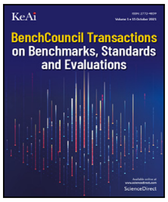
Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/



Research article

CpsMark+: A scenario-oriented benchmark system for office desktop performance evaluation in centralized procurement via simulating user experience

Yue Zhang, Tong Wu [*]

National Institute of Metrology, China

ARTICLE INFO

Keywords:
Computer benchmarks
Hardware performance evaluation
User experience
Scenario-oriented workloads
Centralized procurement

ABSTRACT

Rapid business expansion of various companies has placed growing demand on office desktops recent decades. However, improper evaluation of system performance and inexplicit awareness of practical use conditions often hamper the efforts to make a consummate selection among multiple alternatives. From the perspective of end users, to optimize the evaluation process of desktop performance in centralized procurement, we present CpsMark+, a coherent benchmark system that evaluates office desktop performance based on simulated user experience. Specifically, CpsMark+ includes scenario-oriented workloads portraying representative user behaviors modeled from the cooperative workflow in modern office routines, and flexibly adapted metrics properly reflecting end-user experience according to different task types. The contrast experiment between state-of-the-art benchmarks demonstrates high sensitivity of CpsMark+ to various hardware components, e.g., CPU, and high repeatability with a Coefficient of Variation less than 3%. In a practical case study, we also demonstrate the effectiveness of CpsMark+ in simulating user experience of tested computer systems under modern office-oriented scenarios for improving the quality of office desktop performance evaluation in centralized procurement.

1. Introduction

Computer performance used to be easily indicated by their hardware configurations. As computer architecture grows more sophisticated, nevertheless, using specifications as a metric will give an incomplete picture of overall computer performance in many practical scenarios [1]. Such an evaluation method is biased and thus cannot catch up with the rapid improvement of computer performance brought by thriving design philosophy. In addition, rapid expansion of computer markets makes it more difficult to identify the system performance.

The above obstacle gives rise to the use of various computer benchmarks. However, most existing benchmarks are unable to meet the performance evaluation requirements in centralized procurement of office computers. Micro and kernel benchmarks are constructed by repeating monotonous operations or running pivotal algorithms from synthetic workloads. These benchmarks merely reflect partial performance of a certain component in a specific system and are primarily utilized by researchers or manufacturers to pursue innovative computer design. While some newer benchmarks, e.g., Business Applications Performance Corporation's SYSmark and Futuremark's PCMark, mainly consist of common business application workloads and are more representative of commercial use, while they fail to offer an overall and scenario-oriented evaluation for general end-user experience [2]. Furthermore, they are not open-source benchmarks, thus the opacity of scoring methodology and workload operations impairs their fairness and transparency, which are essential for centralized procurement.

To address the limitations of SYSmark and PCMark, CpsMark 1.0 [3], an open-source benchmark for microcomputers was developed. However, the design philosophy of CpsMark 1.0 is not user-oriented but emphasizes workload capacity. As a result of such design philosophy, in practice, users complain that workload characterization is biased, and metric measurements are inflexible. In addition, its benchmark methodology is not designed with adequate consideration for office scenarios.

Moreover, it is difficult to precisely grasp the specific needs of end users, let alone individual preferences, especially in centralized procurement. Such inaccessibility makes it unwarranted to formulate the performance evaluation process and limits rational utilization of existing computer benchmarks.

This paper aims to solve the above problems, as well as systematically optimize the process of utilizing benchmarks to evaluate the office desktop performance in centralized procurement. Specifically, we have redeveloped CpsMark+, a novel and coherent benchmark

[*] Corresponding author.
E-mail addresses: zhyue@nim.ac.cn (Y. Zhang), wut@nim.ac.cn (T. Wu).

https://doi.org/10.1016/j.tbench.2023.100084
Received 8 June 2022; Received in revised form 28 December 2022; Accepted 1 January 2023
Available online 5 January 2023

system that builds a bridge between system performance and simulated user experience in intended usage scenarios, i.e., daily working scenario in modern office. Extensive experiments on multiple real-world tested systems demonstrate high sensitivity and repeatability of CpsMark+ results. Then we used CpsMark+ as a substitute for hardware specifications in quantitatively evaluating the overall computer performance of responsive bids in a real case of centralized procurement. Experimental results show that user experience ratings of the desktops selected by better benchmark score are significantly higher than those selected by the original bid evaluation method, which indicates the effectiveness of CpsMark+ in simulating user experience under modern office-oriented scenario for office desktop performance evaluation in centralized procurement.

The rest of this paper is organized as follows: Section 2 reviews related work and provides our motivation for developing CpsMark+. Section 3 summarizes the challenges in evaluating office computer performance in modern office scenario for centralized procurement. Section 4 describes our methodology and process in developing CpsMark+, as well as extensive experiments for evaluating and comparing CpsMark+ with other related works. Section 5 presents a case study of centralized procurement where we demonstrate the effectiveness of using CpsMark+ as a computer benchmark to simulate user experience in daily office scenario for desktop performance evaluation. Section 6 concludes our work and elicits possible research directions in the future.

2. Background

2.1. Existing benchmarks and metrics

We have reviewed some related works proposed for computer performance evaluation, while most of them have limitations in benchmarking office desktops under modern office scenario for centralized procurement or have not even been designed for commercial use.

SYSmark 2018 [4] adopts real-world third-party software as workloads to evaluate overall computer performance and is widely applied in commercial markets. Usage scenarios are modeled in the form of subjectively grouped job nature like productivity and creativity, which cannot describe cooperation across tasks in a common workflow. In terms of the workloads, most of them are designed to be CPU-intensive and place little pressure on GPU and storage system, making the evaluation insensitive to graphics and I/O performance that might be cared by end users in daily use. Further, system responsiveness and program start-up are isolated and measured by specific applications, thus weakening the realistic reference value of benchmarking results.

PCMark 10 [5] reports an overall score calculated by the geometric mean of tested metrics for the inclusive workloads within each test group. The geometric mean returns a normalized score that treats the performance of each workload equally. This scoring methodology outputs a balanced result of performance evaluation, which neglects the diversity of importance of different workloads and is unable to describe real user experience in a specific scenario.

Phoronix Test System [6] is an open-source and extensible benchmark system that evaluates comprehensive performance of multiple platforms. It includes hundreds of test programs covering a wide range of applications to evaluate various metrics. Nevertheless, the contributors provide little information about benchmarks' logic and internals, especially on how each system is tagged and applied for specific components [7]. Moreover, the benchmark system has numerous functionally overlapping programs for identical system parts and requires complicated dependencies, which makes them too generic and inefficient to be used in centralized procurement.

There are other benchmarks targeting specific application domains. 3DMark [8] mainly describes real-time gaming performance of graphic cards, its dependence of frame rate as the only metric limits further uses in other fields [2]. SPEC CPU 2017 [9] contains a series of floating-point and integer algorithms extracted from the kernel of compute-intensive applications to evaluate the computing performance of CPUs. The workloads are synthetic and biased, making them more suitable for simulative experiments in academic research and industrial development of processors. The Stanford SPLASH benchmark system [10] evaluates parallel algorithms for shared-memory multiprocessors with real scientific workloads, which is of little use for office routines. Micro benchmarks such as STREAM [11] and lmbench [12] solely test single metric like memory bandwidth or latency of individual hardware component through monotonous program operations, which makes them disregard resource allocation and coordination of mixed workload manipulations within the entire computer system [13].

2.2. Our motivation for upgrading CpsMark+

To address the mentioned limitations of SYSmark and PCMark, we released the microcomputer benchmark CpsMark 1.0 in 2014, which evaluates processor performance based on a series of CPU-intensive workloads abstracted from typical computing scenarios [3].

However, the design of CpsMark 1.0 mainly focuses on workload capacity, instead of reflecting end-user experience. The workload operations are designed to be CPU-intensive and isolated from each other, thus it cannot reflect overall performance and user experience in real scenarios, which by contrast, requires workloads to be coherent and interactive. The scoring methodology treats each workload equally and neglects diverse importance of them in practical tasks. In addition, the third-party software used as workloads and the operating system (Windows 7) supported by the benchmark is obsoleted in burgeoning computer-related markets. Generally, these drawbacks merely make CpsMark 1.0 a simple technical reference for an individual customer, while it is powerless to help make purchase decisions according to actual requirements in centralized procurement of office computers.

Over the last few years, the role of benchmarks has been in the spotlight in purchasing computers. Some organization like Bitkom, a Germany's digital association, has proposed the use of benchmarks in tendering of computers [14]. Intel has also recommended some existing benchmarks as the criteria for screening the shortlist from bidders [1]. Inspired by such evolving roles of benchmarks, we redesigned CpsMark 1.0 to a coherent benchmark system by utilizing simulated end-user experience under office-oriented working scenario for better performance evaluation in centralized procurement and finally developed CpsMark+ in 2019.

3. Challengs in evaluating the system performance of office desktops

3.1. Evolution of computer architecture and usage

Researchers and consumers used to compare the performance of diverse computer systems by merely inspecting their hardware specifications. Latency and throughput used to be typical metrics that served us well in computer performance evaluation, since only the size and the content of input data could affect the processing speed of applications at that time [2]. For the sake of performance evaluation, better hardware always led to higher throughput and lower latency so that computer architecture was merely an inorganic combination of individual components.

As computer architecture and usage grow more sophisticated, simple information of computer configurations hardly predicts the program performance in disparate scenarios explicitly [15]. Such transform gradually gives rise to the thriving of numerous benchmarks, which are a system of objective test programs that return the normalized test score compared with the baseline platform by running a series of identical applications or other computer operations. These benchmarks are generally designed to mimic a particular type of workloads on a constant computer system, by which people can be able to compare the performance of alternative computers under the specific working circumstance.

Nevertheless, modern computer applications increasingly interact with humans, the physical world, and each other—often simultaneously. Some new types of computing tasks like heterogeneous computing [16], for example, can classify different subtasks based on the embedded code segments and automatically assign them to the most suitable computing resources for efficient execution so that the total time consumption of the entire task is minimized. Many tasks operate in parallel and compete for resources internally, which might be a stochastic process and lead to dynamic results. Complicated interactions among tasks, hardware, and humans make it difficult to describe the entire performance of a given system according to a single task or even multiple tasks executed in isolation [2]. Generally, the overall performance of modern computer systems is not solely a function of individual hardware and executed applications, but an intricate integration of hardware architecture, the pattern of software execution and resource allocation, and how humans interact with computer systems [17].

3.2. Obstacles to capturing usage requirements in centralized procurement

Effective evaluation process of computer performance must be built upon an explicit awareness of the intended usage scenario of tested systems, nevertheless, which is especially difficult to obtain for office desktops.

Evaluation of office desktop performance is often massively required in centralized procurement, which is a long and strenuous process where only the opinion of authorities dominates the purchase decision-making. Hence, the decision-making process is usually distant from real stakeholders [18], e.g., the internal customers or the external clients in the case of outsourcing work. The principal of procurement and bidding documents are intensively formulated by management and hardly reflect how procurement items are intended to be used in practice.

Even in the case of individual purchase, compared to traditional electronic products, information of potential usage for modern desktops is still not easy to be directly referenced in the process of performance evaluation, due to the all-round functions and flexible use of modern computers. A game enthusiast who is keen on 3D games, for instance, might also pay attention to computational performance required by a software engineer. Hence, it is hard to capture explicit usage requirements of modern office desktops, which impedes the effective evaluation of system performance, highlighting the importance of how computer benchmarks can precisely reflect end-user experience in specific scenarios.

3.3. Difficulties in reflecting real user experience

In various business domains, a questionnaire is one of the most direct ways to obtain user experience and satisfaction, while like many other similar surveys, it can merely be conducted after the durable use of real end users in practice, which makes it less time-efficient in helping vendors improve their products before releasement or serving as reference when customers are selecting new products. In the field of computers, the rise of various benchmarks solves part of the above problems, however, huge challenges lie in how to precisely reflect user experience without manual intervention.

For a specific computer product, the usage of different potential customer groups could be divergent, which requires an accurate match between benchmark workloads and actual user behaviors. Also, each user may have a different standard in evaluating computer performance, depending on the using habit or product dependency. This phenomenon will influence the perceived user experience without any doubt, and thus requires more considerate design of metrics and scoring methodology. Finally, it is not possible to consummately reflect user experience of computer products with any individual benchmark, because the possible over-specific design will cause the benchmark over-fitting and makes it less applicable for wider use. Therefore, the trade-off between pertinence and universality of benchmarks is also pivotal.

4. The CpsMark+ benchmark tool

In this section, we describe relevant criteria, methodology, and process for developing CpsMark+ in detail. We also carry out analytical and comparative experiments with respect to typical characteristics of computer benchmarks.

4.1. Criteria and design features

Researchers have been theoretically exploring the art of building a consummate benchmark [19,20]. Kistowski et al. [20] assert that all standardized benchmarks are subject to a group of universal criteria, e.g., relevance, repeatability, fairness, and verifiability, which are proved to be necessary. However, in each domain, the criteria are expected to include additional features specific to individual benchmark, depending on its goal, intended usage scenario or other considerations.

The essence of benchmarking office computer performance under daily working scenarios for centralized procurement is to properly evaluate computer systems from a perspective of user experience and describe system performance according to specific purchase demand. In this paper, we propose following benchmark criteria that guide the design of CpsMark+'s features:

- Applications and software manipulations should be scenario-oriented to reflect real user behaviors. Particularly, in centralized procurement, end users can hardly have significant influence on the purchase decision made by authorities, hence the workloads should be closely correlated to behaviors or intended usage that are of interest to end customers in many aspects, e.g., the workload characterization and the input data set.
- Cooperation and diverse importance across tasks should be described. End users usually do not have equal performance requirements for all tasks or even applications involved in an individual task. In practice, if several applications operate towards a common task or purpose, sequence and coherence of them will impact the general working efficiency, since the acceleration of some applications might be more beneficial than that of others.
- Design of metrics should be flexible and account for nonlinearity, which means that composite metrics should not weigh all applications equally. Considering complicated usage of modern desktops, desired metrics of different workloads may vary. In terms of human interaction, for example, a human cannot perceive faster response time beneath some threshold. While for some other tasks, the diversity of execution time on various systems can be ignored.
- The benchmark should be open-source and vendor-neutral. Development of closed-source benchmarks is likely to be manipulated by certain vendors through biased workload design, leading to suspicion [21] and loss of credibility. An open-source benchmark enables public supervision and guarantees the fairness of benchmark results, which is significantly crucial in centralized procurement.

4.2. Entire development process and benchmark framework

Unlike most computer benchmarks, CpsMark+ is designed to be used in centralized procurement, where one single benchmarking result could affect the purchase and use of a specific product for crowds of employees. Hence, during the development process, it is more beneficial to follow an iterative and incremental strategy, instead of top-down principles that formulate schemes at an early stage to make

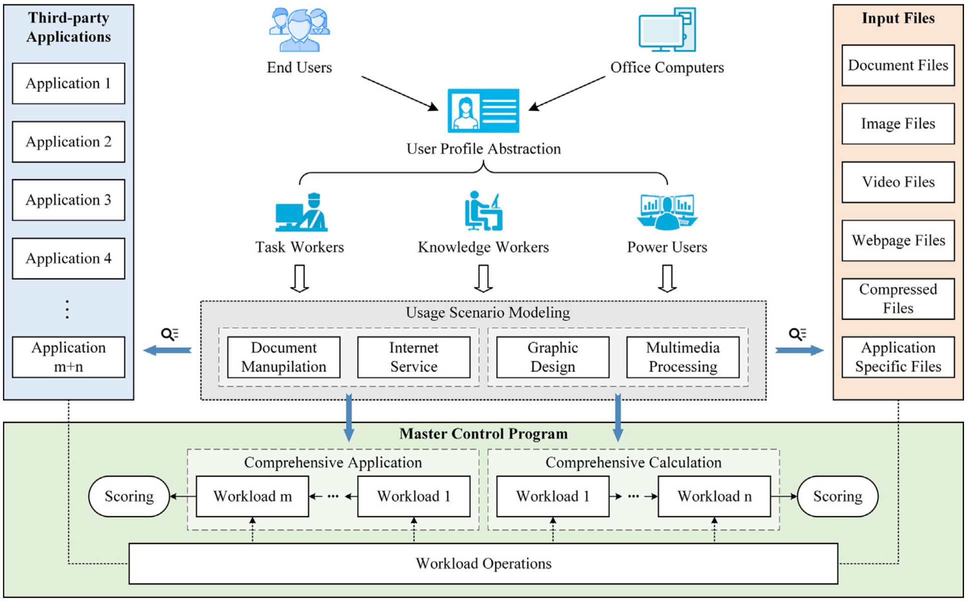


Fig. 1. The main software components and the overall benchmark framework of CpsMark+.

subsequent design right on track. We divide the entire development process into phases, which are associated with relevant checkpoints to guarantee the accomplishment. Within each phase, requirements are elicited from various end users through market research or consultation, then representatives are selected to give feedback on the outcomes of decision-making and implementation. We improve our work based on the feedback and repeat such procedures for each phase. Based on the criteria proposed in Section 4.1, the main software components of CpsMark+ and its overall benchmark framework are depicted in Fig. 1.

CpsMark+ benchmark tool contains three components:

- The automatic setup program, which installs third-party applications and the Master Control Program (MCP) in batches. MCP is responsible for benchmark execution, including test initialization, resource extraction, data integrity check, workload execution, log recording, metric measuring and calculation, and report generation.
- The resource package, including the input files of workload operations.
- The third-party application package, which contains the setups of all third-party applications.

The source code of MCP is maintained online at https://github.com/wanghong3116/CpsMarkPLUS, which is still under further improvement and subject to change. The resource and third-party application packages have been uploaded on the website of National Metrology Data Center of China, which can be accessed online through https://jc.nmdc.ac.cn/view-40-609748.html. Note that CpsMark+ only supports Microsoft Windows 10.

We have not integrated input files, workload applications and the MCP into a unitary package as most commercial benchmarks, which makes our work transparent and easy to be maintained. For the first use of CpsMark+, the trial version of each third-party application is automatically installed on the tested computer system and configured by the execution of an automatic setup program. Likewise, each workload runs independently in the form of complete software, the corresponding application is not merged into the MCP and only receives instructions synchronously from the background of tested computer systems. Such design reduces the influence of the MCP on system performance and enables a clear view of workload conditions provided by logs.

The MCP is devised as a serial layout and contains two separate test modules. Users can initialize the number of iterations to run for eliminating fluctuation of benchmark results. Composed of a sequence of orderly executed workloads, each module independently generates a synthetic score that reflects the performance of inclusive workloads. There is an automatic reboot of the tested computer system between the two modules for eliminating the impacts of varying system status (e.g., cache) on module independence.

4.3. *Workloads*

CpsMark+ has two independent modules for simulating user experience perceived in modern office scenarios, i.e., Comprehensive Application (CA) and Comprehensive Calculation (CC), which can be optionally selected and run independently during the test. Each of them has a series of workloads executed in a specific order. In this section, we will introduce design and characterization of the workloads within each module in detail.

4.3.1. *User profile abstraction of office computers*

Chen et al. [22] point out that benchmarks are expected to be associated with real application domains and mirror practical demands in subsistence. Although a large employer may have numerous user segments, appropriate classification could minimize complexity and throw more light on exploring the performance requirement of specific user segment. For the daily usage of desktop computers in modern office scenarios, we abstract the profiles of end users from the perspective of occupation and profession described in Table 1.

Since CpsMark+ has been designed for commercial evaluation of desktop computers used in modern office scenarios, the user profiles summarized in Table 1 exclude those working in laboratories, R&D centers, factories, or telecommuting. In this paper, we mainly focus on most knowledge workers and some part of power users.

4.3.2. *Usage scenario modeling and application selection*

Employers in a specific department of a company are likely to engage in fixed routine work, thus the performance requirement of a specific task in a homogeneous work section should be more emphasized in centralized procurement of office computers. To highly correlate the design of workloads with the oriented usage scenario of tested computers, we focus on exploring the usage models of intended end users working in daily office scenarios.

According to the abstracted user profiles of office computers, we cluster the usage models into four groups of common office scenarios

Table 1
The profiles of computer end users.

User category	Representative occupations	Performance requirement
Task workers	• Customer service • Front desk consultation • Bank clerks • Data entry specialist • Human resource	• Basic document operations • A single OS-level application • Simple connectivity needs • Static 2D graphics • Few computing occasions
Knowledge workers	• Most students • Teachers and professors • Company administrators • Financial advisors providing multiple advice • Product managers presenting prototypes from multi-angle	• Content creation • Frequent web browsing • Moderately complex application • Moderate scientific computing • Variable multimedia processing like graphics and video • Adequate memory
Power users	• Multimedia designers making high-definition video • Professional architects engaged in complex modeling • Physicians examining delicate 3D medical images	• Complex content creation • Intensive video and 3D graphics processing • Heavy CPU computing • Fast system response • Smooth running of applications

Table 2
The application selection of workloads.

Module	Usage scenario	Application	Version
Comprehensive application	Document manipulation	Microsoft® PowerPoint Microsoft® Word Microsoft® Excel Adobe® Acrobat WinRAR	2016 (16.0.4266.1003) 2016 (16.0.4266.1003) 2016 (16.0.4266.1003) DC (19.010.20091) 5.91 (64-bit)
	Internet service	Google® Chrome Microsoft® Outlook	73.0.3683.75 2016 (16.0.4266.1003)
Comprehensive calculation	Graphic design	Autodesk® AutoCAD Adobe® Photoshop	2018 (22.0.49.0) CC 2019 (20.0.1)
	Multimedia processing	Autodesk® 3ds Max Adobe® Premiere Adobe® After Effects HandBrake	2018 (20.0.0.966) Pro CC 2019 (13.0) CC 2019 (16.0) CLI 1.3.0

based on their overall functions within a specific workflow, i.e., document manipulation, Internet service, graphic design, and multimedia processing, which are described as follows:

- The document manipulation scenario contains multiple manipulations towards the documents in common formats, which are involved in most cases of modern business.
- The Internet service scenario mainly includes web browsing and email creation, which are usually auxiliary means in resource acquisition and information communication.
- The graphic design scenario refers to visual expression of ideas and information through the combination of symbols, pictures, and text, which is crucial for product presentation tasks like poster production.
- The multimedia processing scenario relates to utilizing computers for digitizing and integrating graphics, sound, video and other media information in a specific interactive interface, which is widely applied in consulting, marketing and management.

As for workload applications, we select desktop-level office applications based on the metric of popularity. According to the investigation report of office software markets in China by Chinaiern [23], our software market experts select popular and typical applications for each usage scenario in modern office, which are summarized in Table 2.

Since sufficient time is required for workloads to be developed and validated, versions of some applications are not the latest when CpsMark+ was released. In addition, the intended applications of CpsMark+ are the most widely used version instead of the latest one. While some application like WinRAR is up to date because it is feasible to be instantly updated by end users.

4.3.3. Test module construction

While specific selection of usage scenarios ensures high representativeness of the benchmark, grouping applications with similar performance dependencies from various usage scenarios can easily provide an all-sided picture portraying integral performance required by end customers and enhance the usability of the benchmark. Hence, we merge the usage scenarios into two separately running and scored modules as follows:

- Comprehensive Application (CA) module includes the scenarios of document manipulation and Internet service, which reflect light and middleweight use by task or knowledge workers in most business workplaces, where end users might pay more attention to overall performance, response, and smoothness throughout regular use.
- Comprehensive Calculation (CC) module includes the scenarios of graphic design and multimedia processing, which reflect heavyweight use by power users skilled in professional fields, where end users possibly focus on the execution efficiency of CPU-intensive or GPU-intensive computing tasks.

Within each module, in addition to similar performance dependencies, the usage scenarios are highly correlated and tend to appear in a common workflow under daily office scenarios. Further, each usage scenario is given a different weight based on the sum of metrics measured from inclusive workloads. Such approach can ensure a direct and close connection between benchmark results and computer performance required by end users.

4.3.4. Workload components and design details

To reflect the user experience of office computers in modern office, workloads should be not only scenario-oriented but also capable of simulating user behaviors. Therefore, the workload of CpsMark+ is more than a concept of application automation, but a logical integration of three elements: the input data set extracted from the resource package, the workload operations performed on the input data set through the applications executed by the MCP, and the generated output.

For each workload, the input data set is chosen to functionally reproduce the resources or materials that might be used by end users in modern office scenarios. Specifically, we select raw digital contents or semi-finished project files that are mainly non-structured data such as texts, images, videos, webpages, and other application-specific files, e.g., 3dsMax scene files.

Then we explore basic operating units that frequently appear in the routine use of applications and integrate them into a series of workload operations that can accomplish a common task. We guarantee the completeness of workloads via designing diversified operations that independently generate finished files as output for each application. Moreover, there is no random process in the MCP so that the generated output is uniquely determined by the input data set and the workload operations.

The workload operations of the CA module are briefly described in execution order as follows:

- **Google Chrome.** Simulate users to browse webpages and switch between tabs. Webpages are accessed through locally configured network services. The webpages contain text, pictures, JS (JavaScript) scripts, and flash.
- **Microsoft PowerPoint.** Set the new template style and create slides. Input texts and adjust character formats, alignment, and font size. Add pictures, captions, and typeset. Insert tables and charts with filled data. Browse slides.
- **Microsoft Word.** Input characters, modify titles and character formats, split paragraphs, set the directory, insert pictures, create tables and charts, input data.
- **Microsoft Excel.** Generate and organize data with fixed formula. Classify and enter data under a specific rule. Calculate and sort common statistics. Draw line charts by categories, set titles and styles, adjust size and position. Macro definition and execution.
- **Adobe Acrobat.** Convert PowerPoint, Word, and Excel documents made in previous workloads to PDF files, browse these PDF files page by page.
- **WinRAR.** Compress and decompress mixed files in multiple formats, including images, videos, documents, databases, and log files.
- **Microsoft Outlook.** Simulate users to receive, browse email contents and attachments offline, including Word, Excel, and PowerPoint files. Upload new attachments, edit the body of the email, and reply.
  The workload operations of the CC module are briefly described in execution order as follows:
- **Adobe Photoshop.** Use the PSD (Photoshop Document) file to make a vertical poster. Separate target area from the source material and design the layout of layers. In new layers, set titles and captions, add a logo, and adjust its size, coordinates, and transparency. Combine all layers, virtualize the background and merge them into a large picture.
- **Autodesk AutoCAD.** Use the DWG file to draw distributed structure diagrams of buildings. In the main framework, draw structure and vector identification of each area, add coordinates, and mark the size. Change colors of layers and use different line styles. Design wiring, draw pipeline distribution, and flow direction.
- **Autodesk 3ds Max.** Design a 3D model of a whale. Develop the 3D framework, color the texture, add lighting effects, make reflections and shadow effects by calculating light source position, incidence angle, and reflection angle. Produce motion trajectories and movements of the whale model, render segmental frames of action sequences.
- **Adobe Premiere.** Clip and splice source video materials, add lens transition and subtitles, synthesize sound effects, render, and preview the output video.
- **Adobe After Effects.** Add particle explosion effects, render the firework explosion animation sequence of 1800 frames and 30 FPS.
- **HandBrake.** Convert the H.264 encoded source video with 4K resolution to the H.256 encoded target video with 2K resolution, the container format is MP4. Hardware acceleration will be leveraged if enabled.

Within each module, the workloads are executed in the order specified above. The format or even the content of the generated output for some specific applications is identical to that of the input data set for subsequent applications. Such design enables test modules to describe cooperation across tasks throughout a common workflow. For example, the workloads of the CA module simulate the following coherent user behaviors: resource preparation via the Internet, content creation, document processing, and email delivery.

4.4. Metric design and test implementation

Although work efficiency is a pervasive metric in most benchmarks that evaluate computer performance [24] and is widely referenced in helping customers making decisions, unitary metric design may not tell the true story of user experience for the following reasons.

First, people do not have equal performance requirements for all tasks or even for the same portion of an individual task, so that user experience is usually diversified and varying. For instance, professional designers in an advertising agency might pay more attention to the time consumption of multimedia processing, while the user experience of office secretaries is closely related to the response speed and the fluency of frequent document operations.

Second, the perception of user experience is nonlinear and difficult to quantify. In terms of human interaction, humans cannot perceive faster response time beneath a certain threshold, hence further acceleration of the task will not bring better user experience. For example, a frame rate that exceeds the support of a monitor will no longer improve the user experience of a graphics task, while in this case the program execution could be accelerated by a better GPU.

As a result, in the context of CpsMark+, we define work efficiency as the time consumption for systems under test to complete all operations related to user experience within a specific workload, i.e., application launching, input files loading, and basic operating units, which are outlined in Section 4.3.4. Then we take the defined work efficiency as the metric of CpsMark+ and focus on how it can be measured to properly describe the user experience of tested desktops in modern office scenarios.

4.4.1. Method of sampling

To guarantee the pertinence of the metric, CpsMark+ adopts multiple methods to sample the work efficiency of tested computer systems, depending on various workloads. Such a flexible approach can differentiate the user experience by matching the usages of applications with their performance requirements. To be more specific, we predefine runtime as the time spent by each basic operating unit that actively uses system resources, while response time is the time interval between task activation and task completion. The sampling methods are illustrated in Fig. 2.

In terms of the workloads in the usage scenarios of document manipulation (WinRAR excluded) and Internet service, basic operating units are numerous and densely distributed with lightweight resource consumption. Some intervals of them consist of events irrelevant to the evaluation of user experience e.g., temporary retention of screen display, timer interference, which will have an adverse influence on the effectiveness of workloads if they are included in the metric.

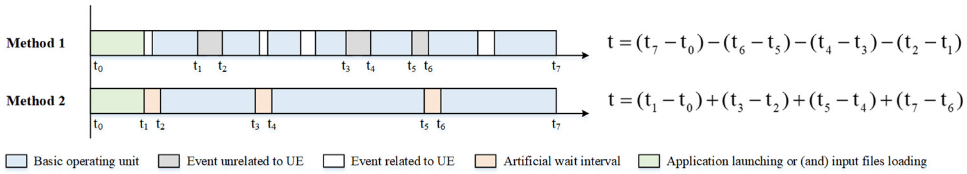

Method 1
t₀ t₁ t₂ t₃ t₄ t₅ t₆ t₇
t = (t₇ − t₀) − (t₆ − t₅) − (t₄ − t₃) − (t₂ − t₁)
Method 2
t₀ t₁ t₂ t₃ t₄ t₅ t₆ t₇
t = (t₁ − t₀) + (t₃ − t₂) + (t₅ − t₄) + (t₇ − t₆)
Basic operating unit
Event unrelated to UE
Event related to UE
Artificial wait interval
Application launching or (and) input files loading

Fig. 2. The two methods of sampling the designed metrics.

However, too many samplings of basic operating units will accumulate the sampling error and cause frequent switches between transient-state and steady-state of program process, which might interfere with the system performance. Hence, we sample the start timestamps and the end timestamps of the entire task and calculate its response time, i.e., $t_7 - t_0$ in Method 1, then we sample the time intervals of irrelevant events and subtract them from the response time as the metric of these workloads.

For the other workloads of CpsMark+, their basic operating units are relatively sparse and have a high concentration of resource consumption. These basic operating units are time-consuming and contribute most of the entire task. In this case, the user experience of end users is more susceptible to the execution speed of a single operation. To accurately measure the runtime, we artificially add extra short waits, e.g., $t_2 - t_1$ in Method 2, between the heavyweight operating units to reset the resource consumption. Finally, we sum the sampled runtime of each basic operating unit as the metric of these workloads.

4.4.2. UI-level vs. API-level automation

Benchmark implementation has a great impact on the test results of the designed metric. There are two primary approaches to automate the execution of workloads, i.e., UI-level and API-level [25,26]. Some benchmarks leverage automated scripts like AutoIt to initiate and navigate applications by simulating mouse clicks or keystrokes [25]. The duration of each task is measured when the completion of the task is detected by application-specific methods. Such an approach mimics practical human interaction at UI level, nevertheless, it instead impedes the accurate reflection of user experience for performance evaluation. Although the estimation of user experience is somewhat subjective, it should be highly relevant to how well computer systems react to or execute the instructions of real end users, however, which might be distorted by a contradictory combination of simulated user behaviors and computer-based metrics.

We choose independent APIs or invoke them from application communication standards, e.g., Component Object Model, to automatically control the execution of each workload. In this case, launching of applications, loading of input files, and basic operating units are implemented through a set of functions, methods, and procedures contained in selected APIs or standards. Compared to the UI-level implementation, our decision to choose API-level implementation provides some tangible benefits as follows:

- **Reduction of irrelevant time measured as metrics.** On the one hand, it takes UI-level implementation a large amount of time to detect the completion of tasks according to the returned signals. For instance, automated scripts may wait for the application to show a pop-up window or may wait for a dialog box to disappear, which requires accurate technical identification. Such a judgment process based on automated scripts is quite time-consuming and significantly falls behind the completion of tasks as perceived by end users. On the other hand, some workload operations themselves take much time for automated scripts to perform. For example, text input might be simulated by continuous keystrokes at a fixed speed, which has identical time consumption on all tested computers. This prolonged simulation accounts for a large proportion of the designed metric and makes the results of measurement diluted by what end users do not value.
- **Less resource consumption and higher test efficiency.** Although some UI-level automated frameworks of benchmarks claim to be lightweight and have little influence on performance, they still consume more computing and memory resources than API-level automation [27]. In addition, API-level automation requires fewer codes to perform and does not need to deal with interface elements. This attribute makes performance evaluation a faster and compact test process and further reduces the overall resource consumption.
- **Greater stability in testing and maintenance.** UI-level automation sometimes gets stuck or goes into endless loops due to UI complexities. For instance, a mouse cursor might miss certain buttons due to the change of resolution, or an unexpected window display may lead to wrong recognition. Some applications are event-driven and can easily enter idle states if there are no users interacting with them [2]. By contrast, API-level automation can guarantee the exact execution of each workload operation and help ease maintenance difficulties brought by external factors [28], e.g., frequent updates of application versions.

4.4.3. Pipeline of metric testing

In CpsMark+, test of the designed metric for a specific workload is performed through the MCP and follows a similar pipeline across all workloads as shown in Fig. 3.

More concretely, for the Nth workload, the MCP first decompresses the resource package and extracts the exclusive input files to a specified location, then an MD5 [29] check is performed towards them to ensure the data integrity. If the MD5 check fails, the test will abort and return to the initialization phase, otherwise, the MCP will move forward to the application execution phase depicted as the dashed rectangle in Fig. 3, where the designed metric T_N is tested. When all the workload operations are finished, an MD5 check is performed towards the generated output. Finally, after a five-second countdown, if there is no user input to interrupt the test, i.e., mouse clicks on the pause button, the MCP will proceed for the next workload until the entire benchmark is completed.

It is worth noting that for the workloads in the usage scenario of document manipulation and Google Chrome, the applications are launched through direct open of the input files, while for the workloads in the usage scenarios of graphic design, multimedia processing, and Microsoft Outlook, the input files are loaded after separate launch of the applications. As a crucial factor affecting the user experience, the speed of application launching is a good indicator of memory and storage performance.

4.5. Scoring methodology

The scoring methodology of benchmarks integrates test results of the designed metric and generates quantified scores that evaluate the overall performance of computer systems. For a commercial benchmark used in centralized procurement, the scoring methodology should provide accurate estimation of the user experience for tested computers to help authorities choose better products from alternatives. For CpsMark+, the design of its scoring methodology meets the following criteria:

- The resulting score does not have significant fluctuation and can remain steady given a constant computer system.

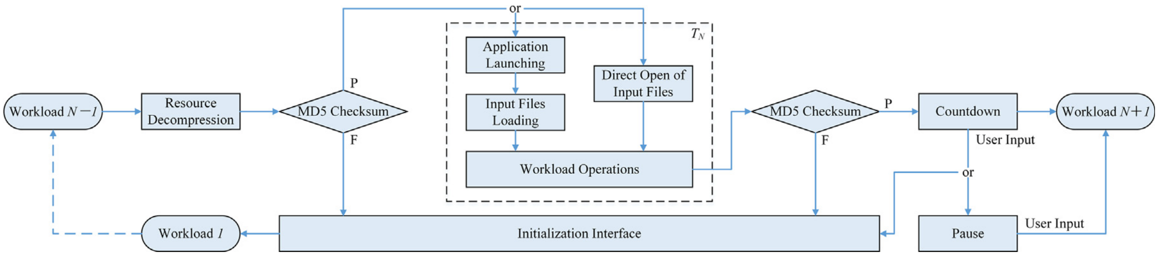


Fig. 3. Intra-workload and inter-workload pipelines of metric testing in CpsMark+.

- The resulting score can sufficiently differentiate the user experience of tested computers with diverse performance.
- The pair-wise relationship between the resulting scores from different computer systems is neutral to the calibration method and the specification of the baseline platform.

Concretely, for each module, we sum the tested metric of each included workload executed on the tested computer system and compare it with the sum of workload metrics tested on the baseline platform. We calculate the ratio value of these two sums and round it to the nearest integer. In this case, a higher score indicates better performance. To be more specific, given the ith module and the number of included workloads N_i, T_j and t_j are the tested metric of the jth workload executed on the tested computer system and the baseline platform, respectively. The resulting score for module i is calculated as follows:

$$S_i = \left[1000 \cdot \frac{\sum_{j=1}^{N_i} t_j}{\sum_{j=1}^{N_i} T_j} \right]$$

Note that we do not take the geometric mean of each score as the overall rating, which places equal weight for each module [30]. Instead, we reserve and separate the score so that end users can flexibly customize the weight of each module when they refer to the benchmark results according to diversified requirements. Within each module, the sum of each tested metric reflects cooperation across workloads and different performance dependencies of them.

4.6. Baseline platform and calibration

As the datum point of the evaluation framework, baseline platforms are prerequisite for most benchmarks. Judicious choice of the baseline platform is of great significance for the resulting score. For instance, an exorbitant configuration of the baseline platform will lead to low sensitivity and weak differentiation of benchmarks, while an inferior one may cause poor repeatability. Hence, at the time of development of CpsMark+, we study the mainstream configurations of office computers purchased in centralized procurement and determine the following configuration for the baseline platform based on performance requirements of the workloads in CpsMark+:

- CPU Model: Intel® Core™ i3-9100 (4 cores, 3.60 GHz, 6 MB L3 cache)
- Graphics: Intel® UHD Graphics 630
- RAM: Kingston® ValueRAM™ 8 GB DDR4 2400 MHz
- Storage: Western Digital® WD Blue™ 1 TB SATA III HDD (6 GB/s, 7200 RPM)
- Chipset: Intel® Z390
- Display Resolution: 1920 × 1080
- OS: Microsoft® Windows® 10

Specifically, to calibrate t_j, we build the baseline platform with brand new parts according to the above hardware configurations and perform a clean installation of the selected operating system. Then we run both modules of CpsMark+ on the baseline platform for 5 independent iterations, the workload-wise calibration t_j is calculated as the median value over the tested metrics of the jth workload from the five runs. Note that since the baseline platform is not a finished product of a computer manufacturer, it is illogical to integrate the tested metrics of all workloads within each module as the module-wise calibration of the baseline platform.

4.7. Benchmark characterization

In this section, we analyze some basic characteristics of CpsMark+ from the perspectives of sensitivity and repeatability, which are two widely used criteria of typical computer benchmarks. Specifically, we have performed extensive test experiments with CpsMark+ on multiple assembled computer systems. Then we analyze the sensitivity of tested module performance to varying hardware characteristics. We also explore the repeatability of workload performance under a constant computer system and stable test environment.

4.7.1. Experimental setup

We alter five different hardware characteristics of a predefined datum point to build the tested computer systems, including the number of CPU cores, CPU frequency, graphics card, storage device, and system memory, which are crucial factors in determining user experience. For each hardware characteristic, we select four configurations with significant pairwise performance differences. They are denoted as Config 1 to Config 4 in ascending order of performance. The detailed configurations of each hardware characteristic are listed in Table 3.

For the configurations of the CPU characteristic, instead of using different processor models, we stick to the CPU model of the datum point and enable different CPU frequencies or numbers of CPU cores by changing BIOS settings. For the configurations of the graphics card, we use the same brand of discrete graphics cards to ensure consistency of graphics drivers and available physical memory. For the configurations of system memory, we all adopt the single-channel mode and only change the memory size of the datum point. The configuration of the datum point is listed as follows:

- CPU Model: Intel® Core™ i7-9700K (8 cores, 3.60 GHz, 12 MB L3 cache)
- Graphics: Nvidia® GeForce® GTX 750
- RAM: Kingston® ValueRAM™ 4 GB DDR4 2666 MHz
- Storage: Seagate® Barracuda® 1TB SATA III HDD (6 GB/s, 5400 RPM)
- Chipset: Intel® Z390
- Display Resolution: 1920 × 1080
- OS: Microsoft® Windows® 10

Notably, for all the experiments in this section, we disable common auxiliary optimization technologies, e.g., Turbo Boost, Hyper-Threading, and Hardware Acceleration, to better highlight the influence of different configurations under various hardware characteristics on benchmark performance from a static perspective. These auxiliary optimization technologies can be enabled in the practical use of CpsMark+.

Table 3
The hardware characteristics and related configurations.

Hardware characteristic	Configuration 1	Configuration 2	Configuration 3	Configuration 4
CPU cores	2-Core	4-Core	6-Core	8-Core
CPU frequency	2.0 GHz	2.5 GHz	3.0 GHz	3.5 GHz
Graphics card	Nvidia GeForce GTX 750	Nvidia GeForce GTX 980	Nvidia GeForce GTX 1080	Nvidia GeForce RTX 2080Ti
Storage device	Seagate Barracuda 1TB SATA III 5400RPM HDD	Western Digital Blue 1TB SATA III 7200 RPM HDD	Samsung 860 EVO 250GB SATA III SSD	Samsung 970 PRO 512GB NVMe M.2 SSD
System memory	4 GB	8 GB	16 GB	32 GB

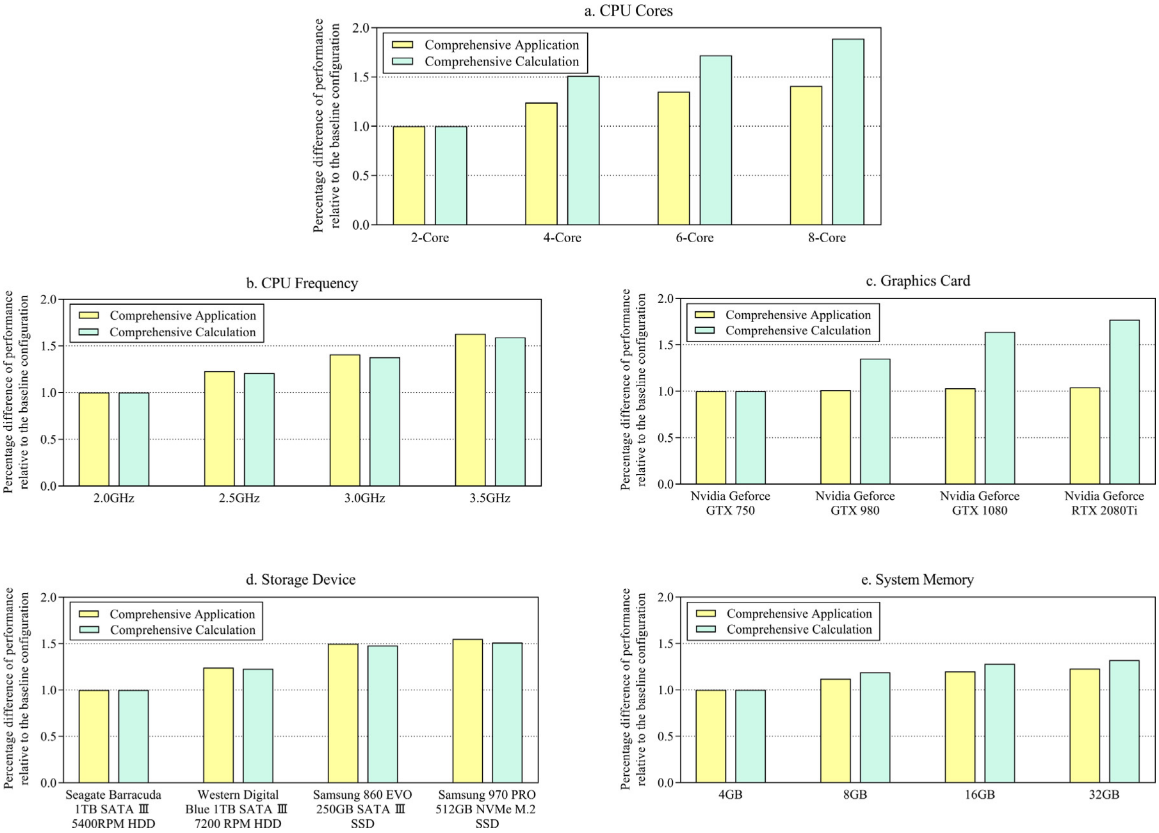


Fig. 4. The sensitivity of the module performance to various hardware characteristics.

4.7.2. *Sensitivity analysis*

Through evaluating the module performance and the workload performance on tested systems with different levels of configurations, we can explore the sensitivity of CpsMark+ scores to various hardware characteristics. To get strict test results, except for the hardware characteristic under test, the other components of a certain configuration remain identical to the components of the datum point. Specifically, we run CpsMark+ on each configuration for 20 independent iterations with a system reboot and a 15-min interval between each run. In each iteration, we sum the tested metrics of the included workloads for each module, then the average of the sums is adopted as the module performance on a certain configuration. Finally, for each hardware characteristic, we calculate the inverse ratio of the module performance tested on the other three configurations to the module performance tested on the first configuration, i.e., base configuration, respectively. The sensitivity of the module performance and the workload performance of CpsMark+ to various hardware characteristics are shown in Fig. 4 and Table 4, respectively.

Based on the module performance evaluation depicted in Fig. 4, we notice that both modules have a high sensitivity to CPU cores and CPU frequency, the module performance steadily increases as the configurations improve, indicating that both modules can make full use of CPU resources and be significantly affected by more CPU cores and higher CPU frequency. The CC module has a significantly higher sensitivity to the graphics card, the best configuration performs 1.77 times better than the base configuration, while there is no significant difference in the performance of the CA module, which indicates that better graphic cards cannot lead to significant performance improvement of the CA module. Rotation speed and storage media of hard disks also have a great influence on the performance of both modules, since the workloads involve application launching and many I/O operations, while drive interface and protocol contribute less to the module performance. Both modules are relatively less sensitive to system memory than the other hardware characteristics, which indicates that larger size of system memory will bring least significant improvement of both module performance compared to better configurations under other hardware characteristics.

We also notice that the sensitivity of the CC module to most hardware characteristics is higher than the sensitivity of the CA module, since the workloads in the CC module are heavier and have more resource consumption. In addition, as the configurations improve, the growth rate of the module performance slows down, especially for the best configurations, because when configurations exceed some requirement bottleneck of the entire workloads, extra improvement of a single hardware characteristic cannot yield much performance growth.

Table 4
The sensitivity of the workload performance to various hardware characteristics.

CPU cores	Chrome	PowerPoint	Word	Excel	Acrobat	WinRAR	Outlook	Photoshop	AutoCAD	3ds Max	Premiere	After effects	HandBrake
Config 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Config 2	1.18	1.21	1.19	1.27	1.23	1.25	1.19	1.48	1.51	1.55	1.49	1.52	1.56
Config 3	1.35	1.37	1.35	1.41	1.36	1.39	1.34	1.73	1.72	1.74	1.69	1.75	1.71
Config 4	1.41	1.43	1.42	1.47	1.44	1.45	1.42	1.89	1.87	1.92	1.91	1.93	1.93
CPU frequency	Chrome	PowerPoint	Word	Excel	Acrobat	WinRAR	Outlook	Photoshop	AutoCAD	3ds Max	Premiere	After effects	HandBrake
Config 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Config 2	1.20	1.22	1.21	1.25	1.23	1.26	1.19	1.18	1.19	1.22	1.24	1.23	1.22
Config 3	1.37	1.38	1.39	1.44	1.40	1.43	1.38	1.33	1.35	1.37	1.36	1.34	1.34
Config 4	1.63	1.65	1.64	1.68	1.64	1.67	1.62	1.59	1.57	1.63	1.61	1.58	1.61
Graphics card	Chrome	PowerPoint	Word	Excel	Acrobat	WinRAR	Outlook	Photoshop	AutoCAD	3ds Max	Premiere	After effects	HandBrake
Config 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Config 2	1.01	0.99	1.00	1.02	1.01	1.01	0.98	1.36	1.34	1.37	1.35	1.34	1.36
Config 3	1.03	1.00	1.01	1.01	0.99	1.02	1.02	1.66	1.65	1.68	1.66	1.63	1.65
Config 4	1.05	1.04	1.04	1.03	1.02	1.03	1.01	1.77	1.79	1.77	1.75	1.78	1.76
Storage device	Chrome	PowerPoint	Word	Excel	Acrobat	WinRAR	Outlook	Photoshop	AutoCAD	3ds Max	Premiere	After effects	HandBrake
Config 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Config 2	1.26	1.25	1.27	1.23	1.24	1.27	1.25	1.25	1.22	1.24	1.25	1.23	1.21
Config 3	1.48	1.47	1.49	1.51	1.50	1.51	1.47	1.46	1.45	1.48	1.49	1.47	1.48
Config 4	1.54	1.55	1.53	1.53	1.55	1.54	1.56	1.51	1.52	1.49	1.50	1.48	1.47
System memory	Chrome	PowerPoint	Word	Excel	Acrobat	WinRAR	Outlook	Photoshop	AutoCAD	3ds Max	Premiere	After effects	HandBrake
Config 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Config 2	1.17	1.16	1.17	1.15	1.18	1.14	1.15	1.22	1.19	1.21	1.18	1.22	1.20
Config 3	1.23	1.22	1.25	1.24	1.23	1.22	1.22	1.28	1.27	1.29	1.26	1.25	1.30
Config 4	1.24	1.21	1.26	1.23	1.26	1.23	1.24	1.31	1.29	1.33	1.32	1.28	1.33

As for the sensitivity of the workload performance of CpsMark+ to each hardware characteristic, based on the workload performance evaluation depicted in Table 4, we have observed the same trend as the sensitivity of the module performance. Generally, the performance of all the workloads is highly sensitive to the number of CPU cores, CPU frequency, and the storage devices. The performance of the workloads that require massive GPU-intensive computing, e.g., AutoCAD and Premiere, is more sensitive to graphics cards, compared to the relatively lightweight workloads, e.g., Microsoft Office. However, the performance of some workloads in the CA module, e.g., Excel and WinRAR, is more sensitive to CPU frequency and storage devices, which might be resulted from frequent float point calculations in the RAM and massive document I/O operations in disks triggered by these workloads. We also find out that the performance improvement of most workloads is not significant once the size of system memory reaches 8 GB, which is likely to be the requirement threshold for the workload software to run smoothly.

4.7.3. *Repeatability analysis*

The repeatability of CpsMark+ is evaluated according to the fluctuation of the module performance and the workload performance tested on the identical computer systems. We leverage Coefficient of Variation (CV), the ratio of the standard deviation to the mean, to indicate the degree of performance fluctuation [31]. To be more specific, for all the experiments under each hardware characteristic, we calculate the CV of the module performance and the workload performance evaluated on the same configuration over 20 independent iterations, respectively. Finally, we aggregate the CV of the module performance under each hardware characteristic and calculate the average CV of the workload performance evaluated on each level of configurations. The results are shown in Fig. 5.

As we can see from the results depicted in Fig. 5, the CV of the module performance under all the hardware characteristics is less than 3%, while the CV of the workload performance under each level of the four configurations is less than 2.5%, which indicates that the overall benchmark results of CpsMark+ are stable and have high consistency under the identical tested computer systems and environment.

Furthermore, for each hardware characteristic, we mark the CV of the module performance under Config 1 and Config 4, respectively. It turns out that except for the results of the CC module under the hardware characteristic of CPU frequency, the best configuration will cause the highest CV of the module performance, while the worst configuration will lead to the lowest CV. Combining the CV of the workload performance with each other, we can conclude that the stability of benchmark results will be improved if the performance of tested configurations exceeds the performance requirements of workloads. As a result, the CV of the module performance under the hardware characteristics of the CPU cores and CPU frequency is relatively high, since CPU frequency of only 2.0 GHz or 2 CPU cores might significantly encumber the performance of tested computer systems.

We also notice that the CV of the heavyweight workload performance is generally higher than the CV of the lightweight workload performance, which is consistent with the previous conclusion, the possible reason is that heavyweight workloads will greatly occupy system resources and lead to unexpected disturbance caused by resource competition between complex program instructions. While Google Chrome is an exception, since the value of its tested metric is relatively small so that it is susceptible to the fluctuation of repetitive experiments. Another finding is that the module performance and the workload performance tested on better configurations will become less volatile, as the performance of high-level configurations might greatly exceed the requirements of workload software. Overall, our benchmark methodology ensures that CpsMark+ is of high sensitivity to provide stable and reliable evaluation results.

4.8. *Comparative evaluation against competing benchmarks*

In this section, we mainly focus on quantitative and qualitative comparison between CpsMark+ and two commonly used computer benchmarks in commercial field, i.e., SYSmark 2018 and PCMark 10. We explain the experimental and the analytical results in detail, which further highlight the strength and the design philosophy of CpsMark+.

4.8.1. *Quantitative comparison*

For quantitative comparison, we compare CpsMark+ with SYSmark 2018 and PCMark 10 with respect to the sensitivity and the repeatability of the module performance under various hardware characteristics. We do not select other metrics, e.g., test duration and power consumption, since SYSmark 2018 and PCMark 10 are not open-source

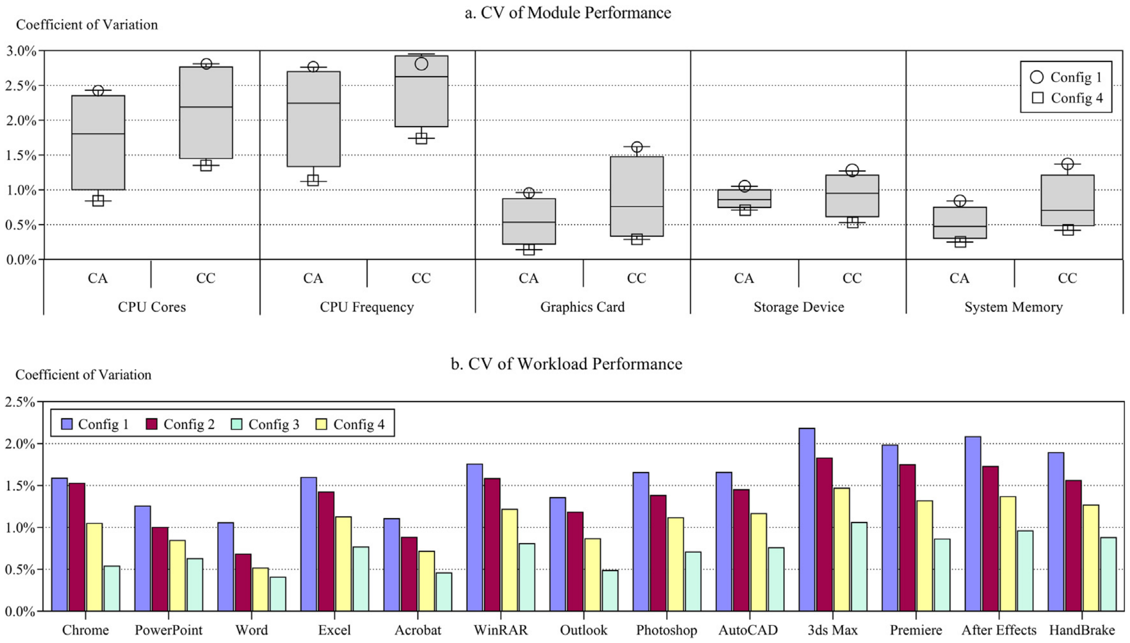


Fig. 5. The repeatability of the module/workload performance under various hardware characteristics.

benchmarks and do not have built-in functions to precisely measure these metrics, which as well makes it impossible to compare the sensitivity and the repeatability of them at a finer granularity, e.g., the level of workload performance. In addition, sensitivity and repeatability are the universal metrics for comparing different benchmarks, even if they possess diverse construction methodologies and usages.

Specifically, we follow the same experimental setup as described in Section 4.7. The modules of SYSmark 2018 include Productivity, Creativity, and Responsiveness, while the modules of PCMark 10 include Essentials, Creativity, and Digital Content Creation. The detailed information about SYSmark 2018 and PCMark 10 is available on their official websites, respectively. Note that in this section, among the three benchmarks, we only compare the sensitivity and the repeatability of the modules that evaluate system performance in similar usage scenarios. Average sensitivity and repeatability (CV in percentage) of the module performance for the three compared benchmarks are summarized in Table 5.

In terms of the sensitivity results depicted in Table 5, among the three modules that evaluate system performance related to document editing and Internet surfing, i.e., the CA module of CpsMark+, the Productivity module of SYSmark 2018, and the Productivity module of PCMark 10, the Productivity module of SYSmark 2018 has the highest sensitivity to all the configurations under each hardware characteristic, since it includes some workloads that have relatively high consumption of system resources, e.g., AutoIT and Shotcut, while the CA module of CpsMark+ has the second highest sensitivity, which is close to the sensitivity of the Productivity module of SYSmark 2018. Among the three modules that evaluate system performance related to multimedia processing and graphics design, i.e., the CC module of CpsMark+, the Creativity module of SYSmark 2018, and the Digital Content Creation module of PCMark 10, the CC module of CpsMark+ is most sensitive to all the hardware characteristics, especially graphics cards, which indicates CpsMark+ can sensitively reflect performance improvement of better GPUs in digital and multimedia processing tasks.

In terms of the repeatability results depicted in Table 5, among the three modules that evaluate system performance related to document editing and Internet surfing, i.e., the CA module of CpsMark+, the Productivity module of SYSmark 2018, and the Productivity module of PCMark 10, the CA module has the highest repeatability, i.e., the lowest CV, across all configurations under each hardware characteristic, while the Productivity module of SYSmark 2018 has the lowest repeatability, i.e., the highest CV. This result is attributed to the UI-level automation of SYSmark 2018, which introduces massive unstable and delayed interactions, e.g., clicking dialog windows. Among the three modules that evaluate system performance related to multimedia processing and graphics design, i.e., the CC module of CpsMark+, the Creativity module of SYSmark 2018, and the Digital Content Creation module of PCMark 10, likewise, the CC module has the highest repeatability, i.e., the lowest CV, across all configurations under each hardware characteristic, which is attributed to the relatively lightweight workloads and the smooth API-level automation of CpsMark+.

Generally, in terms of the modules that evaluate system performance in similar usage scenarios, CpsMark+ exhibits the highest repeatability against state-of-the-art commercial benchmarks, i.e., SYSmark 2018 and PCMark 10, while it also possesses the second highest sensitivity to the hardware characteristics tested in this experiment, which is close to the sensitivity of SYSmark 2018.

4.8.2. *Qualitative comparison*

In this section, we empirically conduct some qualitative comparison of the three benchmarks from the perspectives of workload characterization and scoring methodology.

Firstly, the Responsiveness module of SYSmark 2018 and the Essentials module of PCMark 10 contain a large amount of irrelevant workload operations that cannot precisely simulate user experience perceived in practical usage scenarios of tested computer systems, which is not consistent with the primary attribute of CpsMark+ and accounts for the reason why we exclude them from the above quantitative comparison.

To be more specific, the Responsiveness module of SYSmark 2018 solely measures the response time of program initialization, its workloads consist of a series of sequential application starts and shutdowns, which however, cannot reflect the practical use case in daily office routines and will over amplify the influence of storage devices on the overall performance evaluation based on user experience. On the contrary, each workload of CpsMark+ reflects a common workflow frequently adopted in modern office scenarios and collectively forms typical tasks that are fluent in nature, which exactly justifies our benchmark principal of simulating user experience. Moreover, the Essentials module of PCMark 10 contains the playback of a video with fixed

Table 5
Average sensitivity and repeatability of the module performance for the compared benchmarks.

	CpsMark+ (sensitivity/repeatability)		SYSmark 2018 (sensitivity/repeatability)			PCMark 10 (sensitivity/repeatability)		
CPU cores	CA	CC	Productivity	Creativity	Responsiveness	Essentials	Productivity	Digital content creation
Config 1	1.00/2.43	1.00/2.81	1.00/3.76	1.00/4.97	1.00/4.28	1.00/3.15	1.00/2.78	1.00/4.15
Config 2	1.24/2.15	1.51/2.66	1.28/3.52	1.43/4.35	1.35/3.83	1.19/2.86	1.20/2.62	1.44/3.72
Config 3	1.35/1.46	1.72/1.72	1.41/3.04	1.68/4.06	1.38/3.51	1.30/2.34	1.34/2.17	1.68/3.08
Config 4	1.41/0.84	1.89/1.35	1.47/2.17	1.81/3.87	1.40/3.04	1.33/1.77	1.37/1.56	1.81/2.54
CPU frequency	CA	CC	Productivity	Creativity	Responsiveness	Essentials	Productivity	Digital content creation
Config 1	1.00/2.54	1.00/2.87	1.00/3.88	1.00/5.12	1.00/4.35	1.00/3.25	1.00/2.91	1.00/3.97
Config 2	1.23/2.76	1.21/2.95	1.26/4.02	1.16/5.11	1.13/4.21	1.15/3.11	1.15/2.63	1.13/4.16
Config 3	1.41/1.95	1.38/2.38	1.48/3.34	1.32/4.53	1.19/3.96	1.32/2.48	1.33/2.24	1.32/3.52
Config 4	1.63/1.12	1.59/1.74	1.71/2.73	1.57/4.17	1.22/3.48	1.47/1.93	1.49/1.75	1.49/3.23
Graphics card	CA	CC	Productivity	Creativity	Responsiveness	Essentials	Productivity	Digital content creation
Config 1	1.00/0.96	1.00/1.62	1.00/2.35	1.00/4.16	1.00/3.26	1.00/1.79	1.00/1.48	1.00/3.35
Config 2	1.01/0.64	1.35/1.07	1.04/2.14	1.35/3.25	1.12/2.74	1.02/1.28	1.01/0.81	1.32/2.41
Config 3	1.03/0.43	1.64/0.45	1.05/1.85	1.60/2.64	1.14/2.21	1.03/0.82	1.02/0.59	1.58/1.77
Config 4	1.04/0.14	1.77/0.29	1.05/1.56	1.75/1.83	1.15/1.77	1.03/0.56	1.04/0.37	1.71/1.46
Storage device	CA	CC	Productivity	Creativity	Responsiveness	Essentials	Productivity	Digital content creation
Config 1	1.00/1.05	1.00/1.27	1.00/2.47	1.00/3.47	1.00/2.95	1.00/1.87	1.00/1.52	1.00/2.58
Config 2	1.24/0.88	1.23/0.84	1.29/2.15	1.18/3.12	1.47/2.72	1.24/1.39	1.21/1.07	1.18/2.21
Config 3	1.50/0.84	1.48/1.06	1.53/2.02	1.39/2.85	1.83/2.49	1.39/1.28	1.36/0.89	1.37/1.84
Config 4	1.55/0.71	1.51/0.53	1.65/1.97	1.47/2.34	2.25/2.11	1.47/1.35	1.42/0.81	1.43/1.57
System memory	CA	CC	Productivity	Creativity	Responsiveness	Essentials	Productivity	Digital content creation
Config 1	1.00/0.84	1.00/1.37	1.00/2.20	1.00/3.84	1.00/3.09	1.00/1.85	1.00/1.67	1.00/2.94
Config 2	1.12/0.51	1.19/0.65	1.17/1.75	1.19/2.98	1.23/2.45	1.08/1.26	1.11/1.25	1.14/2.06
Config 3	1.20/0.44	1.28/0.76	1.26/1.58	1.29/3.25	1.25/2.06	1.15/0.99	1.17/0.95	1.25/1.71
Config 4	1.23/0.25	1.32/0.42	1.33/1.33	1.34/2.77	1.26/2.23	1.19/0.63	1.22/0.71	1.30/1.45

duration, thus massive time consumption is included in the calculation of test metrics, which nevertheless, will dilute the contribution of better hardware characteristics to the performance improvement of this module and further reduce the benchmark sensitivity.

Secondly, for each module of PCMark 10, the scoring methodology takes the geometric mean over the test metrics of inclusive workloads, which returns a normalized score that treats the performance of each workload equally and neglects different importance of various workload operations in daily office scenarios. By contrast, as described in Section 4.5, for each module of CpsMark+, the scoring methodology takes the weighted sum over the test metrics of inclusive workloads, which emphasizes the influence of heavy or durable workload performance on simulated user experience and ignores the importance of trivial workload operations that are less involved in the routines of end users.

5. Case study performance evaluation of office desktops using CpsMark+ in a vendor-neutral tendering

In this section, we aim to demonstrate the effectiveness of CpsMark+ in simulating user experience under office-oriented working scenarios for better office desktop performance evaluation in practical centralized procurement. Specifically, in a vendor-neutral tendering of desktop computers for a Chinese company, the tendering was divided into two separate batches with different bid evaluation methods. For the second batch, we combined the original bid evaluation method prepared for the first batch with benchmark scores from CpsMark+ to formulate a new bid evaluation method. The original and the new bid evaluation methods were then independently adopted in the above two tendering batches, respectively. After one-year use of the wining desktops selected by the two bid evaluation methods, we independently investigated the user experience of end users from each tendering batch and collected their ratings. The results show that the desktops purchased in the second batch have significantly higher ratings for user experience, which indicate that the workloads of CpsMark+ can precisely simulate user experience perceived by end users working in modern office-oriented scenarios and enable more targeted performance evaluation for desktops with the above usages.

5.1. *Brief introduction of tendering*

At the beginning of 2020, a large digital marketing agency in China initialized a centralized procurement to purchase desktop computers for the employees from a functional department and a business department, which are denoted as A and B, respectively. For innovating the traditional tendering policy and validating the effectiveness of CpsMark+, within each department, the procurement was arranged as two separate batches of vendor-neutral tendering with different bid evaluation methods, which are denoted as 1 and 2. The basic information of the four tendering batches are listed in Table 6. Then during the next year, the employees of each department were divided into two groups to use the desktop computers purchased in the two tendering batches, respectively.

Note that in addition to the bid evaluation methods for final decision-making among shortlisted alternatives, we also clarified the minimum technical requirements to preliminarily screen candidates from all bidders, which were based on the standard and high-performance configurations in Bitkom's guideline for IT procurement [14], i.e., Vendor-neutral Tendering of Desktop Computers.

5.2. *Improvement of bid evaluation methods*

Main difference between the two tendering batches lay in the bid evaluation methods, which were adopted by bid evaluation committee to determine the best bidding product. In this case study, to help authorities purchase desktop computers with better end-user experience at a certain cost and further validate the effectiveness of CpsMark+, we decided to partly replace the straightforward hardware-based scoring rules in the original bid evaluation method with benchmark scores from CpsMark+ to develop the new bid evaluation method.

5.2.1. *The original bid evaluation method*

The old bid evaluation method consists of 3 sections with a total of 100 points, i.e., the commercial section, the technical section, and the price section. The final score for a certain bid is the sum over the score for each section. Specifically, the score for the commercial section is the direct sum over the score for each included item (0–1 point per

Table 6
Basic information of the four tendering batches.

Tendering batches	Purchase quantity	End users	Primary responsibilities
1A	39	Department A (Functional)	Supportive market research & analysis
2A	39		
1B	46	Department B (Business)	Marketing related service of FMCG
2B	46		

Table 7
The weight of each item within the technical section.

	CPU	Motherboard	Monitor	Memory	Storage	Graphics
1A	15	9	4	11	15	13
1B	20	8	7	11	13	18

item). The score for the technical section is the weighted sum over the score for each included item, which is the weighted average over ratings for various metrics ranked by the importance (0–1 point per metric). Detailed information of the items within each section are listed as follows:

1. Commercial section (3/3 points for 1A/1B)
   - Quality of bid response documents.
   - Efficiency of logistics and query systems.
   - Quality of after-sales service.
2. Technical section (67/77 points for 1A/1B)
   - CPU. Metrics: craftsmanship, number of cores, base frequency, size of L3 cache, Thermal Design Power.
   - Motherboard. Metrics: chipset, expansion slots, structure, BIOS, power supply.
   - Monitor. Metrics: screen size, resolution, brightness, panel type, ports.
   - Memory. Metrics: DDR generations, capacity, operating frequency, CAS latency.
   - Storage. Metrics: HDD/SSD, capacity, rotation speed (for HDD), interface, disk buffer.
   - Graphics. Metrics: integrated/discrete, craftsmanship, architecture, GPU frequency (for discrete graphics).

   The score for the jth item is calculated as follows:

   $$Score_j = w_j \cdot \frac{\sum_{i=1}^{n_j}[r_j(i) \cdot \left(1 - \frac{i-1}{n_j}\right)]}{\sum_{i=1}^{n_j}(1 - \frac{i-1}{n_j})}$$

   where n_j is the number of metrics for the jth item, $r_j(i)$ is the rating (0–1 point) for the ith metric of the jth item, w_j is the weight of the jth item predefined by domain experts, which is listed in Table 7.
3. Price section (30/20 points for 1A/1B)
   The lowest quotation among all the bids that meet the minimum technical requirements is defined as the Negotiated Base Price (NBP), then the price section score for a certain bid is the product of the price coefficient and the ratio of the NBP to its quotation. The price coefficients for the tendering batches of 1A and 1B are 30 and 20, respectively.

5.2.2. *The new bid evaluation method*

In terms of the new bid evaluation method for the tendering batches of 2A and 2B, we introduce benchmark scores from CpsMark+ to replace any items related to system performance in the original technical section, i.e., all the items except for the Monitor item. The weight of the Monitor item and the weights of the commercial and the price sections remain constant. To maintain a total score of 100 points, the benchmark score weights for the tendering batches of 2A and 2B are 63 and 70, respectively.

To calculate the absolute benchmark score from CpsMark+, unlike the item weights within each section in the original bid evaluation method, the weight of the CA/CC module is not predefined by domain experts from the bid evaluation committee, instead it is assigned as the average value of the survey results from the real end users of both departments. The weights of the CA/CC module for the tendering batches of 2A and 2B turn out to be 0.71/0.29 and 0.12/0.88, respectively. Then the absolute benchmark score from CpsMark+ for each tendering batch is defined as follows:

$$Score_{cps} = \sqrt[\sum_{i=1}^{2} w_i]{\prod_{i=1}^{2} {s_i}^{w_i}}$$

where w_i is the weight of the ith module, s_i is the median score of the ith module over 5 independent tests on a certain bidding product.

To scale the absolute benchmark score from CpsMark+ for better reflection of relative performance among various bidding products, we adopted a similar strategy as in the price section. Specifically, the best absolute benchmark score among all the bids that meet the minimum technical requirements is defined as the Negotiated Maximum Performance (NMP), then the final benchmark score for a certain bid is the product of the benchmark score weight and the ratio of its absolute benchmark score to the NMP. Finally, the score for the new technical section is the direct sum over the final benchmark score and the score for the Monitor item.

5.3. *Effects of introducing benchmark scores from CpsMark+*

To evaluate the effects of introducing benchmark scores from CpsMark+ as part of the new bid evaluation method, for the winning bids purchased in the tendering batches of 1A/1B and 2A/2B, we performed a comparative analysis towards the one-year user experience rated by the respective end users.

5.3.1. *Evaluation protocols of user experience*

We first formulated the explicit evaluation protocols for rating user experience of office desktops in modern office scenarios. The ISO 9241 standard [32] of human–computer interaction defines usability as “the extent to which a product can be used by specific users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use”. We defined user experience of the winning bids in a similar way as the usability defined in the ISO 9241 standard. Since for all the bids that meet the minimum technical requirements, the effectiveness of the products in fulfilling the tasks specified by the tenders is guaranteed, we mainly focused on the following two metrics:

(1) Efficiency, i.e., the user-perceived time consumption for software and applications to achieve specified goals. The rating is scaled as “very efficient” (5 points), “somewhat efficient” (4 points), “neutral” (3 points), “somewhat inefficient” (2 points), or “very inefficient” (1 point).

(2) Smoothness, i.e., the user-perceived overall smoothness in daily use of software or applications, including jank, launching speed, delay, and response to instructions. The rating is scaled as “very smooth” (5 points), “somewhat smooth” (4 points), “neutral” (3 points), “somewhat unsmooth” (2 points), or “very unsmooth” (1 point).

In terms of the rating items, we surveyed each department to find out the software or applications frequently used by most end users

Table 8
The rating items and corresponding weights.

	MySQL	Excel	Power BI	Photoshop	Premiere	After effects	Internet explorer	Word	PowerPoint	Lark [33]
1A/2A	0.18	0.19	0.13	0	0	0	0.09	0.16	0.12	0.13
1B/2B	0	0	0	0.22	0.18	0.25	0.14	0.10	0.05	0.06

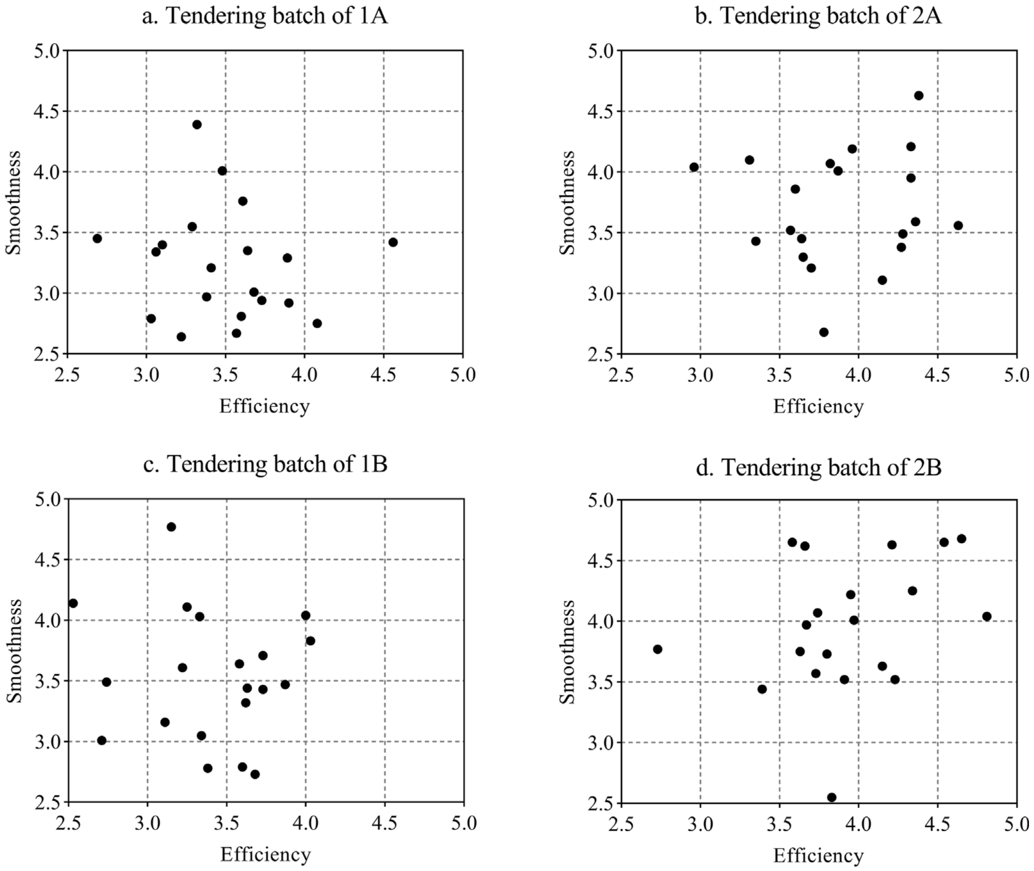


Fig. 6. The distributions of user experience ratings for the winning bids.

within one year after the procurement. Then we gave them different weights according to the average hours of use over the entire department, which are listed in Table 8.

For each tendering batch, i.e., 1A, 2A, 1B, and 2B, we randomly invited 20 end users from the corresponding group of their department to independently rate the user experience of the desktop computers purchased in this tendering batch. The questionnaires adopted for rating the user experience are similar as CSAT [34]. For each desktop computer, the total score for each metric of the user experience is the weighted sum over the metric ratings for all the items.

5.3.2. Evaluation results

The distributions of user experience ratings for the winning bids from the four tendering batches are shown in Fig. 6. As we can see from the results, for both metrics of the user experience, the ratings from all surveyed end users are between 2.5 and 5 points. Specifically, the ratings for both user experience metrics of the winning bids from the tendering batches of 1A/1B are mostly between 2.5 and 4 points, while the ratings from the tendering batches of 2A/2B are mostly between 3 and 4.5 points, which indicates that the user experience of the desktop computers selected by the new bid evaluation method is improved to some extent.

Table 9 shows some descriptive statistics of the above user experience ratings and the average quotation for the desktops purchased from each tendering batches. For the tendering batches of 1A/2A, the efficiency and the smoothness ratings for the winning bids are 3.51/3.90 points and 3.23/3.69 points, with an increase of 11.11% and 14.24%, respectively. For the tendering batches of 1B/2B, the efficiency and the smoothness ratings for the winning bids are 3.40/3.93 points and 3.53/3.96 points, with an increase of 15.59% and 12.18%, respectively.

Although the rating results of user experience demonstrate the effectiveness of CpsMark+ in identifying office desktops with better user experience under modern office-oriented scenarios, the analysis so far has only told part of the story for evaluating the effects of introducing benchmark scores from CpsMark+ in centralized procurement, since pricier bids generally tend to deliver better system performance, which will cause a much higher budget. To this end, we also consider the average quotation for the winning bids from each tendering batch, which is 5316/5562 CNY and 6465/6948 CNY for the tendering batches of 1A/2A and 1B/2B, with an increase of 4.63% and 7.47%, respectively. Note that in this paper, charges for other services, e.g., logistics and insurance, are excluded from the average quotation. Apparently, the higher average quotation of the winning bids leads to more significant increase of user experience ratings. This result demonstrates that the new bid evaluation method based on benchmark scores from CpsMark+ can help authorities select the bid with better user experience and higher cost-effectiveness in the centralized procurement of office desktops.

5.3.3. Statistical analysis

In this case study, we randomly selected 20 end users from each tendering batch for higher survey efficiency and minimizing the rating

Table 9
Descriptive statistics of the user experience ratings and the average quotation for the winning bids.

		1A	2A	1B	2B
Efficiency	Mean (%)	3.51 (70%)	3.90 (78%)	3.40 (68%)	3.93 (79%)
	95% confidence interval	[3.32–3.71]	[3.69–4.10]	[3.18–3.61]	[3.70–4.15]
Smoothness	Mean (%)	3.23 (65%)	3.69 (74%)	3.53 (71%)	3.96 (79%)
	95% confidence interval	[3.02–3.45]	[3.47–3.91]	[3.28–3.78]	[3.71–4.22]
Average quotation per computer, CNY		5316	5562	6465	6948

Table 10
Results of the *p*-value in significance tests (at a 5% significance level).

		1A	2A	1B	2B
Efficiency	Normality	0.8978	0.5410	0.1805	0.5569
	Homogeneity of variance	0.8486		0.8741	
	Student’s t-test	0.0070		0.0001	
Smoothness	Normality	0.1643	0.8280	0.6373	0.0643
	Homogeneity of variance	0.9769		0.8455	
	Student’s t-test	0.0035		0.0165	

deviation due to the subjective evaluation of user experience. Hence, we perform further statistical analysis to explore potential significant changes of user experience ratings within the whole populations from the tendering batches of 2A/2B. The results of significance tests are shown in Table 10.

According to the Shapiro–Wilk test, the normality for all the distributions of user experience ratings is accepted, which indicates that user experience of the winning bids from each tendering batch is concentrated within a certain range. Then we conduct a two-tailed F test to infer the homogeneity of variance between user experience ratings from 1A and 2A, as well as 1B and 2B. Specifically, all the results accept the null hypothesis, which is possibly attributed to the similar responsibilities of employees from the same department.

The results of the student’s t-test also infer a significant change of user experience ratings within the whole populations from the tendering batch of 2B. Specifically, the p-value of the student’s t-test for efficiency ratings from the tendering batches of 1B/2B is just 0.0001, which suggests that a significant change of user experience perceived by all the employees from the tendering batch of 2B exists with a large probability. The possible reason is that system performance of the winning bids from the tendering batch of 2B breaks through requirement bottleneck of the routine tasks in department B.

5.3.4. *User experience of items excluded from CpsMark+*

Although we have seen significant improvements in user experience of the winning bids selected by the new bid evaluation method, the rating items for user experience evaluation partly overlap with the workloads of CpsMark+. Without loss of generality, we conduct a comparative analysis to further validate the effectiveness of CpsMark+ in simulating user experience of tested computer systems with respect to software or applications that are not included in its workloads.

Specifically, for each rating item that is not adopted as the workload application of CpsMark+, we collect and average its user experience metrics over the winning bids selected by the original and the new bid evaluation methods, respectively. The results are shown in Fig. 7 and Table 11.

According to the above results, under the workloads that are not included in CpsMark+, user experience of the office desktops selected by the new bid evaluation method also improves by varying degrees. For example, in terms of heavy workloads, the average ratings for efficiency and smoothness of MySQL increase by 22.95% and 26.56%, respectively. The similar trend of user experience improvement is also observed with respect to more lightweight workloads, e.g., Power BI, Internet Explorer, and Lark. These results suggest that the workloads of CpsMark+ are sufficiently representative for simulating user experience of tested computer systems perceived under a wide range of workloads.

6. Conclusions

This paper presents CpsMark+, a scenario-oriented benchmark system that quantitatively evaluates the overall performance of office desktops in centralized procurement. Considering the proposed challenges in benchmarking desktops under practical usage scenarios for centralized procurement, the workloads of CpsMark+ are designed to be scenario-oriented and can simulate user experience of tested computer systems perceived by end users working in modern-office scenarios. The metrics testing and the scoring methodology are flexibly adjusted based on each individual workload. Extensive experiments on multiple real-world tested computer systems demonstrate high sensitivity and repeatability of benchmark scores from CpsMark+, compared to SYSmark 2018 and PCMark 10. From the perspective of end users,

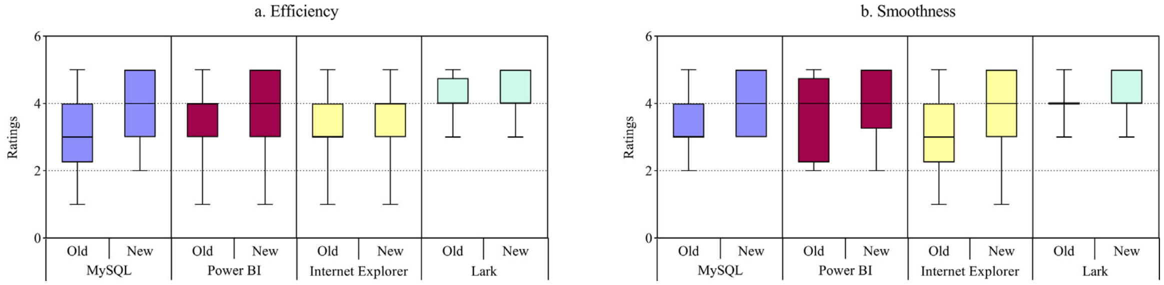


Fig. 7. User experience ratings for software or applications absent in CpsMark+.

Table 11
Average user experience ratings for software or applications absent in CpsMark+.

		MySQL	Power BI	Internet Explorer	Lark
Efficiency	Old bid evaluation method	3.05	3.50	3.43	4.20
	New bid evaluation method	3.75	3.80	3.60	4.28
Smoothness	Old bid evaluation method	3.20	3.55	3.20	3.98
	New bid evaluation method	4.05	3.95	3.55	4.30

in a practical centralized procurement of office desktops, by replacing the original bid evaluation method with benchmark scores from CpsMark+ and comparing user experience ratings for the winning bids selected by the two bid evaluation methods, we also demonstrate the effectiveness of using CpsMark+ to simulate user experience of tested systems in modern-office scenarios for better evaluation of office desktop performance in centralized procurement.

Our work provides a general idea to design computer benchmarks used in other usage scenarios and helps further explore the benefits of introducing benchmark scores in traditional bid evaluation methods for centralized procurement of office desktops. In the future, we will focus on designing parallel workloads that contain more complex interactions and involving other metrics, e.g., battery life or energy efficiency.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key R&D Program of China under grant number 2018YFF0212106. We are thankful to the purchasing manager in the centralized procurement and all the employees that participated in the user experience survey.

References

[1] U. Norf, The role of benchmarks in the public procurement of computers, 2019, https://www.intel.co.uk/content/dam/www/public/us/en/documents/white-papers/role-of-benchmarks-white-paper.pdf.

[2] S.M. Pieper, J.M. Paul, M.J. Schulte, A new era of performance evaluation, Computer 40 (9) (2007) 23–30.

[3] Release of CpsMark 1.0, 2014, https://read01.com/4aAj54.html#.YWL7O9pBxPa.

[4] SYSmark 2018, 2018, https://bapco.com/products/sysmark-2018/.

[5] PCMark 10, 2017, https://benchmarks.ul.com/pcmark10.

[6] Phoronix test system, 2022, https://www.phoronix-test-system.com/.

[7] A. Martin, V. Marangozova-Martin, Automatic benchmark profiling through advanced trace analysis, in: European Conference on Parallel Processing, Springer, Cham, 2016, pp. 63–74.

[8] 3Dmark, 2022, https://benchmarks.ul.com/3dmark.

[9] J. Bucek, K.D. Lange, J.v. Kistowski, SPEC CPU2017: Next-generation compute benchmark, in: Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, 2018, pp. 41–42.

[10] S.C. Woo, M. Ohara, E. Torrie, J.P. Singh, A. Gupta, The SPLASH-2 programs: Characterization and methodological considerations, ACM SIGARCH Comput. Archit. News 23 (2) (1995) 24–36.

[11] J.D. McCalpin, Stream benchmark, 1995, Link: www.cs.virginia.edu/stream/ref.html#what, 22(7).

[12] L.W. McVoy, C. Staelin, Lmbench: Portable tools for performance analysis, in: USENIX Annual Technical Conference, 1996, pp. 279–294.

[13] G. Lu, X. Lin, R. Zhou, Mbench: Benchmarking a multicore operating system using mixed workloads, in: BPOE, Springer, Cham, 2015, pp. 50–63.

[14] F. Felicia, Vendor-neutral tendering of desktop computers, 2019, https://www.itk-beschaffung.de/sites/beschaffung/files/2021-03/200128_lf_vendor-neutral-tendering-of-desktop-computers_en.pdf.

[15] A. Tarvo, S.P. Reiss, Using computer simulation to predict the performance of multithreaded programs, in: Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering, 2012, pp. 217–228.

[16] S. Mittal, J.S. Vetter, A survey of CPU–GPU heterogeneous computing techniques, ACM Comput. Surv. 47 (4) (2015) 1–35.

[17] Y. Wang, V. Lee, G.Y. Wei, D. Brooks, Predicting new workload or CPU performance by analyzing public datasets, ACM Trans. Archit. Code Optim. (TACO) 15 (4) (2019) 1–21.

[18] S. Vagstad, Centralized vs. decentralized procurement: Does dispersed information call for decentralized decision-making? Int. J. Ind. Organ. 18 (6) (2000) 949–963.

[19] K. Huppler, The art of building a good benchmark, in: Technology Conference on Performance Evaluation and Benchmarking, Springer, Berlin, Heidelberg, 2009, pp. 18–30.

[20] J. v. Kistowski, J.A. Arnold, K. Huppler, K.D. Lange, J.L. Henning, P. Cao, How to build a benchmark, in: Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering, 2015, pp. 333–336.

[21] U. Gordon, PCWorld, in: AMD Accuses BAPCo and Intel of Cheating with Sysmark Benchmarks, 2016, https://www.pcworld.com/article/419213/amd-accuses-bapco-and-intel-of-cheating-with-sysmark-benchmarks.html.

[22] Y. Chen, F. Raab, R. Katz, From tpc-c to big data benchmarks: A functional workload model, in: Specifying Big Data Benchmarks, Springer, Berlin, Heidelberg, 2012, pp. 28–43.

[23] In-depth research of China office software market, 2021, http://www.chinaiern.com/baogao/scbg/2953763.shtml?bd_vid=6645286086633733527.

[24] A. Crolotte, Issues in benchmark metric selection, in: Technology Conference on Performance Evaluation and Benchmarking, Springer, Berlin, Heidelberg, 2009, pp. 146–152.

[25] T. Nguyen, P. Calyam, R.B. Antequera, Benchmarking in virtual desktops for end-to-end performance traceability, in: 2015 IFIP/IEEE International Symposium on Integrated Network Management, IM, IEEE, 2015, pp. 1268–1273.

[26] S. Taheri, L.A. Beni, A.V. Veidenbaum, A. Nicolau, R. Cammarota, J. Qiu…, M.R. Haghighat, WebRTCbench: a benchmark for performance assessment of webRTC implementations, in: 2015 13th IEEE Symposium on Embedded Systems for Real-Time Multimedia (ESTIMedia), IEEE, 2015, pp. 1–7.

[27] B. Daniel, Q. Luo, M. Mirzaaghaei, D. Dig, D. Marinov, M. Pezzè, Automated GUI refactoring and test script repair, in: Proceedings of the First International Workshop on End-To-End Test Script Engineering, 2011, pp. 38–41.

[28] T.E. Vos, P.M. Kruse, N. Condori-Fernández, S. Bauersfeld, J. Wegener, Testar: Tool support for test automation at the user interface level, Int. J. Inf. Syst. Model. Des. (IJISMD) 6 (3) (2015) 46–83.

[29] R. Rivest, The MD5 message-digest algorithm (No. rfc1321), 1992.

[30] P.J. Fleming, J.J. Wallace, How not to lie with statistics: the correct way to summarize benchmark results, Commun. ACM 29 (3) (1986) 218–221.

[31] A.G. Bedeian, K.W. Mossholder, On the use of the coefficient of variation as a measure of diversity, Organ. Res. Methods 3 (3) (2000) 285–297.

[32] T. Jokela, N. Iivari, J. Matero, M. Karukka, The standard of user-centered design and the standard definition of usability: analyzing ISO 13407 against ISO 9241-11, in: Proceedings of the Latin American Conference on Human–Computer Interaction, 2003, pp. 53–60.

[33] Lark, 2022, https://www.larksuite.com.

[34] V. Mittal, C. Frennea, Customer Satisfaction: A Strategic Review and Guidelines for Managers, in: MSI Fast Forward Series, Marketing Science Institute, Cambridge, MA, 2010.

Yue Zhang born in 1995, research assistant. His main research includes IT benchmarks, AIops. Now he is a Ph.D. candidate in Renmin University of China.

Tong Wu born in 1975, associate research fellow. His main research includes IT benchmarks, computer architecture and EMC. Now he works in National Institute of Metrology, China.