The problem of retrieving records at random from a very large archival data base has not previously been effectively soluble by data processing techniques. This paper describes such an application, in which a large bank in the United Kingdom uses the IBM 3850 Mass Storage System for storing and retrieving customer account statements.

National Westminster Bank mass storage archiving

by C. M. Gravina

National Westminster Bank is one of the largest clearing banks in the United Kingdom, with approximately 2500 branch offices which service nine million customer accounts. Two computer centers provide a centralized accounting service for all branches, including on-line inquiry to a 10 000-megabyte data base containing account information.

The system stores all account transactions until the customer receives a statement, or printed copy of his ledger. Statements are printed at frequencies ranging from daily to every six months, depending on customer requirements, and also are printed on demand. The Bank is legally obliged to retain copies of all statements for at least ten years and receives a significant number of customer queries about statements printed up to three years previously.

The archiving problem

The Bank prints approximately 100 million statement sheets per year and receives up to 8000 requests for statement information every day. Duplicate copies originally were filed in the branch offices, but it soon became apparent that the expense of filing and retrieving statements was high in terms of staff time, paper cost, and storage space. The inquiry process was particularly time consuming, as many statement sheets often had to be searched for but one missing item.

microfilm bureaus

A branch terminal network was implemented in 1970, and the Bank took that opportunity to centralize statement storage using microfilm. The statements were written to microfilm cassettes using computer-output-to-microfilm equipment. The cassettes were stored in one of two microfilm bureaus, which were linked to the branch network by a number of typewriter terminals. The bureaus had equipment for rapidly retrieving microfilm cassettes, and also high-speed film viewers and printers. When a branch sent an inquiry to one of the bureau terminals, it was processed manually.

This centralized microfilm-archiving system provided significant savings and removed the problem of paper storage in the branches. In addition, the total system costs were easier to quantify than if they had been distributed among 2500 branches.

Implementation of the system raised several new problems, however. The first was the need to index the stored records. When paper was used, a customer folder contained the account history in chronological order, and any new sheets were collated with the old ones. This procedure was not possible with microfilm, as one cassette could hold only records produced on a given date, and the records for one account would be scattered among many cassettes. Record retrieval required knowledge of the storage date and of the cassette and frame numbers within that date. This problem was partly solved by using an on-line microfilm index, and bureau inquiries had a microfilm reference added automatically by the system. However, the references could be held on line for only a limited period before they themselves had to be microfilmed to save storage space.

The second problem was that users no longer had control over their own data. They had to rely on a remote system which was operated manually and thus was subject to human error. Requests could be delayed or lost through system failure; and branches were tempted to submit hard-copy requests to back up specific requests.

Moreover, it soon became apparent that archiving costs still were high and were increasing, with more than 200 people employed in the processing of inquiries. For these reasons the Bank pursued alternative solutions, and in October 1974, when IBM announced the 3850 Mass Storage System (MSS),[1] National Westminster Bank set up a team to study the feasibility of using the MSS for archiving.

the 3850 Mass Storage System

The storage medium in the 3850 is two-inch-wide magnetic tape, rolled inside plastic cartridges. Each cartridge can hold 50 million bytes of data, and two cartridges form a logical volume. The cartridges are stored in library units, each of which can hold more than 200 000 megabytes of data.

Access to MSS data is in three steps. First, a cartridge is picked from the library and loaded into a data recording device. This step typically takes about ten seconds. The cartridge picking rate for a 3850 library unit cannot exceed about 400 cartridges per hour, but up to eight data recording devices can be configured so that cartridges can be processed in parallel. In the second step, the tape is searched and the requested data staged to disk storage in logical 250 000-byte cylinders. The staging rate is about one cylinder per second, plus several seconds for search time. The third step is accessing of the records on disk by the central processing unit, with a response time measured in milliseconds.

Evaluation of the 3850 for archiving

The objectives of the Bank's evaluation were to see whether use of the 3850 could reduce the high costs associated with archive inquiries and to determine whether customers would receive better and more reliable service. The first step was to configure a practical system from known application requirements.

size of the data base

The Bank generates about 25 000 million bytes of customer ledger data (in compacted format) each year. For legal reasons, as stated above, this data must be held for ten years, even though inquiries virtually cease after three years. It was tentatively decided to use the 3850 for storing data and handling inquiries on a three-year basis, and in addition to keep microfilm backup copies to satisfy the legal requirement for long-term storage.

inquiry system

The inquiry rate was expected to exceed the 3850 cartridge processing rate by a factor of three or more; thus it would be necessary to batch inquiries so that several requests could be satisfied with every cartridge processed. Response time, then, would depend on the number of cartridges awaiting processing, rather than on the number of outstanding requests. Processing times could be improved by using one of the more powerful 3850 models, which allow several cartridges to be processed in parallel.

The critical design parameters were thought to be:

- Levels of inquiry priority
- Response times required for each priority
- Mean inquiry rate
- Inquiry distribution by time
- Inquiry distribution across the data
- Number of record retrievals per inquiry.

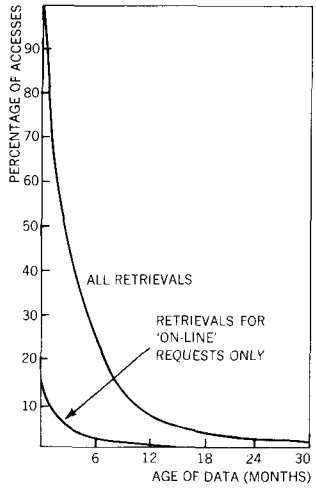
Inquiries received by the microfilm bureaus were surveyed in order to measure these parameters, and if possible to ascertain the

Bank's real needs in providing efficient service to its customers. An inquiry from a terminal in one of the Bank's branches might request an on-line response giving, for example, an account balance on a certain day or details of a check transaction. Branches can also request duplicate copies of a customer's statement, which are printed and delivered overnight.

The on-line response time was found to be noncritical provided that the inquiry itself was acknowledged immediately. A response time of less than 30 minutes would satisfy most users and would permit all inquiries to be returned the same day they were received, as the branch network remains open for an hour after the inquiry cut-off time at 1600 hours. Hard-copy requests have to be processed by 1900 hours to meet overnight delivery schedules.

The microfilm bureaus handled up to 8000 requests per day, of which 75 percent were for hard copy. It was not possible to determine how many of the latter were duplicate requests, which would not be made in a computerized system, nor how many would change to on-line requests. Given the long response time, it was not necessary to establish the peak-second or peak-minute rate, but only that of the peak hour, which was about 20 percent of the daily rate. There was a long-term growth trend, and there were peaks at year-end and at tax-year-end. Inquiry rates could be increased substantially by external factors such as tax legislation.

Figure 1 Record retrieval distribution—percentage of accesses as a function of data age



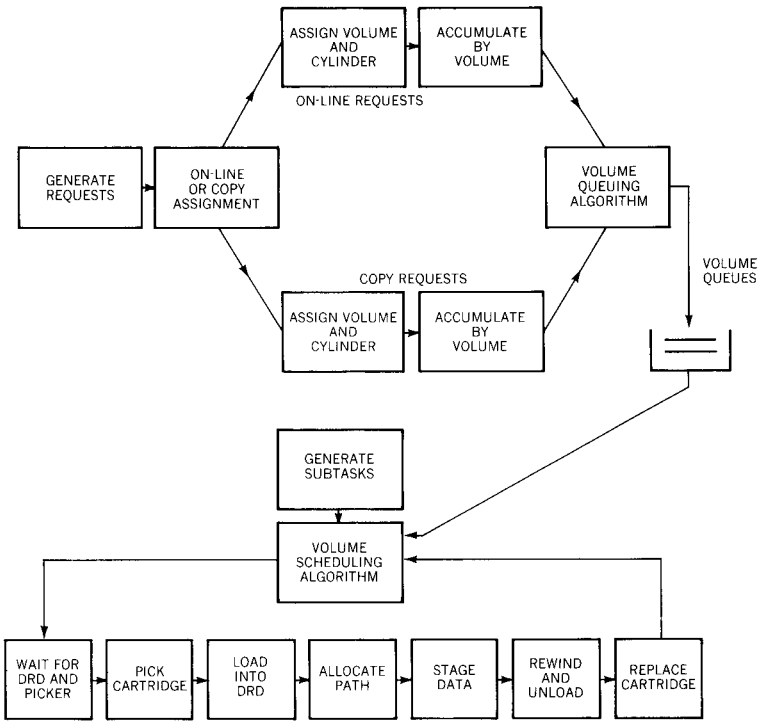
Requests were not distributed randomly across the data, but were found to decrease with increasing data age, as would be expected (see Figure 1). Note that the 25 percent of on-line requests generated only 16 percent of the record retrievals, indicating that branches were requesting hard copy whenever they wished to search for an item across several statements. Hard-copy requests were received for data up to three years old, but virtually all on-line requests could be serviced by data less than one year old. This fact is significant in that most activity is confined to a small part of the total data base. (The use of direct-access storage for the most recent data was considered, but was not economically justifiable in comparison with mass storage.)

Some requests could be satisfied by a single record retrieval—for example, a request for an account balance on a given date—but some requests could involve searching through many hundreds of records for a missing item. There were 2.1 retrievals for each average request; requests for large searches were controlled by limiting the number of records per request, so that users were forced to make searches in stages.

3850 configuration process

The performance of the 3850 is controlled by the following configuration alternatives:

Figure 2 Application simulation model



NOTE: DRD = DATA RECORDING DEVICE

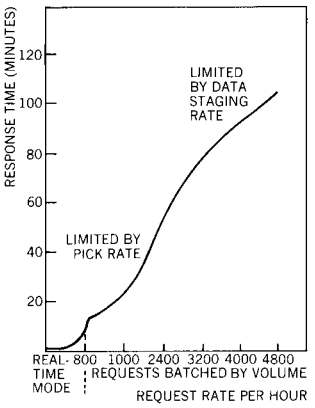
- The number of libraries (one or two), which determines the cartridge picking rate and also affects total storage capacity.
- The number of data recording devices per library, which determines the number of cartridges that can be processed in parallel.
- The number of staging paths to disk from each library.
- The number of staging drives.

These parameters must be determined from the size of the data base, the inquiry rate, the distribution of requests across the data, and the response times required. MSS storage capacity is unlikely to be a significant factor, though for small data bases, disk storage may be more attractive economically than mass storage. The staging path and drive requirements can be calculated readily from the inquiry rate and again are unlikely to be critical at the relatively low data rates needed by inquiry applications.

The cartridge picking rate and requirements for data recording devices can be estimated by iterative methods from the inquiry rates and probability distribution. Performance is also affected by volume and request processing algorithms, however, and is best optimized by modeling techniques, as the factors involved tend to be interactive.

A simulation model of the Bank's application was developed, with assistance from IBM's General Products Division Laboratory in Boulder, Colorado, in defining the 3850's performance characteristics (see Figure 2). The model was used to determine a suitable MSS configuration, and also to show the sensitivity of that configuration to increasing request rates, changing request distributions, varying numbers of volume-processing subtasks, and 3850 component failures. The model was written using the General Purpose Simulation System (GPSS).[2]

Figure 3 Performance simulation of random retrievals from a 45 000-megabyte MSS data base



The simulation runs showed that response times of less than two minutes were possible for inquiry rates of less than 300 per hour (see Figure 3). As the inquiry rate increased, the response time increased rapidly because the 3850 could no longer handle inquiries in real time, and a queue of volumes awaiting processing began to build up. With further increases in the inquiry rate, the response time leveled off because there was a greater probability that a required volume already would be in the processing queue.

To optimize performance for this particular application, three alternative methods of volume processing were modeled. The goal was to find a compromise between the need for fast response to on-line requests and the need to meet end-of-day processing deadlines for hard-copy requests. The methods modeled and the results were as follows:

- The first request for a given volume would place the volume on a processing queue. When the volume reached the head of the queue, all requests received for that volume would be processed. Maximum response time on the chosen configuration was estimated to be 60 minutes for this option, but end-of-day processing time was relatively short.
- Only on-line requests would be satisfied during the day, and all hard-copy requests would be processed at the end of the day. Estimated on-line response time was five minutes, but end-of-day processing time was four hours, which was unacceptable.
- Volumes receiving on-line requests would be given priority. However, all requests for these volumes would be processed. This option restricted the active number of volumes to about one third of the total. Response times were less than 30 minutes, and end-of-day processing was about an hour. This option best met the application requirements and was adopted.

After considering the effects of possible hardware failures, the probability of growth in the inquiry rate, and the potential for other applications, a mass storage system was designed to handle an inquiry load for one National Westminster computer center of up to 2400 retrievals per hour. The system comprised two 3850 Model A2 libraries and 1600 megabytes of disk storage, with two

sets of cartridge pickers, eight data recording devices, and four staging paths. It provided 200 000 million bytes of storage space.

Archive systems design

Archive inquiry systems typically differ from other applications in having very large data bases, very low inquiry rates, and relatively long response times. Some of the resulting design problems are:

- The need to segment the data base into manageable units for processing and data recovery.
- The need to establish checkpoints to prevent loss of inquiries, as the user cannot distinguish between a long response time and a system failure.
- The need for application independence. Though the application is nominally part of a teleprocessing system, it can still be processing a request backlog when the network is closed down at the end of the day. As an application of lower priority, it should not adversely affect the network's performance or reliability.

The systems designing of the Bank's application was carried out in three phases—the data base, the indexing system, and the inquiry system.

data-base design

Archive data, by definition, is never updated but always added to. Once stored, records must be protected against being updated or deleted. Any data loss must be localized and readily recovered from by using off-site backup copies.

The Bank's data base was expected to grow by 500 megabytes per week to a peak size of 75 000 megabytes after three years. The oldest records would then be dropped at regular intervals. The data base would be segmented into 100-megabyte data sets, one per logical volume of two MSS cartridges. A backup copy of each data set would be held on another mass storage volume in a secure bunker. Experience has shown that whole-volume backup is the best approach with mass storage. Because of a high level of data redundancy with automatic error correction, there is virtually no loss of individual records due to media failure. Cartridges or volumes can suffer physical damage, however.

addition of new records

New records can be added to a data base either by collating them to keep each customer's records in the correct sequence and inserting them directly, or by writing them at the end of the data base.

Direct insertion would simplify the indexing system and permit requests to be satisfied by a single retrieval. However, the in-

sertion of many records requires a high processing overhead, and the whole data base might have to be backed up after each update. Requests would be distributed over the whole data base rather than concentrated on a small part, and there would be some risk of record loss or corruption during the insertion process.

Addition of records at the end of the data base would require the creation of a complex indexing system but would simplify backup procedures and optimize inquiry performance.

A compromise design would be possible in which records would be collated in order within periodic batches. At the end of each period, a new batch would be added to the data base. The batch size would be limited largely by the updating and backup processing times acceptable to the installation.

It was decided to update the Bank's data base in weekly batches of about 500 megabytes. Records are always available for inquiry, as the on-line data base is itself reorganized once a week.

data structure

A bank's customer ledger sheets are difficult records to organize into a data base. They must be located by the customer's account number within a data set keyed by storage date. Consecutive records for a given account may be created at intervals ranging from one day to six months, or at random intervals requested by the customer. One record can consist of a single ledger sheet or more than a thousand, and individual sheets are of various lengths.

When retrieving a record, the requestor usually does not know the precise record creation date, so the system must find records that span a range of dates. In optimizing the performance of a mass storage archiving system, there is the additional requirement of having to know the physical location (volume and cylinder) of a record before retrieval. To avoid degrading performance, it is also advisable to prevent associated records from overflowing across volume and cartridge boundaries. MSS performance considerations also make it necessary to avoid data structures that require more than one access to retrieve the desired record. Thus it was decided to store data in the form of simple ledger-sheet records, each approximately 250 bytes long.

Reducing the size of a data base can improve inquiry response times by reducing the number of cartridges being processed, so the Bank's records were compacted by:

- Removing blanks
- Removing decimal points from amounts
- Packing numeric fields and stripping leading zeroes
- Storing alphameric data in six-bit code

- Storing dates as binary digits
- Using flags to replace descriptors
- Avoiding data duplication by storing common fields, such as the branch name, only once per cylinder.

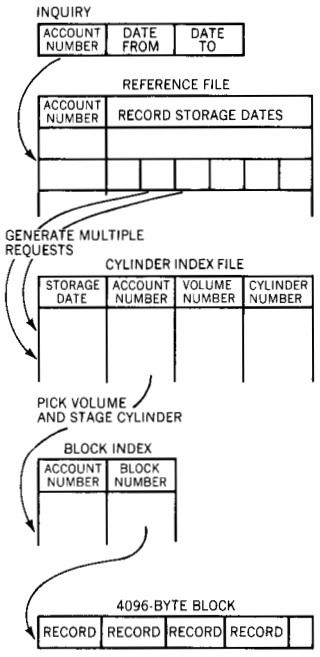
The access method chosen was VSAM (the Virtual Storage Access Method)[3] which supports record retrieval by relative byte address within a data set. The relative byte address is provided by either a VSAM index or a user-written index.

indexing system

The data indexing method used in this application had to satisfy data requests of the form *retrieve all records for account X between dates Y and Z,* where the precise storage date of each record and the number of records are unknown. The index files had to be small enough to be held on direct-access storage and be updated easily.

A two-level indexing system was adopted. The first level, the reference file, provided the storage date of every ledger record in the MSS. The second level, the cylinder index file, was used to locate a given record using a key consisting of the storage date concatenated to the account number.

Figure 4 Data indexing system



The indexing system is shown diagrammatically in Figure 4. The reference file contains a record for every account number, with a two-byte entry for every statement from which the storage date can be calculated. This file is updated every week and will grow to nearly 600 megabytes after three years. Probably it would be possible to segment the file by age to reduce updating time and on-line storage space. The older entries would be stored in the MSS, and a new scheduling algorithm would be needed to minimize staging.

The cylinder index file, which occupies about seven megabytes of disk storage, provides an index to the whole MSS data base. It gives better performance than the data-set indexing available with standard access methods, as it removes the need to stage and open an index data set each time a volume is processed. The cylinder index points to the volume and cylinder that contain a desired record, thereby also providing the information needed for data-set allocation and MSS performance optimization. Note that storage dates do not span volumes or cartridges. A new set of volumes is used for each weekly storage run to minimize backup and recovery problems.

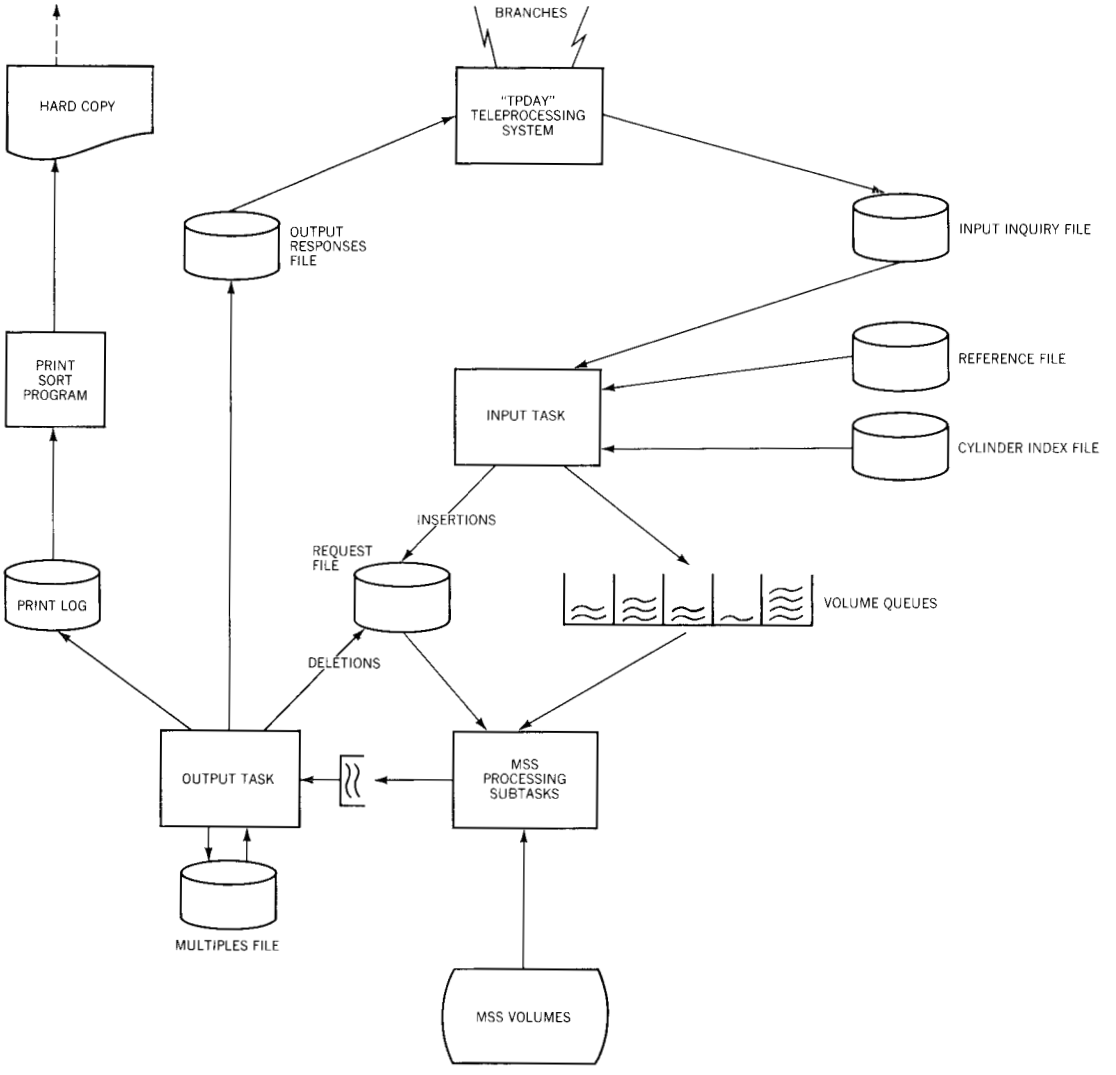
The cylinder index must be complemented by block index records in every data cylinder. These records contain pointers to each data block in the cylinder. The cylinder index file also contains flags to indicate when a customer's records overflow a cylinder boundary. These flags ensure that both cylinders are staged by the MSS.

A number of constraints must be considered in designing an MSS-based inquiry system. Though nominally a real-time application, such a system has such long response times that it must be isolated from the main teleprocessing system and provided with its own message logging and restarting facilities. Requests for data access across a large number of volumes must be generated from incoming inquiries, then stored until needed and retrieved in batches by volume. The volumes themselves must be processed in a logical order within several levels of priority, and a response to the user must be composed from the multiple records obtained, over perhaps 30 minutes, from several different volumes. User demands for printed output must be processed randomly during the day. It is necessary to sort the output and print the records in batches during the evening to meet different delivery deadlines. Lastly, the application must attempt to optimize throughput by taking account of unique MSS characteristics.

Figure 5 outlines the Bank's inquiry system as designed to take into account the above constraints. The system has the following key characteristics:

- *Application isolation* is obtained by using two disk files for communication with the Bank's teleprocessing system for inquiries and responses. The two applications can be started up and closed down independently, and they can run in the same or different processors.
- *Request generation* is performed by the input task. For each inquiry, a number of requests are generated as a result of accessing the reference file. Each request has volume and cylinder information added from the cylinder index file and is then stored on the request file in volume and cylinder sequence. If more than one request is generated for an inquiry, a base record is written to the multiples file to warn the output task that several output records are to be expected.
- *Volume queues,* in the form of multiple logical FIFO (first-in-first-out) queues, are provided by an in-core table. The highest priority is given to volumes for which on-line requests are scheduled, and any of several lower priorities are used for hard-copy requests. When a request is generated, the queuing table is checked to see if the volume is on the correct queue. A volume is always placed on the queue for its highest-priority request. On a system restart, the volume queues can be rebuilt by processing the request file sequentially.
- The *output task* processes records retrieved from the MSS by generating a suitable on-line response or a hard-copy print image. Then it deletes the request from the request file. The multiples file is used to accumulate multirecord responses, which are processed only after all records have been collected.
- *Printing* is initiated by an operator's command at the end of the day. It is done in batches, each level of printing priority

Figure 5 MSS application flow diagram



being cleared by the MSS so that output for branches far from the computer center is dispatched earlier than output for nearby branches.

MSS processing subtasks

Full MSS utilization can be obtained by employing multiple subtasks, each subtask being responsible for processing one mass storage volume (two cartridges) at a time. Each data recording device can process one cartridge, so the Bank's MSS, as installed with eight data recording devices, can process at least four volumes simultaneously. In practice it was found advisable to use six subtasks to load the MSS, as it is not always necessary to access both cartridges of a volume.

When a subtask is ready for work, it selects the volume at the top of the highest-priority queue and reads all requests for that volume from the request file. The volume is inhibited from being picked by another subtask while being processed. Any further requests have to wait until the volume is selected again by a subtask.

data-set processing

Normal data processing applications allocate data volumes to disk drive units for the duration of a job. The 3850 provides a large number of addressable virtual units, on which mass storage volumes can be logically mounted. Even the 96 units available to the application were insufficient, however, as 750 volumes sometimes have to be mounted during a day. To overcome this limitation, the volumes are allocated dynamically as needed, using the dynamic allocation function provided by the MVS operating system.[4] The allocation process reserves a unit, mounts the volume, and checks the volume label.

Only one volume can be allocated at a time, however. This limitation presented a potential performance bottleneck, as the MSS can take up to 20 seconds to mount a volume. The problem was solved by modifying the MVS allocation function to provide unit reservation only. Therefore several volumes can be mounted at the same time, and volume label checking is deferred until the data set is open.

MSS performance optimization

MSS performance in an inquiry application tends to be limited by the cartridge picking rate. Thus it is worthwhile to ensure that when a volume is processed, each cartridge is picked only once. Normally this can be achieved by sending the data-cylinder-stage order (ACQUIRE) to the MSS when the volume-mount order is sent.

The processing method used for each subtask is designed to optimize MSS performance by pre-staging data before a volume is processed. The method is summarized below:

- Select a volume for processing
- Select a virtual unit
- Demount any volume mounted on that unit
- Allocate the unit to the new volume
- Mount the volume in order to stage the volume label
- Issue ACQUIRE for cylinders wanted on cartridge 1
- Issue ACQUIRE for cylinders wanted on cartridge 2
- Wait for completion of cartridge 1 staging
- Open the data set (and check volume label)
- Wait for completion of cartridge 2 staging
- Process the requests for the volume selected
- Close the data set and de-allocate the unit
- Select the next volume.

The 3850 manages space on staging drives in such a way that newly staged data overwrites the oldest inactive data. If a wanted cylinder is still on a staging drive from a previous access, it will not be restaged. Thus if requests are found to be concentrated on relatively small regions of the data base, MSS performance can be improved by increasing the staging space.

The inquiry distribution for the Bank's application indicates that data cylinders rarely will be kept on disk long enough to be reused by other requests, as the 800 pages provided by the disk configuration would be overwritten every 20 minutes at the peak inquiry rate of 2400 accesses per hour. The volume labels of the more frequently accessed volumes, however, may be reused often enough to avoid deletion and restaging, providing a small improvement in performance.

If the data base is small, data reuse could result in a major performance improvement. Performance might then be affected by other resources such as the job catalog, the mass storage volume inventory, and the MSS control tables, all of which are held on disk storage.

data recovery

It is rare for mass storage volumes to become unusable because of cartridge damage or for records to become inaccessible. This is fortunate, as the backup copy of an archive file is likely to be stored off-site in a secure, fireproof vault, and recovery may take a long time. Volumes may become temporarily unavailable, however, if a cartridge is retained by a failed accessor or data recording device. Recovery usually can be effected within a half hour, which is almost within the allowable request response time. There is no need to send error messages to users: The application automatically flags the volume with *soft error* status, which prevents it from being processed.

The soft error is reset by the operator when the volume becomes available. Unusable volumes are changed to *hard error* status, which causes error messages to be sent in reply to affected requests. Those requests are then purged from the system.

Implementation

Designing, coding, and testing this application took five man-years of effort over ten months' elapsed time. To provide a substantial data base, data collection started twelve months before the rest of the project began and six months before creation of the data base could start. A systems test package was written to provide comprehensive checking of the data-base creation programs before they were used, as the later correction of any missed errors would have required a massive file re-creation effort. The

tests included a record-by-record check of the indexing system, as well as conversion of data-base records to their unloaded form for comparison with the input records.

It was difficult to provide suitable stress testing for the large MSS configuration before the application went on line. However, two weeks of parallel running resolved most of the hardware and software problems. During this phase, inquiries were received from branches and were processed by the MSS, but only the microfilm bureaus sent responses.

Conclusions

The application went on line in June 1977 and handled the full inquiry load a few weeks later. Performance was analyzed after six months to determine the system's effect on users. Response time was found to be better than expected, with many inquiries being processed in less than five minutes, as the mean record retrieval rate usually was below 1000 records per hour. The fast response apparently was due to two factors: there was a noticeable reduction in hard-copy requests, which often had been made as backup for on-line requests, and MSS availability was better than expected. A certain amount of redundancy had been allowed for in the MSS configuration but was rarely needed. Data availability also proved to be excellent, with no cases of unreadable records and only one or two cases of mechanical damage to cartridges.

It is not yet known what the inquiry rate will be for data more than three years old. That rate will determine the residual work load of the microfilm bureau. However, the primary objectives of the application—substantial cost savings and improved service to customers—have been amply fulfilled.

ACKNOWLEDGMENTS

My thanks to the colleagues with whom I was privileged to work on this project, in particular Mr. B. Pfeiffer, project leader, and Mr. F. C. Rogers, senior programmer, and to the National Westminster Bank for permission to publish this paper.

CITED REFERENCES

1. *Introduction to the IBM 3850 Mass Storage System (MSS)*, IBM Systems Library, order number GA32-0028, IBM General Products Division, MSS Support Center, Department 25E, P.O. Box 1900, Boulder, Colorado 80302 (July 1975).
2. *General Purpose Simulation System V Introductory User's Manual*, IBM Systems Library, order number SH20-0866, IBM Corporation, Technical Publications Department, 1133 Westchester Avenue, White Plains, New York 10604 (August 1971).

3. *OS/VS Virtual Storage Access Method (VSAM) Programmer's Guide*, IBM Systems Library, order number GC26-3838, IBM Corporation, General Products Division, Programming Publishing—Department J57, 1501 California Avenue, Palo Alto, California 94304 (April 1976).
4. *OS/VS2 Systems Programming Library: Job Management*, IBM Systems Library, order number GC28-0627, IBM Corporation, Programming Systems Publications, Department D58, Building 706-2, P.O. Box 390, Poughkeepsie, New York 12602 (January 1976).

Reprint Order No. G321-5079.