White Paper

Public Sector Performance Benchmarking

The Role of Benchmarks in the Public Procurement of Personal 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 because specifications do not necessarily predict performance.

Co-Authors

Bryon Wirtz
Government Segment Sales Manager, CCG Sales

Joab Paiva
Independent Technology Strategist & Consultant

Table of Contents

Strategies to Achieve Optimal Value from Public Funding 1
Emerging Trends Impacting Purchasing Decision-Making 1
Using Benchmarks to Assess Performance 2
- System Performance 2
- Battery Life 2
- AI Capabilities and Performance 3

Which Benchmark Should I Use? 3
About Benchmark Developers 4
Intel's Recommended Benchmarks 5
- Productivity 5
- Professional 5
- AI & Machine Learning 5
- Battery Life 5

Other Benchmarks for Evaluating PC Platforms 6
Best Practices 7
Why Intel Cares About Benchmarks 7
Appendix: Recommended Benchmarks for Content Creation 7

Strategies to Achieve Optimal Value from Public Funding

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

Procurement authorities are most likely to achieve optimal value for public money if public tender documents (including the technical specifications) are written to give equal access to interested bidders. This will maximize the number of compliant bids from which procurement authorities can choose.

When purchasing computers, procurement authorities should identify the features of the computer system they would like to purchase, considering the person who will be using the PC, 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.

Emerging Trends Impacting Purchasing Decision-Making

While the goal, simply stated, is to choose the best solution for the price, many factors and priorities influence the decision. Currently, a number of emerging trends figure into determining which solution is "best" for a specific entity.

The growing impact of cyber threats is one such influential trend. In 2025, the annual cost of cybercrime is expected to reach USD $10.5 trillion.[1] AI is supercharging cyber threats, enhancing the efficacy of cyber criminals while arming defenders with powerful tools to preempt and counteract evolving threats. Understandably, cybersecurity's importance has increased to match its impact. It has become a top priority in device purchase decisions.

Geopolitical uncertainty is increasing risk and altering how organizations and governments make purchase decisions. Decision-makers are prioritizing flexible and reliable supply chains to minimize risk and ensure continuity.

In recent years, heightened environmental awareness, regulatory pressures, and the potential for cost savings have elevated sustainability to a key procurement criterion. Many decision-makers are willing to invest more in devices that support their sustainability objectives.

Beyond security, supply chain reliability, and sustainability, there is a growing understanding that every phase of the PC life cycle, from design to manufacturing, impacts a device's total cost of ownership. For example, the ability to manage and optimize PCs remotely reduces unnecessary resource consumption, which can lead to lower operational costs as well as reducing environmental impact.

As employees become increasingly mobile and maintaining PC fleets gets more complex, the ability to remotely and securely manage devices is gaining importance for other reasons as well. PC fleet stability and comprehensive remote manageability tools keep organizations up and running and agile, allowing them to scale and reach their goals faster. These tools also help organizations to recover from unexpected outages such as the 2024 "Blue-Screen Friday" event that brought down millions of computers worldwide. The collective realization that the PC fleet is a part of critical infrastructure and is integral in a disaster recovery strategy highlights the importance of resilience in device procurement strategies.

Perhaps most significantly, the advent of AI has brought new opportunities for increasing productivity and improving outcomes. While AI workloads have been popular in server and cloud, they have been migrating to the PC to optimize cost, performance, security, and enhance personalization while keeping data safely on the device. In response, public sector organizations are gearing up for the next wave of computers with AI-accelerated applications and capabilities that will significantly impact how devices are used, secured, and managed. Powerful and efficient PCs, optimized for evolving AI workloads and use cases, enable higher-quality work to be executed faster in content creation, data visualization, design, research, and more. As AI and machine learning become ubiquitous in software, organizations that invest in AI PCs can take advantage of the transformative opportunities they offer.

Using Benchmarks to Assess Performance

With the emergence of AI and multiple processing units on a PC, evaluating device performance has become increasingly complex. In the past, physical metrics such as frequency and core counts were used as proxies for performance. However, given the growing number of factors to consider, these metrics alone no longer provide an accurate representation. For this reason, Intel recommends using commercial benchmarks to measure system performance, battery life, and AI capabilities. To make an informed purchasing decision, it is essential for buyers to conduct benchmarks that reflect their unique usage patterns and requirements.

It's important to note that not all benchmarks represent overall system performance in an enterprise environment. Some measure specific aspects of a system that may or may not be relevant to an organization, so care must be taken when weighing how to use benchmark scores.

System Performance

Benchmarks are specialized computer programs 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 then 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.

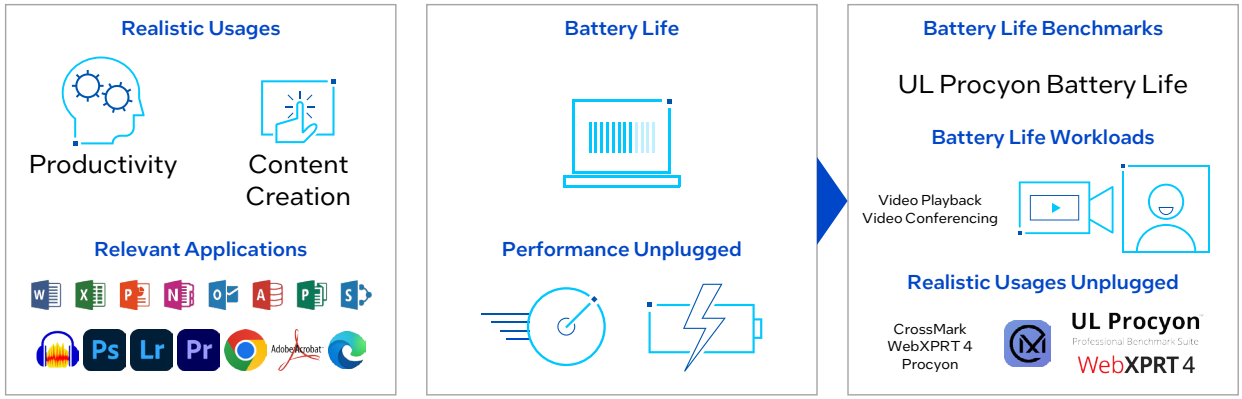
Evaluating products using performance-based benchmarks, rather than processor numbers or clock speeds, can lead to better-informed decisions. Application-based benchmarks reflecting the usage models can provide a solid framework for comparing the performance of computing products to be deployed within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

In recent years, battery life and AI performance have become increasingly important components of overall system performance. New measures have been developed to assess these attributes in real-world scenarios.

Battery Life

Battery life benchmarks aim to measure battery life performance meaningfully, taking into account the fact that different workloads and system-level configurations, such as screen size and brightness, can impact performance. These benchmarks take a multifaceted approach, measuring not only the battery life a user obtains while performing specific tasks but also the system performance.

Recommended Way to Measure Battery Life



Real Applications & Usages

Measure battery life in the context of tasks that make sense.

Battery Life & Performance

Measure not only the battery life you get while performing those tasks, but also the performance.

Multifaceted Approach

Use benchmarks like UL Procyon Battery Life, which uses real applications and has different workload scenarios for battery life.

AI Capabilities and Performance

Growing AI adoption has led to new device selection considerations and to the development of new benchmarks that measure AI performance. These benchmarks cover different AI capabilities and areas of focus, including computer vision, image generation, and large language models (LLMs), to name just a few.

A new type of PC —called an AI PC— is capable of running AI on the PC instead of sending workloads to the cloud. This requires three engines: a central processing unit (CPU), a graphics processing unit (GPU), and a neural processing unit (NPU), specifically designed for AI tasks. Each has a unique role to play in providing the optimal balance of power and performance. The CPU provides bursts of power for fast response and the lowest latency for AI operations. The GPU excels at peaks such as periodic throughput-intensive operations. An integrated NPU runs AI inferencing on the local PC at lower power than the CPU or GPU, helping to extend battery life. These three engines, combined in an "xPU" strategy, deliver distinct strengths and capabilities to enhance performance and enable new AI use cases.

Identifying the right AI PC calls for benchmark tests that measure the performance of all three compute engines together in the system. Similarly, verifying that a system supports different frameworks and data precision types is now a key step in the procurement process.

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 deploying a computer system that is different from what the organization requires, and in some cases, may even lead to discrimination against specific vendors and their products.

It is important to keep in mind that different benchmarks will be better suited to testing different components of a computer system. Free benchmarks may be appropriate for basic performance testing for certain use cases. Crossmark, for example, is easy to install and may sometimes meet an organization's needs. By contrast, Procyon may require up to 15 hours to complete; however, the level of accuracy and detail provided by the Procyon benchmark allows for a much deeper analysis. Generally, for more in-depth analysis or specialized testing, paid benchmarks may be a better fit. Using a combination of benchmarks or one appropriate for testing a variety of usage scenarios ensures a more holistic view of how a system will perform when engaged in tasks that employees actually do on their PCs.

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.

In terms of choosing one or more benchmark tests, procurement authorities should consider these factors:

1. Are the tests real and relevant, and do they use popular applications? Are they a true representation of the device's actual intended usage?
2. Are they reproducible, not predicting an outcome, and up to date?
3. Are they reliable, with unbiased and quality input[2], or are they meant to showcase a specific feature that may not be as relevant to the intended use?

Then, it is important to assess what type of benchmark will provide the most pertinent information for a given set of circumstances. These types are currently available in the market:

System – most likely application-based benchmarks evaluate the overall performance (response time) of a system for a defined usage model reflected in the benchmarking tool, including complete applications the representative user would typically run.

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

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

Relevant and representative: Government procurement authorities should choose a benchmark or a combination of benchmarks that measures 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 that is different from what is needed.

Recognized 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 ill-suited 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 consider how new features (e.g., AI acceleration) affect performance. Using an outdated benchmark to compare two systems may provide an inaccurate performance comparison. For example, consider a case where one system contains new hardware optimizations to boost performance. If the benchmark software is not up to date, it may not be able to take advantage of the newest hardware and may report inaccurate results.

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. On the other hand, software adoption of machine learning and artificial intelligence technologies has seen an increasing ramp in recent years in client computing, and that could be a consideration to future-proof system procurement. In cases where future technologies are considered, a best practice is to choose a platform that is widely compatible and supports existing infrastructure to get the highest ROI potential.

Real-world testing: Synthetic benchmarks that measure the capabilities of one component in isolation can be helpful for software developers to optimize new software versions, but they can create an expectation of performance that enterprise software applications are not realistically able to achieve. PassMark offers a good example of this phenomenon. PassMark scores are based on synthetic tests that take into account hardware specs and theoretical calculations on potential performance. These tests may yield results that look great on paper, but they don't necessarily translate into real-world performance for tasks such as business or professional workloads.

Ground rules of computer design and comparison

(Computer Architecture: A Quantitative Approach, Fifth Edition, Hennessy & Patterson)

Measuring performance and mainstream usages

"Our position is that the only consistent and reliable measure of performance is the execution time of real programs, and that all proposed alternatives to time as the metric or to real programs as the items measured have eventually led to misleading claims or even mistakes in computer design."

Reporting performance

"The guiding principle of reporting performance measurements should be reproducibility – list everything another experimenter would need to duplicate the results."

About Benchmark Developers

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

- Nonprofit benchmarking consortium (examples are BAPCo, SPEC)
- Nonprofit open-source community (examples are Geekbench and Principled Technologies)
- For-profit independent benchmark vendor (an example is UL Benchmarks)
- Smaller for-profit developers (examples are Primate Labs, PASSMARK)

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 on 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 member. Examples of such consortiums include BAPCo, SPEC, and EEMBC.

Website links:

- BAPCo: http://www.bapco.com/
- SPEC: https://www.spec.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 that they seek. For example, UL Benchmarks solicits input from its user community, although company executives make the final decisions.

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

Intel's Recommended Benchmarks

The performance benchmarks recommended by Intel for mainstream personal computers, at the time of this white paper's publication, are listed in the following table. They span four categories, or pillars, pertaining to system performance: productivity, professional software, AI and machine learning, and battery life. Intel recommends using a basket (carefully chosen combination) of benchmarks to evaluate overall performance.

Real-World Commercial Benchmarks

Productivity	Professional	AI & Machine Learning	Battery Life
Procyon® Office Productivity	**Spec WST 4.0**	**MLPerf Client**	**Procyon Office Productivity Battery Life**
The benchmark from UL Solutions uses Microsoft Office applications to measure Windows PC and Apple Mac performance in office productivity tasks.	The benchmark targets the diverse needs of engineers, scientists, and developers who rely on workstation hardware for daily tasks by including a comprehensive set of tests	This is a benchmark for Windows and macOS, focusing on client form factors in ML inference scenarios like AI chatbots, image classification, etc.	This benchmark uses Office apps to simulate a working day. The benchmark uses Word, PowerPoint, Excel and Outlook to run workloads in a loop and can factor in different screen brightness levels.
Web XPRT4	**Procyon® Photo Editing**	**Procyon® AI Text Generation**	**Teams 3x3**
WebXPRT will run a series of tests to measure your browser's HTML and Javascript handling, as well as tasks like photo manipulation and face detection.	This multi-platform benchmark measures performance using Adobe® Lightroom® Classic and Adobe® Photoshop® in a typical photo editing workflow that includes batch processing and image retouching.	This benchmark provides a more compact and easier way to repeatedly and consistently test AI performance with multiple LLM AI models.	A "Teams 3x3 battery life benchmark" refers to a test that measures how long a laptop battery lasts while actively participating in a Microsoft Teams video.
CrossMark	**Procyon® Video Editing**	**Procyon AI Image Gen**	
CrossMark is an easy to run native cross-platform benchmark that measures the overall system performance and system responsiveness using models of real-world applications.	The UL Procyon Video Editing Benchmark uses Adobe Premiere Pro in a typical video editing workflow.	This benchmark provides a consistent, accurate, and understandable workload for measuring the inference performance of on-device AI accelerators.	

Evaluating performance using baskets of these real-world benchmarks can point the way to the device that will best meet an organization's specific needs. Benchmark descriptions, organized by pillar, follow:

Productivity

Procyon® Office Productivity – The Procyon office productivity benchmark from UL Solutions uses Microsoft Office applications to measure Windows PC and Apple Mac performance in office productivity tasks. The benchmark workloads are built on relevant, real-world tasks using Microsoft Word, Excel, PowerPoint, and Outlook (v. 2.7.1108.64; Office version: v.16.0.17628.20110).

Web XPRT4 – WebXPRT 4 Chrome (v118) is a benchmark from Principled Technologies that measures JavaScript/ HTML5 performance using web applications based on real-world usages, like Photo Enhancement, Organize Album Using AI, Stock Option Pricing, Encrypt Notes, and OCR Scan, Sales Graphs, and Online Homework. It produces results for each of the test scenarios, plus an overall score.

CrossMark – CrossMark (v1.0.1.95_HotFix) is a benchmark from the BAPCo consortium that is an easy-to-run, native cross-platform benchmark that measures the overall system performance and system responsiveness using models of real-world applications.

Professional

SPECworkstation® 4.0 –The benchmark targets the diverse needs of engineers, scientists, and developers who rely on workstation hardware for daily tasks by including a comprehensive set of tests. A complete benchmark run estimates system performance by measuring 23 workloads and over 80 subtests to generate scores for seven industry verticals and four hardware subsystems. The benchmark reports top-level scores in two categories: industry verticals and hardware subsystems.

Procyon® Photo Editing – The Procyon photo editing benchmark from UL Solutions uses real-world applications to test performance. This multi-platform benchmark measures Windows PC and Apple Mac computer performance using Adobe® Lightroom® Classic and Adobe® Photoshop® in a typical photo editing workflow that includes batch processing and image retouching.

Procyon® Video Editing – UL Procyon benchmarks use real applications to test performance whenever possible. The UL Procyon Video Editing Benchmark uses Adobe Premiere Pro in a typical video editing workflow.

AI & Machine Learning

MLPerf Client – The goal of the MLPerf Client working group is to produce machine-learning benchmarks for client systems such as desktops, laptops, and workstations based on Microsoft Windows and other operating systems. The MLPerf Client benchmarks are scenario-driven, focusing on real end-user use cases using LLM-focused performance tests.

Procyon® AI Text Generation – Provides a compact and easy way to repeatedly and consistently test AI performance with multiple LLM AI models.

Procyon® AI Image Generation – The Procyon AI Image Generation Benchmark provides a consistent, accurate, and understandable workload for measuring the inference performance of powerful on-device AI accelerators such as high-end discrete GPUs. This benchmark was developed in partnership with multiple key industry members to ensure it produces fair and comparable results across all supported hardware.

Battery Life

Procyon® Battery Life – The Procyon Battery Life Benchmark continues the Battery Life Profile concept that UL Solutions pioneered in PCMark 10. Instead of producing a single number, the UL Procyon battery life profile provides a broad view of battery life across different scenarios, including video playback, idle, and office productivity.

Teams 3x3 – This is Microsoft Teams video conferencing with 9+1 participants for AI-powered experiences (background effects and automatic framing), when available, or in-app blur on systems without the ability to perform Windows Studio Effects. Tested using Microsoft Teams (work or school) classic version (1.6.00.28557).

Other Benchmarks for Evaluating PC Platforms

The additional benchmarks below may be useful in specific scenarios but present some limitations as they do not test full system performance in a comprehensive way. For this reason, they are not included in the recommended list.

CINEBENCH R20 CPU – *3D-Modeled Frame Rendering*

CINEBENCH R20 CPU is published by MAXON, a for-profit company known for a 3D modeling application called Cinema 4D. The CINEBENCH CPU test renders a single frame from a 3D-modeled movie using all available processing threads. 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. CINEBENCH no longer tests the GPU. For Windows, CINEBENCH R20 supports Win64.

Geekbench 5 CPU – *Synthetic*

Geekbench 5 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 5 supports Win64.

UL Benchmarks PCMark 10 - *Windows Semi Synthetic Test*

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.

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

Mobile Mark 30

MobileMark 30 is an application-based, performance-qualified battery life benchmark designed to assist users in making PC purchasing decisions. The new productivity scenario has been designed to best capture the modern mobile office user who seeks to remain productive while on the go.

Best Practices

Benchmark results can come from multiple sources and be interpreted differently. Intel recommends running benchmarks directly, whenever possible, to guarantee the most consistent and realistic results, always keeping in mind the end-user of the device and the intended use cases. Aspects to be aware of when looking at benchmark databases include the following:

- **What is the system hardware configuration?** System configuration can significantly impact scores, even more so for benchmarks that examine a specific workload on a specific device.
- **How was the benchmark run?** Benchmarks can be run from different OS images with different applications running in the foreground and background. Other settings, like screen resolution and battery/performance, can greatly impact results.
- **How is the database populated?** Suppose anyone can submit results to a database. In that case, the average score is likely shifted lower because benchmarks will be run in a wide variety of unknown configurations, leading to scores spreading out and generally shifting lower.
- **How many results are available?** Too few results may result in skewed results due to a small sample size's ability to influence the average. Many results also tend to shift scores lower as described above.

Why Intel Cares About Benchmarks

As a trusted advisor and global technology leader, Intel strives to provide expertise and unbiased guidance to help public sector organizations accurately compare PCs and understand their strengths and weaknesses in different areas. The goal is to ensure that organizations find the right devices to meet their unique needs. Intel's comprehensive benchmark recommendations reflect our commitment to transparency.

Appendix: Recommended Benchmarks for Content Creation

Light Content Creation

Applying filters, creating HDR photos, photo organization. These workloads are light-threaded and scale up with higher CPU turbo frequencies.

Photo Editing

- Procyon Photo Editing
- PugetBench for Adobe® Photoshop® CC
- PugetBench for Adobe® Lightroom® Classic
- CrossMark Creativity sub-score

Heavy Content Creation

3D image and video rendering, video encoding, and exporting. These workloads are usually heavy-threaded and make use of all CPU cores available to finish the task.

Video Editing, Rendering, and Encoding

- Procyon Video Editing
- Blender
- PugetBench for Adobe Premiere CC
- PugetBench for Adobe After Effects

Sources

[1] Blake Hall, "How AI-driven fraud challenges the global economy -and ways to combat it," World Economic Forum, Jan 16, 2025, https://www.weforum.org/stories/2025/01/how-ai-driven-fraud-challenges-the-global-economy-and-ways-to-combat-it/#:~:text=Jan%2016%2C%202025,is%20growing%20Image%3A%20Getty%20Images.

[2] Product–Neutral Tendering of Notebooks," BITKOM, 2022, https://www.bitkom.org/sites/main/files/2023-09/ICT-Procurement-Product-Neutral-Tendering-of-Notebooks-2022.pdf.

Printed in USA 0226/BW/MIM/PDF Please Recycle 366410-001