9: SBus の占有率

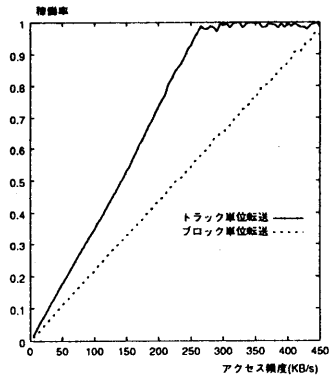

図 10: ディスクの稼働率

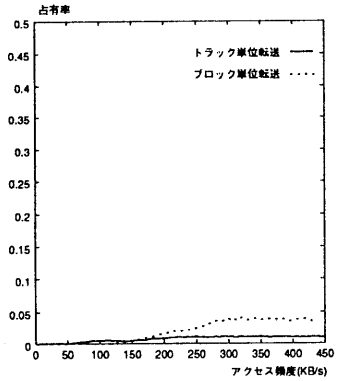

図 11: STAFF-Link の占有率

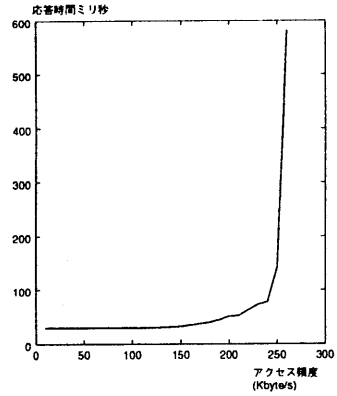

図 12: ライトアクセスの応答時間

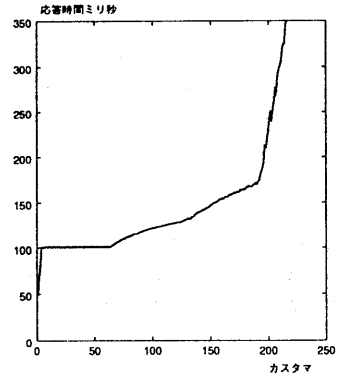

図 13: VOD での 1 トラック当りの応答時間

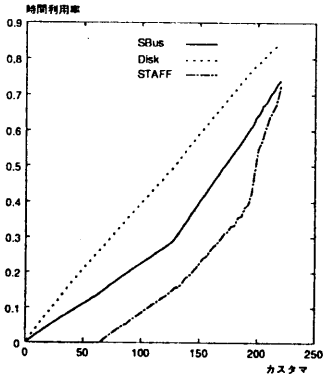

図 14: VOD での各ファシリティの時間利用率

スクに対しての転送方法を改善することが有効であると考えられる.

4 おわりに

本論文では，二つの異なったアクセスパターンを用い JUMP-1 のディスク入出力サブシステムの評価をシミュレーションにより行った．このシミュレーションにより本システムのデータ転送サイズに対する特性及び VOD におけるアクセスに対して 200 人のカスタマ同時にビデオ再生の要求を満たせることを示した．また，本システムにおける SBus でのボトルネックを明らかにした.
今後は，アクセスパターンにトレースデータを用い本システムの特性をより明らかにしていくと共にモデルと開発中の実機に対する整合性を高めていくつもりである.

謝辞
本研究の一部は文部省科学研究費(重点領域研究(1)課題番号 0423510 「超並列ハードウェア・アーキテクチャの研究」)によります.

参考文献

1) 文部省重点領域研究「超並列原理に基づく情報処理基本体系」第 6 回シンポジウム予稿集,pp4-42-4-49,Mar 1995.
2) 文部省重点領域研究「超並列原理に基づく情報処理基本体系」第 3 回シンポジウム予稿集,pp257-279.Sep 1993.
3) 中條 拓伯, 松田 秀雄, 金田 悠紀夫, “超並列計算機におけるワークステーションクラスタ・ファイルシステム”,情報処理学会計算機アーキテクチャ研究会報告 ARC-107-24,Jul 1994.
4) 岡田 勉, 中條 拓伯, 松本 尚, 小畑 正貴, 松田秀雄, 平木 敬, 金田 悠紀夫, “超並列計算機 JUMP-1 における入出力サブシステムのアクセス方式”,情報処理学会計算機アーキテクチャ研究会報告 ARC-107-23, Jul 1994.
5) M.H.マクドゥガル著, 小林誠訳, シミュレーションによるコンピュータシステムの性能評価：テクニックとツール, 工学社, Apr 1990.