TECHNICAL MEMORANDUM

65

The ECMWF grid point model simulator

D. Dent, J.K. Gibson and N. Storer

Research Department

August 1982

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

CONTENTS

1. INTRODUCTION

This technical memorandum describes a simulation program which is used to study the CPU utilisation and input/output processes of the ECMWF gridpoint forecast model. It enables various input/output configurations to be investigated with a minimum of coding effort and is capable of reproducing accurately the performance of the operational forecast model, executing on the CRAY-1.

2. THE PURPOSE OF A SIMULATOR

To be of real value, a simulator must:

a. be capable of simulating a given task or tasks accurately with respect to the resources simulated,

b. be flexible and adaptable to a much greater degree than the task or tasks simulated.

If the above criteria are satisfied, the resulting simulator may be used to study the performance of the simulated tasks, both in their current configuration, and, by adapting the simulator, in alternative configurations. The adaptation of the simulator would normally be a trivial task compared with the alteration of the simulated process.

In the specific case of the forecast model simulator, the processes simulated are CPU time and input/output. Results obtained with this simulator have shown that the forecast model's behaviour can be simulated with a high degree of accuracy. Input/output configurations have been studied with the aid of this simulator, and optimal configurations identified. Such studies have involved minimal changes to the simulator code, but have reflected changes that would have involved major coding changes to the forecast model. Where optimal input/output configurations have been identified by the simulator, incorporation of the appropriate changes into the forecast model have confirmed the accuracy of the simulated results.

3. THE ECMWF GRID POINT MODEL

3.1 The ECMWF grid point model contains a semi-implicit adiabatic scheme, employing a leapfrog time integration with a time filter ensuring added stability; a parameterisation scheme for the sub-grid scale processes, and a radiation scheme. An operational resolution of 192 points per latitude line, 48 lines of latitude between equator and poles, and 15 levels in the vertical is employed. Problems associated with the diminishing spatial separation of grid points as the poles are approached are resolved by applying a diffusion scheme which also acts as a space filter.

3.2 The leap-frog time integration leads to a requirement to have data available in four time-related states:

a. t - 1 values
b. t values, unfiltered
c. t values, partly filtered
d. t + 1 values

This leads naturally to the work-file structure illustrated in Figure 1.

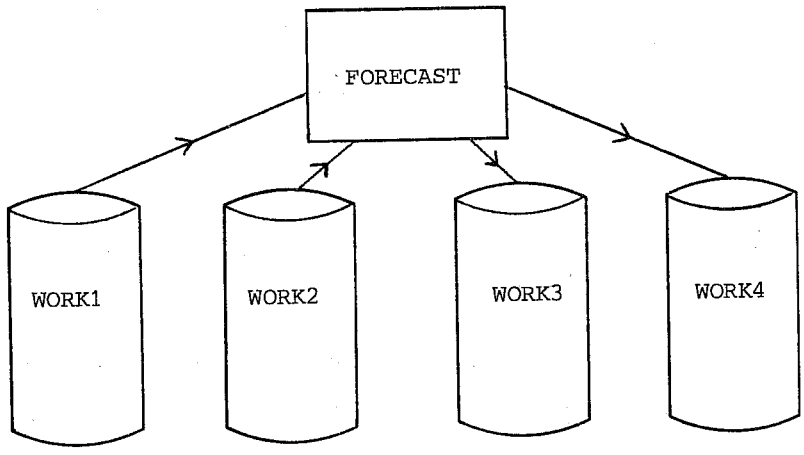


Fig. 1 Forecast Work Files

The 4 main work files contain:

WORK1	-	t - 1 values (input)	
WORK2	-	t values (input)	
WORK3	-	t values (filtered)	(output)
WORK4	-	t + 1 values	(output)

3.3 Each model time step requires:

a. a North to South scan of the work files, during which dynamical changes are computed, and changes computed by the implicit computation, the parameterisation scheme, and the radiation scheme are applied.

b. The solution of a set of Helmholtz equations at the completion of the data scan, involving a transformation to Fourier space, Gaussian elimination, and a reverse transform from Fourier space back to grid point space.

c. On completion, the exchange of the work files so that the output files from the current step become the input files for the next step, and vice versa.

3.4 Radiation time steps are taken every 48th model time step in the current operational forecast. A radiation time step involves the computation of the radiation increments to be added each subsequent model time step, requiring an additional scan from North to South of the work files. In practice, this involves a considerable extra CPU overhead each radiation step.

3.5 Post-processing steps are required at points in the forecast where results are to be written. A post-processing step involves, in addition to the requirements of a normal model time step:

a. the writing of 2 restart files, equivalent in size and similar in content to the input work files WORK1 and WORK2;

b. the writing of 4 work files to be used as input by a subsequent post-processing job;

c. extra computation required to produce the data written to the post-processing work files (e.g. σ to pressure transformations, etc.).

3.6 A large amount of memory is reserved within the model code to avoid the need to use an input/output scheme for the Helmholtz solution (3.3 b. above). Various input/output schemes are available to enable this reserved memory to be released. Such schemes involve a forward and reverse pass through the data, and are usually avoided where possible due to the adverse consequences such "backward" processing inflict on computer efficiency.

4. THE GRID POINT MODEL SIMULATOR

4.1 Simulator capabilities

The grid point model simulator:

a. Can be loaded with a specified proportion of the forecast model CPU loading (i.e. 0 CPU loading, 100% normal, 200% normal, etc.).

b. Writes dummy model work files.

c. Simulates the adiabatic part of a normal model time step, writing 2 work files, and reading 2 work files line by line, with a CPU loading similarly located to the CPU loading of the forecast.

d. Simulates the semi-implicit Helmholtz equation solution part of the normal model time step. Options are available to perform simulated input/output, or model the "in core" approach used in the operational model.

e. Simulates the additional input/output and CPU loading necessary to perform a forecast radiation step.

f. Writes dummy post processor work files for the post processor simulator (see h.) to read.

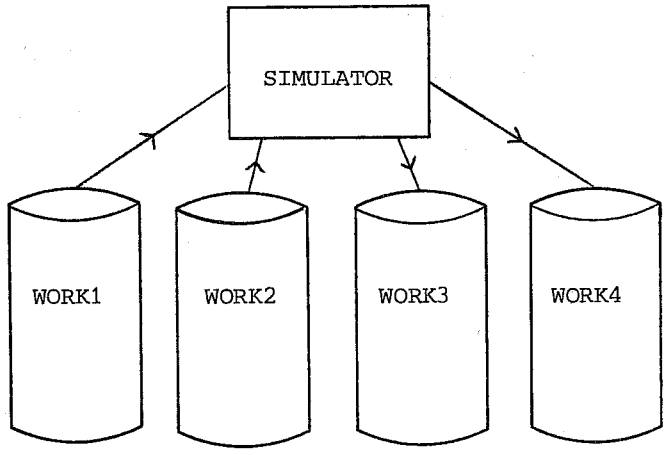
g. Writes dummy restart files.

h. Simulates the post processor job by reading the work files and providing the appropriate CPU loading.

i. Allows various alternative forms of forecast work files:

i. the current 4 work file scheme.

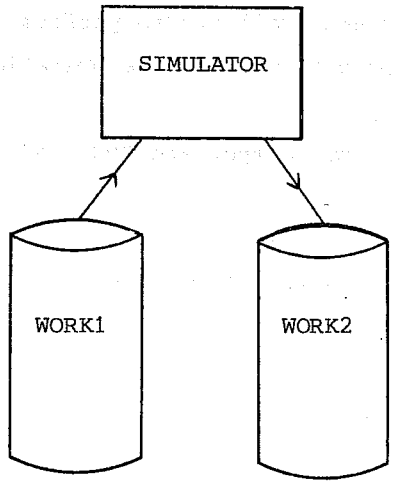
Figure 2 below illustrates the input/output configuration. During alternate time steps data is read from WORK1 and WORK2, and written to WORK3 and WORK4.



To conform to current operational requirements, WORK1,2,3,4 are each 97 records of length 17408 words (34 blocks of 512 words)

Fig. 2

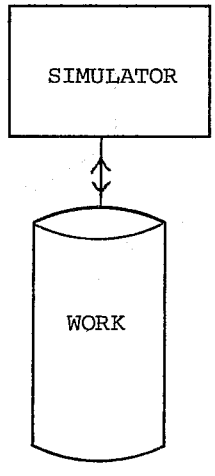
ii. 2 work files, one input, one output, with double length records. This results in longer transfer times per record, but fewer calls to I/O routines. Thus some system I/O overheads are reduced.



WORK1 and WORK2 have 97 records, each of length 34816 words.

Fig. 3

iii. 1 work file, with double length records, read and written using random access techniques.



WORK has 97 records, each of length 34816 words.

Fig. 4

4.2 Simulator interface

The simulator is initiated by means of a procedure called MSIM. All of the options available are selected by means of keyword parameters as follows:

```
MSIM(
        TYPE    =MODEL/PPROC
       ,SCHEME  =INCORE/I-O
       ,NWORK   =4/2/1
       ,CPULOAD=100
       ,STEPS   =3
       ,WKLEN   =17408
       ,SILEN   =2910
       ,RUNPP   =0
       ,PROCESS=ADIA:SEMI:RAD:PPROC
       ,TIME    =0    :0    :0   :0
       )
```

PARAMETERS (ALL OF WHICH ARE KEYWORD)

NAME	DESCRIPTION
TYPE	'MODEL' - means that the system to be simulated is the grid-point model.
	'PPROC' - means that the system to be simulated is the post-processing job.
	(DEFAULT: OMITTED IMPLIES TYPE=MODEL)
SCHEME	'INCORE'- means that the semi-implicit solution to the Helmholtz equations is performed without having to resort to extra I/O.
	'I-O' - means that the semi-implicit solution to the Helmholtz equations is performed using an I/O scheme, which involves forward and reverse scanning of datasets.
	(DEFAULT: OMITTED IMPLIES TYPE=INCORE)
NWORK	'4' - means that the normal I/O scheme, using 4 workfiles is to be used.
	'2' - means that an I/O scheme, using only 2 workfiles, each with double-length records is to be used.
	'1' - means that an I/O scheme, using only 1 workfile, with double-length records, which are read/written randomly, is to be used.
	(DEFAULT: OMITTED IMPLIES NWORK=4)
CPULOAD	The value is a number which represents the percentage of model CPU that is to be used in the simulation. Zero is a valid value.
	(DEFAULT: OMITTED IMPLIES CPULOAD=100)
STEPS	The value is a number which represents the number of timesteps over which the simulation is to be performed.
	(DEFAULT: OMITTED IMPLIES STEPS=3)
	(" 'STEPS' " STEPS=961) (10-day forecast +1)

WKLEN — The value is a number which represents the length of the records for the workfiles.

(DEFAULT: OMITTED IMPLIES WKLEN=17408)

SILEN — The value is a number which represents the length of the records for the semi-implicit files, used if 'SCHEME=I-O).

(DEFAULT: OMITTED IMPLIES SILEN=2910)

RUNPP — The meaning of this keyword depends upon whether 'TYPE=PPROC' or 'TYPE=MODEL'.
For 'TYPE=PPROC':
The value is a number which represents the amount of CPU-time that the post-processing job is to simulate. The units are 1/100ths. of a second. So 100 means 1 second. A value of zero means that there will be no simulated CPU time.

For 'TYPE=MODEL':
The value is a number which determines whether or not the post-processing job is to be disposed to the input queue. If the value is non-zero then the post-processing job will be submitted. A zero value will inhibit the running of the post-processing job.

(DEFAULT: OMITTED IMPLIES RUNPP=0)
(" 'RUNPP' " RUNPP=2300)

PROCESS — The values are character strings which indicate which parts of the model are to be simulated.

'ADIA' - means that the basic adabatic part of the model is to be simulated.
'SEMI' - means that the semi-implicit part of the model is to be simulated.
'RAD' - means that the radiation part of the model is to be simulated.
'PPROC' - means that the post-processing part of the model is to be simulated. (The post-processing job itself is not spun-off if 'RUNPP=0).
'ALL' - means that all of these parts of the model are to be simulated.

(DEFAULT: OMITTED IMPLIES PROCESS=ALL)
(I.E. PROCESS=ADIA:SEMI:RAD:PPROC)

TIME — The values represent the amount of CPU-time (in 1/100ths of a second) which is to be simulated for the various parts of the model. The 1 to 4 values have a 1:1 correspondence to the values of the 'PROCESS' parameter.
0 has a special meaning.
Instead of actually meaning 0 CPU-time, it indicates that a default value is to be used. That default value will have been based upon timings taken from an actual grid-point model run.

(DEFAULT: OMITTED IMPLIES TIME=0:0:0:0)

EXAMPLE

```
JOB(JN=XYZMDL,US=ABC,T=1000,M) ** MODEL SIMULATION **
ACCOUNT(  ...ETC...  )
ACCESS(DN=MSIM,ID=ECMWF)
ASSIGN(DN=FT10,U,DV=DD-19-20)
ASSIGN(DN=FT11,U,DV=DD-19-32)      --  MODEL WORKFILES
ASSIGN(DN=FT12,U,DV=DD-19-40)      --  (UNBLOCKED I/O)
ASSIGN(DN=FT13,U,DV=DD-19-52)
*.
ASSIGN(DN=FT24  ,DV=DD-19-53)
ASSIGN(DN=FT25  ,DV=DD-19-41)      --  SEMI-IMPLICIT FILES
ASSIGN(DN=FT26  ,DV=DD-19-33)      --  USED IF 'SCHEME=I-0'
ASSIGN(DN=FT27  ,DV=DD-19-21)
*.
MSIM(TYPE=MODEL,PROCESS=ALL,RUNPP=1)
/EOF
JOB(JN=XYZPPJ,US=ABC,T=30  ,M) **PPROC SIMULATION**
ACCOUNT(  ...ETC...  )
ACCESS(DN=MSIM,ID=ECMWF)
ACCESS(DN=FT30,PDN=SY7PPROC30,ID=ECMWF,UQ)
ACCESS(DN=FT31,PDN=SY7PPROC31,ID=ECMWF,UQ)    -- CREATED BY 'TYPE=MODEL'
ACCESS(DN=FT32,PDN=SY7PPROC32,ID=ECMWF,UQ)    -- SIMULATION JOB.
ACCESS(DN=FT33,PDN=SY7PPROC33,ID=ECMWF,UQ)
ASSIGN(DN=FT30,BS=22)
ASSIGN(DN=FT31,BS=22)
ASSIGN(DN=FT32,BS=22)
ASSIGN(DN=FT33,BS=22)
*.
MSIM(TYPE=PPROC,RUNPP)
*.
DELETE(DN=FT30)
DELETE(DN=FT31)
DELETE(DN=FT32)
DELETE(DN=FT33)

                                   END OF JOB
```

4.3 Output

Printed output from a simulator run is limited to a summary of the parameters selected. Timing information is extracted from the job logfile using the wall clock time (printed to the nearest second) and the accumulated CPU time (printed to tenths of milliseconds). Entries are made into the logfile when each time step is complete so that it is easy to examine the timing characteristics of the execution.

4.4 Comparison of simulator v forecast

The figures presented in this section were produced from data accumulated by the standard CRAY System Performance Monitor. This is part of the operating system which measures various parameters (in this case every minute), and stores them for later analysis. In each of the figures, the upper graph was produced by an actual operational forecast, run on Wednesday 25 August 1982 between 00.30 and 01.01, the lower graph was produced by a simulation, run on Tuesday 31 August 1982 between 06.30 and 07.01. As can be seen, the 2 sets of graphs are very similar both in overall shape

and also quantitively.

On these graphs, radiation timesteps occur at 36, 47, 57 minutes. Post-processing timesteps occur at 37, 48, 58 minutes. Post-processing jobs occur at 40,50,60 minutes.

For the "CPU UTILISATION" graph:

EXEC + TASKS - are the operating system

USERS - are the model and post-processing jobs

BLOCKED - is the amount of time spent waiting for I/O to complete.

IDLE - is the amount of idle time.

The increased I/O blocked time which occurs after the post-processing jobs have completed (i.e. at 41 & 52 minutes) on the operational forecast run, are caused by verification/diagnostic jobs, which are not simulated.

For the "CENTRAL MEMORY USAGE" graph:

CPU EXEC - is the memory integral for the execution phase of the model and post-processor.

CPU WAIT - is the memory integral when either the model or post-processor could use the CPU but the other job is already using it.

I/O WAIT - is the memory integral when no job is executing, waiting for I/O to complete.

UNUSED - is the free-memory integral.

For the "CPU UTILISATION BY TASK" graph:

SCP - is the Station Call Processor, which handles the links to the front-ends.

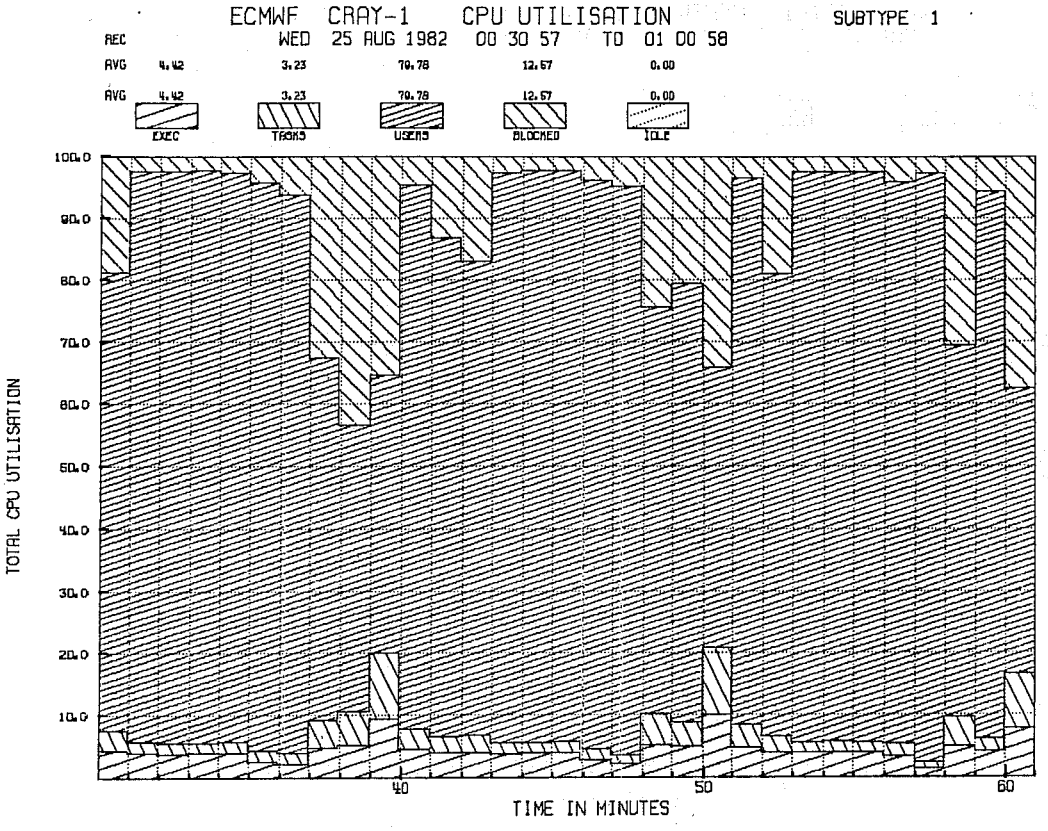
UEP - is the User Exchange Processor, which handles user calls to the operating system.

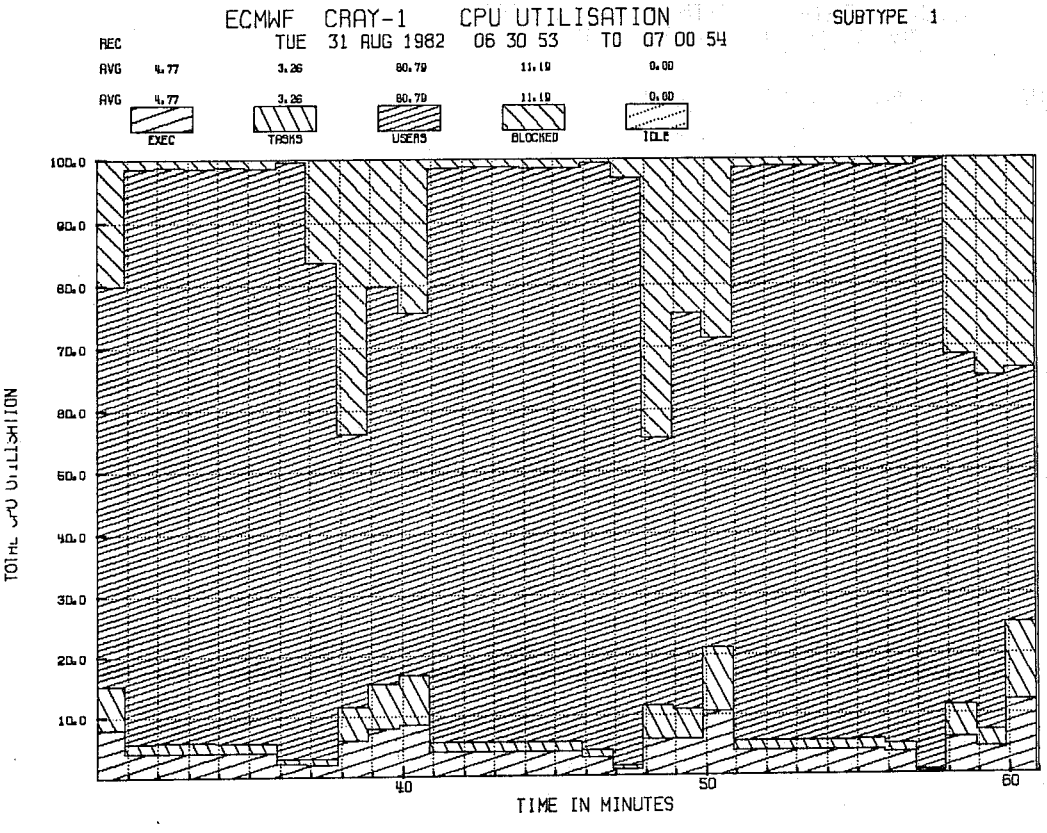
DQM - is the Disk Queue Manager, which performs the I/O queue handling.

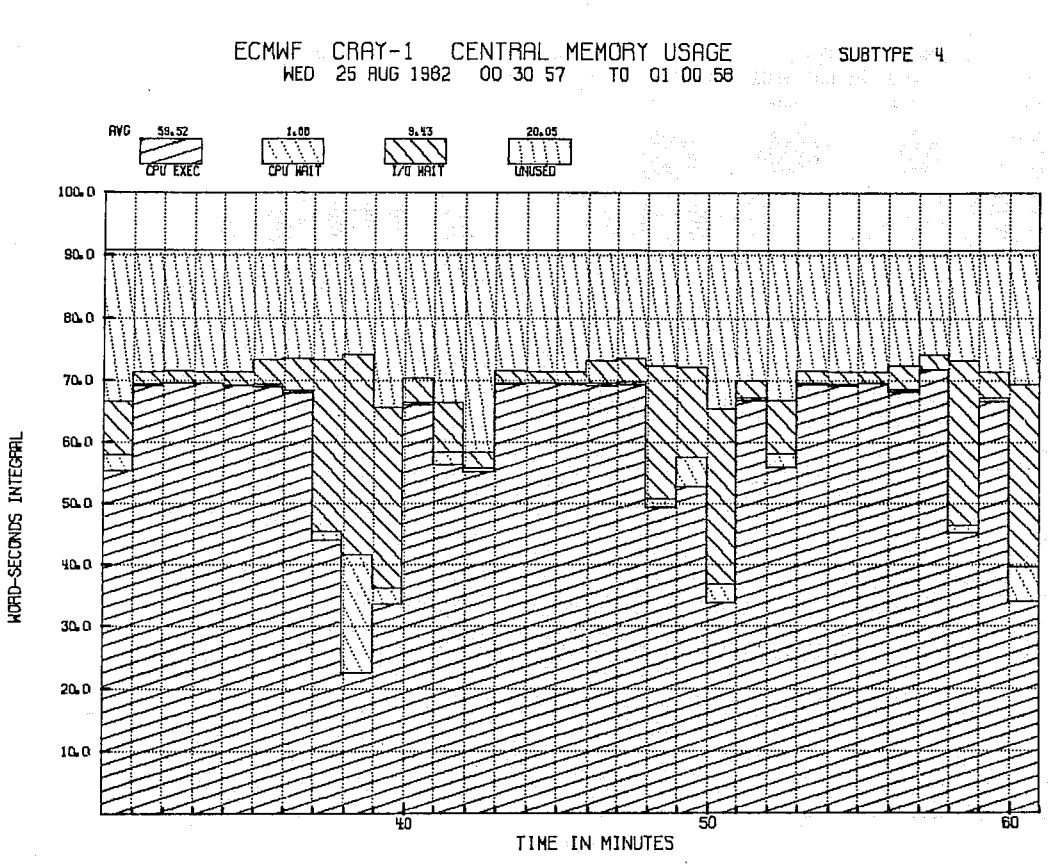
JSH - is the Job Scheduler.

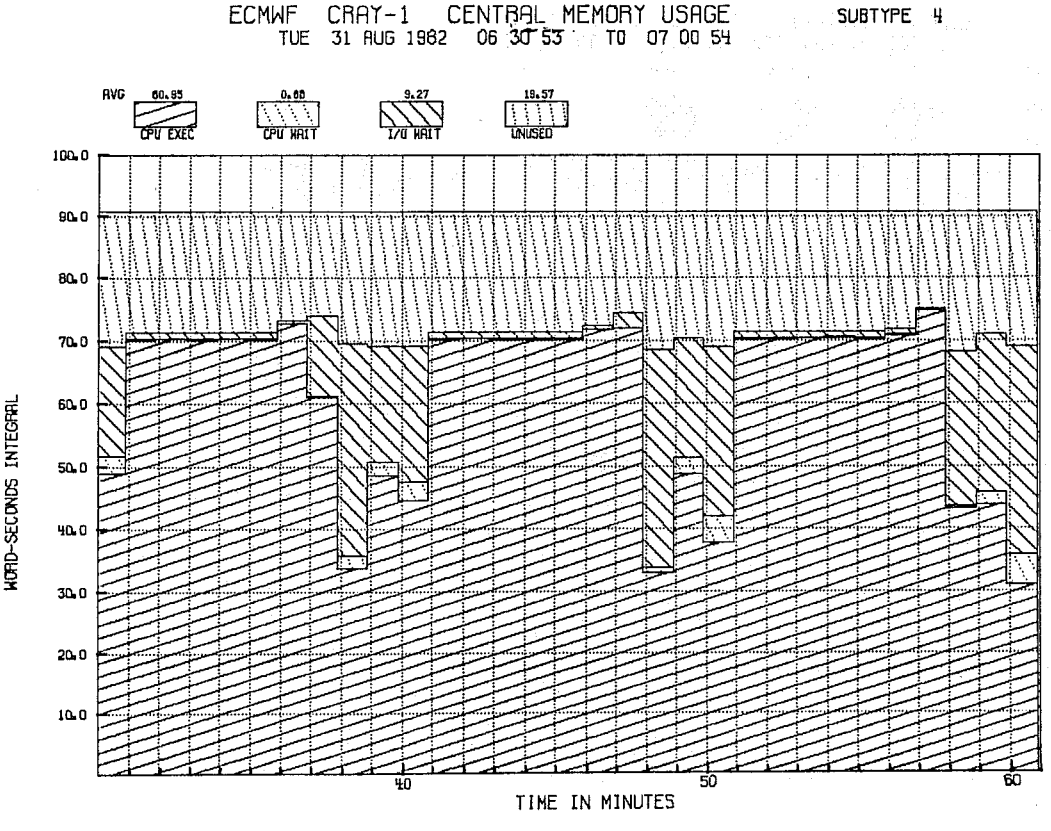
For the "DISK CHANNEL USAGE" graph:

The usage on channel 5 is 3% lower for this run of the simulator because some datasets which were placed on drives on this channel during the operational run have been spread over the other 3 channels.

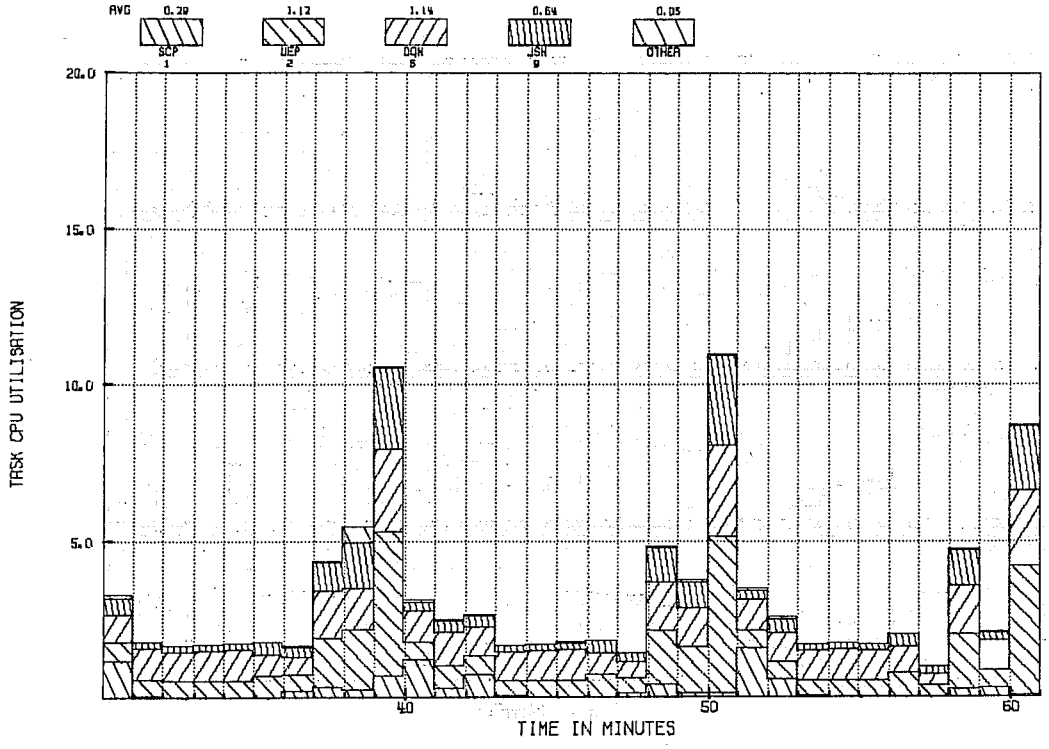

ECMWF CRAY-1 CPU UTILISATION SUBTYPE 1
REC WED 25 AUG 1982 00 30 57 TO 01 00 58
AVG 4.42 3.23 79.78 12.57 0.00
AVG 4.42 3.23 79.78 12.57 0.00
EXEC
TASKS
USERS
BLOCKED
IDLE
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
TOTAL CPU UTILISATION
40
50
60
TIME IN MINUTES


ECMWF CRAY-1 CPU UTILISATION SUBTYPE 1
REC TUE 31 AUG 1982 06 30 53 TO 07 00 54
AVG 4.77 3.26 80.79 11.19 0.00
AVG 4.77 3.26 80.79 11.19 0.00
EXEC
TASKS
USERS
BLOCKED
IDLE
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
40
50
60
TIME IN MINUTES

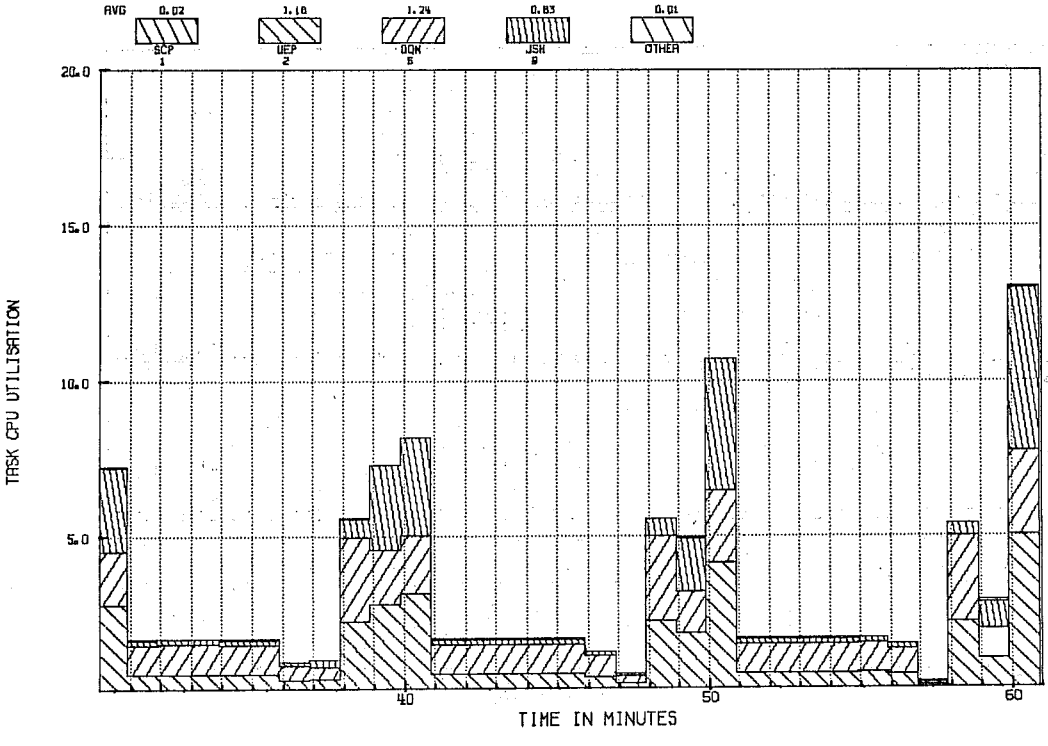

ECMWF CRAY-1 CENTRAL MEMORY USAGE
WED 25 AUG 1982 00 30 57 TO 01 00 58
SUBTYPE 4
AVG
59.52
CPU EXEC
1.00
CPU WAIT
9.43
I/O WAIT
20.05
UNUSED
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
WORD-SECONDS INTEGRAL
40
50
60
TIME IN MINUTES


ECMWF CRAY-1 CENTRAL MEMORY USAGE
TUE 31 AUG 1982 06 30 53 TO 07 00 54
SUBTYPE 4
AVG
60.85
CPU EXEC
0.88
CPU WAIT
9.27
I/O WAIT
19.57
UNUSED
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
WORD-SECONDS INTEGRAL
40
50
60
TIME IN MINUTES

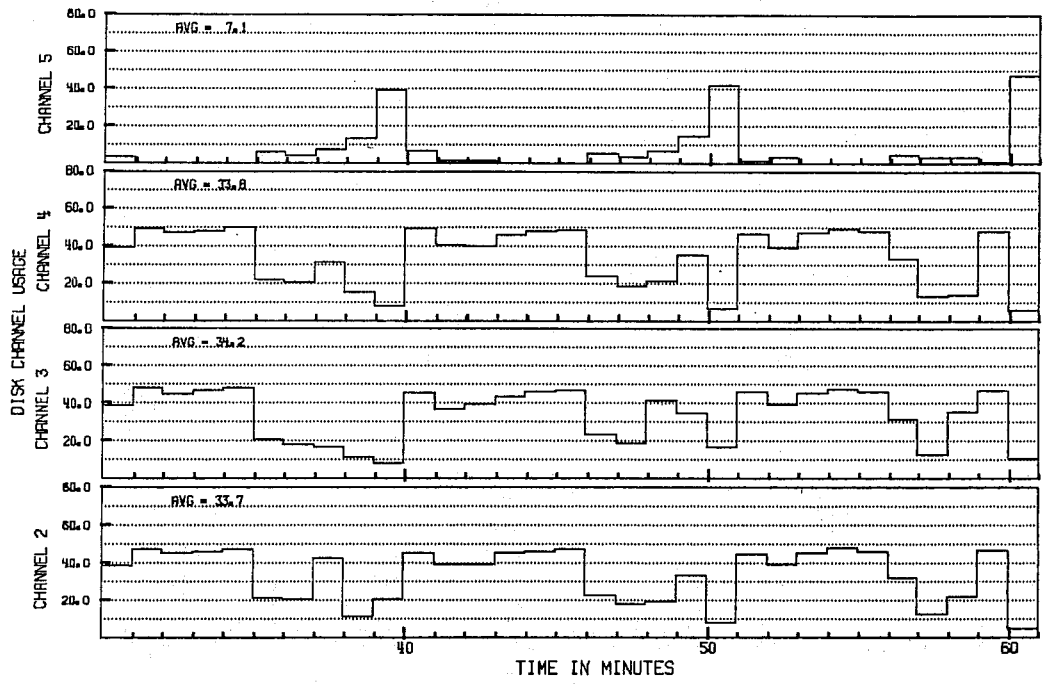
ECMWF CRAY-1 CPU UTILISATION BY TASK SUBTYPE 1
WED 25 AUG 1982 00 30 57 TO 01 00 58



ECMWF CRAY-1 CPU UTILISATION BY TASK SUBTYPE 1
TUE 31 AUG 1982 06 30 53 TO 07 00 54



ECMWF CRAY-1 DISK CHANNEL USAGE SUBTYPE 6
WED 25 AUG 1982 00 30 57 TO 01 00 58



ECMWF CRAY-1 DISK CHANNEL USAGE SUBTYPE 6
TUE 31 AUG 1982 06 30 53 TO 07 00 54