Title

SEARCHING THE NUCLEAR SCIENCE ABSTRACTS DATA. BASE BY USE OF THE BERKELEY MASS STORAGE SYSTEM

Authors
Saith, Gloria L.
Herr, J. Joanne.

Publication Date
1971-02-01

To be presented at the Association
of Scientific Information Dissemination
Centers (ASIDIC), Semiannual Meeting,
Feb. 23-25, 1971, Washington, D. C.

UCRL-20286 c2
Preprint

SEARCHING THE NUCLEAR SCIENCE ABSTRACTS DATA BASE BY USE OF THE BERKELEY MASS STORAGE SYSTEM

Gloria L. Smith and J. Joanne Herr

February 1971

AEC Contract No. W-7405-eng-48

LAWRENCE RADIATION LABORATORY
UNIVERSITY of CALIFORNIA BERKELEY

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Searching the Nuclear Science Abstracts Data Base by Use of the Berkeley Mass Storage System

Gloria L. Smith and J. Joanne Herr

Lawrence Radiation Laboratory
University of California
Berkeley, California 94720

February 1971

Introduction

The Lawrence Radiation Laboratories (LRL) at Berkeley and Livermore, California, are doing scientific research using computer facilities based on the largest computers made by Control Data Corporation (the CDC-7600 and CDC-6600), and the largest mass memory units available to date. The memory device was developed under a special contract with the International Business Machines Corporation, and is called the IBM-1360 Photodigital Storage System (Chipstore). The Livermore system has an on-line capacity of a trillion bits of data. The Berkeley memory is a third as large, and is used chiefly for the storage of data used to track and identify atomic particles in the Laboratory's high energy physics program. Recent work of the Information Research Group (IRG) has shown that the Mass Storage System (MSS) offers a speedy and inexpensive method of information retrieval from 4½ years of the Nuclear Science Abstracts data base.

Film Chips, the Basic Storage Media for MSS

The smallest physical unit of the system is a piece of high-resolution photographic film (35×70 mm) called a "chip." Nearly 5 million bits of information can be packed on a single chip. The chips fit into slots in a plastic box (about the size of a cigarette pack), and the boxes fit into compartments of movable trays resembling a stack of egg crates.

Recording, Processing, and Reading action of the IBM 1360

Data from the CDC-6600 computer are recorded on the silver halide film chips by means of a concentrated beam of electrons, shot from small tungsten filaments in the turret of an electron gun. For data recording, an individual blank chip is positioned in a vacuum chamber, and the information is written by repeated sweeps of the electron beam across the chip surface in boustrophedonic fashion.* The electron beam, in effect, "paints" lines of data as combinations of dark and clear spots, corresponding to the zeroes and ones of the binary language of the computer. Each chip is divided into 32 frames, and lines of data are recorded a frame at a time at a rate of more than one-half million bits per second. It takes about 18 seconds to record a chip, and less than 2½ minutes to process it as it moves through the film-developing station on a special transport mechanism. There, the chips progress automatically through a sequence of developing, stopping, fixing, washing, and drying. Up to eight chips can be at various stages of developing at a time. After processing, chips are automatically packed in plastic cells, 32 chips to a box.

Error correction-detection codes are checked for validity at a read station, and if a chip is not recorded properly it is discarded and the same data are read onto a new chip. The boxes are moved pneumatically by means of an air blower system interconnecting the writer, the file, and the reader. At the reading station the requested chip is picked from its storage cell and read by a high-speed flying-spot scanner. As the scanner moves back and forth across the data lines, the beam of light is transmitted through a clear spot on the chip, and blocked by a dark spot. A photomultiplier behind the chip senses and amplifies the transmitted light pattern, and electronic circuitry converts the signals to binary bits for delivery to the computer. The total time to select an individual box, transport it to a reader, and select a

*From the Greek bous, ox + strephein, to turn. It writes alternate lines of code in opposite directions.

chip to be read is less than 5 seconds.

Boxes of chips can be taken out of the system or re-inserted manually; this means that the potential size of the data base is virtually limitless. To keep track of all the data that are on-line there is a small process control computer in the chipstore equipment that controls all the hardware actions.

CDC-854 Disk Pack Drive for MSS Tables

A separate control unit is a CDC-854 Disk Pack, which holds all the tables and indexes to the data written into the chipstore. Although it has a record of high reliability its contents are dumped on to tape daily as a precautionary measure. Since the pack is removable the disk pack drive can be repaired and serviced without disturbing the tables.

CDC-6600 Computer Complex

The Berkeley Mass Storage System is connected to a CDC-6600 digital computer, multiprogrammed to allow up to 64 jobs to reside in core memory at once and share the central processing unit (CPU). The CDC-6600 has 131 072 words of 60-bit central core memory (CM). One of Berkeley's 10 peripheral processing units (PPU's) controls a high-speed data-channel connection between the chipstore and the 6600, acting as the system monitor and operator interface.

Software Design

One important objective in the design of the system was to make the interaction with the MSS easy for the user. A flexible read mechanism allows the user to read data either from the chipstore or from more conventional media such as magnetic tapes and system disk files. It is also possible to intermix data from the two random-access devices (chipstore and disk) and the linear device (magnetic tape). A user can label his data with a two-level hierarchy of names for data sets and subsets. The language used is Fortran IV.

Summary of Statistics Relevant to IRG Applications

The four main components of the MSS, as shown in Fig. 1, are:

(1) Data storage by IBM-1360 Photodigital Storage System (Chipstore).

(2) Table and index storage by CDC-854 Disk Pack.

(3) Processing by CDC-6600 computer.

(4) Easy-to-use software.

Some significant sizes and speeds are:

Chipstore's on-line capacity is 3.3×10^{11} bits of data recorded on photographic chips stored in boxes.

System holds 2250 boxes (equivalent to 2750 full reels of tape).

Box access time is 3 sec.

Average read speed is

1.1×10^{6} data bits/sec (~ 4 sec/chip).

Each chip holds 75 000 60-bit words (3/4 million 6-bit characters).

An issue of NSA fits on 6 to 7 chips.

Each box holds 32 chips (equivalent to about 5 issues of NSA, or 1-1/3 full tapes).

Figure 2 is a diagram of the chips, showing that they are divided into a rectangle of 4 frames by 8 frames, 32 in all. Figure 3 shows the size of the 1360 storage unit in relation to people. It occupies about as much floor space as 5 or 6 desks. Further information about the Berkeley chipstore can be found in a paper by Penny et al.[1] and a report by Metcalf.[2] The basic paper by Kuehler and Kerby of IBM describes the hardware in detail.[3]

The three big advantages for our use of this mass memory are the fast-read operations, the random-access capability, and the high reliability in recovery of archival data. Operating experience has shown that the read error rate is less than 1/60 of the error rate on magnetic tape.

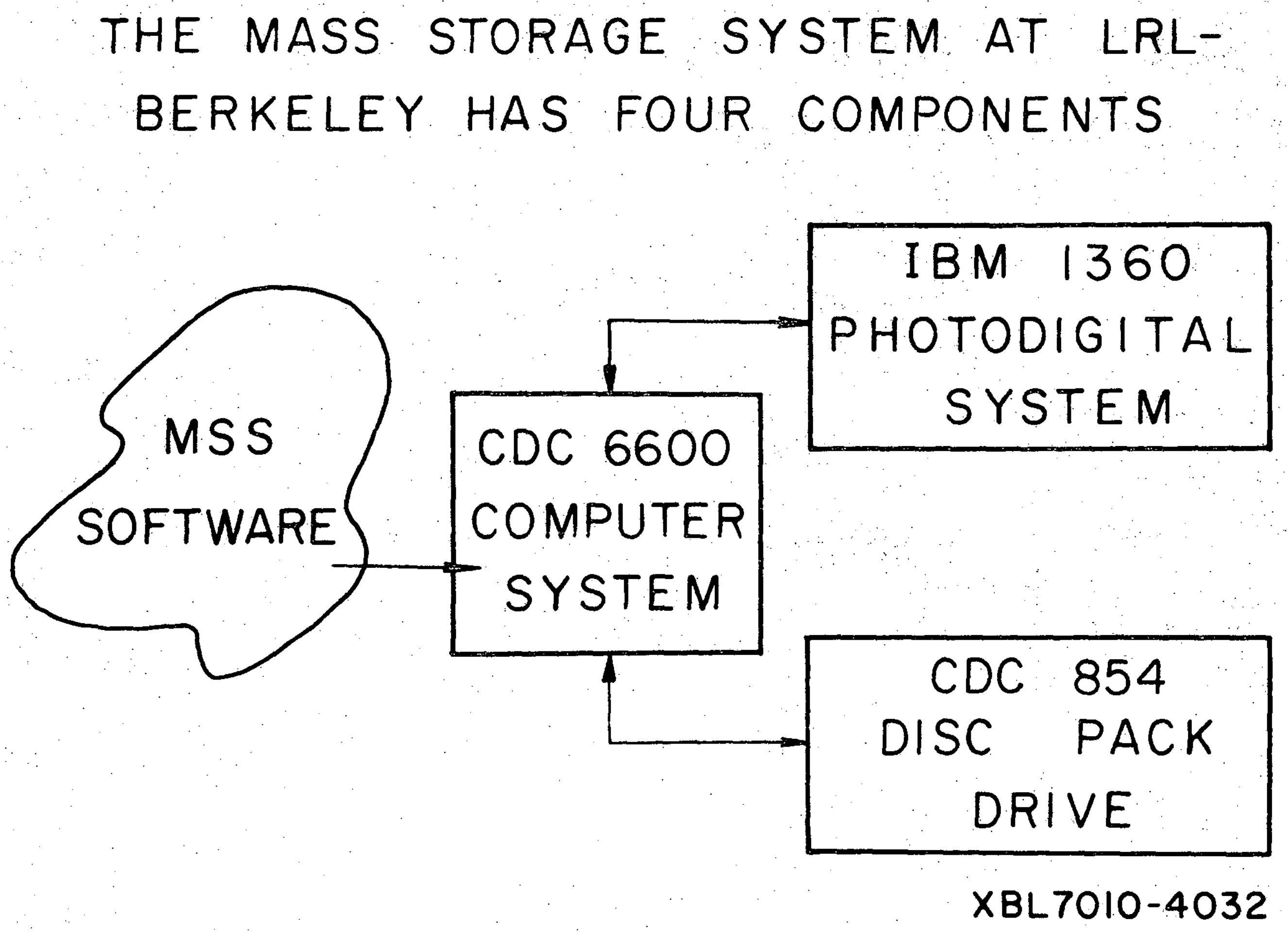

THE MASS STORAGE SYSTEM AT LRL-BERKELEY HAS FOUR COMPONENTS
MSS SOFTWARE
CDC 6600 COMPUTER SYSTEM
IBM 1360 PHOTODIGITAL SYSTEM
CDC 854 DISC PACK DRIVE
XBL7010-4032

Fig. 1. Main Elements of the MSS.

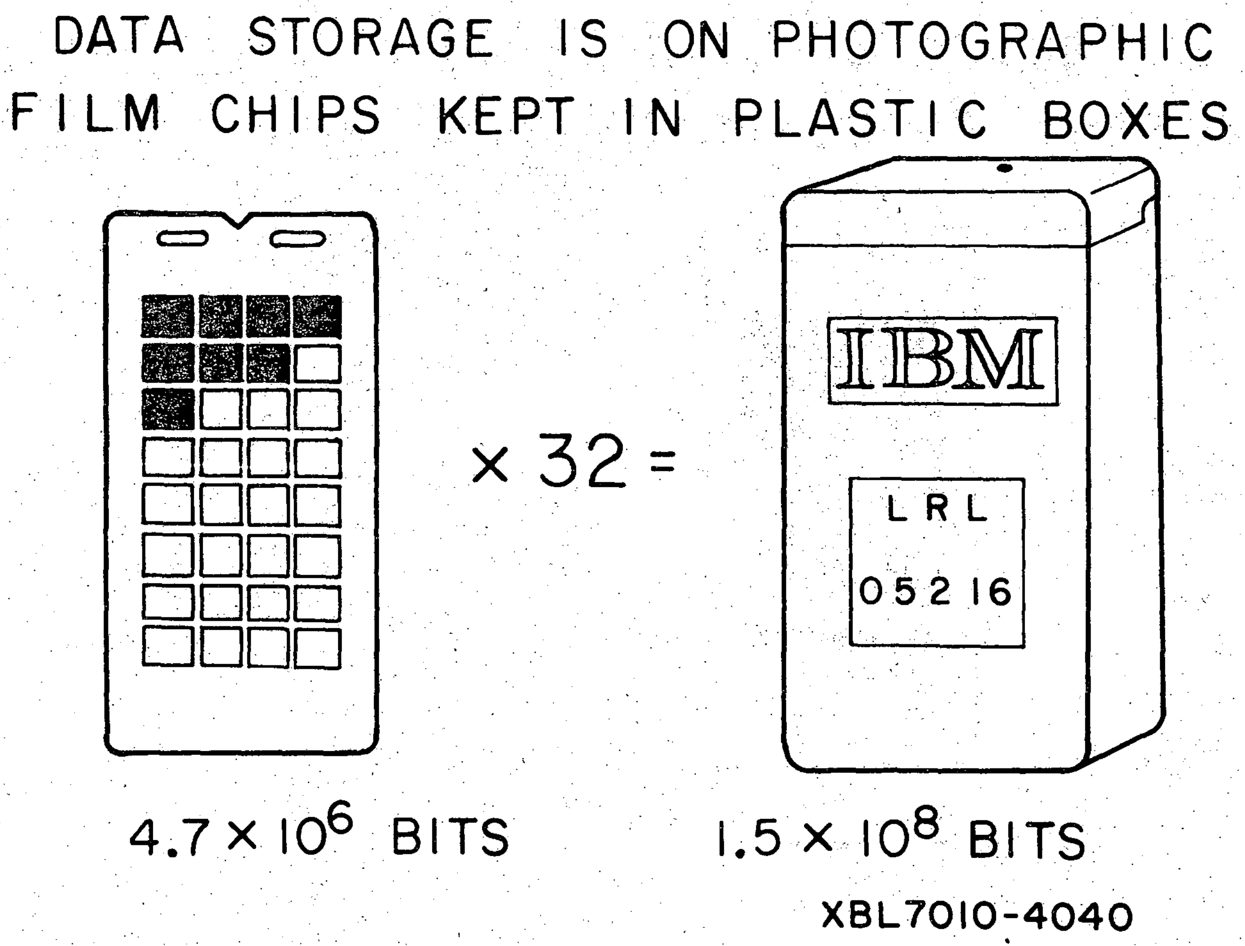
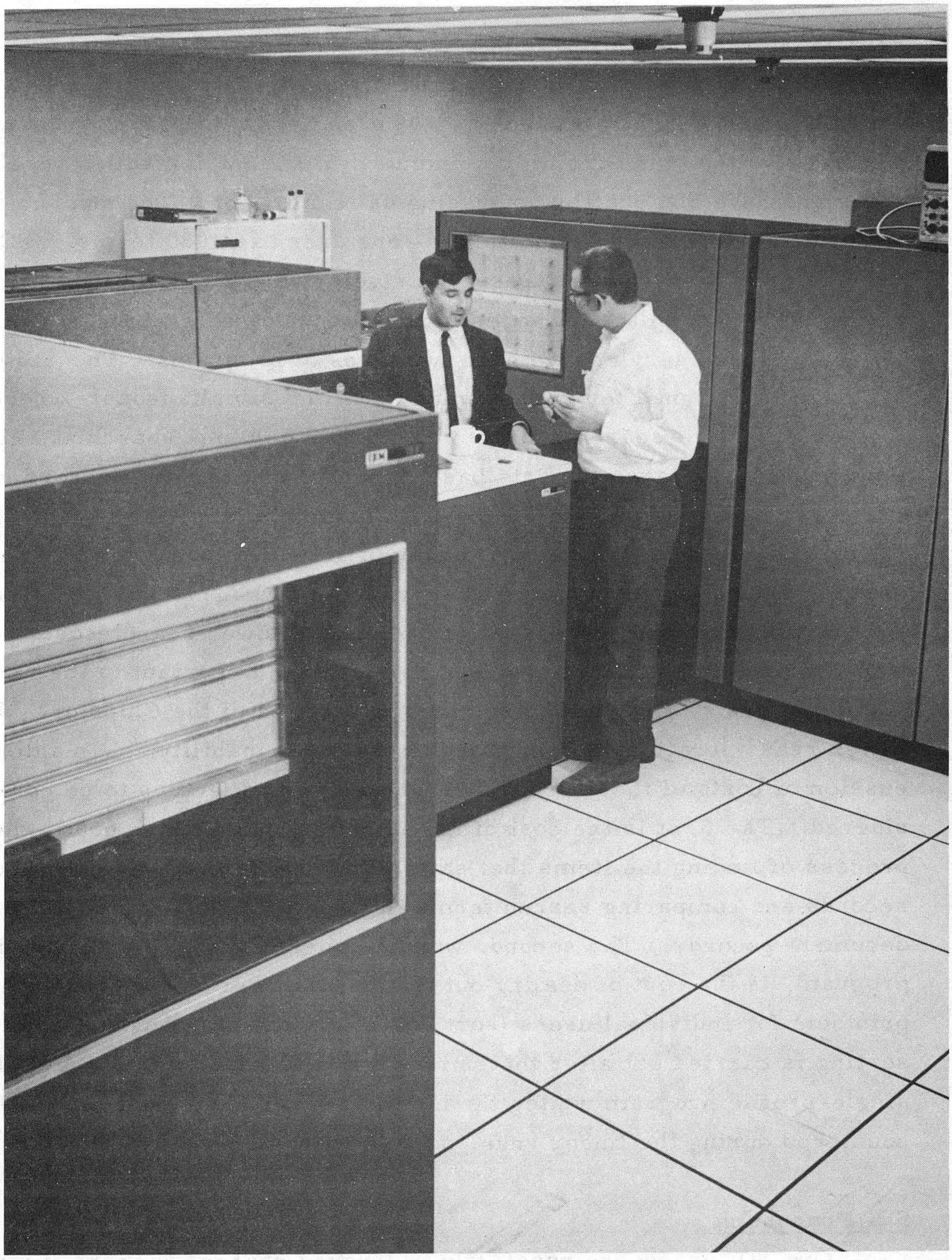

DATA STORAGE IS ON PHOTOGRAPHIC FILM CHIPS KEPT IN PLASTIC BOXES
× 32 =
IBM
LRL
05216
4.7 × 10^6 BITS
1.5 × 10^8 BITS
XBL7010-4040

Fig. 2. Basic Storage Units.



XBB 689-5627

Fig. 3. Overview of Chipstore Area.

IRG Use of the Chipstore

Since 1967 we have been operating an SDI system using the Nuclear Science Abstracts tapes. At present, there are 118 users who receive the twice-monthly printouts. We use a controlled vocabulary and have SDI profiles averaging 41 terms each. The number of terms allowed is virtually unlimited—one user has 1080 terms. SDI service, and its associated research studies, are our major effort, but we also accumulate the tapes for retrospective searching.

For these services we have two types of programs. The SDI program is designed for running many profiles simultaneously, whereas the retrospective-search program can handle only one search at a time. Both programs search the data base linearly, that is, each record is examined in accession number order to determine if it satisfies any of the search questions. There is a way of skipping over large segments of data, as will be shown later when we look at NSA categories. The last part of this paper discusses our efforts to lower search costs and still provide high-quality service. Particularly important in the reduction of costs is our use of two characteristics of the Chipstore–the faster read operations and the random-access capability.[4] In a discussion of costs of these programs, there are two factors to be considered. The first is the cost of the actual search (search cost)—the process of finding the items that satisfy the queries by reading in the records and comparing search terms in the profile with terms in the document records. The second, which applies only to the multiprofile program, is the cost of dealing out the hits in order to prepare the printouts for individual users (sort cost). In the SDI program, the sorting is carried out after the entire data base has been scanned. The single-profile program writes the hits onto a print tape as they are encountered during the linear scan, so no sorting is required.

SDI Processing

Our SDI procedure constantly undergoes changes to make it more efficient. This is part of the mission of the research group. Because at present the major cost in the SDI program is the comparison of terms to determine hits, we didn't expect that using the Chipstore would make a large difference in the search cost. But we found that the faster reads of the Chipstore and its random-access capability reduced the computer cost of an average SDI search 20%, from 42.5 cents to 34.0 cents per profile.

For retrospective searches we expected more dramatic changes, since the costs of these searches are due in large part to the process of reading in the data base. A comparative test confirmed our suspicions. A search of 4 years of NSA, carried out in part with tapes and in part with the Chipstore, illustrates the magnitude of the effect. There were 18 terms in the search, and it retrieved about 7 hits per issue. The average cost per issue for the tape searches was $3.53, whereas the average for the Chipstore was $0.90, or about a quarter of the cost of the tape searches.

We can save up retrospective searches to run with the multiprofile program--that is, the program that is used for SDI runs. In this way, the cost of reading the data base is shared among several searches, but a price is paid in sorting the hits. Still, if all the data base must be read, the sort cost is not greater than the money saved by reading the data base only once for several users. But what about the single retrospective search in a very narrow field? Most retro searches are of this type. We have found that the random-access capability of the Chipstore can be used with these searches in a very simple way based on NSA subject categories.

File Inversion for NSA Categories

In addition to indexing terms, the NSA indexers have also assigned subject categories to the items in NSA. We have inverted the file on the 77 NSA subsections (subject categories). This is a rather easy file inversion, since the items in any one subsection in a single issue are contiguous--this is because the subsections are used to arrange the material for the printed NSA. So our inverted file has only one entry per issue for each subsection–its address range. It's like a table of contents. To do a retrospective search, the subsections are specified and the linear search is carried out on only that part of the data base.

Table I shows the distribution of the data base records among the NSA sections; the largest section occupies only 21% of the total data base. A closer examination of this large section, Table II, shows the distribution at the next level of specificity--the subsection--which is available to us for restricting the retro searches. The amount of material scanned can be limited by section or by subsection, and both are useful for the retro searches.

A recent search will illustrate the cost savings found by limiting a search by subject category. It was a 41-term search that was limited to the Nuclear Physics section, which occupies 9.4% of the data base. It was carried out on 18 issues and produced 136 hits. The cost, has the search been saved and run with four others on the multiprofile program, would have been about $8.30. The category restriction gave us two big advantages: it permitted us to run the search immediately (without holding it to combine with others), and it reduced the cost by 75% to $1.99.

For a long time we've used the NSA categories to restrict both SDI and retrospective searches. Before we were using the Chipstore, such restriction did not influence the cost–but it was a very useful way to improve the quality of the output by defining the subject area of interest. Now the technique can be used to dramatically decrease the cost of the searches. We have found few subject searches that could not be limited in a useful way by category. We have taken good advantage of the Chipstore's random-access capability, but without the large cost of inversion of the file on the indexing terms. The average cost of inverting one issue by categories alone is $0.78 for the Chipstore.

Table I. Distribution of data base among NSA subject categories (sections)
Volume 24, Issues 1-24 -- 64355 Records

Section Number	Section Name	Number of Records	Percent of Total
20.00	Chemistry	9528	14.8
22.00	Earth Sciences	1161	1.8
24.00	Engineering	2717	4.2
26.00	Instrumentation	2812	4.4
28.00	Life Sciences	8405	13.1
30.00	Metals, Ceramics, and Other Materials	7212	11.2
32.00	General Physics	13442	20.9
34.00	High Energy Physics	6299	9.8
36.00	Nuclear Physics	6064	9.4
38.00	Reactor Technology	6563	10.2

Table II. Distribution of data base among NSA subject categories
Detail of section 32.00: General Physics

Volume 24, Issues 1-24 -- 64 355 Records for all of NSA;
13 442 Records for Section 32.00-33.00: General Physics (20.9% of total)

Subsection Number	Subsection Name	Number of Records	Percent of Total
32.00	General	179	0.3
32.10	Astrophysics	3158	4.9
32.20	Atomic and Molecular Physics	2108	3.3
32.30	Cosmic Radiation	264	0.4
32.40	Direct Energy Conversion	216	0.3
32.50	Fluid Physics	196	0.3
32.60	Geophysics	1182	1.8
32.70	Low-Temperature Physics	1079	1.7
32.80	Plasma and Thermonuclear Physics	2686	4.2
32.90	Shielding	273	0.4
33.00	General	59	0.1
33.10	Solid-State Physics	1694	2.6
33.20	Theoretical Physics	348	0.5

Summary and Future Possibilities

Thus far, we have used the faster reads and random-access capability of the Chipstore to reduce the cost of our SDI program by about 20%, and to produce a very low cost retrospective-search program very simply and without doing a full file inversion. Now, there are two routes we could take. We could extend the category gimmick to the SDI service, or we could go immediately to a fully inverted file. Since we now have enough SDI users to recover the cost of a complete file inversion, we are investigating techniques to accomplish this. With a fully inverted file, retrospective searches will be even less expensive than with the category-restricted linear search and the restriction will again be only a device for improving the quality of the output. Moreover, with an inverted file we will be in a position to begin working on an on-line search capability.

References

1. S. J. Penny, R. Fink, and M. Alston-Garnjost, "Design of a Very Large Storage System," AFIPS Conference Proceedings, 37, 45-51 (1970 Fall Joint Computer Conference, Houston, Nov. 17-19, 1970).
2. M. Metcalf, The Berkeley Mass Storage System, Lawrence Radiation Laboratory Report UCID-3479, Sept. 1970.
3. J. D. Kuehler and H. R. Kerby, "A Photodigital Mass Storage System," AFIPS Conference Proceedings, 29, 735-742 (1966 Fall Joint Computer Conference, San Francisco, Nov. 7-10, 1966).
4. J. J. Herr, Comparison of Efficiencies of Various Retrieval Programs on the CDC-6600 Computer, Lawrence Radiation Laboratory Report UCRL-20285, Feb. 1971.

LEGAL NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

TECHNICAL INFORMATION DIVISION
LAWRENCE RADIATION LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720