Living with 400 I/O's per second

A Robelle Tutorial
FLORUG HP Performance Training Seminar
Hutchinson Island, Florida

February 15, 2001

robelle solutions technology

1

The slowest link in the performance chain will always be physical mechanisms, i.e. disk and tape drives. Where CPU transactions are measured in millions or billions of instructions per second and logical memory I/O's measured in tens of thousands per second, physical I/O's are still measured in dozens, or perhaps hundreds.

Neil Armstrong, Senior Programmer in Robelle's R&D lab, explains software and database strategies for improving I/O performance, but he also details his explorations of HP's new PCI backbone and what it suggests for increased I/O performance.

Robelle Solutions Technology Inc.
Suite 201, 15399 - 102A Avenue
Surrey, B.C. Canada V3R 7K1

Toll-free: 1.888.robelle
Telephone: 604.582.1700
Fax: 604.582.1799
E-mail: support@robelle.com
Web: www.robelle.com

For Techies

References

What's Inside

	Page
Introduction	2
Ch..Ch..Ch. Changes	3
IO. IO It's off to Disc we go	4
Let's Get Physical	5
Measurements Please	8
Memory Manager	9
We Got Class	11
Promise of PCI	12
How's That for Performance	15

robelle
solutions technology

2

This paper will discuss new versions of MPE and Image and their impact on the HP 3000 and how the new A and N-class servers can meet the demands of the new expanded limits within MPE.

For Techies

References

Introduction

- New File System Capabilities
  - Large Files
- New Version of Image
  - Larger Jumbo Datasets
  - Large non-Jumbo Datasets
- New Servers
- New I/O Subsystem

robelle
solutions technology

3

Over the last eighteen months, the MPE arena has seen a flood of system enhancements. With the introduction of MPE/iX 6.5, we have the ability to create large fixed files up to 128 Gb. This limit, for all practical purposes, was just a practical one: how to test the larger files! I'm told that in order to increase the limit, it was a single line change and a re-compile of the file system.

With the release of Express 2 of MPE/iX 6.5, the C.09.02 version of IMAGE was released. This version changed the internal record number format in a way that removes the limit of 80Gb for a Jumbo dataset. For all the gory details see my article on our web site at

`http://www.robelle.com/tips/big-image2.html`

There have also been discussions of having these huge datasets be a single file, rather than several files in the HFS space (as in the case with the current Jumbo datasets).

In conjunction with all of this software development, another project was under way at HP CSY. The result of this work was announced on February 1st, with the new A and N Class servers. These new servers required that the entire I/O subsystem be re-written.

We will discuss some of the ramifications of these new enhancements and the promise of the new servers in terms of IO performance.

For Techies

References

How MPE uses Disc IO

- Files and Pages
- Swapping from Disc to Memory
- Posting from Memory to Disc

robelle solutions technology

4

Files on disc are "mapped" into "pages" of virtual memory, which are swapped into and out of RAM memory. For example, a run of Suprtool requires that as you execute the code in the program file, those pages must be in memory (requires a disc read). The program's data stack also must be in memory, but since it is modified it requires a read in and a write out (when swapped out to make room for something else).

Another example. As your program references or read records, IMAGE, behind your back, figures out which page the records are in and swaps the page in to memory. If you modify the record, the page is marked as dirty and EVENTUALLY the page is posted to disc.

If the dataset is many times larger than memory then the system will have to swap portions (or pages) of the file in and out of memory to suit the needs of the system. It has to keep track of which pages are no longer being used, which are dirty and need to be posted.

The less swapping of pages into and out of real memory, the faster the response time and throughput of the system.

For Techies

References

How Fast Are Discs?

- Discs are Physical Media
- Inherently slower
- Positioning Measurements vs Transfer Measurements
- Measurement Scale
- Disc Speed vs Clock Speed

robelle
solutions technology

5

Discs are a physical medium and as such. there are certain physical limitations imposed. There are two types of disc performance measurement for any device; these are Positioning Measurements and Transfer Measurements.

Positioning Measurements are those measurements associated with the physical movements of the disc in order to perform the read or write operations. The typical Positioning Measurements are Seek Time, Latency, Settle Time, Command Overhead Time and Access Time.

Transfer Measurements are those associated with getting the data from the heads, once they are positioned, into the disc drive's internal buffers and subsequently over the interface and into memory.

Many of the physical measurements are expressed in milliseconds, which, in comparison to the clock speed of a processor, is an enormous amount of time. For example the seek time for an extremely fast SCSI drive is about 4msec, while the high end N-class servers are 550 Mhz, which if driven at it's theoretical maximum will execute half a million instructions in one msec.

For Techies

References

Positioning Measurements

- Positioning Measurements
  - Seek Time
  - Latency (Speed of Disc)
  - Settle Time
  - Command Overhead Time
- Key Measurements
  - Seek Time
  - Latency

robelle
solutions technology

6

Seek Time is is the time required for the read/write heads to move between tracks over the surfaces of the platters: this is a physical movement with certain overhead associated with it. Typically, the number reported for seek time is the average seek time which is the movement from one random track to any other,

Disc speeds are traditionally reported in revolutions per minute. Some typical numbers reported these days are typically 5600, 7200 and even up to 10000 rpm. The higher this number, the lower the latency of the disc. Latency is the time that the read/write heads take to move to the correct sector once the heads are positioned on a given track. Latency is usually reported as average latency which is the amount of time it takes the disc to turn one half rotation. Average Latency can be calculated using a simplified formula of 30000/Spindle speed. Using this the average Latency for a disc spinning at 10000 rpm is 3.0 msec.

The *settle time* specification (sometimes called *settling time*) refers to the time required, after the actuator has moved the head assembly during a seek, for the heads to stabilize sufficiently for the data to begin to be transferred.

Command overhead refers to the time that elapses from when a command is given to the disk until something actually starts happening to fulfill the command

For Techies

The other 2 Seek Time metrics are Track-to-Track, (the amount of time to move between adjacent tracks) and Full Stroke, which is the amount of time to seek the entire width of the disc from the innermost to the outermost.

Sometimes disc specs report a Worst Case Latency number, which is the amount of time it takes for one full rotation.

References

Transfer Measurements

- Transfer Measurements
  - Internal Media Transfer Rate
  - Head Switch Time
  - Cylinder Switch Time
  - Internal Sustained Transfer Rate

robelle
solutions technology

7

The *internal media transfer rate* of a drive (often just called the *media transfer rate* or the *media rate*) is the actual speed that the drive can read or write bits on the surface of the platte. It is normally quoted in megabits per second, abbreviated Mbit/sec.

Switching between heads in a cylinder is a purely electronic process. However, it still requires time, called the *head switch time*.

Cylinder switch time occurs when the drive finishes with all the data on a given cylinder and needs to switch to the next one. This normally only occurs during fairly long reads or writes, since the drive will read all the tracks in a cylinder before switching cylinders. A cylinder switch is slower than a head switch, because it moves the actuator assembly, around 2 to 3 milliseconds.

For real-world transfers, we want the rate at which the drive can transfer data sequentially from multiple tracks and cylinders. This is the drive's *sustained transfer rate* (sometimes the *sequential transfer rate*), abbreviated *STR*.

An example: let's say we want to read a 4 MB file from a hard disk that has 300 sectors per track in the zone where the file is located; that's about 0.15 MB per track. If the drive has three platters and six surfaces and if this file is stored sequentially, it will on average occupy 26 tracks over some portion of 5 cylinders. Reading this file in its entirety would require (at least) 25 head switches and 4 cylinder switches.

For Techies

STR is based upon the drive's media transfer rate, but includes the overheads required for head switch time and cylinder switch time. STR is measured in *bytes*, not *bits*, and includes only data, not the overhead part of each sector or track.

STR is most relevant for reflecting the drive's performance when dealing with largish transfers.

References

How Can I Measure

- Glance/iX
- SOS
- DiskPerf
- Suprtool Test Environment
  - Set Stat on
  - Run Fflush (available at www.allegro.com)
  - Check with Klondike

robelle
solutions technology

8

You can measure a system's disc performance in a number of ways and I typically use all of them. When starting on a new machine, I use Glance/iX. It is usually available and tells me things such as how much memory is available, the type of system it is, and how the system is generally configured. It also shows rates of disc I/O for the system. SOS is a similar third-party product which will also gives disc I/O metrics.

There is a new product from Allegro called DiskPerf, which lets you check the relative speeds of your disk drives. DiskPerf identifies older, slower disks, measures the true cost of your RAID level, and helps identify hardware and configuration problems.

When doing Suprtool tests, we try to insure that the I/O subsystem is stressed, as opposed to memory. I always turn the Statistics feature of Suprtool on with the Set Stat On command. Before a test, I flush the test files out of memory with a utility called Fflush. I then check that the file I am about to read doesn't have any pages in memory using the Klondike utility. This way I am certain that all tests on a file are the same, or at least I am narrowing the playing field. Once the tests are complete, I can review the statistics and look at the progress messages to look for any problems and compare the total number of records read against the overall Wall Time.

For Techies

References

Glance/iX: HP

Suprtool:
http://www.robelle.com

DiskPerf, Fflush:
Http://www.allegro.com

Klondike:
http://www.lund.com

The Memory Manager

- Prefetch
  - Bring pages into memory directly
  - Work with >4Gb files
  - Span Space-ID's properly
- 6.5 problems and tradeoffs

robelle solutions technology

9

For the past eight years or so Suprtool has been happily doing what is known as prefetch which, generally provides a performance gain of 7 to 15%. Suprtool calls the "prefetch_" system routine, so that while Suprtool is busy on other things, the system is theoretically bringing the pages needed next into memory, ahead of the call to Fread. (Note: even Fread calls "prefetch_")

A lot of effort and testing went into Suprtool for the 6.5 large-file feature, to insure that prefetch would still work properly. The key to testing was to insure that when a file spanned a Space-ID (went beyond 4Gb) the global pointer that pointed to the space-id would decrement to the adjacent space-id properly. This was done by watching the virtual address and the offset into the space. The first incarnation was incorrect and we looped around and began prefetching from the start of the file again, I noticed that the extract became approximately 30% slower.

This was done months prior to the official launch of MPE/iX 6.5. Shortly after the release, we a customer reported that Suprtool was at least four times slower in a serial extract. Experiments showed that if calls to prefetch were turned off, the problem was not as bad, but the performance was still not great. After much investigation and work on my part and the part of HP's Craig Fairchild we narrowed down the problem to a particular MPE patch. The patch changed what happened to pages of memory when we were finished with them.

For Techies

References

The Memory Manager Continued

- ROC vs MakeAbsent
  - Release Overlay Candidates

10

When prefetching pages of a file directly into memory, it is a good idea to clean up after yourself and release these pages back to the system.

You can do this by calling an internal routine which marks these pages as being release overlay candidates. The MPE patch that impacted performance on Suprtool made these pages immediately disappear from memory.

HP reversed the effects of this patch with another patch. Recently HP called and asked me to make some experimental changes to how Suprtool deals with pages after Suprtool is done with them and the work is still on-going.

For Techies

References

N-Class and A-Class

- New Chips
- Faster Memory Controllers
  - 8.5GB/s memory bus
  - 105 ns CPU to Memory Latency
- Faster System Bus
  - 4.3GB/s over two buses
- Hot Swap Power Supplies
- Hot Plug PCI Slots

robelle
solutions technology

11

On February 1st 2001, HP announced the new N and A-class servers. The goal of the design was to provide huge amounts of processor, memory and I/O Bus bandwidth, balanced by very low latencies between the CPUs, memory and I/O. This is delivered by PA-8500 or PA-8600 chips, 4.3 GB/s system bus bandwidth, across two system buses and up to 8.5 GB/s memory bus bandwidth, shared across up to four memory buses. The I/O bandwidth can be up to 6.4 GB/s aggregate, shared across 24-266 MB/s I/O channels and with a CPU to memory to latency of 105 nanoseconds.

These servers also have the capability for Hot Swapping of Power Supplies, fans and PCI cards.

The bottom line of all this is that these servers are faster than many of us have ever seen in the HP 3000 arena.

For Techies

References

The Promise of PCI

- DeFacto Standard
- Optimized
- Incredible Throughput
- Twin Channel PCI boards
- “Make Each Disc Access Count”

robelle
solutions technology

12

PCI is now the optimized, industry-standard method of connecting a computer to a SCSI bus that was developed to meet the high demands for increased I/O. What is more interesting is the fact that each master controller has 12-single byte wide buses. The main two master controllers allow for 10 Twin Turbo slave I/O controllers. The main system (Bus 0) controller has a multifunction Core I/O that is responsible for I/O thru a 10/100 Base T port, RS-232C, LAN Console, Remote console, local serial port, and an Ultra 2 SCSI port.

The bottom-line is that the overall architecture provides for an extremely fast transfer of data across the I/O subsystem. This does however, place even more emphasis on the performance of disc drives. This once again makes the motto coined by Bob Green some years ago, of "Make each disc access count", just as relevant as it was ten years ago.

Every time you access the disc, you want to retrieve as much information as possible from the disc to this incredibly fast /IO system. This is of course the model that Suprtool uses for it’s high performance; it reads large chunks of data from disc and attempts to move the data as few times as possible.

A product that may become crucial for these large N-class machines is DeFrag/X which will defragment your discs. The can help your drives make fewer physical movements and keep the I/O pump primed.

For Techies

References

Defrag/X:
http://www.lund.com

The Suprtool Effect

- How Fast
- Invited to Test
  - the frustrating first time
  - the extremely successful second time
- Some Results
  - Glance reported 400 IO's per second

13

In mid-2000 I was invited to HP CSY to do initial testing on a new N-class Server along with two other third-party software vendors. Once we were selected, I researched all about the N-class architecture by reading thru the N-Class White Paper that was written in April, 1999 when the 9000 N-Class servers were released.

Testing began well enough with the restoring of our test suite and running of some initial tests. Then problems began with the disc drives that were in the system and we saw full SCSI bus resets for every disc in the disc array. It was slightly frustrating, because at times we saw tremendous I/O rates followed by no I/O when the SCSI bus reset.

A few months later in December 2000, we were invited back for more certification testing. This time, without SCSI resets, I was able to run my standard Suprtool test suite as well as do some huge extracts and sorts on a sample AMISYS database.

Giving me this kind of access to my own N-Class Server with 2Gb of memory and over 200 Gb of disc is like leaving an entire Grade 2 class alone in a candy store. While doing the first huge tests of extracting over 5Gb of data from a dataset, I ran Glance/iX and watched the I/O rates and I was amazed to see them exceed 400 per second. The initial results were very impressive.

For Techies

References

Extract Performance

- Performance of a straight extract
- 11,899,340 records
- 464 bytes per record
- 5,521,293,760 bytes
- 398 seconds
- 27,745,194.7739 bytes per second

The first test that I performed in an attempt to measure the IO performance was to do a straight extract of data from a jumbo dataset to a large file. The jumbo dataset was real world data 5.5 Gb in size, which was extracted in 398 seconds, which amounted to about 28 million bytes being exchanged per second.

Previously I had done testing of large files on a 997/400 with a similar disc configuration and the highest rate of bytes per second had only been 3,161,353 bytes per second.

This is a huge difference which can be directly attributed to the PCI bus and the fast system bus.

For Techies

References

Sort Performance

- Raw Sort Performance
- Sort 11,899,340 records
  - 464 byte record
  - 10 byte key value
  - worst case sort
- 1133 Seconds

robelle
solutions technology

15

The next case that was tested was with Raw sort performance which was equally impressive, as we sorted the same 5.5 Gb file by a 10 byte key value, which had already been sorted in descending order so that we taxed both IO and CPU.

The difference again was astounding as the N-class server was able to process 8 million bytes per second and the 997/400 was only able to process 987,000 bytes per second.

For Techies

References

If Selection Performance

- Selection via Table feature
  - Table Feature loads values in memory
  - Read thru dataset and do comparison on 10 byte key value
  - Extract 1/10th of data
- 11,899,340 records
- 448 seconds

robelle
solutions technology

16

The final case that astounded me the most was using the Table feature of Suprtool which loads a file into a mapped file then reads an entire dataset and compares against the key value in the mapped file. The N-class server processed an astounding 24 million bytes per second. I can only theorize that this is due to the fact that the table file was in memory completely and the Very Low Latency Memory combined with the high IO bandwidth really makes a huge difference for this type of extract.

Unfortunately I didn't have data to compare against during my testing on the 997/400.

For Techies

References

Summary

- Extremely High IO Bandwidth
- Disc Performance Weakest Link
- Same Rules Apply

robelle
solutions technology

17

In summary, the new hardware is extremely impressive and will more than meet the needs for most if not all e3000 applications. The old rules still apply though of making each disc access count and reading as much data as you can with each fread. Another good rule is to pay attention to disc balancing and keeping your discs healthy.

For Techies

References