WHITE PAPER

Public Sector
PC Procurement

The Role of Benchmarks in the Public Procurement of Computers

As computing technology has advanced, it has become more difficult to compare the performance of various computer systems simply by looking at their specifications. Since specifications do not necessarily predict performance.

Author(s)

Ulrich Norf

Achieve 'Value for Public Money'

This white paper is intended to provide an overview of benchmarks to government procurement authorities working on technical specifications for computer systems (desktops, notebooks and servers) in public tenders.

Procurement authorities are most likely to achieve 'Value for Public Money' if public tender documents (including the technical specification) are written to give equal access to interested bidders, which will maximize the number of compliant bids for the procurement authorities to choose between.

When purchasing computers, procurement authorities should identify the features of the CPU and the computer system they would like to purchase, considering the intended use of the products and the available budget. They should consider all the features brought to the market by research and technological innovation in the manufacture of computers, with a goal of identifying the features that best meet their requirements and help users achieve their goals.

Using benchmarks to assess performance

As computing technology has advanced, it has become more difficult to compare the performance of various computer systems simply by looking at their specifications. For example, a processor that operates at a higher clock frequency may not necessarily deliver more computational power. Since clock speed (frequency) does not necessarily predict performance when comparing processors with different architectures, from different manufacturers, tests (called benchmarks) have been developed to compare product performance. Relying only on physical parameters to compare different products may not result in a well-informed decision.

Price, performance and power are important considerations when purchasing IT equipment for government agencies. Since public funds are being used, computer platforms should be purchased at the best possible price, with the maximum performance, and with the highest energy efficiency to obtain the lowest total cost of ownership. Regarding performance, many public procurement agencies use system specifications such as CPU frequency and core count, cache size, RAM type and amount, and HDD/SSD type and capacity. This information is readily available from the system and processor suppliers, and is an easy way to compare various computer platforms. The problem is that using specifications as a metric of performance gives an incomplete picture of overall computer platform performance – what about user experience, platform responsiveness and energy efficiency? None of these can be determined by system and processor specifications.

Table of Contents

Benchmarks are specialized computer program that run on the systems under evaluation. The benchmarking program executes a series of standard tests and trials simulating particular workloads on the system, and generates a final performance score. The performance score provides a snapshot of system performance on the workloads measured, which enables an objective, data-driven comparison.

Evaluating products using performance-based benchmarks, rather than processor numbers or clock speeds, can lead to better informed decisions. Application-based benchmarks based on usage models can provide a solid framework for comparing performance of computing products to be deployed within government agencies.

Which benchmark should I use?

There are numerous performance benchmarks available, and it is not always easy for a procurement authority to choose the most appropriate benchmark(s) for a specific tender.

Choosing an inappropriate benchmark could result in buying a computer system which is different from what the organization requires, and in some cases may even lead to discrimination against specific vendors and their products.

Regardless of which benchmark is chosen, it is extremely important to set up and follow a rigorous methodology when using performance benchmarks. Variations in the way a benchmark is run may lead to the results being unreliable and not comparable, which may even result in a challenge to the contract award.

The first aspect to consider is the type of benchmark. There are four types currently available in the market:

System benchmarks evaluate the overall performance of a system for a defined usage model.

Component benchmarks measure the performance of individual components, such as the CPU, memory, or graphics card.

Synthetic benchmarks are used for isolating performance of certain features of certain hardware components.

Application-based benchmarks measure a defined usage model, including complete applications the representative user would typically run. All actions are scripted, including thinking time, mistakes and multitasking.

For any of these types, a good performance benchmark should always have the minimum attributes described below:

Relevant and Representative: Government procurement authorities should choose a benchmark or a combination of benchmarks which measure performance using tests representative of the actual everyday use for which the system is intended.

If the benchmark is not relevant, procurement authorities may risk purchasing a product which is different from what is needed.

Use of "Future" Technologies: New IT purchases are usually aimed toward better performance, newer technology and forward compatibility. However, predicting the future is a hard thing to do, and government IT purchases should be focused on what is available in the mainstream today.

Forward-looking technologies such as OpenCL have been the "wave of the future" for over 10 years now. The reality is that it is not a mainstream technology today. So, if there is no plan to update software over an extended period of time, then adopting new technologies such as OpenCL should not be a concern.

Recognised and Built with Stakeholder Input: Procurement authorities should choose performance benchmarks developed and maintained by well-recognized industry consortia, i.e. independent non-profit standardization bodies with wide industry representation. Benchmarks from consortia have been openly developed and broadly debated, have well-defined methodologies, and are therefore generally objective, impartial, reliable, reproducible and widely accepted.

Privately developed benchmarks may have varying levels of stakeholder involvement in the development process. Input is sought, but the developer makes the decision about the benchmark measurements. Furthermore, there is typically less transparency in the methodology used in these benchmarks, which could make them unsuited to compare the performance of different computer systems, especially in the context of public tenders, because variations in the methodology influence the scores and make the resulting performance scores non-comparable.

Up-to-date: Procurement authorities should always use the most recent version available of any given benchmark.

Good performance benchmarks are continuously updated, and new benchmarks are regularly introduced to keep pace with development and innovation in the computer industry. A benchmark that is not up-to-date will not take into account how new features (e.g. multithreading) affect performance.

Using an outdated benchmark to compare two systems may provide an inaccurate performance comparison. For example, assume a case where one system offers real performance benefits with improved technology, and a second, older system does not. If the outdated benchmark does not recognize and measure the technology improvement, the system with the old technology may actually score higher than the system with the new, better performing technology. In short, outdated benchmarks may disadvantage the most recent, innovative and better performing products.

About benchmark developers

Benchmark developers are categorised in order of transparency and openness as below:

- Non-profit benchmarking consortium (examples are BAPCo, SPEC and EEMBC)
- Non-profit open source community (an example is Principled Technologies)
- For-profit independent benchmark vendor (an example is Futuremark)
- Smaller for-profit developers (an example is Primate Labs)

The benchmark development process is different for each type of organization. For example, each member of a benchmarking consortium is a stakeholder and may assist with coding, provide feedback to the development,

and/or participate in pre-release testing. Each member of a benchmarking consortium also has a vote toward which features will be included in the final released benchmark. Intel is part of these consortiums as a contributor to the development of these benchmarks like any other partner members. Examples of such consortiums include BAPCo, SPEC and EEMBC.

Website links:

BAPCo	http://www.bapco.com/
SPEC	http://www.spec.org/
EEMBC	http://www.eembc.org/

In the case of an open source community or independent benchmark vendor, the final decisions as to which features are included in a final released benchmark are typically made by the executives of the organization. Independent benchmark vendors vary in the level of user input they seek. For example, Futuremark solicits input from its user community, although company executives make the final decisions.

Small companies tend to be the least transparent of all of the benchmark developers, because they typically do not have resources to solicit user feedback during the benchmark development process.

Performance Benchmarks Recommended by Intel

The performance benchmarks recommended by Intel for mainstream PC's and servers, at the time of publication of this white paper, are listed in the following Appendix. See below for detailed descriptions of Intel's recommended and other benchmarks, along with benchmark assessment disclosures.

Intel's Recommended Benchmarks

Recommended Benchmarks for Mainstream PC

Benchmark	Usage Model	Windows* Support	Publisher
CrXPRT* 2015	Chromebook* Performance Qualified Battery Life	N/A (Chrome OS*)	Principled Technologies (PT)
HDXPRT* 2014	Windows Media Editing	Win64	Principled Technologies
TouchXPRT* 2016	Windows Light Media Editing	UWP-x86, UWP-x64 & UWP-ARM	Principled Technologies
WebXPRT* 3	Web Applications on a Specified Browser	N/A (browser)	Principled Technologies
MobileMark* 2018	Windows Performance Qualified Battery Life	Win64	Business Applications Performance Corporation (BAPCo)
SYSmark* 2018	Windows Desktop Applications	Win64	Business Applications Performance Corporation
APPmark* 2018	Cross-Platform Light Productivity, Media & BL	UWP-x64, UWP-ARM	Business Applications Performance Corporation
SPEC CPU* 2017	Compute Intensive Application Performance	Win64	Standard Performance Evaluatic Corporation (SPEC)
3DMark* 2.6	DirectX* and OpenGL* ES Gaming	Win32 & Win64 (Time Spy supports Win64, Night Raid supports UWP-ARM)	UL/Futuremark

Benchmark Descriptions:

3DMark 2.6

3DMark* 2.6 is published by UL/Futuremark, a for-profit company owned by Underwriters Laboratories (UL). 3DMark tests DirectX* and OpenGL* ES Gaming performance using stand-alone benchmarks: Time Spy (DX 12), Night Raid (DX 12), Fire Strike (DX 11), Sky Diver (DX 11), Cloud Gate (DX 10) and Ice Storm (DX 9 / OGL ES). Each benchmark produces metrics on graphics, physics (CPU), combined and overall performance. For Windows*, 3DMark 2.6 supports both Win32 and Win64 – except for Time Spy, which supports Win64. Night Raid supports UWP-ARM.

APPmark 2018

APPmark* 2018 is published by the Business Applications Performance Corporation (BAPCo), a benchmarking consortium. APPmark tests Cross-Platform Light Productivity, Media & Battery Life performance using real-world scenarios: Productivity, Creativity and Video Playback. With the exception of Video Playback, each scenario produces individual metrics that roll up to an overall score. The Battery Life Rating is calculated by iterating on all three scenarios, along with idle time, in a battery rundown scenario. For Windows*, APPmark 2018 supports UWP-x64 and UWP-ARM.

CINEBENCH R15

CINEBENCH* R15 CPU is published by MAXON, a for-profit company known for a 3D modelling application called Cinema 4D*. The CINEBENCH CPU test renders a single frame from a 3D-modelled movie using up to 256 processing threads, if available. This is not relevant to PC, because mainstream applications such as Microsoft Office* do not utilize every available thread of execution. The CPU test produces a CINEBENCH points (cb) score. For Windows*, CINEBENCH R15 supports Win64. Note that CINEBENCH is currently not recommended for evaluating mainstream PC platforms.

CoreMark 1.2

CoreMark* 1.2 is published by the Embedded Microprocessor Benchmark Consortium (EEMBC), a benchmarking consortium. CoreMark tests Processor Core Performance using common integer algorithms: matrix manipulation, linked list traversal (exercises a common use of pointers), state machine operation (common use of data dependent branches) and cyclic redundancy check. CoreMark produces a single score, and multiple instantiations can be run to obtain scores for multicore performance. CoreMark supports Win32 and Win64.

CrXPRT 2015

CrXPRT* 2015 is published by Principled Technologies (PT), an open source community and host of the BenchmarkXPRT* development forum. CrXPRT tests Chromebook* Performance Qualified Battery Life using real-world scenarios: Photo Effects, Face Detection JS, Offline Notes, Stocks Dashboard, DNA Sequence Analysis, 3D Shapes and Photo Collage. Each scenario produces individual metrics that roll up to an overall score. The battery life test produces an estimated battery life expressed in hours and minutes, a separate performance score, and a frames per second (FPS) rate for the HTML5 gaming component. CrXPRT 2015 is supported on all Chromebooks and is available via the Chrome Web Store.

Geekbench 4

Geekbench* 4 is published by Primate Labs, a for-profit company. Geekbench CPU tests processor performance using a suite of synthetic component workloads that cover Cryptography, Integer, Floating Point and Memory functions. Geekbench CPU produces separate Single-Core and Multi-Core scores. For Windows*, Geekbench 4 supports Win64.

Note that Geekbench is currently not recommended for evaluating mainstream PC platforms.

HDXPRT 2014

HDXPRT* 2014 is published by Principled Technologies (PT), an open source community and host of the BenchmarkXPRT* development forum. HDXPRT tests Windows* Media Editing using real-world scenarios: Edit Photos, Convert Videos and Edit Music. Mainstream applications used in the scenarios include Adobe Photoshop* Elements, Apple iTunes* and CyberLink MediaEspresso*. Each scenario produces individual metrics that roll up to an overall score. For Windows, HDXPRT 2014 supports Win64.

MobileMark 2018

MobileMark* 2018 is published by the Business Applications Performance Corporation (BAPCo), a benchmarking consortium. MobileMark tests Windows* Performance Qualified Battery Life using real-world scenarios: Office Productivity, Creativity and Web Browsing. Mainstream applications used in the scenarios include Microsoft Office*, Adobe Creative Cloud* and Google Chrome*. Each scenario produces an individual Performance Qualification Rating and Battery Life Rating. Idle time is included in the battery rundown iteration. For Windows, MobileMark 2018 supports Win64.

PassMark PerformanceTest 9

PerformanceTest* 9 CPU Mark is published by PassMark Software, a for-profit company. CPU Mark tests processor performance using a suite of synthetic component workloads: Integer Math, Floating Point Math, Prime Numbers, Extended Instructions (SSE), Compression, Encryption, Physics, Sorting and CPU Single Threaded. Each synthetic test produces individual metrics that roll up to an overall "CPU Mark" score. For Windows*, PerformanceTest 9 CPU Mark supports Win32 and Win64. Note that PassMark is currently not recommended for evaluating mainstream PC platforms.

PCMark 10

PCMark* 10 is published by UL/Futuremark, a for-profit company owned by Underwriters Laboratories (UL). PCMark tests Windows* Everyday Computing performance using stand-alone benchmarks: PCMark 10, PCMark 10 Express and PCMark 10 Extended. The PCMark 10 benchmark tests Essentials, Productivity and Digital Content Creation scenarios, producing metrics for each scenario that roll up to an overall score. PCMark 10 supports Win64. Note that PCMark is currently not recommended for evaluating mainstream PC platforms.

PCMark Vantage

PCMark* Vantage HDD Suite is published by UL/Futuremark, a for-profit company owned by Underwriter Laboratories (UL). Because it is out-of-date and was originally developed for Microsoft Windows* Vista, PCMark Vantage is not recommended for performance evaluations on modern hardware platforms. However, the HDD Suite continues to be used by the storage industry and technical press because the tests can show differentiation within a single class of storage media, either HDD or SSD. For this reason, the PCMark Vantage HDD Suite continues to be used prudently to showcase storage alternatives.

SPEC CPU2006

SPEC CPU*2006 is published by the Standard Performance Evaluation Corporation (SPEC), a benchmarking consortium. SPEC CPU tests Compute Intensive Application Performance using integer and floating point subtests based on real programs. SPECint*_base2006 and SPECfp*_base2006 measure how fast a processor completes a single integer or floating point compute task. SPECint*_rate_base2006 and SPECfp*_rate_base2006 measure throughput, or how many integer or floating point compute tasks a processor can accomplish in a given amount of time. For Windows*, SPEC CPU2006 supports Win32 and Win64. SPEC CPU2006 was retired by SPEC on January 9, 2018.

SPEC CPU2017

SPEC CPU*2017 is published by the Standard Performance Evaluation Corporation (SPEC), a benchmarking consortium. SPEC CPU tests Compute Intensive Application Performance using integer and floating point subtests based on real programs. SPECspeed*2017_int_base and SPECspeed2017_fp_base measure how fast a processor completes a single integer or floating point compute task. SPECrate*2017_int_base and SPECrate2017_fp_base measure throughput, or how many integer or floating point compute tasks a processor can accomplish in a given amount of time. For Windows*, SPEC CPU2017 supports Win64.

SYSmark 2018

SYSmark* 2018 is published by the Business Applications Performance Corporation (BAPCo), a benchmarking consortium. SYSmark tests Windows* Desktop Applications performance using real-world scenarios: Productivity, Creativity and Responsiveness. Mainstream applications used in the scenarios include Microsoft Office*, Adobe Creative Cloud* and Google Chrome*. Each scenario produces individual metrics that roll up to an overall score. An Energy Consumption Test is also included that aggregates the total energy consumed during each scenario in watt hours (Wh), as measured by a compatible logging power meter. For Windows, SYSmark 2018 supports Win64.

TouchXPRT 2016

TouchXPRT* 2016 is published by Principled Technologies (PT), an open source community and host of the BenchmarkXPRT* development forum. TouchXPRT tests Windows* Light Media Editing using real-world scenarios: Beautify Photos, Blend Photos, Convert Videos for Sharing, Create Music Podcast and Create Slideshow from Photos. Each scenario produces individual metrics that roll up to an overall score. For Windows, TouchXPRT 2016 is supported on UWP-x86, UWP-x64 and UWP-ARM.

VRMark 1.2

VRMark* 1.2 Orange Room is published by UL/Futuremark, a for-profit company owned by Underwriters Laboratories (UL). VRMark Orange Room tests Virtual Reality Gaming performance using a scenario designed for currently available, mainstream VR headsets and PC platforms. VRMark provides an overall scenario score, however in order for the score to be valid, the PC must maintain a consistent frame rate of 109 FPS or higher without dropping frames. For Windows*, VRMark 1.2 Orange Room supports Win64.

WebXPRT 3

WebXPRT* 3 is published by Principled Technologies (PT), and open source community and host of the BenchmarkXPRT* development forum. WebXPRT tests Web Applications on a Specified Browser using real-world scenarios: Photo Enhancement, Organize Album Using AI, Stock Option Pricing, Encrypt Notes and OCR Scan, Sales Graphs and Online Homework. Each scenario produces individual metrics that roll up to an overall score. WebXPRT 3 is supported on all modern, HTML5-enabled web browsers.

Benchmark Assessment Disclosure:

UL/Futuremark PCMark* 8

Benchmark suite for evaluating Windows-based desktop and notebook platforms with two newly added off-the-shelf applications tests, an integrated battery life test and support for OpenCL (GPGPU).

PCMark 8 target usages include Home (common tasks at home), Creative (advanced home computer tasks), Work (tasks for an office environment), Storage (SSD, HDD and hybrid drive performance) and Applications (Adobe CS6 and Microsoft Office applications).

In case of PCMark 8 Applications tests, Futuremark has adopted a new strategy of using installed off-the-shelf applications as workloads. Relevant Adobe and Microsoft applications must be purchased and installed in order to run the PCMARK 8 Applications tests. The specific versions of these applications are not set by Futuremark, which raises concerns over consistency, comparability and reproducibility.

PCMark 8 does not generate an overall "PCMark" score. Rather, it is made up of 6 individual metrics: PCMark 8 Home, PCMark 8 Creative, PCMark 8 Work, PCMark 8 Storage and PCMark 8 Applications – Adobe CS6 and PCMARK 8 Applications – Microsoft Office and battery life testing. This can lead to biased marketing of a specific subtest versus looking at the overall performance of a processor using the entire suite of PCMark 8 subtests.

The scores from this benchmark can be misleading, as a "PCMark" score could actually mean very different things depending on which PCMark 8 subtest is being shown, and on top of which run mode (conventional or accelerated) was selected.

In Conventional mode the workloads do not use OpenCL. This reflects how the majority of software works today, with the score providing a performance baseline.

"Run Accelerated" Accelerated mode allows workloads to use OpenCL acceleration, anticipating the way software will work in the future.

"Conventional" versus "Accelerated" is confusing and misleading, and most users will opt for Accelerated mode without giving much thought to the fact that OpenCL does not represent a mainstream usage model. In today's software ecosystem, the reality is that only a very narrow range of end user applications have been programmed to benefit from OpenCL enabling, and more importantly, a very narrow range of workloads actually take advantage of OpenCL acceleration.

UL/Futuremark PCMark* 10

Benchmark suite for evaluating Windows-based desktop and notebook platforms with support for OpenCL (GPGPU).

Myth: PCMark 10 Extended represents productivity. The additional scenario in PCMark 10 Extended is derived from 3DMark* Fire Strike – a scenario designed for high-performance gaming PCs. Not clear why the competition refers to a benchmark that includes a gaming scenario as "productivity" in their marketing collateral. By definition, a productivity usage model does not include gaming scenarios such as 3DMark Fire Strike.

Myth: PCMark uses real applications. Although LibreOffice* is a real application, it is pre-installed only on Linux* distributions. LibreOffice is open source, and is compiled and built by UL/Futuremark using proprietary parameters – it is not commercial shrink-wrapped software available to everyone.

Myth: PCMark 10 is a well-proportioned benchmark. Benchmarks should represent the activities that most users perform on their PC devices. Mainstream applications such as Microsoft Office* benefit from fewer, more powerful CPU cores – the GPU workloads in PCMark 10 do not represent what most PC users do. Out of 63 tests, 15 are OpenCL* enabled – a technology used mainly for image and video processing.

Myth: OpenCL* is the wave of the future. OpenCL has been around for over a decade and is great for predictable, uniform tasks such as image and video processing. It is almost completely useless for interactive applications that do not benefit from parallelization, such as Microsoft Office*. There is not one line of OpenCL code in Microsoft Office, so why focus on OpenCL?

Myth: Futuremark is an independent organization. In 2014, Futuremark was acquired by the world's largest independent testing company, Underwriters Laboratories (UL). Known mainly for fire and safety compliance testing, UL is likely to leverage Futuremark's software into a variety of benchmarking services. Futuremark used to be an independent benchmarking company, but now is part of a testing services corporation. UL has productized several services around Futuremark's software.

Recommended Benchmarks for Servers

For servers, Intel recommends the performance benchmarks listed in the table below, classified by the workload measured:

WORKLOAD	PERFORMANCE BENCHMARK
Enterprise Resource Planning Customer Relationship Management	• SAP Standard Application Benchmarks* • Oracle Applications Standard Benchmark* • TPC-C* • TPC-E*
Virtualized Environments IaaS (Infrastructure as a Service)	• SPECvirt_sc*2013 • VMmark2.5* • TPC-VMS*
Virtual Desktop Infrastructure Virtual Hosted Desktop	• Login VSI* • View Planner*
Java*	• SPECjbb*2013 • SPECjEnterprise*2010
Online Transaction Processing	• TPC-C* • TPC-E* • HammerDB
Data Warehousing / Data Mart Data Analysis / Data Mining	• TPC-H*
Energy Efficiency	• SPECpower_ssj*2008 • SAP* Server Power benchmark • SPECvirt*_sc2013_PPW (Server and Storage Performance Per Watt) • SPECvirt*_sc2013_ServerPPW • VMmark2.5 (with Server Power or with Server and Storage Power)
Decision Support	• SAP H (BW-EML) • TPC-DS* • TPC-H*
Big Data Analytics	• BigBench • HiBench
Email	• Microsoft* Exchange* Load Generator
Cloud Storage	• COSbench
Enterprise Storage	• SAN (Storage Area Network) • IOMeter* • Remote FIO (iSCSI) • Local FIO (backend characterization) • NAS (Network Attached Storage): • Remote FIO
Workstation	• SPECCPU*2006 • Catalyst Autocad* • SPECapc creo* • 3DSmark* • Ansys* Fluent, Mechanical • ABAQUS* • 3DMARK* • SPECviewperf*
Technical Computing	1. Application Specific Benchmarks **CAE** • LS-DYNA* **Computation Fluid Dynamics** • OpenFOAM* **Digital Content Creation** • Embree* Ray Tracing

WORKLOAD	PERFORMANCE BENCHMARK
Technical Computing	**Energy** • 2D-FFT • 3D-FFT **Financial Services** • Monte Carlo RNG* • Monte Carlo Sim* • BlackScholes* • STAC-A2* • Binomial Options* **Life Sciences** • Gromacs* • MPI Hmmer* • LAMMPS* • BWA* • NAMD* **Manufacturing** • ANSYS Mechanical* • miniFE* • GTC-P* (need PPPL license agreement) **Weather** • WRF* **Physics** • ASKAP tHogbomClean* 2. Component Benchmarks • DGEMM • LINPACK • STREAM • SGEMM • SPECCPU*2006 • SPECMPI*2007 • SPECOMP*2012