

Presentation #4015
Performance Metrics "Cheat Sheet"
by David Hesser, Hewlett-Packard Company
(404)648-3773

In the several years, as I have focused on performance and performance tools calls here in the Response Center, I have been asked on a number of occasions, when I have made an observation or a recommendation, "Is that documented anywhere?"

Unfortunately, the answer, in many cases, is no. The documentation for some of the performance metrics on the MPE side of the house has gotten somewhat out of date during the past several years, especially as we have grown into larger and larger systems with user counts approaching and in some cases, exceeding 1,000. This is an effort to explain how some of the metrics in the SCOPE log files and in GLANCE can be used to make reasonable performance decisions.

These explanations are not out of the lab, and they should not be taken as the gospel according to HP. They are simply some of my own personal observations, developed after looking at a few hundred systems over the past couple of years.

Also, please put these thoughts into the context of time. When we talk about GLANCE, we are talking about a relatively tight granularity...30 seconds to 1 minute, but over a relatively short time frame. Not many of us have the patience to watch a GLANCE screen for hours or days at a time. It is a real-time performance diagnostic, designed to answer the question "What is going on with my system right now?"

With the SCOPE collector, we are looking at a granularity of 5 minutes to an hour, but our time horizon can be much larger...a day, a week, a month, or even a year. The "averages" we get from SCOPE can be very valuable, especially if they are high. Remember, in order to get a high average, you are either running consistently at that level, or else you are running above it for the half time. Put another way, when averages reach 80%, there is probably some time at 90 or 100% to offset lower values. The lower values don't "help" performance as much as the higher values "hurt" it.

Finally, please understand that my numbers are typically considered "liberal", especially by marketing types. The reason is that I am trying to address a performance bottleneck, not save a customer a few dollars in equipment costs. Besides, which costs a company more...another 128 MB of memory or 100+ online users waiting for a response?

CPU...how much is too much? When should we upgrade?

The number 80% will be seen throughout this article, pretty much anytime a resource server is not involved. If you go back into the world of queuing theory, you will see that if a server is busy 80% of the time servicing user request, it will continue operating at pretty close to its peak efficiency. Beyond that point, the overhead operations (dispatching, handling interrupts, etc.) becomes a greater and greater factor. If we look at the fields that are available from the SCOPE log files, we could say that CPU_SESSION and CPU_JOB represent the work being done for the user, CPU_SYSTEM, CPU_DISPATCH, CPU_ICS, and CPU_IDLE represent the time the computer is not doing work either because it has no work to do or because it does not have the required pages in memory to do it. These same fields are available from within GLANCE by looking at the CPU DETAIL screen.

Performance Metrics "Cheat Sheet"
4015-1

If we see that the CPU used for sessions and jobs is CONSISTENTLY exceeding that magic 80% level, we can probably say that the system is bottlenecking on CPU. But is that truly a problem? Now

we come back to the statement that performance, like beauty, is in the eye of the beholder. Remember that jobs will take all available CPU, since they typically never have to wait for user input. The CPU bottleneck that we will be seeing will be meaningful only if there are user complaints about performance, or if faster turnaround on batch jobs is required. If there are response time problems with specific interactive sessions, we might consider Workload Mgr. as a tool that could make more CPU available to that class of user. If batch throughput is the issue, then we probably need to make more CPU power available (translation...upgrade).

An alternative way of looking at the CPU would be to group the system CPU usage (CPU_SYSTEM, ICS, DISPATCH, and MEMMGR) with the CPU_SESSION metric. If that number is CONSISTENTLY above the 75 to 80% level, then we have a CPU bottleneck, and, since batch work is not in the equation, that bottleneck probably needs to be addressed now.

One quick message here to those who have multi-processor systems with greater than 4 CPU's ...we get a number of calls requesting that "BYCPU" metrics for processors 5 through "n" be added to the SCOPEXL metrics. While that is certainly a reasonable request, the data probably would not be of significant use. The 3000 uses an SMP (symmetric multi-processing) algorithm to balance CPU usage among all CPU's, and I would expect to see that the BYCPU metrics for processors 5 through "n" would be very much the same as for processors 1 through 4. This would be a much more valid request for the HP 9000 family which uses processor affinity as part of its CPU assignment algorithm.

MEMORY...how much do we need, how much do I have? How do I tell if I need more?

Let's take the second question first...there is a bug in the MEMMAP program of SYSDIAG on 5.0 that MAY not show you your correct memory size. Use the number for the GLANCE MEMORY DETAIL screen. That number is correct. The rule of thumb I use when configuring memory is as follows:

20 MB for the resident OS

16 more MB if any appreciable networking...VT sessions, SHAREPLEX, SNA, etc.

1-2 MB per session...

1 MB if 3GL (COBOL, PASCAL), VPLUS, TurboIMAGE, NM, etc.

2 MB if 4GL (COGNOS, ORACLE)M SQL, CM, etc.

2-4 MB per active job...same criteria as for sessions.

All of this is with a minimum of 64 MB. Some people say a minimum of 96 MB.

What metrics do I use to look at memory. Historical guidelines say to use memory manager CPU and memory disc i/o's or page fault rate. Today, I tend to use memory manager CPU (CPU_MEMMGR), along with the clock cycle rate (MEMMGR_CLOCK_CYCLES), and system library page faults (NMLIB_FAULTS & CMLLIB_FAULTS). All of these metrics are available both within GLANCE and from the SCOPE log files. The reason that I do not like to use the system page fault rates any more is twofold. First, it would appear that memory mapped files (TurboIMAGE datasets, etc.) will tend to show up as page faults and memory mgr. i/o's.

Performance Metrics "Cheat Sheet"

4015-2

Arguably, an increase in memory size can cut down on those i/o's so I don't have much of a problem with that. The problem with this i/o metric comes into play on very large systems with a large and dynamic user community. By large, I mean in excess of 200 to 250 users. By dynamic, I mean lots of logons and process creations within logons. A process creation, by its very nature involves a certain number of page faults. Many of the pages that are involved in these page faults literally did not exist just seconds earlier. Because of that, we could have a system with "gobsabytes" of memory, and if the process creation rate is still high, so also will be page fault rate.

Bottom line, then, is that if I see a high page fault rate, especially a high transient page fault rate, in the Memory Detail screen of GLANCE, I will tend to go back to the GLOBAL screen to get a feel for the process creation activity before using this metric to make a judgment on memory size. Neither GLANCE nor SCOPE have a specific metric that represents process creation rate, although it can be calculated using exported SCOPE metrics. If I feel that I can contribute high memory management activity to new process creations, then the addition of memory will probably not affect my system performance.

DISC SPACE...How much for TRANSIENT? How much should I have available? Does fragmentation make a difference? How should I configure PERM & TRANSIENT?

Again, let's take the second questions first. My rule of thumb is pretty simple...when free space gets to 15%, it's time to start thinking about doing something, when it gets to 10% it's time to start doing it.

So how much TRANSIENT space do I need. My experience on 5.0 leads me to believe that the system will use about 6-7 MB (25 to 27 K-sectors) per "user", where a user would be either a job or session. Jobs and sessions that use process handling and sessions that are logged on using VT over the LAN will tend to push that number up, while sessions that spend time looking at the colon prompt without going into a program or subsystem will tend to push that number down. Also, the greater the number of users on the system, the less that MPE will factor into that number, and vice versa.

Given that knowledge, how should we configure our drives. Well, user volumes are easy...just configure them at 1000, since there will never be transient space used on a user volume. Of course, if you want to keep a safety margin, you might configure 980 or something along that line, but part of system management is disc space management. Ldev 1 must have around 350 K-sectors, "reserved" for transient space during the boot process. That is around 25% of a 7933 drive...ever wonder where those default numbers of 75 75 came from? At any rate, configure Ldev 1 accordingly, keeping at least 350,000 sectors "reserved" for transient. On an EAGLE drive, that would be 83 83...on a 2GB drive and above, it can be around 95 95 (give or take a percentage point). Once you get beyond ldev 1, figure out how much space you want to reserve...for 250 users and 20 jobs, for example, I will want to reserve around 7 million sectors for transient space. On a system with 10 2GB drives then (78 million sectors), I might make a first cut at configuring PERM and TRANSIENT at around 91 91. That would "reserve" 9% for transient, 9% for PERM, and leave the other 82% up for grabs, first come, first served. That is assuming, by the way that all 10 of those drives are in the system volume set.

And what difference does disc fragmentation make on my system? There is no doubt that makes much less of a difference on MPE/iX system that it did (or does) on MPE/V systems.

Performance Metrics "Cheat Sheet" 4015-3

MPE/iX files can have virtually an unlimited number of different sized extents, while MPE/V files were limited to 32 extents, with all extents, except the last one, the same size. Fragmentation does make a difference, though, at least in my experience here at the RC. One of the most effective pieces of MPE/iX in terms of i/o performance is the prefetch algorithm. This allows the operating system to decide if a process is going to need more data in the same general vicinity of the disc, and to read in a larger than requested amount of data to cut down on the need for future physical i/o's. The problem here, though, is that the prefetch algorithm will not read across an extent boundary. If disc fragmentation, then, is causing us to create files with many small extents, the disc prefetch algorithm will be severely hampered in its effectiveness when it comes time for us to read the file. This will result in higher physical i/o rates and potentially higher pause for disc metrics.

So what do we look for in terms of fragmentation. Again, a rule of thumb that I use is to look at two metrics in the DISCFREE A output, the largest free area or LFA, and the total free space or TFA. The break point that I look for in disc fragmentation is basically when LFA is 50% or more of TFA. This is somewhat of an arbitrary value, I understand. I look at systems where LFA is 5% or less of TFA, and those are "obviously" fragmented. On the other hand, I've seen systems where LFA is 90% of TFA, and they're in great shape. I just picked the 50% value as an indicator. Preemptively running VOLUTIL CONTIGVOL or one of the third party disc defrag tools (Lund and Tymmlabs have tools, I think) can help keep the discs in a healthy state. There are SCOPE metrics that address disc space issues, TOTAL_FREE, LARGEST_FREE, PEAK_TRANSIENT, etc., and these may be useful, but they are system-wide metrics and are not collected on a per-disc basis. DISCFREE A and DISCFREE C continue to be my tools of choice for collecting information on disc fragmentation and overall disc space availability.

DISC I/O Rates...how do we recognize an i/o bottleneck

This is probably one of the tougher nuts to crack on today's 3000. The reason is that a perceived i/o bottleneck could, in reality, be a memory bottleneck. Some folks like to look at i/o rates and do some extrapolation to see if the system's disc configuration can handle the load. I prefer to look at the disc utilization metric. Disc i/o rates can be deceiving, especially in these days of f/w SCSI, HPFL, and (heaven forbid) HPIB. The utilization metric does not take drive speed into account, though. It is simply a metric of what percent of the time the drive was involved in an i/o operation. If we think of the disc as an "i/o" server, that takes us back to the 80% number again...if a disc has a utilization percentage consistently of 75 to 80% or greater, then we are probably seeing an i/o bottleneck forming around that disc. If those are shown to be memory mgmt i/o's then the root cause of the bottleneck could be memory, but we'd have to take a closer look.

By the way, I recently had a question about why you would see memory mgmt i/o's on a user volume. Remember that i/o's to and from memory mapped files will show up as memory mgmt i/o's as well as paging in of program file code pages. Those files could be on a user volume as well as a system volume. There is more to memory mgmt i/o's than just transient space.

Going back to disc utilization...you can get those metrics from GLANCE in the DISC DETAIL screen. Or you can get them from SCOPE. The data item name of DISC_UTILIZATION in the global report file will give you the utilization of the busiest drive on the system. If that number is low, then you probably are not experiencing any disc bottlenecks. If that number is high, you don't know if the problem is one of a single drive running high or if there are more drives running high, and this is just the highest. In order to get more info, you will need to explore the global item, @DISC_UTILIZATION, which will allow you to get utilization metrics for each disc (up to 64) on the system.

Performance Metrics "Cheat Sheet"

4015-4

Other SCOPE fields that can point towards a potential i/o bottleneck include CPU_PAUSED and DISCQUEUE. CPU_PAUSED can also be seen in GLANCE, but the DISCQUEUE metric can only be accessed from the SCOPE log files.

This is probably a good time to discuss the disc array dilemma...Disc arrays were invented for several reasons, but none of them were higher performance. I like to point to a ridiculous example like comparing a 5.4 GB array with 10 537 MB Eagle drives. If we assume that the 10 Eagles were appropriately separated on a number of channels, in other words, only the individual drives were the limiting factor in i/o's, we could comfortably run 300 i/o's per second through those 10 drives with their 10 controllers, but we can only run about 50 i/o's per second through the single controller of the 5.4 MB drive...a perfect opportunity for an i/o bottleneck to develop.

A Final Note...

A lot of times you may see where you can add 5% more performance by doing this or that. Before you invest in the time and effort, though, make sure that you are getting a good return. A 10% "performance improvement" may sound like a lot, but when you think about it, it's the difference between a 5 second response time, and a 4.5 second response time. It's the difference between a one hour batch run and completing the same run in 54 minutes. It's the difference between 200 users and 220 users. Some of those differences may be significant, but other may not be. Just be sure that the results you are looking for can be met by the performance improvements you are proposing.

A Final, Final Note...

The metrics that are made available to us through SCOPE collector can provide us some excellent insight into the resource utilization on our HP 3000 systems. Many users and system managers have only used LaserRX to display this information, but, while LaserRX does focus on some of the major metrics, it barely scratches the surface when

looking at the kinds of data that can be available. When I teach a customer performance class, I regularly ask students for performance and resource utilization questions that they would like to have answered as they apply to their systems. The students' initial "limiting" question typically is "what kind of data is available?" When they get beyond that limit and start thinking "outside the box," I get questions such as:

1. What is the average CPU per transaction for the interactive programs that were run yesterday?
2. How many logons and logoffs were done during the past 24 hours?
3. What was the system's process creation rate during times of peak CPU and disc utilization?
4. What were the average i/o rates by i/o channel?
5. What is the relationship between top disc utilization, disc queue length, and pause for disc?
6. What is the breakdown of CM to NM CPU usage on the system?

The answers to some of these questions are more obvious than others, but each of them can be answered using data from the SCOPE log files. Some (question 2, for example) can be answered from other sources (system log files), and some (question 4, for example) need input from other sources, but the data is there, available by exporting to an ASCII file and then reporting or downloading to the PC graphics package of your choice, or by extracting and feeding into PerfView.

