USING HARDWARE MONITOR TO SOLVE PROBLEMS ON HP3000 - III

Ivan Loffler
Senior DP Staff Analyst
GTE Service Corporation
P. O. Box 1548 - F071
Tampa, FL 33601

I. ABSTRACT

This paper describes a case study concerned with solving a severe response time problem on an HP3000 large data base system. A major tool used on this project was the hardware monitor.

The paper states the problem, the objectives, and describes the problem solution. To arrive at a solution it was necessary to define levels of utilization of the troubled system, for which the hardware monitor was used directly. Also, the application of capacity planning guidelines is described here, especially how the guidelines were used for the definition of system utilization.

Also, software tools such as: SOO, SHOWME, and Event Monitoring Facility were used. These tools were calibrated for accuracy (i.e., capture ratio) and for overhead by comparison to hardware monitor data.

The results of this project, including conclusions and recommendations are contained in this paper.

II. INTRODUCTION

This paper describes the problems and the solutions as they occurred in one of our data centers within GTE. The main problem was an extremely long response time (up to 30 minutes) experienced by users of a large data base system, run on a Hewlett-Packard HP3000, Series III minicomputer.

The configuration consists of the CPU, multiplexor channel, selector channel, 1 megabyte of main memory, 8 disk drives, 2 tape drives and about 30 terminals connected via a multiplexor channel. The disk drives are attached to the system via the selector channel[1] (See Figure 1).

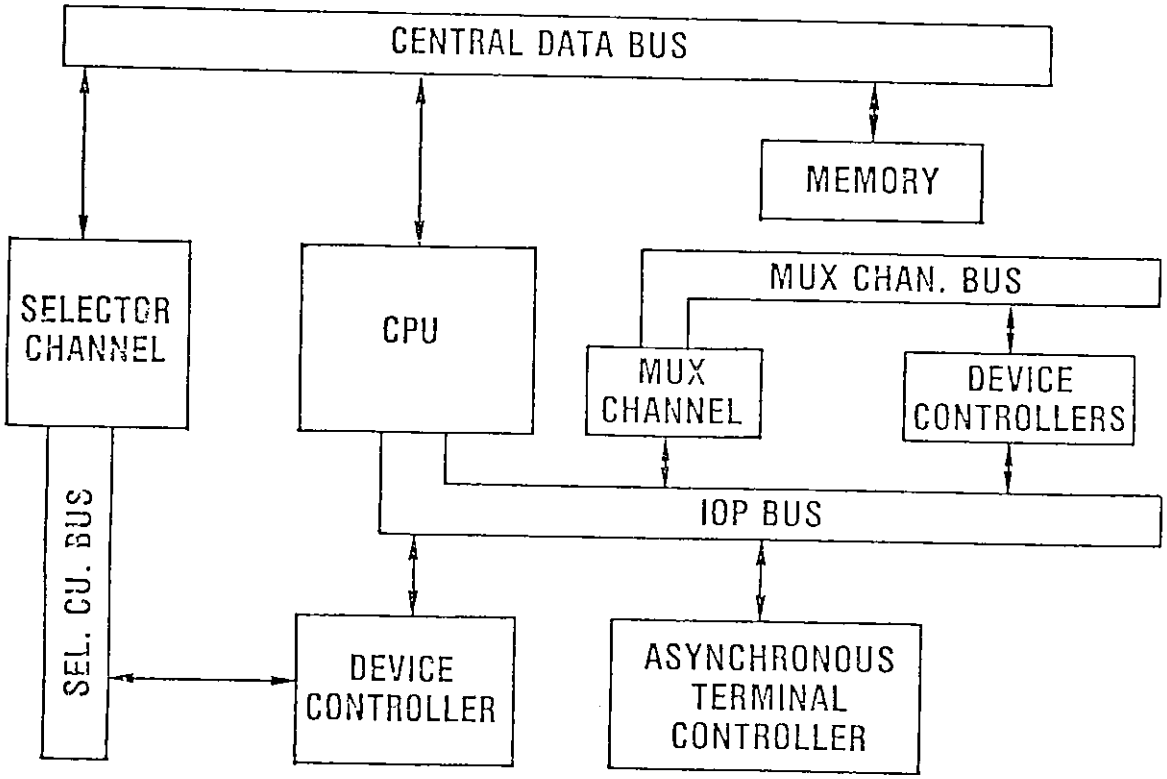

FIGURE 1

The hardware is controlled by the MPE-III operating system. The data base is supported by the IMAGE subsystem.

Also provided in this paper is an analysis of the problems and a discussion on selected diagnostic tools. A description of a major tool used in this project, the hardware monitor, is also included. The paper lists solutions designed to rectify the problem.

III. DEFINITION OF PROBLEMS

The HP3000 data center started to experience a severe degradation of the response time on its large online data base system. This system has been designed to keep track of the telephone circuits and related equipment, including trunk and facilities. It provides for such details as circuit orders and line assignments. The system requires an interaction with the engineering department, and has to absorb intensive data entry and frequent inquiries.

The response time delay of up to 30 minutes impaired the productivity of the operators, clerks, and engineers, who use the system. This resulted in delays in equipment ordering and could result in delays in service to telephone company customers.

After the management of the center recognized this as an emergency, several technical groups were involved and an informal task force was formed with an objective to improve the service levels of the system. The group included application development people, vendor's hardware and software engineers, the data center technical support and a corporate planning group, of which the author is a member.

This task force divided the investigation into appropriate areas of interest. The development team looked at the way the application design might be improved, while the Hewlett-Packard people investigated the load on the system using the Event Monitoring Facility, a software measurement tool. The planning group brought in the Tesdata hardware monitor to measure the hardware components of the system, to compare their utilization against the recently developed capacity guidelines for minicomputers, to calibrate all software tools used in this project, and to observe the impact of remedial changes on the investigated system.

IV. HARDWARE MONITORING[2]

Since the hardware monitor became a major diagnostic tool in this project, it deserves more explanation. The hardware monitor is a measurement device, independent of the monitored system called the host. It is electrically attached to the host's backplane pins and collects specific discrete electronic signals. Because it does not interact with any part of system's software, it is totally independent of host activities and thus does not impose any overhead on the measured host.

Figure 2 describes the hardware monitor in a block diagram. The electronic signals, such as CPU Busy, Selector Channel Busy, and Byte Count are captured by high speed probes and brought to the hardware monitor's capturing registers via a system of cables and concentrators. The signals are processed inside the hardware monitor by its internal Boolean logic circuitry. The hardware monitor's internal minicomputer organizes the signals into the memory in the form of data. The data are either displayed online on a CRT or recorded on a magnetic tape or disk for postprocessing and for printing of reports.

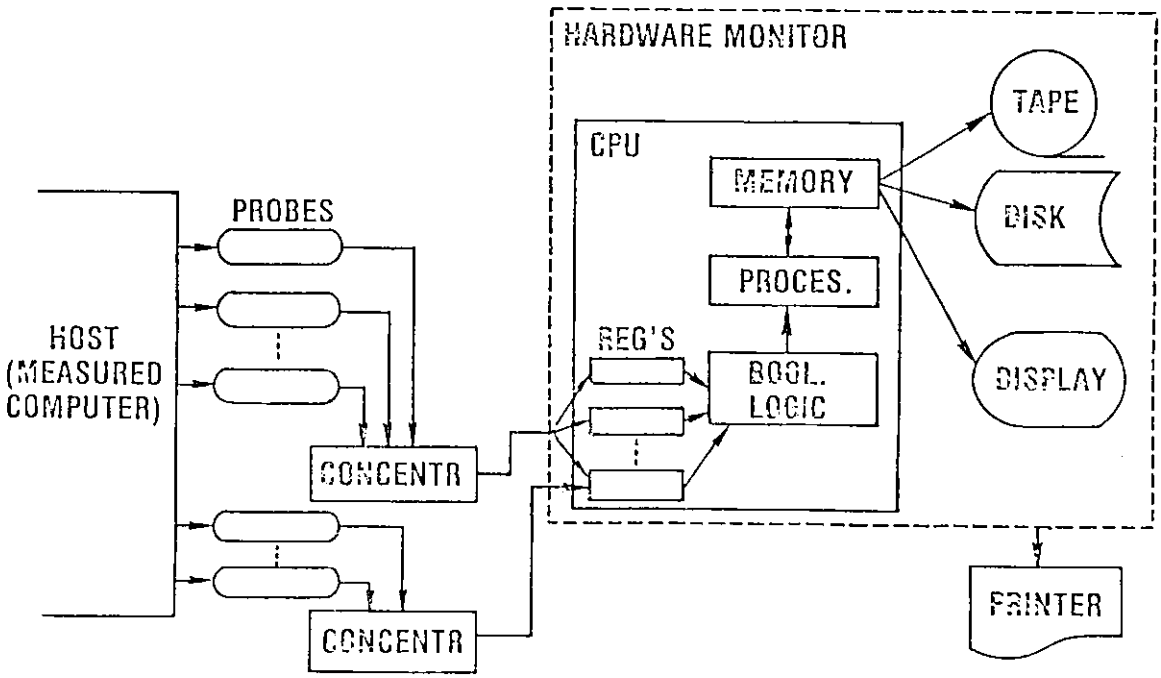

FIGURE 2

Basic information required about the system being measured is called the System Profile[3]. It basically consists of:

- CPU Busy
- CPU in Privilege State (Overhead)
- Multiplexor Channel Busy
- Selector Channel Busy
- Instruction Count
- Disk Drive Busy

The hardware monitor is the best suited device for capturing this information, since it does not skew the data by its own processing overhead, as software monitors do. Also, the Boolean logic of the hardware monitor allows for composite measures such as CPU/Selector Overlap, CPU Only, System Busy, and CPU Only (See Figure 3).

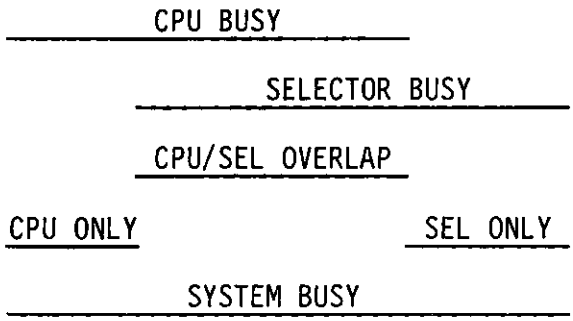

FIGURE 3

These measures are available from the hardware monitor only. They are crucial for the investigation of system performance. For example, the CPU/Selector Channel Overlap is commonly low on machines operating under the MPE-III operating system. This results in inefficient operation and decreases the usable capacity of the system.

V. ANALYSIS OF PROBLEMS

Analyzing the collected data from all the available tools, such as the hardware monitor, Event Monitoring Facility, SOO and SHOWME software monitors, and also from the physical observation of data center operations, there were several apparent areas for improvement:

1. Over 60% of the disk I/O activity was concentrated on one disk drive, creating a high amount of conflicts and long queues on this particular drive. Since this drive also contained a system data set, the situation was complicated even further by adding to the response time delay.

2. The system swapping amounted to 7.2 swaps of 8 KBytes per CPU second. Comparing this number with our Capacity Planning Guidelines, the swapping (or paging in IBM terminology) should not exceed 10 pages per second on a comparably sized IBM machine. Since every HP swap equals to two IBM pages, the maximum paging was 14 per seconds, exceeding the guidelines by 40%.

3. The CPU activity was low, compared to the amount of I/O activity. This suggests an I/O bound workload. Since some percentage of the CPU Busy time has to be attributed to the I/O processing, it was necessary to concentrate the tuning effort on the I/O area.

4. Some transactions contain up to 71 disk I/O's. This obviously is a design and data base deficiency, and it is adding considerably to the response time delay. There is an obvious correlation between the amount of I/O processed and the response time. The disk (Model Number 7920) timings are described in Table 1.

Average Random Seek Time:	25.0 ms
Average Rotational Delay:	8.3 ms
Data Transfer Time (937.5kB/Sec):	8.5 ms
Total per Average 8K Block	41.8 ms

TABLE 1

41.8 ms is the time required to process one I/O within the disk drive. There are some delays added: the contention for the drive, controller and channel, the CPU involvment (which may not always be overlapped with the I/O operation), and the memory contention. These factors may very well increase the time for processing an I/O to 1 second.

5. Number of active users was observed at a maximum of 28, which exceeded the previously recommended maximum of 15 active terminals. Previous monitoring of the simulated application system by the hardware monitor determined that on the average, one terminal would utilize the system at about 6% of its usable capacity.

6. The HP3000-III does not have adequate capacity, using current versions of utility support software, to reorganize our very large data base in a timely fashion. A vendor's software engineer estimated that this task would take about 10 days. Currently, secondaries account for 30% of the records in portions of the data base. This fact is resulting in unnecessary disk I/O operations. Decreasing and maintaining a reasonable amount of secondaries requires more disk space, which is an area where maximum capacity is also being approached. If this capability to reorganize remains unresolved, it will keep the system on a continually degrading course.

7. Some miscellaneous observations were also made; such as, very long searches for certain transactions (resulting in a high number of I/O operations), and inefficient algorithms for resolving synonyms in the data base (resulting in a 30% level of secondary records).

VI. SOLUTIONS

1. Disk I/O Contention

Table 2 describes the number of disk I/O's on each disk drive at the beginning of our investigations. These numbers were reported by the Event Monitoring Facility.

Drive	Disk I/Os	%
1	6,331	62.0
2	1,056	10.3
3	54	0.5
5	953	9.3
65	200	2.0
66	927	9.1
67	511	5.0
7	180	1.8
TOTAL	10,212	100.0

TABLE 2

The obvious disparity created overutilization of drive 1, resulting in contention, queues, and long delays on this drive.

This drive contained the system data set plus some application data sets. The trivial system response time (i.e., pressing the RETURN key) was about 30 seconds.

Reorganization of disk data sets, with an objective to decrease contention was recommended and implemented. Drive 1 was dedicated to the system data set. The result was an immediate response to the RETURN key. Since the production response time was of a magnitude higher, the result of this improvement was not measurable.

2. Memory Contention

The contention for space in the main memory leads to swapping. Each swap results in at least one I/O operation. In the heavily I/O bound workload the 7.2 swaps per second (measured by the Event Monitoring Facility) contributes to the I/O load.

The suggested remedy was to increase the size of memory. The original system had 1 Mbyte of main memory. Thanks to Hewlett-Packard's cooperation we were able to experiment with two 0.5 Mbyte increases.

The system with 1.5 Mbyte of memory decreased the amount of swapping to 22% of the 1 Mbyte swapping level. The second 0.5 MByte decreased the remaining swapping in half, but since swapping had already been decreased considerably, the second memory increment did not provide much improvement in performance.

Also, the hardware monitor utilization measurements supported these results. Table 3 describes all three steps. The CPU Busy and the Selector Channel Busy are basic measures. The third measure, Utilization Level[4], is an arithmetic sum of the first two measures. The Utilization Level was calibrated previously, and it was found that the maximum laboratory achievable value is 180 (MPE-III). The practical maximum utilization of the system was found to be 135.

Memory Size	CPU Busy	Selector	Utilization Level
1 Mbyte	61.8%	64.2%	126.0
1.5 Mbyte	49.2	49.2	98.4
2 Mbyte	43.2	45.0	88.2

TABLE 3

Although these comparisons were not based on results of a rigorous benchmark, an effort was made to compare data taken during periods of comparable workload execution. The type of transactions and number of users was comparable, and all measures were made on the same day, during the same shift.

The memory size testing was performed after some tuning had already been implemented, thus the original high utilization (over 130) was not reached.

3. I/O Operation

The high amount of I/O operation was attributed to the following:

- Disk I/O contention, as discussed previously.
- High swapping rate.
- Very large database operations.

The application design team implemented changes to alleviate the I/O bound workload. They were:

- Multiple copies duplication was moved from the disk to memory.
- Optional restriction of conditional search to a smaller subset.
- Elimination of unnecessary summary reports displayed on a terminal.
- Ability of the user to bring less information to the screen.

The results of this effort are shown in Table 4.

Date	CPU Busy	Selector	SEL/CPU Ratio
Jan. 6, 1981 (prior to changes)	58.3%	60.8%	1.04
Feb. 10, 1981 (after changes)	43.9	42.8	0.97

TABLE 4

4. System's Capacity

The capacity of HP3000 was exceeded by the operation of this data base. Based on previous monitoring on a simulated workload, the "Utilization Level" increased 9 units with each added terminal. This would permit a maximum load of 15 active terminals. It was observed in actual practice that as many as 28 users were active.

Having up to 28 active terminals on the system resulted in contention of all the system's resources (except Multiplexor Channel), and delays in long search response time of up to 30 minutes.

One of the initial changes, and perhaps the most significant to date, was to split the workload over 2 shifts. After this change, the number of users on each shift usually do not exceed 15. This decreased the response time for transactions requiring long searches to a maximum of 30 seconds (from the original maximum of 30 minutes). Other types of transactions (those not requiring long searches) now required only a few seconds.

VII. CALIBRATION MEASUREMENTS

1. MPE-III/MPE-IV Comparison

During the hardware monitoring sessions, we also discovered a shortcoming of the system which influences its capacity. It is the lack of the ability of the operating system MPE-III to overlap CPU and I/O activity. The HP engineers were aware of this problem and claimed that the MPE-IV operating system should rectify it.

Table 5 compares the hardware monitoring measurements of MPE-III and preliminary released MPE-IV operating systems. A single stream serial batch benchmark consisting of five steps was used for comparison. The first step is CPU bound, the second to fourth steps are CPU and I/O mixed with increasing I/O portion, and the fifth step is I/O bound.

	MPE-III			MPE-IV		
STEP	CPU	SEL	OVERLAP	CPU	SEL	OVERLAP
1	41.54s	8.04s	0.00s	41.54s	6.70s	0.00s
2	63.65	18.76	9.38	61.64	18.76	10.05
3	50.92	34.84	25.46	48.91	32.83	26.13
4	55.61	50.25	20.77	54.27	48.24	20.77
5	60.97	65.66	28.14	56.95	63.65	29.48
TOTAL	272.69s	177.55s	83.75s	263.31s	170.18s	86.43s

TABLE 5

The same benchmark job consumed 3.5% less CPU time and 4.5% less selector channel time, when run under the MPE-IV operating system. However, the most significant improvement is the ability of MPE-IV to overlap the CPU and the selector channel activities. The formula (1) describes the method of calculation of the Overlap Factor:

$$\frac{\text{CPU + SEL}}{\text{OVERLAP}} = \text{Overlap Factor} \qquad (1)$$

The improvement of the Overlap Factor under the MPE-IV operating system was 7.1%. Since the overlap consists of two values (CPU and Selector) only one half of the Overlap Factor improvement should be considered for improvement in the system's capacity.

Since this benchmark is serial in nature and does not heavily load the system, the improvement is marginal. Therefore, another measurement of the system, heavily loaded with a series of engineering test programs is described in Table 6. This time an average percentage utilization is described in time samples.

NUMBER OF PROGRAMS	MPE-III			MPE-IV		
	CPU	SEL	OVERLAP	CPU	SEL	OVERLAP
1	41%	65%	9%	70%	47%	47%
2	50	66	16	98	57	56
3	50	78	26	99	79	78
4	49	90	41	99	57	56
AVERAGE	48%	75%	23%	92%	60%	59%

TABLE 6

In this case, the system under MPE-IV could provide an average of 24% more capacity, and 39% in an extreme case (3 programs). The overlap factor calculated by using formula (1) improved by a factor of 2 above MPE-III.

It should be noted, that all measurements were conducted on a pre-released version of the MPE-IV operating system using a benchmark technique. The performance results on the final supported version in a production environment might be different.

2. "SHOWME" Calibration[4]

The CPU time measured by the hardware monitor is considered an absolute measure. This allows it to calibrate any software package.

Table 7 describes the CPU times of the hardware monitor (H/M) and the SHOWME software for both MPE-III and MPE-IV operating systems. Again all measurements were taken while the above described benchmark job was executed.

	MPE-III			MPE-IV		
STEP	H/M	SHOWME	CAPTURE RATIO	H/M	SHOWME	CAPTURE RATIO
1	41.5s	39s	0.94	41.5s	38s	0.92
2	63.6	60	0.94	61.6	57	0.93
3	50.9	45	0.88	48.9	40	0.82
4	55.6	47	0.85	54.3	41	0.76
5	61.0	50	0.82	56.9	42	0.74
TOTAL	272.6s	241s	0.88	262.2s	218s	0.83

TABLE 7

SHOWME has 88% capture ratios under MPE-III and 83% under MPE-IV operating systems. These factors must be used whenever SHOWME is used to define the actual CPU utilization.

VIII. SUMMARY

It is obvious that we did not measure all parts of the system with the hardware monitor, the main storage and buses for example. This would require much more effort in installation, operation, and analysis and additional benefits would be marginal. Also, the urgency of the project required swift action.

The hardware monitor was not the only tool used in this project. There were software monitors, accounting data, physical observation and chiefly the dedication and effort of all personnel involved, which helped to solve this difficult and complex problem. However, the hardware monitor did play a major role in pointing out otherwise obscure bottlenecks and inefficiencies in a timely fashion, and thus saving many manhours, which would have been used otherwise. It also helped to calibrate the software monitors so they can be used in the future instead of the hardware monitor, which is an expensive and labor intensive tool.

The capacity of the HP3000-III and IMAGE/3000 to maintain the large data base remains a problem. This problem should be resolved since it has the potential to become worse with regular operation of the data base.

The effort to improve the service level of the HP3000-III system is not yet completed. The improvement of response time from the maximum of 30 minutes to 30 seconds was only the first step. The next step will be capacity planning with an objective to maintain an acceptable service level as the workload grows. Installation of MPE-IV, further changes in the application, extension of main memory to 2 Mbytes, and possible installation of the Model 44 are some of the available actions for future consideration.

References

(1) Hewlett-Packard: "OEM Computer Products Catalog; 5922-0151(D)," November 1980.

(2) Tesdata Systems Corporation: "MS58 Computer Performance Measurement System Reference Manual", McLean, Va.

(3) Buzen, J. P: "A Survey of System Tuning Tools and Techniques; System Tuning", Infotech State-of-the-Art Report, pp. 229 - 241, Bershire, England, 1977.

(4) Loffler, I: "Capacity Planning for Minicomputers", INTELCOM 80, Los Angeles, Ca., November 1980.

(5) Wenig, R. P: "Effective Use and Application of Minicomputers", IMS, Inc., Framingham, Ma., 1979.