

IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems

IEEE Computer Society

Developed by the
Artificial Intelligence Standards Committee

IEEE Std 2937™-2022

STANDARDS

IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems

Developed by the
Artificial Intelligence Standards Committee (C/AISC)
of the
IEEE Computer Society

Approved 16 June 2022

IEEE SA Standards Board

Abstract: Artificial intelligence (AI) computing differs from generic computing in terms of device formation, operators, and usage. AI server systems, including AI server, cluster, and high-performance computing (HPC) infrastructures are designed specifically for this purpose. The performance of these infrastructures is important to users not only on generic models but also on the ones for specific domains. Formal methods for the performance benchmarking for AI server systems are provided in this standard, including approaches for test, metrics, and measure. In addition, the technical requirements for benchmarking tools are discussed.

Keywords: artificial intelligence (AI), artificial intelligence computing, artificial intelligence server system, IEEE 2937™, performance benchmarking, performance benchmarking tool

The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2022 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published 14 September 2022. Printed in the United States of America.

IEEE is a registered trademark in the U.S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

PDF: ISBN 978-1-5044-8854-9 STD25525

Print: ISBN 978-1-5044-8855-6 STDPD25525

IEEE prohibits discrimination, harassment, and bullying.

For more information, visit <https://www.ieee.org/about/corporate/governance/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system or otherwise, without the prior written permission of the publisher.

Important Notices and Disclaimers Concerning IEEE Standards Documents

IEEE Standards documents are made available for use subject to important notices and legal disclaimers. These notices and disclaimers, or a reference to this page (<https://standards.ieee.org/ipr/disclaimers.html>), appear in all standards and may be found under the heading “Important Notices and Disclaimers Concerning IEEE Standards Documents.”

Notice and Disclaimer of Liability Concerning the Use of IEEE Standards Documents

IEEE Standards documents are developed within the IEEE Societies and the Standards Coordinating Committees of the IEEE Standards Association (IEEE SA) Standards Board. IEEE develops its standards through an accredited consensus development process, which brings together volunteers representing varied viewpoints and interests to achieve the final product. IEEE Standards are documents developed by volunteers with scientific, academic, and industry-based expertise in technical working groups. Volunteers are not necessarily members of IEEE or IEEE SA and participate without compensation from IEEE. While IEEE administers the process and establishes rules to promote fairness in the consensus development process, IEEE does not independently evaluate, test, or verify the accuracy of any of the information or the soundness of any judgments contained in its standards.

IEEE makes no warranties or representations concerning its standards, and expressly disclaims all warranties, express or implied, concerning this standard, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. In addition, IEEE does not warrant or represent that the use of the material contained in its standards is free from patent infringement. IEEE standards documents are supplied “AS IS” and “WITH ALL FAULTS.”

Use of an IEEE standard is wholly voluntary. The existence of an IEEE Standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard.

In publishing and making its standards available, IEEE is not suggesting or rendering professional or other services for, or on behalf of, any person or entity, nor is IEEE undertaking to perform any duty owed by any other person or entity to another. Any person utilizing any IEEE Standards document, should rely upon his or her own independent judgment in the exercise of reasonable care in any given circumstances or, as appropriate, seek the advice of a competent professional in determining the appropriateness of a given IEEE standard.

IN NO EVENT SHALL IEEE BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Translations

The IEEE consensus development process involves the review of documents in English only. In the event that an IEEE standard is translated, only the English version published by IEEE is the approved IEEE standard.

Official statements

A statement, written or oral, that is not processed in accordance with the IEEE SA Standards Board Operations Manual shall not be considered or inferred to be the official position of IEEE or any of its committees and shall not be considered to be, nor be relied upon as, a formal position of IEEE. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that the presenter's views should be considered the personal views of that individual rather than the formal position of IEEE, IEEE SA, the Standards Committee, or the Working Group.

Comments on standards

Comments for revision of IEEE Standards documents are welcome from any interested party, regardless of membership affiliation with IEEE or IEEE SA. However, **IEEE does not provide interpretations, consulting information, or advice pertaining to IEEE Standards documents.**

Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments. Since IEEE standards represent a consensus of concerned interests, it is important that any responses to comments and questions also receive the concurrence of a balance of interests. For this reason, IEEE and the members of its Societies and Standards Coordinating Committees are not able to provide an instant response to comments, or questions except in those cases where the matter has previously been addressed. For the same reason, IEEE does not respond to interpretation requests. Any person who would like to participate in evaluating comments or in revisions to an IEEE standard is welcome to join the relevant IEEE working group. You can indicate interest in a working group using the Interests tab in the Manage Profile & Interests area of the [IEEE SA myProject system](#).¹ An IEEE account is needed to access the application.

Comments on standards should be submitted using the [Contact Us](#) form.²

Laws and regulations

Users of IEEE Standards documents should consult all applicable laws and regulations. Compliance with the provisions of any IEEE Standards document does not constitute compliance to any applicable regulatory requirements. Implementers of the standard are responsible for observing or referring to the applicable regulatory requirements. IEEE does not, by the publication of its standards, intend to urge action that is not in compliance with applicable laws, and these documents may not be construed as doing so.

Data privacy

Users of IEEE Standards documents should evaluate the standards for considerations of data privacy and data ownership in the context of assessing and using the standards in compliance with applicable laws and regulations.

¹ Available at: <https://development.standards.ieee.org/myproject-web/public/view.html#landing>.

² Available at: <https://standards.ieee.org/content/ieee-standards/en/about/contact/index.html>.

Copyrights

IEEE draft and approved standards are copyrighted by IEEE under US and international copyright laws. They are made available by IEEE and are adopted for a wide variety of both public and private uses. These include both use, by reference, in laws and regulations, and use in private self-regulation, standardization, and the promotion of engineering practices and methods. By making these documents available for use and adoption by public authorities and private users, IEEE does not waive any rights in copyright to the documents.

Photocopies

Subject to payment of the appropriate licensing fees, IEEE will grant users a limited, non-exclusive license to photocopy portions of any individual standard for company or organizational internal use or individual, non-commercial use only. To arrange for payment of licensing fees, please contact Copyright Clearance Center, Customer Service, 222 Rosewood Drive, Danvers, MA 01923 USA; +1 978 750 8400; <https://www.copyright.com/>. Permission to photocopy portions of any individual standard for educational classroom use can also be obtained through the Copyright Clearance Center.

Updating of IEEE Standards documents

Users of IEEE Standards documents should be aware that these documents may be superseded at any time by the issuance of new editions or may be amended from time to time through the issuance of amendments, corrigenda, or errata. An official IEEE document at any point in time consists of the current edition of the document together with any amendments, corrigenda, or errata then in effect.

Every IEEE standard is subjected to review at least every 10 years. When a document is more than 10 years old and has not undergone a revision process, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE standard.

In order to determine whether a given document is the current edition and whether it has been amended through the issuance of amendments, corrigenda, or errata, visit [IEEE Xplore](#) or [contact IEEE](#).³ For more information about the IEEE SA or IEEE's standards development process, visit the IEEE SA Website.

Errata

Errata, if any, for all IEEE standards can be accessed on the [IEEE SA Website](#).⁴ Search for standard number and year of approval to access the web page of the published standard. Errata links are located under the Additional Resources Details section. Errata are also available in [IEEE Xplore](#). Users are encouraged to periodically check for errata.

Patents

IEEE Standards are developed in compliance with the [IEEE SA Patent Policy](#).⁵

Attention is called to the possibility that implementation of this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. If a patent holder or patent applicant

³ Available at: <https://ieeexplore.ieee.org/browse/standards/collection/ieee>.

⁴ Available at: <https://standards.ieee.org/standard/index.html>.

⁵ Available at: <https://standards.ieee.org/about/sasb/patcom/materials.html>.

has filed a statement of assurance via an Accepted Letter of Assurance, then the statement is listed on the IEEE SA Website at <https://standards.ieee.org/about/sasb/patcom/patents.html>. Letters of Assurance may indicate whether the Submitter is willing or unwilling to grant licenses under patent rights without compensation or under reasonable rates, with reasonable terms and conditions that are demonstrably free of any unfair discrimination to applicants desiring to obtain such licenses.

Essential Patent Claims may exist for which a Letter of Assurance has not been received. The IEEE is not responsible for identifying Essential Patent Claims for which a license may be required, for conducting inquiries into the legal validity or scope of Patents Claims, or determining whether any licensing terms or conditions provided in connection with submission of a Letter of Assurance, if any, or in any licensing agreements are reasonable or non-discriminatory. Users of this standard are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. Further information may be obtained from the IEEE Standards Association.

IMPORTANT NOTICE

IEEE Standards do not guarantee or ensure safety, security, health, or environmental protection, or ensure against interference with or from other devices or networks. IEEE Standards development activities consider research and information presented to the standards development group in developing any safety recommendations. Other information about safety practices, changes in technology or technology implementation, or impact by peripheral systems also may be pertinent to safety considerations during implementation of the standard. Implementers and users of IEEE Standards documents are responsible for determining and complying with all appropriate safety, security, environmental, health, and interference protection practices and all applicable laws and regulations.

Participants

At the time this standard was completed, the AI System and Application Test Working Group had the following entity membership:

Wei Bao, Chair
Fang You, Vice Chair
Yibo Zhang, Secretary

<i>Organization Represented</i>	<i>Name of Representative</i>
Beihang University	Zhongzhi Luan
Beijing Baidu Netcom Science Technology Co., Ltd.	Liang Xiong
China Electronics Standardization Institute	Wei Bao
Huawei Technologies Co., Ltd.	Fang You
Inspur Electronic Information Industry Co., Ltd.	Shaohua Wu
Intel Corporation	Haining Wang
Institute of Software Chinese Academy of Science	Lingzhong Meng
Lenovo Group Limited	Xiaoling Zhou
Megvii Technology Limited	Yibo Zhang
NVIDIA Corporation	Ritika Borkar
Tencent	Haitao Liu
YITU Technology	Chunhao Zhao

The Working Group gratefully acknowledges the contributions of the following participants. Without their assistance and dedication, this standard would not have been completed.

Ritika Borkar	Chaoming Liang	Shaohua Wu
Deepak Canchi	Haitao Liu	Zhao Xia
Xiaoqi Cao	Ada Lu	Liang Xiong
Ramesh Chukka	Zhongzhi Luan	Yang Xu
Jian Dong	Chao Ma	Yunzhi Xue
Kefeng Fan	Jingqing Mei	Yuze Yang
Chao Huang	Lingzhong Meng	Qi Zhang
Baihong Li	Hongzhi Tao	Qun Zhang
Rengang Li	Haining Wang	Chunhao Zhao
Yanhua Li	Kai Wang	Xiaoling Zhou
	Wei Wen	

The following members of the entity Standards Association balloting group voted on this guide. Balloters may have voted for approval, disapproval, or abstention.

1 st Cycle Corporation	Beijing RealAI Technology Co.,	Guangdong OPPO Mobile
AI Speech Co., Ltd.	Ltd.	Telecommunications Corp.,
Alipay (China) Technology Co.,	China Electronics	Ltd.
Ltd.	Standardization Institute	Haier Group Corporation
Beihang University	China Southern Power Grid Co.,	Hangzhou Qulian Technology
Beijing Baidu Netcom Science	Ltd.	Co., Ltd.
Technology Co., Ltd.	CloudWalk Technology	Honor Device Co., Ltd.
Beijing Knowledge Atlas	Dell	Huawei Technologies Co., Ltd.
Technology Co., Ltd.	Dreamerse	Inspur Electronic Information
	General Electric	Industry Co., Ltd.

Institute of Biomedical
Engineering, Chinese
Academy of Medical
Sciences & Peking Union
Medical College
Institute of Software Chinese
Academy of Science
Intel Corporation
JD.com, Inc.
Lenovo Group, Ltd.
Megvii Technology, Ltd.
Nanjing Research Institute of
Next-Generation Artificial
Intelligence Co., Ltd.

NVIDIA Corporation
NXP Semiconductors
SenseTime Group, Ltd.
Shandong Computer Science
Center
Shandong Institute of
Standardization
Shanghai Huagongbao
E-commerce, Ltd.
Shanghai Jiao Tong University
Shanghai Platform for Smart
Manufacturing Co., Ltd.
Siemens Corporation
State Grid Corporation of China

Tencent
The Boeing Company
Xiaomi Communications Co.,
Ltd.
YITU Technology
Yokosuka Telecom Research
Park, Inc.
Zhejiang Lab
Zhejiang University
ZTE Corporation

When the IEEE SA Standards Board approved this standard on 16 June 2022, it had the following membership.

David J. Law, *Chair*
Ted Burse, *Vice Chair*
Gary Hoffman, *Past Chair*
Konstantinos Karachalios, *Secretary*

Edward A. Addy
Ramy Ahmed Fathy
J. Travis Griffith
Guido R. Hiertz
Yousef Kimiagar
Joseph L. Koepfinger*
Thomas Koshy
John D. Kulick

Johnny Daozhuang Lin
Kevin Lu
Daleep C. Mohla
Andrew Myles
Damir Novosel
Annette D. Reilly
Robby Robson
Jon Walter Rosdahl

Mark Siira
Dorothy V. Stanley
Lei Wang
F. Keith Waters
Karl Weber
Sha Wei
Philip B. Winston
Daidi Zhong

*Member Emeritus

Introduction

This introduction is not part of IEEE Std 2937-2022, IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems.

Artificial intelligence (AI) server systems, including AI server, AI server cluster, and AI high-performance computing infrastructures, are the main forms of computing devices providing specific computing capabilities for AI computing for domain applications. The performance of an AI server system is crucial to the effectiveness and efficiency of AI application in domain vertical applications.

However, the technical solutions of AI server systems are diverse, and the abilities of domain users vary in term of applying parallel training and inference with data in various scales. These issues create difficulties for users in promoting the actual performance of AI server systems. Tests, observations, and analysis on the performance bottlenecks of AI server systems are needed, as well as important prerequisites of AI application efficiency optimization.

For the performance test of AI server systems, this standard specifies a systematic set of approaches with rules, selected workloads, metrics, and requirements on toolkits. It can be applied by all organizations, including AI computing device manufacturers, server manufacturers, AI service providers, developers, and end users of AI server systems, for the performance test and promotion for AI server systems.

Contents

1.	Overview.....	11
1.1	Scope.....	11
1.2	Word usage	11
2.	Normative references	11
3.	Definitions, acronyms, and abbreviations.....	12
3.1	Definitions.....	12
3.2	Acronyms and abbreviations.....	13
4.	AI server system performance benchmarking framework	14
4.1	Overview.....	14
4.2	Modes of test.....	14
4.3	Information of a test.....	15
5.	Training test.....	16
5.1	Overview.....	16
5.2	Scenarios.....	18
5.3	Metrics and measure approach.....	22
5.4	Requirement on toolkit.....	28
6.	Inference test.....	30
6.1	Overview.....	30
6.2	Workload.....	31
6.3	Scenarios.....	32
6.4	Metrics and measure approach.....	35
6.5	Requirement on toolkit.....	42
	Annex A (informative) Utilization ratio	43
	Annex B (normative) Rules for the selection and use of power measurement devices	45
	B.1 Power meter	45
	B.2 BMC.....	46
	Annex C (informative) Effective computing ability.....	49
	C.1 Effective computing ability of training	49
	C.2 Effective computing ability of inference.....	49
	Annex D (informative) Extra computing jobs for “mixed” query arrival mode.....	50
	D.1 File operation	50
	D.2 Balance binary tree search.....	50
	Annex E (informative) Bibliography.....	51

IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems

1. Overview

1.1 Scope

This specification defines the methodology for performance testing of artificial intelligence (AI) server systems, including AI servers, AI server clusters, and AI high-performance computing systems.

1.2 Word usage

The word *shall* indicates mandatory requirements strictly to be followed in order to conform to the standard and from which no deviation is permitted (*shall* equals *is required to*).^{6,7}

The word *should* indicates that among several possibilities one is recommended as particularly suitable, without mentioning or excluding others; or that a certain course of action is preferred but not necessarily required (*should* equals *is recommended that*).

The word *may* is used to indicate a course of action permissible within the limits of the standard (*may* equals *is permitted to*).

The word *can* is used for statements of possibility and capability, whether material, physical, or causal (*can* equals *is able to*).

2. Normative references

The following referenced documents are indispensable for the application of this document (i.e., they must be understood and used, so each referenced document is cited in text and its relationship to this document is explained). For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments or corrigenda) applies.

⁶ The use of the word *must* is deprecated and cannot be used when stating mandatory requirements, *must* is used only to describe unavoidable situations.

⁷ The use of *will* is deprecated and cannot be used when stating mandatory requirements, *will* is only used in statements of fact.

This standard does not contain normative references.

3. Definitions, acronyms, and abbreviations

3.1 Definitions

For the purposes of this document, the following terms and definitions apply. The *IEEE Standards Dictionary Online* should be consulted for terms not defined in this clause.⁸

artificial intelligence (AI) server: Server with specific design and implementation that provides computing capabilities for artificial intelligence applications in an information system.

NOTE—Server for generic computing supports AI computing by AI acceleration instruction set, acceleration circuit, acceleration module, and so on. This kind of server is a type of AI server.⁹

artificial intelligence (AI) server cluster: Group of connected AI servers, which is regarded externally as one server to jointly provide dedicated processing capabilities for artificial intelligence applications in an information system.

artificial intelligence (AI) server system: Collectively, denomination of artificial intelligence server, artificial intelligence server cluster, and artificial intelligence high-performance computing facilities.

concurrency valve: Minimal rate of incoming query occurrences over seconds.

NOTE 1—Queries are sent to an AI server system for process in a certain inference test.

NOTE 2—In an inference test in rush query arrival mode, a period of time exists in which the rate of incoming queries per second is no less than a defined concurrency valve value.

NOTE 3—Example: In an inference test in rush query arrival mode, if the concurrency valve value is 2^{10} queries/s, the number of incoming queries per second is not less than 2^{10} .

high-performance computing (HPC) system: Computing system that is designed, sold, and optimized to execute highly parallel applications.

NOTE 1—Computing devices, such as general-purpose computing on graphics processing units, neural-network processing units, field-programmable gate arrays, or other accelerating processors, are used in HPC systems for achieving high processability, specifically for those tasks that need large-scale parallel computing.

NOTE 2—HPC systems consist of multiple clustered servers (primarily for increased computational capability, high-speed interprocess interconnects, and large and high bandwidth memory capability).

NOTE 3—HPC systems may be purposely built or assembled from more commonly available computer systems.

NOTE 4—When an HPC system contains accelerating processors (see NOTE 1) and is built for the purpose of processing artificial intelligence (AI) tasks, such as training machine learning models or inference, it is called an “AI HPC.”

performance: Measurable results of a system under test.

NOTE 1—In the context of an artificial intelligence (AI) server system performance test, quantitative results are measured to represent performance.

NOTE 2—Performance can be measured or computed based on one or more defined metrics (such as time consumption, speed, etc.) that are critical to or required by users, such that the concerned behaviors or characteristics of an artificial intelligence server system can be represented.

⁸ *IEEE Standards Dictionary Online* is available at: <http://dictionary.ieee.org>. An IEEE account is required for access to the dictionary, and one can be created at no charge on the dictionary sign-in page.

⁹ Notes in text, tables, and figures of a standard are given for information only and do not contain requirements needed to implement this standard.

preprocessing: Part of an inference process, which transforms raw data into a state such that a specific machine learning model can process.

NOTE 1—Preprocessing includes data normalization, scaling, per-example arithmetic calculation, image padding, format change, signal sampling, framing, analog-digital conversion, and so on.

NOTE 2—Preprocessing takes use of the above approaches whenever needed depending on the dataset and machine learning (ML) model, to make input data to be processable by an ML model.

postprocessing: Part of an inference process, which transforms the results out of a machine learning algorithm or model to the form that is accepted by a user or a system.

query: Data packet containing test samples sent by the tester to the system under test.

system under test (SUT): System that processes a test item given by the tester and returns the result in a test.

test sample: Data sample used by a test.

test system: System used by a tester to conduct tests and measure performances.

NOTE 1—A test system consists of hardware, software, and data necessary to drive a system under test (SUT) to run a test.

NOTE 2—A test system does not refer to the machine learning software toolkit or accelerating library necessary to run a task required by a test.

tester: Organization that organizes, designs, and/or conducts a performance test.

NOTE—A tester is different with a test participant who interacts with a SUT in the run of a test.

timing: Operation to get and return the current timestamp of a system under test.

utilization ratio: Portion of the time period, in which the artificial intelligence (AI) server system's accelerating processors are executing training or inference instructions.

3.2 Acronyms and abbreviations

AI	artificial intelligence
AUC	area under curve
BCE	binary cross entropy
BLUE	bilingual evaluation understudy
DLRM	deep learning recommendation model
FP16	half-precision floating-point format
FP32	single-precision floating-point format
FPGA	field-programmable gate array
GPGPU	general-purpose computing on graphics processing unit
HPC	high-performance computing
ID	identifier
LAMB	layer-wise adaptive moments optimizer for batch training
LARS	least angle regression
MAP	mean average precision
MIOU	mean intersection over union
ML	machine learning

NPU	neural-network processing unit
RAM	random access memory
SGD	stochastic gradient descent
TFLOPS	tera floating-point operations per second
TOP1 ACC	TOP1 accuracy
WER	word error rate

4. AI server system performance benchmarking framework

4.1 Overview

To examine the performance of an AI server system, under a specified dataset, developers, users, testing institutions, enterprise, or other tested parties independently select a model and use AI servers, AI server clusters, or AI high-performance computing facilities for training and conducting an inference test. Based on the business needs, different modes of test can be selected and applied.

4.2 Modes of test

4.2.1 Open mode

4.2.1.1 Open training

Given the training dataset and requested accuracy requirement, the open training task shall be performed with an AI server system (to be tested) to train a selected model to meet the accuracy on the test dataset.

4.2.1.2 Open inference

Given the test dataset and accuracy criteria, the open inference task shall be performed with an AI server system (to be tested) to run the inference process with a model provided by the tested party. The accuracy of the provided model meets the given accuracy requirement on the given test dataset.

4.2.2 Close mode

4.2.2.1 Closed training

Given the training dataset, target model structure, and accuracy criteria, closed training shall use an AI server system (to be tested) to run modeling and optimization algorithms to obtain the target model that meets the accuracy threshold on the given test set.

4.2.2.2 Closed inference

Given the test dataset, reference model, and accuracy criteria, closed inference shall use an AI server system (to be tested) to run the inference process of the model to obtain the inference result that meets the accuracy threshold on the given test set.

4.2.3 Scene

According to actual application scenarios, the scene types of an AI server system performance test are as follows:

- a) Generic specific: For common problems, refer to or use publicly available models and datasets to complete model training and inference tasks.

- b) Domain specific: For industry problems, use special models and datasets to complete model training and inference tasks.

When performing a test, the models in this standard can be selected by a tested party.

4.3 Information of a test

For a test, the tested party shall provide the following information to the tester:

- a) Name of the tested party.
- b) Test ID (granted by a tester, for distinguishing tests and their runs).
- c) Division type (0 for close, 1 for open).
- d) Whether this test is generic or not (0 for generic, 1 for domain specific).
- e) Whether this test is a training test or not (0 for training, 1 for inference).
- f) Test item index (5.2 and 6.3).
- g) Result submission timestamp in format [yyyy:MM:dd HH:mm:ss] (recorded by toolkit at tester side).
- h) Type of the tested target (0 for AI server, 1 for cluster or data center).
- i) Number of AI servers [valid when item h) is 1].
- j) Information of each type of AI server [including model number, computing ability (in TFLOPS), and number of accelerating processors for this test].
- k) Information of software components used by the performance test, including but not limited to the following:
 - 1) Names and versions of deep learning (or machine learning) frameworks and libraries.
 - 2) When a used component is published in an open-source platform (such as github), the identifier (such as one of URL and commit hash) to that component shall be provided.
- l) Communication protocol and maximal bandwidth.
- m) Topology of the targeted AI server system to be tested (0 for single AI server, 1 for leader/follower, 2 for ring, 3 for tree-like, and 4 for others).
- n) Operating system information (name, kernel name, and version).
- o) Whether to test with virtualized computing or container components (0 for apply, 1 for not apply).
- p) Virtualized computing or container component information (name and version) [valid when item n) is 0].
- q) Whether, or not to apply changeable mini-batch size (0 for no, 1 for yes) and batching scheme.
- r) Value of mini-batch size [positive integer and valid when item p) is 0].
- s) Name of optimizer for training [valid when item e) is 0].
- t) Whether to apply mixed precision during training [valid when item e) is 0].
- u) Whether to apply autoML for test [valid when item e) is 0 and item c) is 1].
- v) Whether to apply a certain parallelism mode [0 for not, 1 for model parallelism, 2 for data parallelism, 3 for mixed parallelism, and 4 for other type of parallelism modes; valid when item e) is 0].

- w) Whether to apply asynchronous parameter update [0 for not use, 1 for use; valid when item e) is 0 and item c) is 1].
- x) Whether to apply weight sparsification [0 for not use, 1 for use; valid when item e) is 1 and item c) is 1]. The name of the sparsification approach should be provided.
- y) Whether to apply quantization [0 for not use, 1 for use; valid when item e) is 1]. The name of the quantization approach should be provided.
- z) The method chosen for measuring power should be reported with all power metrics.
- aa) Name, version, and source address of dataset.

5. Training test

5.1 Overview

5.1.1 Process of training

The training process refers to the establishment of a machine learning model, based on a machine learning algorithm, by using training data. The agreement in this clause shall be used to guide a training test. A training process consists of the following phases:

- a) **Preparatory phase:** To conduct a training test, the training dataset shall be achieved before the actual run of training. When necessary, format exchange and encapsulation can be performed to satisfy the possible technical requirement of specific platforms.
- b) **Running phase:** In this phase, the actual run of training happens. For this purpose, a tested party shall implement and run a claimed test item, satisfying the corresponding criteria (see 5.2). The implementation of training code includes training data preprocess and loading, training, validation on entire validation dataset (every epoch unless specified differently under 5.2), format conversion, and persistence of trained model. The result of such implementation and run is the trained model. During training, necessary intermediate information shall be recorded and computation of metrics shall be performed. For a training test item, several consecutive runs of training targeting the same model are required (see 5.2). Average of metrics values among selected runs is helpful to reduce the uncertainty caused by the use of randomization in certain optimization processes. Runs are selected by dropping one slowest and one fastest run from the set of training runs. A nonconverging run is treated as the slowest run and dropped from metric calculation. A test result is invalid if more than one nonconverging run exists. Logs of runs are required to trace the training processes, which helps testers to determine the compliance of tests.
- c) **Report phase:** In a formal training test, the completion of the running phase triggers the report of result (5.1.3). A report of result refers to the transfer of the result from test party to tester via an approach agreed to by both.

5.1.2 Rules of training

For the best chance of fairness, the test process of training shall adhere to the following rules:

- a) During a test of training, the following actions shall not be taken:
 - 1) To use a training dataset different from the definition in 5.2 unless the training set is not specified.
 - 2) To apply transfer learning or pretraining of model mechanisms, for close division or open division.
 - 3) To change the formation of SUT in running phase.
 - 4) To delete any part of training data for a test item in close division, except the remains of which volume is less than one batch size.