Benchmarks - Are they Really Useful?

A Boyanich, S P Maj
Department of Computer Science
Edith Cowan University
Western Australia
iso9660@yahoo.com

Abstract

Benchmarking is an important commercial tool, which can be used for quantifying the performance of computer and network equipment. Furthermore, benchmarks are potentially valuable as part of curriculum studies in computer and network technology. Benchmarks may be of value to support the understanding of different architectural features and their effect on equipment performance. In effect the benchmarking results may provide a tangible metric that can be related directly not only to various architectural features but also the interactions between different levels in the memory hierarchy, hence providing a method of quantifying different performances of differing computer architectures and configurations. In this context a wide range of benchmarks were tested using the criteria of: repeatability, comparability, consistency, use of meaningful units etc. The criteria selected are based on the fundamental principles of measurement science. Our results clearly indicated that different benchmarking suites gave significantly different results for the same equipment. Also each benchmark failed to give consistent results, when compared to other benchmarks, even on identical equipment. Furthermore many of the benchmarks provided performance data in arbitrary units that were difficult to relate to expected changes in performance. In effect every benchmark tested failed to meet the evaluation criteria. The authors offer an alternative benchmarking method that is designed to meet such criteria and experimental work to date indicates some success using this new metric. This paper presents results of this work and gives recommendations regarding the use of benchmarks in computer education courses.

1. Introduction

Benchmarking is a term used to describe the process of testing either a PC or a selected PC module (e.g. Hard Disc) and obtaining a metric representing the associated performance so as to be used as a comparison between other similar devices. It is simply a test to compare performance that may be used to aid the selection of equipment. There is currently a wide range of benchmarking programs readily available. Primarily these standards fall into one of three categories - trade magazines, standards organizations such as SPEC and TPC (Ideas International) and finally individuals. There is a wide range of Benchmarks and often a collection of them is used, a possible advantage being that others

may compensate the weakness of any one Benchmark in the suite. Benchmark programs considered directly relevant to a typical single user, multi-tasking environment running a de facto standard suite of 32 bit applications include: AIM Suite III, SYSmark, SPEC 2000 and Ziff-Davis PC Benchmark. Consumer magazines use Benchmark suites to evaluate PC's and publish their results (Table 1) *[1]*. However results such as these are only of limited value and raise more questions than answers. By example, what difference in performance can a user expect if the bench mark value result is higher by 1 or 2 units or by a factor of 10 or more? Furthermore, what difference in performance would a user expect between an IBM Aptiva EQ3 (Business Disk WinMark 98 value of 939) and a Gateway G6 300 (Business Disk Win Mark 98 value of 1,380)? More significantly what units are used and are the scales linear, logarithmic or hyperbolic?

Intel marks their microprocessors with a part number and the maximum rated clock speed. Furthermore they publish a special series of Benchmarks called the Intel Comparative Microprocessor Performance Index (iCOMP) that can be used as a relative gauge of microprocessor performance. However, approximately half of the PCs (under $1,000) now sold in the USA are likely to have Advanced Micro Devices (AMD) microprocessors. For many AMD microprocessors the model designation does not correspond with the associated clock speed. For example, the AMD K5 PR133 has a clock speed of only 100Mhz. Cyrix, IBM, SGS Thompson, and AMD jointly developed the alternative P Rating (PR) system. This Benchmark is based on the Winstone de facto standard windows Benchmark suite.

Table 1: PC Benchmark Suite

Benchmark	Gateway G6 300	IBM Aptiva EQ3
Business Winstone 98	20.63	18.33
CD-ROM WinMark 98 : Overall	1,556.67	1,350
CPUmark 32	772	550.33
Business Disk WinMark 98	1,380	939
High-End Disk WinMark 98	3,783.33	2,736.67
Business Graphics WinMark 98	93.13	105.67
High-End Graphics WinMark 98	146	130

The use of benchmarks and benchmarking can be an essential component of computer architecture curriculum *[2]*. In effect benchmarking results may provide a tangible metric that can be related directly not only to various architectural features but also the interactions between different levels in the memory hierarchy. However, the expectation of benchmarks are that they meet the requirements defined by measurement science.

1. Measurement and Measurement Science

In an attempt to obtain selection criteria that may be used to evaluate benchmarks in the fundamentals of measurement science were considered. Measurements are used to make, exchange, sell and control objects. According to Barney,

> *'A measurement is the process of empirical objective assignment of numbers to properties of objects or events in the real world in a way such as to describe them' [3]* p.138

History has many examples of measures in the search for useful standards such as faedm, finger-width and pint. Early Egyptians defined one finger-width as a zebo and established an associated simple, reproducible and denary scale of standard measurements. It is significant that human dimensions were used as the basis of one of the first standards. Some Medieval units formed the basis of the English Imperial System, which though based on human dimensions did not use a decimal scaling system. The use of non-uniform scales in the Imperial System required different conversion factors to convert between units. Lord Kelvin is reported to have said that,

"?*the English Imperial System of units are absurd, ridiculous, time wasting, brain destroying.*" According to Barney it is essential to be able to express quantities in their, '?*commonly occurring magnitudes without the use of fractions or very large numbers and to be able to measure these quantities simply and accurately'.* The European Metric System of measurements (Systems International - SI) is based on the decimal number system From the three, independent fundamental units of Mass, Length and Time other units may be derived e.g. area, velocity, etc. The SI system meets the ancient requirements of simplicity and ease of use due to reasonably sized units. Measurements, therefore, require the definition of units, the establishment of standards, the formation of scales for the comparison of measured quantities and ease of use. It can therefore be argued that any measurement standard, to be of practical value to PC users, must therefore have the following characteristics:

- Use fundamental units which allow other units to be derived
- Units based on the decimal scaling system
- Units relevant to human dimensions or perceptions
- Be useful and allow different systems to be measured

We therefore used these criteria to evaluate a range of different benchmarks.

3. Experimental Design and Platforms

Two Complex Instruction Set Code (CISC) architecture PCs and two Sun Microsystems Sun SPARC devices were used for during the experimental work. The CISC specifications were as follows:

CISC System One
CPU: Intel Pentium-II (Stock 350Mhz, 3.5x 100Mhz FSB)
Front Side Bus Speed for Test Series A: 100Mhz (CPU @ 350 - Stock)
Front Side Bus Speed for Test Series B: 133Mhz (CPU @ 466 -Over clocked)
RAM: 256mb
Hard Disk Drive: Western Digital WDC WD205BA (20gb ATA33)
IDE Chipset: Intel PIIX-4 BX
SCSI Chipset: None
Video Card: nVidia RivaTNT-1 v3400, 16mb Video Ram, Memory Clock: 120Mhz, 3d Engine Clock: 105Mhz
Sound Card: Creative Labs PCI64

CISC System Two
CPU: Intel 80486DX (Stock 40Mhz)
Bus Speed for Test Series A: (CPU @ 40Mhz - Stock)
Bus Speed for Test Series B: (CPU @ 50Mhz - Over clocked)
RAM: 24mb
Hard Disk Drive: SCSI Seagate Hawk ST32151N (2.1gb SCSI-2 Fast @10mbit/sec)
IDE Chipset: None
SCSI Chipset: Adaptec aic2940
Video Card: S3-805 with a GenDAC, 2mb Video Ram
Sound Card: None
Network: None

RISC System One
Make: Sun Microsystems SPARCstation 2
Architecture: Sun4c - 4/75
Manufacturers Name: "Calvin"
CPU: Fujitsu CY7C601 (CPU @ 40Mhz - Stock)
Bus Speed: 20Mhz Sbus
RAM: 64mb
Hard Disk Drive: SCSI Seagate Hawk ST32151N (2.1gb SCSI-2 Fast @10mbit/sec)
IDE Chipset: None
SCSI Chipset: NCR

Video Card: Sun TurboGX
Sound Card: Sun Internal Proprietary
Network: Sun/AMD Lance

RISC System Two
Make Sun Microsystems SPARC Station 20 (SS20)
Architecture: Sun4m - 20/xx
Manufacturers Name: "Kodiak"
CPU0: Ross/Fujitsu RT200DW-200/512 (HyperSPARC @ 200Mhz)
CPU1: Ross/Fujitsu RT200DW-200/512 (HyperSPARC @ 200Mhz)
(CPU1 was disabled during testing)
Bus Speed: 25Mhz/64-bits wide Sbus
RAM: 224mb
Hard Disk Drive: SCSI Seagate Hawk ST32151N (2.1gb SCSI-2 Fast @10mbit/sec)
IDE Chipset: None
SCSI Chipset: NCR
Video Card: Sun TurboGX
Sound Card: Sun Internal Proprietary
Network: Sun/AMD Lance

Using these platforms the Ziff Davis, SPEC CPU95 and lmbench benchmark suites were tested and evaluated. In addition to this Microscope and Norton disk benchmarks were used to test two Integrated Drive Electronics (IDE) hard disk drives. During all the testing particular attention was given to the requirements of basic measurement science i.e. repeatability, accuracy, systematic and random errors etc. Accordingly all these tests were run continuously and repeated on separate occasions and all initial readings were discarded as part of a 'run-in' period to allow the system to stabilize.

3. Results

Two Integrated Drive Electronics (IDE) hard disc drives were tested under identical experimental conditions using Microscope and Norton Utilities (Table 2). The results clearly show that both software tools give comparable average seek times. However, there was a significant difference in the data transfer rates for the same hard disc drive tested by different software under identical experimental conditions. By example, for the same hard disc drive Microscope gave a result of 1.18Mbytes/s, which is significantly different from the value of 2.6Mbytes/s given by Norton Utilities.

The *lmbench [4]* suite of operating system micro-benchmarks provides a collection of programs for cross-platform comparisons. Brown and Seltzer modified lmbench to *'increase its flexibility and precision, and to improve its methodological and statistical operation'.[5]*

Table 2 : Hard Disc Drives

Benchmark program	Quantum LPS 420A	Seagate ST 3144A
Microscope Ave Seek Time (ms)	11	13
Microsocpe Data Transfer Rate (Kb/s)	670	700
Norton Ave Seek Time (ms)	12.2	14.7
Norton Data Transfer Rate (Kb/s)	1617	990

Certainly lmbench has made a significant impact on benchmarking with subsequent versions addressing known defects. The same PC was tested using CacheChk, lmbench, Microscope, Norton Utilities and two in-house benchmark programs. Units used by these programs, measuring the same device (e.g RAM memory or Hard disk) included Mbytes/s, nanoseconds, microseconds, nanosecond/byte, latency etc. Furthermore, some benchmarks measured track-to-track seek time however others only gave figures for the average seek time. Not only were direct comparisons between benchmark suites problematic but also different benchmark programs gave very different results for the same device. When the results obtained from different suites were converted to common units significant differences were evident. Different suites enable different features on both the device and interface to the device, which has a significant effect on the final result. The inference is that the results are derived by different methods and that this difference may be a source of confusion for a user

For both CISC platforms the higher performance PC consistently gave a higher benchmark value. Any increase in clock speed of both systems gave a proportional increase in performance of each device. By example, for the Ziff Davis Winbench99/CPUmark99 an increase of the CPU and front side bus speeds from 350/100Mhz to 466/133Mhz respectively gave a proportional increase in the benchmark results (+/- 2.3%). Any change in hardware performance was reflected in the results for applications tested. However no recognisable units were given for this or any of the tests in the suite. In attempt to make benchmark results that may be more meaningful to a user or student Ziff Davis provide results for applications that include: Photoshop, Premiere, MicroStation SE MP etc.

Again no meaningful units were supplied which could be directly related back to a common metric of bits and bytes per second.

Using the SPEC benchmark suite gave comparable results. However it should be noted that the new SPEC2000 suite has replaced the SPEC95 suite. The problem here is that the two systems do not provide any meaningful units and appear to give very different results for an identical computer system (table 3).

Table 3 System: Sun Enterprise 3500/4500

Suite: CPU2000	#CPU	Base	Peak
CINT2000(86)	1	198	212
CFP2000(90)	1	246	261
CINT2000 Rates(64)	1	2.3	2.46
CFP2000 Rates(65)	1	2.85	3.02

[6]

Suite: CPU95	#CPU	Result	Baseline
CINT95	1	18.3	14.9
CFP95	1	30.1	26.5
CINT95 Rates	1	165	134
CFP95 Rates	1	267	239

[7]

3. Discussion

According to Seltzer, Krinsky, Smith and Zhang,

> '*Most performance analysis today uses either micro benchmarks, or standard macro benchmarks (eg SPEC, LADDIS, the Andrew benchmark). However, the results of such benchmarks provide little information to indicate how well a particular system will handle a particular application. Such results are, at best, useless, and, at worst, misleading'.*
> *[8]*

According to Hennessy and Patterson, there are some commonly held misbeliefs about benchmarks, in particular Million Instructions Per Second (MIPS) and Million Floating-Point Operations Per Second (MFLOPS) *[2]*. In an attempt to obtain more meaningful

Benchmarks specialist interest groups also evaluate equipment using specific applications (Table 4). The Unix-based batch application environment is well served by Benchmark suites (SPEC/CINT95) suitable for both commercial and scientific establishments. However, according to Lee, Crowley, Baer, Anderson and Bershard *[9]* :

> *'However, most of the world's personal computers and workstations run some flavour of the Microsoft Windows operating system on Intel x86 processors. Moreover, most of these computers run personal productivity and entertainment applications rather than engineering or server workloads'*

Again according to Lee et al :

> *'Our results show that desktop applications differ from SPEC95 integer applications in two important ways: First desktop applications have a bigger instruction working set size due to the fact that they tend to call a greater number of distinct functions. Second, the desktop applications tend to execute many more indirect calls'*

Table 4 Specialist Interest Group Benchmarks

	Socket 7 Intel TX based AB-AXS	Socket 7 AMD 640 based Shuttle HOT603	Slot 1 LS Based Aopen AB6L
Norton	92.9	94.5	114.4
CPU Mark 32	535	536	606
FPU/WinMark	762	759	1200
Final Reality	2.37	2.38	2.88
Quake 2	23.9	24.1	
Turok Demo	57.0	56.4	68.9
Incoming	23.66	24.4	

As a relative guide Benchmarks are an aid to selection, however, all of these results must be interpreted and many questions still remain. Questions include:

- Some of the figures are to two decimal places. By example, the IBM Aptiva EQ3 has a Business Winstone 98 figure of 18.33. What is the basis for this number of significant figures?
- What difference in performance can a user expect if the bench mark value result is higher by 1 or 2 units or by a factor of 10 or more? For example, what difference in performance would a user expect between an IBM Aptiva EQ3 (Business Disk WinMark 98 value of 939) and a Gateway G6 300 (Business Disk Win Mark 98 value of 1,380)? What difference in performance can a user expect from a Pentium 100 (iCOMP 90) and a Pentium 200 (iCOMP 142)? Are the scales linear, logarithmic, hyperbolic?
- How does the iCOMP rating compare to the PR rating?
- What units are used? Does each Benchmark define their own (unspecified) units?
- As a user, how is the difference in performance manifested and perceived?
- Why does the same device, when tested under identical experimental conditions, give significantly different results? .
- How can heterogeneous devices be compared, e.g. how can the performance of a hard disc drive be compared to a microprocessor?

In an attempt to address at least some of these concerns and hence provide a more suitable pedagogical framework a nodal model of PC equipment has been proposed by Maj *[10]*. This B-Node model uses the common units of either Mbytes/s or Frames/s which are directly relevant to user perceived performance. It is possible to model all devices using this B-Node model and hence directly compare heterogeneous devices. According to Maj,

> *'Using this conceptual framework the PC is considered as a series of nodes whose performance is measured by bandwidth (Frames/s). Using the standard compulsory ECU course evaluation questionnaire the unit was highly rated by students. Furthermore, a more detailed study was conducted to investigate student experience of the nodal concept. From an enrolment of eighty students, forty were given questionnaires. Thirty-six students thought the nodal concept should be taught. Thirty-five students thought that this concept helped them understand computer technology. Thirty-five students thought that using a common unit (Frames/s) helped in evaluating PC devices.'*

Conclusions

A wide range of benchmarks were tested using the criteria of: repeatability, comparability, consistency, use of meaningful units etc. The criteria selected are based on the fundamental principles of measurement science. Each benchmark suite was self consistent but consistency was not maintained between different products. Typically no meaningful

or common units were provided. When units were provided there was no consistency between products. The lack of or use of arbitrary units made it difficult to relate performance to user expectations. In effect every benchmark tested failed to meet the evaluation criteria. Hence these benchmarks cannot be used for a consistent model for student understanding There is the possibility that for more advanced studies students could approximate the inconstancies by understanding the criteria on which the se benchmarks operate However, this is far beyond the scope of an introductory level course.

The authors offer an alternative modelling technique, B-Nodes, that is designed to meet at least some of these criteria. Experimental work to date indicates some success using this new metric. The results from using the B-Node model as a pedagogical framework were extremely positive.

References

1. Hardware, *Hardware - Consuming Interest*, in *Australian Personal Computer*. 1998. p. 117-122.
2. Hennessy, J.L. and D.A. Patterson, *Computer Architecture A Quantitative Approach*. 1996, San Francisco: Morgan Kaufmann.
3. Barney, G.C., *Intelligent Instrumentation - Microprocessor Applications in Measurement and Control*. 1985, Exeter: Prentice/Hall International (UK).
4. McVoy, L. and C. Staelin, *LmBench*. 1998, BitMover.
5. Brown, A.B. and M.I. Seltzer. *Operating System Benchmarking in the Wake of Lmbench: A Case Study of the Performance of NetBSD on the Intel x86 Architecture*. in *The 1997 ACM Sigmetrics International Conference on Measurement and Modelling of Computer Systems*. 1997. Seattle, Washington, USA.
6. SPEC, *All Published Spec CPU2000 Results*. 2000, SPEC.
7. SPEC, *All Published SPEC CPU95 Results*. 1995, SPEC.
8. Seltzer, K., et al. *The Case for Application-Specific Benchmarking*. in *The 1999 Workshop on Hot Topics in Operating Systems*. 1999. Rio Rico, AZ, USA.
9. Lee, D.C., et al., *Execution Characteristics of Desktop Applications on Windows NT*. ACM SIG Computer Architecture, 1998. **26**(3).
10. Maj, S.P., D. Veal, and P. Charlesworth. *Is Computer Technology Taught Upside Down?* in *5th Annual Conference on Innovation and Technology in Computer Science Education*. 2000. University of Helsinki, Finland: The Association of Computing Machinery.