This paper was subsequently published by IEEE Micro with adjustments to match the style of the publication.
https://ieeexplore.ieee.org/document/9623444 0272-1732 © 2021 IEEE, Digital Object Identifier 10.1109/MM.2021.3112911

How Many VAXes Fit in the Palms of Your Hands?

Exploring an old benchmark on a new CPU chip

John L. Henning,Oracle, Nashua, NH, USA

Background: Benchmarking in the 1970s and 1980s

During the early history of microprocessors, benchmarks were of interest to customers and important to marketing, but results and methods were not comparable. For example, a 1985 Performance Summary [1] from a vendor of popular minicomputers contains results from a variety of benchmarks, with a variety of weaknesses:

Instruction timing: A 10-page table provides instruction timing for 7 models, ranging from 0.096 for a bit clear to 9007 for a CISC polynomial. (Although not stated, the unit is presumably microseconds.) Concerns: (1) The assembly-language benchmark program is not provided. (2) Customers would not know whether their own applications primarily use fast or slow instructions. (3) Comparisons are provided among the single vendor's systems, but no comparisons are provided to other systems. (4) Even if such comparisons were available, they would not be very meaningful given architectural differences.

Fortran benchmarks such as GAUSS, HANOI, HUGHES, PRIME: Times are provided for 50 Fortran benchmarks on 5 computer models. Because these are written in a higher-level language, they may be more meaningful than the assembly-language benchmark of individual instructions. Concerns: (1) Although the benchmarks are claimed to be "industry standard", it would have been useful to include a reference to where they may be found. (2) The text says that multiple benchmarks were modified "to reduce variability" but does not define what that means. (3) It is not clear whether times can be compared to times seen on systems from other vendors, especially given that there were modifications. (4) It is noted that some vendors may have "omitted operating system overhead in their quotes" of the benchmark results. (5) Some of the benchmarks had "dead code" which was optimized away when the compiler recognized that it served no useful function. For two of the benchmarks, the result was a run time of zero seconds.

Transaction processing: A 40-page chapter provides information about several transaction processing workloads, comparing 6 computer models while varying load. Concerns: (1) The benchmarks were not available to customers. (2) They exercise the vendor software environment, and are not portable. Therefore, it is impossible to do comparisons.

Whetstone and Dhrystone: For these benchmarks, references are provided to versions of source code. Concerns: (1) As noted by Reinhold Weicker, the author of Dhrystone [2], both of these benchmarks are "synthetic": they collect and measure program fragments. As such, they may miss important characteristics of real applications. (2) Both had several popular versions, and it was not always clear which version a vendor quoted. (3) Dhrystone - unlike many previous benchmarks - includes a series of run rules, written by Weicker [3]. However, Dhrystone did not have a mechanism for enforcement of the rules or for peer review of results.

Figure 1 A model of a VAX 11/780, which defined performance of 1.0 for the original SPECmark

SPECmark

At the time that Weicker was publishing his Overview of Common Benchmarks [2], the Standard Performance Evaluation Corporation was just coming into existence and publishing its first results. Weicker noted that "SPEC's goal is to collect, standardize, and distribute large application programs".

The initial benchmark suite was termed "SPECmark" (later known as SPECmark89 or SPEC CPU 89). SPECmark improved comparability because: (1) Application programs provide more meaningful data than synthetic kernels and instruction timings. (2) SPEC controls the source code, thereby reducing ambiguity as to what is measured. (3) The benchmarks were ported to multiple environments. (4) Run rules constrain practices that are allowed. (5) Reporting rules require that sufficient information is provided so that results can be reproduced. (6) For results published by SPEC, testers are required to submit their results for peer review.

Perhaps most importantly, (7) SPECmark checks whether or not the program obtained acceptable answers. Many benchmarks omit this step, and without it, results may be meaningless. As a witty computer scientist observed: "I can make it run as fast as you like if you remove the constraint of getting correct answers." [4]

SPECmark performance was calculated relative to the performance of the "reference system", the then well-known VAX 11/780, which defined performance of 1.0. Figure 1 shows a model of a VAX 11/780. The actual size of the CPU cabinet on the left side of the model is about 150 x 120 x 75 cm (60 x 45 x 30 in.), weighing 500 kg (1100 lb.), with a power requirement of 6225 W [12]. For example, the result disclosure page for the SPARCstation 330, excerpted in Figure 2, shows that this 1989 deskside system was already over 10x as fast as the 1978 dual-refrigerator-sized VAX.

SPEC Benchmark Release 1.0 Summary

Sun Microsystems, Inc. SPARCstation 330

RESULTS: Benchmark No. & Name	SPEC Reference Time (seconds)	SPARCstation 330 Time (seconds)	SPEC Ratio
001. gcc	1482	107.6	13.8
008. espresso	2266	195.9	11.6
013. spice 2g6	23951	2152.6	11.1
015. doduc	1863	225.2	8.3
020. nasa7	20093	1800	11.2
022. li	6206	552.8	11.2
023. eqntott	1101	87.7	12.6
030. matrix300	4525	314.7	14.4
042. fpppp	3038	232.9	13.0
047. tomcatv	2649	351.5	7.5
Geometric Mean	3867.7	343.7	11.3

Hardware	
Model Number:	SPARCstation 330
CPU:	25 MHz CYC7C601 (IU)
FPU:	25 MHz SPARC FPC/FPU
Cache Size:	128KB (I+D)
Memory:	32 MB
Disk Subsystem:	327 MB, SCSI disk
Network Interface:	Ethernet
Software	
O/S Type and Rev:	SunOS 4.0.3
Compiler Rev:	Sun Fortran 1.2
Other Software:	None
File System Type:	SunOS 4.0.3
Firmware Level:	ROM Rev 3.0
System	
Tuning Parameters:	None in use
Background Load:	None
System State:	Single User

Tested in: Sept. 1989 | **By: SMI, WSD Perf.** | **Of: Mountain View, CA** | **SPEC License # 006**

Figure 2 A 1989 result using the original SPECmark. In 1989, a single chip provided >10x the performance of the VAX 11/780

SPECmark Rating for a Modern System

An attempt was made to discover the SPECmark rating of a contemporary system using a contemporary compiler. Results for a single core of an Oracle Cloud system with Intel Xeon Gold 6354 [5] using GCC 10.2 are shown in Figure 3.

SPECmark® Release 1.2b Summary

Oracle, Inc. VM.Optimized3.Flex (Intel Xeon Gold 6354, 3 GHz)

RESULTS: Benchmark No. & Name	SPEC Reference Time (seconds)	ORACLE VM.Optimized3.Flex Time (seconds)	SPEC Ratio
001.gcc	1482	.1494	9,921
008.espresso	2266	.2984	7,595
013.spice2g6	23951	3.5935	6,665
015.doduc	1863	.0990	18,826
020.nasa7	20093	.7519	26,722
022.li	6206	.3774	16,446
023.eqntott	1101	.1321	8,335
030.matrix300	4525	.0907	49,886
042.fpppp	3038	.0943	32,203
047.tomcatv	2649	.0455	58,245
Geometric Mean	(Estimated - see notes)		17,826

Hardware	
Model Number:	X9-2c
CPU:	2 x Intel Xeon Gold 6354 @ 3 GHz
FPU:	Integrated
Cache Size:	L1: 32 KB I + 48 KB D on chip per core L2: 1.25 MB I+D on chip per core L3: 39 MB I+D on chip per chip
Memory:	512 GB; more info in notes
Network Interface:	Ethernet
Software	
O/S Type and Rev:	Oracle Linux Server release 7.9
Compiler Rev:	gcc / gfortran 10.2.1-11.1.0.1)
Other Software:	None
File System Type:	ext4
Firmware Level:	FW 5.0, BIOS 3.4 66030400, Apr-2021
System	
Tuning Parameters:	Oracle defaults (see notes)
Background Load:	747 default processes, 0.00 load avg.
System State:	Multi-user

Tested in: July 2021 | **By: Oracle** | **Of: Burlington, MA** | **SPEC License #: 006**

Notes: This is a RETIRED benchmark suite from 1989, modified for compatibility with modern compilers.
The changes were NOT APPROVED by SPEC. All results must therefore be considered to be ESTIMATES.

Benchmark tuning:

```
001.gcc1.35    -g -m32 -O1
008.espresso   -g -m64 -Ofast   -march=native
013.spice2g6   -g -m32 -O3 -ffast-math
015.doduc      -g -m64 -O3      -march=native
020.nasa7      -g -m64 -Ofast
022.li         -g -m64 -Ofast -flto
023.eqntott    -g -m64 -O3      -march=native
030.matrix300  -g -m64 -O3      -march=native
042.fpppp      -g -m64 -Ofast   -march=native
047.tomcatv    -g -m64 -Ofast   -march=native
```

Memory detail:

```
 512 GB (13 x 32 GB 2Rx4 PC4-2400T-R
       + 3 x 32 GB 2Rx4 PC4-2666V-R running at 2400)
```

Oracle Linux installation defaults oracle-database-preinstall-19c in sysctl.conf:

```
fs.aio-max-nr = 1048576
fs.file-max = 6815744
kernel.panic_on_oops = 1
kernel.sem = 250 32000 100 128
kernel.shmall = 1073741824
kernel.shmmax = 4398046511104
kernel.shmmni = 4096
net.core.rmem_default = 262144
net.core.rmem_max = 4194304
net.core.wmem_default = 262144
net.core.wmem_max = 1048576
net.ipv4.conf.all.rp_filter = 2
net.ipv4.conf.default.rp_filter = 2
net.ipv4.ip_local_port_range = 9000 65500
```

Figure 3 Year 2021 system measured with SPECmark

The system achieved an estimated SPECmark rating of 17,826. This result is termed an "estimate" because of changes to SPECmark that have not been approved by SPEC, including:

- Modify timing to use `perl Time::HiRes` instead of `/bin/time`.
- Include appropriate header files, such as `stddef.h, string.h, errno.h.`
- Use `stdargs.h` instead of `varargs.h`.
- Resolve symbol clashes.
- Adjust `static` and `extern`.
- Fix argument types where these led to incorrect answers.
- Attempt to fix compiler warnings which may be relevant to compiling in 64-bit mode. Due to time constraints, this attempt was cut short, and as can be seen in the notes section of Figure 3, some benchmarks were compiled in 32-bit mode.

The attempt to use this long-retired benchmark demonstrated additional ways in which SPEC has improved comparability over the years. (8) Starting with SPEC CPU 2000, all benchmark tuning is placed in a single config file which is published with the result. (9) The original SPECmark had several benchmarks which read no input files. This is dangerous because if too much is known at compile time, ultimately, a benchmark may be reduced to a print statement. (10) Although SPEC CPU prefers benchmarks that are derived from real applications, several SPECmark benchmarks are sufficiently small [6] that they appear to be kernels. Later SPEC CPU releases refreshed the suites with new applications and new versions of old applications, leading to much larger source code, as shown in Figure 4 and in the description page for SPEC CPU 2017 [7].

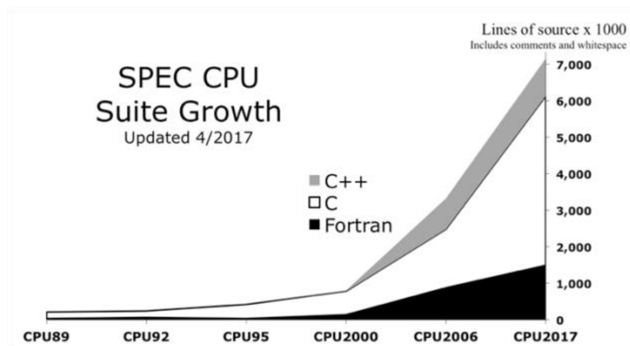

Figure 4 SPEC CPU growth [8]

Unofficial SPECrate89 Throughput

Although Figure 3 provides an estimated SPECmark value for the Oracle system, the test used only 1 core on a 36-core system. It would be interesting to have a measure of full-system CPU performance. Over the years, SPEC CPU defined several throughput-oriented SPECrate metrics for multiple processors. [9] [10] Although the definitions have varied in the major releases of SPEC CPU, all of them include:

- Multiple identical copies are started.
- The observed time is from the start of the first copy to completion of the last copy.
- The SPECrate metric is inversely proportional to the observed time.

SPECmark® Throughput - UNOFFICIAL

RESULTS: Benchmark No. & Name	SPEC Reference Time (seconds)	ORACLE VM.Optimized3.Flex Time (seconds)	Copies	Unofficial Throughput
001.gcc	1482	.1931	36	276,357
008.espresso	2266	.3856	36	211,546
013.spice2g6	23951	3.7602	36	229,308
015.doduc	1863	.1648	36	407,013
020.nasa7	20093	.8594	36	841,664
022.li	6206	.5381	36	415,183
023.eqntott	1101	.1845	36	214,798
030.matrix300	4525	.1963	36	829,945
042.fpppp	3038	.1718	36	636,760
047.tomcatv	2649	.1610	36	592,499
Geometric Mean	(Estimated and UNOFFICIAL - see notes)			408,134

Oracle, Inc.
VM.Optimized3.Flex
(Intel Xeon Gold 6354, 3 GHz)

Hardware

Model Number:	X9-2c
CPU:	2 x Intel Xeon Gold 6354 @ 3 GHz
FPU:	Integrated
Cache Size:	L1: 32 KB I + 48 KB D on chip per core L2: 1.25 MB I+D on chip per core L3: 39 MB I+D on chip per chip
Memory:	512 GB; more info in notes
Network Interface:	Ethernet

Software

O/S Type and Rev:	Oracle Linux Server release 7.9
Compiler Rev:	gcc / gfortran 10.2.1-11.1.0.1)
Other Software:	None
File System Type:	ext4
Firmware Level:	FW 5.0, BIOS 3.4 66030400, Apr-2021

System

Tuning Parameters:	Oracle defaults (see notes)
Background Load:	747 default processes, 0.00 load avg.
System State:	Multi-user

Tested in: July 2021	By: Oracle	Of: Burlington, MA	SPEC License #: 006

Notes: This is a RETIRED benchmark suite from 1989, modified for compatibility with modern compilers.
The changes were NOT APPROVED by SPEC. All results must therefore be considered to be ESTIMATES.
The 1990s methods of throughput calculation differ from the definition used on this page, which anachronistically uses the CPU 2017 method:

```
number of copies * (time on the reference machine / time on tested system)
```

Benchmark tuning:

```
001.gcc1.35     -g -m32 -O1
008.espresso    -g -m64 -Ofast   -march=native
013.spice2g6    -g -m32 -O3 -ffast-math
015.doduc       -g -m64 -O3      -march=native
020.nasa7       -g -m64 -Ofast
022.li          -g -m64 -Ofast -flto
023.eqntott     -g -m64 -O3      -march=native
030.matrix300   -g -m64 -O3      -march=native
042.fpppp       -g -m64 -Ofast   -march=native
047.tomcatv     -g -m64 -Ofast   -march=native
```

Memory detail:

```
512 GB (13 x 32 GB 2Rx4 PC4-2400T-R
      + 3 x 32 GB 2Rx4 PC4-2666V-R running at 2400)
```

Oracle Linux installation defaults "oracle-database-preinstall-19c" in sysctl.conf:

```
fs.aio-max-nr = 1048576
fs.file-max = 6815744
kernel.panic_on_oops = 1
kernel.sem = 250 32000 100 128
kernel.shmall = 1073741824
kernel.shmmax = 4398046511104
kernel.shmmni = 4096
net.core.rmem_default = 262144
net.core.rmem_max = 4194304
net.core.wmem_default = 262144
net.core.wmem_max = 1048576
net.ipv4.conf.all.rp_filter = 2
net.ipv4.conf.default.rp_filter = 2
net.ipv4.ip_local_port_range = 9000 65500
```

Figure 5 Throughput of a contemporary system measured with the 1989 benchmark + (anachronistically) the 2017 method of throughput calculation.

- The SPECrate metric is proportional to the number of copies. (Exception: SPECmark89 v1.2b reported the number of copies, but did not multiply by them.)

Beyond the above list, the definitions have varied, usually by including additional constant factors that were intended to cause the reported results to fall within a desired range.

In the interest of providing some measure of full system SPECmark performance, the year 2017 method of calculating throughput was employed [11], which is simply:

```
ncopies * reftime / observed time
```

where the `reftime` is the time for a single copy on the reference system – in this case, the VAX 11/780. The results are shown in Figure 5, which is marked "Unofficial" because it not only uses the unapproved changes of Figure 3, it also uses an anachronistic method of calculating the throughput.

Summary

How Many VAXes Fit in the Palms of Your Hands? If you hold one contemporary Xeon Gold 6354 in each palm, you hold the processing power of over 400,000 VAX 11/780s.

[1] The 1985 Performance Summary is the third edition of a glossy, typeset, well-organized document with 164 pages and many tables and graphs. It includes work by multiple performance groups at a now-defunct computer manufacturer. The full title is not provided here because it is labeled "For Internal Use Only", although one suspects that customers may have routinely seen copies or excerpts.

[2] Reinhold P. Weicker, An Overview of Common Benchmarks, Computer, Volume 23, Issue 12, December 1990

[3] Reinhold P. Weicker, "Dhrystone benchmark: Rationale for version 2 and measurement rules," SIGPLAN Notices, vol. 23, no. 8, pp. 49–62, Aug. 1988. https://github.com/Keith-S-Thompson/dhrystone/blob/master/v2.0/README or https://www.netlib.org/benchmark/dhry-c

[4] Richard Hart, Digital Equipment Corporation CSE Performance Group, personal communication. 1982. https://www.spec.org/cpu2017/Docs/overview.html section Q2

[5] Intel Corporation, "Intel Xeon Gold 6354 Processor." https://ark.intel.com/content/www/us/en/ark/products/212460/intel-xeon-gold-6354- processor-39m-cache-3-00-ghz.html

[6] John L. Henning, SPEC CPU Suite Growth: An Historical Perspective, Computer Architecture News, Vol. 35, No. 1 - March 2007 https://www.spec.org/cpu2006/publications/SIGARCH-2007-03/01_cpu_suite_growth.pdf

[7] Standard Performance Evaluation Corporation, "SPEC CPU 2017 Documentation Index" https://www.spec.org/cpu2017/Docs/ section "Benchmarks"

[8] The figure is originally from [6] and was updated at https://www.spec.org/cpu2017/Docs/overview.html section Q19

[9] Alexander Carlton, CINT92 and CFP 92 Homogeneous Capacity Method, SPEC Newsletter, Vol. 4, No. 2, June 1992, https://www.spec.org/cpu92/specrate.txt

[10] John L. Henning, SPECrate2006: Alternatives Considered, Lessons Learned, Conference: Computer Performance Evaluation and Benchmarking, SPEC Benchmark Workshop 2009, Austin, TX, USA, January 25, 2009. Proceedings, available at https://blogs.oracle.com/jhenning/spec-benchmark-workshop-2009

[11] Standard Performance Evaluation Corporation, "Q15. What are SPECspeed and SPECrate metrics?," SPEC CPU 2017 Overview https://www.spec.org/cpu2017/Docs/overview.html section Q15

[12] "VAX hardware handbook," Digital Equipment Corporation, Maynard, MA, USA, vol. 1, p. A–14, 1986. http://www.bitsavers.org/pdf/dec/vax/handbook/VAX_Hardware_Handbook_Volume_1_1986.pdf

John L. Henning is currently a performance engineer at Oracle, Nashua, NH, USA, and has been the Secretary for the SPEC CPU Subcommittee since 1998. His first performance optimization experience was in 1973 when he trimmed an eight-hour DOS/360 job to 45 minutes by avoiding the use of magnetic tape. Contact him at: john dot henning at oracle dot com or john at spec dot org