INTRODUCING THE HP ON-LINE PERFORMANCE TOOL (OPT/3000)

Robert L. Mead Jr.
Member of Technical Staff
Hewlett-Packard Company
Computer Systems Division

Robin P. Rakusin
Product Manager
Hewlett-Packard Company
Computer Systems Division

Clifford A. Jager
Project Manager
Hewlett-Packard Company
Computer Systems Division

INTRODUCTION

The question of whether or not a computer system is being effectively utilized is often difficult, if not impossible, to answer. Equally difficult can be the identification of a bottleneck when the performance of a system is less than expected. These difficulties typically arise due to a lack of information on which to base a judgement or decision. Even in those situations where information is available, it is often the case that information is incomplete, or possibly inaccurate or misleading, thus forcing the analyst to make a "best guess" as to the true situation. When detailed and complete information is available, it is frequently difficult to separate the useful information from the vast amount of data provided. In this paper we describe an interactive software product designed specifically to aid in the analysis of HP 3000 computer system performance, and which addresses the problems just described.

This product, the HP On-line Performance Tool (OPT/3000), is Hewlett-Packard's first performance measurement software product, and can be used to identify performance problems or bottlenecks, to characterize the workload on an HP 3000, to collect information required for capacity planning activities, to analyze system table configurations, and in some cases, to tune the performance of individual applications. OPT/3000 provides information in 23 separate interactive displays in the following areas: CPU utilization and memory management activity, memory usage, I/O traffic, program and process activity, and system table usage. Although each display is designed to be quickly and easily understood, the assumption is made that the user has been trained on the internal operation of MPE IV, the newest version of the HP 3000 Multiprogramming Executive operating system. OPT/3000 is designed to operate in conjunction with MPE IV and can be used on any HP 3000 Series II, Series III, Series 30, Series 33, or Series 44.

This paper presents an overview of the HP On-Line Performance Tool, and discusses some intended applications of OPT/3000. The information reported by OPT/3000 is also reviewed in-depth, as well as the techniques used to obtain the information.

OVERVIEW OF OPT/3000

The HP On-line Performance Tool is a software product that provides performance related information in an interactive environment. As mentioned earlier, OPT/3000 can generate 23 different displays containing performance related information, in addition to seven menu displays. These displays are grouped into six categories, called display contexts, each of which is associated with a different type of system resource. The six contexts are: Memory, CPU/Memory Management, I/O, Process, System Tables, and Global (a little bit of everything). Within each context, displays are available at successively greater levels of detail. This structure allows the user to progress from summary level information to more detailed information as the situation requires. In many cases, the summary level information is sufficient.

Once a display has been generated, it is automatically updated at periodic intervals, with the length of the time interval under control of the user. A display can also be updated upon demand, simply by entering a carriage return. All commands within OPT/3000 consist of a single ASCII character, and a different set of commands are available in each display context. Certain global commands are available in all contexts. In addition, the pound sign character (#) is used as an escape character to access a set of control operation commands. These commands perform such operations as changing the current display context and suspending the updating of the current display. With this simple user interface, the generation of a different display within the current context is accomplished via a single keystroke, and the generation of a new display within a different context with a minimum of three keystrokes. Menu displays are available within each context, and list the commands available within that context.

An extensive on-line help facility is also available as an integral part of OPT/3000. With this facility, documentation explaining any command or display can be quickly displayed. In many cases, interpretation guidelines are also provided to aid in the identification of performance problems.

OPT/3000 utilizes the features of the HP 264x series of terminals to generate displays with a graphical format, where practical. The terminal features used include the four available video enhancements (blinking, inverse video, underlining, half-bright), the line drawing character set, and the cursor addressing capabilities. OPT/3000 automatically checks to verify that an appropriate terminal is being used, and warns the user if an incompatible terminal is in use.

A hard copy of any display can be generated on the line printer (device class LP) with a single keystroke. The hard copy displays are similar in layout to the interactive displays, but some reformatting is necessary to convey the same information, due to the lack of video enhancements on a line printer (e.g. paper cannot blink).

Although the HP On-line Performance Tool is primarily designed for interactive use, it can be executed in batch mode to collect summary information about system activity. These summary reports can be used to provide data for capacity planning activities, and can be generated interactively as well. Once activated, the summary reports are generated independent of the interactively generated displays.

There is no limit on the number of copies of OPT/3000 which can be executing simultaneously. OPT/3000 obtains much of its information via a new internal measurement interface facility incorporated within MPE IV. This facility maintains a set of measurement counters accessible by multiple users. Additional information concerning the measurement interface, and the techniques used by OPT/3000 to collect information, will be discussed in a subsequent section.

APPLICATIONS OF OPT/3000

There are several anticipated uses for the HP On-line Performance Tool. Among these uses are the identification of performance problems and bottlenecks, the analysis of system table configurations, characterization of the system workload, capacity planning, and performance tuning of applications. Each of these activities utilizes some or all of the capabilities of OPT/3000. We will now briefly discuss each of these application areas before describing in some detail the information provided by OPT/3000.

The ability to quickly move between displays and the variety of information available through OPT/3000 facilitates its use in identifying performance problems and bottlenecks. In particular, it is expected that a system clearly bottlenecked by CPU, memory, or I/O will be quickly identified. OPT/3000 can also be used to determine if disc accesses are unbalanced between multiple drives. Poorly behaved application programs can also be identified, in terms of programs which use excessive numbers of files and extra data segments and those which waste stack space.

A second application area is that of system table configuration analysis. Inappropriately configured system tables can degrade system performance, either by wasting memory if the tables are unnecessarily large, or by causing processes to delay while waiting for an entry in a table that is configured too small. In the latter case, system failures may also result if the table size is exceeded. OPT/3000 allows the user to quickly identify those tables which are not properly configured, and to determine (through utilization statistics) a more appropriate value.

The characteristics of the workload on an HP 3000 can be determined using OPT/3000. The names of all active or allocated programs on the system can be easily determined, as well as the users of each program. The CPU usage, disc I/O rate, and memory usage characteristics of an individual application can be determined if the application is running stand-alone on the system.

The summary reports which can be generated by OPT/3000 in either batch or interactive mode can be used to provide data for capacity planning activities. These reports indicate the CPU usage of the system, memory management activity, and the I/O traffic on individual discs, line printers, and magnetic tapes. The information used to generate the summary reports can also be logged to an OPT/3000 log file (disc or tape), which could be processed to provide input for generating plots. In this manner, OPT/3000 can gather trending information that can be used to determine when additional peripherals or systems are needed, or to detect changes in the day-to-day processing load.

Although OPT/3000 is oriented towards the measurement and analysis of the system as a whole, it can be of some value when tuning the performance of individual applications. In particular, OPT/3000 can provide information relating to an application's usage of files and

extra data segments, plus detailed information about the application's use of its stack. CPU usage information is also available.

INFORMATION PROVIDED BY OPT/3000

As mentioned earlier, the HP On-line Performance Tool provides information in six different display contexts: global, memory, CPU/memory management, I/O, process, and system tables. In this section we describe the types of information available in each context. In general, the information provided by OPT/3000 can be divided into two basic classes. The first class of information shows the state of some aspect of the system at the moment the display update is generated. Examples of this class of information include the current contents of main memory and the current list of active programs. The second class of information summarizes activity within the system during some interval. CPU utilization and disc I/O rates are examples of this second class of information. In most cases, this summary class information is reported for two types of intervals: the interval between the previous display update and the current display update, and the interval encompassing all update intervals since the start of OPT/3000 execution. These two intervals are herein referred to as the current interval and the overall interval, respectively. The user can clear all totals associated with the overall interval at any time in order to start a new overall interval. We will now discuss the information available in each of the display contexts.

Global Context

The global context is automatically entered upon execution of OPT/3000, and it provides summary level information concerning CPU usage, memory utilization, disc I/O rates, and process activity. The generation of summary reports is also controlled from the global context. The two displays in the global context can be used to quickly determine potential problem areas (e.g. memory bottleneck), and then more detailed displays in the other contexts used to isolate and verify the problem at hand. The global context can also be used to monitor general system activity, in order to detect fluctuations in resource usage. The CPU and disc I/O information summarizes the activity for both the current and overall intervals, whereas the memory and process information describes the situation at the time of the update.

Memory Context

The memory context consists of eight displays, and provides information related to the usage of main memory and segment sizes. Three of the displays provide information related to the current contents of memory and the remaining displays consist of histograms depicting distributions of segment sizes or free areas in memory. The highest level display concerned with the contents of memory allows the user to determine the current percentage of memory containing code segments, stacks, and extra data segments. Additionally, the user can determine the type of code and data segments in memory. For example, the user can determine the percentage of the extra data segment memory usage that is due to IMAGE/3000, KSAM/3000, the file system, or system tables. Likewise,

code segment memory usage is separated into segments originating from program files, and those from segmented libraries.

The user is also able to generate an image of the current contents of main memory, either for all of memory or for a single bank (64K words). This image indicates the type of each segment (e.g. file system data segment, stack, program file code segment), the approximate size of the segment (in either 1K or 64 word increments), and other miscellaneous information about the segment (e.g. is it locked or frozen?, is it an overlay candidate?). These images are generated by utilizing the display enhancement capabilities of the HP 264x series of terminals, and consist of a sequence of alternating white and gray rectangles (generated using inverse video and half-bright). Each rectangle represents an individual segment.

The histogram displays depict the distribution of segment sizes, in either 1K or 512-word increments. The highest level display depicts separate distributions for code, stack, and extra data segments. The remaining four histogram displays generate higher resolution histograms for each of the above three segment types, plus one for free areas in memory. The histograms are generated using the line drawing character set of the terminal, so as to provide maximum resolution.

CPU/Memory Manager Context

The CPU/memory manager context includes three displays with information related to CPU usage and memory management activity. The highest level display provides information about both CPU and memory management activity, while the remaining two displays provide more detailed information about each of these areas. All information provided in this context is of the interval summary class, with information for both the current and overall intervals.

The CPU information provided allows the user to determine the percentage of time the CPU is in various states, as well as the rate at which processes are being allowed to execute in the CPU. The reported CPU states include CPU busy executing processes, CPU time for memory management, CPU time on background memory "garbage collection", CPU on overhead processing (e.g. handling interrupts, dispatcher time), CPU waiting for user disc I/O to complete, CPU waiting for memory management I/O to complete, and CPU idle. Information for process launches and process preemptions is reported as the number of occurrences of the event per second (i.e. as a rate). Reported rates include the current interval, the overall interval, and the maximum rate observed in a single interval since the start of the overall interval. These three rates are depicted with a one-line bar on the terminal screen, utilizing inverse video and half-bright inverse video to indicate the current and maximum rates, plus an asterisk to denote the mean rate over all intervals.

Event rate information is also reported for memory management activity in this context. The events reported include memory allocation, memory management disc I/O write, memory management disc I/O read, release code

segment from memory, and release data segment from memory. Information is also available concerning how the memory manager satisfies requests for absent segments. When a segment absence fault occurs in MPE IV, the algorithm used by the memory management routines can terminate with one of five possible outcomes, ranging from recovering the segment from the list of overlay candidates to temporarily postponing the request to avoid thrashing. OPT/3000 shows the percentage of memory allocation attempts terminating with each of the five outcomes. These percentages are shown for both the current and overall intervals, utilizing a bar with alternating white and gray areas.

I/O Context

The I/O context provides four displays regarding I/O completion rates for discs, line printers, and magnetic tapes. The highest level display indicates the I/O completion rate per second for each type of device, for both the current and overall intervals. The remaining three displays provide more detailed information about individual devices within each device category. These displays indicate the completion rate for three types of I/O operations (read, write, and control) on each individual device. This information can be used to determine if the I/O traffic is balanced between the devices on the system, or to identify times of peak activity.

Process Context

The process context includes four displays with information concerning process and program activity on the system at the time the display is updated. The highest level display provides information about all active or allocated programs. This information includes the fully-qualified program file name, the size of the program file in words, the number of segments in the program, the number of current users of the program, and limited working set information.

Once the above display has been generated, a second level display can be used to determine more detailed information about each process sharing a program file (or for system processes or command interpreter processes). The more detailed information includes the user name and account of the user, the process number (PIN), the size of the process stack in words, the CPU time used by the process, the number of open files and extra data segments, and the job/session number.

Additional information about a specific process can then be obtained by generating a third level display. This display contains all of the information present in the second level display, plus more detailed information about how the process is utilizing its stack space (e.g. the size of the DL area, size of the global data area). Also included are the names of all open files, a list of son processes, and a list of explicitly obtained extra data segments and their sizes.

Some of the information reported in the second and third level displays could be used to circumvent the security aspects of MPE. For this reason, these two displays cannot be generated by all users of OPT/3000. The security provisions within OPT/3000 allow a user with either system manager (SM) or operator (OP) capability to generate these two displays for any program file. Any other user can only generate the displays for a program file if they are the creator of the file, or are the account manager for the account in which the program file resides.

The remaining display in the process context provides information about the number of processes in various states. For example, the total number of processes waiting for blocked I/O, number of processes waiting for RINs, and the number of processes in the dispatch queue are reported. This information indicates the state at the time the display is updated, and no averages or totals over time are reported.

System Tables Context

The system tables context contains two displays indicating the current and maximum utilization of configurable system tables. One display provides only the current and maximum utilizations in a graphical format, using inverse video and half-bright inverse video bars. For almost all tables reported, the maximums are for the time since the last system warmstart. For the remaining tables, the maximum is that observed by OPT/3000. The second display provides more detailed information in a tabular format. This detailed information includes the configured number of entries, entry size, and maximum utilization observed by OPT/3000, as well as the current and maximum table utilizations.

MEASUREMENT TECHNIQUES

The HP On-line Performance Tool obtains the information used to generate its displays from two basic sources. The first source is the new internal measurement interface facility within MPE IV, and the second is internal MPE data structures and tables. The measurement interface provides all information related to CPU usage, memory management activity, and I/O traffic. All other information reported by OPT/3000 is obtained by examining internal MPE tables and data structures.

The measurement interface facility in MPE IV provides OPT/3000 with a formal mechanism for accessing instrumentation within MPE IV. When the facility is enabled by OPT/3000, the measurement interface obtains an extra data segment to be used as a set of counters. This segment is then locked and frozen in memory and its location stored in a global cell. As events occur, the appropriate counters within the extra data segment are incremented by code within MPE IV, and accessed in a consistent manner by OPT/3000. The CPU state time information is maintained in a similar fashion. OPT/3000 determines the activity during an interval by comparing the current sample to the previous sample, and computing the change in each counter. A count of the number of processes that have activated the interface is maintained by MPE IV. When the count falls to zero, the extra data segment is released and the instrumentation disabled. This mechanism allows multiple copies of OPT/3000 to use the same shared instrumentation. As MPE continues to evolve, both the measurement interface facility and OPT/3000 will be

modified to reflect any changes within MPE.

The overhead within MPE for maintaining the counters has been determined to be approximately 0.3 to 0.8 percent of available CPU time, depending upon the amount of activity within the system. OPT/3000 can collect data from the extra data segment, and update all of its internal totals with the change in each counter in approximately 40 milliseconds. As can be seen from this data, the measurement interface facility provides a very low overhead method for obtaining performance information.

All other information reported by OPT/3000 must be obtained by examining internal MPE data structures and tables. The information concerned with the current contents of main memory is obtained by scanning all of memory, examining each region and sub-region header (these are similar to the memory links in MPE III). The segment size histograms are produced by processing the segment tables. The information concerning program files and processes is obtained by examining the loader segment table directory, process control block table, and the process control block extension area in the stack of a process. System table utilization information is partially obtained by examining information maintained in the header portion of each table.

The overhead required to gather any of the information just mentioned varies depending upon the system configuration. In general, the CPU time required to collect the necessary information and update any display ranges from 300 to 800 milliseconds, depending upon the display.

This normally translates into a total CPU overhead for OPT/3000 ranging from 1 to 3 percent of available CPU time, depending upon the displays generated and the frequency of display updates. An update interval of 15 seconds is the default used, and results in overhead in the 1 to 2 percent range.

SUMMARY

The HP On-line Performance Tool is part of Hewlett-Packard's integrated approach towards offering HP 3000 users alternatives in performance measurement analysis. In addition to OPT/3000, the first HP 3000 software performance measurement product, a new System Performance Evaluation package and a new MPE Internals and System Performance Analysis course are being offered for HP 3000 users.

The recently introduced HP 3000 System Performance Evaluation Consulting package offers an alternative to OPT/3000 for HP 3000 users who want system performance analysis conducted by HP Performance Specialists. These Specialists have in-depth training on the internals of MPE and on the performance characteristics of the HP 3000. They also have a number of HP-supplied software tools at their disposal, such as OPT/3000, IOSTAT, and the MPE IV Data Collection Program (MPEDCP), for collecting and analyzing performance measurement information on the HP 3000.

A new MPE Internals and System Performance Analysis training class is being offered in conjunction with the HP On-line Performance Tool. The first part of the course discusses the areas of MPE IV that are necessary for understanding the performance measurement information presented in OPT/3000, in particular, the new MPE IV memory manager, the dispatcher, scheduler and I/O areas in MPE IV, and the process structures. The second part of the course reviews the interrelationships of the performance measurement variables discussed in the first part, and presents operational guidelines for OPT/3000. In addition, case study workshops will be used to share HP Performance Specialist techniques and experiences with class participants.