TECHNICAL MEMORANDUM

69

Benchmark exercises on Cray X MP

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

Operations Department

November 1982

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

1. INTRODUCTION

The paper ECMWF/TAC-NCF(82)2 defines a set of benchmarks to be executed on a prototype Cray X-MP. This paper describes the tests carried out and tabulates the results so obtained. The tests were completed at Cray Research Incorporated's facility at Mendota Heights, Minneapolis, during the period 8th to 22nd October, 1982.

2. ACKNOWLEDGEMENT

Considerable assistance and support was provided by many of the staff of Cray Research, for which grateful acknowledgement is made. In particular, the personal support and long unsociable hours worked by Mike Brown and Peter Sydow were especially appreciated.

3. SUMMARY OF JOBS USED IN THE BENCHMARK

3.1 The grid point forecast model

The operational version of ECMWF's forecast model requires:

a. work files to accommodate data which cannot be retained in memory
b. that files be written during "write-up" steps as input to a post processing job
c. that a post processing job be run to convert the "written-up" data into the form required by ECMWF's operational suite.

It is usual to use 4 work files, 2 for input and 2 for output, during a model time step, whereas logically a single input file and a single output file would suffice. The increased number of files is simply a device to spread the input/output load across 4 disk controllers. As data is required to be read several lines ahead of data to be written back, it would also be feasible to use a single work data set provided an efficient random input/output scheme were available. Thus 3 configurations of the forecast were considered:

(i) 4 work files, each of 1.7 megawords
(ii) 2 work files, each of 3.4 megawords. Since double length records are used in this case, the number of input/output requests is halved
(iii) 1 work file of 3.4 megawords accessed at random. Again, double length records are used.

All three input/output configurations were available. In the case of the 4 and 2 work file versions, about 94% of the input/output could, in theory, be overlapped by CPU processes. For the 1 work file version Cray software does not permit the overlapped read and write to proceed concurrently. In consequence only 47% of the input/output could, in theory, be overlapped without considerable re-coding of the forecast.

The forecast, though performing a considerable amount of input/output, is highly CPU bound. Exceptions are the steps which write up data for subsequent post processing. Such "write up" steps write an additional 2 to 3 megawords of data to disk. As this data is not written to pre-allocated continguous disk space, times for these steps can vary considerably. Account of this variation should be taken when comparing different tests.

3.2 The post processing job

The post processing job which converts the "written up" data to the form required by ECMWF's operational suite is predominantly input/output bound. For the reasons given above, a considerable variation in times can be expected from one run to another.

3.3 The grid point model simulator

The ECMWF grid point model simulator is described in Technical Memorandum No.65. It is capable of simulating both the forecast model and the post processing job. The main purpose of using the simulator was to enable a study to be made of jobs running two at a time, utilising the two CPUs of the X-MP. Memory limitations prevented real jobs from being run in parallel; the simulator was not so restricted. In consequence it was possible to simulate two forecasts running in parallel using various devices for workfiles; it was also possible to run a simulated forecast with a real analysis.

3.4 The ECMWF analysis

The ECMWF analysis is not a simple program - it should be viewed as a series of job steps involving both programs and data manipulation. The version tested involved:

a. copying 19 input files
b. compilation of 11 source codes
c. copying first guess sigma value, estimated errors, pressure co-ordinate persistence, initialised sigma persistence, climatological data, etc.
d. sigma to pressure conversion (program)
e. interpolation (program) to de-stagger data
f. statistics of first guess (program)
g. copying observations, observation errors, etc.
h. pre-GAP stage of analysis (program)
i. copying observation correlations
j. GAP data checking phase of analysis (program)
k. GAP data assimilation phase of analysis (program)
l. POST GAP stage of analysis (program)
m. produce humidity first guess errors (program)

n. copy humidity first guess errors
o. steps h to l are repeated for humidity analysis
p. copy POST GAP error file
q. re-normalise (program)
r. copy pressure co-ordinate analysis for plotting
s. copy pressure co-ordinate increments, normalised analysis errors, first guess, unconverted observations, first guess errors, etc., for plotting
t. interpolate to staggered grid (program)
u. statistics of analysis (program)
v. pressure to sigma conversion (program)
w. stratospheric correction (program)
x. copying all the main files to a single data set
y. extraction of extremes (program)
z. end of job.

In consequence, a considerable time is taken during the analysis job simply moving data from disk to disk. Such processes, as explained in 3.1 above, could not be expected to produce constant timings. Also, since the types of disks attached to the X-MP are essentially the same as those attached to the Cray 1A at ECMWF, little if any improvement in timing could be expected from such processes. In order to confine the benchmark test as far as possible to the program steps of the analysis, it was decided that analysis times would be taken from the beginning of step d. to the end of step w., omitting the initial data copies and compilations, and the final copying of all the main files.

4. TESTS USING THE FORECAST

4.1 The Forecast Duration Tested

A 48-hour forecast was run initially. This showed that apart from the first 3 steps, a 24-hour forecast is typical of any 24-hour forecast period for the forecast model used. Further tests were, in consequence, confined to 24-hour forecasts, with 5 write-up steps at 6-hourly periods throughout the forecast (0,6,12,18 and 24 hours). Forecasts were normally followed by five job steps, each containing the post processing job to process the written-up data from one of the write-up steps.

4.2 The Forecast Configurations Tested

Fig. 4.1 contains the results of one forecast run at ECMWF on Cray 1A and six forecasts run on the Cray X-MP. Previous investigations have shown that the 4 work file configuration is optimal for input/output to disks. Investigations into an optimal configuration using buffer memory resident (BMR) data were not

FORECAST CONFIGURATION			TIME FOR FORECAST		AVERAGE TIME PER POST PROCESSING JOB		TIME FOR FORECAST +5 POST PROCESSING JOBS	
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS	CPU	CLOCK	CPU	CLOCK	CPU	CLOCK
4	DISK	ECMWF CRAY 1-A	1125	1311	22.2	72	1236	1671
4	DISK	X-MP	630	770	17.2	44.8	716	994
1	BMR	X-MP	630	737	17.2	42.2	716	948
2	SSD	X-MP SYNCHRONOUS I/O	630	686	17.2	43.6	716	904
2	SSD	X-MP ASYNCHRONOUS I/O	630	666	17.2	46.6	716	909
4	SSD	X-MP SYNCHRONOUS I/O	630	702	17.2	43.0	716	919
1	SSD	X-MP SYNCHRONOUS I/O	630	688	17.2	44.0	716	908

Fig. 4.1 Forecast and post processing times

NOTE: BMR = Buffer memory resident data
SSD = Solid state storage device
All times are in seconds.

FORECAST CONFIGURATION			CLOCK TIMES (SECONDS)							
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS	F/C MINUS WRITE UP STEPS	WRITE STEP 1	WRITE STEP 2	WRITE STEP 3	WRITE STEP 4	WRITE STEP 5	SUM OF WRITE UP STEPS	TOTAL TIME FOR F/C
4	DISK	ECMWF CRAY 1-A	1131	31	36	36	40	37	180	1311
4	DISK	X-MP	662	20	25	23	22	18	108	770
1	BMR	X-MP	640	21	20	20	20	16	97	737
2	SSD	X-MP SYNCHRONOUS I/O	592	21	19	20	19	15	94	686
2	SSD	X-MP ASYNCHRONOUS I/O	571	21	19	20	20	15	95	666
4	SSD	X-MP SYNCHRONOUS I/O	607	21	20	19	19	16	95	702
1	SSD	X-MP SYNCHRONOUS I/O	593	21	19	19	20	16	95	688

Fig. 4.2 Breakdown of forecast times

NOTE: BMR = Buffer memory resident data SSD = Solid state storage device All times are in seconds

possible, as only 4 megawords of buffer memory were available. The 1 work file test possible with BMR allows only up to 47% of the input/output to be overlapped by CPU processing (see 3.1 above). The initial tests using the solid state storage device (SSD) were carried out using a synchronous driver, with the result that no overlapping of input/output with CPU processing was possible. Towards the end of our time on the X-MP a new, asynchronous driver for the SSD was made available, and the 2 workfile SSD forecast test was re-run.

4.3 Presentation of Forecast Results

In Fig. 4.1 the times for each forecast and its associated post processing jobs have been recorded separately, allowing the forecast performance to be assessed without the variability introduced by the disk input/output associated with the post processing jobs.

4.4 Break-down of Forecast Times

As the forecast times given in Fig. 4.1 contain write-up steps which are subject to the variability introduced by input/output to disk, they have been broken down into their component parts in Fig. 4.2. This figure shows the net time for each forecast when times for write-up steps have been removed, together with the times for each write-up step.

4.5 X-MP to Cray 1 Ratios for Net Forecast Clock Times

Each forecast was run stand-alone, leaving one of the X-MP^{S} two CPU^{S} idle throughout its run. Using the first column of times in Fig. 4.2 as a basis for comparison, the following ratios of X-MP times in terms of Cray 1 times were obtained:-

a) 1.71 using 4 disk work files.

b) 1.77 using 1 BMR random work file.

c) 1.86 using 4 SSD work files.

d) 1.91 using 1 SSD random work file.

e) 1.98 using 2 SSD work files.

4.6 Some Practical Implications of the Results

The difference in time taken when using the asynchronous SSD driver compared to the synchronous driver indicates that the cost of non-overlapped input/output to SSD is only 3.6% of the total time taken. The use of non-overlapped input/output would enable the forecast to reduce the amount of main memory used by at least 1/3 of its current use, as much space is taken by double buffering techniques, retaining values in core to avoid input/output, etc. From the ratio

FORECAST CONFIGURATION			TIME FOR FORECAST		AVERAGE TIME PER POST PROCESSING JOB		TIME TO COMPLETE F/C PLUS 5 P.P. JOBS	
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS	CPU	CLOCK	CPU	CLOCK	CPU	CLOCK
1 + 3	SSD	DEGRADED FORECAST WITH EXTRA WORK FILES FOR HELMHOLTZ SOLUTION	632	754	17.6	84	720	838

FORECAST CONFIGURATION			CLOCK TIMES (SECONDS)							
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS	F/C MINUS WRITE UP STEPS	WRITE STEP 1	WRITE STEP 2	WRITE STEP 3	WRITE STEP 4	WRITE STEP 5	SUM OF WRITE UP STEPs	TOTAL TIME FOR F/C
1 + 3	SSD	DEGRADED FORECAST WITH EXTRA WORK FILES FOR HELMHOLTZ SOLUTION	659	18	21	20	20	16	95	754

Fig. 4.3 Test with degraded forecast and post processing

NOTE: SSD = Solid state storage device.
4 of the 5 post processing jobs executed in parallel with the forecast.
All times are in seconds.

of maximum input/output rates obtained in later tests (see 7.4) it can be deduced that the corresponding cost of non-overlapped input/output for BMR would be 24% of the total time taken.

4.7 Special Test using a Degraded Forecast and Post Processing Job

An additional test was performed in which the forecast was configured to run in a degraded mode. A version of the forecast was used where the right hand sides of the Helmholtz equations are written to work files instead of being retained in memory. During the Helmholtz Solution these files are read backwards. This could have been achieved using random access on SSD, but would have required code changes too extensive to be performed in the time available. Thus the existing scheme, which involves double BACKSPACE operations with their associated high cost in overheads was used. For this test, the post processing jobs were degraded to use only a small amount of memory at the expense of considerably increased input/output. With this configuration, it was possible to submit the post processing jobs from the forecast, and to allow forecast and post processing jobs to share memory, using both CPU^{S}. Thus 4 post processing jobs completely overlapped the forecast, while the final post processing job ran stand alone after the forecast had completed. This used one CPU fully throughout the test, whereas the second CPU was idle for part of the time. The full details of this test are contained in Fig. 4.3.

In this case the clock time for the forecast plus 5 post processing jobs represents the time from the beginning of the forecast to the end of the last post-processing job. Expressing this time as a ratio of X-MP time in terms of Cray 1 time a figure of 2.01 is obtained.

4.8 X-MP to Cray 1 Ratios for CPU Times

The CPU times taken by various sections of the forecast and post-processing on both Cray 1A and Cray X-MP have been broken down and compared in Fig. 4.4

4.9 Accuracy of Results

The results obtained from forecasts on the X-MP did not completely agree with those obtained at ECMWF. It was thought that this was probably due to minor computational differences arising from the use of different library routines. To establish that X-MP hardware was not responsible for these differences, an X-MP forecast binary was transferred and run on Cray 1s serial number 42. The results of this forecast agreed exactly with the results obtained from the Cray X-MP.

SECTION OF FORECAST OR POST PROCESSING BEING TIMED	CPU TIME (CRAY 1A)	CPU TIME (CRAY X-MP)	RATIO (1A/XMP)
NORMAL FORECAST TIME STEP	8.49	4.61	1.84
FORECAST WRITE-UP STEP	26.34	16.50	1.60
WRITE-UP OVERHEAD	17.85	11.89	1.50
FORECAST RADIATION STEP	74.31	43.22	1.72
RADIATION OVERHEAD	65.82	38.61	1.71
24 HOUR FORECAST	1125.18	630.00	1.79
POST PROCESSING JOB	22.17	17.15	1.29
5 x POST PROCESSING JOBS	110.85	85.75	1.29
24 HOUR F/C + 5 x PP JOBS	1236.03	742.48	1.66

Fig. 4.4 CPU times taken by components of the forecast and post processing

5. TESTS USING THE SIMULATOR

5.1 Simulations run as Stand Alone Jobs

The tests using the forecast described in Figs. 4.1 and 4.2 were repeated as stand alone jobs using the simulator. This enabled a comparison to be made between the performance of the simulator and that of the real forecast. In addition, some configurations were tested by means of the simulator. In all cases the job limit was set to one, and the post processing jobs ran one at a time after the completion of the forecast simulation. As with the real forecast, write-up steps and post processing simulations exhibited variability due to input/output being performed to non-contiguous disk data sets. The results of these tests are presented in Figs. 5.1 and 5.2, which may be compared directly with Figs. 4.1 and 4.2. It should be noted that the initial 3 steps of the real forecast involve slightly more data manipulation than the initial simulator set-up. Also a small amount of disk input/output associated with radiation steps in the forecast (there are 3 radiation steps in a 1-day run) was not simulated. In consequence, simulated times are slightly less than real times.

5.2 Simulations run as Pairs of Jobs simultaneously

A second set of simulations were performed to investigate the performance of the Cray X-MP when handling two jobs simultaneously, each using one of the two available CPU^{S}. This was done by setting the job limit to 2, and starting both simulated forecasts at the same time. At each forecast write-up step a post processor simulation was submitted. When the first simulated forecast completed, it was replaced by one of the queued post processing jobs. In this way there were always two jobs in execution at any one time, with the possible exception of the last post processing job. Fig. 5.3 shows the times taken by each simulated forecast, together with the times taken by the simulated post processing jobs. The final column in Fig. 5.3 shows the time taken for each test to complete, and represents the time from the beginning of the pair of forecasts to the end of the last post processing job run.

5.3 Breakdown of Results from Pairs of Simulations

The results from pairs of forecasts have been broken down in Fig. 5.4 to provide a presentation similar to that of Figs. 4.2 and 5.2. It is important to note the effects of disk contention on the write-up steps. Care was taken to pre-allocate two separate sets of contiguous work files spread over 8 separate disks, to enable two simulated forecasts using disk work files to proceed without contention. As only 8 disks were available in total, the data written up at write-up steps could not always be directed to disks not involved in other input/output operations. In consequence, write-up step times vary from 18 seconds

SIMULATED FORECAST CONFIGURATION			TIME FOR FORECAST		AVERAGE TIME PER POST PROCESSING JOB		TIME FOR FORECAST + 5 POST PROCESSING JOBS	
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS	CPU	CLOCK	CPU	CLOCK	CPU	CLOCK
4	DISK	X-MP	629	779	17.2	42.2	715	990
1	BMR	X-MP I/O AS FOR FORECAST	629	735	17.3	54.5	716	1007
2	SSD	X-MP SYNCHRONOUS I/O	629	666	17.3	40.9	716	871
2	SSD	X-MP ASYNCHRONOUS I/O	629	651	17.3	54.2	716	922
4	SSD	X-MP SYNCHRONOUS I/O	629	686	17.2	41.8	715	895
1	SSD	X-MP SYNCHRONOUS I/O	629	679	17.2	41.5	715	877
1	BMI	X-MP I/O OVERLAPPED AS WOULD BE IF F/C RE-CODED	629	709	17.3	40.9	716	913
4	SSD	X-MP ASYNCHRONOUS I/O	630	649	17.3	41.8	717	895
1	SSD	X-MP ASYNCHRONOUS I/O	629	664	17.3	54.2	716	935

Fig. 5.1 Simulated forecast and post processing times

NOTE: BMR = Buffer memory resident data
SSD = Solid state storage device
All times are in seconds

SIMULATED FORECAST CONFIGURATION			CLOCK TIMES (SECONDS)							
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS	F/C MINUS WRITE-UP STEPS	WRITE STEP 1	WRITE STEP 2	WRITE STEP 3	WRITE STEP 4	WRITE STEP 5	SUM OF WRITE UP STEPS	TOTAL TIME FOR F/C
4	DISK	X-MP	675	20	21	21	21	21	104	779
1	BMR	X-MP I/O AS FOR FORECAST	622	21	23	23	23	23	113	735
2	SSD	X-MP SYNCHRONOUS I/O	578	18	17	18	17	18	88	666
2	SSD	X-MP ASYNCHRONOUS I/O	555	19	19	18	19	21	96	651
4	SSD	X-MP SYNCHRONOUS I/O	587	19	20	20	20	20	99	686
1	SSD	X-MP SYNCHRONOUS I/O	581	19	20	20	20	19	98	679
1	BMR	I/O OVERLAPPED AS WOULD X-MP BE IF F/C RE-CODED	619	18	18	18	18	18	90	709
4	SSD	X-MP ASYNCHRONOUS I/O	560	18	18	18	18	17	89	649
1	SSD	X-MP ASYNCHRONOUS I/O	568	18	19	19	20	20	96	664

Fig. 5.2 Breakdown of simulated forecast times

NOTE: BMR = Buffer memory resident data
SSD = Solid state storage device
All times are in seconds

SIMULATED FORECAST CONFIGURATION				CLOCK TIMES (SECONDS)		
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS		TIME TAKEN BY FORECAST	TIME TAKEN BY 5 PP JOBS	TIME TAKEN TO COMPLETE TEST
4	DISK	ECMWF	IN THE LAST COLUMN ONLY TIMES HAVE BEEN DOUBLED TO SHOW TIMES FOR 2 F/C	1311	360	3342
4	DISK	X-MP	RUN TOGETHER	926	230	1176
4	DISK			947	220	
1	SSD	X-MP	RUN TOGETHER	687	226	923
1	SSD			689	226	
1	BMR	X-MP	RUN TOGETHER	845	266	1125
4	DISK			855	260	
1	SSD	X-MP	RUN TOGETHER	752	349	1172
4	DISK			856	358	

Fig. 5.3 Simulations of pairs of forecasts using dual CPUs

SIMULATED FORECAST CONFIGURATION			CLOCK TIMES (SECONDS)		
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS	TIME TAKEN BY FORECAST	TIME TAKEN BY 5 PP JOBS	TIME TAKEN TO COMPLETE TEST
1	SSD	X-MP RUN TOGETHER	703	182	986
1	BMR		810	210	
2	SSD	X-MP RUN TOGETHER	754	286	1124
4	DISK		895	278	
2	SSD	X-MP RUN TOGETHER	710	217	994
1	BMR		819	210	

Fig. 5.3 (cont.) Simulations of pairs of forecasts using dual CPUs

SIMULATED FORECAST CONFIGURATION			CLOCK TIME (SECONDS)							
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS	F/C MINUS WRITE-UP STEPS	WRITE STEP 1	WRITE STEP 2	WRITE STEP 3	WRITE STEP 4	WRITE STEP 5	SUM OF WRITE UP STEPS	TOTAL TIME FOR F/C
4	DISK	ECMWF CRAY 1-A	1131	31	36	36	40	37	180	1311
4	DISK	X-MP RUN TOGETHER	770	29	41	26	26	34	156	926
4	DISK	X-MP RUN TOGETHER	732	33	63	56	29	34	215	947
1	SSD	X-MP RUN TOGETHER	582	21	20	20	23	21	105	687
1	SSD	X-MP RUN TOGETHER	581	21	20	20	24	23	108	689
1	BMR	X-MP RUN TOGETHER	662	33	33	23	36	26	151	813
4	DISK	X-MP RUN TOGETHER	719	36	23	24	26	25	134	853
1	SSD	X-MP RUN TOGETHER	608	20	24	21	41	28	144	752
4	DISK	X-MP RUN TOGETHER	752	20	20	24	20	20	104	856

Fig. 5.4 Breakdown of times for simulated pairs of forecasts

SIMULATED FORECAST CONFIGURATION			CLOCK TIME (SECONDS)							
NUMBER OF WORK FILES	WORK FILE LOCATION	COMMENTS	F/C MINUS WRITE-UP STEPS	WRITE STEP 1	WRITE STEP 2	WRITE STEP 3	WRITE STEP 4	WRITE STEP 5	SUM OF WRITE UP STEPS	TOTAL TIME FOR F/C
1	SSD	X-MP RUN TOGETHER	592	37	18	18	19	19	111	703
1	BMR		697	34	21	20	19	19	113	810
2	SSD	X-MP RUN TOGETHER	577	34	31	44	42	31	177	754
4	DISK		769	36	22	22	22	24	126	895
2	SSD	X-MP RUN TOGETHER	593	39	19	19	20	20	117	710
1	BMR		707	36	19	20	19	18	112	819

Fig. 5.4 (Cont.) Breakdown of times for simulated pairs of forecasts

to 63 seconds. Similar considerations account for much of the variability of the times taken by post processing jobs in Fig. 5.3.

5.4 Throughput Ratios obtained from Pairs of Simulations

Taking the times tabulated in the last column of Fig. 5.3, only tests using SSD or BMR workfiles gave Cray 1A throughput ratios in excess of 3, the best being the case where both simulations used SSD (3.62). The tests involving at least one simulation with disk resident work files returned ratios in the range 2.84 to 2.97. In all of these cases it is evident that disk contention during the write-up steps and the execution of the post processing simulation was sufficiently degrading to cause a ratio that would have been in excess of 3 to fall below 3.

5.5 The "CPU Hang" Problem

In addition to the problem of disk contention, a second problem having a detrimental effect on the performance of pairs of jobs executing simultaneously was observed. While such jobs were running, their CPU utilisation was continually monitored using a display available on an input/output subsystem (IOS) monitor screen. Occasionally, one job was observed to pause "waiting for CPU" while one CPU would remain "100% idle", resulting in a delay in that job's processing. This problem appears to be a job scheduler problem, and is being addressed by Cray Research. Steps were taken when running tests to minimise the problem through manual interference, but it could not be completely removed.

5.6 An Assessment of the "CPU Hang" Problem Overheads

An attempt was made to assess the degradation caused by the "CPU hang" problem described in 5.5. For a normal Cray 1 job, system overheads can be calculated by

a) adding up the figures given at the end of the job for "Time Spent Executing in CPU",
"Time Waiting to Execute", and
"Time Waiting for I/O"

b) subtracting this total from the clock time for the job

This was done for several "stand alone" simulations run one at a time. For these simulations, disk I/O variability was removed by suppressing write-up steps and post processing simulations. The system overhead was computed for the tests run "stand alone". Several pairs of simulations were then submitted, and the computation repeated for each job separately. It was supposed that the difference in time a) as defined above when obtained from a stand alone test

SIMULATED FORECAST CONFIGURATION			EXECUTING IN CPU + WAITING TO EXECUTE + WITING FOR I/O	CLOCK TIME FOR JOB	"PROBLEM" OVERHEAD	SYSTEM OVERHEAD
NUMBER OF WORK FILES	WORK FILE LOCATION	C O M M E N T S				
4	DISK	DISKS ON BIOP	690	702	-	12
4	DISK	DISKS ON DIOP	692	694		2
1	SSD		577	586	-	9
1	BMR		632	641	-	9
1	BMR	RUN TOGETHER	675	690	43	15
4	DISK	RUN TOGETHER	707	727	17	20
1	SSD	RUN TOGETHER	587	601	10	14
4	DISK	RUN TOGETHER	684	705	-7	21
1	SSD	RUN TOGETHER	581	593	4	12
1	BMR	RUN TOGETHER	632	644	0	13

Fig. 5.5 System and "problem" overheads

compared to that obtained when the test was run together with another job would give some indication of the "problem" overhead. The results of these tests are given in Fig. 5.5.

6. TESTS USING THE ANALYSIS

6.1 The Version of the ECMWF Analysis used

The version of the ECMWF Analysis tested was similar to that used in the ECMWF Operational Data Assimilation cycle.

6.2 Presentation of Results

The clock times obtained by various Analysis configurations are presented in Fig. 6.1. Analysis times were extracted from the first "LDR" loader step to the end of the "SFC" stratospheric correction step (i.e. from step d) to the end of step w) as defined in 3.4 above). This enabled variable length compilations, and spurious job steps at the end associated with the disposal of results to "PT" tape to be excluded. Analyses were successfully run using

a) all files on disk

b) random workfiles on SSD

c) random workfiles BMR.

A variety of file configurations were tried in order to establish optimum positioning. Since the Analysis code is sensitive to compiler changes, a large block compiler (CFTI5K) was built to obtain similar optimisation to that obtained at ECMWF. Some tests were performed with the new asynchronous SSD driver when it became available.

6.3 CPU Times for Analysis Job Steps

Figure 6.2 presents comparative CPU times for the 4 major analysis job steps, together with X-MP - Cray 1 ratios.

6.4 Analyses and Simulated Forecasts run together

For these tests the Analysis was configured to run making use of up to 4 megawords of BMR or SSD space, other files being written to disk. Various input/output configurations were selected for simulated forecasts, and tests were performed running real Analyses and simulated forecasts simultaneously. Details of the results of these tests are given in Fig. 6.3. It should be noted that these tests were configured to use no more than 7½ megawords of SSD or BMR. Further savings could have been achieved had more SSD or BMR storage been used.

ANALYSIS CONFIGURATION	CLOCK TIMES
ECMWF CRAY-1	745
XMP:-	
FILES ON SSD:-	461 (CFT)
	451 (CFT 15K)
	454 (CFT 15K + NEW DRIVER)
FILES ON HALF SSD:-	468 (CFT 15K)
	472 (CFT 15K + NEW DRIVER)
FILES ON BMR:-	499 (CFT 15K)
FILES ON DISK:-	618 (CFT 15K)
½ SSD WITH F/C RUNNING ON SSD	493 (CFT 15K)
BMR WITH F/C ON SSD	527 (CFT 15K)
½ SSD WITH F/C ON DISK	529 (CFT 15K)
BMR WITH F/C ON DISK	621 (CFT 15K)

Fig. 6.1 Analysis times

JOB STEP OF THE ANALYSIS BEING TIMED	CPU TIME (CRAY 1A)	CPU TIME (CRAY X-MP)	RATIO (1A/XMP)
PRE-GAP	14.2	9.7	1.46
GAP (DATA CHECKING MODE)	155.6	115.6	1.35
GAP (ANALYSIS)	259.8	180.5	1.44
POST GAP	14.5	10.5	1.35

Fig. 6.2 CPU times taken by principal job steps of the analysis

DETAILS OF TEST	ANALYSIS TIME	ANALYSIS RATIO	FORECAST TIME	FORECAST RATIO	COMBINED RATIO
ANALYSIS + FORECAST AT ECMWF	745	1.0	1311	1.0	1.0
ANALYSIS:- USING ½ SSD FORECAST:- USING SSD	493	1.51	711	1.84	3.35
ANALYSIS:- USING ½ SSD FORECAST:- USING DISKS	529	1.41	964	1.36	2.87
ANALYSIS:- USING BMR FORECAST:- USING SSD	527	1.41	691	1.90	3.30
ANALYSIS:- USING BMR FORECAST:- USING DISKS	621	1.20	918	1.43	2.63

Fig. 6.3 Analysis and simulated forecasts together

NOTE: In all the above cases the analysis completed before the forecast. The job limit was set to 2 and the outstanding post processing simulations successively executed as the second job. This "extra work" was counter balanced by some idle time on one CPU late in each forecast.

7. INPUT/OUTPUT TESTS

7.1 Introduction

A series of short tests were performed to examine the rate of input/output that could be achieved on the various devices available. Fig. 7.1 illustrates the configuration of the various devices.

7.2 The Input/output Tests

A simple FORTRAN program was written to write and read large blocks of data. The data was directed to each type of device, then jobs were run in pairs to assess the maximum data rate that could be obtained.

The jobs performing I/O to SSD and buffer memory were separately run with a job performing I/O to a single disk, to study the degradation in I/O on the fast devices when performed in conjunction with I/O on a slower device.

7.3 Presentation of Results

The results of these tests are presented in Figs. 7.2 and 7.3. A transfer rate 6.67 times that for BMR was obtained for SSD.

It should be noted that no attempt was made to optimise disk I/O by ensuring that the data be written to contiguous disk space.

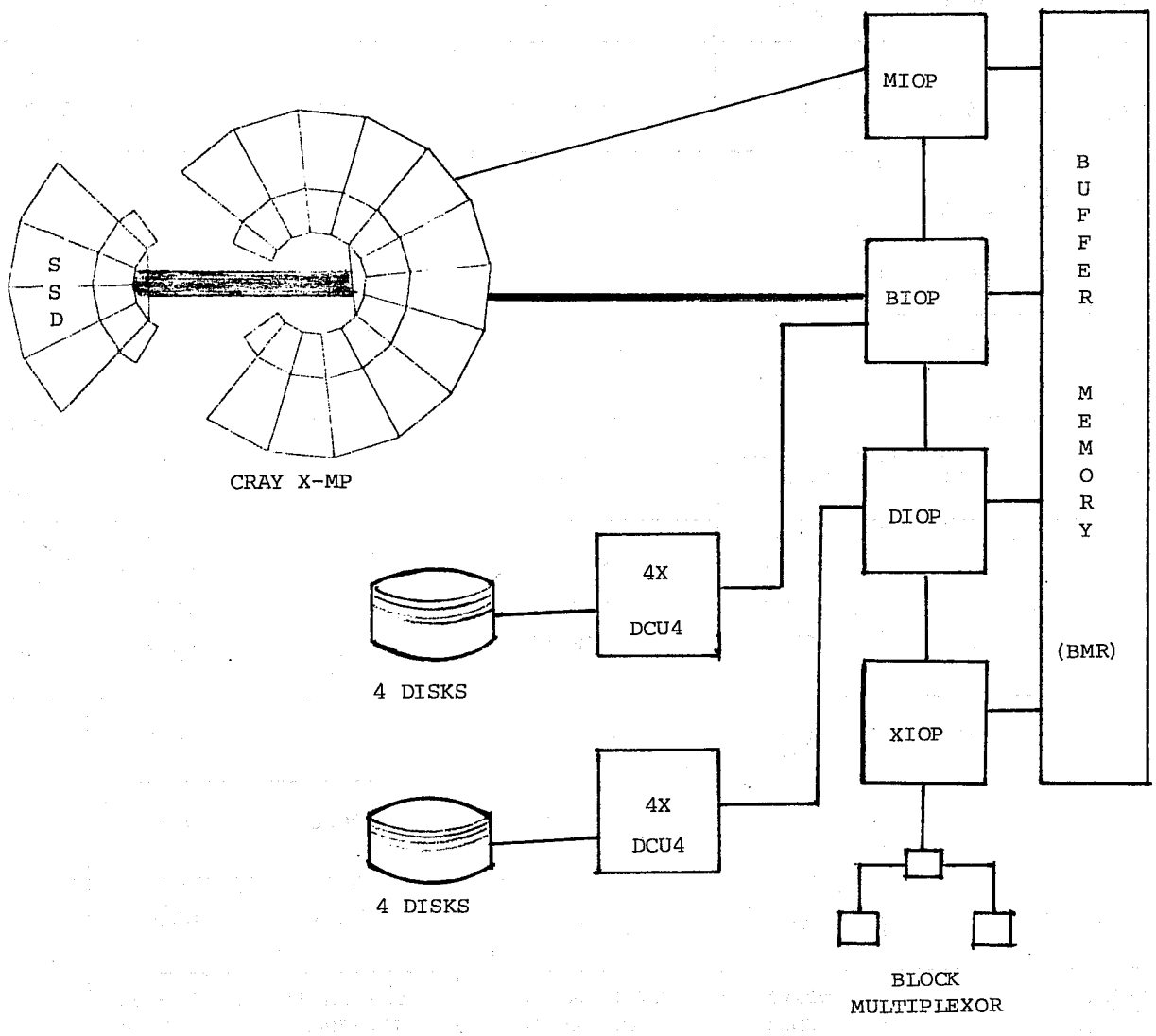

Fig. 7.1 X-MP configuration

DESCRIPTION OF TEST	RECORD LENGTH	NUMBER OF TRANSFERS	NUMBER OF M.WORDS TRANS.	TIME TAKEN	TRANSFER RATE		
					MW/SEC	MBYTE/SEC	MBIT/SEC
1 JOB - ALL I/O TO SSD	384000	1600	614.4	20	30.72	245.76	1966.08
2 JOBS IN PARALLEL - ALL I/O TO SSD	384000	3200	1228.8	33	37.24	297.89	2383.12
1 JOB - ALL I/O TO BMR	384000	1600	614.4	113	5.44	43.50	347.98
2 JOBS IN PARALLEL - ALL TO BMR	384000	3200	1228.8	220	5.58	44.68	357.47
1 JOB - ALL I/O TO 1 DISK	768000	800	614.4	1247	0.49	3.94	31.53
2 JOBS IN PARALLEL: I/O TO 4 DISKS VIA DIOP	384000	32 x 80 =2560	983.04	589	1.67	13.35	106.82
2 JOBS IN PARALLEL: I/O TO 4 DISKS VIA BIOP	384000	32 x 69 =2208	847.87	589	1.44	11.52	92.12
2 JOBS IN PARALLEL: COMBINED RATES			1830.92		3.11	24.87	198.94
TRANSFER RATES PER DISK FOR ABOVE JOBS: DIOP DISKS					0.42	3.34	26.71
TRANSFER RATES PER DISK FOR ABOVE JOBS: BIOP DISKS					0.38	2.88	23.03
SSD I/O (IN PARALLEL WITH DISK I/O)	384000	1600	614.4	21	29.25	234.06	1872.46
BMR I/O (IN PARALLEL WITH DISK I/O)	384000	1600	614.4	141	4.36	34.86	278.88

Fig. 7.2 Input/output tests

I/O TARGET	I/O RATE (M.BYTES/SECOND)	
	STAND ALONE JOB	2 COPIES OF JOB
SSD	245.76	297.89
BMR	43.50	44.68
SINGLE DISK	3.94	
8 DISKS	One job performing I/O to 4 disks via DIOP (A2)	24.87
4 DISKS (A2)		13.35 (3.34 per disk)
4 DISKS (A1)	One job performing I/O to 4 disks via BIOP (A1)	11.52 (2.88 per disk)

Fig. 7.3 INPUT/OUTPUT RATES

8. 16 BANK MEMORY TESTS

8.1 Introduction

In order to test the impact of a Cray X-MP with memory arranged in 16 rather than 32 banks, the following programs were run with the X-MP re-configured as a ½m word, 16 bank machine.

8.2 The post processing job

The degraded post processing job (the only part of the forecast/analysis code that would conveniently run in a ½ m word machine) was run stand alone in one CPU, and in both CPUs together. Results are contained in Fig. 8.1.

8.3 The special kernal

A "kernal", designed to use memory as heavily as possible, was run on both the 32 and 16 bank machines. It demonstrates the worst possible case of degradation likely between two jobs running simultaneously.
The results are contained in Figure 8.2, and demonstrate a degradation of 25%.

CONFIGURATION	CPU TIME	TOTAL TIME
32 BANK STAND ALONE	17.6	84
16 BANK STAND ALONE	17.7	92
16 BANK 2 JOBS TOGETHER	17.7 17.7	90 157

Fig. 8.1 16 bank test using post processing job

CONFIGURATION	CPU TIME
32 BANK STAND ALONE	58.9
16 BANK STAND ALONE	60.8
16 BANK 2 JOBS TOGETHER	70.3 80.2

Fig. 8.2 16 bank "kernal" test

References

ECMWF/TAC-NCF(82)1 Introduction to the Cray X-MP and description of bench marking exercises.

ECMWF/TAC-NCF(82)2 Plans for further bench mark exercises for Cray X-MP.

Dent, D., Gibson, J.K., Storer, N., 1982 The ECMWF grid point model simulator. ECMWF Technical Memorandum No.65.