

中华人民共和国国家标准

GB/T XXXXX—XXXX

人工智能 计算中心 计算能力评估

Artificial intelligence — Computing centres — Computing capacity assessment

点击此处添加与国际标准一致性程度的标识

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX—XX—XX 发布

XXXX—XX—XX

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 I

引言 I

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 缩略语 2

5 总则 2

 5.1 评估对象 2

 5.2 评估内容概述 3

 5.3 评估框架 3

 5.4 分类原则 4

6 评估指标 4

 6.1 算力规模 4

 6.2 组网规模 5

 6.3 存储规模 5

 6.4 训练性能 5

 6.5 推理性能 6

 6.6 通信性能 7

 6.7 存储性能 7

 6.8 计算效率 8

 6.9 长稳运行时间 8

 6.10 故障恢复时间 8

 6.11 计算可用度 9

7 评估方法 9

 7.1 总则 9

 7.2 算力规模 9

 7.3 组网规模 10

 7.4 存储规模 10

 7.5 训练性能 10

 7.6 推理性能 10

 7.7 通信性能 10

 7.8 存储性能 11

 7.9 计算效率 12

7.10 长稳运行时间 12

7.11 故障恢复时间 12

7.12 计算可用度 123

附录 A （资料性）各类型智算中心的计算能力指标示例 14

附录 B （资料性）智算中心计算能力评估工具（AISBench-AICC）示例 15

参考文献..... 15

前 言

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。

本文件由全国信息技术标准化技术委员会(SAC/TC28)提出并归口。

本标准起草单位：

本标准主要起草人：

引 言

本文件的发布机构提请注意，声明符合本文件时，可能涉及到XXXXXX的专利的使用。专利名称如下：
本文件的发布机构对于该专利的真实性、有效性和范围无任何立场。

该专利持有人已向本文件的发布机构承诺，他愿意同任何申请人在合理且无歧视的条款和条件下，就专利授权许可进行谈判。该专利持有人的声明已在本文件的发布机构备案，相关信息可以通过以下联系方式获得：

联 系 人：XX

通讯地址：XX

邮政编码：XX

电子邮件：XX

电话：+8610-XX

传真：+8610-XX

网址：XX

请注意除上述专利外，本文件的某些内容仍可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

人工智能 计算中心 计算能力评估

1 范围

本文件规定了人工智能计算中心计算能力的评价指标，给出了评估方法。

本文件适用于人工智能计算中心计算能力的评估，也为人工智能计算中心规划、设计、建设和运维提供参考依据。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867—2022 信息技术 人工智能 术语
GB/T 42018—2022 信息技术 人工智能 平台计算资源规范
GB/T AAAAA -AAAA 人工智能 服务器系统性能测试方法
GB 40879—2021 数据中心能效限定值及能效等级
GB 50174—2017 数据中心设计规范

3 术语和定义

GB/T 41867-2022和GB/T 42018-2022界定的以及下列术语和定义适用于本文件。为了方便使用，以下重复列出了GB/T 41867—2022和GB/T 42018-2022中的某些术语和定义。

3.1

人工智能计算中心 artificial intelligence computing centers

智算中心

一种能够为多用户提供人工智能计算服务、数据容纳的结构或结构组。

注1：人工智能计算中心使用信息技术、电信网络设备提供人工智能计算任务中涉及的数据存储、处理、迁移和人工智能计算加速等功能，一般可提供深度学习模型训练计算、推理计算、模型微调等人工智能计算服务。

注2：人工智能计算中心包括计算、存储、网络等硬件设备，以及算力调度管理组件、深度学习框架、加速库等用以支撑人工智能计算的必要软件组件。

3.2

性能 performance

人工智能计算中心运行计算任务时，可被测量的特性。

注：性能可基于一个或多个参数（如运行时间、能耗、吞吐率、有效计算能力、每秒浮点运算次数、故障恢复时长等）的测量或计算获得，以表征在某设备（组）中运行的某技术过程的行为、特性及效率。

3.3

计算节点 computing node

人工智能计算中心实施人工智能计算的部件。

注：计算节点一般是人工智能加速器、人工智能加速卡、人工智能服务器或计算中心中其他任何表征人工智能计算性能的部分。

3.4

人工智能计算能力 artificial intelligence computing capability

执行人工智能任务或支持人工智能运算功能的能力。

4 缩略语

下列缩略语适用于本文件：

AI：人工智能（Artificial Intelligence）

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）

FP64：双精度浮点数（Double-precision Floating-point format）

INT8：8位整型数（8-bit Integer）

I/O：输入输出（Input/Output）

IOPS：每秒输入输出量（Input/Output Per Second）

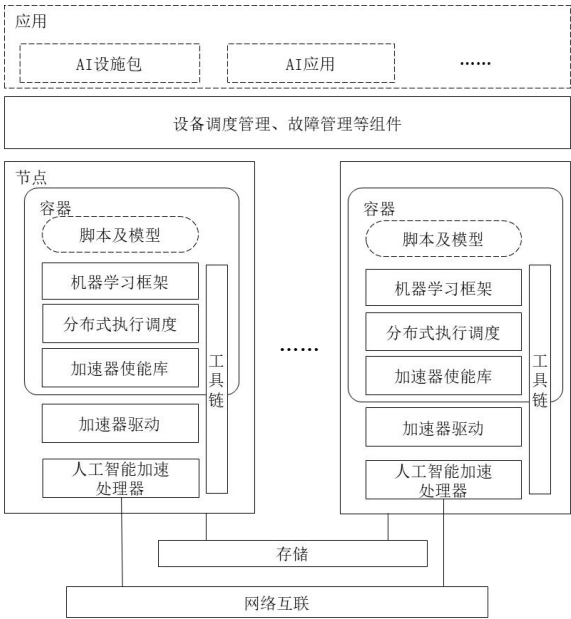
MTTR：平均故障恢复时长（Mean Time To Repair）

MTBF：平均无故障工作时长（Mean Time Between Failures）

5 总则

5.1 评估对象

智算中心满足高性能和高可靠的人工智能计算需求，为多个用户提供人工智能计算、数据处理等服务。用户使用人工智能计算资源时，一般通过网络将计算任务部署到计算中心的特定部分（节点或一组节点），在一定的计算能力配额之内，执行计算任务。智算中心（架构图见图1）的计算能力由其中包含的人工智能加速器、网络互联部件、数据存储设备、相关配套软件及应用实现共同决定。



注 1：图中用虚线框表示不属于智算中心必须具备的组件。

注 2：图中的运行环境一般指容器或其他能支撑应用运行的必要组件。

图 1 智算中心架构图

5.2 评估内容概述

智算中心的计算能力评估涉及多方面的考量，除了硬件资源本身能提供的物理算力规模外，还包括是否通过网络互联、任务调度、计算优化等技术有效整合物理资源，以及是否能稳定、持续、高效地将计算资源转化为服务人工智能任务的计算能力。本文件定义智算中心计算能力的评估指标体系，第5章提出计算能力评估的参考框架和分类指导，第6章规定了各维度的评估指标，第7章规定了评估方法。

5.3 评估框架

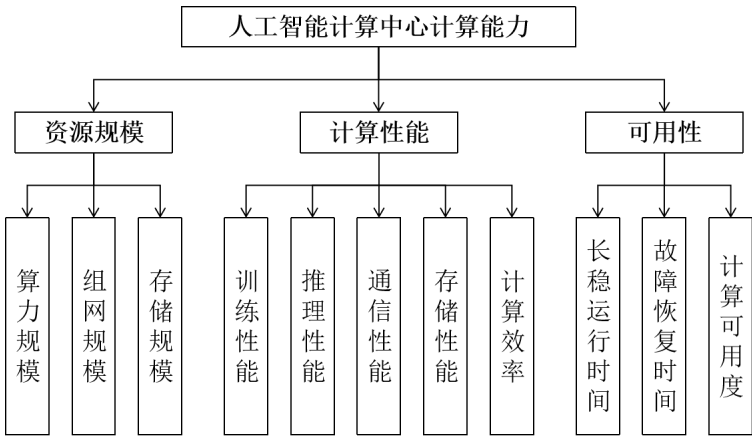


图 2 智算中心计算能力评估框架

智算中心计算能力评估主要包括三个维度，其中每个维度下又划分了评估域，每一个评估域下包含一个或一组评估指标：

a) 资源规模：

- 1) 资源规模是智算中心的固有属性;
 - 2) 资源规模由AI加速处理器、存储和网络设备、计算节点可扩展性,及基础设施建设、供电需求、物理空间等属性综合决定;
 - 3) 资源规模评估主要从算力规模、组网规模、存储规模三个评估域进行评估:其中,算力规模主要评估智算中心整体计算能力(如,所有计算节点总通用算力和所有计算节点总人工智能算力),组网规模主要评估组网数量、网络带宽等能力,存储规模主要评估整体存储系统性能、存储带宽等能力。
- b) 计算性能:
- 1) 计算性能指智算中心运行人工智能计算任务时可被测量的特性;
 - 2) 计算性能由集群网络通信设计、存储系统I/O能力、算力规模和算法优化等因素共同决定;
 - 3) 通过一组或多组任务场景下的时间效率、资源利用率和能力指标体现,以表征某一智算中心的计算特性和效率。
- c) 可用性:
- 1) 可用性是指智算中心计算能力可被用户使用的程度;
 - 2) 可用性通过计算任务长稳运行和故障恢复两部分指标体现;
 - 3) 可用性由计算中心部件设备及链路的可靠性、AI计算中心冗余备份等系统可靠性设计及故障快速恢复机制等因素共同决定。

5.4 分类原则

依据资源规模、计算性能和可用度三个维度综合评估计算能力,智算中心划分为5个类别(各类别对应的计算能力指标示例见附录A),分别适用于不同的应用场景和用户需求,分类原则如下:

- a) Cat1: 计算资源在数百TFLOPS至数PFLOPS(FP16),一般用于中小型企业日常处理的基础型AI任务,满足百亿以下参数大模型的训练,如垂直领域的行业大模型;
- b) Cat2: 计算资源在数十PFLOPS(FP16),一般用于大型企业日常处理更复杂的AI任务,满足百亿到千亿区间参数大模型的训练,如互联网智能推荐系统、实时数据分析和更复杂的AI模型;
- c) Cat3: 计算资源在数百PFLOPS(FP16),主要用于地区级算力需求,其为多个企业和机构提供共享的计算资源,满足千亿到万亿区间参数大模型的训练以及小规模租赁业务,支持区域内的大规模AI模型训练和推理需求,如城市智能交通系统、智慧城市管理;
- d) Cat4: 计算资源在数EFLOPS(FP16),主要用于国家级算力需求,满足万亿到十万亿区间参数大模型的训练,其用于支持全国范围内的大规模AI模型训练和应用以及大规模租赁业务,通常用于国家级重大项目和战略性研究,如国家级语言模型训练、基因组研究中的AI应用;
- e) Cat5: 计算资源在数十EFLOPS及以上(FP16),主要用于跨区域级算力需求,满足超十万亿参数大模型的训练,支持最前沿的AI研究和超大规模模型的训练,如超大规模语言模型、宇宙模拟中的AI应用。

6 评估指标

6.1 算力规模

智算中心中所含所有人工智能处理器的规格算力之和。算力规模 CA_B 的定义见公式(1),量纲和适用关系见表1:

$$CA_B = [CA_{INT8}, CA_{FP16}, CA_{BF16}, CA_{FP32}] \quad (1)$$

式中:

CA_{INT8}——智算中心中所含所有人工智能加速处理器对INT8精度计算能力之和；
CA_{FP16}——智算中心中所含所有人工智能加速处理器对FP16精度计算能力之和；
CA_{BF16}——智算中心中所含所有人工智能加速处理器对BF16精度计算能力之和；
CA_{FP32}——智算中心中所含所有人工智能加速处理器对FP32精度计算能力之和。
注1：除上述列出的示意精度外，测试时可根据被测系统支持的范围进行调整，增加额外支持的精度。
注2：智算中心的规格算力应包括所有用于人工智能加速计算的处理单元。

表 1 智算中心规格算力量纲和适用关系

算力类型	量纲
单位时间内四分之一精度整型操作数	OPS@INT8
单位时间内半精度浮点操作数	FLOPS@FP16
单位时间内布瑞恩半精度浮点操作数	FLOPS@BF16
单位时间内浮点操作数	FLOPS@FP32

注1：BF16由1比特符号位，8比特指数，7比特尾数组成。
注2：除上述列出的精度外，测试时可根据被测系统支持的范围进行调整，增加额外支持的精度。

6.2 组网规模

智算中心组网中处理节点的数量，表示网络所能连接计算部件的规模。

6.3 存储规模

智算中心存储规模包括总容量和可用容量：
a) 总存储容量：存储系统提供的存储盘物理容量总和；
b) 可用存储容量：存储系统可用于存放数据的容量总和。

6.4 训练性能

6.4.1 实际吞吐率

智算中心实际吞吐率（又称有效算力）见表2，反映智算中心对特定训练作业的处理能力。

注：对视觉类测试，单位是images/s；对自然语言处理类测试，单位是sentences/s；对自然语言生成类测试，吞吐率可为定长输入（句中单词或字的个数）、输出条件下，每秒处理的语素数量，单位是tokens/s。

表 2 智算中心实际吞吐率

指标	说明
单负载模型训练任务吞吐率	AI计算中心达到训练任务的测试中止条件时的数据量和任务执行时间的比值。
（可选）AI计算中心多模型训练吞吐率综合加速比	AI计算系统在给定任务集合S上，实际吞吐率与每任务基线吞吐率之比的加权几何平均

6.4.2 训练时间

训练时间是智算中心在特定数据集上训练一个模型使其达到目标准确率或完成特定训练轮数时所用的时间（不包括预处理和模型加载时间），总体训练用时及其细化指标见表3。

表3 智算中心训练时间

指标	说明
总体训练用时 T_T	从训练开始读入数据，到模型训练完毕、完成在非电易失性存储器上的持久化，所使用的总时长。
数据读入用时 $T_L^{b,e}$	为训练目的，训练数据被读入加速器内存（使用通用计算环境、加速器缺少或不配置存储时，可为主存），达至可用状态，所使用的时间
训练启动用时 T_v	多加速器训练时，从训练开始指令到所有加速器都被分配并开始执行训练任务所经历的时长
训练用时 T_{TR}^{cd}	从训练开始命令调用到训练退出 ^a 之间的时间间隔
第 i 个（ i 为正整数）训期运行用时 $T_{EP(i)}$	训练过程第 i 次遍历（使用）训练集所用的时间
第 j 次（ j 为正整数）验证用时 $T_{V(j)}$	第 j 次使用验证数据集试运行当前模型，得出当前模型准确率等指标值的过程
模型格式转化用时 T_{CV}	训练完毕后，将结果模型转化为要求格式所耗费的时间
模型持久化用时 T_p	将加速器内存中的模型读出，并完整写入非电易失性存储所用的时间
节点间通信时延 T_{NC}	源节点开始发送数据至目标节点完全接收数据的用时
注：假设训练数据已封装为机器学习框架能够处理的格式。 ^a 训练退出可有多种充分条件。正常结束训练退出条件，如测试集准确率门限等。 ^b 数据并行时，数据读入用时为数据读入开始至所有工作节点都完整获得所需数据的总用时（含网络传输用时）。 ^c 训练任务的用时包含数据预处理用时。 ^d 表中受测系统无法统计的时间，不作要求。 ^e 数据读入过程可伴随训练同步发生，时间计入训练用时。	

6.5 推理性能

6.5.1 并发吞吐量

智算中心线上系统服务每秒能成功处理的用户请求数量，表征系统在高并发情况下的人工智能计算任务处理能力和性能。

6.5.2 推理时延

智算中心多次处理推理任务或提供推理服务的间隔时间，时间单位为毫秒（ms），推理任务总延时而其细化指标见表4。

表4 智算中心推理时延

指标	说明
推理任务总延时 ^a T_I	多次连续推理端到端总延时

单样本推理延时 T_{IN}	被测者对某样本推理的开始时间与结束时间的差
首语素延时 $T_{f token}$	被测者收到样本时间与被测者发送第一个语素时间的差
下个语素平均延时 $T_{n token}$	被测者发送某一个语素与发送下一个语素时间差的平均值
^a 因作业到达模式不同，推理总延时 T_I 可能包括被测者等待作业的间隔时间	

6.6 通信性能

6.6.1 总线平面传输带宽

在通过总线协议连接的智算中心网络系统中，从一个计算节点通过总线架构到另一个计算节点（包括AI加速处理器之间，或通用处理器与AI加速处理器之间）可以传输的最高数据速率，单位为bps，表征处理器之间的数据传输能力。

注：通过总线协议以内存语义等协议特征实现通信的传输平面称为总线平面。

6.6.2 网络平面传输带宽

在智算中心网络系统中，从一个计算节点到另一个计算（存储）节点进行通信的最高传输速率，单位为bps，表征该节点或服务器能够向外发送数据的传输速率，是衡量网络节点通信性能的重要指标。

注：通过网络协议以网络语义等协议特征实现通信的传输平面称为网络平面。

6.6.3 总线平面传输时延

在通过总线协议连接的智算中心网络系统中，通过总线架构直接连接的两个计算节点（包括AI加速处理器之间，或通用处理器与AI加速处理器之间）的最短通信时间和最长通信时间，单位为ms，表征总线平面节点间数据传输的速率，反映节点间通信性能。

6.6.4 网络平面传输时延

在智算中心网络系统中，从一个网络节点到另一个网络节点进行通信的最短时间和最长时间，单位为ms，表征网络中节点间数据传输的速率，反映网络延迟性能。

6.6.5 带宽收敛比

智算中心网络架构中的所有下行接口的总带宽与所有上行接口的总带宽的比值。带宽收敛比反映了计算中心的网络通信性能。

注：带宽收敛，是指数据报文在计算中心网络架构的传输过程中，由于网络架构、网络设备等非故障原因而不能实现无阻塞交换。在带宽收敛时，网络设备会有部分端口会被拥塞，丢弃部分数据报文。

6.6.6 网络带宽

用一截面将网络划分为对等的两半时，或者两个节点数目都相同的子网时，穿过该截面的最大传输率。二分网络带宽是说明沿等分网络最大通信带宽的性能指标，表征了网络吞吐率能力。

6.7 存储性能

6.7.1 存储 IOPS

智算中心存储系统中所有存储节点瞬时能够处理的输入输出操作总次数。IOPS评估存储系统在高负载情况下的处理能力，反映系统的响应速度和效率。

6.7.2 存储带宽

智算中心存储系统中所有存储节点或存储设备的读/写带宽进行合成所得到的总带宽。存储带宽表征存储系统在特定时间内能够支持的最大数据传输速率。

6.8 计算效率

6.8.1 算力利用率

单位时间内在智算中心硬件上执行的浮点运算次数和计算设备的规格计算能力之比，用于衡量在指定负载下的软硬件协同优化能力，见公式7。

$$\text{算力利用率} = \frac{\text{FLOP}_{\text{step}}/t}{N * \text{FLOPS}_{\text{theory}}} \quad (7)$$

式中：

t ——单位耗时，单位秒s；

N ——智算中心中AI加速器数量；

$\text{FLOP}_{\text{step}}$ ——模型单步（优化）浮点运算次数；

$\text{FLOPS}_{\text{theory}}$ ——AI加速器理论每秒浮点运算次数。

6.8.2 多卡线性度

智算中心开展多卡并行计算时，同一计算任务从N个卡扩展到M个卡上的平均吞吐率的比值，见公式8。

$$\text{多卡线性度} = \frac{TP_N}{(TP_M * N)/M} = \frac{TP_N * M}{TP_M * N} \quad (8)$$

式中：

N, M ——计算任务卡数从N个扩展到M个， $M > N$ ；

TP_N ——在N个卡上执行计算任务的总吞吐率；

TP_M ——在M个卡上执行计算任务的总吞吐率。

6.9 长稳运行时间

模型训练过程中，两次业务中断之间的业务运行时长，表征系统的可靠性，反映系统在无中断状态下的持续工作的能力。

6.10 故障恢复时间

6.10.1 定义

智算中心在执行特定任务时，某部分或整体从发生故障到恢复正常运行所需要的平均时间，见公式5和6。

$$t_F = t_{F2} - t_{F1} \quad (5)$$

$$\overline{T_F} = \frac{1}{M} \sum T_F \quad (6)$$

式中：

T_{F1} ——发生同类故障而无法继续执行任务的时点；

T_{F2} ——每次故障被修复，任务重新获得执行的时点；

M ——发生同类故障的次数， $M \geq 3$ ；

T_F ——故障恢复用时，单位为毫秒ms。

6.10.2 故障分级

智算中心故障分为如下三个等级，应对不同等级的故障分别进行统计和处理：

a) 警告类故障：不影响AI服务器短时间正常运行、不会造成训练任务中断的故障，如AI服务器单个风扇故障、单个电源故障等可以快速替换部件恢复人工智能服务器正常运行的故障；

b) 紧急类故障：无法保证AI服务器短时间内正常运行但训练任务尚未中断，需尽快排查处理的故障，如AI加速器等关键部件温度超过阈值，单路电力系统故障，环境温湿度轻微超出阈值等经过处理可快速恢复且过程中没有造成训练中断或者人工智能服务器异常宕机的故障；

c) 严重类故障：造成AI服务器宕机或训练任务中断的故障。如OS宕机、AI加速器掉卡、网络端口异常断开、存储系统无法正常读写、电力中断、环境温湿度大幅超出阈值无法支持AI服务器正常运行等造成训练异常终止的故障。

注：修复方法一般包含自动修复和手动修复。

6.11 计算可用度

计算资源可用度是智算中心的正常运行时间占总运行时间的百分比。可用度表征了系统能够正常运行并提供服务的程度。见如下公式：

$$\text{计算可用度} = \frac{T_0}{T_0 + T_1} * 100\% \quad (7)$$

式中：

T_0 ——平均无故障工作时间；

T_1 ——平均故障恢复时间。

7 评估方法

7.1 总则

第7章所规定的评估方法，宜使用AISBench-AICC工具进行实施，工具说明和使用流程见附录B。

7.2 算力规模

智算中心的算力规模，应按以下方法测量评估：

- 对任一服务器单机；
- 关闭系统节能设置，保持所有核心同时运行；
- 核间通信带宽不低于测试负载容量；

- d) 按表1的规定，定义不同精度下，特定尺寸的矩阵乘法（元素为0或非0），并估计完成1次矩阵乘加所需计算量，矩阵尺寸应能完全使用核尺寸；
- e) 使用单核执行矩阵乘法，若干次获得平均时间；
- f) 求单核在特定精度下，单次矩阵乘的计算量与平均时间的商，作为单核标称计算能力；
- g) 累计服务器内所有该精度单核标称计算能力，作为单服务器特定精度标称计算能力；
- h) 累计所有服务器，在该精度下的标称计算能力。

7.3 组网规模

智算中心的组网规模指标，应按以下方法测量评估：

- a) 准备必要的网络拓扑测试工具或软件；
- b) 使用网络拓扑测试工具测量网络节点的数量，包括服务器、交换机、路由器等设备的数量。

7.4 存储规模

智算中心的存储规模指标，应按以下方法测量评估：

- a) 确定存储集群中所有存储设备的总物理容量，统计所有存储盘的物理容量总和，记录为存储系统的裸容量；
- b) 考虑存储配置的冗余和数据保护机制（如RAID），计算存储系统的可用容量，即实际可用于数据存放的容量。

7.5 训练性能

7.5.1 实际吞吐率

智算中心的训练任务的实际吞吐率应按照GB/T AAAAA-AAAA中关于AI服务器集群训练过程实际吞吐率的测量方法评估。

7.5.2 训练时间

智算中心训练相关用时应按照GB/T AAAAA-AAAA中关于AI服务器集群训练相关用时的测量方法评估。

7.6 推理性能

7.6.1 并发吞吐量

智算中心的并发吞吐量指标，应按以下方法测量评估：

- a) 编写测试脚本模拟实际工作负载或直接使用系统服务核心代码段作为负载；
- b) 设置测试时间 t 需大于一个请求的完整响应时间，从较低的并发数开始，逐渐增加并发用户数或请求频率，观察系统响应；
- c) 记录各并发量级测试阶段的关键性能指标，如总吞吐量 s 等；
- d) 计算各并发量级的吞吐量，选择其中的最大值作为系统服务并发吞吐量。

7.6.2 推理时延

智算中心的推理时延应按照GB/T AAAAA-AAAA中关于AI服务器集群推理时间测量方法评估。

7.7 通信性能

智算中心的通信性能指标，应按以下表7所示方法测量评估。

表 4 通信传输效率测量方法

指标	测量方法
总线平面传输带宽	a) 根据均匀分布的概率模型或其他指定概率模型,选取m对同一总线协议连接系统内的AI加速处理器,选取n对同一总线协议连接系统内通用处理器与AI加速处理器, m和n要满足合理采样次数。 b) 对每一组采样处理器使用带宽测量工具进行测量; c) 计算m对AI加速处理器之间带宽的均值作为AI加速处理器之间的总线平面传输带宽;计算n对通用处理器和AI加速处理器之间带宽的均值作为通用处理器和AI加速处理器之间的总线平面传输带宽。
网络平面传输带宽	a) 根据均匀分布的概率模型或其他指定概率模型,选取p对网络系统中的节点,确保采样覆盖不同拓扑路径,以满足合理的测试覆盖范围。 b) 使用网络性能测试工具进行带宽测量,记录节点间的最大传输速率,重复测试不少于三次,确保数据可靠性。 c) 计算p对节点间带宽的平均值,作为网络平面传输带宽的测量结果。这个值反映节点或服务器的最大数据传输能力。
总线平面传输时延	a) 根据均匀分布的概率模型或其他指定概率模型,选取m对同一总线协议连接系统内的AI加速处理器,选取n对同一总线协议连接系统内通用处理器与AI加速处理器, m和n要满足合理采样次数。 b) 对每一组采样处理器使用时延测量工具测量最短时间和最长时间; c) 分别计算m+n对采样处理器的最短时延均值和最长时延均值,作为总线平面点对点时延。
网络平面传输时延	a) 根据均匀分布的概率模型或其他指定概率模型,选取q对网络系统中的节点进行时延测试,确保覆盖不同区域的节点, q需满足合理的采样要求。 b) 使用时延测量工具,测量每对节点之间的最短通信时间和最长通信时间,分别执行多次测量(至少10次),确保数据的稳定性和准确性。 c) 分别计算q对节点之间的最短时延均值和最长时延均值,作为网络平面点对点时延的测量结果,用于表征网络节点间的数据传输延迟。
带宽收敛比	a) 准备必要的带宽测量工具或软件; b) 使用带宽测量工具获得每台服务器与汇聚交换机间的上行带宽,与每台服务器与接入交换机间的下行带宽; c) 计算所有服务器下行带宽之和与上行带宽之和,并计算二者的比值得到收敛比。
网络带宽	a) 准备必要的网络带宽测试工具或软件; b) 使用网络带宽测试工具/软件测试通信系统网络带宽,执行网络带宽测试; c) 在同一网络平面内,对每台计算设备,进行m($3 \leq m \leq 10$)次数网络带宽测量,记录其中最高网络带宽; d) 对所有计算设备的最高网络带宽求和作为网络带宽。

7.8 存储性能

智算中心的存储性能指标,应按以下表7所示方法测量评估。

表 5 存储性能指标测量方法

指标	测量方法
存储系统IOPS	a) 明确存储系统影响因素，如磁盘类型、RAID级别、缓存大小、块大小、操作系统和文件系统等； b) 使用IOPS测试工具测量存储系统I/O性能，执行I/O测试； c) 进行m（ $3 \leq m \leq 10$ ）次I/O测试，记录测试过程中的IOPS； d) 取记录的IOPS最大值作为存储系统IOPS。
存储带宽	a) 对聚合链路内任一存储服务器； b) 使用带宽测试工具测量带宽，执行数据传输测试； c) 对每一存储服务器进行m（ $3 \leq m \leq 10$ ）次数据传输测试，记录其中最高传输速率； d) 对每一存储服务器的最高传输速率求和作为聚合带宽。

7.9 计算效率

7.9.1 算力利用率

智算中心的算力利用率指标，应按以下方法测量评估：

- a) 使用性能测试工具，向计算中心提交一系列模型计算任务，确保任务充分覆盖所有计算资源；
- b) 监控计算集群在任务执行过程中的资源使用情况，包括CPU、GPU、内存等；
- c) 记录每个计算任务的实际执行时间和资源使用量（浮点运算次数）；
- d) 统计该计算中心的卡数和单卡芯片理论算力；
- e) 计算算力利用率。

7.9.2 多卡线性度

智算中心的多卡线性度指标，应按以下方法测量评估：

- a) 确定测试的AI加速器数量与型号，配置测试环境，确保一致性和稳定性；
- b) 选择合适的测试工具和工作负载，如大模型训练任务，分别在单卡、多卡环境中运行相同的任务；
- c) 记录不同卡数量下的性能数据，如计算时间、吞吐量等，收集每种配置的运行结果；
- d) 计算多卡相对于单卡性能的提升比例，绘制性能提升曲线，分析线性度。

7.10 长稳运行时间

智算中心的长稳运行时间指标，应按以下方法测量评估：

- a) 启动计算中心工作一段时间进行数据收集，通过故障记录系统或人工记录等方式，统计业务中断的发生时间、故障类型和恢复时间；
- b) 对于每次业务中断，记录其发生时间与上一个业务中断恢复时间之间的间隔；
- c) 计算长稳运行时间。

7.11 故障恢复时间

智算中心的平均故障恢复时间指标，应按以下方法测量评估：

- a) 明确故障范围与类型，如服务器硬件故障、网络中断、数据损坏、电力中断、软件故障等；

- b) 启动计算中心工作一段时间进行数据收集，通过故障记录系统或人工记录等方式，监测并记录故障的发生时间、故障类型和恢复时间；
- c) 计算所有故障恢复时间的平均值，作为故障恢复时间。

7.12 计算可用度

智算中心的计算可用度指标，应在长稳运行时间和故障恢复时间得出评估结果后，按照6.11定义计算得出。

附 录 A

(资料性)

各类型智算中心的计算能力指标示例

按照计算能力评估指标，各类型智算中心对应的量化要求如表A.1所示：

表 A.1 计算能力指标要求

评估维度	评估域	评估指标	1 类	2 类	3 类	4 类	5 类
规模	算力规模	规格算力 (FLOPS@FP16)	>1000T	>10P	>100P	>1E	>10E
	组网规模	网络规模	百卡以上	百卡级	千卡级	万卡级	十万卡级
	存储规模	存储容量	TB 级	TB 级	PB 级	PB 级	PB 级
性能	训练性能	实际吞吐率	--	--	--	--	--
		训练时间	--	--	--	--	--
	推理性能	并发吞吐量	--	--	--	--	--
		推理延时	--	--	--	--	--
	通信性能	总线平面传输带宽	100GB/s	200GB/s	400GB/s	400GB/s	800GB/s
		网络平面传输带宽	100GB/s	200GB/s	200GB/s	400GB/s	400GB/s
		总线平面传输时延	100ns	100ns	100ns	100ns	100ns
		网络平面传输时延	500ns	500ns	500ns	500ns	100ns
		网络带宽	200GB/s* 数十卡	200GB/s* 百卡	200GB/s*万 千卡	200GB/s* 万卡	200GB/s* 十万卡
	存储性能	存储 IOPS	100k	1000k	10000k	100000k	1000000k
		存储带宽	100GB/s	1000GB/s	10TB/s	100TB/s	1000TB/s
	计算效率	算力利用率	--	--	--	--	--
		多卡线性度	--	--	--	--	--
可用性	长稳运行	平均无中断工作时间	>15 天	>7 天	>5 天	>1 天	>12 小时
	故障恢复	平均故障恢复时长	<30 分钟	<30 分钟	<1 小时	<1 小时	<1 小时
	可用度	计算资源可用度	>95%	>90%	>90%	>85%	>85%

附录 B
(资料性)
智算中心计算能力评估工具 (AISBench-AICC) 示例

B.1 工具说明

智算中心计算能力评估工具套件 (AISBench-AICC) 是人工智能计算系统性能测试工具。AISBench-AICC适用于大型人工智能服务器集群、智算中心的规模、性能、可用性等指标评估，兼容主流人工智能加速器类型，如CPU、GPU、NPU等，兼容主流深度学习软件框架。AISBench-AICC是本标准的配套评估工具。

B.2 工具使用流程

AISBench-AICC使用流程见图B-1：

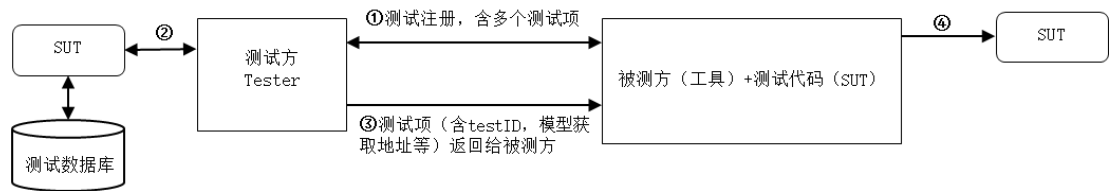


图 B.1 AISBench-AICC 使用流程

- 其中：
- a) 由测试方（测试机构）与被测方确立测试项；
 - b) 测试方在本地测试数据库注册测试项，并生成测试ID；
 - c) 测试方返回每个测试项的ID给被测方，被测方编辑、调试测试代码，在正式测试脚本中使用ID；
 - d) 被测方启动正式测试，测试方探知并监视测试、获得测试结果。

参 考 文 献

- [1] Nikolas et. al. Elasticity in Cloud Computing: What It Is, and What It Is Not[C]. Proceedings of the 10th International Conference on Autonomic Computing (ICAC 2013). CA, US: 2013.
- [2] Nikolas et. al. Ready for Rain? A View from SPEC Research on the Future of Cloud Metrics[R]. Technical Report SPEC-RG-2016-01, SPEC Research Group – Cloud Working Group, Standard Performance Evaluation Corporation (SPEC), 2016.
- [3] Liu, W. ; et al. SSD: Single Shot MultiBox Detector[J]. Computer Vision – ECCV 2016. LNCS, vol9905. Springer. 2016: 21–37.
- [4] Ren, S. ; et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137–1149.
- [5] Chen, L. C; et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. CoRR abs/1706.05587. arXiv. 2017.
- [6] Devlin, J. ; et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. CoRR abs/1810.04805. arXiv. 2018.
- [7] Amodei, D. ; et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin[C]. Proceedings of the 33rd International Conference on Machine Learning (ICML), NY, US: JMLR W&CP vol 48, 2015: 1–28.
- [8] Zhou, X. ; et al. EAST: An Efficient and Accurate Scene Text Detector[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), HI, US: IEEE, 2017: 2642–2651.
- [9] Schroff, F. ; et al. FaceNet: A Unified Embedding for Face Recognition and Clustering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), MA, US: IEEE, 2015: 815–823.
- [10] Vaswani, A. ; et al. Attention Is All You Need[J]. abs/1706.03762. arXiv. 2017.
- [11] Cheng, H. T. ; et al. Wide & Deep Learning for Recommender Systems[C]. Proceedings of the 1st Workshop on Deep Learning for Recommender Systems: ACM, 2016: 7 – 10.
- [12] Wu, Y. ; et al. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation[J]. CoRR abs/1609.8144. arXiv. 2016.
- [13] GB/T 25000.22-2019 系统与软件工程系统与软件质量要求和评价(SQUARE) 第22部分:使用质量测量[S]
- [14] GB/T 17166-2019 能源审计技术通则[S]
- [15] Giladi, R. and Ahituv, N. ; SPEC as a Performance Evaluation Measure[J]. Computer, 1995, 28(8): 33–42.
- [16] GB/T 41867—2022 信息技术 人工智能 术语[S]
- [17] GB/T 42018—2022 信息技术 人工智能 平台计算资源规范[S]
- [18] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023. <https://arxiv.org/pdf/2302.13971.pdf>
- [19] Du Z, Qian Y, Liu X, et al. Glm: General language model pretraining with autoregressive blank infilling[J]. arXiv preprint arXiv:2103.10360, 2021. <https://arxiv.org/pdf/2103.10360>
- [20] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern

recognition. 2022: 10684–10695.

http://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf

[21] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 586–595.

http://openaccess.thecvf.com/content_cvpr_2018/papers/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.pdf

[22] Dosovitskiy A, Brox T. Generating images with perceptual similarity metrics based on deep networks[J]. Advances in neural information processing systems, 2016, 29.

<https://proceedings.neurips.cc/paper/2016/file/371bce7dc83817b7893bcdeed13799b5-Paper.pdf>
