BenchCouncil Transactions on Benchmarks, Standards and Evaluations 1 (2021) 100012

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

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

Call for establishing benchmark science and engineering

Jianfeng Zhan

Institute of Computing Technology, Chinese Academy of Sciences, China

ARTICLE INFO

Keywords:
Benchmark science and engineering
Origin and evolution
Measurement standard
Standardized data set
Standard benchmark hierarchy
Consistent benchmarking
Meta-benchmark

ABSTRACT

Currently, there is no consistent benchmarking across multi-disciplines. Even no previous work tries to relate different categories of benchmarks in multi-disciplines. This article investigates the origin and evolution of the benchmark term. Five categories of benchmarks are summarized, including measurement standards, standardized data sets with defined properties, representative workloads, representative data sets, and best practices, which widely exist in multi-disciplines. I believe there are two pressing challenges in growing this discipline: establishing consistent benchmarking across multi-disciplines and developing meta-benchmark to measure the benchmarks themselves. I propose establishing benchmark science and engineering; one of the primary goals is to set up a standard benchmark hierarchy across multi-disciplines. It is the right time to launch a multi-disciplinary benchmark, standard, and evaluation journal, TBench, to communicate the state-of-the-art and state-of-the-practice of benchmark science and engineering.

1. The origin and evolution of the benchmark term

Benchmarking is common practice in all industries, and indeed in many areas of life [1]. For example, an Olympic sprinter or fund manager or IT product manager may compare themselves against a benchmark or a close competitor to evaluate their performance. Unfortunately, the benchmark term independently evolves in multi-disciplines and has related but different implications. This section investigates the origin and evolution of the benchmark concept.

I find that the modern benchmark concept (close to its current definition) first appeared in measurement science [2] in the form of bench mark (two words separated by a space). For example, in geodesy, a bench mark is a mark whose height, relative to datum, has been determined by leveling—the operation to measure differences in height between established points relative to a datum [3]. Later, this concept is extended into multi-disciplines.

In the computer discipline, one of the earliest benchmarking effort [4] dated back to 1962 in the Auerbach Corporation's Standard EDP Reports. Joslin defined this benchmarking effort as "a routine used to determine the speed performance of a computer system" [4]. The reports included reporting performance data using typical benchmark tasks – many basic functions – but based on the vendor's published data without stipulating that the benchmark must run on the system under test. Around 1965, Joslin [5] stated that the most important question in computer evaluation should be "how long will it take this system to process my workload (my computer application)?". This exploring methodology produced the concepts of workload modeling, application benchmark, synthetic benchmarks, and standard benchmark, which are still used nowadays [4]. These concepts seem abstract, not directly related to the bench mark concept, though having some connections. The primary reason may be that the computer is a new thing at that time.

The followings are simple explanations of these concepts. Workload modeling is selecting a representative sample set of programs from the entire real workloads [4], which is a critical factor ensuring the benchmark quality. An application benchmark is a mix of programs to be run on several different computer configurations to obtain comparative performance in terms of handling the specific applications [5]. Because of the difficulty (cost) of porting real applications across different systems, in 1969, Bucholz [6] argued a greater degree of abstraction – a synthetic benchmark to imitate the real application – is necessary to make comparisons across different systems practical. The rising costs of synthetic benchmarks motivated the standardization of benchmarks. In 1976, a group of government and industry personals was formed to ascertain the possibility of a standard benchmark library [7], which was the first try in this regard.

As a general term, in the 1987 edition of the Oxford Reference Dictionary, the benchmark is defined as a surveyor's mark indicating a point in a line of levels, a standard or point of reference [3]. The editors obviously did not consider the benchmark concept that appeared in the computer discipline, but their benchmark definition is similar to that in geodesy we referred at the beginning of this section; Zairi et al. [3] thought this definition is the beginning of today's use of the word benchmark in the management discipline.

E-mail address: zhanjianfeng@ict.ac.cn.
URL: https://www.benchcouncil.org/zjf.html.

https://doi.org/10.1016/j.tbench.2021.100012

Available online 21 December 2021

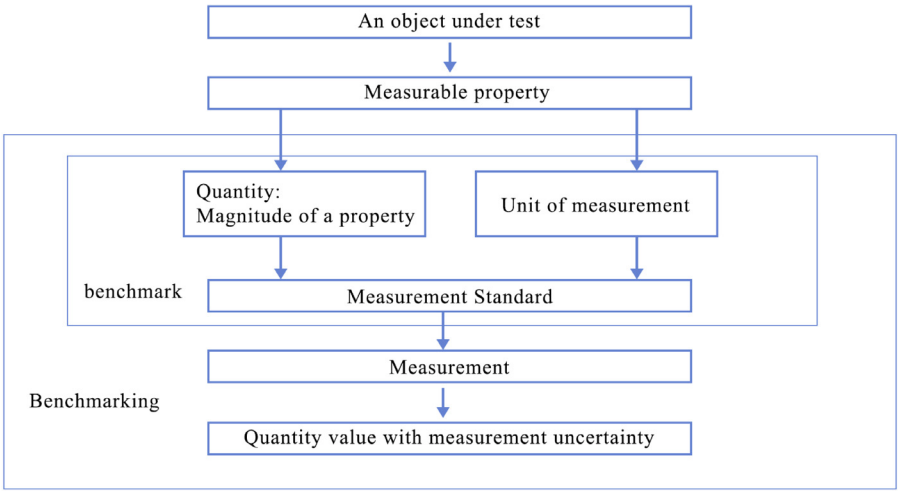

Fig. 1. The interpretation of the first category of the benchmark from the perspective of metrology [8,9].

In the management discipline, the Xerox Corporation was the pioneer of benchmarking [3]: its roots began in 1979, evaluated itself externally through this process which became known as competitive benchmarking. This benchmarking research and practice [3] encompassed an in-depth, ongoing study of best competitors, including detailed reverse engineering of competitor products, technology processes, what they achieved and how they did it, and a tear-down analysis of operating capabilities and features of competing products. This benchmarking practice is very similar to the computer discipline's benchmark-driven performance engineering in terms of the principle. The latter tries to disclose the root causes of the performance bottlenecks of and optimize the computer systems considering the specific workloads.

Gradually, benchmarking was extended as a strategic quality tool to all aspects of the business and progressively integrated into the management process [3]. In this context, Zairi et al. [3] defined it as the continuous process of measuring products, services, and processes against the industry best practices that lead to superior performance.

2. Five categories of benchmarks

This section investigates five categories of benchmarks in multi-disciplines. My intention is not to provide a consistent or unified benchmark definition. Instead, I try to reveal the essence of the benchmarks in five different scenarios. I leave the discussion of consistent benchmarking in the following two sections.

The first category of the benchmark is a measurement standard. In the computer discipline, the Linpack benchmark is of this category, which is widely used to report the performance of a high-performance computer. I provide an interpretation of this category from the perspective of metrology. The Joint Committee for Guides in Metrology (JCGM) [8] defines a measurement standard as a realization of the definition of a quantity, with stated value and associated measurement uncertainty, used as a reference. As shown in Fig. 1, a benchmark realizes the definition of a quantity, the unit of measurement, the measurement methodology, and the reference implementation with stated measurement uncertainty. A quantity is a measurable property of the object under measurement, like length, energy, etc. Benchmarking covers two phases: the design and implementation of the benchmark and measuring the object's properties with the benchmark.

The second one is the representative workloads that run on the systems under measurement. The application benchmarks or synthetic benchmarks in the computer discipline, discussed in Section 1, are of this category. They provide the design input to the system design and implementations. They do not necessarily meet the stringent definition of measurement standards, but they are also used to evaluate systems. For example, in the computer discipline, many deep learning workloads (algorithms) are random with poor repeatability [10,11]. Deep learning is a kind of artificial intelligence (AI) workload. However, they are representative workloads that cannot be overlooked in the system design and implementation.

Generally speaking, the first category of the benchmarks is selected from the second category according to more strict criteria. Fig. 2 explores how to define the representative workloads in the computer discipline. There is increasing freedom from a mathematical problem definition to an algorithm, an intermediate representation, An ISA-specific representation (ISA is short for instruction set architecture), and a micro-architecture representation. Section 3 will further discuss this challenge.

The third is a standardized data set that represents real-world data science problem [12], with defined properties, some of which have ground truth. ImageNet [13] (deep learning benchmark) and MIMIC-III [14] (critical care benchmark) are typical examples. The benchmark of this category is often used to measure against different algorithms. The state-of-the-art algorithm implementation plus the data set usually constitutes the benchmark of the second category.

The fourth is a representative data set, used as a reference. For example, a financial benchmark is an index (statistical measure), calculated from a representative set of underlying data, is used as a reference for financial instruments or contracts [15]. Well-known financial benchmarks include the London Interbank Offered Rate (Libor) and the Euro Interbank Offered Rate [15].

The fifth is the industry best practices in different domains. Benchmarking is the continuous process of searching the industry best practices that lead to superior performance and measuring products, services, and processes against them [3]. The Xerox Corporation pioneered and enhanced this benchmarking process.

3. The challenges

As I elaborate in Section 2, the five categories of benchmarks have a closely connected relationship. However, currently, there is no consistent benchmarking across multi-disciplines. Even no previous work tries to relate those five categories of benchmarks in multi-disciplines. The metrology science paves a foundation for this direction. However, they mainly focus on classical quantities like length, time, and power. Significantly different from those classical quantities, the properties of the objects in the computer, management, or finance disciplines are

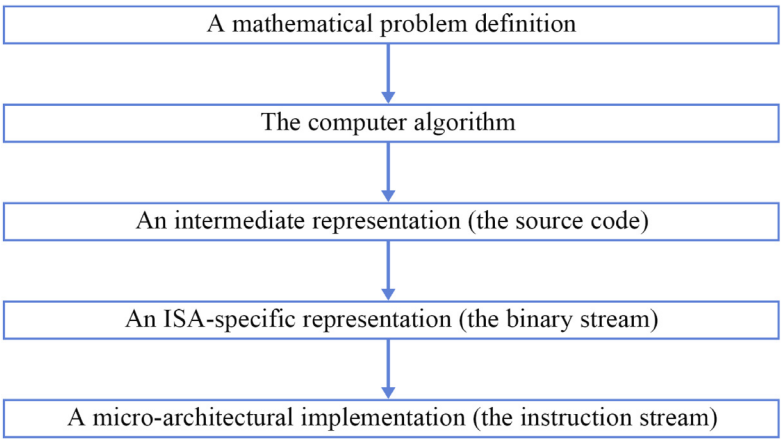

Fig. 2. In the computer discipline, a representative workload, the second category of the benchmarks, is hierarchically defined. From top to down is a mathematical problem definition, an algorithm, an intermediate representation, an ISA-specific representation, a micro-architectural representation. The lower level has more state space. State-of-the practice only analyzes a micro-architectural representation, which is only a subspace or even a point at a high-dimension space [16]. This hierarchy definition can be extended to other disciplines.

greatly affected by its mathematical problem definition and concrete implementation, which raises a serious challenge.

Different observation angles may distort the observable properties. For example, shown in Fig. 2, the quantity value of a computer workload is greatly affected by mathematical problem definitions, concrete algorithms, different ISA and micro-architecture implementations.

I further take the first category of benchmarks as an example to demonstrate the importance of tackling this challenge. Measuring "Quantum Supremacy" against the classical supercomputer is a fundamental issue. Google's "Quantum Supremacy" declaration in 2019 [17] stated that the Sycamore superconductive quantum computer (200 s) is over a billion times faster than the state-of-the-practice Summit system in 2016 [18] (10,000 years) in the task of measuring and simulating one million samples. However, in 2021, a group of scientists and engineers declared, on the Sunway Supercomputer [19], they reduced the classical simulation sampling time of Google Sycamore to 304 s, from the previously claimed 10,000 years through both algorithmic and architecture innovations.

The speed up – the ratio of the quantity values of two different kinds of systems – definitely will change wildly in the future. Understanding the benchmark very well under a hierarchy like that defined in Fig. 2 is a priority before correctly interpreting the implication of the speed up, or else it will mislead the scientific community. The situation may become much complex in the other disciplines, as a clear hierarchy definition is also a luxury. Establishing consistent benchmarking across multi-disciplines is very challenging.

The other challenge is how to measure the benchmarks themselves. Previous work has a preliminary discussion on this issue. For example, in the computer discipline, the characteristics of a (good) benchmark, i.e., representative [4,20], relevance, reproducible, fair, verifiable, repeatable, and economical are discussed in [21,22]. However, most of those properties are subjective. We need a meta-benchmark to evaluate those benchmarks.

I take the representative characteristic as an example; the current theory and practice cannot convince the community that this topic is seriously treated. From the perspective of mathematics, it is necessary to establish a mathematical foundation and consider the meaning of representative in a high dimension space. Unfortunately, in practice, the benchmarking methodology seems ad-hoc. For example, it is reported that there are 6.8 million apps in the leading app stores [23]. How does the community infer the mobile phone market's representative workloads (and benchmarks)?

4. The proposal

I believe that it is necessary to establish benchmark science and engineering; one of the goals is to set up standard benchmark hierarchy across multi-disciplines. There are two reasons. First, there is a natural hierarchy in different categories of benchmarks. As we discussed in Section 2, the first benchmark category is selected from the second category according to more strict criteria. Second, through this hierarchy, we can tackle the challenge of the rising cost of benchmarking. For example, we can put more resources on the primary benchmarks while relating the other benchmarks to the primary benchmarks through traceability.

Fig. 3 is my proposal. The most important is to keep benchmarking consistently, and the following measures will help achieve the target: (1) the unified definition of base quantity and units of measurement; (2) the realization of quantities and units of measurement with different accuracy (and hence cost) levels; (3) the traceability and calibration across the standard benchmark hierarchy. Traceability [8] is a property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty.

At the first tier, the international community needs to define the fundamental benchmarking principle and realize the base quantity, unit of measurement, primary measurement standard, which is the reference of all other benchmarks. The second tier is the first and second categories of the benchmarks. They will reuse the definitions and realizations of base quantity and unit of measurement from the first tier. Meanwhile, the definition and realization of derived quantity and unit of measurement are necessary.

The third tier is the second and fourth categories of the benchmarks. The community often needs to revisit and ponder the mathematical or data problem definitions to provide state-of-the-art and state-of-the-practice implementations. The fourth tier is the fifth category of the benchmarks. As it searches for the best practice, keeping an eye on the advancement of all hierarchies is necessary.

5. TBench: the venue for benchmark science and engineering

I think it is the right time to launch a new journal, BenchCouncil Transactions on Benchmarks, Standards, and Evaluations (in short, TBench). It will provide a venue to communicate and tackle the challenges mentioned above as there is no multidisciplinary and interdisciplinary journal on this area. I only noticed in the management discipline a closely related journal named Benchmarking: An International Journal.

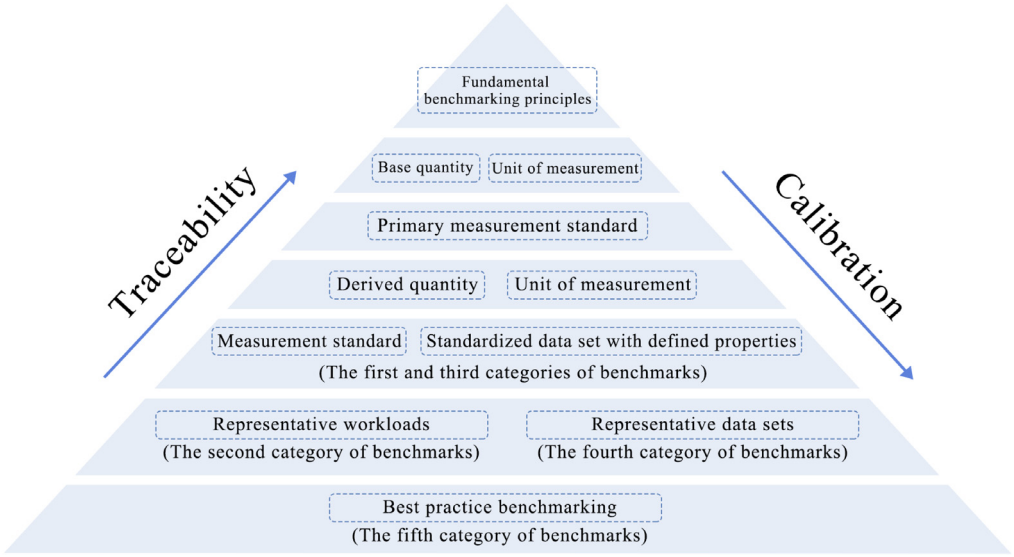

Fig. 3. The standard benchmark hierarchy proposal.

The vital importance of a new journal is to guarantee that high-quality submissions receive high-quality reviews promptly. According to the past experiences in the other reputable journals and conferences in the computer discipline, which is my primary background, I have some considerations.

In the computer discipline, a journal paper often cannot receive consistent and timely reviews compared with other top-tier conferences. For example, different associate editors invite reviewers from uncertain sources to handle papers with large deviations. Instead, a program committee meeting provides comparatively consistent reviews at a top-tier conference.

Another issue is the significant delay. Overall, the average turnaround of handling a paper is from three months to a year. Some journals reject most submissions at the disposal of a staff who does not understand its content to speed up the process and reduce the external review load. That will harm our community for two reasons. First, the value of peer review is to provide constructive feedback, which is the stone of our scientific community. Second, it will result in the abuse of editor rights. The last issue is most journals adopt a single-blind review, which prevents fair review.

To resolve the above issues, I enact the following plans. (1) Consistent and reliable reviews. In addition to about thirty founding editors or editors, similar to the program committee member of a conference, we will invite approximately 30 associate editors (Junior researchers with Ph.D. degrees). The associate editor is similar to the external review committee member of a conference. A team of founding editors, editors, and associate editors will provide the basis for consistent and reliable reviews.

(2) Fast-track peer review. The editor-in-chief (EIC) will read each paper's abstract and introduction. Suppose the team thinks this is a high-quality paper with high impact potential. In that case, they will invite three editors to have a timely review, including possible remote discussion and make a final decision within three weeks. The team will ask one editor and two associate editors to review the other papers. Overall, the team will finish one round of decisions within one month.

(3) A double-blind review process. One member of the EIC team without conflict of interest (COI) is responsible for checking COIs, while the other EIC and editor, who do not know the authors' identities, make a final decision. Each published article is reviewed by a minimum of three independent reviewers using a double-blind peer-review process. The identities of the reviewers are not known to the authors, and the reviewers also do not know the identities of the authors.

Acknowledgments

I am very grateful to many persons' contributions to TBench, especially Prof. Dr. Tony Hey for discussing the TBench plan, Dr. Lei Wang for discussing and proofreading this article, Mr. Shaopeng Dai for compiling the references, Mr. Qian He for drawing the figures, Mr. Zhengxin Yang for discussing the metrology related work, Ms. Chitra Krishnamoorthy, Ms. Divyaa Veluswamy, and other KeAI and Elsevier staffs for publishing TBench. Without all of you, launching TBench is impossible.

References

[1] A. Clare, Performance evaluation, in: The CFA Institute Investment Foundations, 2014, pp. 173–205.
[2] S.S. Stevens, et al., On the Theory of Scales of Measurement, Bobbs-Merrill, College Division, 1946.
[3] M. Zairi, P. Leonard, Origins of benchmarking and its meaning, in: Practical Benchmarking: The Complete Guide, Springer, 1996, pp. 22–27.
[4] B.C. Lewis, A.E. Crews, The evolution of benchmarking as a computer performance evaluation technique, MIS Q. (1985) 7–16.
[5] E.O. Joslin, Evaluation and performance of computers: application benchmarks: the key to meaningful computer evaluations, in: Proceedings of the 1965 20th National Conference, 1965, pp. 27–37.
[6] W. Buchholz, A synthetic job for measuring system performance, IBM Syst. J. 8 (4) (1969) 309–318.
[7] D.M. Conti, Findings of the Standard Benchmark Library Study Group, (500–538) Sept. of Commerce, National Bureau of Standards, Institute for Computer Sciences and Technology, 1978.
[8] I. BIPM, I. IFCC, I. IUPAC, O. ISO, The international vocabulary of metrology—basic and general concepts and associated terms (VIM), 3rd edn. JCGM 200: 2012, in: JCGM (Joint Committee for Guides in Metrology), 2012.
[9] R.N. Kacker, On quantity, value, unit, and other terms in the JCGM international vocabulary of metrology, Meas. Sci. Technol. 32 (12) (2021) 125015.
[10] F. Tang, W. Gao, J. Zhan, C. Lan, X. Wen, L. Wang, C. Luo, Z. Cao, X. Xiong, Z. Jiang, et al., Aibench training: balanced industry-standard AI training benchmarking, in: 2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), IEEE, 2021, pp. 24–35.
[11] Z. Jiang, W. Gao, F. Tang, L. Wang, X. Xiong, C. Luo, C. Lan, H. Li, J. Zhan, HPC AI500 V2. 0: The methodology, tools, and metrics for benchmarking HPC AI systems, in: 2021 IEEE International Conference on Cluster Computing (CLUSTER), IEEE, 2021, pp. 47–58.
[12] MIT, Automl benchmark datasets, 2021, https://openml.github.io/automlbenchmark/benchmark_datasets.html Accessed December 2, 2021.
[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[14] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.
[15] IOSCO, Financial Benchmarks, Technical Report, 2013.
[16] L. Wang, X. Xiong, J. Zhan, W. Gao, X. Wen, G. Kang, F. Tang, Wpc: Whole-picture workload characterization across intermediate representation, isa, and microarchitecture, IEEE Comp. Archit. Lett. (2021).
[17] F. Arute, K. Arya, R. Babbush, D. Bacon, J.C. Bardin, R. Barends, R. Biswas, S. Boixo, F.G. Brandao, D.A. Buell, et al., Quantum supremacy using a programmable superconducting processor, Nature 574 (7779) (2019) 505–510.
[18] J. Wells, B. Bland, J. Nichols, J. Hack, F. Foertter, G. Hagen, T. Maier, M. Ashfaq, B. Messer, S. Parete-Koon, Announcing Supercomputer Summit, Technical Report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2016.
[19] Y. Liu, X. Liu, F. Li, H. Fu, Y. Yang, J. Song, P. Zhao, Z. Wang, D. Peng, H. Chen, et al., Closing the "quantum supremacy" gap: achieving real-time simulation of a random quantum circuit using a new Sunway supercomputer, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2021, pp. 1–12.
[20] F. Pan, W. Wang, A.K. Tung, J. Yang, Finding representative set from massive data, in: Fifth IEEE International Conference on Data Mining (ICDM'05), IEEE, 2005, pp. 8–pp.
[21] 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.
[22] K. Huppler, The art of building a good benchmark, in: Technology Conference on Performance Evaluation and Benchmarking, Springer, 2009, pp. 18–30.
[23] Statista, Number of apps available in leading app stores, 2021, https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/, accessed at Dec 2, 2021.

Dr. Jianfeng Zhan is a Full Professor at Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), and University of Chinese Academy of Sciences (UCAS), and director of the Software Systems Labs, ICT, CAS. He received his B.E. in Civil Engineering and MSc in Solid Mechanics from Southwest Jiaotong University in 1996 and 1999, and his Ph.D. in Computer Science from Institute of Software, CAS, and UCAS in 2002. His research areas span from Chips, Systems to Benchmarks. A common thread is benchmarking, designing, implementing, and optimizing a diversity of systems. He has made substantial and effective efforts to transfer his academic research into advanced technology to impact general-purpose production systems. Several technical innovations and research results, including 35 patents, from his team, have been adopted in benchmarks, operating systems, and cluster and cloud system software with direct contributions to advancing the parallel and distributed systems in China or even in the world. He has supervised over ninety graduate students, post-doctors, and engineers in the past two decades. Dr. Jianfeng Zhan founds and chairs BenchCouncil and serves as the Co-EIC of TBench with Prof. Tony Hey. He has served as IEEE TPDS Associate Editor since 2018. He received the second-class Chinese National Technology Promotion Prize in 2006, the Distinguished Achievement Award of the Chinese Academy of Sciences in 2005, and the IISWC Best paper award in 2013, respectively.