Use of Monitors in FORTRAN:  
A Tutorial on the Barrier, 
Self-scheduling DO-Loop, 
and Askfor Monitors 

by 

E. L. Lusk and R. A. Overbeek
Use of Monitors in FORTRAN: A Tutorial on the Barrier, Self-scheduling DO-Loop, and Askfor Monitors

E. L. Lusk and R. A. Overbeek
Mathematics and Computer Science Division

June 1985
CONTENTS

1. Introduction ................................................................................................... 1
2. Example 1: Self-scheduling DO-loops............................................................... 3
3. Barrier Synchronization.................................................................................. 9
5. The Concept of Monitors ............................................................................. 21
6. An Introduction to the Askfor Monitor ......................................................... 25
7. A More Complex Example of the Askfor Monitor ........................................... 31
8. A Simple Grid Computation ......................................................................... 36
9. Locks ............................................................................................................. 47
10. The Fetch-and-Add ..................................................................................... 48
11. Running the Package .................................................................................. 50
12. Conclusion ................................................................................................ 50
References..................................................................................................... 51
Use of Monitors in FORTRAN:

A Tutorial on the Barrier,

Self-scheduling DO-Loop,

and Askfor Monitors

E. L. Lusk and R. A. Overbeek
Argonne National Laboratory
Argonne, Illinois 60439

ABSTRACT

A set of macro libraries has been developed that allows programmers to write portable FORTRAN code for multiprocessors. This document presents, in tutorial form, the macros used to implement three common synchronization patterns: self-scheduling DO-loops, barrier synchronization, and the askfor monitor.

1. Introduction

Monitors have been used for many years as a basis for process synchronization[1, 3, 4, 5, 6, 9]. After some experimentation, we have found that portable programs for use on multiprocessors can be written using monitors implemented by means of macros. Furthermore, once a programmer is familiar with the use of the macros, they offer a relatively convenient mechanism for writing programs that require the synchronization of multiple processes.

Two of our earlier publications[7, 8] discussed the basic concept of macros and briefly described their use. We are now writing a series of tutorials that more thoroughly explain how to use our macros to implement a variety of common synchronization patterns. This document is the first of the series and is intended for anyone who wishes to write portable FORTRAN for multiprocessors. Our earlier reports are not prerequisites to this document, but can be be accessed for more of the implementation details.

This work was supported in part by National Science Foundation grant MCS82-07496 and in part by the Applied Mathematical Sciences Research Program (KC-04-02) of the Office of Energy Research of the U.S. Department of Energy under Contract W-31-109-Eng-38.
One problem faced by a person beginning to code for multiprocessors is a choice of "synchronization primitives". A large number of primitives have grown up from applications in systems programming, telecommunications, and shared databases. The set of proposed alternatives may well appear bewildering to the uninitiated, and arguments between advocates of the different primitives can become quite heated. Unquestionably, the particular synchronization primitives that you select will have important consequences. Choosing primitives that are too "low-level" will almost inevitably lead to avoidable errors. On the other hand, utilizing primitives that are too "high-level" can preclude efficient implementation on many machines.

This second point is somewhat controversial, so let us briefly amplify our position. The use of a rendezvous mechanism of the sort included in Ada will (we believe) prevent the efficient implementation of an algorithm based on small-granularity parallelism. That is, the cost of any synchronization mechanism determines a lower bound on the size of the "work units" that are done in parallel.

We believe that monitors offer a reasonable conceptual framework for writing FORTRAN code for multiprocessors, without introducing overhead that unnecessarily restricts the class of algorithms that can be implemented.

Before offering a precise definition of what we mean by a monitor, we present a few programming examples that illustrate common synchronization patterns. These examples can then be used as a basis for forming a precise notion of the concept. Please note that our examples are (for the most part) relatively trivial. They are intended to clarify synchronization patterns and use of our macros; they are not considered of interest in themselves. The examples all illustrate "small-granularity" parallelism. That is, the computational tasks that are executed in parallel are very small. However, exactly the same synchronization patterns would be used in programs that capitalized on "large-granularity" parallelism.

To make our discussion concrete, we may occasionally discuss some aspect of a specific machine. We usually reference the Denelcor HEP, although our techniques are equally applicable to most multiprocessors that feature a number of processors operating on a globally shared memory.

In the programs covered in this tutorial, we use lowercase names to represent macro invocations. In addition, after each listing we briefly comment on each example.
2. Example 1: Self-scheduling DO-loops

Our first program illustrates a number of important concepts. The task is to read in two vectors of integers, add the two vectors, and write out the resulting vector. To illustrate the use of parallelism, we coordinate a number of processors that add different elements of the vectors in parallel. That is, one process computes \( C(1) = A(1) + B(1) \), while a separate process adds \( A(2) \) and \( B(2) \). (Adding two integers is a very small task; that is why our example illustrates small-granularity parallelism. If, instead, each element in the vector \( C \) were set to the smallest prime number greater than the product of the corresponding elements in the \( A \) and \( B \) vectors, the granularity would be considerably larger.)

The program used to add the two vectors is as follows:

```fortran
* THIS PROGRAM DEMONSTRATES THE "SELF-SCHEDULING DO-LOOP"
* SYNCHRONIZATION PRIMITIVE. IT READS IN TWO VECTORS, ADDS THEM,
* AND PRINTS THE RESULT.
*
PROGRAM ADDTWO
*
newproc(SLAVE)
*,
*
COMMON AREA VARIABLES
*
INTEGER A(1000), B(1000), C(1000)
INTEGER NPROCS, N
COMMON /MAINC/ A, B, C, N, NPROCS
*
******************************************************************************
*
* INITIALIZE THE ENVIRONMENT
*
******************************************************************************
```

```fortran
* DECLARE THE VARIABLES AND COMMON TO SUPPORT THE MONITOR
*
******************************************************************************
*
gsdec(GS)
*
INTEGER I
*
******************************************************************************
*
* INITIALIZE THE ENVIRONMENT
*
******************************************************************************
```
* initem

* INITENV

*****************************************************************************

* INITIALIZE THE SELF-SCHEDULING DO-LOOP MONITOR

*****************************************************************************

* gsinit(GS)

*****************************************************************************

* READ IN THE NUMBER OF PROCESSES TO RUN IN PARALLEL

*****************************************************************************

* READ (5,10) NPROCS

10 FORMAT(I4)

WRITE(6,20) NPROCS

20 FORMAT(' NPROCS = ',I4)

*****************************************************************************

* READ IN THE TWO INPUT VECTORS

*****************************************************************************

* READ (5,10) N

READ (5,10) (A(I), I = 1,N)

READ (5,10) (B(I), I = 1,N)

*****************************************************************************

* CREATE THE SLAVE PROCESSES

*****************************************************************************

* DO 30 I=1,NPROCS-1

create(SLAVE)

30 CONTINUE

* CALL WORK

* WRITE (6,10) (C(I), I = 1,N)

STOP

END
THE SLAVE PROCESSES JUST CALL THE WORK SUBROUTINE TO ADD UP ELEMENTS UNTIL THE END OF THE VECTOR IS REACHED. THE PROCESSES THEN EXIT (WHICH IS ASSUMED TO DESTROY THEM).

SUBROUTINE SLAVE
CALL WORK
RETURN
END

THE WORK SUBROUTINE JUST CAUSES A PROCESS TO GRAB AVAILABLE SUBSCRIPTS UNTIL ALL OF THE WORK HAS BEEN COMPLETED. AT THAT POINT THE SUBROUTINE EXITS. NOTE THAT IF THERE IS A SINGLE PROCESS (I.E., NO SLAVES), THE ALGORITHM STILL WORKS JUST FINE.

SUBROUTINE WORK
COMMON AREA VARIABLES
INTEGER A(1000), B(1000), C(1000)
INTEGER NPROCS, N
COMMON /MAINC/ A, B, C, N, NPROCS

gsdec(GS)
INTEGER I

10 CONTINUE

CLAIM THE NEXT AVAILABLE SUBSCRIPT (RETURNED IN I)

getsub(GS, I, N, NPROCS)

IF (I .EQ. 0) GO TO 20
A number of features of this program require discussion:

1. The user creates a COMMON area that will be shared between the processes. At this point, there is not complete agreement between vendors as to how to distinguish between COMMON that is common only to routines in one process (i.e., multiple copies exist for multiple processes) and COMMON that is globally shared. In our examples, we assume that COMMON is globally shared. To protect yourself, you could code a small macro that simply converts some word like "globcommon" to the appropriate wording for whichever machine you intend to run on. For experiments, this hardly seems necessary. For large production code development, such measures seem only prudent.

2. The program uses two macros, \texttt{initenv} and \texttt{env}, to perform initialization functions appropriate to a given environment. On the Denelcor HEP, we use this to read in the maximum amount of time (in minutes) that the program should execute. The initialization code starts a process that abnormally terminates after that time period. This fairly minor courtesy reduced dramatically the contention problems for access to the machine (people previously would "hang" for extended periods until an operator cancelled their jobs).

3. The \texttt{gsdec} macro is used to define a monitor that can be used to coordinate a "self-scheduling DO-loop". We will show you exactly what that means later in the code; for now all you really need to know is that the \texttt{gsdec(GS)} is used to generate the COMMON blocks required to coordinate processes to work on a self-scheduling DO-loop. It should be used
in your main program and in any routines that include `gsinit` or `getsub` macro invocations. The GS is an arbitrary two-character "name" of the monitor. It can be anything, but the names used in the `gsdec`, `gsinit`, and `getsub` invocations must agree. Note that we included `gsdec(GS)` in the main routine and in the WORK subroutine.

4. Line 41 contains `gsinit(GS)`, which initializes the variables that will be used to maintain proper synchronization between the processes. It must be invoked before more processes are created (i.e., before the other processes that will utilize the monitor begin to try acquiring subscript values from it).

5. Lines 49-52 represent the code that reads in the number of processes desired and prints it back out. It is quite useful to make the number of processes an input value. It is highly desirable that the code run properly with a value of 1, which allows the code to be (more or less) debugged on a uniprocessor.

6. Lines 70-72 create the extra processes that have been requested. Line 71 illustrates the use of the `create` macro. In this case, no parameters are passed to the "slave" processes. Execution of the `create` will cause a new process to be created, with execution beginning at the routine specified (in this case, SLAVE). Line 10 uses the `neuproc` macro to declare SLAVE as a routine name which will be assigned to a created process.

7. The created processes all start at SLAVE, which simply calls WORK. After creating the "slaves", the master process immediately calls WORK. If you examine the subroutine WORK, you will see the code that all of the processes are simultaneously executing. The "heart" of the self-scheduling DO-loop is given in lines 116-128. Each process executes the macro command `getsub(GS,I,N,NPROCS)`. Think of this command as executing instantaneously. This is, of course, not true; but because it is a monitor command it is indivisible. Only one process will be allowed to execute the command at any point in time. The result of the command is to return in I the next available subscript in the range 1-N. The first parameter is just the name of the monitor (to connect the `getsub` with the uses of the `gsdec` and `gsinit` macros), and the last parameter is the number of processes that will be competing for the subscripts. Each process claims a subscript (line 123). If there are no more, I will be returned as 0 (causing a GO TO 20). Otherwise, the subscript is processed (line 127), and the process returns to claim another available subscript. This continues until the processes exhaust the subscripts in the range 1-N.
There are a few fine points worth noting. First, no process will be returned a value of 0 before all processes have reached the point where they are requesting a subscript and there is none available (that is why NPROCS must be a parameter to the request -- it allows the monitor command getsub to be able to determine when all of the processes have completed their last tasks). This ensures that no process will actually return before the last subscript has been processed.

Second, it is assumed that when the slave processes exit from SLAVE (after returning from WORK) that they will "die"; furthermore, it is assumed that the slaves may exit either before or after the master returns (to line 74) to complete its functions. This assumption may, in fact, not be valid for some implementations (which would make the code fail to port properly). Please note that our development of these tools is still at a preliminary state. Once we have run experiments on several more machines, we are confident that a (more or less) complete set of macros will evolve leading to a high probability of portability.

Let us now pause and consider why this pattern of synchronization is called a "self-scheduling DO-loop". Note that the illustrated logic is used to replace a DO-loop of the following form (which would occur in the uniprocessing version of the code):

```
DO 10 1 = 1,N
    <body of loop>
10 CONTINUE
```

The pattern will work only when executions of the body of the loop can be done in parallel. As an exercise to see whether or not you fully grasp this pattern, we suggest that you consider the case in which you have nested loops that can all be run in parallel. That is, consider the following pattern:

```
DO 10 1 = 1,N
    DO J = 1,M
        <body of loop, which operates on A(I,J)>
    10 CONTINUE
```

Assume that the different instances of the loop body can all compute in parallel. Try to construct the appropriate logic using two monitors named G1 and G2 to implement such logic for parallel execution.
3. **Barrier Synchronization**

Now let us introduce a second pattern of synchronization called *barrier* synchronization. This type of synchronization is quite simple: the execution of

\[\text{barrier}(B_1, N)\]

will cause a process to "hang" in the monitor command until N processes are all "hanging". Then they are all released from the barrier and allowed to continue.

One common use of barrier synchronization occurs when multiple processes must be coordinated to perform a sequence of tasks, where the individual tasks must be executed strictly in order. To illustrate, suppose that we wished to perform a two-stage computation. The first stage is just as in our previous example, where C is set to A + B. The second stage involves creating a new vector D, where each value in D is a 0 or 1. The values in D are 0 exactly when the corresponding value in C is identical to the value that precedes it in C. That is, if A = (1,4,2,5,6) and B = (3,2,4,2,1), then C would be set to (4,6,6,7,7) and D would then be set to (1,1,0,1,0). This example is obviously contrived. Clearly, both C and D could be created in a single loop. But suppose that the creation of C is a complex task calling for parallelism and that the creation of D could then be done most efficiently after creating C. In this case, the following code might be used:

```
* THIS PROGRAM READS IN TWO VECTORS A AND B. IT THEN CREATES
* C = A + B. THEN D IS CREATED AS A VECTOR IN WHICH EACH ELEMENT
* IS SET AS FOLLOWS: D(I) = 0 IFF (C(I) = C(I-1). THOSE ELEMENTS
* OF D WHICH ARE NOT SET TO 0 ARE SET TO 1. D(1) IS ALWAYS SET
* TO 1.
*****

PROGRAM GETDUPS
*
newproc(SLAVE)
*
* COMMON AREA VARIABLES
*
INTEGER A(1000), B(1000), C(1000), D(1000)
INTEGER NPROCS, N
COMMON /MAINC/ A, B, C, D, N, NPROCS
*```
19     gsdec(G1)
20     gsdec(G2)
21     bardec(BA)
22     * INTEGER I
23     *
24     initenv
25     *
26     ************************************************************
27     *
28     * INITIALIZE THE MONITORS
29     *
30     ************************************************************
31     *
32     *
33     gsinit(G1)
34     gsinit(G2)
35     barinit(BA)
36     *
37     READ (5,10) NPROCS
38     10 FORMAT(I4)
39     WRITE(6,20) NPROCS
40     20 FORMAT(' NPROCS = ',I4)
41     *
42     ************************************************************
43     *
44     * READ IN THE TWO INPUT VECTORS
45     *
46     ************************************************************
47     *
48     READ (5,10) N
49     READ (5,10) (A(I), I = 1,N)
50     READ (5,10) (B(I), I = 1,N)
51     *
52     DO 30 I=1,NPROCS-1
53        create(SLAVE)
54     30 CONTINUE
55     *
56     CALL WORK
57     *
58     WRITE (6,40)
59     40 FORMAT(' THE VALUES IN THE C VECTOR ARE AS FOLLOWS:')
60     WRITE (6,10) (C(I), I = 1,N)
61     WRITE (6,50)
62     50 FORMAT(' THE VALUES IN THE D VECTOR ARE AS FOLLOWS:')
63     WRITE (6,10) (D(I), I = 1,N)
64     STOP
65     END
THE SLAVE PROCESSES JUST CALL THE WORK SUBROUTINE TO ADD UP
ELEMENTS UNTIL THE END OF THE VECTOR IS REACHED. THE PROCESSES
THEN EXIT (WHICH IS ASSUMED TO DESTROY THEM).

---------------------------------------------------

SUBROUTINE SLAVE

CALL WORK
RETURN
END

---------------------------------------------------

THE WORK SUBROUTINE PERFORMS A 2-STAGE COMPUTATION. FIRST,
C = A + B IS COMPUTED. THEN A BARRIER IS USED TO MAKE SURE THAT
NO PROCESS BEGINS THE SECOND STAGE UNTIL THE FIRST STAGE IS
COMPLETED. THE SECOND STAGE THEN USES ANOTHER SELF-SCHEDULING
DO-LOOP TO CALCULATE D.

---------------------------------------------------

SUBROUTINE WORK

* COMMON AREA VARIABLES

INTEGER A(1000), B(1000), C(1000), D(1000)
INTEGER NPROCS, N
COMMON /MAINC/ A, B, C, D, N, NPROCS

gsdec(G1)
gsdec(G2)
bardec(BA)

INTEGER I

CONTINUE
CONTINUE
getsu(sub(G1, I, N, NPROCS))
IF (I .EQ. 0) GO TO 20
C(I) = A(I) + B(I)
GO TO 10
Again, several points should be mentioned:

1. Two self-scheduling DO-loops are used in WORK. We used two monitors G1 and G2. Actually, since the `getsub` command reinitializes itself before releasing all of the processes at the end, a single monitor would suffice. We used two to illustrate the use of similar monitors with distinct names.

2. The barrier monitor is declared with `bardec(BA)` in a way completely analogous to the declaration for self-scheduling DO-loops. Similarly, `barinit(BA)` initializes the barrier.

3. The `barrier(BA,NPROCS)` in line 115 sets a barrier between the two stages of computation.

4. Since self-scheduling DO-loops do not release any process with a subscript of 0, the barrier is actually unnecessary.
In some sense, the above example is awful. It uses two monitors (G1 and G2) where one would suffice, and it uses an unnecessary barrier. However, it should help you to become more familiar with the statements required to declare, initialize, and use monitors. Furthermore, it does illustrate in a fairly simple manner how a barrier can be set.

4. A More Challenging Example

Let us proceed to a more challenging example. The next program constructs vectors in descending order and then sorts them into ascending order using a Shell sort algorithm. It acquires timings for sorting vectors of sizes 100, 1000, and 10000. The Shell sort is interesting in its own right and requires some explanation.

To see how a Shell sort works, one should first understand how a simple insertion sort works. Consider the vector (2,-1,1,6,10,-8,18,0). To sort this vector into ascending order, one could proceed from left to right starting with the second element and inserting each element into the sorted list to its left. Thus, -1 would be examined first and inserted into the list to its left (which contains only the value 2); the result would be (-1,2,1,6,10,-8,18,0). Then the third element would be examined and inserted into the sorted list to its left, producing (-1,1,2,6,10,-8,18,0). The fourth element would be left unchanged, as would the fifth; they are already in correct order. Processing the sixth element results in (-8,-1,1,2,6,10,18,0). The seventh element would be left unchanged. Finally, processing the last element would give (-8,-1,0,1,2,6,10,18). This process of inserting each new element into the sorted list to its left is the essential idea behind insertion sorting.

The problem with simple insertion sorting is that a large number of comparisons may be required to move a small number that occurs late in the list into its correct location. The Shell sort corrects this with a rather ingenious idea. Again, consider the vector (2,-1,1,6,10,-8,18,0). Now suppose that the sort were to take place in two stages. In the first stage, the original list of 8 elements is viewed as four interleaved lists. The first list is composed of the first element and the fifth (i.e., 2 and 10); the second list is (-1,-8); the third is (1,18); and the fourth is (6,0). During the first stage, each of these lists is ordered using a simple insertion sort, giving (2,-8,1,0,10,-1,18,6). Clearly, fairly long moves can occur during the first stage. Then the resulting list can be sorted with a simple insertion sort.

This notion becomes quite significant when the original vector is long. Of course, we can carry the idea further and use more stages with gradually diminishing numbers of lists, as long as the last stage treats the whole vector as a single list.
This basic notion led to the following algorithm for sorting a vector V containing N elements:

procedure shellsort(V, N)
    compute h(1), h(2), ... h(i+2) such that h(1)=1,
    h(k) = (3*h(k-1)) + 1, h(i+1) < N, and h(i+2) > N.
    set gap = h(i)
    while (gap > 0) do
        set nextel = gap + 1
        while (nextel <= N) do
            sortel = nextel - gap
            while (sortel > 0) and (V(sortel) > V(sortel+gap)) do
                exchange V(sortel) and V(sortel+gap)
                set sortel = sortel - gap
            enddo
            set nextel = nextel + 1
        enddo
        set gap = (gap-1)/3
    enddo
endproc

The sort algorithm is somewhat complex, but it does represent a fairly good sort algorithm that is easily implemented.

The above algorithm can clearly take advantage of multiprocessing on all but the last pass (i.e., on all executions of the main loop with a "gap" size greater than 1). This is achieved by allowing separate processes to sort the
interleaved lists. The program that achieves this result using self-scheduling DO-loops and barrier synchronization is as follows:

```fortran
1  define(rmde,REAL)
2  ************************************************************
3  *
4  * THIS PROGRAM DEMONSTRATES THE "BARRIER" AND "SELF-SCHEDULING DO-LOOP"
5  * SYNCHRONIZATION PRIMITIVES. IT FILLS IN A VECTOR (A) WITH VALUES IN
6  * DESCENDING ORDER. THEN IT USES A SHELL SORT (SEE KNUTH'S 3RD VOLUME
7  * ON SORTING AND SEARCHING ALGORITHMS) TO SORT THE VALUES INTO
8  * ASCENDING ORDER. TIMES ARE ACQUIRED FOR TABLE SIZES OF 100, 1000, AND
9  * 10000.
10  *
11  ************************************************************
12
13  PROGRAM SRTPGM
14  *
15  newproc(SLAVE)
16  *
17  * COMMON AREA VARIABLES
18  *
19  mode A(10000)
20  INTEGER NPROCS, N, GAP
21  LOGICAL PGDONE
22  COMMON /MAINC/ GAP, A, PGDONE, N, NPROCS
23  *
24  gsdec(GS)
25  bardec(B1)
26  *
27  INTEGER I,J
28  *
29  initenv
30  *
31  ************************************************************
32  *
33  * INITIALIZE THE BARRIER AND SELF-SCHEDULING DO-LOOP MONITORS
34  *
35  ************************************************************
36  *
37  gsinit(GS)
38  barinit(B1)
39  *
40  PGDONE = .FALSE.
41  *
42  READ (5,10) NPROCS
```
10 FORMAT(I4)
WRITE(6,20) NPROCS
20 FORMAT( ' NPROCS = ',I4)

DO 30 I=1,NPROCS-1
   create(SLAVE)
30 CONTINUE

THE MAIN LOGIC JUST FILLS IN THE TABLE AND Sorts IT.
TIMINGS ARE TAKEN FOR TABLES OF 100, 1000, AND 10000.

N = 10
DO 50 I=1,3
   N = 10 * N
   CALL FILL
50   clock(J)
   T1 = J
   CALL SORT
51   clock(J)
   T2 = J - T1
   WRITE(6,40) N, T2
40   FOR1IAT(' SIZE = ',I5,' TOTAL TIME = ',E12.5)
50 CONTINUE

ONE LAST CALL TO LOOP IS REQUIRED TO FREE THE OTHER PROCESSES

FROM THE BARRIER (SO THEY CAN EXIT).

PGDONE = .TRUE.
CALL LOOP(0)
STOP
END
THE FOLLOWING LITTLE ROUTINE JUST FILLS THE VECTOR WITH VALUES IN DESCENDING ORDER.

SUBROUTINE FILL

mode A(10000)
INTEGER NPROC, N, GAP
LOGICAL PGDONE
COMMON /MAINC/ GAP, A, PGDONE, N, NPROC

INTEGER I

DO 10 I = 1, N
A(I) = (N - I) + 1.0
10 CONTINUE

RETURN
END

THE SLAVE PROCESSES JUST HANG ON THE BARRIER IN THE "LOOP" AND HELP WHEN A TABLE IS TO BE SORTED.

SUBROUTINE SLAVE

CALL LOOP(1)
RETURN
END

THE SORT ROUTINE IS EXECUTED BY THE MASTER PROCESS. IT JUST CALCULATES THE RADIX FOR EACH PASS OF THE SHELL SORT, AND JOINS THE SLAVE PROCESSES WHEN WORKING ON EACH PASS. THE RADIX VALUES ARE H0, ..., H2, H1: H0 IS 1; H1 IS (3*H(1-1) + 1); H(T+2) >= N. SEE KNUUTH FOR ARGUMENTS IN FAVOR OF THESE VALUES.

SUBROUTINE SORT

mode A(10000)
INTEGER NPROC, N, GAP
LOGICAL PGDONE
COMMON /MAINC/ GAP, A, PGDONE, N, NPROCS
*
INTEGER I1, I2, I3
*
I1 = 1
I2 = (I1 * 3) + 1
I3 = (I2 * 3) + 1
10 CONTINUE
11 IF (I3 .GE. N) GO TO 20
12 I1 = I2
13 I2 = I3
14 I3 = (I2 * 3) + 1
15 GO TO 10
20 CONTINUE
GAP = I1
30 CONTINUE
31 IF (GAP .LE. 0) GO TO 40
32 CALL LOOP(0)
33 GAP = (GAP - 1) / 3
34 GO TO 30
40 CONTINUE
RETURN
END

******************************************************************************
* THE LOOP ROUTINE IS THE CODE REQUIRED TO COORDINATE THE NPROCS PROCESSES AS THEY EXECUTE ONE PASS OF A SHELL SORT. NOTE THE BARRIER AT THE TOP, WHICH IS USED TO CAUSE THE PROCESSES TO WAIT FOR THE VECTOR TO BE SET UP AND THE INCREMENT CHOSEN. THEN A SELF-SCHEDULING DO-LOOP IS USED TO ALLOCATE SUBSCRIPTS. NOTE THAT THE MASTER PARTICIPATES IN THIS LOGIC, SO THE PROGRAM CAN BE RUN WITH NPROCS SET TO 1.
******************************************************************************

SUBROUTINE LOOP(WHO)
INTEGER WHO
mode A(10000)
INTEGER NPROCS, N, GAP
LOGICAL PGDONE
COMMON /MAINC/ GAP, A, PGDONE, N, NPROCS

Again, several features of this program merit discussion:
1. First, line 1 is an example of a very simple macro. It states that every occurrence of the word "mode" will be replaced by the word "REAL". This macro allows a programmer to easily alter features of a program. We use the UNIX macro processor `m4`. Any macro processor will work equally well, but `m4` is a standard UNIX utility and will become commonly available. Because our approach to coding requires you to code simple macros, we encourage you to gain familiarity with its use. We hope that all of the required details will be adequately covered in this document, but for more details you should consult the UNIX documentation.

2. The variable PGDONE is set to TRUE when there are no more vectors to sort. It is referenced by the slave processes when they fall through the barrier in LOOP. It is used as a signal to the slaves to destroy themselves.

3. The main routine just constructs vectors to sort (by calling FILL) and accumulates the time required to sort each table. Note the use of the `clock` macro to read the system clock (lines 61 and 66). In environments that do not support a system clock, the macro generates code to return a meaningless value (as is done, for example, with the macros designed for use on UNIX in a uniprocessing environment).

4. In this code the routine LOOP is multiprocessed logic. Since the master process prepares computational tasks (i.e., sets up vectors to sort, and creates the parameters for each stage of a sort), it will exit LOOP after each stage of a sort. The slaves, on the other hand, always wait in LOOP for the next stage of a sort (they hang on the barrier in LOOP, waiting for the master to set up the problem and join them at the barrier). This type of logic requires that each process know whether it is a slave or the master. This is achieved by passing a parameter (0 for the master; 1 for a slave) whenever the LOOP subroutine is called (lines 82, 117, and 155).

5. The slaves are created and wait in LOOP for the master to join them. Before joining them, the master process (in the routine SORT) calculates the "gap" required on the next stage of a sort. The master calls LOOP with decreasing "gap" lengths until the a vector is completely sorted (i.e., until after calling LOOP with GAP set to 1).

6. The logic of LOOP is based on the idea that each process selects one of the "interleaved lists" and sorts it into order. For each stage of a sort, there will be GAP lists to sort.

This sort certainly does not fully utilize all available parallelism. In fact, the last stage of each sort (with GAP=1) is handled by a single process. However, on
large vectors it does give a speedup roughly equal to half the number of available processors.

5. The Concept of Monitors

We have deferred discussion of exactly what we mean by a monitor to allow you to become familiar with a number of the details involved in our basic approach to implementing multiprocessing algorithms. It is now time to analyze the concept of monitors in some detail.

In our examples of self-scheduling DO-loops and barriers, it was very important that each process be able to view execution of getsub and barrier commands as indivisible operations. The implementation of these commands involves data structures (generated by gsdec and bardec macros) that are shared among a number of processes. The essential problem here is to make sure that access and updates are carefully synchronized to prevent destructive interference among the processes. Sections of code that can safely be executed by only a single process at a time are called critical sections.

For each set of shared data structures, there is a set of operations that represent critical sections. In the case of a self-scheduling DO-loop, there is a single critical section (claiming the next subscript) associated with the data structures. In more complex cases, there may be a number of critical sections.

A monitor is a conceptual abstraction composed of three distinct parts:

1. The data that are shared,
2. The operations that represent critical sections associated with the shared data, and
3. The code required to initialize the shared structures.

Thus, when we speak of using a monitor to coordinate activity based on some shared data, we mean that the programmer carefully defines the shared data, the operations on the shared data, and the operations required to initialize the data. The details of how monitor operations are implemented to prevent destructive interference represent the central difficulty. When a process has begun to execute a critical section (i.e., when it enters a monitor operation), we refer to this as "having entered the monitor". Only one process may be "in a monitor" at any given point. This is achieved conceptually by setting a lock at the point where the operation is entered and releasing the lock at the point where the operation is exited. Thus, in a self-scheduling DO-loop, the operation getsub is basically

set a lock
claim the next subscript (which is a shared variable)
release the lock.

In the case of the self-scheduling DO-loop, the gsdec(GS) defines the shared data in a COMMON block; The gsinit(GS) initializes the lock to "not set" and the subscript to 1. However, in the case in which there are no more available subscripts, we decided to cause the process issuing the getsub to "hang" until all of the remaining processes were also requesting a subscript. Only then will the getsub release them all (with a value of 0 to indicate that they were not successful in claiming a subscript). Conceptually, we "hang" a process as follows.

When a process is in a monitor and cannot complete the operation until a later point, it is delayed. This means that the process relinquishes the monitor and is placed in a queue. A delay queue can contain any number of "hung" processes, in arbitrary order. Each delayed process can be reactivated only by another process issuing a continue. It is important that only one process ever be "in a monitor". Thus, the process that issues the continue (which can be done only when it is in the monitor) exits from the monitor at the same instant that the (arbitrarily chosen) delayed process is reactivated (and regains control of the monitor).

Summarizing, we have identified four basic monitor operations:

menter(<monitor-name>): this operation is the first instruction in a monitor operation. It sets a lock, which ensures that the process can enter the monitor operation only in the case in which no other process is active in any operation associated with <monitor-name>.

mexit(<monitor-name>): this operation is the last instruction in a monitor. It releases the lock set by menter(<monitor-name>).

delay(<monitor-name>,<queue-number>): this operation causes the process that is in the monitor to be delayed in the queue designated by <queue-number>. The process loses control of the monitor.

continue(<monitor-name>,<queue-number>): this operation causes the active process to immediately exit from the monitor. If there are any delayed processes in the designated queue, one of them will be activated (thus, inheriting the lock set by the process leaving the monitor). Otherwise, the lock will be released by the exiting process.

With these definitions in hand, we can now give the exact logic of the getsub monitor:
getsymb(<monitor-name>,<returned-subscript>,<max-val>,<num-proc>)

menter(<monitor-name>)

if (the shared subscript is less than or equal to <max-val>) then

set <returned-subscript> to the value of the shared subscript
increment the shared subscript

else

set <returned-subscript> to 0

if (fewer than (<num-proc> - 1) processes are in delay-queue-1)

delay(<monitor-name>,1)

endif
reset the shared subscript to 1
continue(<monitor-name>,1)
endif

mexit(<monitor-name>)
Once this logic is understood, the logic for barrier synchronization becomes trivial:

\begin{verbatim}
barrier(<monitor-name>, <num-proc>)

menter(<monitor-name>)

if (fewer than (<num-proc> - 1) processes are in
delay-queue-1)

delay(<monitor-name>, 1)

endif

continue(<monitor-name>, 1)

mexit(<monitor-name>)
\end{verbatim}

We do not intend to introduce you to all of the features available in \texttt{m4} for implementing such monitors as macro invocations. However, you will need to understand the essential features of \texttt{m4} in order to work with our more complex monitors (and to code your own monitor operations, when necessary). Therefore, let us introduce you to \texttt{m4} by presenting a macro definition of the \texttt{barrier} operation:

\begin{verbatim}
define(barrier, 
[ menter($1) 
   IF ($1 < 1 .LT. ($2 - 1)) THEN 
       delay($1, 1) 
   ENDIF 
   continue($1, 1) 
   mexit($1)]
)
\end{verbatim}

The \texttt{define} command is used to create new macros. It contains two arguments -- the name of the macro and what should be generated by the macro. In this simple example, "barrier" is the name, and the text surrounded by the square brackets is the replacement text. \texttt{m4} can be used to search for strings of the form "barrier(a1,a2)", which represent invocations of the macro. Whenever such a string is found, the replacement text will replace it. The only complexity involves the arguments of the invocation: wherever \$1 occurs in the replacement text, the first argument of the macro invocation will be used, and similarly for any number of arguments. Thus,
barrier(BA,NPROCS)

would generate

menter(BA)
IF (BAC1 .LT. (NPROCS - 1)) THEN
delay(BA,1)
ENDIF
continue(BA,1)
nexit(BA)

This code would then be expanded to actual FORTRAN using definitions of the basic monitor operations that we supply for different machines.

Before leaving this example, one more comment is required. The variable $1C1 occurs in the macro definition (and gets expanded in our example to BAC1). This variable, by definition, contains the number of processes delayed in the first delay queue for the monitor.

6. An Introduction to the Askfor Monitor

In our studies of implementing algorithms for multiprocessors, we have become increasingly aware of the significance of a single synchronization pattern. It appears with surprising frequency in both numeric and non-numeric applications. The pattern involves the concept of "a pool of tasks", where processes "ask for a task to work on" and "add tasks to the pool". In some sense, this is exactly what a self-scheduling DO-loop does, except that for the DO-loop processes never add new tasks to the pool. In the case of the DO-loop, the pool of problems is represented by a subscript; if the shared subscript is I and the maximum value is N, then the pool of remaining tasks to be parcelled out is the subscript values from the current value of I to the maximum value N. This fairly simple special case, however, is not sufficient for many applications. Frequently, the pool of remaining problems cannot be represented by a single subscript; in fact, the structure required to represent the pool is often quite complex.

Let us first give a more precise description of the general pattern:

1) A sequence of computational tasks (i.e., problems) must be solved. We shall refer to these as the "major" tasks $T_1$, $T_2$, ....

2) Each major task $T_i$ may be decomposed into one or more minor tasks $t_{i1}$, $t_{i2}$, ....

3) A minor task may itself be decomposed.
4) At any point in the computation, the solution of a minor task may result in a solution for the current major task. Thus, the current major task is thought of as "unsolved" until either a subcomputation produces a solution or until all subcomputations are completed. We refer to this latter situation as a solution by exhaustion.

Let us illustrate one instance of this pattern that we found in writing a routine to perform QR-factorization of a sequence of matrices. In this case, each major task amounts to computing the desired factorization. We use Householder's algorithm for performing the factorization. It is not important that you understand what this algorithm does; it is necessary only to understand the following synchronization requirements:

1. The first step in factoring a matrix is to create the reflection for column 1. This reflection can then be applied to all remaining columns. In general, you create the reflection for column k after the previous k-1 columns have all had their reflections created and applied to column k.

2. A reflection for a column k can be applied to all columns to the right of k. However, the reflections must be applied in order. That is, the reflection created from column 1 must be applied to column 3 before the reflection created from column 2 is applied to column 3.

In this case, the pool of remaining tasks includes two types of minor tasks: create a reflection for column i and apply the reflection for column i to column j. Initially, the pool will contain only the minor task "create the reflection for column 1". The solution of the first minor task will cause a number of new minor tasks (to apply the reflection to each of the remaining columns) to be added to the pool. Processes claim minor tasks and add them to the pool until the pool is exhausted, representing completion of the major task. In this particular case, all major tasks are solved by exhaustion (and are solved "successfully"). This is frequently not the case. The challenge involved in setting up the problem is how to represent the "pool of problems" in a reasonably efficient manner. This is a problem-dependent aspect of the synchronization: different problems utilize widely different representations of the problem pool. Thus, special-purpose logic will be needed to "claim a problem from the pool" or to "insert a problem into the pool".

We have found this pattern of synchronization, which we will call the askfor monitor (since processes "ask for" the next available task to perform), to be quite difficult to implement properly. The fact that problem-dependent code is required to manage the pool of outstanding tasks makes the use of our askfor monitor more complex than using the barrier or getsub monitors. We begin, therefore, with the simplest of settings — our first example that adds two vectors...
to produce a third. In this case, since it is a self-scheduling DO-loop, the pool of problems can be managed by a single shared subscript. The code is as follows:

```fortran
1 define(getprob,
2     [IF (SUB .LE. 2) THEN
3         $1 = SUB
4         SUB = SUB + 1
5         $3 = 0
6     ENDIF]
7 )
8
9
10 * THIS PROGRAM DEMONSTRATES THE "ASKFOR" SYNCHRONIZATION PRIMITIVE.
11 * IT READS IN TWO VECTORS, ADDS THEM, AND PRINTS THE RESULT.
12 *
13 ******************************************************************************
14 *
15     PROGRAM ADDTWO
16 *
17     newproc(SLAVE)
18 *
19 * COMMON AREA VARIABLES
20 *
21     INTEGER A(1000), B(1000), C(1000)
22     INTEGER NPROCS, N
23     COMMON /MAINC/ A, B, C, N, NPROCS
24 *
25 ******************************************************************************
26 *
27 * DECLARE THE VARIABLES AND COMMON TO SUPPORT THE MONITOR
28 *
29 ******************************************************************************
30 *
31     adec(MC)
32 *
33     INTEGER SUB
34     COMMON /POOL/ SUB
35 *
36     INTEGER I
37 *
38 * INITIALIZE THE ENVIRONMENT
39 *
40 ******************************************************************************
41 *
initenv

*-----------------------------------------------
* INITIALIZE ASKFOR MONITOR
*-----------------------------------------------
ainit(M0)
SUB = 1

*-----------------------------------------------
* READ IN THE NUMBER OF PROCESSES TO RUN IN PARALLEL
*-----------------------------------------------
READ (5,10) NPROCS
10 FORMAT(I4)
WRITE(6,20) NPROCS
20 FORMAT(' NPROCS = ',I4)

*-----------------------------------------------
* READ IN THE TWO INPUT VECTORS
*-----------------------------------------------
READ (5,10) N
READ (5,10) (A(I), I = 1,N)
READ (5,10) (B(I), I = 1,N)

*-----------------------------------------------
* CREATE THE SLAVE PROCESSES
*-----------------------------------------------
DO 30 I=1,NPROCS-1
create(SLAVE)
30 CONTINUE

*-----------------------------------------------
CALL WORK
*-----------------------------------------------
WRITE (6,10) (C(I), I = 1,N)
STOP
END
THE SLAVE PROCESSES JUST CALL THE WORK SUBROUTINE TO ADD UP ELEMENTS UNTIL THE END OF THE VECTOR IS REACHED. THE PROCESSES THEN EXIT (WHICH IS ASSUMED TO DESTROY THEM).

SUBROUTINE SLAVE

CALL WORK
RETURN
END

SUBROUTINE WORK

* COMMON AREA VARIABLES

INTEGER A(1000), B(1000), C(1000)
INTEGER NPROCS, N
COMMON /MAINC/ A, B, C, N, NPROCS
adec(MD)

INTEGER SUB
COMMON /POOL/ SUB

INTEGER I, RC

10 CONTINUE

CLAIM THE NEXT AVAILABLE SUBSCRIPT (RETURNED IN I)
To understand this program, the reader should compare it to the version using self-scheduling DO-loops. Several points are worth mentioning:

1. First, note that the programmer explicitly gives the shared COMMON used to manage the pool of processes (lines 31-32). He must also initialize his representation of the pool (line 51).

2. The _getsub_ operation has been replaced with an invocation of the _askfor_ monitor operation (line 136):

   ```
   askfor(MO, RC, NPROCS, getprob(I, N, RC),)
   ```

Here, the first argument (MO) is the name of the monitor. The second argument (RC) is a variable set by the monitor operation to indicate whether a problem could be taken from the pool (set to 0 on successful acquisition of a problem; otherwise it will get set to 1). The third argument (NPROCS) gives the number of processes sharing the pool. The fifth argument gives a macro invocation that can be used to generate the problem-dependent logic required to try to claim a problem from the pool. There is a missing sixth argument which is used to generate the code required to reinitialize the pool; since we are only adding one pair of vectors, there is no need to reinitialize the pool.

The intrinsic complexity involved in the _askfor_ monitor is based on the fact that the user has to provide a macro to generate the logic to claim tasks (and the logic to reinitialize the pool, if a sequence of major tasks are to be processed). The logic to claim a problem is generated by the macro _getprob_, which is defined in lines 1-7 of the program. Note that _getprob_ does not generate a monitor operation; rather, it generates a block of code in the middle of the _askfor_ monitor operation. This is why it does not begin with an _merror_ nor end with a _mexit_.

7. A More Complex Example of the Askfor Monitor

Many applications of the askfor monitor are substantially more complex than our last example. Frequently, the complexity originates in the representation of the pool of problems. However, there can be other sources of complexity:

1. Some problems can be "solved" by any arbitrary process working on a minor task. All processes must (on their next request for a task) be notified of "problem end" by a positive value in the return code. For example, suppose that the object of our previous problem were to determine whether the sum of any two corresponding elements in A and B exceeded 100. In this case, a solution by exhaustion (return code of 1) would indicate that no values did exceed 100. On the other hand, any process that found a sum that exceeded 100 could "post" end-of-problem, using the actual sum as the return code (actually any value greater than 1 would be adequate). Posting an end-of-problem condition is accomplished with the probend macro.

2. Some programs must compute solutions to a sequence of major tasks. This requires the inclusion of an extra parameter on the askfor invocation giving a macro that can be used to reinitialize the monitor. Furthermore, each process should receive the end-of-problem return code exactly. Once a process has received such a return code, it can ask for another task (which is what slave processes normally do) or itself go and acquire another major task to solve (which is what the master process would normally do).

3. In the cases where a sequence of major tasks are to be solved, some means must exist for signalling "end of program" to processes. This is done by invoking the progend macro. In this case, every process that has been delayed or requests another task will receive a return code value of -1.

To see exactly how these concepts come into play, let us consider a modified version of our earlier example. In this example, a sequence of pairs of input vectors are processed. For each pair, the question

"Does the sum of any corresponding elements exceed 100?"

is answered. Between problems, the reset logic just resets the shared subscript to 0. The subscript is set to 1 when a new pair of vectors are ready to be processed.

```plaintext
define(getprob,
     IF (SUB .GT. 0) THEN
     IF (SUB .LE. $2) THEN
```
$1 = \text{SUB} \quad \text{SUB} = \text{SUB} + 1 \quad \text{SUB} = 0 \quad \text{ENDIF} \quad \text{ENDIF}$

$3 = 0$

define(reset, [SUB = 0])
define(probstart, [\text{meenter(MO)}] SUB = 1 \quad \text{continue(MO,1)} \quad \text{mexit(MO)}])$

$\text{..................................} \quad \text{....................................}$

"\text{THIS PROGRAM DEMONSTRATES THE "ASKFOR" SYNCHRONIZATION PRIMITIVE.}"
"\text{IT READS IN PAIRS OF VECTORS. FOR EACH PAIR, IT DETERMINES}"
"\text{WHETHER OR NOT THE SUM OF ANY TWO CORRESPONDING ELEMENTS}"
"\text{EXCEEDS 100.}"

"\text{PROGRAM CHKTWO}"

newproc(SLAVE)

\text{COMMON AREA VARIABLES}

INTEGER A(1000), B(1000)\quad \text{INTEGER NPROCS, N}

\text{COMMON /MAIN/ A, B, N, NPROCS}

\text{DECLARE THE VARIABLES AND COMMON TO SUPPORT THE MONITOR}

adec(MO)

INTEGER SUB\quad \text{COMMON /POOL/ SUB}

INTEGER I, J, NPROCS
* INITIALIZE THE ENVIRONMENT

** *initialize

* INITIALIZE ASKFOR MONITOR

```fortran
initenv
```

* READ IN THE NUMBER OF PROCESSES TO RUN IN PARALLEL

```fortran
READ (5,10) NPROCS
```

* CREATE THE SLAVE PROCESSES

```fortran
DO 30 I=1,NPROCS-1
  create(SLAVE)
  CONTINUE
```

* READ IN THE NUMBER OF PAIRS TO PROCESS

```fortran
READ (5,10) NPROBS
```

```fortran
DO 40 J = 1,NPROBS
```
* READ IN THE TWO INPUT VECTORS

*-----------------------------------------------

* READ (5,10) N
READ (5,10) (A(I), I = 1,N)
READ (5,10) (B(I), I = 1,N)

* probstart
CALL WORK(0)

40 CONTINUE
progend(MO)

* SLAVE PROCESSES

*-----------------------------------------------

* SUBROUTINE SLAVE

* CALL WORK(1)
RETURN
END

* WORK SUBROUTINE

*-----------------------------------------------

* SUBROUTINE WORK(WHO)
INTEGER WHO

* COMMON AREA VARIABLES

INTEGER A(1000), B(1000)
INTEGER NPROCS, N
COMMON /MAINC/ A, B, N, NPROCS

* adec(MO)
INTEGER SUB  
COMMON /POOL/ SUB  
INTEGER I, RC, SUM  
10 CONTINUE  

CLAIM THE NEXT AVAILABLE SUBSCRIPT (RETURNED IN I)  

askfor(MO,RC,NPROCS,getprob(I,N,RC),reset)  

IF (RC .NE. 0) GO TO 20  

SUM = A(I) + B(I)  
IF (SUM .GT. 100) THEN  
   probend(MO,2)  
ENDIF  

GO TO 10  

CONTINUE  

IF ((RC .NE. -1) .AND. (WHO .EQ. 1)) GO TO 10  
IF (WHO .EQ. 0) THEN  
   WRITE(6,30) RC  
   WRITE(6,30) RC  
ENDIF  
RETURN  
END  

Note the following points:

1. The `getprob` logic must now check to make sure that there is an active problem. If the logic does not alter RC, then the requesting process will be delayed until `probstart` is executed. Here `probstart` is an actual monitor operation. It issues a `continue` to activate any processes that might be waiting for work to arrive.

2. If a process recognizes an end-of-problem condition, it still must go through the `askfor` requesting a task. Our logic insists on returning the end-of-job return code to each of the cooperating processes (and waits for them all to be delayed, so that it can release them all with the
appropriate return code).

3. The WHO variable is used in the WORK routine to distinguish the master process, which prints the return code and goes to get the next problem.

4. The second operand on the probend monitor operation is the return code to be passed back to each process. It should always be greater than 1.

The reader should note that the logic to support the full askfor monitor is somewhat complex. For a detailed description, you should consult[8].

8. A Simple Grid Computation

The next programming example that we present represents a prototypical grid problem. The essential characteristics of this class of grid problems are as follows:

1. The computation involves a cellular space. That is, a grid exists that divides the space into cells.
2. Each cell has a state characterized by one or more numeric values.
3. A neighborhood function exists that defines the set of neighbors for a given cell.
4. Time is thought of as a discrete sequence $t_0, t_1, t_2, \ldots$. There exists a transition function that defines the state of a cell at time $t_{i+1}$ in terms of the state of the cell and its neighbors at time $t_i$.

In our example, we have chosen a particularly simple case. We have a two-dimensional space in which the state of each cell is represented by a single floating-point value $\phi(x,y)$. We assume that $\phi$ is given on the boundary cells, and we arbitrarily assign 0 to the interior cells. The neighbors of a cell are just the four cells that share the bounding faces of the cell. The transition function is defined as follows:

Cells on the boundary have constant states. For each nonboundary cell, the state at time $t_{i+1}$ is just the average of the states of the neighboring cells at time $t_i$.

This example, while simple, does have some physical significance. As time progresses, the values in the cells converge to a solution of Laplace’s equation $\nabla^2 \phi = 0$, with the original values of $\phi$ on the boundary.

A grid problem of the sort we are considering here obviously has a great deal of exploitable parallelism. We use two grids in memory, where the states of cells are updated from one grid to the other and then back again. This doubles the memory requirement for the problem, but greatly simplifies the
synchronization logic and allows for more parallelism than if the grid were updated in place. We based our implementation on the ASKFOR monitor, where each computational task is to update all cells in one slab of the grid by one time step. A slab is composed of all cells that have the same x-coordinate value. Thus, the logic of the master process for this computation is as follows:

**master process:**

Initialize the monitor to manage the pool of problems

Acquire the number or processes to be utilized

Create the slave processes (which wait for work to be added to the pool of tasks)

Initialize the values in the specified cube and enter the slabs in the grid into the pool of problems

Call **work** to aid in the computation

Print the result

**work:**

Invoke the ASKFOR monitor to acquire a computational task from the pool (i.e., a slab to be updated)

While (a task was acquired)

update the slab (i.e., compute the values for the next time step for the specified slab)

Call **postprob** to update the status of the newly updated slab and to add any "released" slabs to the pool of tasks

Invoke the ASKFOR monitor to acquire another task
postprob: (monitor enter)

Update the status of the slab just processed (adding one to the number of time steps computed for the slab)

If the next slab to the left (i.e., the slab with an x-coordinate of one less than the just processed slab) can now be updated, but could not be updated before the current slab was updated, then add the left slab to the pool of tasks that are ready for computation.

If the next slab to the right can now be updated, but could not be updated before the current slab was updated, then add the right slab to the pool of tasks that are ready for computation.

If the current slab can be updated again, add it to the pool of available tasks.

end (monitor exit)

The code to implement this logic is as follows:

```c
1 define(getprob,
2   [CALL GETPRB($1,$2)]
3 )
4   PROGRAM CUBE
5   newproc(SLAVE)
6 *
7 *
8 * THE FUNCTION OF THIS PROGRAM IS TO APPROXIMATE THE VALUE OF A
9 * FUNCTION 'PHI' SATISFYING BOUNDARY CONDITIONS
10 *
11 * PHI(X,Y,Z) = X * X - Y * Y + X * Y * Z
12 *
13 * FOR (X,Y,Z) ON THE BOUNDARY OF THE CUBE. THE VALUE AT AN INTERIOR
```
* POINT IS APPROXIMATED AS THE AVERAGE VALUE OF THE NEIGHBORING
* POINTS.

* COMMON AREA VARIABLES

* REAL PHI,A(20,20,20),B(20,20,20)
* INTEGER STATUS(20),PQ(21),N,NPROCS,PQBEGIN,PQEND,XDIM,YDIM,ZDIM
* COMMON /POOL/ A,B,STATUS,PQ,N,NPROCS,PQBEGIN,PQEND,XDIM,YDIM,ZDIM

* DECLARE THE VARIABLES AND COMMON TO SUPPORT THE MONITOR

* adec(MO)

* DECLARE THE WORKING VARIABLES

* INTEGER I,J,K,X,Y,Z

* INITIALIZE THE ASKFOR MONITOR POOL. INITIALLY EACH ELEMENT OF THE
* QUEUE REPRESENTS A SLICE (CONSTANT X COORDINATE) OF THE CUBE ON
* WHICH TO APPROXIMATE THE VALUE OF THE FUNCTION PHI.

* INITIALIZE THE ENVIRONMENT

* initenv

* GET THE DIMENSIONS OF THE CUBE

* READ (5,20) XDIM,YDIM,ZDIM

* GET THE NUMBER OF ITERATIONS TO PERFORM

* READ (5,80) N

* READ (5,100) NPROCS

* INITIALIZE THE INTERIOR OF THE CUBE TO ZERO

* DO 110 I=2,XDIM-1
DO 120 J=2,YDIM-1
  DO 130 K=2,ZDIM-1
  A(I,J,K) = 0
  CONTINUE
120 CONTINUE
110 CONTINUE

* INITIALIZE THE BOUNDARY OF THE CUBE
*
* THE FACES X = 1 AND X = XDIM
* DO 140 J=1,YDIM
  DO 150 K=1,ZDIM
    A(1,J,K) = PHI(1,J,K)
    B(1,J,K) = A(1,J,K)
    A(XDIM,J,K) = PHI(XDIM,J,K)
    B(XDIM,J,K) = A(XDIM,J,K)
  CONTINUE
140 CONTINUE

* THE FACES Y = 1 AND Y = YDIM
* DO 160 I=1,XDIM
  DO 170 K=1,ZDIM
    A(I,1,K) = PHI(I,1,K)
    B(I,1,K) = A(I,1,K)
    A(I,YDIM,K) = PHI(I,YDIM,K)
    B(I,YDIM,K) = A(I,YDIM,K)
  CONTINUE
160 CONTINUE

* THE FACES Z = 1 AND Z = ZDIM
* DO 180 I=1,XDIM
  DO 190 J=1,YDIM
    A(I,J,1) = PHI(I,J,1)
    B(I,J,1) = A(I,J,1)
    A(I,J,ZDIM) = PHI(I,J,ZDIM)
    B(I,J,ZDIM) = A(I,J,ZDIM)
  CONTINUE
180 CONTINUE
* INITIALIZE THE PROBLEM QUEUE
DO 200 I=2,xdim-1
   PQ(I-1) = I
200 CONTINUE

* PQBEGIN = 1
* PQEND = XDIM-1

* INITIALIZE THE STATUS VECTOR
DO 210 I = 1,xdim
   STATUS(I) = 0
210 CONTINUE

* INITIALIZE THE MONITOR
ainit(MO)

* CREATE THE SLAVE PROCESSES
DO 220 I = 1,nprocs-1
   create(SLAVE)
220 CONTINUE
CALL WORK
progend(MO)

IF (MOD(N,2) .EQ. 0) THEN
   CALL PRCUBE(A)
ELSE
   CALL PRCUBE(B)
ENDIF
STOP
END

*************

PHI FUNCTION
FUNCTION PHI(X,Y,Z)
    INTEGER X,Y,Z
    PHI = (X * X) - (Y * Y) + (Z * Z)
    PHI = 1
RETURN
END

SUBROUTINE SLAVE
CALL WORK
RETURN
END

SUBROUTINE WORK

COMMON /POOL/ A,B,STATUS,PQ,N,NPROCS,PQBEQ,PQEND,XDIM,YDIM,ZDIM

COMMON /POOL/ X

INTEGER RC,X
* DECLARE THE VARIABLES AND COMMON TO SUPPORT THE MONITOR

      ado (MO)
* CONTINUE
10 ask for (MO, RC, NPROCS, getprob(X, RC),)
     WRITE(6,872) X, RC
*     FORMAT(' X = ',I4, ' RC = ',I4)
*     IF (RC .NE. 0) GO TO 100
*     IF (MOD(STATUS(X),2) .EQ. 0) THEN
*         CALL COMP(A,B,X)
*     ELSE
*         CALL COMP(B,A,X)
*     ENDIF
*     CALL POSTPB(X)
  GO TO 10
* 100 CONTINUE
  RETURN
  END

* COMPUTE SUBROUTINE

SUBROUTINE COMP(P, Q, X)

REAL P(20,20,20), Q(20,20,20)
INTEGER X

* COMMON AREA VARIABLES

REAL A(20,20,20), B(20,20,20)
INTEGER STATUS(20), PQ(21), N, NPROCS, PQBEGIN, PQEND, XDIM, YDIM, ZDIM
* COMMON /POOL/ A, B, STATUS, PQ, N, NPROCS, PQBEGIN, PQEND, XDIM, YDIM, ZDIM
*
INTEGER I, J, K

* WRITE(6,873) X

873 FORMAT( ' ENTERING AREA: X = ', I4)

DO 10 J = 2, YDIM - 1
   DO 20 K = 2, ZDIM - 1
      Q(X, J, K) = (P(X - 1, J, K) + P(X + 1, J, K) +
                     P(X, J - 1, K) + P(X, J + 1, K) +
                     P(X, J, K - 1) + P(X, J, K + 1)) / 6.0
   CONTINUE
10 CONTINUE

RETURN
END

WRITE(6,873) X
FORMAT(' ENTERING AREA: X = ', I4)

DO 10 J = 2, YDIM - 1
   DO 20 K = 2, ZDIM - 1
      Q(X, J, K) = (P(X - 1, J, K) + P(X + 1, J, K) +
                     P(X, J - 1, K) + P(X, J + 1, K) +
                     P(X, J, K - 1) + P(X, J, K + 1)) / 6.0
   CONTINUE
10 CONTINUE

RETURN
END
SUBROUTINE QPROB

INTEGER X

* COMMON AREA VARIABLES

REAL A(20,20,20), B(20,20,20)
INTEGER STATUS(20), PQ(21), NPQ, PQEND, XDIM, YDIM, ZDIM
COMMON /POOL/ A, B, STATUS, PQ, N, NPQ, PQEND, XDIM, YDIM, ZDIM
COWAON /POOL/ A, B, STATUS, PQ, N, NPQ, PQEND, XDIM, YDIM, ZDIM

* DECLARE THE VARIABLES AND COMMON TO SUPPORT THE MONITOR

adec(MO)

PQ(PQEND) = X
PQEND = PQEND + 1
IF (PQEND .GT. 21) PQEND = 1
RETURN
END

SUBROUTINE POSTPROB

SUBROUTINE POSTPB(X)

INTEGER X

* COMMON AREA VARIABLES

REAL A(20,20,20), B(20,20,20)
INTEGER STATUS(20), PQ(21), N, NPRCS, PQBEG, PQEND, XDIM, YDIM, ZDIM
COMMON /POOL/ A, B, STATUS, PQ, N, NPRCS, PQBEG, PQEND, XDIM, YDIM, ZDIM

* DECLARE THE VARIABLES AND COMMON TO SUPPORT THE MONITOR

adeC(MO)

implement(MO)

STATUS(X) = STATUS(X) + 1
IF (X .EQ. 2) STATUS(1) = STATUS(X)
IF (X .EQ. XDIM-1) STATUS(XDIM) = STATUS(X)

IF (STATUS(X) .LT. N) THEN
  IF ((X .GT. 2) .AND.
    - (STATUS(X-2) .GE. STATUS(X)) .AND.
    - (STATUS(X) .EQ. STATUS(X-1))) THEN
    CALL QPROB(X-1)
  ENDIF
  IF ((X .LT. XDIM-1) .AND.
    - (STATUS(X+1) .EQ. STATUS(X)) .AND.
    - (STATUS(X) .LE. STATUS(X+2))) THEN
    CALL QPROB(X+1)
  ENDIF
  IF (((STATUS(X-1) .EQ. STATUS(X)) .AND.
    - (STATUS(X) .EQ. STATUS(X+1))) THEN
    CALL QPROB(X)
  ENDIF
ENDIF
exit(MO)
RETURN
END
9. Locks

Before ending this discussion of basic synchronization primitives, some discussion of "locks" should be included. Consider the problem of computing a dot product of two vectors A and B, storing the result in the integer I. If a self-scheduling DO-loop is used, processes are allocated distinct subscripts. However, the problem remains of exactly how to synchronize the addition of values into the integer I. The basic idea behind the use of monitors would suggest that a monitor be established for the variable I, with the single operation of adding a value to the variable. This is a perfectly reasonable way to handle the problem.

However, in some problems it becomes necessary to selectively gain access to a variety of data items and then to update them in fairly complex patterns. This can be done most easily using "locks". A lock amounts to a "monitor enter"
operation, while an "unlock" is the corresponding "monitor exit" operation. When utilizing locks, there is no notion of delaying the process. The process simply gains control of the lock, performs the critical section, and releases the lock.

To allow this style of programming, we have included four macros in our standard package:

- **lockdec(N)** declares the variables required to support N locks. These locks are based on an array (named LO) of monitors.
- **lockinit(N)** initializes the N locks.
- **lock(J)** acquires the Jth lock. This amounts to an `mnenter(LO, J)` operation, which acquires entry to the Jth monitor of the array named LO.
- **unlock(J)** releases the Jth lock.

To update the shared variable I, one would then use

```plaintext
lock(1)
I = I + some value
unlock(1)
```

where the single lock would be declared with

```plaintext
lockdef(1)
```

and initialization would be accomplished with

```plaintext
lockinit(1)
```

In those cases where a number of locks must be held before updating a set of values, the programmer should always acquire the locks in some fixed order (e.g., increasing order of lock number). This technique is called a standard allocation pattern and will guarantee that a deadlock will not occur between two processes attempting to claim overlapping sets of locks. Thus, if process 1 requires locks 1 and 2, while process 2 requires locks 2, 3, and 1, a deadlock could occur unless the fixed order of locking is observed (forcing process 2 to claim lock 1 before lock 2, and lock 2 before lock 3).

10. The Fetch-and-Add

There has been a great deal of interest in the use of the fetch-and-add as a primitive operation in a multiprocessing environment[2]. For many applications, it can be used to eliminate critical sections. It seems quite likely that, as
people contemplate the potential advantages of machines with thousands of processors, the significance of the fetch-and-add operation will turn out to be substantial.

To allow researchers to develop codes for machines that support fetch-and-add on multiprocessors that do not (or to allow portability from systems that do to those that do not), we have written a set of macros that allow one to code using the fetch-and-add. On systems that do not support the operation as a primitive, the macros generate the code to support it as a monitor operation (with the implied overhead associated with the critical section).

The macros to support fetch-and-add are as follows:

\[ \text{fadec}(<\text{monitor-name}>, <\text{type}>) \] is used to generate the variables and COMMON area required to support the monitor. Here <\text{type}> should be INTEGER or REAL, giving the type associated with the value maintained by the monitor.

\[ \text{fainit}(<\text{monitor-name}>) \] initializes the monitor, setting the associated value to 0.

\[ \text{fadd}(<\text{monitor-name}>, <\text{old}>, <\text{new}>, <\text{incr}>) \] returns the "old" value in <\text{old}>, increments it with <\text{incr}>, and returns the "new" value in <\text{new}>. The value associated with the monitor takes on the "new" value.

Thus, the macros might be used as follows:

\[ \text{fadec}(F1, \text{INTEGER}) \]
\[ \text{fainit}(F1) \]
\[ \text{fadd}(F1, \text{OLDI}, \text{NEWI}, J) \]

Here the integer value associated with F1 is incremented by the fetch-and-add. The original value is returned in OLDI, and the new value in NEWI.

This method really does not convey the full power of fetch-and-add, since the operation is designed to be applicable on any word in memory. If you think of its use in manipulating linked lists, then the above version of the macros is really quite inadequate. To somewhat improve the situation, we allow you to declare and manipulate an array of monitors, rather than a single monitor. When this is done, the subscript of the desired monitor (which can be thought of as a "cell" address) is used as the last argument. For example,
fadec(F1, INTEGER, 100)
fainint(F1, 100)

fadd(F1, OLDI, NEWI, J, CELL)

declares and utilizes 100 monitors. Here the \textit{fadd} includes an extra argument, which is the subscript of the monitor for which the operation applies.

11. Running the Package

To prepare a program for execution on some particular machine, you must expand the macro definitions into versions for the desired machine. Typically, you would keep one file of macro definitions for all monitors. This file would not include the basic macro package (which includes \textit{menter} and \textit{mexit}). The basic macro package is machine-specific. User-defined macros specific to a given program are normally included at the front of the program. Following these conventions,

\texttt{m4 <basic-macro-library> <standard-monitors> <user-source> >! <dest>}

can be used to create the desired source in \texttt{<dest>}. We can supply versions of the \texttt{<basic-macro-library>} for both UNIX uniprocessors (for debugging on a uniprocessor system) and IIEP FORTRAN. The \texttt{<standard-monitors>} code at this point includes the monitors that we have discussed in this document, as well as a number of other standard monitors (which we hope to discuss in future reports).

12. Conclusion

This tutorial focuses on the use of three of our macro packages -- those for the self-scheduling DO-loop, barrier, and askfor monitors. We have not discussed all features of these macros, but we have covered their most common uses. We believe, however, that the monitors have wide applicability in more demanding contexts.

If you have any comments or suggestions, we would appreciate hearing from you.
References


Distribution for ANL-84-51 Rev. 1

Internal:

K. L. Kliewer
A. B. Krisciunas
E. L. Lusk (200)
P. C. Messina
R. A. Overbeek (65)
D. M. Pahis
T. M. Woods (2)
G. W. Pieper

ANL Patent Department
ANL Contract File
ANL Libraries
TIS Files (6)

External:

DOE-TIC, for distribution per UC-32 (173)
Manager, Chicago Operations Office, DOE
Mathematics and Computer Science Division Review Committee:
  J. L. Bona, U. Chicago
  T. L. Brown, U. of Illinois, Urbana
  S. Gerhart, Wang Institute, Tynsboro, MA
  G. H. Golub, Stanford U.
  W. C. Lynch, Xerox Corp., Palo Alto
  J. A. Nohel, U. of Wisconsin, Madison
  M. F. Wheeler, Rice U.
D. Austin, ER-DOE
J. Greenberg, ER-DOE
G. Michael, LLL
C. E. Oliver, ER-DOE