ICS 35.080
CCS L 80

中华人民共和国国家标准化指导性技术文件

GB/Z 328—2026

智能计算集群效能指标与测试方法

Effectiveness metrics and testing methods for intelligent computing clusters

2026-10-30 发布

国家市场监督管理总局
国家标准化管理委员会 发布

目　　次

前言 …… Ⅲ
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　缩略语 …… 1
5　效能指标体系框架 …… 2
6　集群测试指标 …… 3
6.1　训练效能 …… 3
6.2　推理效能 …… 8
7　测试方法 …… 9
7.1　训练效能 …… 9
7.2　推理效能 …… 12
8　测试报告 …… 14
附录 A（资料性）　倍率系数 r 选取经验 …… 16
附录 B（资料性）　测试基准模型参考 …… 17
附录 C（资料性）　测试报告模板 …… 18
附录 D（资料性）　测试框架说明 …… 20
参考文献 …… 21

前　言

本文件为规范类指导性技术文件。

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由全国智能计算标准化工作组(SAC/SWG 32)提出并归口。

本文件起草单位：之江实验室、阿里云计算有限公司、中国电信集团有限公司、中兴通讯股份有限公司、中电信数智科技有限公司、中国信息通信研究院、蚂蚁科技集团股份有限公司、成都超算中心运营管理有限公司、中国软件评测中心(工业和信息化部软件与集成电路促进中心)、浪潮电子信息产业股份有限公司、超聚变数字技术股份有限公司、中移(苏州)软件技术有限公司、海光信息技术股份有限公司、浙江大华技术股份有限公司、京东科技信息技术有限公司、联想(北京)有限公司、上海超级计算中心、上海曦智科技股份有限公司、深圳思谋信息科技有限公司、中国科学院计算机网络信息中心。

本文件主要起草人：杨非、蒙贵云、闫俊、张卫星、李振廷、任银垠、高丰、沈培、曹科杰、崔娜妮、李响、张宇、张乾、郭智慧、李永辉、黄强、何维、宇文梦柯、郭晓辉、汪可、周新中、滕耘、杨朋霖、孔维生、韩冬、杨四雄、张丹丹、孟怀宇、马刚波、吴璨。

智能计算集群效能指标与测试方法

1 范围

本文件规定了智能计算集群的效能指标体系框架，涵盖训练效能和推理效能的关键指标，描述了相应的测试方法，并给出了测试报告。

本文件适用于智能计算集群的设计、建设及运维等环节的效能测试与指标分析。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

计算节点 computing node

由计算设备构成，运行计算任务的基础单元。

[来源：GB/T 46572—2025，4.2]

3.2

智能计算集群 intelligent computing cluster

由多个计算节点、存储节点、网络节点组成的集合，用于执行大规模计算和数据处理任务。

3.3

效能 effectiveness

智能计算集群在给定运行条件和负载场景下，执行计算任务过程中所体现的计算性能、资源利用效率、系统协同的综合水平。

3.4

基准模型 benchmark model

用于效能测试的标准模型。

4 缩略语

下列缩略语适用于本文件：

BF16：布瑞恩半精度浮点数(Brain Floating-point)

FLOPS：每秒浮点运算次数(Floating Point Operations Per Second)

FP16：半精度浮点数(Half-precision Floating-point Format)

FP32：单精度浮点数(Single-precision Floating-point Format)

GPU：图形处理器(Graphics Processing Unit)

IOPS：每秒输入/输出操作次数(Input/Output Operations Per Second)

INT8：8 位整型数(8-bit Integer)