AMD RYZEN™ PRO 5000 SERIES PROCESSORS: MEASURING WHAT MATTERS – BENCHMARK CONSIDERATIONS FOR COMMERCIAL PC PURCHASES

PERFORMANCE IS EVERYTHING— SO HOW DO WE MEASURE IT?

While many factors impact the PC selection process, performance is almost always #1. It doesn't matter how big the organization is, or which products or services it delivers, businesses buy technology to make people and processes more productive.

Even as data security concerns have gotten more attention in the past few years, performance is still the guiding force in decision-making. Whether putting out an RFQ or running their own internal evaluation, companies rely on specification or test metrics to figure out which PC is right for their needs.

Gone are the days when companies were mostly driven by price, settling for bulky PCs or heavy mobile devices that had "good enough" performance. Today's critical digital foundations need every advantage for success. This means businesses are willing to invest more in performance that will pay much greater returns.

This paper will explore different techniques and strategies for evaluating performance and the pros and cons for each. Using one of the benchmark-based strategies, two modern processors will be evaluated as an example of how to use a comprehensive performance review to better understand the how the two CPUs compare.

OLD SCHOOL METRICS: RUSTY AND UNRELIABLE

Traditionally, companies have used various physical specifications, such as processor frequency and cache size, to set a necessary baseline for PC performance. Unfortunately, these are an incomplete, and often inaccurate, way to assess performance versus actual application workloads.

USING FREQUENCY

There are two problems with using frequency as a meaningful measurement of performance.

First, two identical processors operating on the same frequency can yield dramatically different performance levels. This is due to the efficiency of underlying architectural implementation, measured in Instructions Per Clock (IPC), and is thus invisible to basic spec comparisons.

Secondly, frequency is not a constant for most modern processors. This is especially true for today's notebook PCs, where frequency is constrained by thermal considerations. Additionally, frequencyr will vary dramatically based on everything from the task being performed to the number of cores in use.

A CLOSER LOOK AT THE MATH

RFQs will often use processor frequency as a way to measure expected performance. It's a very inaccurate technique that mostly persists as an artifact from the early days of the PC market, and a quick glance at a basic performance equation quickly shows us why it's incomplete.

CPU T IME = I * IPC * T	
I	Number of instructions in program
IPC	Instructions Per Clock
T	Clock cycle time

While "T," or the inverse of the processor frequency is a key factor, the "IPC" or average number of instructions per clock has an equal impact. Why?

As processors become more sophisticated with super-scaler designs, improved cache, and deep instruction pipelines, they can execute more than one instruction per clock cycle. Silicon designers considering multiple potential options often find that the best design involves a reduction in the number of clock cycles required to execute a set of instructions, rather than simply increasing the frequency of the clock.

For example, assume an application process requires a billion (1x109) instructions to complete and a particular processor has an average IPC of 3.7 and frequency of 2Ghz.

To complete this task would require 1x109 x 3.7 x 0.5 x 10-9, or **1.85 seconds**

Now let's assume a second processor, with a different architecture, operates at a frequency that is 25% higher, but with an average IPC that is fractionally higher at 4.7.

To complete this task would require 1x109 x 4.7 x 0.4 x 10-9, or **1.88 seconds**

This means the second processor is 2% slower, despite having a 25% faster frequency. As this shows, while frequency can be used to predict performance of two parts with the same design, it is too simplistic to compare parts from different families.

FREQUENCY VS. THERMALS

With notebook processors especially, the idea of boost and base frequencies is an oversimplification—the actual frequency at any given time will depend on workload, power-mode, and platform thermal design. For example, the higher the boost frequency, the more heat is produced, and the thermal design of the notebook—including fan size and speed, and also the manufacturer's design constraints for skin temperature—can mean that achieving peak boost will reduce the sustained performance over time.

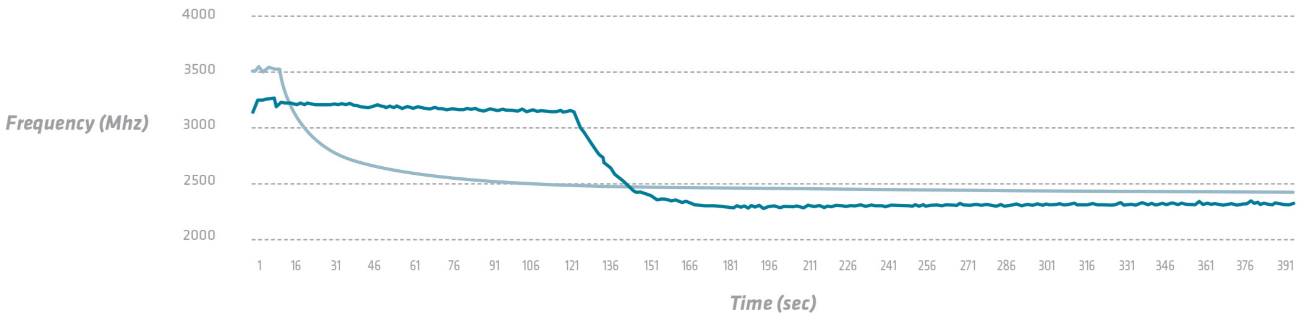
Figure 1 shows the frequency of two processors, with different designs, over time.

- The first processor has a higher boost frequency and quickly jumps to this maximum in the first few seconds. This is also true for the second processor, although its maximum is not as high.
- However, after about 15 seconds, the first processor rapidly reduces its operating frequency in response to the building thermal load, while the second processor sustains a higher frequency for minutes before throttling lower.

The total amount of work done is the area under the curve and so, for longer, more complex tasks where users often have to wait for results, the second processor with lower boost frequency quickly overtakes the first, completing more work over time. But frequency specifications alone give no indication of this real-world behavior.

Processor Frequency Over Time

Figure 1



A BETTER APPROACH: ROBUST, REAL-WORLD TESTS

Modern applications are highly complex, with sophisticated underlying algorithms and data access patterns. The measured efficiency of a processor—its IPC—can vary substantially not only between applications, but also between different workloads in a single application. Many workloads also involve displaying graphics on screen and reading or writing data to and from local or network storage, and the accelerators for these operations can also dramatically impact performance. CPU performance is a poor and incomplete metric; how can we do better?

USER TESTS

One of the best ways to evaluate PC performance is to conduct individual real-world tests that are:

- focused on everyday computing tasks
- designed to mimic real working environments
- built around real-world file and data needs

These tests will do a significantly better job of predicting individual future user satisfaction with their PC and give decision-makers a more accurate performance picture than any published benchmark.

However, this approach is not without its disadvantages. These tests are time-consuming to design and run and can create challenges for decision-makers who rely on measuring performance in a consistent, reliable, and unbiased manner.

CUSTOMIZED APPLICATION SCRIPT TESTING

Beyond individual user testing, the next-best approach is for in-house developers to aggregate suggested workloads from a variety of different classes of users to create customized application scripts that can deliver application performance metrics that map to real user priorities. While this can improve the consistency of measurements and provide repeatable results, it requires substantial and careful upfront investment and doesn't always scale well from one PC generation to the next.

COMPOSITE BENCHMARKS

Given the complexity of customized test design, many companies rely on industry-standard PC benchmarks to evaluate system performance. Rather than using a single metric, however, companies should aim to build a broader, more comprehensive picture of performance by building a composite score across several benchmarks.

BENEFITS AND DISADVANTAGES

Figure 2 compares three different approaches to evaluating PC performance—benchmarks, application scripts, and user evaluations—and shows how the results have different levels of business relevance.

WHAT MAKES A GOOD BENCHMARK?

There are commonly two types of benchmarks used to evaluate PC performance: "synthetic" and "application-based." Both can be useful in the decision process, although individual benchmarks can often have undesirable attributes. Following a general principle of using multiple benchmarks together can mitigate these issues, providing a more dependable picture of performance.

A good benchmark should be as transparent as possible, with a clear description of both what is being tested and how testing is accomplished. In the case of application-based benchmarks, this allows buyers to understand whether workloads being used match their own organization's usage. Without this transparency, it's reasonable to worry that benchmarks are being crafted in favor of a specific manufacturer or processor.

PC Evaluation Strategy

Figure 2

	Pros	Cons
SELF-EVALUATION Evaluate the systems with commonly-used applications and workloads within the organization 1	Very specific to actual usage. Best gauge of user expected performance.	Time consuming and challenging for some organizations to implement.
BUSINESS SCRIPT Time to completion script for common commercial applications 2	Good reflection of user expected performance.	Difficult to create and set up test script.
BENCHMARKS Use a wide selection of industry-standard benchmarks to factor out any issues with a test 3	Easiest and provides some comparative value.	May not provide good insight into user-performance experience.

NOT ALL APPLICATION-BASED BENCHMARKS ARE EQUAL

The tests in application-based benchmarks should represent the workloads that are most relevant to the organization. For example, if 30-50% of a benchmark comes from applications that are seldom used in a commercial setting, then that score is probably not relevant.

Consider the benchmark in **Figure 3**, which is based primarily on consumer-type workloads that are rarely present in an office environment. Therefore, this benchmark would likely not be useful to most commercial organizations.

Some application-based benchmarks measure the off-the-shelf performance and may not represent either the actual deployed version of the application or recent performance optimizations or updates. This is where synthetic benchmarks become very useful.

WHAT ABOUT SYNTHETICS?

While application benchmarks show how well a platform is optimized for specific versions of certain applications, they are not always a good predictor of what new application performance will look like. Unlike application-based benchmarks, synthetic benchmarks measure the overall performance potential of a specific platform.

For example, many video conferencing solutions use multiple CPU cores to perform functions such as virtual backgrounds. Synthetic benchmarks that measure the multi-threaded capability of a platform can help predict how well a platform can deliver this new functionality.

A Cautionary Note

It is important to avoid using a narrow measure of performance to drive synthetic benchmarks. Individual processors, even in the same family, can vary in how they handle even a small piece of code.

Synthetic Benchmark Sub-scores

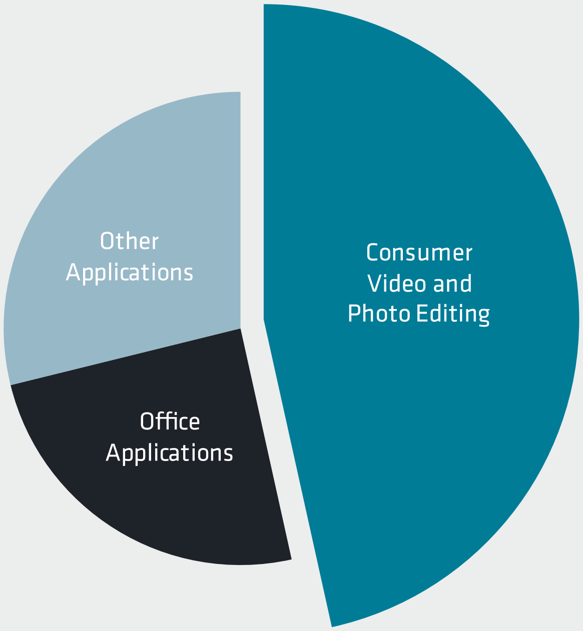
With synthetic benchmarks built on narrow measurement sub-scores, it is important to not overly weight these results. Decision-makers should instead evaluate the composite score, which exercises a broader set of processor functions.

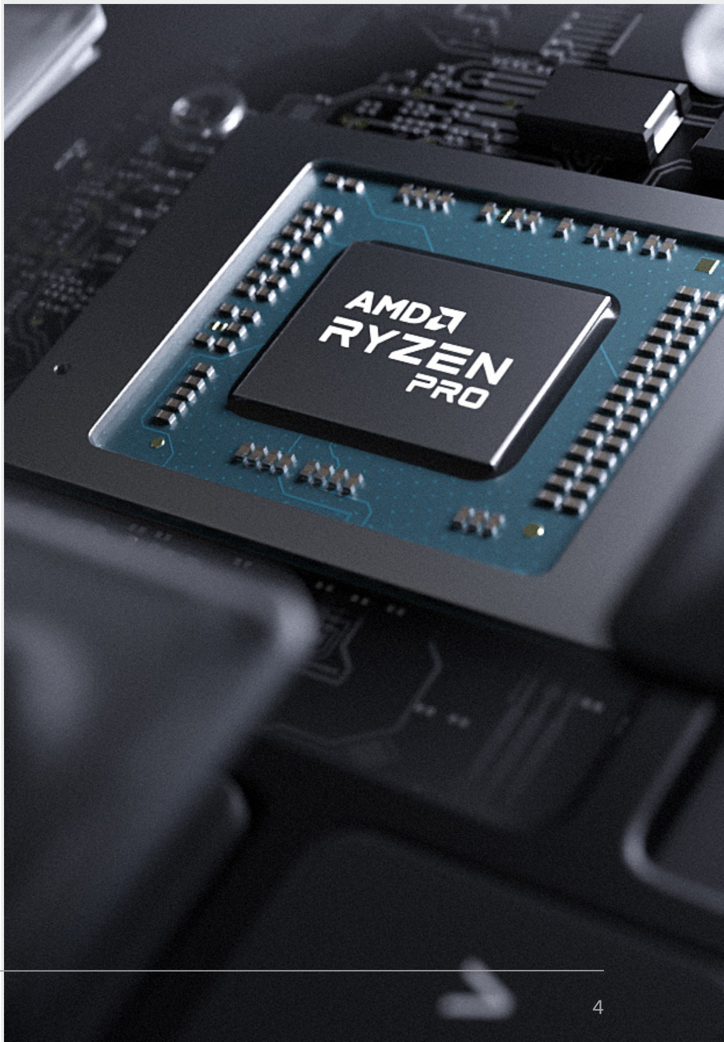
One example is memory sub-scores. These can measure latency and throughput but may not factor in other elements such as a processor cache design that reduces the impact of memory latency on overall application performance.

Figure 4 shows two example processors for which the results for the memory sub-score do not reflect the results from the synthetic benchmark's composite score nor a system level benchmark score. A

Benchmark Composition

Figure 3

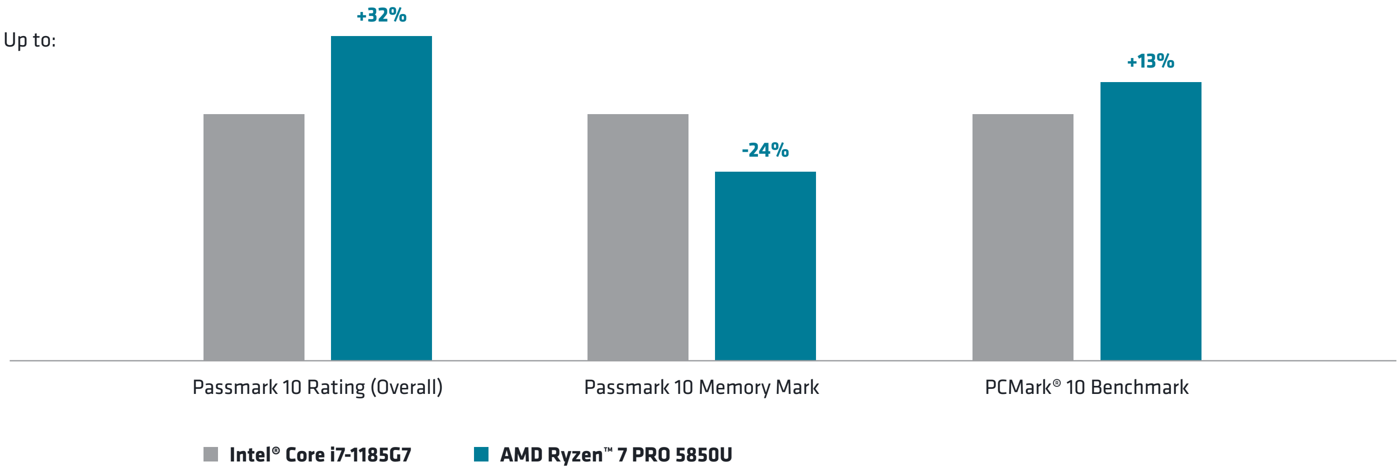




synthetic benchmark score should capture several individual tests, executing more lines of code and exercising different workloads, and then provide an overall composite score. This enables a much broader view of the platform's performance.

As we can see in **Figure 4**, synthetic sub-scores can deviate significantly from the composite score or even other system level benchmarks. Sub-scores which exercise less of the capabilities and provide a more limited view can still be utilized, just with less emphasis vs. the composite score.

Synthetic Benchmark Sub-Score vs. Composite and Application Score *Figure 4*



MEASURING FOR A MULTITASKING WORLD

Application benchmarks have a hard time simulating the desktop workload of a modern, multitasking office worker, because applications rarely run alone, but multiple applications at once add a larger margin of test error.

Synthetic benchmarks that measure the raw multi-threaded processing power of a platform are a good proxy for the demands of today's multi-tasking users.

Combine both for clarity

A best practice is to consider both application-based and synthetic benchmark scores together. By combining scores using a geometric mean, you can account for the different score scales of different benchmarks. This provides the biggest, clearest picture of performance for a specific platform, considering both today's application requirements and ensuring capacity for what comes tomorrow.

OTHER IMPORTANT CONSIDERATIONS

Benchmarks are an important part of a system evaluation, but shouldn't obscure other considerations.

- Measured benchmark performance can vary by operating system (OS) and application version. Ensure that these versions match what's in use inside your environment.
- Other conditions can impact scores, such as background tasks, room temperature, and OS features such as virtualization-based security (VBS) enablement. Tests must reflect these details.

Some users may use relatively niche applications and functions not covered by the benchmarks. Consider augmenting benchmark scores with user measurements and correlating them with synthetic benchmark scores.

MANAGING MEASUREMENT ERROR

Any measurement will have a margin of "measurement error," that is, how much it varies from one test to another. Most benchmarks have an overall measurement error in the 3-5% range.

This is driven by a variety of factors, including the limitations of measuring time, the “butterfly effect” of minor changes in OS background tasks, and other nuances.

Did You Know?

When a new version of a benchmark is introduced, t often has little or no correlation to the previous version.

New workloads, applications, and measurement calculations are included, creating a completely different score. These calculations may also measure only limited workflows.

It’s critical to understand these new core components to ensure they still reflect relevant, real-world usage.

One way to overcome this error would be to measure results five times, discard the highest and lowest scores, and take the mean of the remaining three scores.

Building it into the purchasing process

It is important to consider measurement error when setting requirements in purchase requisitions.

If a score of X correlates well with user satisfaction, then the requisition should stipulate that score should be within [X-Epsilon, Epsilon] where epsilon is the known measurement error. When epsilon is not known, it is reasonable to assume it is in the 3-5% range of the target score.

PUTTING TIME MEASUREMENT IN CONTEXT

Some benchmarks compute a score by measuring tasks that are only milliseconds or even less in length, where humans typically deal with timeframes in the seconds and up range. These benchmarks then apply a mathematical weighting formula to the very short task and derive a computed numerical score that may imply a result far outside of what a typical person would perceive.

For example, if a task takes 5ms on one processor and 6ms on another, the benchmark will report that the first is 20% faster, even without any weighting factor applied. However, a real user is not very likely to notice this difference in actual time required to process the function.

Of course, the responsiveness of a computer is always important, but the key is to understand the duration and number of times the function in the test is being measured, any mathematical weighting factor, and how much that would be perceivable to the user.

These measurements can be augmented with tests that are longer in duration, or Time to Completion (TTC) type tests, both of which measure on timescales that are more perceivable and meaningful to users.

CONSIDERING INCONSISTENT APPLICATION PERFORMANCE

Application performance is not a constant and, as such, is always changing. Software venders bring out new versions, change compilers, optimize key functions, or the OS itself changes. As a result, the system performance can change in unexpected ways.

A good example is Microsoft Excel, a common application test used to evaluate PC performance. Microsoft is constantly working to improve performance of Excel, rolling out changes with each update. Their blog on recent changes rolled out in September of 2020 details a wide range of improvements that can significantly impact performance, sometimes even by orders of magnitude.

It’s a great illustration of how “the same” applications performance tests may vary considerably based on which version of the application is being used. And for benchmarks that embed elements of an application, the difference can be even greater. The benchmark may be several years old, using old application code that has little correlation to current performance.

Even with benchmarks that use the installed version of an application, there can be dramatic differences unless the same application versions are used across all tested systems.

Excel performance improvements now take seconds running Aggregation functions - Microsoft Tech Community

BUILDING A MORE THOROUGH BENCHMARK EVALUATION

The AMD Ryzen™ 7 PRO 5850U and the Intel Core i7-1185G7 are two processors that are used in a variety of enterprise class commercial systems, and many companies purchasing laptops would consider them as primary processor candidates. When conducting the testing, the battery should be fully charged, since this process can contribute significant thermal load that impacts performance.

For this evaluation to be thorough, a wide variety of different types of tests, including industry common synthetic benchmarks, application benchmarks, and actual application time to complete tests were used. These include:

1. CineBench R23 (including 1T and nT)
2. Geekbench v5 (including single core and multi-core)
3. PCMark® 10 Overall
4. PCMark® 10 Express
5. PCMark® 10 Extended
6. PCMark® 10 Productivity
7. PCMark® 10 Digital Content Creation
8. PCMark 10 Applications (including Overall, PowerPoint, Word, Excel, and Edge)
9. PCMark® 10 Gimp Cold App Startup
10. PCMark® 10 Gimp Warm App Startup
11. PCMark® 10 APP start
12. Passmark 10 Overall (including sub-scores for CPU Mark)
13. Passmark 9 CPU Mark (which is very often used in commercial tenders)
14. Sysmark 2018 Rating Overall (older but still referenced for some tenders)
15. Sysmark 25 Rating Overall
16. Puget Photoshop Overall Score
17. Puget Photoshop General Score
18. Blender Bench CPU-BMW27 (TTC)-sec
19. Blender Bench CPU-ClassRoom (TTC)-sec

This set of benchmarks is collected from a variety of industry sources and tests different attributes to provide a more comprehensive view of processor and system performance. Assuming a benchmark error can be roughly 3-5%, the following comparisons will consider +/-5% to be roughly equal in performance. Above this, and users will begin to notice a difference.

In addition, a geomean score, the average of the benchmarks used in the analysis, is also provided. This provides an average benchmark score that factors in differences in magnitude between the different benchmarks.

More importantly, averaging multiple different benchmarks from different sources helps address any possible bias or limitation of a particular test by averaging it across many tests.

CPU PERFORMANCE

The first set of synthetic tests will focus mostly on the processor performance itself. These will give an indication of the raw processing power of a particular processor design. These include:

1T or single-thread performance

This measures the ability of a single core on the processor to execute instructions. Some applications are more sensitive to 1T performance.

nT or multi-thread performance

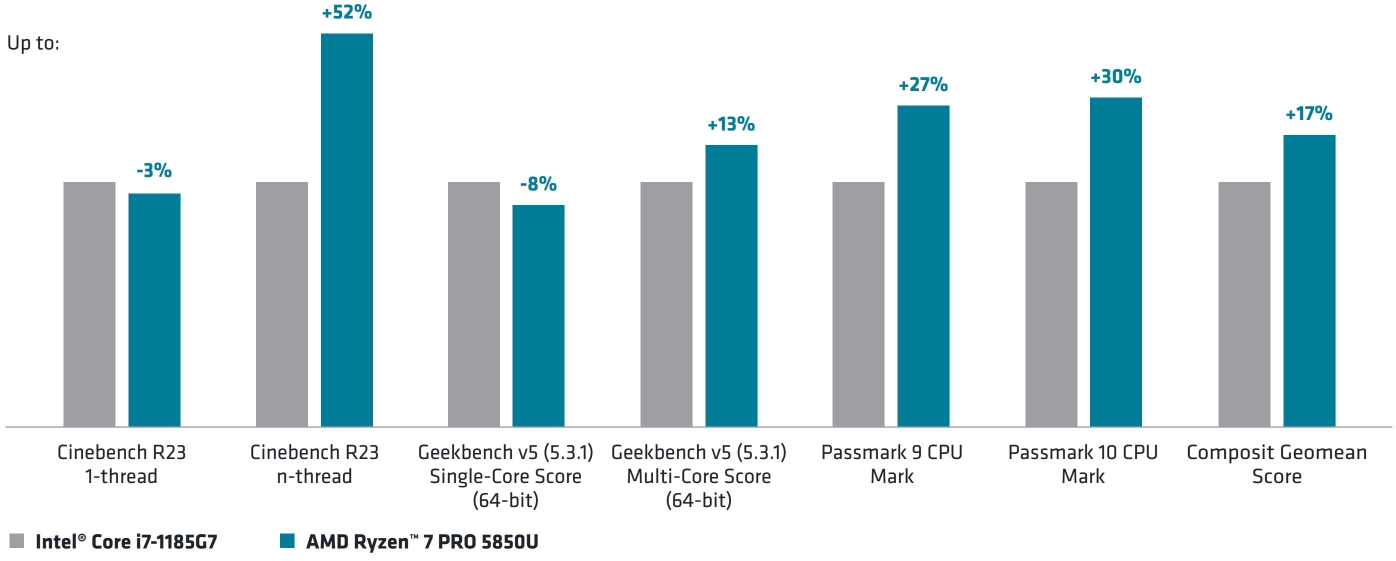
This measures the ability of the processor to use more than one core/thread to execute instructions. To more rapidly scale performance, more cores and threads are added to processor designs.

Application vendors are working to exploit this capability by dividing tasks into multiple code paths that can execute in parallel. Users involved in more complex computing, whether it is large spreadsheet calculation, content creation, or multi-application mega-tasking, benefit from a better nT score.

OVERALL PROCESSOR PERFORMANCE

Overall CPU scores measure both 1T and nT capabilities to give a broad view of overall processor performance. While Passmark 9 is being replaced by the newer Passmark 10 version, it is still widely referenced and therefore included in the analysis.

CPU Performance *Figure 5*



The benchmarks used were the following:

1. CineBench R23 (including 1T and nT)
2. Geekbench v5 (including single core, multi-core)
3. Passmark 10 CPU Mark
4. Passmark 9 CPU Mark

In terms of single-thread or **1T performance, a slight advantage goes to the Intel Core i7-1185G7 processor**. However, on CineBench 1T it is within the 5% margin of error.

In terms of nT or **multi-thread performance** tests as well as the Passmark 9 and Passmark 10 CPU tests, the difference between the two processors is much greater. Here the **Ryzen™ 7 PRO 5850U has a larger advantage and is well outside the 5% range for each of the tests.**

- In this case, the advantage of 8 high-performance cores / 16 threads over the Core i7's 4 cores / 8 threads is apparent. Heavy multi-taskers, users requiring complex computations, frequent users of online collaboration tools, or content creators will have a more productive experience.

The **geomean score** looking at the complete set of processor tests indicates that **the Ryzen™ 7 PRO 5850U enjoys a large margin in overall processor performance with a multi-thread advantage** that will help ensure that even future applications and workflows will have more processing power to utilize.

SYSTEM LEVEL PERFORMANCE

For system level tests, a series of both synthetic and application-based benchmarks were used, including:

1. Passmark 10 Overall
2. PCMark 10 Benchmark
3. PCMark 10 Extended
4. Sysmark 2018 Rating Overall
5. Sysmark 25 Rating Overall

These benchmarks test many different aspects of the system, including processor, memory, storage, responsiveness, and graphics capabilities. As such, they are a reasonably good indication of overall system level performance.

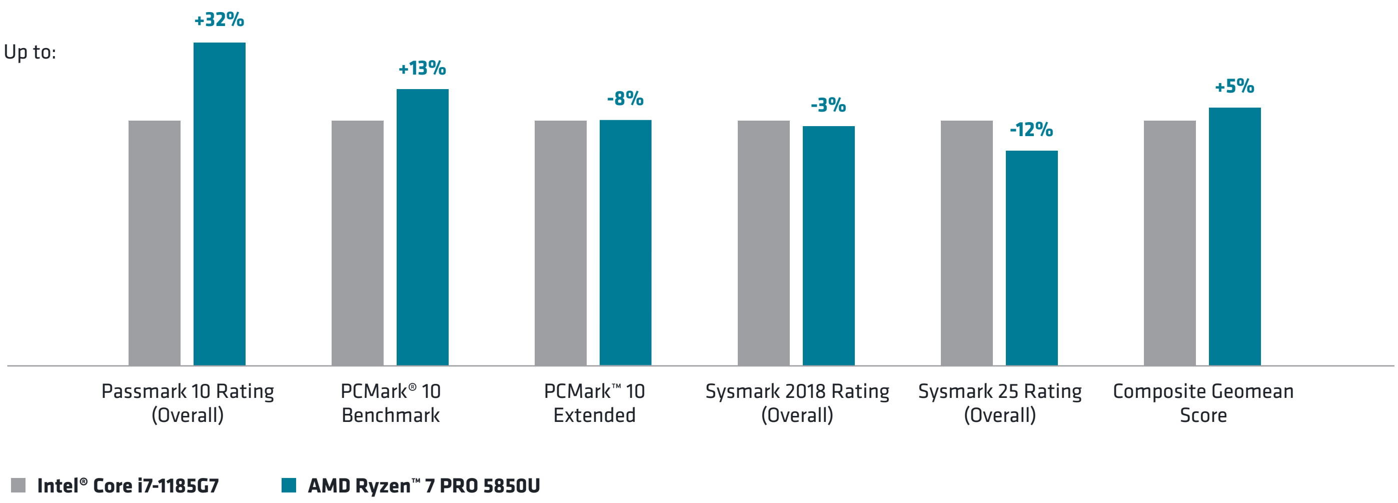
For these results, the two processors are pretty much equal (+/-5%) in two of the tests, indicating a comparable performance. **The Ryzen™ 7 PRO 5850U is significantly outside the 5% margin on several of the tests while the Core i7-1185G7 has the advantage on one of the tests.** This would indicate a **slight system level performance nod to the Ryzen™ 7 PRO 5850U** as noted by a geomean average of +5%.

APPLICATION PERFORMANCE

For application performance, a series of tests were used:

1. PCMark® 10 Productivity Test Group
2. PCMark® 10 APP Performance Overall
3. PCMark® 10 App Performance Word
4. PCMark® 10 App Performance Excel

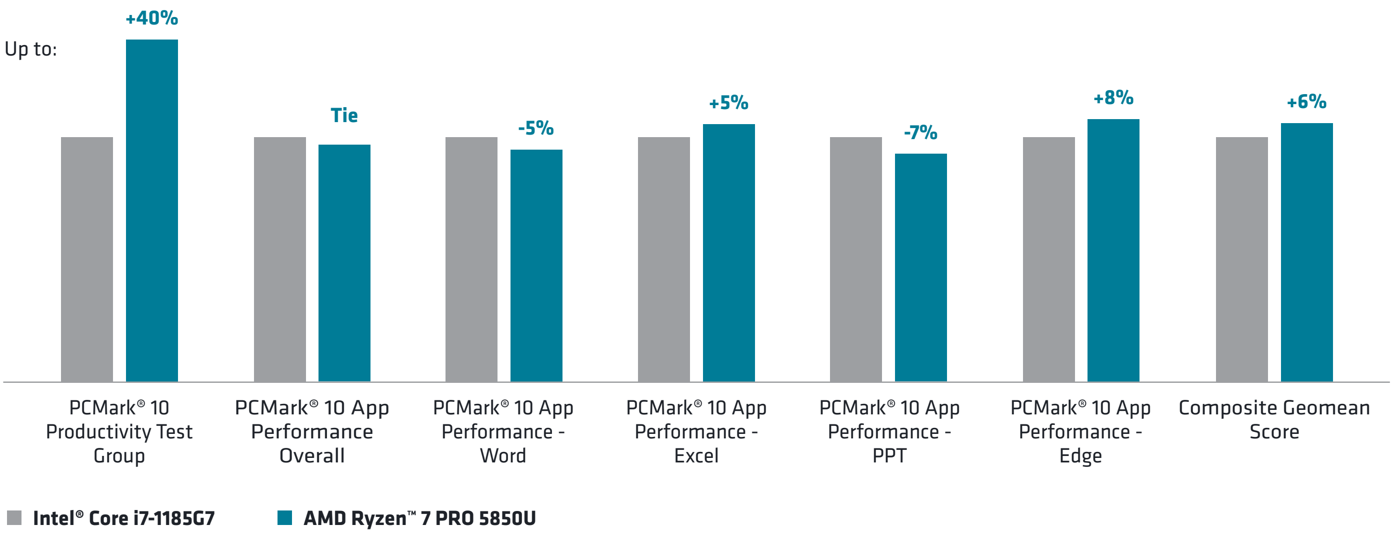
System Level Performance *Figure 6*



5. PCMark® 10 App Performance PowerPoint
6. PCMark® 10 App Performance Edge

Application Performance[1]

Figure 7



FuturMark's PCMark 10 Application tests use the Microsoft Office applications to gauge the level of productivity a user can expect from the system. This is a popular test, and a good gauge of general office productivity. In this set of tests there is an overall score that rolls up the various individual application test scores.

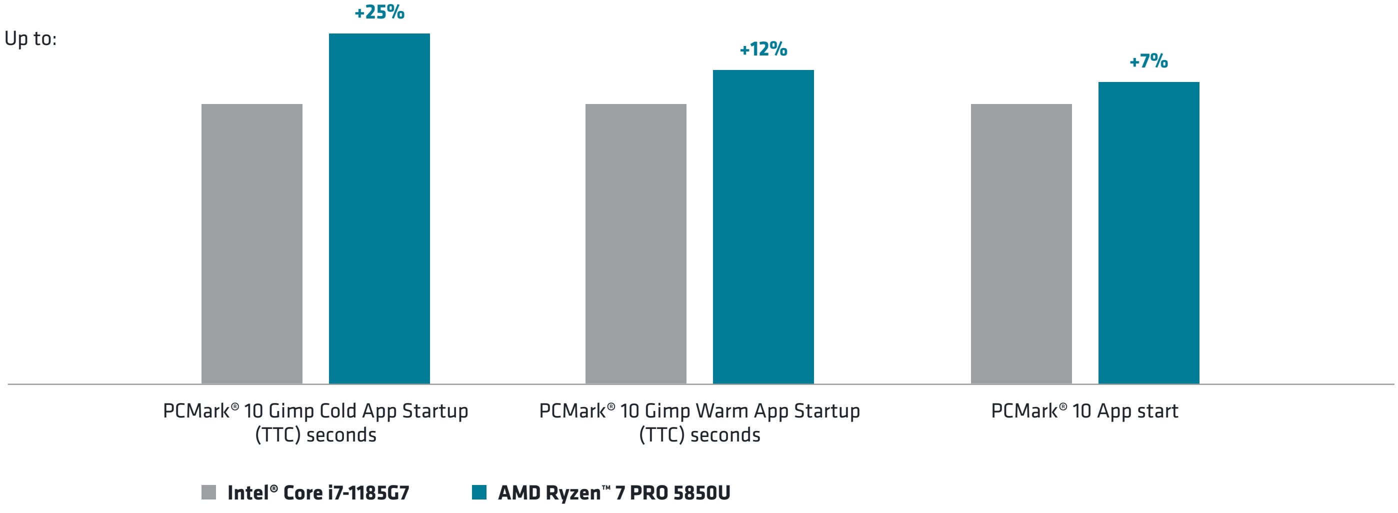
The PCMark 10 Productivity test also evaluates productivity using a different set of open-source productivity applications. These results would indicate that in terms of office productivity, **the Ryzen™ 7 PRO 5850U is slightly ahead of the Core i7-1185G7,** borne out by the geomean average score of +6%.

SYSTEM RESPONSIVENESS

It is not always just raw performance that's important; system responsiveness is also a key indicator of user satisfaction. To test how fast a system can load applications, several startup tests were used, including PCMark® 10 Gimp cold and warm startup and app start.

System Responsiveness

Figure 8



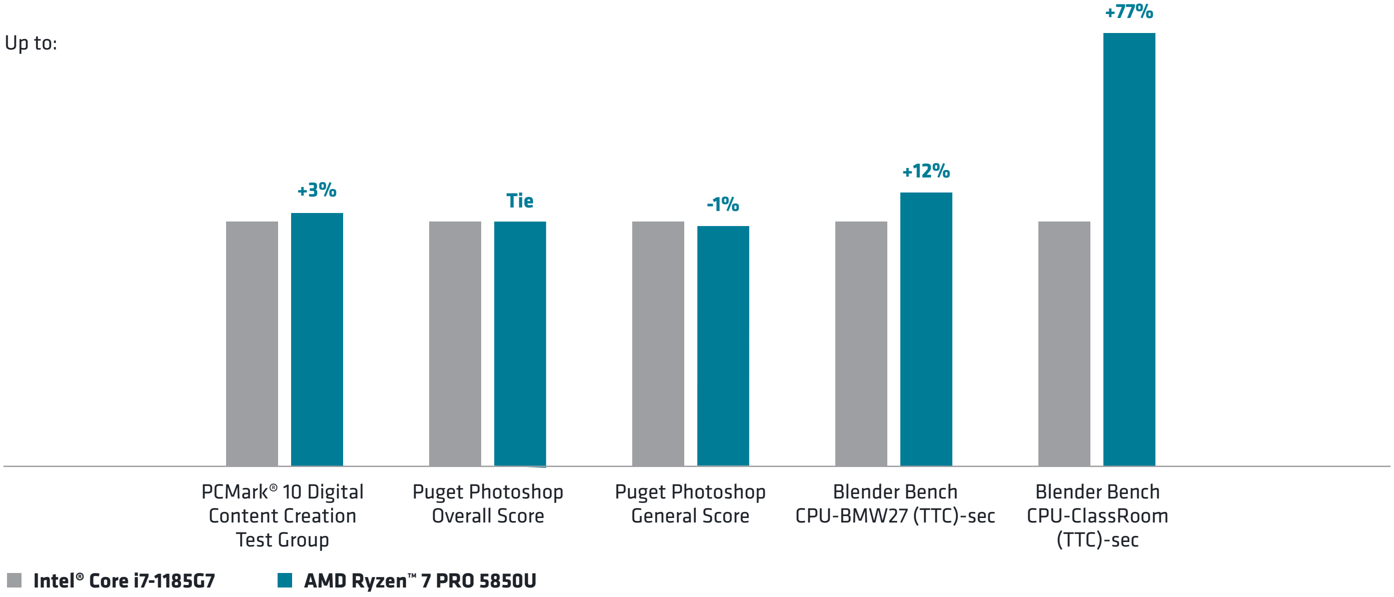
In this situation, the Ryzen™ 7 PRO 5850U shows a more responsive or "snappy" performance.

CONTENT CREATION

The final performance evaluation category is around content creation. This helps to qualify how well the processors can handle heavier or simultaneous workloads, and includes the following tests:

1. PCMark® 10 Digital Content Creation Test Group
2. Puget Photoshop Overall Score
3. Puget Photoshop General Score
4. Blender Bench CPU-BMW27 (TTC)-sec
5. Blender Bench CPU-ClassRoom (TTC)-sec

Content Creation Performance *Figure 9*



In this range of tests, the Ryzen™ 7 PRO 5850U and the Core i7-1185G7 are roughly equal except when multi-thread rendering is involved. In these tests, which can utilize **the Ryzen™ 7 PRO 5850U core and thread advantage, the lead is well beyond the 5% threshold.** For content creation tasks or high simultaneous application use, users would be well served by looking at Ryzen™ 7 PRO 5850U based systems with its high-performance core and thread advantage.

PUTTING IT ALL TOGETHER

Combining analysis across all the different benchmarks helps to give a more complete picture to compare the two processors. With a total of 25 tests and sub-tests used for the evaluation, it helps factor out any limitations of a single specific benchmark as well as test a range of attributes at the processor, system, and application levels.

Again, the analysis considers the two processors to be equal for results that are +/-5%, which is generally the error range for many benchmark tests. The breakdown would be as shown in **Table 1:**

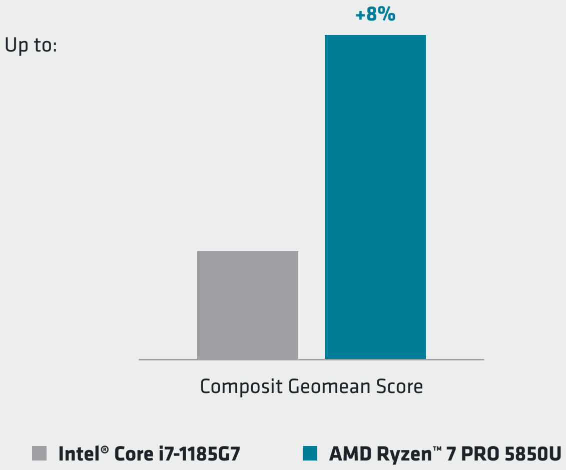
Performance Across All Tests (+/- 5%) *Table 1*

25 TESTS AND SUB-TESTS INCLUDING SYNTHETIC BENCHMARKS SYSTEM/APPLICATION BENCHMARKS	PERCENT OF TESTS WITH RESULTS WITHIN +/- 5% (CONSIDERED EQUAL PERFORMANCE)	PERCENT OF TESTS WITH RESULTS GREATER THAN 5% (CONSIDERED AN ADVANTAGE)
AMD Ryzen™ 7 PRO 5850U	36%	52%
Intel Core i7-1185G7		12%

Further, taking the geomean average for tests which report a computed score, **the Ryzen™ 7 PRO 5850U has an average of 8% better performance across a wide variety of benchmarks.**

Compsite Geomean Score

Figure 10



BENCHMARKS TEST USED FOR GEOMEAN AVERAGE

Cinebench R23 1-thread
Cinebench R23 n-thread
Geekbench v5 (5.3.1) Single-Core Score (64-bit)
Geekbench v5 (5.3.1) Multi-Core Score (64-bit)
Passmark 9 CPU Mark
Passmark 10 Rating (Overall)
Passmark 10 CPU Mark
PCMark® 10 Benchmark
PCMark® 10 Extended
PCMark® 10 Productivity Test Group
PCMark® 10 Digital Content Creation Test Group
PCMark® 10 APP Performance Overall
PCMark® 10 App Performance_Word
PCMark® 10 App Performance_Excel
PCMark® 10 App Performance_PPT
PCMark® 10 App Performance_Edge
Puget Photoshop Overall Score
Puget Photoshop General Score
Sysmark 2018 Rating (Overall)
Sysmark 25 Rating (Overall)

Based on these results with a wide cross-section of different types of benchmarks, it is reasonable to predict that **the AMD Ryzen™ 7 PRO 5850U is going to outperform or at least equal the performance of the Core i7-1185G7 in most situations.**

As such, it would be a top choice for commercial users.

IN CONCLUSION: A SMARTER APPROACH TO BENCHMARKING

Correctly evaluating performance is not a simple, one-dimensional task. There are several techniques that can be used by an organization to determine which system would best meet their needs. Since using a narrow benchmark score may lead to incorrect conclusions, the best overall picture of performance when using the benchmark strategy comes from looking at a wide range of both application-based and synthetic tests.

The final and best step in any evaluation is to allow groups of users to “test-drive” systems in their actual work environment. With a solid benchmark or application script performance used as a foundation, a trial is a final way to ensure users will be satisfied with their experience. Ultimately, this will always be the most effective way to measure real-world performance—by real-world user satisfaction.

Learn more about how the AMD Ryzen™ PRO 5000 series processors are built for a higher standard of real-world productivity.

DISCLAIMER

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions, and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability, or fitness for particular purposes, with respect to the operation or use of AMD hardware, software, or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD’s products are as set forth in a signed agreement between the parties or in AMD’s Standard Terms and Conditions of Sale.