Chinastandards.org

ICS 35.160
CCS L 61

chinastandards.org - PREVIEW

NATIONAL STANDARD OF THE PEOPLE'S REPUBLIC OF CHINA

GB/T 45087-2024

Performance test methods for artificial intelligence server systems

人工智能 服务器系统性能测试方法

Issued on: November 28, 2024 **Implemented on: November 28, 2024**

Issued by: State Administration for Market Regulation;
Standardization Administration of the PRC

Preview: page 1 of 5 shown, 44 in the full document. Complete English PDF, instant download:
https://chinastandards.org/standards/gb-t-45087-2024

Table of Contents

chinastandards.org · PREVIEW

Foreword

This document is in accordance with the provisions of GB/T 1.1-2020 "Guidelines for standardization work Part

1.Structure and drafting rules for standardization documents" Drafting. This document was proposed and coordinated by the National Technical Committee for Information Technology Standardization (SAC/TC28). This document was drafted by: China Electronics Technology Standardization Institute, Huawei Technologies Co., Ltd., Inspur Electronic Information Industry Co., Ltd. Intel (China) Co., Ltd., Pingtou Ge (Shanghai) Semiconductor Technology

Preview: page 2 of 5 shown, 44 in the full document. Complete English PDF, instant download:
https://chinastandards.org/standards/gb-t-45087-2024

Co., Ltd., iFlytek Co., Ltd., H3C Information Technology Co., Ltd. Technology Co., Ltd., Advanced Micro Devices (China) Co., Ltd., Beijing University of Aeronautics and Astronautics, Cambrian Technologies Co., Ltd., Nanjing NARI Ruiteng Technology Co., Ltd., China Southern Power Grid Co., Ltd. Ultra-high Voltage Transmission Company, Sinopec Yingke Information Technology Co., Ltd. Company, Guangdong Research Institute of China Telecom Co., Ltd., Shanghai Enflame Technology Co., Ltd., Institute of Software, Chinese Academy of Sciences, Beijing Ren Technology Development Co., Ltd., Shanghai Qianshi Technology Co., Ltd., Shanghai Supercomputing Center, Shanghai Wenyu Information Technology Co., Ltd., Midea Group Tuan (Shanghai) Co., Ltd., Guoke Chushi (Chongqing) Software Co., Ltd., Shanghai Artificial Intelligence Research Institute Co., Ltd., Sichuan Huakun Zhenyu Intelligent Technology Co., Ltd., Shenzhen Kunyun Information Technology Co., Ltd., China Railway Construction Corporation Limited, China Railway Fifth Survey and Design Institute Group Co., Ltd. Co., Ltd., Southwest University of Science and Technology. The main drafters of this document are. Dong Jian, Xu Yang, Zhang Qi, Wang Waner, Cao Xiaoqi, Huang Jianbin, Liang Zhaoming, Bao Wei, Wu Shaohua, Wang Haining, Lin Xiaodong, Ma Shanshan, Gao Hui, Zhang Yibo, Tao Yumei, Yang Yuze, Zheng Huiping, Liu Rubing, Li Lanbo, Ji Tuo, Luan Zhongzhi, Cheng Guipeng, Huang Xiancui, Jun Mu, Chao Shi, Heng Ye, Ning Wang, Dongqing Liu, Xianxu Li, Chunyu Shi, Jingqing Mei, Lingzhong Meng, Ruiquan Ding, Qiulin Cheng, Geng Wu, Huazhen Yu, Dandan Zhang, Zhong Kaitao, Ren Pei, Fu Xinjie, Hu Yanling, Song Haitao, Bai Shiyu, Liu Dong, Luan Lihong, Li Dong, Zheng Zhong, Yu Wenxin.

The AI server system includes AI servers, clusters, and high-performance computing facilities, and is a platform for various deep learning models. (including large-scale pre-trained models) is the core carrier of training and reasoning, and is the core tool for various industries to use artificial intelligence technology to improve production efficiency. The AI server system is designed to handle AI computing tasks and is very similar to general-purpose servers in terms of architecture, computing methods, and usage. The server systems are quite different, and their test processes, loads, and indicators are all unique. The benchmarking method is proposed, and requirements are put forward for the functionality and fairness of the benchmarking tools. The issuing organization of this document calls attention to the fact that when declaring compliance with this document, it may involve 7.4.2,

7.7.1 and the systemic nature of artificial intelligence servers. The use of patents related to the test method. The issuing organization of this document takes no position on the authenticity, validity and scope of this patent. The patent holder has promised to the issuing agency of this document that he is willing to cooperate with any applicant under reasonable and non-discriminatory terms and conditions. The patent holder's statement has been filed with the issuing agency of this document, and relevant information can be obtained through Get the contact information below. Patent holder. China Electronics Technology Standardization Institute Address. No. 1, Andingmen East Street, Dongcheng District,

Preview: page 3 of 5 shown, 44 in the full document. Complete English PDF, instant download:
https://chinastandards.org/standards/gb-t-45087-2024

Beijing Please note that in addition to the above patents, some of the contents of this document may still involve patents. The issuing agency of this document does not assume the responsibility for identifying patents. responsibility. Artificial Intelligence Server System Performance Testing Method

1 Scope

China's national performance test methods for artificial intelligence server systems. Buying an AI server is unusually difficult to do rationally, because the specifications that vendors quote - peak floating point operations, memory bandwidth, accelerator count - correlate weakly with how fast the machine actually trains or serves a model. The gap comes from everything between the accelerators: the interconnect topology, the memory hierarchy, the host, the storage feeding the data pipeline, and the software stack, any of which can leave the accelerators idle most of the time. So a meaningful comparison has to be made by running representative workloads under defined conditions, which is what this standard specifies. It defines the system performance test modes and describes methods for both training performance and inference performance, and applies to the performance testing and evaluation of AI server systems.

This document defines the server system performance test mode and describes the training performance and reasoning performance test of the artificial intelligence server system. method. This document is applicable to the performance testing and evaluation of artificial intelligence server systems.

2 Normative references

The contents of the following documents constitute the essential clauses of this document through normative references in this document. For referenced documents without a date, only the version corresponding to that date applies to this document; for referenced documents without a date, the latest version (including all amendments) applies to This document.

GB/T 41867-2022 Information Technology Artificial Intelligence Terminology

3 Terms and definitions

The terms and definitions defined in GB/T 41867-2022 and the following apply to this document.

3.1 System under test A system that processes the test jobs given by the tester and returns results that meet the requirements.

Note. The system under test consists of artificial intelligence server system hardware, operator implementation library, machine learning framework software, model compilation

Preview: page 4 of 5 shown, 44 in the full document. Complete English PDF, instant download:
https://chinastandards.org/standards/gb-t-45087-2024

components and other necessary hardware and software.

3.2 tested party An organization or individual that provides the system under test and test information and assists in the implementation of the test.

3.3 reference modelreferencemodel A standardized model for defining system test requirements. [Source: ISO /IEC 14776-414.2009, 3.1.87, modified]

3.4 Timing Get and return the current timestamp of the system under test.

Note. It is assumed that the time of each node in the system under test (3.1) is consistent.

3.5 A server in an information system that can provide high-performance computing and processing capabilities for artificial intelligence applications.

Note 1: AI servers contain computing modules designed specifically for AI computing, providing dedicated accelerated computing capabilities for AI applications. Note

2.Based on general-purpose servers, servers equipped with AI acceleration cards that provide dedicated computing acceleration capabilities for AI applications are called "AI" servers. Compatible with the server".

Note 3: A server designed specifically for AI accelerated computing and providing dedicated AI computing capabilities is called an "AI all-in-one server". [Source: GB/T 41867-2022, 3.1.3, modified]

chinastandards.org - PREVIEW

Preview: page 5 of 5 shown, 44 in the full document. Complete English PDF, instant download:
https://chinastandards.org/standards/gb-t-45087-2024