大容量記憶システムの利用特性と動作解析

金澤正憲，飯田記子，柴山　守，萩原　宏
（京都大学 大型計算機センター）　（京都大学 工学部）

1. はじめに

プロセッサの高速化と主記憶量の増大により大型計算機システムの処理能力は飛躍的に向上し、数多くの利用者に種々のサービスをすることができるようになった。京都大学大型計算機センターでは、システムのレベル・アップにつれて、会話型処理の利用者が急激に増大してきた。会話型処理では、バッチ処理におけるカードの役割をオンライン・ファイル装置が担うことになる（ここでいうオンライン・ファイル装置とは、計算機と直接的に接続され、人手を介さずにアクセスできる装置のことを指す）。更に、利用者自身による磁気テープのハンドリングは煩わしいため、バッチ処理においてもオンライン・ファイル装置を指向するようになってきた。そういう装置の一つとして大容量記憶システム（Mass Storage System，以後ＭＳＳと略す）がある。センターでは、オンライン・ファイル指向に対処するために昨年４月にＭＳＳを導入した。

センターでは、従来から、共用ファイルと呼ばれるＤＡＳＤを利用者に用意している。共用ファイルは同一ボリュームを多数の利用者が共用するが、ＭＳＳでは、ボリューム単位で希望する利用者に割当てる方式を採用し運用している。このような運用の下でＭＳＳの動作解析を行うために、まず、共用ファイルに格納されるデータセットとＭＳＳに格納されるデータセットを利用者がどのように区別して使用しているかを大きさという点から調べた。次に、ＭＳＳを運用する上で必要となる各種のサービスを提供するユティリティであるＭＳＳ ＡＭＳコマンドを利用した簡単なソフトウェア・モニタによりＭＳＳの動作状況を測定した。ここでは、それらの方法と結果について述べる。

2. ＭＳＳの構成と動作環境

センターの計算機システムは、ＦＡＣＯＭ Ｍ200 ４CPUのＴＣＭＰ（Tightly Coupled Multiprocessor）主記憶３２ＭＢを中核として構成されている。オペレーティング・システムは、OSⅣ/F4と呼ばれ、１論理仮想空間の大きさが１６ＭＢの多重仮想記憶方式であり、バッチ処理とＴＳＳ処理を同時に行っている。

利用者用ＤＡＳＤ（共用ファイル）としては、５０００ＭＢ（２００ＭＢ×２５台）*用意され、利用者は必要に応じて、必要な大きさのデータセットを確保することができる。ＭＳＳの容量は１０２ＧＢ（１０２０ボリューム）で、ステージング・ディスクとして２００ＭＢのＤＡＳＤを４台用いる。ＭＳＳは１ボリューム（容量１００ＭＢ）を単位として希望する利用者に割当てて運用される。ＭＳＳのボリュームをＭＳＶ（Mass Storage Volume）と呼ぶ。

測定時において、ＭＳＳの利用は３５０ボリュームであり、計算機システムの

＊測定当時。現在は、７９２５ＭＢ（３１７ＭＢ×２５台）。

負荷は、次のとおりであった。1日当り（約12時間運転）の処理件数はバッチジョブ 3300～4000件、TSSジョブ約2100件、バッチジョブの多重度は12～15、TSSの同時アクティブ端末数は140～160台，アクティブなリモートバッチ端末数は約20。なお、利用者数は約3000である。

3. MSSの動作と解析プログラム（MAP）について

MSSにあるデータセットへアクセスすると、MSSは次のように動作する。

(i) データセットの割当て（Allocation）

あるデータセットを割当てる時、指定されたMSVのVTOCページ*をステージング・ディスク上に読込み（ステージングし）、そのデータセットが存在するかどうかを調べる。このとき、VTOCページはアクティブ状態になる。

(ii) データセットのオープン

データセットのオープン時に、対応するMSVからデータセットをステージングする。このとき、データセットを含むいくつかのページがアクティブ状態になる。

(iii) データセットのクローズ

MSSが特に動作することも、ページの状態が変化することもない。

(iv) データセットの割当て解除

データセットの割当て解除時に、データセットを含むページはインアクティブ状態に変る。VTOCページは、同一MSV内の他のデータセットがすべてアクティブ状態でない時、インアクティブ状態になる。

(v) ジョブ・ステップの終了時

書込み動作により内容の変更されたシリンダのみをMSVへ書き出す（デステージングする）。

(i) (iii)のステージングは、ステージング・ディスク上に該当するページが既にある場合には行われない。また、新しくデータセットを作成する時は、領域割当てのみ行われる。領域割当て時に、ステージング・ディスク上にフリー・ページ**がない場合はインアクティブ・ページから割当てられる。

MSSの動作解析を行うためには、仮想装置にマウントされている（1ページ以上がアクティブである）MSVの動的な変化、ステージング・ディスクの負荷状況、および、ページ状態の遷移状況を知る必要がある。このために、MSS動作解析プログラム（Mss Activity trace Program，MAPと呼ぶ）を開発した。MAPは、図1に示すように、MSSDUMPプログラムとMSSMAPプログラムから構成されている。

MSSDUMPプログラムは、ステージング・ディスク上にあるMSC（Mass Storage Control）テーブルからページとMSVに関する情報を収集するために、サービス・ユティリティ（MSS AMSコマンド）を一定時間毎に起動する。MSS AMSコマンドは収集したデータを予め指定されたデータセットに出力する。MSS DUMPプログラムはオペレータが投入したSTART指令により起動された一つの

---

* VTOCを含むページ。1ページは8シリンダ（約2MB）で割当て時の単位である。
** 仮想ページ（MSVのページ）が割り当てられていないページ。

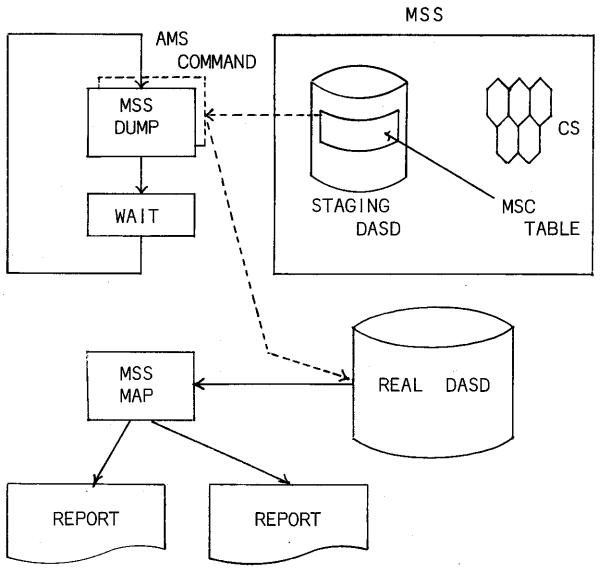


図1　MAPの構成

サブシステムとして動作する。任意のMSS AMSコマンドを起動することが可能であるが、動作解析のためにはDUMPMSSコマンドとTRACEコマンドを用いた。

MSSMAPプログラムは、MSS AMSコマンドによってデータセットに収集されたページやMSVに関する情報をMSV毎に分類・整理し、表やグラフを作成する。種々の目的に合わせたいくつかのルーチンによりMSSMAPプログラムを構成した。MAPによって得られた結果を次節に示す。

4. 測定結果

4.1 利用特性

利用者が共用ファイルとMSVをどのように使い分けているかを調べるために、それぞれにおけるデータセットの大きさを取上げた。図2にMSV上の、図3に共用ファイル上のデータセットの大きさの分布を示す。共用ファイルではデータセットの大きさは1MB以下の割合が多く、100KB以下で既に全体の60%余りを占めている。MSVでは2～10MBの大きさのものが多い。即ち、共用ファイルの10倍前後の大きさのデータセットがMSVに作成されている。ところで、MSSの運用に先立ってセンターは、MSVに格納すべきデータセットの容量は、余り小さくても、また20MBを超えるように大きくても適していないことを利用者に通知した。この結果、MSSの特徴に対して適した使用がなされているものと考えられる。

一方、MSV上のデータセットの内容は、利用者のプログラミング言語はFORTRANが殆んどであることと、データセットのレコード形式から推測して、FORTRANプログラムの実行時のバイナリ形式のデータ（レコード形式がVB, VBS, U）が約60%，ソース・プログラムおよびカード形式のデータ（レコード形式がF,

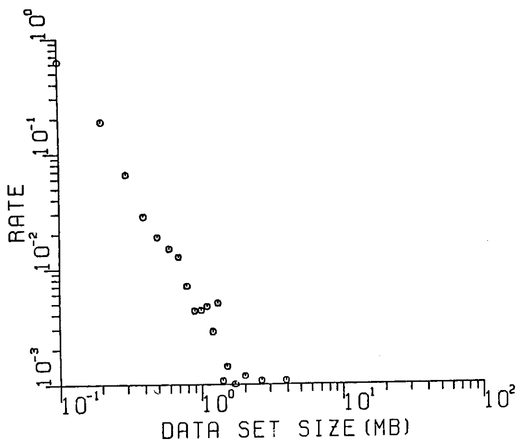


図2. MSV分布

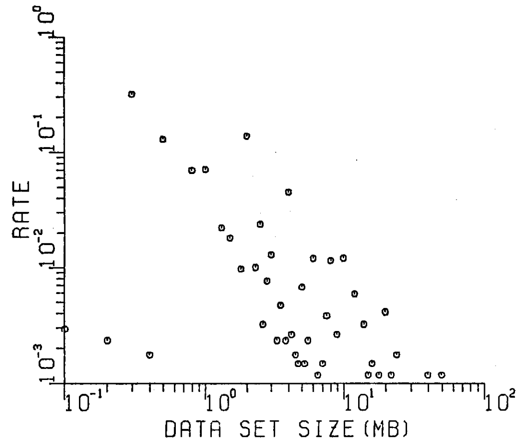


図3. 実用ファイル分布

FB）が30％強であった。データセットの編成については、80％強が順編成、約15％が区分編成であった。

なお、これらの統計量は、MAPではなく、各MSVのVTOCを読込むプログラムを使用して得たものである。

4.2 動作解析

MAPによって測定された結果を示す。図4は、ページの状態をMSV毎に分類・整理して、表示したものの一部である。アクティブ・ページを一つでも含む場合はA，アクティブ・ページがなくてインアクティブ・ページを一つでも含む場合はI，その他は・で表示されている。1文字がMSSDUMPプログラムによる測定周期（10分間）に相当する。このデータから、同時にAの状態にあるMSVの数を求めれば、仮想装置に同時にマウントされている台数を知ることができる。この値は、システム編集において定義する仮想装置の台数に関して有益な情報を与える。

テーブル NO.	ボリューム 通番	ページ状態
24	M002	....
25	M004	III...
26	0331	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..AAIAIIIIAAAIAIAAIIIIIAAAI
27	1189	IIIIIIIIIIIIIIIAAIAAIIIAAIIIIIIIIAAAIAAAIAAAIIIIAAAIAAAAAAAAAAA
28	2468	IIIIIIIIIIIIIIIIIIIIIIIIIIII................................
29	0660	IIIIIIIIIIIIIIIIIIIIIIIII........AAAIIIIIIIIIIIIIIIIIIIIAAIIII
30	2265	.............AAAAAAAIIIIIIIIIIAAIAAAAAAAAAAAAAAAAIAAAAAAAAAAAIIIII
31	B032	..AAIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.............
32	0818	IIIIIIIAAIIIIIAIIIAAAIIIAAAAAAAIAAAAAAAAIIIIIIAAIAAAAAAAIIIIIIIIAAII
33	0733	..................IIIIIIIIIAIIAIIIAIIIIIIIIIIIIIIIIIIIIAAAAAIIIIIIIII
34	0937	....
35	0754	IIIAAIIIIAAAAAAIAAAIIIAAAAIIAIIAAIIIAIIAAAAAAAAAIAIIIIIIIIIIIIIAAIIIII
36	0302	IIIIAIIIIIIIIIIIIIIIIAAAAAIIAIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIAI
37	B033	...AIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.............
38	0105	IIIIIIIIIIIIIIIIIIIAAAIIIIIIIIIIIIIIAIIIIIAAIIAIIIIAIAIIIIIIIIIIIIIIIII
39	1937	......................................AAAAAAIAAIIIAIIIIIIIIIAIIIII
40	0821	IIIIIIIIIIIIIIIIIIIIAAAAIIIIIIAAAIIIIIIIIIIAIIIIIIIIIIIIIIIIIIIIIIII
41	M005	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..................
42	S000	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII...............
43	M001	....
44	M009	IIIIIIAIIAAIIIAAAIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

図4. MSV毎のページ状態

ステージング・ディスク上のページ（実ページ）とMSVに関する運用時における動作状況の測定結果を図5に示す。ステージング・ディスクには200MBの装置が4台用意されているため、実ページは合計388ページある（ただし、MSCテーブル、使用不可ページを除く）。図5によれば、運用開始時の過渡的な部分を除いても約80％が再利用可能なインアクティブ状態にあり、ステージング・ディスクの容量にかなりの余裕が見られる。

次に、ステージングおよびデステージング時の量、所要時間、回数などを表1に示す。更に、DRD（Data Recording Device）で読み書きされる1シリンダ当りの平均転送時間を求めた。平均転送時間とアクティブ・ページの割合との関係は図6に示すように比例的である。一方、平均転送時間は一般的にステージング・ディスクのビジー（使用）率が高くなれば、比例的に長くなる。従って、図6から見て、ビジー率とアクティブ・ページの割合も比例関係にあるとして良い。この結果、ステージング・ディスクのビジー率をモニタリングすることが何らかの理由で難しい場合、アクティブ・ページの割合を用いて、ある程度推測できると思われる。

表1. MSSの動作状況

CASE	LOAD. VOL.	STG. CYL.	STG. TIME	No. of STG.	DSTG. CYL.	DSTG. TIME	No. of DSTG.	VTOC No. of STG.	VTOC No. of DSTG.	ACT. PAGE
1	333	713	3,389	243	622	1,734	95	41	77	9％
2	843	2,138	10,749	686	1,502	4,062	188	82	143	20％
3	1,471	3,531	20,246	1,265	2,618	6,684	239	190	188	33％

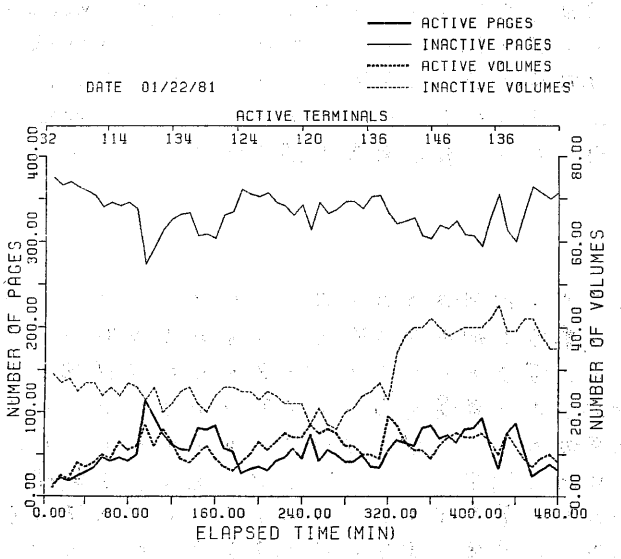


図5. ステージング・ディスク使用状況

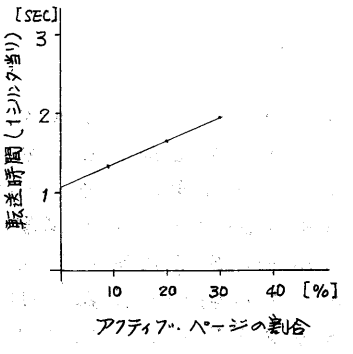


図6. 平均転送時間

データセットの割当て時とオープン時には、MSVのVTOCが必ず参照される。このため、VTOCページがステージング・ディスク上に残っていることは、データセットの割当てと参照時の応答時間の短縮に対して有効に働く。そこで、VTOCページに着目して、それがインアクティブ状態で残存している時間、残存して再利用されるまでの時間について解析した。その結果を図7，図8に示す。それぞれ、アクティブ・ページの割合が、(a)は9%，(b)は20%，(c)は33%の場合である。

図7は、インアクティブとなった時点から他の仮想ページに割当てるためパージされるまでの時間の分布である。平均値と分散から計算した正規分布を図中に実線で示した。(a)，(b)の場合には、比較的正規分布に近い形であるが(c)の場合にはむしろ指数分布に近くなっている。

図8は、インアクティブとなった時点から再びアクティブになる（再利用される）までの時間の分布である。平均値から求めた指数分布を図中に実線で示した。いずれの場合にも、20分以内に再利用される割合が比較的多いことが判った。これは、MSVを貸与するという運用であるために、ある利用者のデータセットは同一のMSVに存在することが殆んどであるということに因るものと考えられる。即ち、ある一つのデータセットを使用した後に、次に同じMSVにある別のデータセットを取扱うことが多いためと推測される。

図7，8の(c)の場合、1時間以上経過するとインアクティブ・ページは殆んどパージされていることが判る。また、再利用される割合は10分以内が40%にも達している。図示していないが、この場合の図5に相当するステージング・ディスクの使用状況を調べると、インアクティブなMSVの数が0になっている部分があった。また、アクティブ・ページが75%になる状況も見られた。これらの結果，(c)の場合はかなりの負荷がかかった状態であると言える。

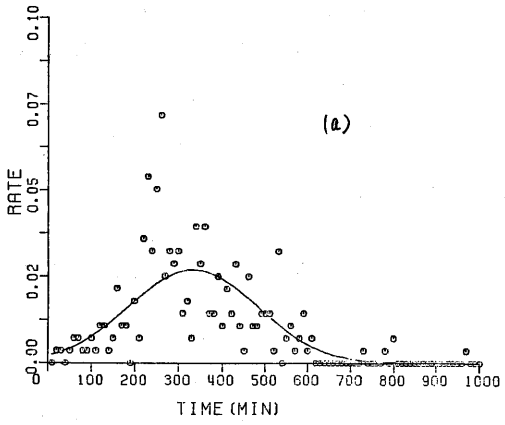
以上の結果は、MSS AMSコマンドのうちのDUMPMSSおよびTRACEコマンドを用いて得られる結果の一部である。DUMPMSSコマンドによって付表1に示すようなMSCテーブルの内容を収集することができる。また、MSSDUMPプログラムから他のMSS AMSコマンドを起動すれば、各種テーブルの情報の検査を行わせたり、MSSのLRU調整値の表示や変更を行うこともできる。

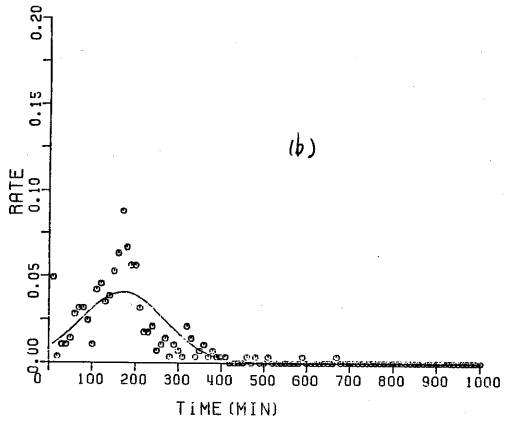
センターに設置されているMSS（FACOM 6450）のハードウェア仕様を付表2に示す。IBM3850と互換性があり、転送速度が高速化されている。

5. おわりに

オンライン・ファイル装置を指向する利用者のニーズにこたえるため、センターではMSSを導入し、ボリューム単位で利用者に割当てる方式で運用している。このような運用の下で、MSSにどの程度の大きさのデータセットが格納されているかを調べた。その結果、比較的小さい容量のものは共用ファイルを、大きいものはMSSを利用しており、利用者がうまく使い分けていることが判った。

MSSの動的な振舞を測定するために簡単なソフトウェア・モニタを開発した。このソフトウェア・モニタは、MSSを運用する上で必要なサービスを提供するMSS AMSコマンドをうまく利用して非常に簡単に作成することができた。さらに、MSS AMSコマンドを使い分けることにより、多様な情報を効率よく収





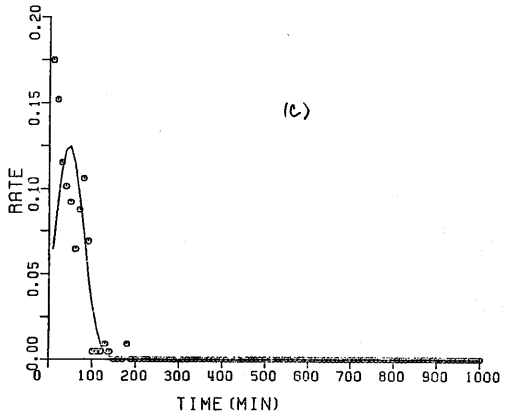
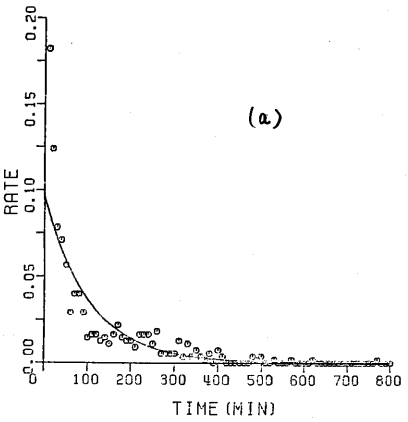
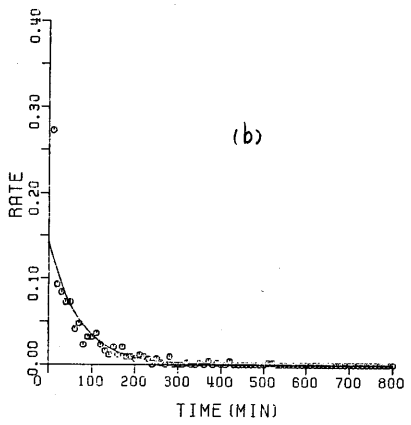


図7. パージされたVTOCページの分布





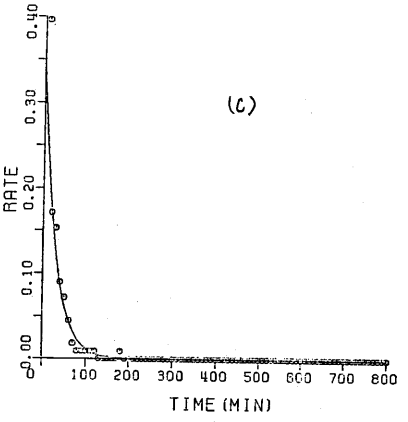


図8. 再利用されたVTOCページの分布

集することができた。また、対象がMSSという動的変化の遅い装置であるために、事象の変化を追うのではなく、一定時間毎の監視によって十分な結果が得られた。

MSSMAPにより分析した結果、現在のシステムはステージング・ディスクの容量に十分余裕があり、再利用の割合も高かった。しかし、アクティブ・ページの割合が30%になると、比較的短い時間内においてもかなりのページがページされることが判った。このことから、アクティブ・ページの割合が50%を越えるとボトルネックが生じ得ると推測されるが、この点に関しては更に高い負荷の時の測定結果を得る必要がある。

MSSの運用に関しては、マイグレーション(migration)ボリュームとしての利用[2]，大容量のデータセット、例えば、データベースでの利用[9]等が考えられる。これらの利用方式の導入によって変化するMSSの動的振舞を監視することは、ステージング・ディスクの台数の検討やSDC(Staging Disk Controller)の構成の検討に有益な情報を与えるであろう。

最後に、ご助言をいただいた本センター北川一助教授、ご協力をいただいたセンターおよび富士通関係各位に謝意を表します。

[参考文献]

1) Boyd, D.L. : Implementing Mass Storage Facilities in Operating Systems, Computer Vol.11, No.2, 40-45, 1978.
2) Considine, J.P. and J.J. Myers : MARC: MVS Archival Storage and Recovery Program, IBM Syst. J. Vol.16, No.4, 378-397, 1977.
3) Hempy, H : IBM 3850 Mass Storage System, Performance Evaluation Using a Channel Monitor, in Computer Performance, K.M. Chandy and M. Reiser, Eds., North-Holland, Amsterdam, 177-196, 1977.
4) Sekino, A. and T. Kitamura : Architectural Considerations of the NEC Mass Data File Subsystem, NCC Vol.48, 557-564, 1979.
5) Tsuruho, S. et al. : Mass Storage Systems Performance Analysis Using a Queuing Model, 3rd USA-JAPAN Computer Conference 320-324, 1978.
6) Wimmer, W. : Über die Massenspeicherhierarchie als Teil eines Dateiverwaltungssystems am DESY-Rechenzentrum, Angewandte Informatik Vol.20, No.9, 381-388, 1978.
7) 伊藤，川田：超大容量記憶装置の動向，情報処理 Vol.19，No.5，465~471，1978
8) 伊藤，：外部記憶装置，情報処理 Vol.21，No.4，350~357，1980
9) 小沢他：大容量記憶システム(MSS)を用いた情報検索の性能について，情報処理学会第22回(昭和56年前期)全国大会講演論文集，1981
10) 柴山，金沢，飯田：京大大型計算機センターにおける大容量記憶システムの動作解析，ibid，1981
11) 富士通：FACOM 6450 大容量記憶システム解説書，マニュアル
12) 富士通：FACOM OSIV/F4 MSS 解説，マニュアル
13) 富士通：FACOM OSIV/F4 MSS AMS コマンド文法書，マニュアル

付表1　MSCテーブル構成表

No.	MSCテーブル
1	構成テーブル
2	診断／EC／オーバレイ領域
3	MSFセルマップ
4	メッセージ・バッファ　テーブル
5	マウンテド　ボリューム　テーブル
6	リカバリ　ジャーナル
7	スケジュール　キュー
8	スクラッチ　カートリッジ　リスト
9	ページ状況テーブル
10	ステージング　ドライブグループ　テーブル
11	トレース　テーブル
12	トランジェント　ボリューム　リスト
13	ベリフィケーション　テーブル
14	仮想ボリューム　アドレス　テーブル
15	仮想ボリューム／ボリューム識別子相互参照テーブル
16	ボリューム　インベントリ　テーブル

付表2　MSSのハードウェア仕様

容量	102 GB
DRC数	2
DRD数	4
最大データ転送速度	1136 KB／秒
平均アクセス時間	9～13 秒
最低アクセッサ時間	4 秒
最低移動時間	5 秒