BNL 29228

CONF-810361--1

MASTER

A Comparison of System-2000 and Scientific Information Retrieval (SIR) in a Specific Scientific Application

Veronica Evans
Atmospheric Sciences Division
Brookhaven National Laboratory
Upton, N.Y. 11973

To be presented at the
ASTUTE Spring 1981 Conference
Austin, Texas
March 25, 1981

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of 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. Reference herein to any specific commercial product, process, or service by 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. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

This research was performed under the auspices of the United States Department of Energy under Contract No. DE-AC02-76CH00016.

By acceptance of this article, the publisher and/or recipient acknowledges the U.S. Government's right to retain a nonexclusive, royalty-free license in and to any copyright covering this paper.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MOW

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of 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. Reference herein to any specific commercial product, process, or service by 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. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Table of Contents

		Page
1.	Introduction	1
2.	Schema Definition	2
3.	Data Input	5
4.	Data Retrieval	8
5.	Data Modification	18
6.	Conclusion	20
7.	Appendix	21

1. Introduction

This paper compares specific procedures of two data base management systems available at Brookhaven National Laboratory, System-2000 (S2K) and Scientific Information Retrieval (SIR). The computer on which these reside is a CDC-6600 and the operating system is NOS-BE level 488.

Since S2K was available at BNL several years prior to SIR, all of our data management work was accomplished on S2K until SIR became available. An S2K data base had been created and implemented for use in an Acid Rain Project. Acid Rain is the term used to describe the increase in the acidity of precipitation caused by the increase in sulfur and nitrogen oxide emissions into the atmosphere. The goal of the Acid Rain Project was to predict which areas of the Eastern United States would be most affected by Acid Rain in the future.

The Acid Rain data base contained geology data and water quality data from lakes and streams in the Eastern U.S. This data was obtained from Federal, State, and Local agency sources, as well as field reports. The data base was a large (23.5 million characters) data base by BNL standards. It was a "live" working data base, and methods of loading, updating, modifying and retrieving the data were established.

For this comparison study a new S2K data base and a SIR data base were created using a subset of the actual raw data from the original Acid Rain data base. Identical data was loaded in both data bases and the procedures followed were duplicated as closely as possible on both. All jobs were run in batch mode to get accurate statistics on computer time and resources.

This paper does not try to compare the two data base management systems exhaustively, but rather discusses the results of certain specific

procedures in terms of computer resources and the ease of implementing these procedures from the user's standpoint.

2. Schema Definition

Since S2K and SIR are both hierarchical data base management systems, the schemas are quite similar (Figure 1). The data consists of state codes owning county/station data owning geology data and water quality data by date.

Logical Diagrams of Schema Definitions

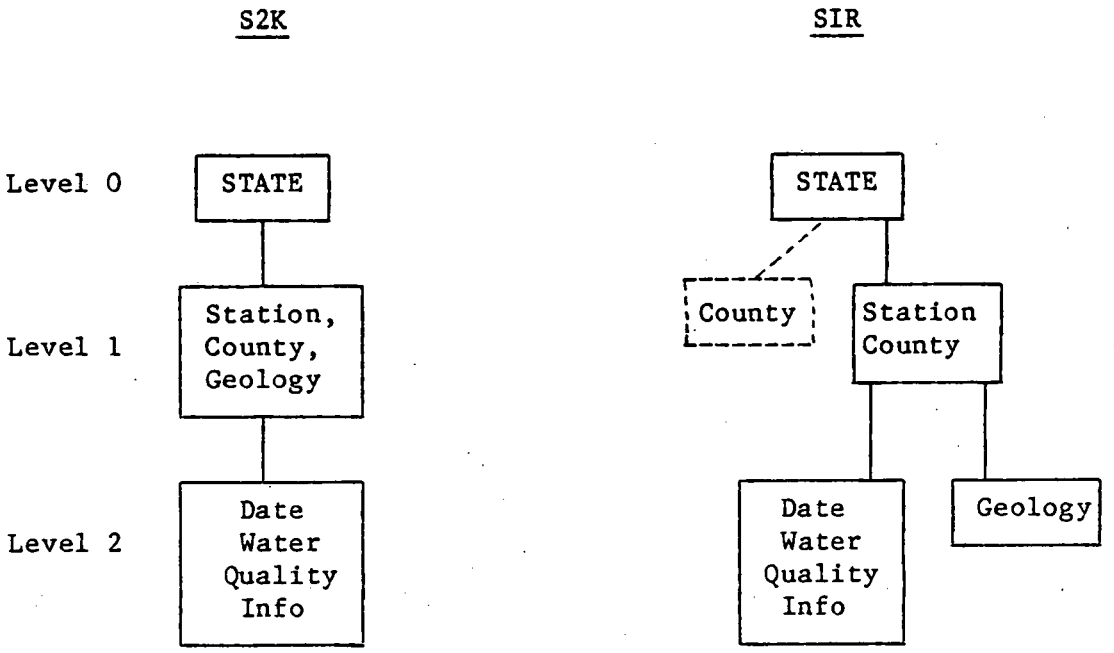


Figure 1

The S2K schema (Figure 2) has state at level 0, county, station and geology information at Level 1, and date and water quality information at Level 2.

```
 1*  STFIP (INTEGER NUMBER 99)
10*  STN RG (RG)
  11*  CTFIP (INTEGER NUMBER 999 IN 10)
  12*  STN CODE (NAME X(8) IN 10)
  13*  STN NAME (NAME X(10) IN 10)
  14*  LAT (NON-KEY DECIMAL NUMBER 99.99 IN 10)
  15*  LON (NON-KEY DECIMAL NUMBER 99.99 IN 10)
  16*  REF1 (NON-KEY NAME XXX IN 10)
  20*  TYPE (NON-KEY NAME X IN 10)
  21*  TYPE1 (NON-KEY DECIMAL NUMBER 999.99 IN 10)
  22*  TYPE2 (NON-KEY DECIMAL NUMBER 999.99 IN 10)
 100*  VALUE RG (RG IN 10)
   110*  DATE (DATE IN 100)
   121*  PHVAL (NON-KEY DECIMAL NUMBER 9.9 IN 100)
   131*  ALKVAL (NON-KEY INTEGER NUMBER 999 IN 100)
   141*  CAVAL (NON-KEY INTEGER NUMBER 999 IN 100)
```

Figure 2.

Key components in the data base are State Code (STFIP), County Code (CTFIP), Station Code (STN CODE), Station Name (STN NAME) and date, C1, C11, C12, C13, C110 respectively.

The SIR schema (Figure 3) defines State as the Case ID with four record types; record 1 contains state and county codes, record 2 contains station information, record 3 contains date and water quality information, and record 4 contains geology data.

```
TASK NAME          CASE DEFINITION
CASE ID            STFIP
N OF CASES         30
RECS PER CASE      5000
MAX INPUT COLS     80
MAX REC TYPES      7
MAX REC COUNT      150000
TASK NAME          RECORD TYPE 1
RECORD SCHEMA      1,STCNTY
SORT RECORDS       CTFIP
DATA LIST          FIXED(1)
                   /1 STFIP 2-3(I)
                   /1 CTFIP 4-6(I)
TASK NAME          RECORD TYPE 2
RECORD SCHEMA      2,STNID
```

```
SORT RECORDS        CTFIP,STNNAME
DATA LIST           FIXED(1)
                    /1 STFIP 2-3(I)
                    /1 CTFIP 4-6(I)
                    /1 STNCODE 10-17 (A)
                    /1 STNNAME 18-27 (A)
                    /1 LAT 28-32(2)
                    /1 LON 33-27(2)
                    /1 REF 38-40(A)
MISSING VALUES      STNCODE(BLANK)/
                    STNNAME(BLANK)/
                    REF(BLANK)/
TASK NAME           RECORD TYPE 3
RECORD SCHEMA       3,DATA
SORT RECORDS        CTFIP,STNNAME,DDATE
DATA LIST           FIXED(1)
                    /1 STFIP 2-3(I)
                    /1 CTFIP 4-6(I)
                    /1 STNNAME 18-27(A)
                    /1 DDATE 54-61(A)
                    /1 PH 64-66(1)
                    /1 ALK 67-69(I)
                    /1 CA 70-72(I)
DATE VAR LIST       DDATE(YYIMMIDD)
MISSING VALUES      PH(BLANK)/
                    ALK(BLANK)/
                    CA(BLANK)/
TASK NAME           RECORD TYPE 4
RECORD SCHEMA       4,GEOL
SORT RECORDS        STNNAME,SRTYPE
DATA LIST           FIXED(1)
                    /1 STFIP 2-3(I)
                    /1 STNNAME 18-27 (A)
                    /1 SRTYPE 41(A)
                    /1 TYPE1 42-47 (2)
                    /1 TYPE2 48-53 (2)
```

Figure 3.

In SIR, the sort records which are similar to keys in S2K are state code, county code, station name and date.

The schema definition for S2K contains component names, keys, and data types. The schema definition for SIR also contains variable names, sort records, and data types. However, an important part of the SIR schema is a description of the data file, i.e., the number of columns on the input file

and the location and format of the variables on the input file. This is an important difference between S2K and SIR. Although preparation of the SIR schema is more involved, the loading of the data in SIR is easier.

3. Data Input

The first set of data consisting of 780 records was loaded and some retrievals were run. The second set of data consisting of 257 records was loaded and more retrievals and some modifications were run. The third set of data consisting of 27024 records was loaded, unloaded and reloaded and more retrievals and modifications were run.

Of the several methods of loading data available in S2K, the creation of a value string was chosen over PLI, because at BNL the central memory requirements of a PLI program make it a slow and expensive type of job to run. The raw data was read and validated by a Fortran Program and put in value string format. It was loaded using the "LOAD" command.

The data was loaded in raw form in the SIR data base using the "Read Input Data" command (Figure 4). The file on which the data resides is rewound each time a new record is read.

```
INPUT MEDIUM      TAPE1
READ INPUT DATA   RECTYPE=1/
                  ERRFILE=TAPE2,REWIND
READ INPUT DATA   RECTYPE=2/
                  ERRFILE=TAPE2,REWIND
READ INPUT DATA   RECTYPE=3/
                  ERRFILE=TAPE2,REWIND
READ INPUT DATA   RECTYPE=4/
                  ERRFILE=TAPE2,NOREWIND
```

Figure 4.

The actual jobs to load the data are very straightforward on both S2K and SIR; however, the writing of the Fortran program in S2K requires a great deal of user time. Data verification checks must be coded in the program. In SIR, the user can put data verification checks in the schema definition and the data will be verified at load time.

The total computer resource units spent in loading the 3 data sets in both data bases are in Figure 5.

Total Resource Units-Load Data

	Data Set 1	Data Set 2	Data Set 3
S2K	58.091	19.514	1369.857
SIR	29.997	13.329	865.673

Figure 5.

It took S2K an average of 1.66 times the resources that SIR required to load these data sets. SIR is an index sequential file. The data is ordered in the way the sort records are ordered. Since the raw data for this data base was ordered by sort records, it allowed the most efficient load of the data by SIR. S2K is an inverted file and has to set up tables and pointers which require more resources to load the data. The figures in the S2K row represent the sum of the resources of the Fortran loader string program plus the resources of the S2K load job.

Storage of the data base on disk is accomplished using 6 files on S2K and 4 files in SIR (Figure 6).

S2K	SIR
TA - Directory	1 - Codebook File
TB - Unique Value Table	2 - Detail File
TC - Overflow	3 - Procedure File
TD - Multiple Occurrence Table	4 - Data File
TE - Hierarchy	
TF - Data	

Figure 6.

A system audit was run on both data bases after the schemas and each data set was loaded (Figure 7). The audit lists the number of physical recurd units (PRU) of each file. A physical record unit is equivalent to 640 decimal characters. The table lists the total number of PRU's in the six S2K files and the four SIR files.

Data Base Storage on Disk
(in Total Physical Record Units)

	Schema	Data Set 1	Data Set 2	Data Set 3	Data Set 3-2
S2K default padding	176	352	400	3128	3232 ←
S2K w/padding on keys					6096
SIR default padding (.5)	35	420	524	6360	12396
SIR .99 padding			312	3444	3460 ←

Figure 7.

All the figures appearing under the column heading Data Set 3-2 represent a reload of Data Set 3 affter a "Remove Tree Entry" was performed on S2K and "Delete Cases" on SIR. The reload was performed to see how the disk files were affected. They were virtually unaffected on S2K but the storage almost doubled on SIR.

SIR defaults to a 0.5 padding factor when loading the data. A 0.99 padding factor was applied when the data base was restored from tape. It cuts down on storage space considerably but has no effect on retrieval time.

S2K defaults to a zero padding factor. When an increased padding factor was applied to the keys in the S2K data base, the storage space on disk was increased by nearly a factor of 2. The padding factor in S2K is concerned only with the number of times a unique value occurs, therefore depending on the data in the data base, a retrieval's performance may or may not be improved.

The two figures with an arrow next to them in column Data Set 3-2 represent the data sets on which the retrievals for this paper were performed.

4. Data Retrieval

The retrievals that will be discussed were chosen because they closely represent the way in which the original Acid Rain data base was used and they are retrievals which can be easily duplicated on S2K and SIR.

The results are plotted using the following four subplots:

(A) The plot in the upper left-hand corner shows Central Memory use in units of 40K octal word seconds, (CM seconds).

(B) The plot in the lower left-hand corner shows Input/Output time, (I/O seconds).

(C) The plot in the upper right-hand corner shows central processor time (CPA seconds).

(D) The plot in the lower right-hand corner shows the total Resource Units for a job, i.e., the sum of (A) + (B) + (C) (RU seconds).

Resource Units are subsequently converted to Computer Charge Units, or CCU's, at BNL.

Note - The x-axis scale on each of the plots is changed for each retrieval to best represent the data. The terms Data Set 1, 2, 3 from now on will refer to the data bases after data sets 1, 2, and 3 respectively, were loaded.

Retrieval 1 - List all Water Quality Data.

```
          USER,RAINRE%
          DBN IS ACRAIN%
S2K       LIST C1,C11,C13,C110,C121,C131,C141,OB,C11,C13,C110
            WHERE C11 EXISTS%
          EXIT%


          GET FILE          ACRAIN
          PASSWORD          RAINRE
          RETRIEVAL
          HEADING           "LISTING OF ALL DATA IN RECORD 3"
SIR       FOR EACH REC      3
          MOV VAR LIST      STFIP,CTFIP,STNNAME,DDATE,PH,ALK,CA
          COMPUTE           WDATE=DATEC(DDATE,"MM/DD/YY")
          WRITE             1X,STFIP(I2),2X,CTFIP(I3),1X,STNNAME(A10),
                            2X,WDATE,1X,PH(F3.1),1X,ALK(I3),1X,CA(I3)
          FINISH
```

Figure 8.

This retrieval simply lists the water quality data. In the SIR data base the data is already ordered by state, county, station name and date, and in S2K the WHERE clause is required to qualify only water quality data, therefore SIR performed this retrieval in less time and used less resources than S2K. This retrieval was run over Data Set 1 (See Retrieval Plot 1-1).

Retrieval 2 - Listing of Water Quality
Data Observations Occurring after 1975

```
          LIST C1,C11,C13,C110,C121,C131,C141,
S2K        OB C1,C11,C13,C110
           WHERE C110 GT 12/31/75%


          RETREIVAL
          HEADING          "STATIONS WITH DATA AFTER 12/31/75"
          COMPUTE          TESTDT = CDATE("12/31/75","MMIDDIYY")
          FOR EACH REC     3
SIR       SELECT REC IF    (DDATE GT TESTDT)
          MOVE VAR LIST    STFIP,CTFIP,STNNAME,DDATE,PH,ALK,CA
          COMPUTE          WDATE=DATEC(DDATE,"MM/DD/YY")
          WRITE            1X,STFIP(I2),2X,CTFIP(I3),1X,STNNAME(A10),
                           2X,WDATE,1X,PH(F3.1),1X,ALK(I3),1X,CA(I3)
```

Figure 9.

The scientist wanted to do time trends studies on the data and wanted to look at only the "current" data in the data base, so our where clause qualification is C110 GT 12/31/75. In SIR the date information is stored as Julian date and had to be converted before being printed. This retrieval was run over all three data sets. The sorting and where clause processing in S2K were less efficient and SIR performed better all three times, by a factor of 2.0, 2.3, and 6.1 respectively (See Retrieval plots 2-1, 2-2, 2-3).

Retrieval 3 - List Minimum Alkalinity by County
for Observations Occurring after 12/31/75

```
        COMPOSE%
        FOR REPORT RET3%
        PHYSICAL PAGE IS 80 BY 0%
        SELECT RECORD IF C131 EXISTS%
S2K     ORDER BY C11,C131%
        FOR C11,
        PRINT R(1,99)C1, R(4,999)C11, L(9,X(10))C13, R(21,999)C131%
        END REPORT%
        GENERATE ALL WHERE C110 GT 12/31/75%
        EXIT%.
```

```
        RETRIEVAL
        HEADING          "MINIMUM ALKALINITY BY COUNTY-POST 12/31/75"
        COMPUTE          TESTDT = CDATE("12/31/75","MMIDDIYY")
        FOR EACH REC     1
        MOVE VAR LIST    STFIP,CTFIP
SIR     PROCESS REC      3,WITH(CTFIP)
        SELECT REC IF    (DDATE GT TESTDT)
        COMPUTE          WDATE=DATEC(DDATE,"MM/DD/YY")
        COMPUTE          SMALL=MINR(ALK)
        END
        WRITE            1X,STFIP(I2),2X,CTFIP(I3),
                         1X,SMALL
```

Figure 10.

The scientist wanted to display the minimum alkalinity by county on a map. In S2K a report writer job was required because the county information was not in a separate repeating group. If it were, "MIN C131 BY C10" could have been used, cutting down the time and resources used to run the job. In SIR, the county code was the first sort-id in record 3, allowing the "MIN" function to be used. This retrieval was run over Data Set 1 and Data Set 3. S2K used 2.4 times the resources SIR did on Data Set 1 and >10 times on Data Set 3 (See Retrieval plots 3-1, 3-2).

Retrieval 4 - List the Minimum Alkalinity
for a Station in a given County

```
        LIST C1,C11,C13,MIN C131 BY C10,OB C13 WHERE
S2K     C11 EQ 301%
        EXIT%
```

```
        RETRIEVAL
        HEADING          "MINIMUM ALKALINITY BY STATION"
        FOR EACH REC     2
        SELECT REC IF    (CTFIP EQ 301)
SIR     MOVE VAR LIST    CTFIP,STNNAME
        PROCESS REC      3,WITH(CTFIP,STNNAME)
        COMPUTE          STNMIN=MINR(ALK)
        END
        WRITE            1X,STFIP(I2),2X,CTFIP(I3),
                         2X,STNNAME(A10), 2X,STNMIN(F8.4)
```

Figure 11.

The scientist wanted to know which station had the lowest alkalinity in a given county. In contrast to Retrieval 3, this time the "MIN" function in S2K could be used. In SIR the "MIN" function was used but since station name was the sort id needed, more records were processed. This accounts for the higher CPA and I/O seconds used by SIR. This retrieval was run over Data Set 1 and S2K used less resources than SIR (See Retrieval plot 4-1).

Retrieval 5 - List only Storet Stations
in a particular County

```
        LIST C1,C11,C13,C12,C131,OB C13 WHERE
S2K     C12 EXISTS AND C11 EQ 1%
        EXIT%


        RETRIEVAL
        HEADING          "LISTING OF ONLY STORET STATIONS"
        FOR EACH REC     2
        SELECT REC IF    ((STNCODE NE "          ") AND (CTFIP EQ 1))
        MOVE VAR LIST    STFIP,CTFIP,STNNAME,STNCODE
SIR     PROCESS REC      3, WITH (CTFIP,STNNAME)
        MOVE VAR LIST    ALK
        END
        WRITE            1X,STFIP(I2),2X,CTFIP(I3),
                         2XSTNNAME(A10), 2X,STNCODE(A10),1X,ALK(I3)
        FINISH
```

Figure 12.

The Acid Rain data came from different sources. Depending on the existence of the Station code, the scientist could tell the source. This is a very straightforward retrieval in both systems with the same "where" clause. This retrieval was repeated over Data Sets 1, 2, and 3. SIR used more I/O and CPA seconds on Data Sets 1 and 2 but used less than half the resources of S2K on Data Set 3 (See Retrieval plots 5-1, 5-2, 5-3).

Retrieval 6 - Calculate the Average Alkalinity for a particular Station in a particular County

```
S2K     LIST C1,C11,C13,C12,AVG C131 BY C10,OB C12 WHERE
        C12 EQ 220003 AND C11 EQ 3%
        EXIT%
```

```
SIR     RETRIEVAL
        HEADING          "AVERAGE ALKALINITY BY STATION"
        FOR EACH REC     2
        SELECT REC IF    ((STNCODE EQ "220003  ") AND (CTFIP EQ 3))
        MOVE VAR LIST    STFIP,CTFIP,STNNAME,STNCODE
        PROCESS REC      3,WITH(CTFIP,STNNAME)
        COMPUTE          AVGG = MEANR(ALK)
        END
        WRITE            1X,STFIP(I2),2X,CTFIP(I3),
                         2XSTNNAME(A10), 2X,STNCODE(A10),1X,AVGG(F7.3)
```

Figure 13.

This retrieval is similar to Retrieval 4, however the where clause specifies a particular station and there is only record printed. The "AVG" function was used on both systems. This retrieval was run on Data Set 1. S2K used slightly more central memory but used less total resource units than SIR (See Retrieval plot 6-1).

Retrieval 7 - Calculate the Average, Maximum and Minimum Alkalinity for a particular Station in a particular County

```
        LIST C11,C12,AVG C131 BY C10,MAX C131 BY C10,
        MIN C131 BY C10,OB C11 WHERE
S2K     C12 EQ 220003 AND C11 EQ 3%
        EXIT%
```

```
        RETRIEVAL
        HEADING          "AVG,MAX,MIN ALKALINITY BY STATION"
        FOR EACH REC     2
        SELECT REC IF    ((STNCODE EQ "220003  ") AND (CTFIP EQ 3))
        MOVE VAR LIST    STFIP,CTFIP,STNNAME,STNCODE
        PROCESS REC      3,WITH(CTFIP,STNNAME)
SIR     COMPUTE          AVGG = MEANR(ALK)
        COMPUTE          BIG = MAXR(ALK)
        COMPUTE          SMALL = MINR(ALK)
        END
        WRITE            1X,CTFIP(I3),1X,STNCODE(A8),AVGG(F7.3),
                         2X,BIG(F7.3),2X,SMALL(F7.3)
```

Figure 14.

This retrieval is similar to the previous one. The functions are performed on one particular station and one record is written. Writing the command is much easier in S2K. This retrieval was run over Data Set 1. S2K used slightly more central memory, but required less total resource units than SIR (See Retrieval plot 7-1).

Retrieval 8 - List the Geology data for two States

```
        LIST C1,C13,C20,C21,C22,OB C1
S2K     WHERE C12 EQ 00000000 AND
         (C1 EQ 23 OR C1 EQ 33)%
        EXIT%
```

```
        RETRIEVAL
        HEADING          "GEOLOGY DATA FOR STATES 23 AND 33"
        SELECT CASE IF   ((STFIP EQ 23) OR (STFIP EQ 33))
        FOR EACH REC     4
SIR     SELECT REC IF    (SRTYPE EQ "S")
        MOVE VAR LIST    ALL
        WRITE            1X,STFIP(I2),2X,STNNAME(A10),2X,
                         SRTYPE(A1),2X,TYPE1(F8.2),
                         2X,TYPE2(F8.2)
        FINISH
```

Figure 15.

The geology data was used for plotting isopleths. Not all states had geology data, therefore these two states were chosen. Since the geology data was in the county/station repeating group in S2K, the where clause qualified on a special station code. Since the geology data was in a separate record in SIR, the qualification was performed on an "SRTYPE" variable. S2K required less I/O, but more total resource units than SIR (See Retrieval plot 8-1).

Retrieval 9 - List Geology Data including
Latitude and Longitude for two States

```
        LIST C1,C12,C13,C14,C15,C20,C21,C22,OB C1
S2K     WHERE C12 EQ 00000000 AND
         (C1 EQ 23 OR C1 EQ 33)%
        EXIT%


        RETRIEVAL
        HEADING          "GEOLOGY DATA FOR STATES 23 AND 33"
        SELECT CASE IF   ((STFIP EQ 23) OR (STFIP EQ 33))
        FOR EACH REC     2
        SELECT REC IF    ((STNCODE EQ "00000000") AND (STNNAME NE
                         "SOILDATA")
        MOVE VAR LIST    ALL
SIR     PROCESS REC      4,WITH(STNNAME)
        SELECT REC IF    (SRTYPE EQ "S")
        MOVE VAR LIST    ALL
        WRITE            1X,STFIP(12),2X,STNNAME(A10),2X,STNCODE(A8),
                         2X,LAT(F8.2),2X,LON(F8.2),2X,
                         SRTYPE(A1),2X,TYPE1(F8.2),
                         2X,TYPE2(F8.2)
        FINISH
```

Figure 16.

This retrieval is very similar to the previous one, however this time we are retrieving two additional variables, latitude and longitude. All the data is in the same repeating group in S2K and the where clause remains the same. In SIR, since the latitude and longitude are in a different record from the geology data, qualification on two records is performed causing SIR to take more I/O, CM and CPA seconds than S2K (See Retrieval plot 9-1).

5. Data Modification

A few simple changes to the data were performed after Data Set 2 was loaded, and after Data Set 3 was loaded.

Modification 1 - Change all Calcium values to 100 wherever County Code is 3, over Data Set 2, and Change all Calcium Values in State 45 to 75 for County Code 5

S2K Data Set 2

```
CHANGE C141 EQ 100* WHERE C11 EQ 03%
```

S2K Data Set 3

```
CHANGE C141 EQ 75* WHERE C11 EQ 5 AND C1 EQ 45%
```

SIR Data Set 2

```
RETRIEVAL       AUTOSET UPDATE
HEADING         "CHANGE CA VALUE TO 100 FOR ALL COUNTY EQ 3"
FOR EACH REC    3
SELECT REC IF   (CTFIP EQ 03)
MOVE VAR LIST   STFIP,CTFIP,CA
COMPUTE         CA = 100
FINISH
```

SIR Data Set 3

```
RETRIEVAL       AUTOSET UPDATE
HEADING         "CHANGE CA VALUE TO 75 FOR STATE 45,
                COUNTY 5"
FOR EACH CASE   LIST = 45
FOR EACH REC    3,WITH(5)
MOVE VAR LIST   STFIP,CTFIP,CA
COMPUTE         CA = 75
FINISH
```

Figure 17.

Whereas S2K requires a single sentence to accomplish this change, SIR requires five or six lines. Both runs took S2K less CM, I/O, and CPA seconds than SIR (See Modification plots 1-1, 1-2).

Modification 2 - Add a calcium value in a particular station where none exists.

```
S2K     ADD C141 EQ 240* WH C110 EQ 09/06/78 AND C13 EQ HOLBROOK P
        AND C11 EQ 19 AND C1 EQ 23% LIST C1,C11,C13,C110,C141 WH SAME%
        EXIT%
```

```
SIR     RETRIEVAL               UPDATE
        HEADING                 "INSERT CA VALUE FOR INDIVIDUAL STATION"
        COMPUTE                 TESTDT = CDATE("09/06/78","MMIDDIYY")
        CASE IS                 23
        RECORD IS               3,(019,"HOLBROOKP",TESTDT)
        COMPUTE                 CTFIP=019; STNNAME="HOLBROOK P"; DDATE=TESTDT
        COMPUTE                 CA = 240
        COMPUTE                 WDATE = DATEC(DDATE,"MM/DD/YY")
        WRITE                   1X,STFIP(I2),2X,CTFIP(I3),1X,CA(I3)
                                  2X,STNNAME(A10),2X,WDATE
        FINISH
```

Figure 18.

This modification was performed over data set 1. The record was printed after the change was performed. The statement in S2K was a single sentence. The modification required at least 6 lines in SIR. S2K took slightly less CPA seconds but used more I/O and CM seconds than SIR (see Modification plot 2-1).

Modification 3 - Change Calcium values to 97 for county code 1 in all states.

```
S2K     CHANGE C141 EQ 97* WHERE C11 EQ 5%
        EXIT%
```

```
SIR     RETRIEVAL               AUTOSET UPDATE
        HEADING                 "CHANGE CA VALUE TO 97 FOR ALL STATES,
                                  COUNTY 5"
        FOR EACH REC            3,WITH(5)
        MOVE VAR LIST           STFIP,CTFIP,CA
        COMPUTE                 CA = 97
        FINISH
```

Figure 19.

This modification was run over data set 3. It is a simple procedure in both systems. S2K took less CM, I/O, and CPA seconds than SIR (see Modification plot 3-1).

6. Conclusions

The conclusions reached in this paper reflect the way in which S2K and SIR were used in this specific application. There are many features of both systems that were not implemented in this comparison study.

Data loading was easier in SIR because the raw data was loaded directly. In S2K the data had to be converted to value string format in a Fortran program. This was time consuming because all data validation checks had to be coded in the program.

S2K and SIR performed retrievals on small sets of data at about the same rate, but SIR's performance over S2K improved as the size of the data base increased.

Updating and modifying the data were generally more efficient in S2K. Due to the inverted structure of S2K, disk storage costs were less in S2K, even after applying the 0.99 padding factor.

The "conversational" commands of S2K make it an easy language to use. SIR's procedures require the writing of a command procedure.

Since the data used were scientific data and since SIR was designed specifically to handle scientific data, one would expect it to perform certain procedures more efficiently than S2K, and it did. However, it is lacking the useful query language that makes S2K such a powerful tool, and no comparisons could have been made in this area.

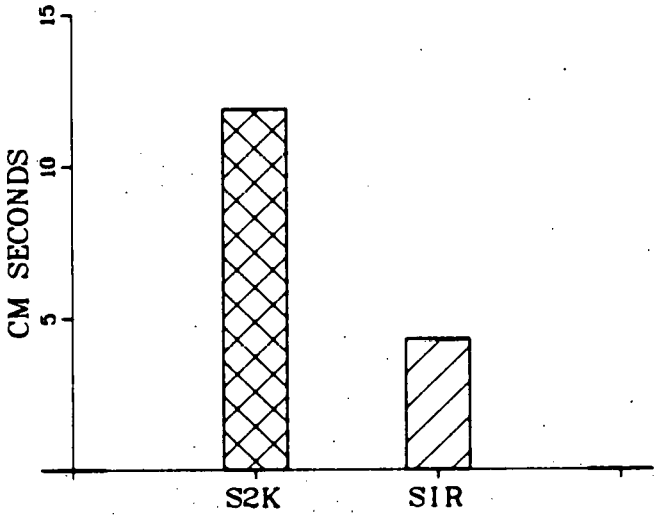
7. Appendix - Retrieval and Modification Plots.

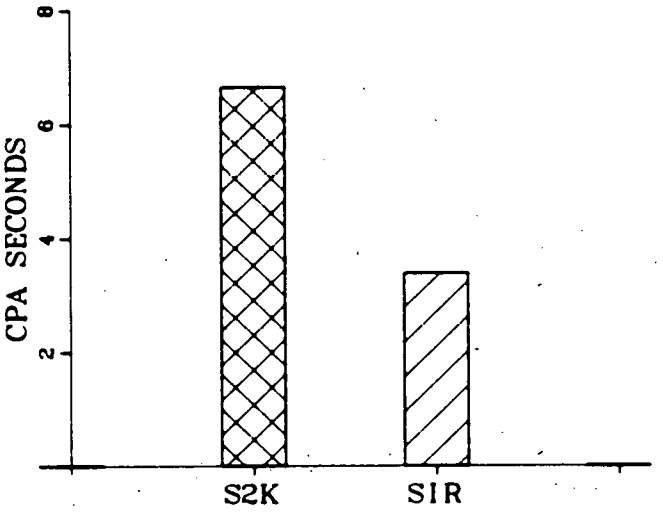
Retrieval	1-1
Retrieval	2-1
Retrieval	2-2
Retrieval	2-3
Retrieval	3-1
Retrieval	3-2
Retrieval	4-1
Retrieval	5-1
Retrieval	5-2
Retrieval	5-3
Retrieval	6-1
Retrieval	7-1
Retrieval	8-1
Retrieval	9-1
Modification	1-1
Modification	1-2
Modification	2-1
Modification	3-1

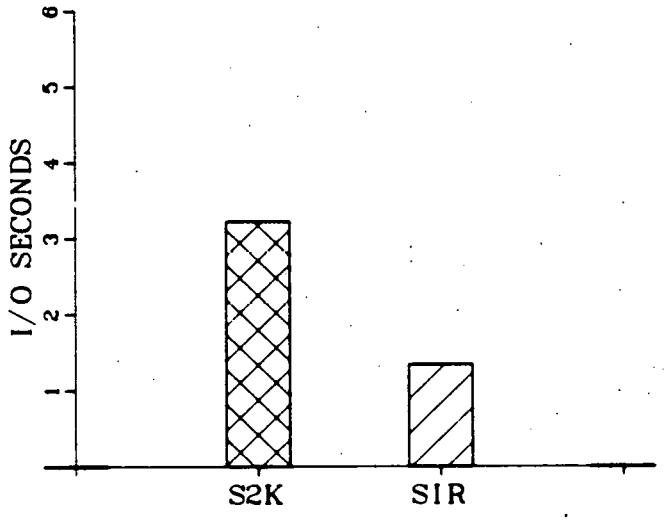
S2K vs SIR

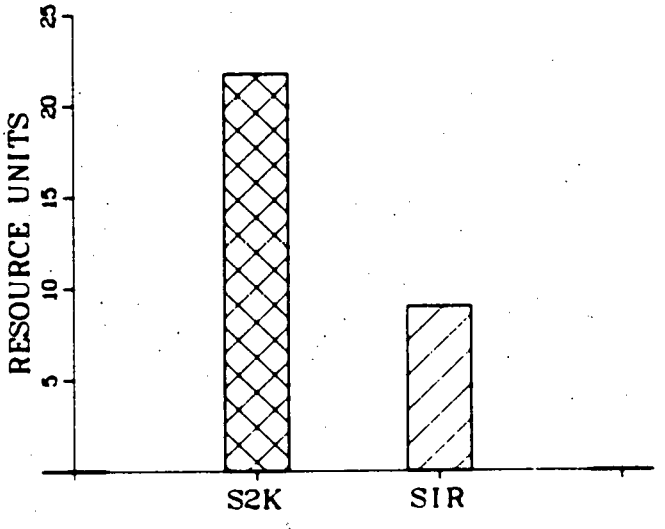
Retrieval 1-1

List All Water Quality Data

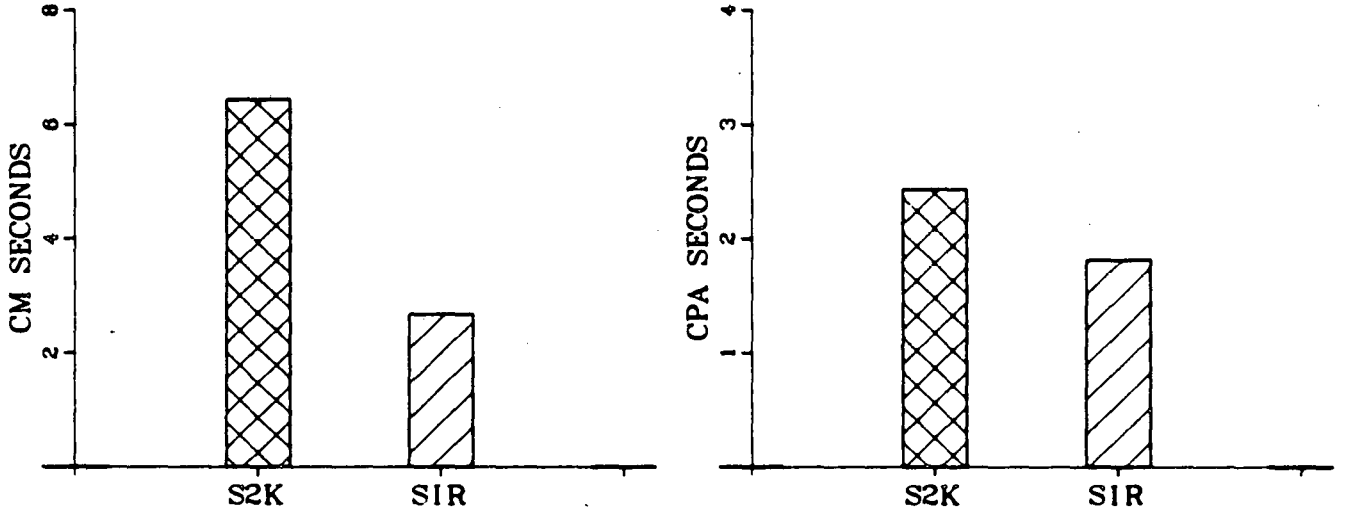


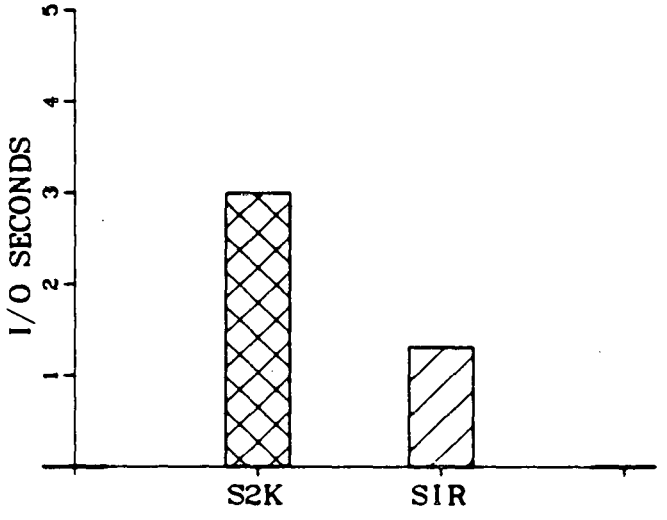


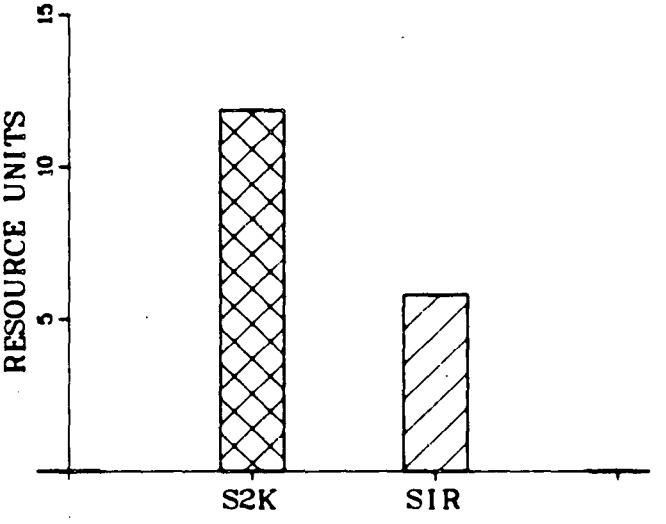




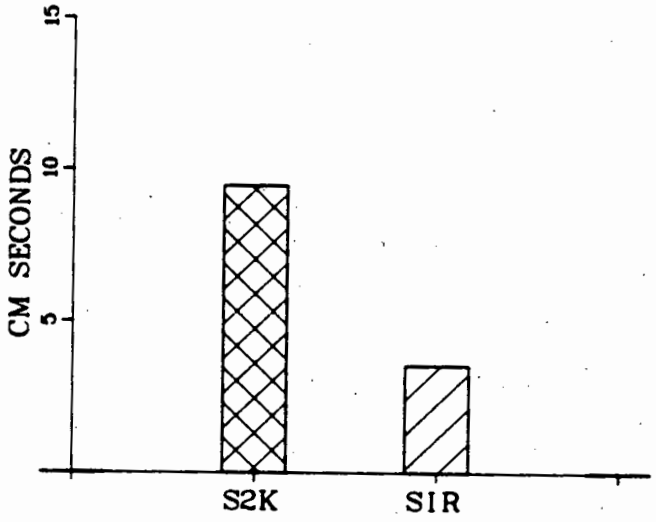
S2K vs SIR
Retrieval 2-1
List Data Occurring After 1975

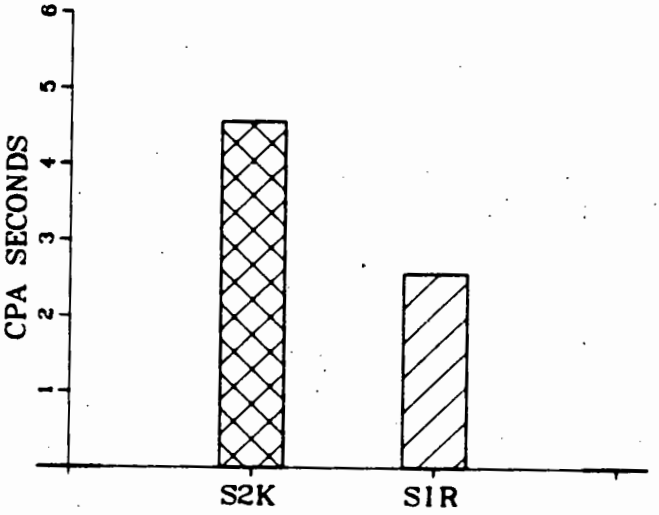


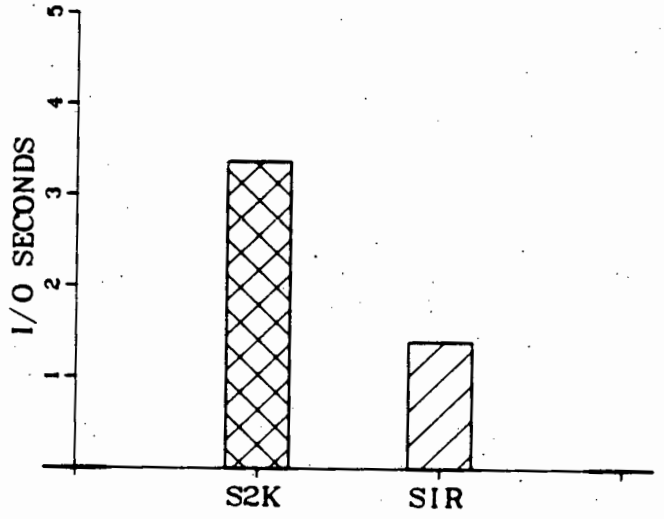


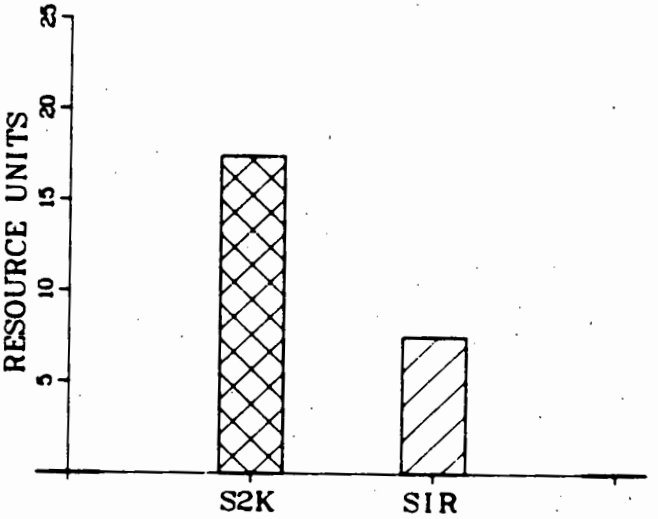


S2K vs SIR
Retrieval 2-2
List Data Occurring After 1975

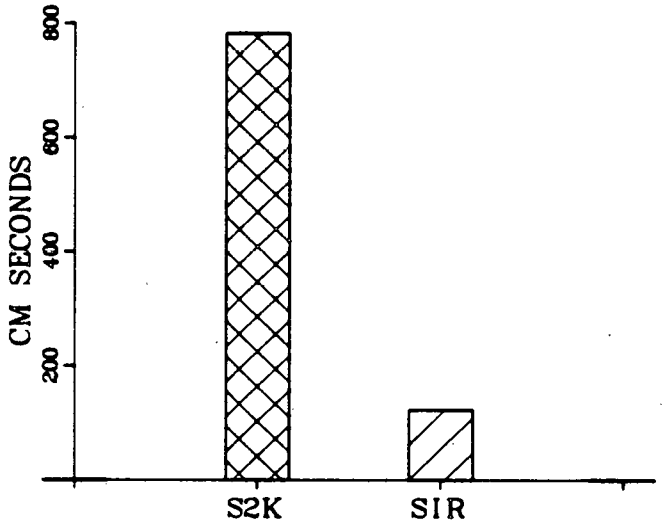


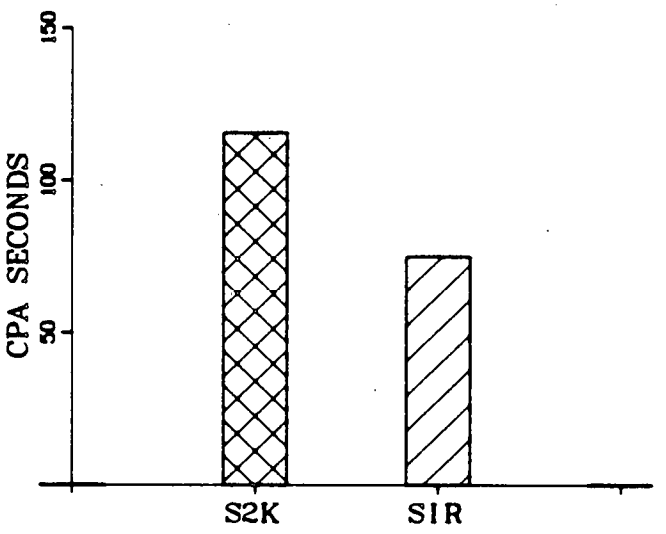


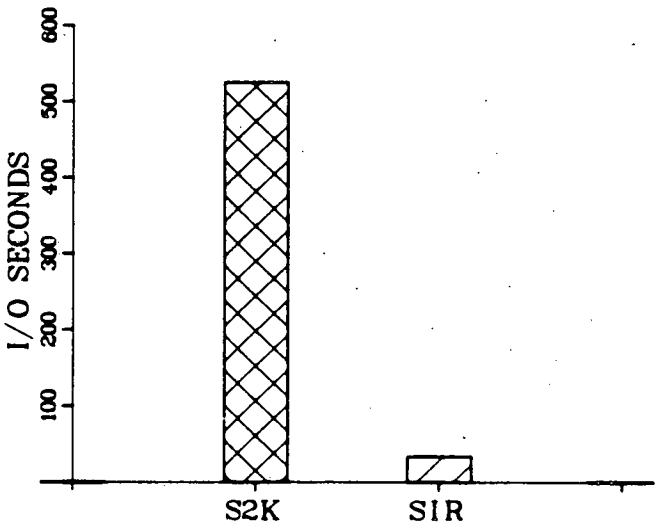


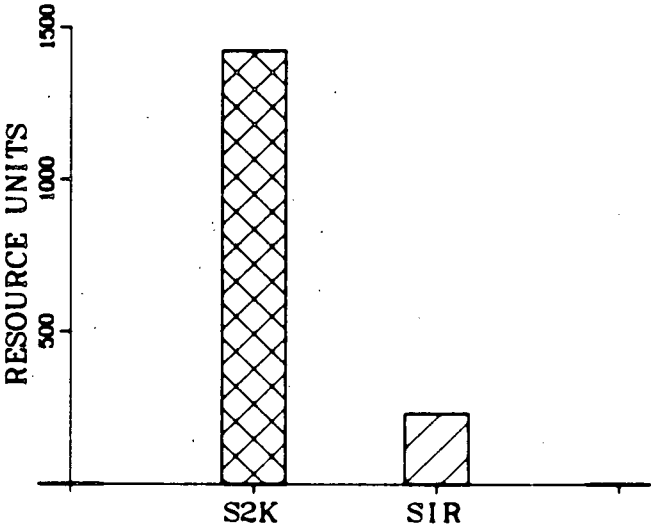


S2K vs SIR
Retrieval 2-3
List Data Occurring After 1975





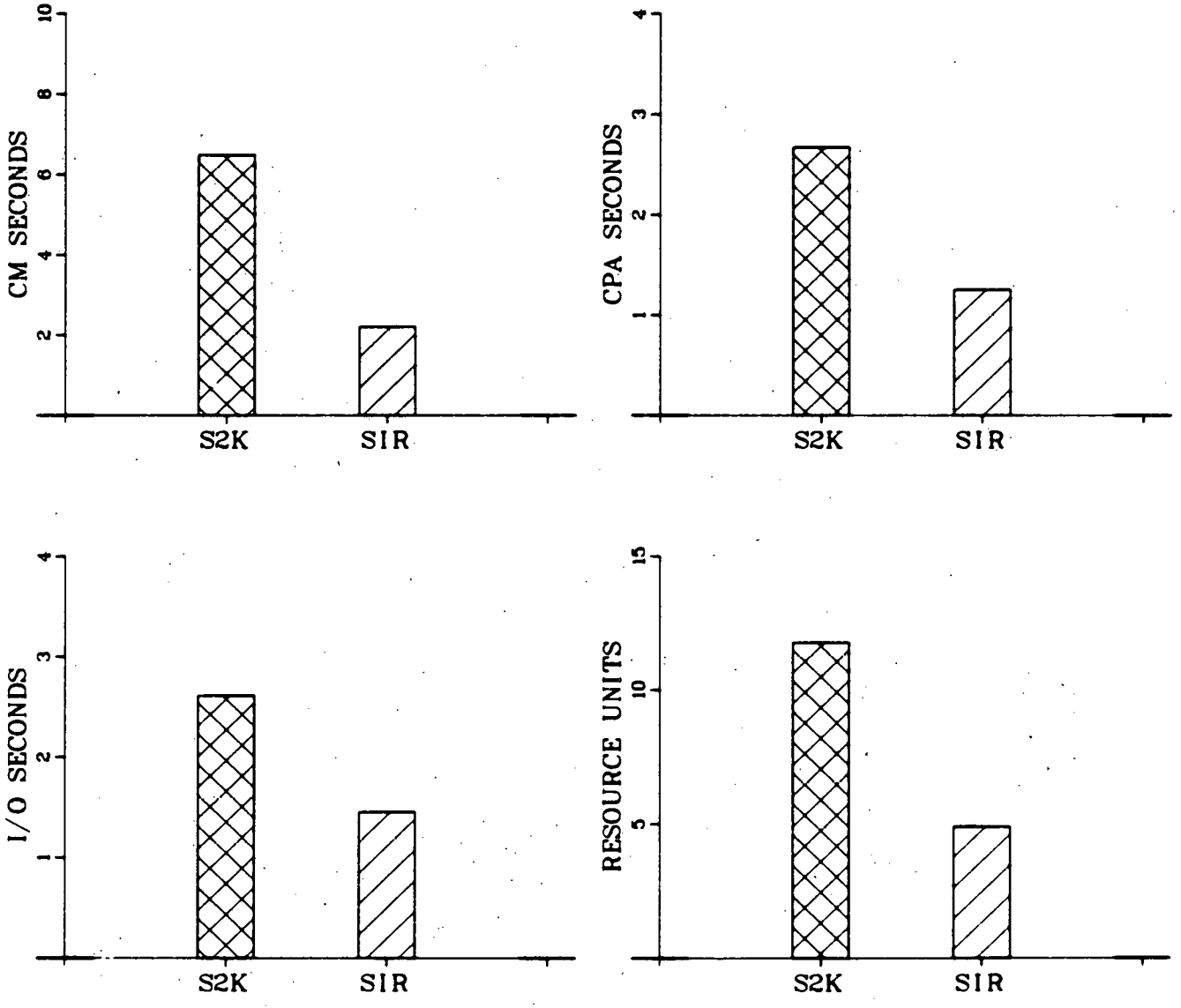




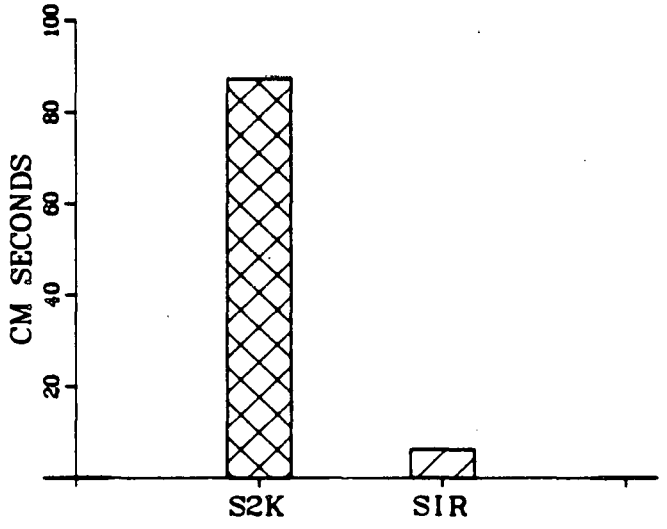
S2K vs SIR

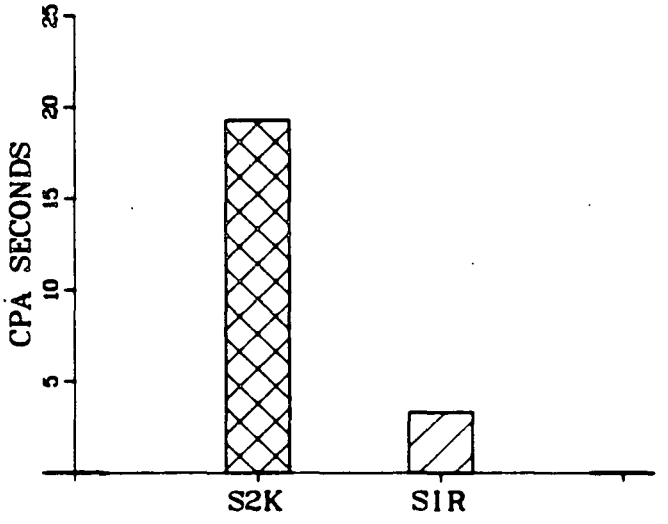
Retrieval 3-1

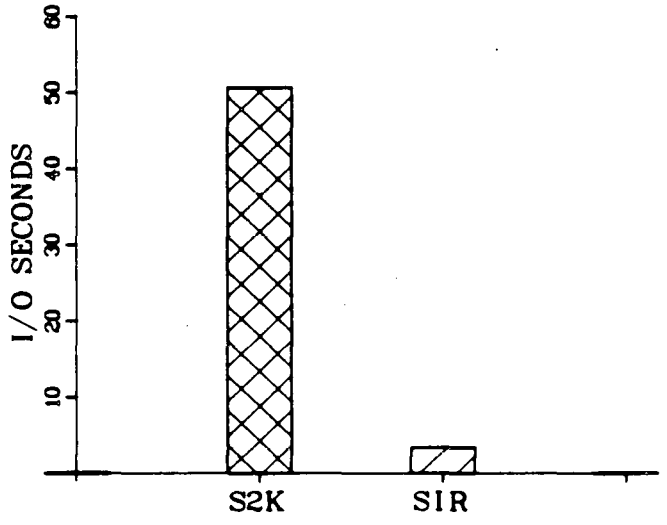
Min Alk By County After 1975

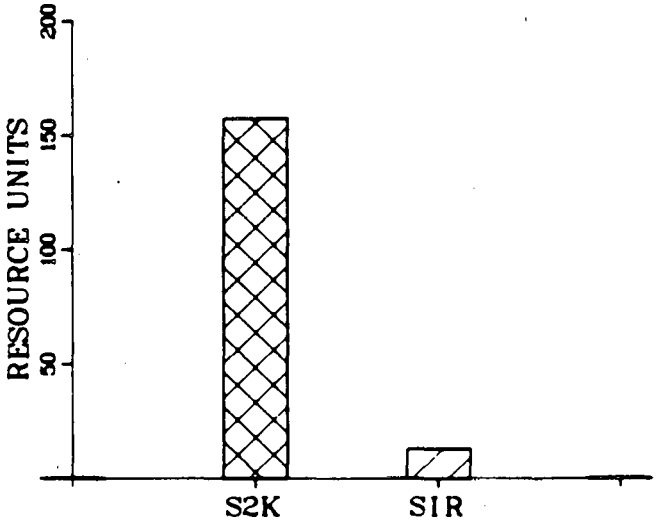


S2K vs SIR
Retrieval 3-2
Min Alk By County After 1975





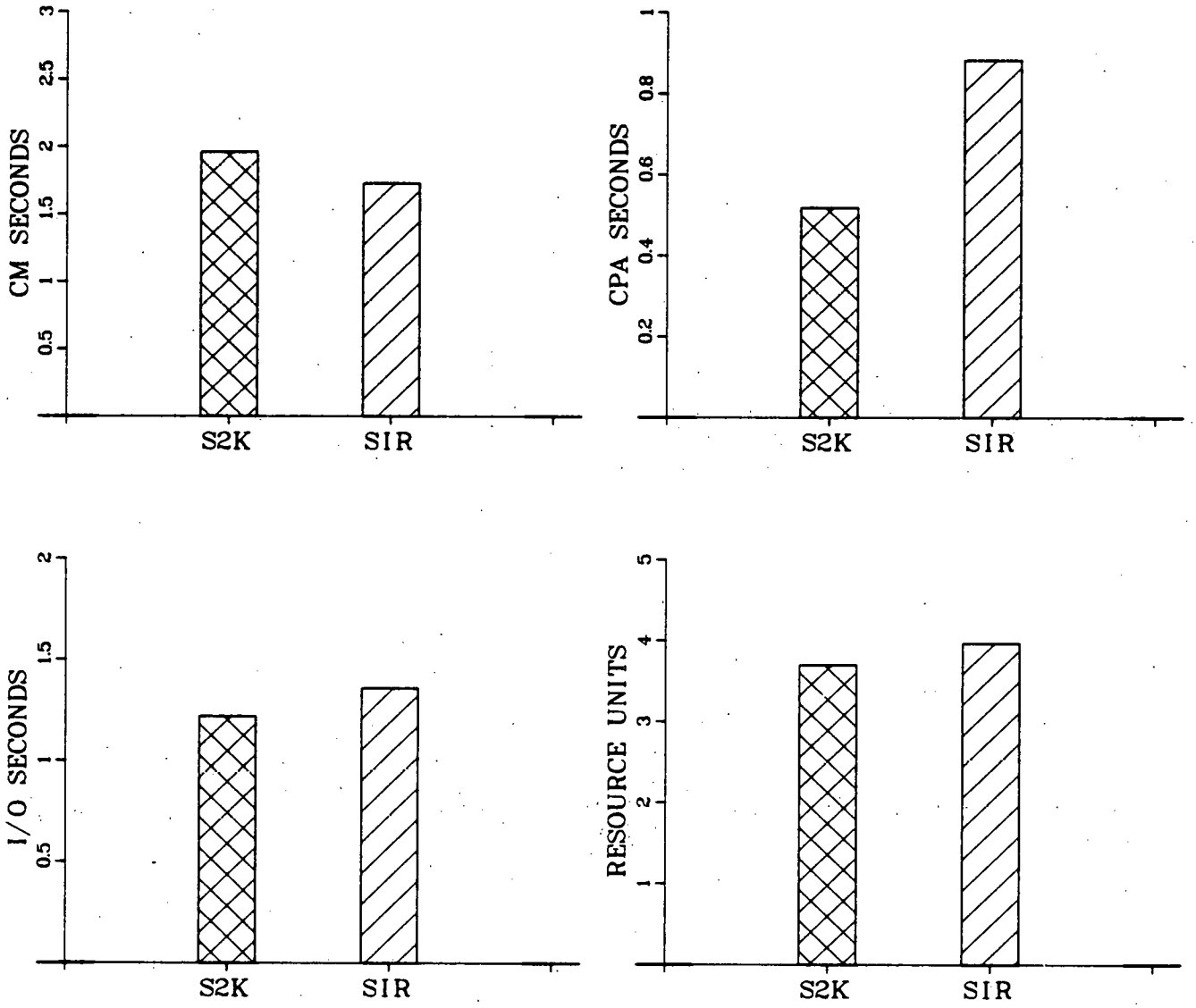




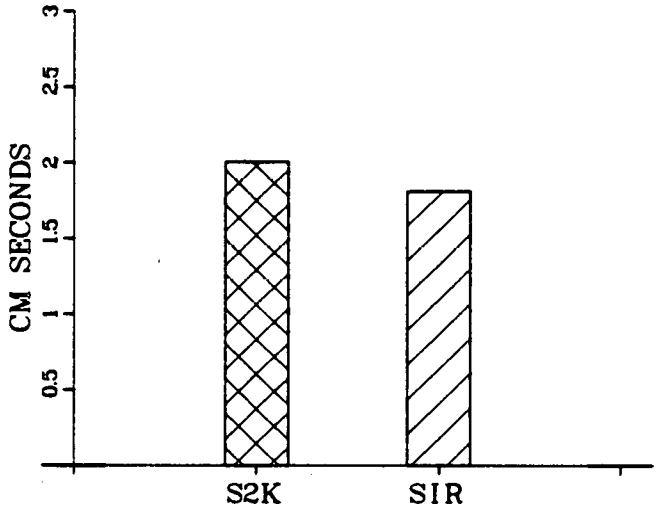
S2K vs SIR

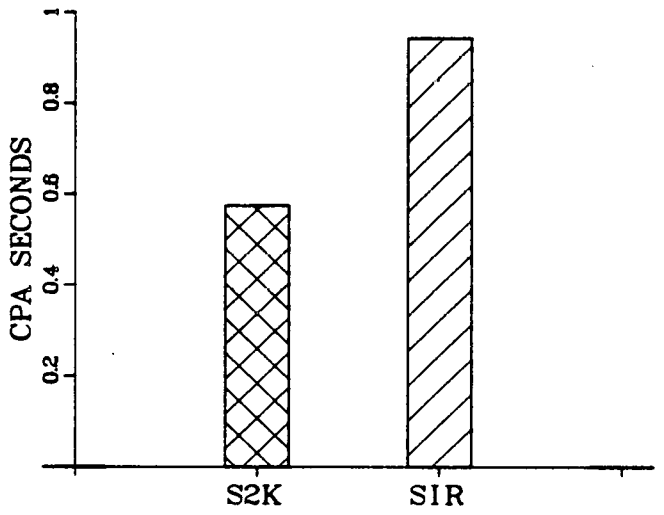
Retrieval 4-1

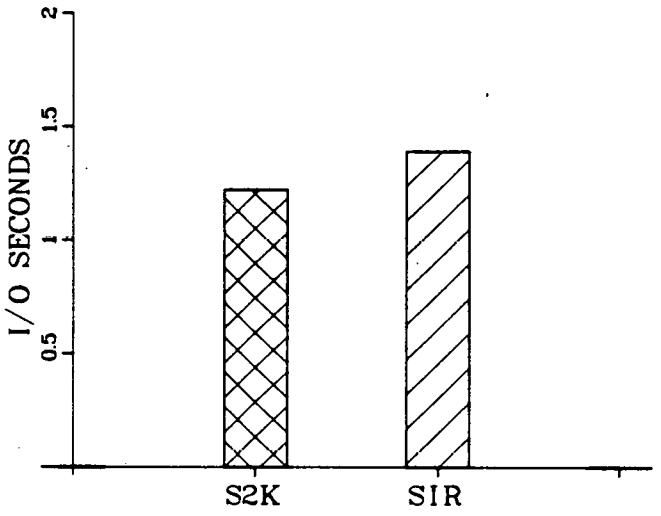
Min Alk By Station

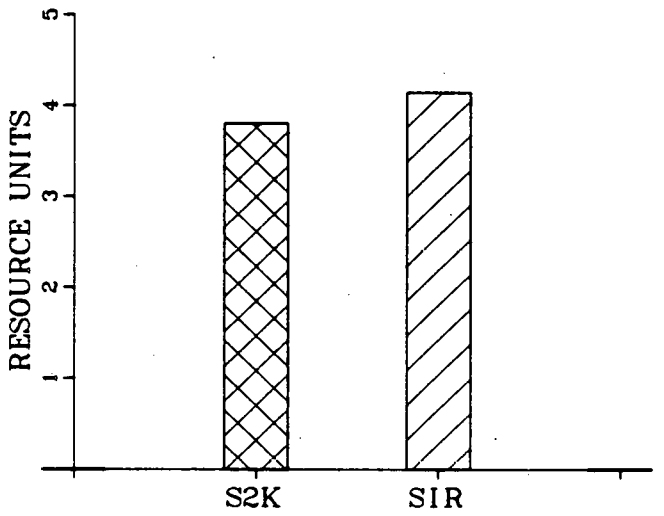


S2K vs SIR
Retrieval 5-1
List Only Storet Stations





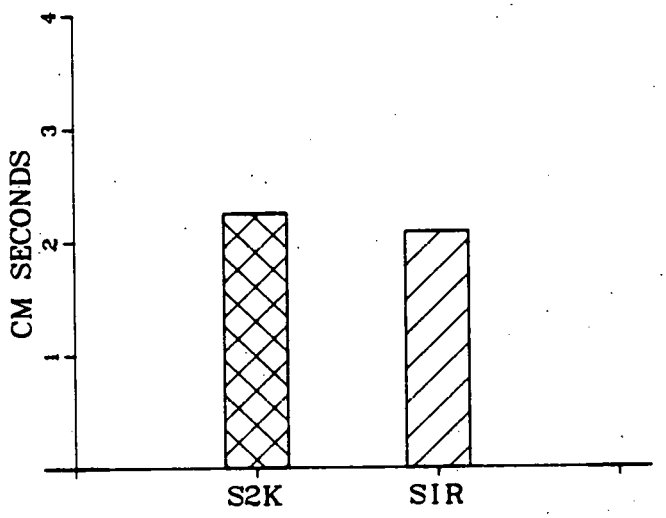


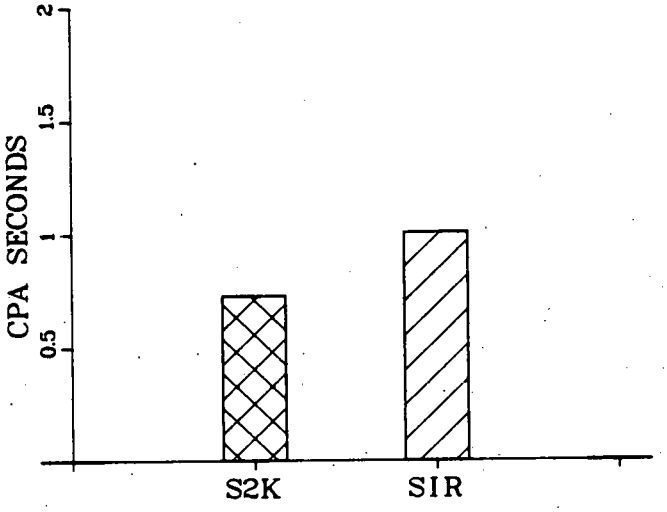


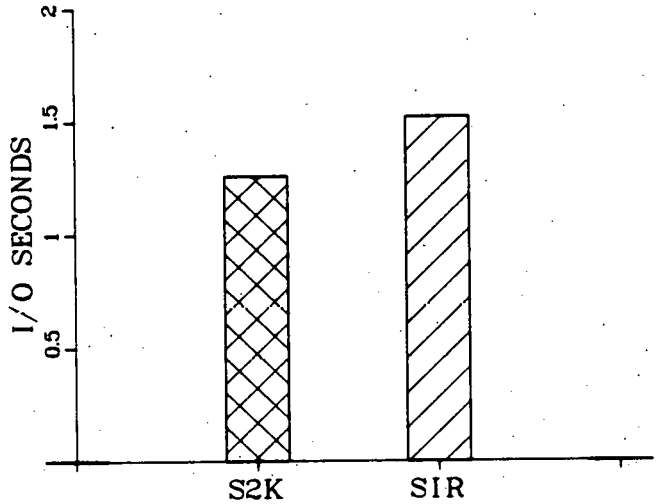
S2K vs SIR

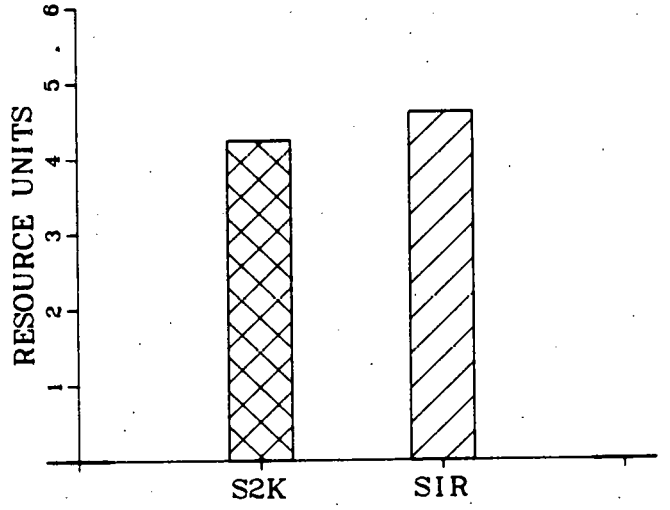
Retrieval 5-2

List Only Storet Stations

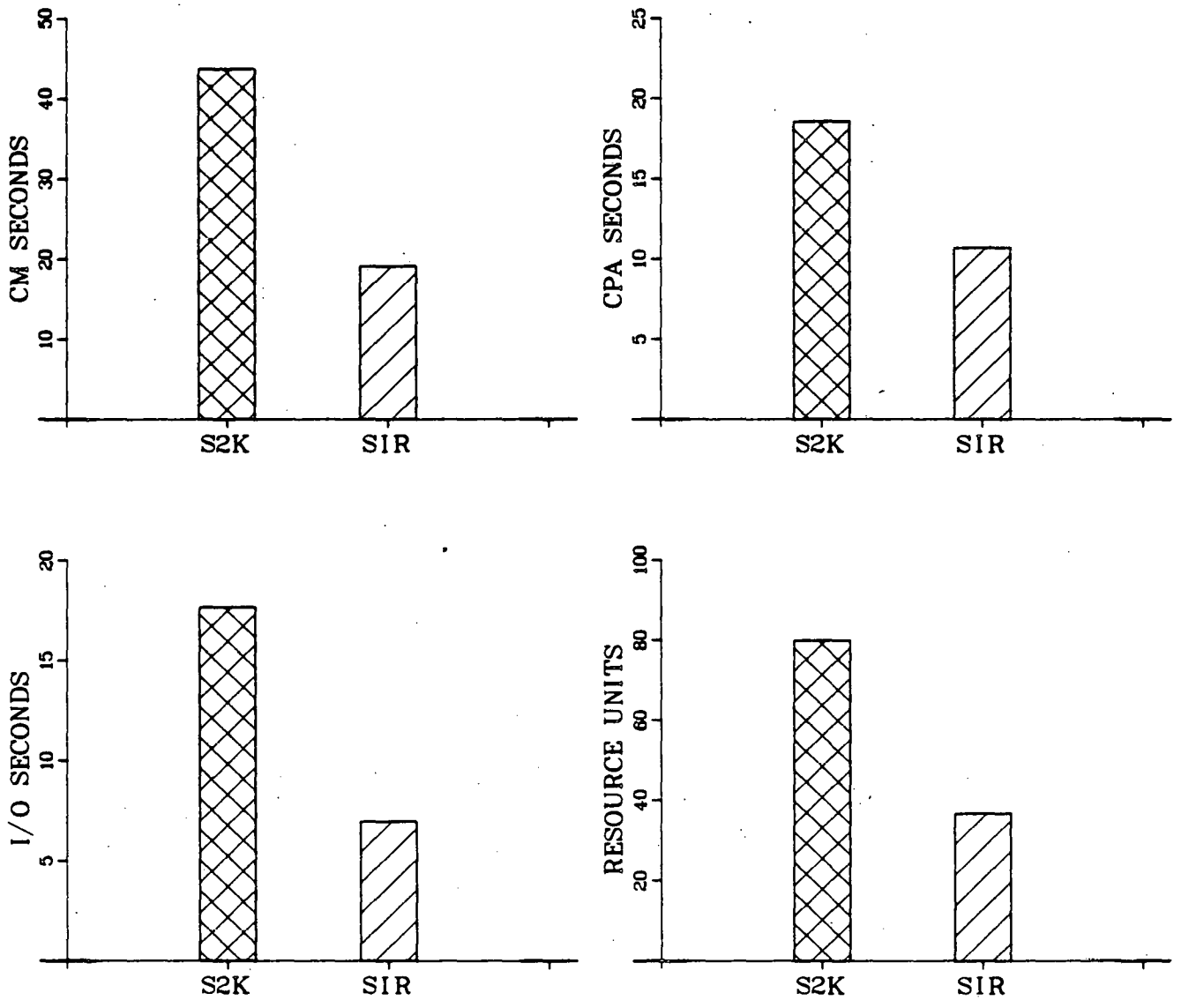




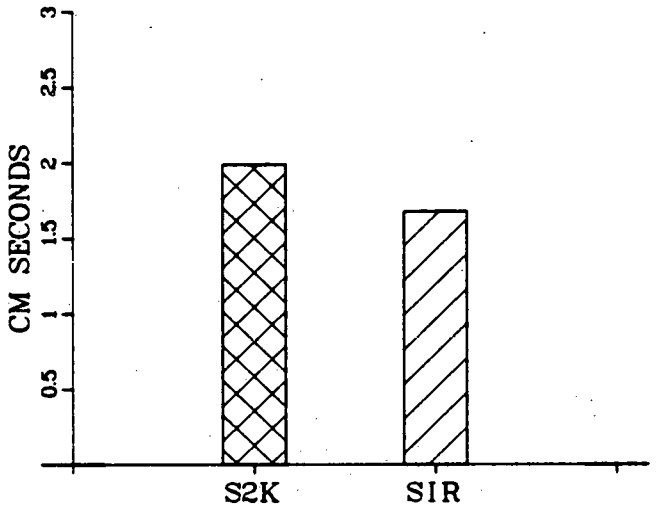


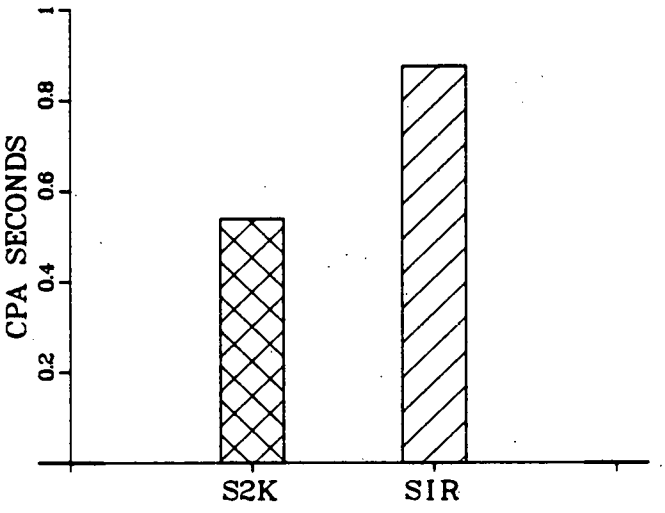


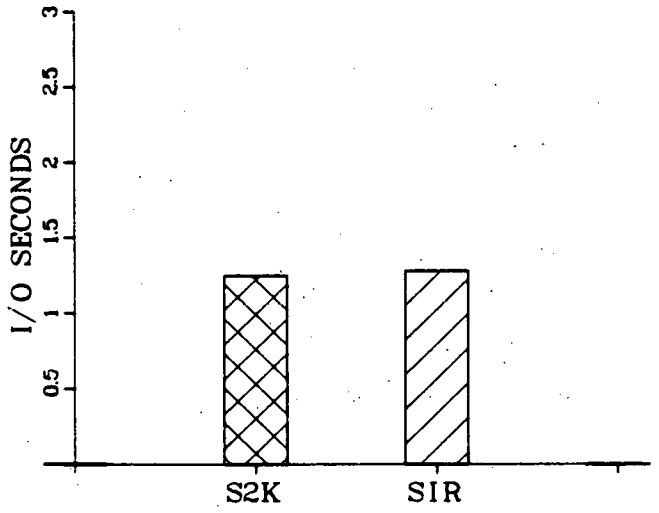
S2K vs SIR
Retrieval 5-3
List Only Storet Stations

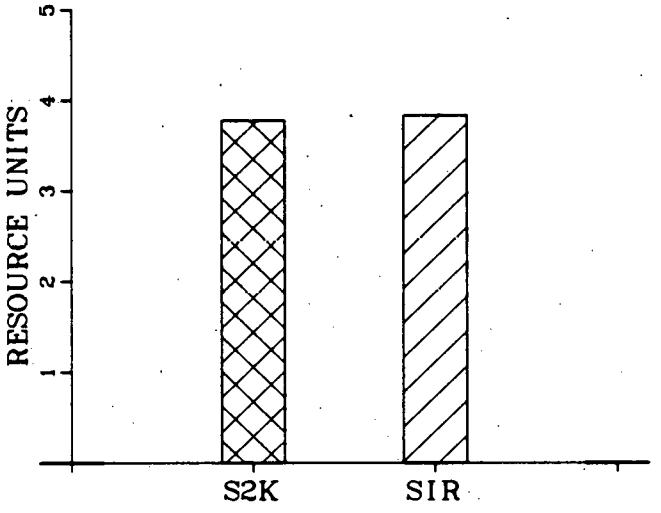


S2K vs SIR
Retrieval 6–1
Avg Alk By Station





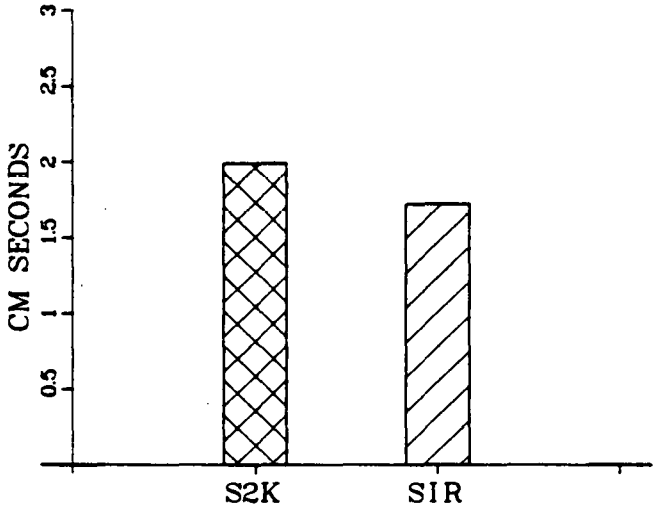


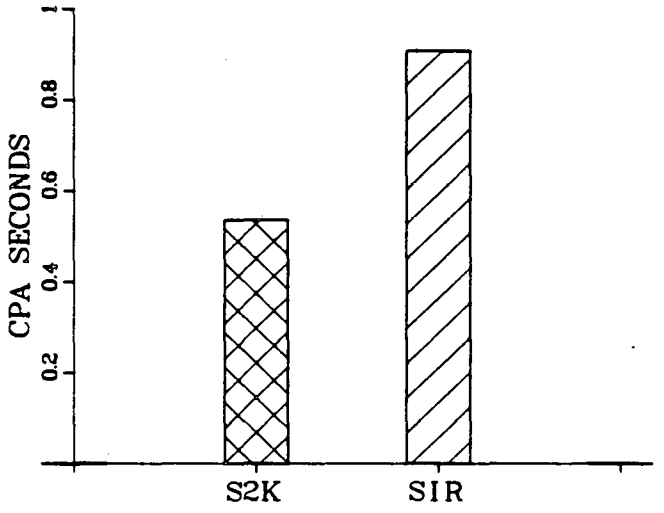


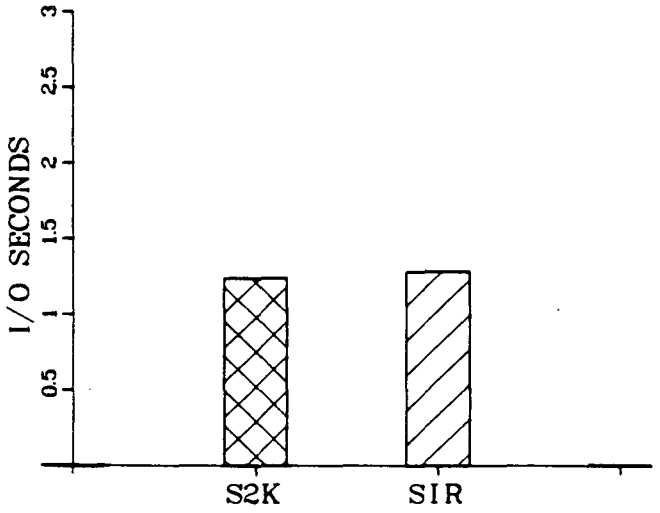
S2K vs SIR

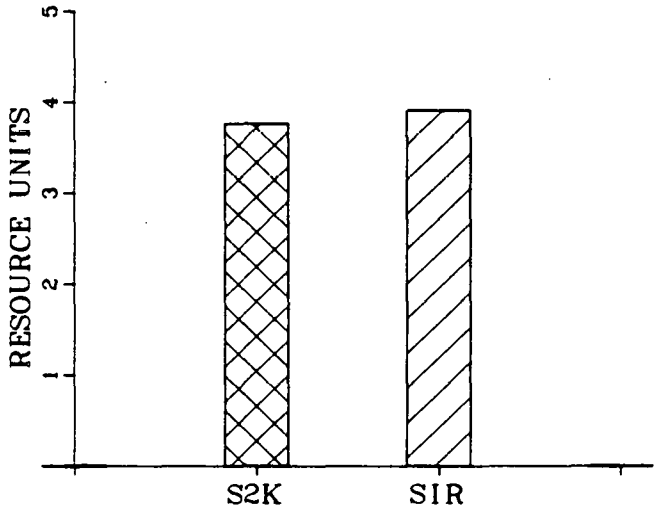
Retrieval 7-1

Avg,Max,Min Alk by Station

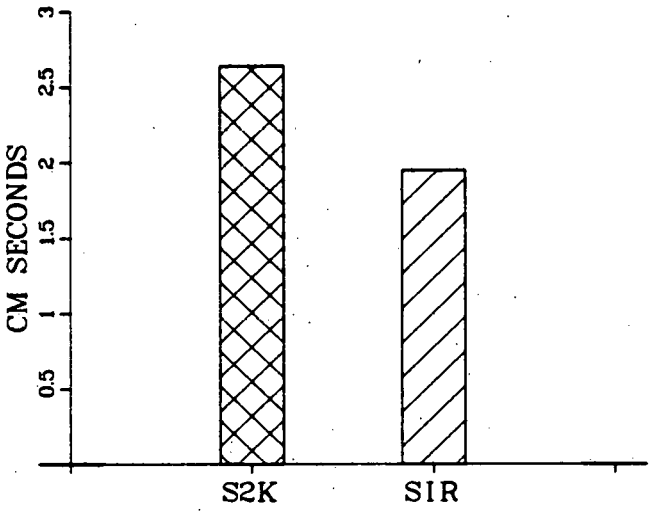


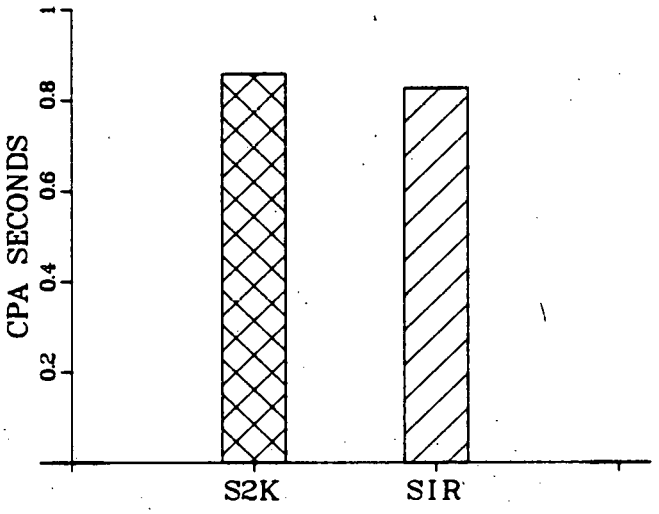


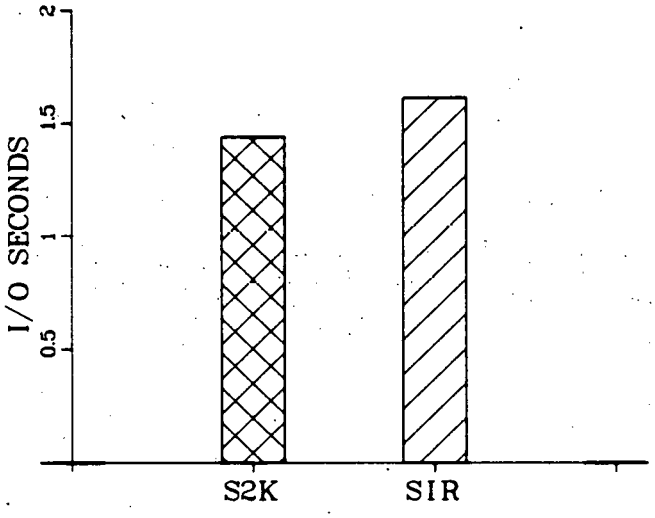


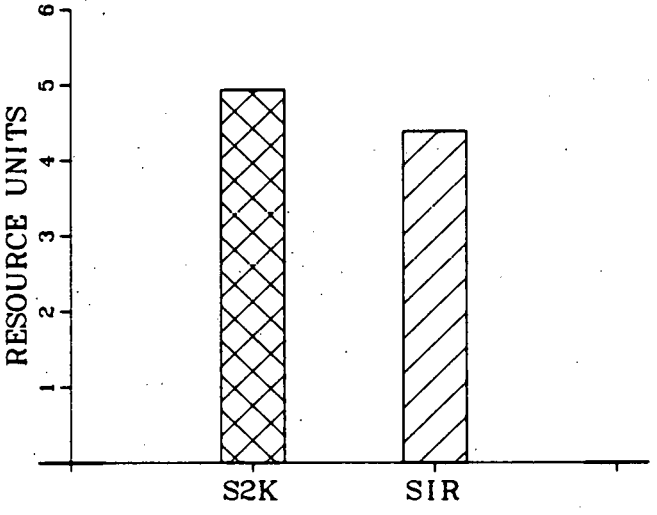


S2K vs SIR
Retrieval 8-1
List Geology Data





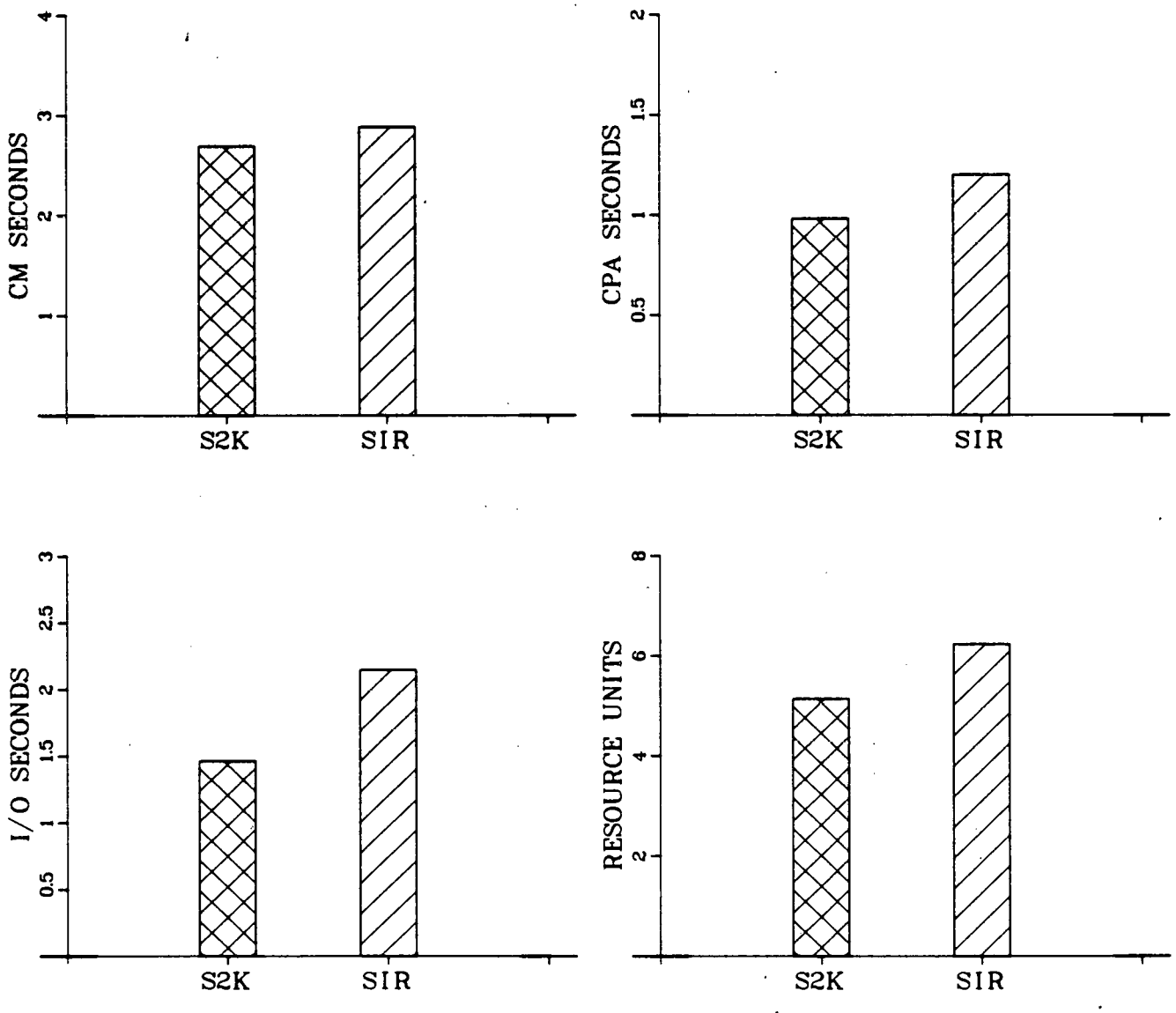




S2K vs SIR

Retrieval 9-1

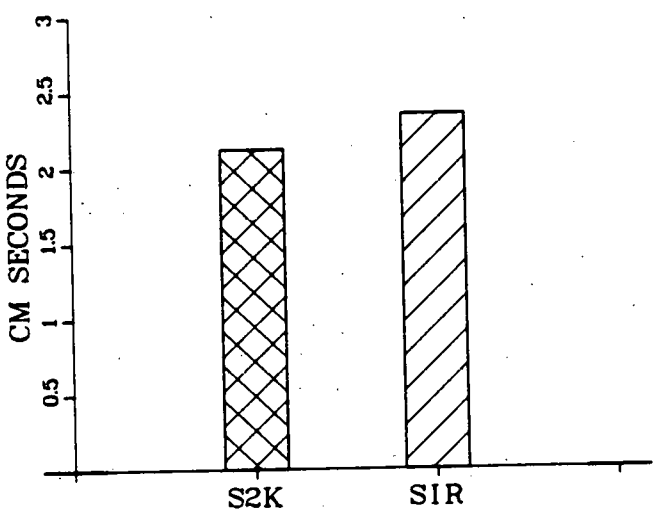
List Geology / Station Data

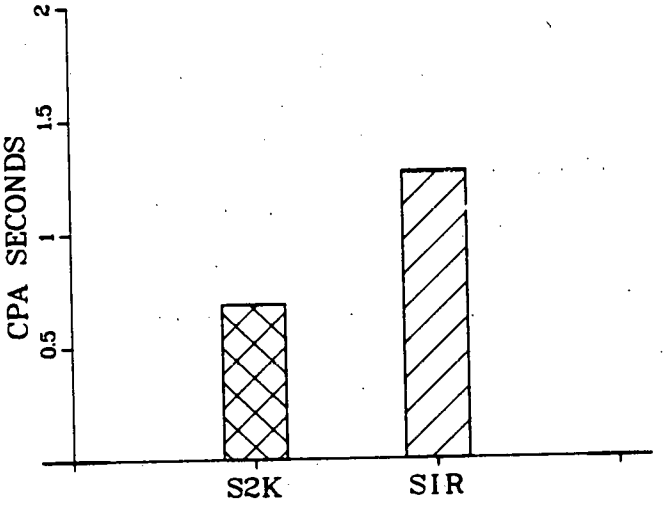


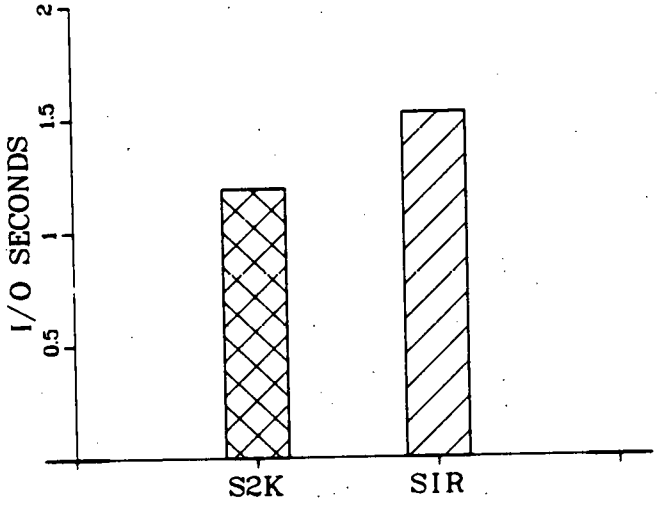
S2K vs SIR

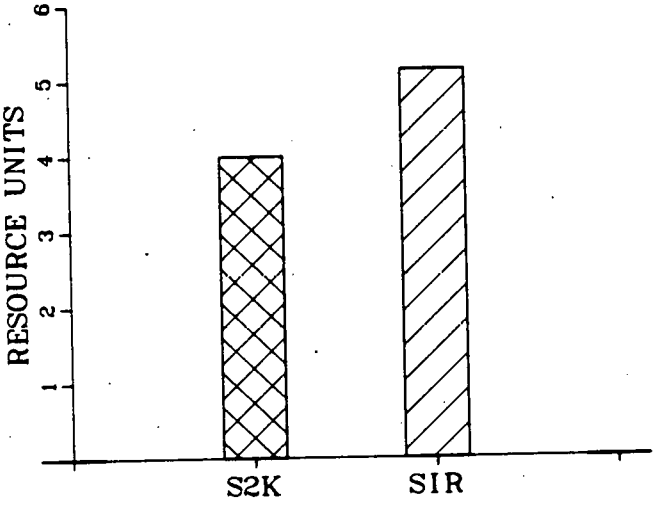
Modification 1-1

Change CA Value For CTFIP

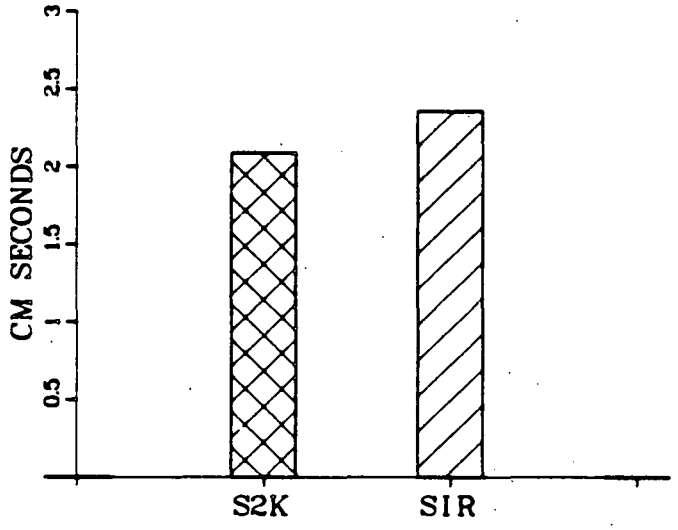


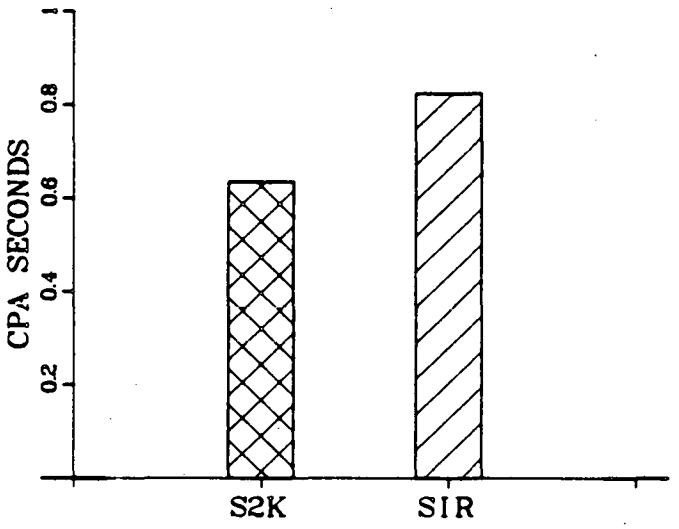


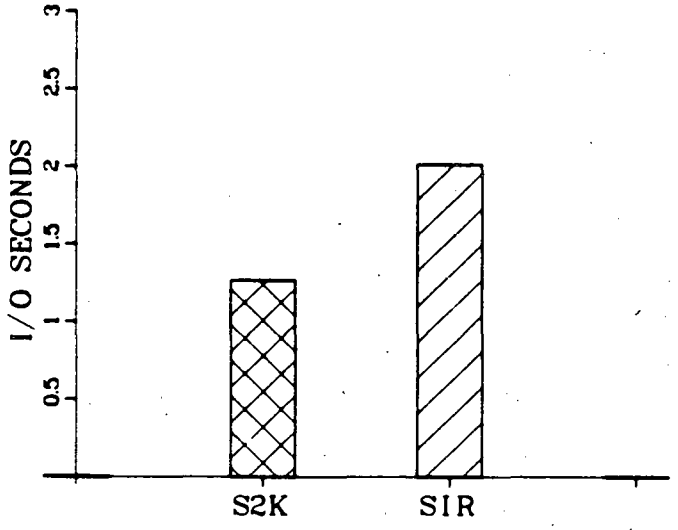


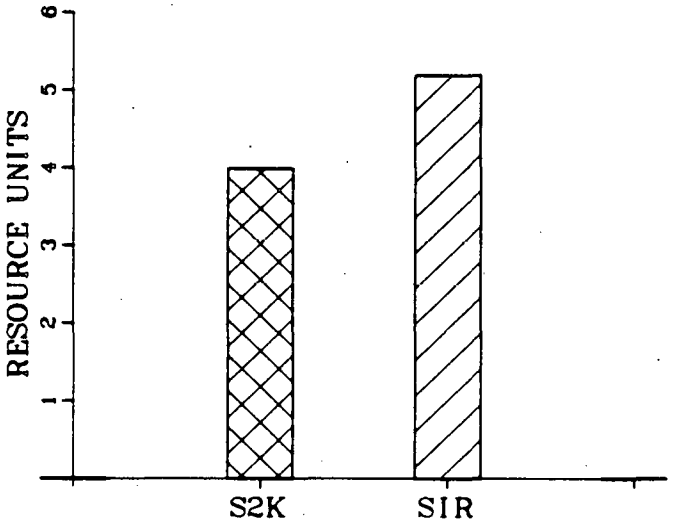


S2K vs SIR
Modification 1-2
Change CA Value For ST/CT





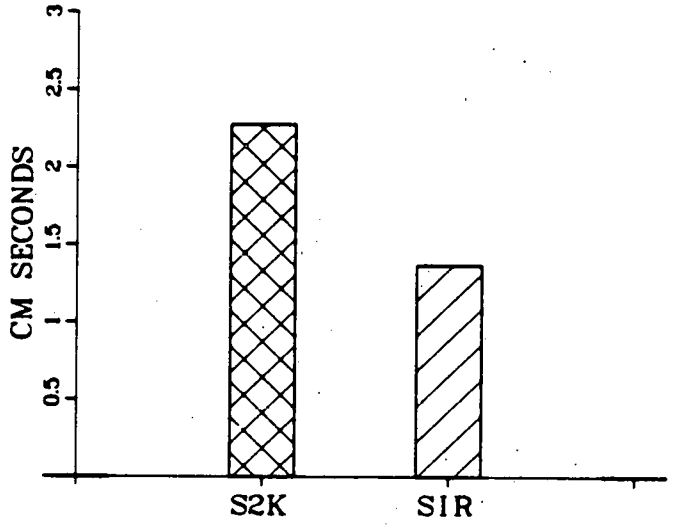


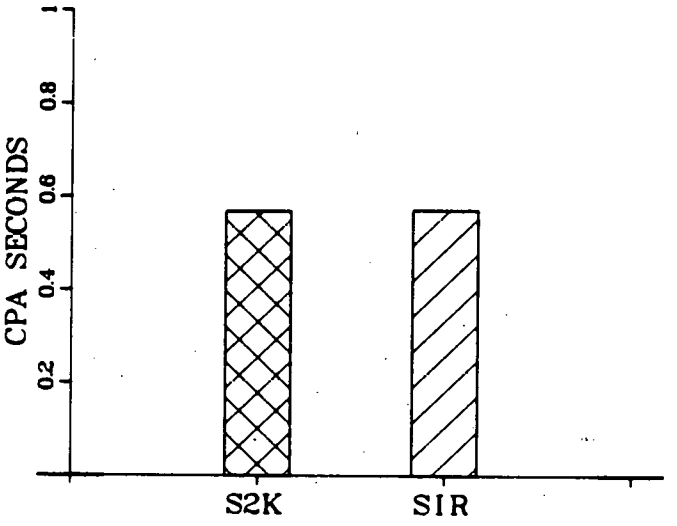


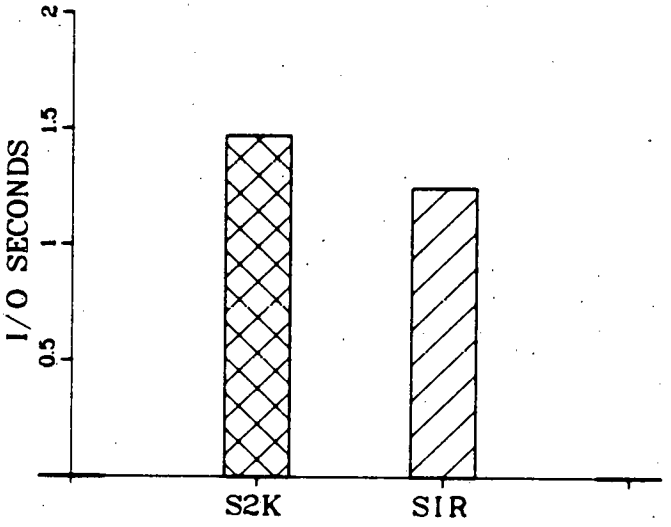
S2K vs SIR

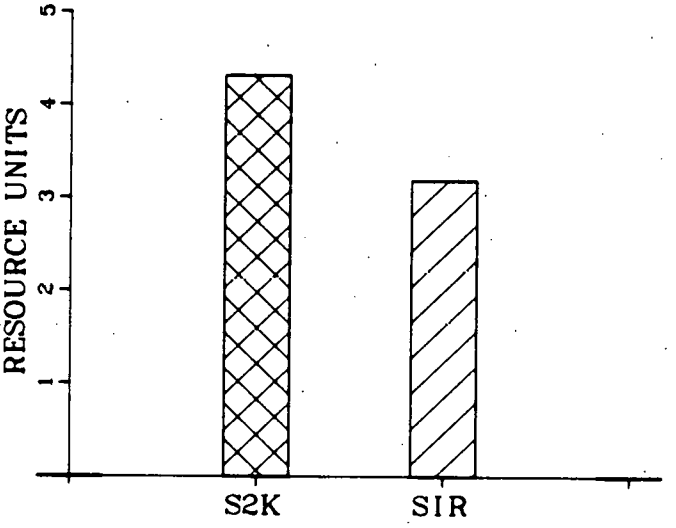
Modification 2-1

Add CA Value to a Station









S2K vs SIR

Modification 3-1

Change CA Value For CTFIP