中国计量测试学会

量学函〔2026〕10号

中国计量测试学会关于《智能算力平台性能测试方法》团体标准征求意见的函

各有关单位:

根据国家标准化管理委员会、民政部印发的《团体标准管理规定》及《中国计量测试学会团体标准管理办法》有关规定，经中国计量测试学会批准立项，由中国信息通信研究院，贵州省计量测试院、博大数据、新华三技术有限公司、贵州师范大学、贵州电网有限责任公司信息中心、贵州省算力科技有限责任公司、中国电信股份有限公司云计算贵州分公司等单位牵头起草的《智能算力平台性能测试方法》团体标准现已完成征求意见稿的编制，为保证标准的科学性、严谨性和适用性，现面向社会广泛公开征求意见。

请各有关单位及专家对上述标准提出宝贵意见和建议，于2026年5月24日前将《征求意见反馈表》反馈至以下联系方式。

联系人：陈龙泉

电话：15001327806

电子邮箱：chenlongquan@caict.ac.cn

附件：1.《智能算力平台性能测试方法》征求意见稿

2.《智能算力平台性能测试方法》编制说明

3.征求意见反馈表

中国计量测试学会

2026年4月24日

附件1

ICS 点击此处添加 ICS 号
CCS 点击此处添加 CCS 号

T/CSMT

团　　体　　标　　准

T/XXX XXXX—XXXX

智能算力平台性能测试方法

Testing method for performance of Artificial Intelligence Computing Platforms

（征求意见稿）

（本征求意见稿完成时间：2025 年 9 月）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中国计量测试学会　　发 布

目　　次

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。

本文件由中国计量测试学会提出、归口并实施。

本文件起草单位：中国信息通信研究院，贵州省计量测试院、博大数据、新华三技术有限公司、贵州师范大学、贵州电网有限责任公司信息中心、贵州省算力科技有限责任公司、中国电信股份有限公司云计算贵州分公司。

本文件主要起草人：　陈龙泉、胡天洋、毛宇博、沈岸平、孙小强、张大元、宋茂江、杨霏、廖蔚松、杨俊实、白旭、万晓兰、陈彦、撒兴杰、张中、李享

本文件为首次制定。

引　言

智能算力平台是一种面向人工智能应用的，提供人工智能算法模型训练与模型运行服务的计算机系统平台，通常包括AI计算终端、AI服务器、AI算力服务器集群或各个集群组成的算力网络等。随着数字经济的快速演进，算力已逐步成为等同于“水、电、煤、汽”的基础性公共资源，形成支撑经济社会发展、赋能千行百业创新的重要新质生产力。尤其在大模型应用的持续驱动下，作为人工智能产业的创新基础，算力的需求量爆炸式增长。而作为算力资源的主要支撑主体——智能算力平台，其性能测试方法不统一，评价标准不一致，针对现有的智能算力平台，无法准确、全面地表征算力供给能力和算力有效使用量，难以继续有效引导算力产业健康发展。无论在科研领域，还是在应用领域，智能算力平台的测量标准化都是一项非常有意义并亟需完善的工作。中国计量测试学会组织本文件的起草，为智能算力平台性能评价提供依据。

智能算力平台性能测试方法

1 范围

本文件规定了智能算力平台性能测试的术语和定义、原理、测试设备和仪器、测试条件、测试方法、数据处理和测试报告等。

本文件适用于AI计算终端、AI服务器、AI算力服务器集群或各个集群组成的算力网络的关键性能测试，具备相同计算功能的平台可参照执行。

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。

YD/T 6048-2024 数据中心算力技术要求和测评方法

3 术语、定义和缩略语

3.1 术语和定义

YD/T 6048-2024界定的以及下列术语和定义适用于本文件。

3.1.1 智能算力平台 Artificial intelligence computing platform

面向人工智能应用，提供人工智能算法模型训练与模型运行服务的计算机系统平台，通常包括AI计算终端、AI服务器、AI算力服务器集群或各个集群组成的算力网络等。

3.1.2 算力 Computational power

算力是智能算力平台对数据进行处理并实现结果输出的一种能力，是衡量平台计算能力的一个综合指标，数值越大代表综合计算能力越强。

3.1.3 算力能效 Computational efficiency

智能算力平台的输出算力与消耗功率的比值,即“智能算力平台单位功率所产生的算力”,是同时考虑计算性能与功率的一种效率。数值越大，代表单位功率的算力越强，效能越高。

3.1.4 综合算力 Comprehensive computational power

智能算力平台不同类型的计算单元输出计算能力的综合评价指标，包括CPU、GPU、ASIC、FPGA等芯片的输出计算能力。

3.2 缩略语

下列缩略语适用于本文件：

AI：人工智能（artificial intelligence）
ASIC：专用集成电路（application specific integrated circuit）
BF16：布瑞恩浮点数（Brain Floating-point）
CPU：中央处理器（central processing unit）
FLOPS：每秒浮点运算次数（floating-point operations per second）
FP16：16位半精度浮点数（half-precision 16-bit floating-point）
FP32：32位单精度浮点数（single-precision 32-bit floating-point）
FPGA：现场可编程逻辑门阵列（field programmable gate array）
GPU：图形处理器（graphics processing unit）

INT4：4位整型数（4-bit integer）
INT8：8位整型数（8-bit integer）
IOPS：每秒读写操作次数（input/output operations per second）
TPU：张量处理器（tensor processing unit）

4 测试原理

智能算力平台的组成如图1所示，通过整合硬件、软件和算法资源，高效提供大规模计算能力，以支持人工智能训练/推理、高性能计算、大数据分析等任务。一般由以下四部分组成：

（1）硬件设备

依托CPU、GPU、TPU、FPGA、ASIC等芯片提供基础计算，并加速人工智能模型的训练和推理的计算单元，是智能算力的基础支撑。

（2）软件框架

提供算法开发、模型训练、推理部署等全链条的支持，通过优化算法和计算流程，提高算力利用效率，常见的软件框架包括TensorFlow，PyTorch、Cafe等，是智能算力的运行环境。

（3）算法优化

通过模型剪枝、量化、蒸馏等技术，减少计算复杂度，提升算力利用效率，是人工智能算力平台的关键优化途径。

（4）数据存储与传输

高效的数据存储和传输机制可减少数据访问延迟和带宽占用，从而提高整体计算效率，是人工智能算力平台的主要交互接口。

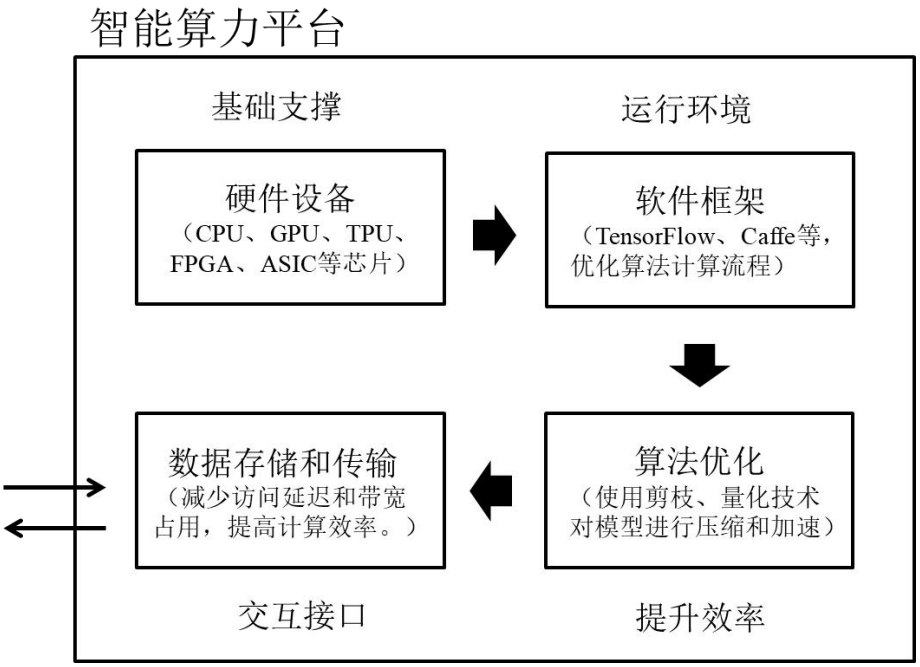

图 1 智能算力平台组成示意图

针对智能算力平台的性能测试，主要是建立统一、可控的标准算力负载环境下，对包含硬件设备、软件框架、算法优化、数据存储与传输的整体系统性能进行测试。其中算力基准测试集软件（Benchmark、CPUZ等）是测试方法的核心，用于对被测平台的算力量值进行实时监测，该软件一般需经计量技术机构进行比对验证后方可用于基础算力的测量。另外AI训练框架软件（Deepspeed等）、AI推理框架软件（Ollama等）用以构建标准的测试数据集用例，以应对训练或推理场景下的测试需求。

5 测试设备和仪器

5.1 测试设备

5.1.1 算力基准测试集软件

测量范围：（0～1.024×10^{18}）FLOPS，最大允许误差：±1%。

5.1.2 AI 推理框架软件

输出范围：（0~510×10^{15}）Tokens，最大允许误差：±1%。

5.1.3 AI 训练框架软件

输出范围：（0~510×10^{15}）Tokens，最大允许误差：±1%。

5.1.4 存储压力测试软件

接口速率：（0~24）Gbps，最大允许误差：±1%。

5.2 测试仪器

5.2.1 数据网络分析仪

支持10/100/1000 M电口，1G/10G/40G/100G/400G/800G bps光口等速率接口，对应数据传输速率最大允许误差：±1%。

支持吞吐量、时延测试功能。

5.2.2 功率计

功率：（0~500）W，最大允许误差：±1%。

6 测试条件

a) 测试环境应整洁，无腐蚀性介质，无影响测量结果的机械振动和电磁干扰；
b) 环境温度 24 ℃~35 ℃，湿度 20%~75%；
c) 电源电压及频率：220（1±10%）V，50（1±2%）Hz。

7 测试程序

7.1 测试前准备

测试前准备应按下列规定执行：

a) 被测平台外观应完好，标识清晰完整，无影响其正常工作的机械损伤；
b) 各部件应安装牢固，附件齐全；
c) 开启电源，被测平台应能正常工作。

7.2 测试方法

7.2.1 基础算力

如图2，基于被测人工智能算力平台的操作系统，安装并适配算力基准测试集软件；运行基准测试集软件，并记录基础算力显示值。

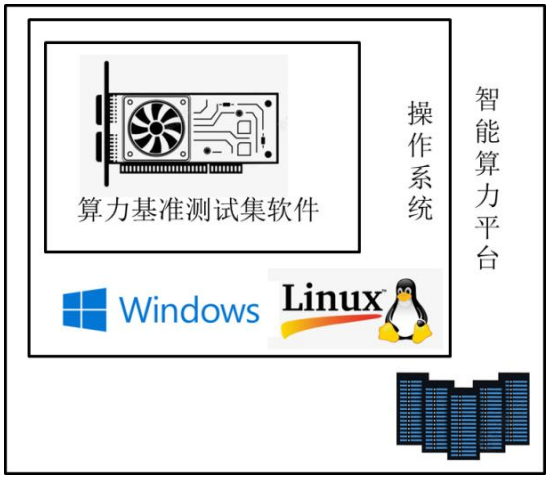

图 2　智能算力平台组成示意图

7.2.2　数据传输

如图3，将被测人工智能算力平台的数据传输端口连接数据网络测试仪，根据被测平台的技术说明书，设置数据网络测试仪的端口速率、工作模式等，保证对应端口之间正常通信。启动吞吐量测试，测量被测平台的数据传输速率，并记录数据网络测试仪上的传输速率显示值。启动时延测试，测量被测平台的数据转发时延，并记录数据网络测试仪上的转发时延显示值。

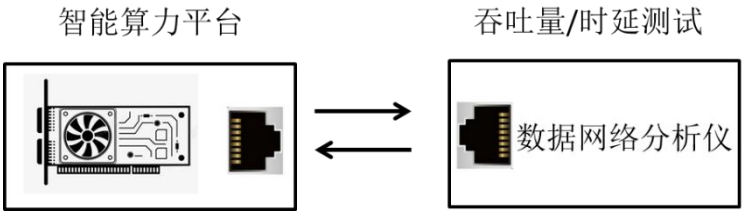

图 3　智能算力平台组成示意图

7.2.3　数据存储

如图4，基于被测人工智能算力平台的操作系统，安装并适配存储压力测试软件；运行存储压力测试软件，并记录数据存储速率显示值。

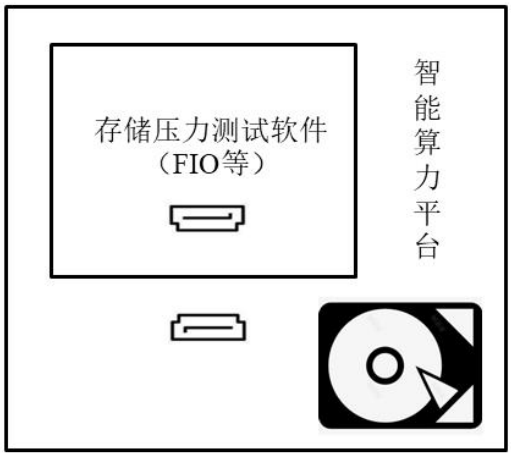

图 4　智能算力平台组成示意图

7.2.4　AI 推理场景算力

如图5，基于被测人工智能算力平台的操作系统，部署AI智能模型以及训练框架；运行AI智能模型训练，同步启动算力基准测试集软件测试AI训练场景下的算力，并记录算力显示值（数据精度类型为FP16/FP32/BF16/INT8）。

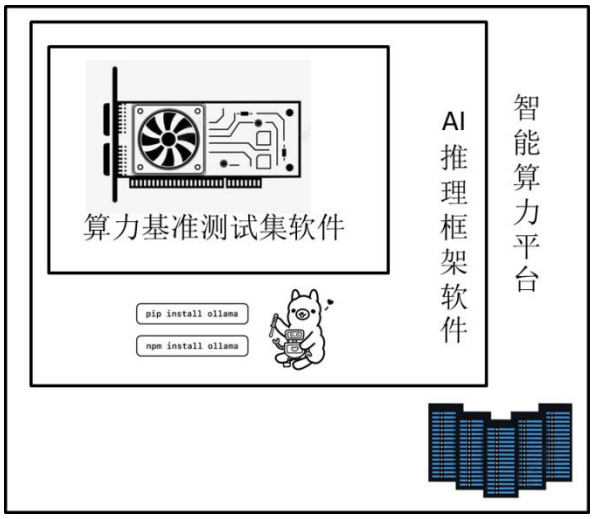

图 5　智能算力平台组成示意图

7.2.5　AI 训练场景算力

如图6，基于被测人工智能算力平台的操作系统，部署AI智能模型以及训练框架；运行AI智能模型训练，同步启动算力基准测试集软件测试AI训练场景下的算力，并记录算力显示值（数据精度类型为FP16/FP32/BF16）。

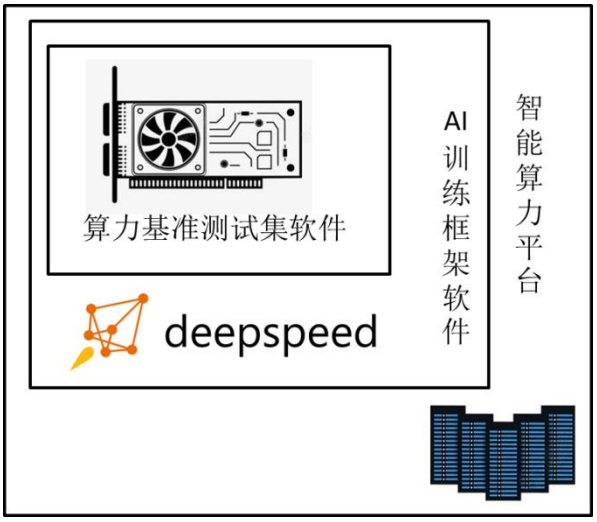

图 6　智能算力平台组成示意图

7.2.6　功耗

如图7，在被测人工智能算力平台不同运行场景下，在电源入口处，使用功率计来测量被测平台在不同算力负载下的功耗，记录功率计显示值。

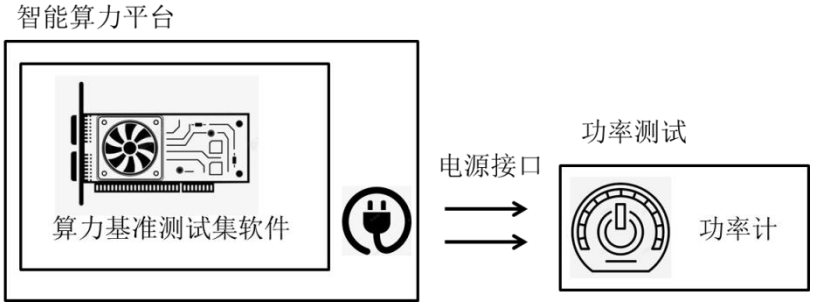

图 7　智能算力平台组成示意图

8　测试数据处理

8.1　综合算力

综合算力应参考基础算力、数据传输、数据存储、AI推理场景算力、AI训练场景算力等参数加权获取，各参数的权重应符合表1的要求。

表1 综合算力计算参数权重列表

序号	类别	归类权重	主要指标P_i	指标权重W_i
1	基础算力	30%	有效算力比例*ECPR*	30%
2	数据传输	20%	吞吐量*Throughout*	10%
			时延*Latency*	10%
3	数据存储	20%	存力规模*Storage_size*	10%
			磁盘读写速率*IOPS*	10%
4	AI推理场景算力	15%	有效推理算力*ICE*	15%
5	AI训练场景算力	15%	有效训练算量*TCE*	15%
6	功耗	0%	功率*P*	0%

其中对单台AI计算终端、AI服务器等中小规模智能算力平台，有效算力比例根据公式（1）计算：

$$ECPR = \frac{MCP}{NCP} \cdots\cdots (1)$$

式中：

NCP——被测智能算力平台的标称算力数值，单位为GFLOPS；

MCP——被测智能算力平台的测量基础算力数值，单位为GFLOPS；

$ECPR$——被测智能算力平台的有效算力比例，用单精度浮点数（FP32）表示，以百分数形式表示。

存力规模、磁盘读写速率通过存储压力测试软件直接获取，根据表2分为5个等级。

表2 数据存储能力分值

指标	一级（0.2）	二级（0.4）	三级（0.6）	四级（0.8）	五级（1.0）
存力规模（TB）	[0,10)	[10,50)	[50,100)	[100,500)	[500,+∞)
磁盘读写速率（MB/s）	[0,1000)	[1000,5000)	[5000,10000)	[10000,50000)	[50000,+∞)

吞吐量、时延通过数据网络分析仪直接获取，根据表3分为5个等级。

表3 数据传输能力分值

指标	一级（0.2）	二级（0.4）	三级（0.6）	四级（0.8）	五级（1.0）
吞吐量（Gbps）	[0,1)	[1,10)	[10,40)	[40,100)	[100,+∞)
时延（us）	[1000,+∞)	[100,1000)	[10,100)	[1,10)	[0,1)

有效推理算力根据公式（2）计算：

$$ICE = \frac{ICP}{NCP} \cdots\cdots (2)$$

式中：

NCP——被测智能算力平台的标称算力数值，单位为GFLOPS；

ICP——被测智能算力平台AI推理场景下的算力数值，单位为GFLOPS；

ICE——被测智能算力平台的有效推理算力，以百分数形式表示。

有效训练算力根据公式（3）计算：

$$TCE = \frac{TCP}{NCP} \cdots\cdots (3)$$

式中：

NCP——被测智能算力平台的标称算力数值，单位为GFLOPS；

TCP——被测智能算力平台AI训练场景下的算力数值，单位为GFLOPS；

TCE——被测智能算力平台的有效训练算力，以百分数形式表示。

智能算力平台的综合算力效率根据公式（4）计算：

$$CCE = \sum_i P_i \times W_i \cdots\cdots (4)$$

式中：

P_i——智能算力平台综合算力参考的各参数测试分值，见表2；

W_i——智能算力平台综合算力参考的各参数权重，见表2；

CCE——被测智能算力平台的综合算力效率，以百分数形式表示。

智能算力平台的综合算力根据公式（5）计算：

$$CCP = CCE \times NCP \cdots\cdots (5)$$

式中：

CCE——被测智能算力平台的综合算力效率，以百分数形式表示；

NCP——被测智能算力平台的标称算力数值，单位为GFLOPS；

CCP——被测智能算力平台的综合算力，单位为GFLOPS。

对AI算力服务器集群或各个集群组成的算力网络等大规模智能算力平台，其有效算力比例根据公式（6）计算：

$$CCP = \sum a_i\, CPU_i + \sum b_i\, GPU_i + \sum c_i\, FPGA_i + \sum d_i\, ASIC_i \cdots\cdots (6)$$

式中：

a_i——某型号CPU服务器数量；

b_i——某型号GPU服务器数量；

c_i——某型号FPGA服务器数量；

d_i——某型号ASIC服务器数量；

CPU_i——大规模某智能算力平台中某型号单台CPU服务器的综合算力；

GPU_i——大规模某智能算力平台中某型号单台GPU服务器的综合算力；

$FPGA_i$——大规模某智能算力平台中某型号单台FPGA服务器的综合算力；
$ASIC_i$——大规模某智能算力平台中某型号单台ASIC服务器的综合算力；
CCP——被测服务器集群或算力网络的综合算力。

8.2 算力能效

被测智能算力平台的算力能效根据公式（7）计算：

$$CPE = \frac{CCP}{P} \quad (1)$$

式中：
P——被测智能算力平台的整体功耗，可在电源入口处测量获得，单位为W；
CCP——被测智能算力平台的综合算力数值，单位为GFLOPS；
CPE——被测智能算力平台的算力能效，单位为GFLOPS/W。

9 测试报告

测试报告应包括下列内容：

a) 设备状态：送样单位，设备名称，设备型号，设备编号等；
b) 测试数据：智能算力平台的基础算力、数据存储、数据传输、AI 推理场景算力、AI 训练场景算力等原始测试数值。
c) 测试结果及分析：对单台AI计算终端、AI服务器等中小规模智能算力平台，应给出根据测试数据判定的综合算力效率、综合算力、算力能效等测试结果；对AI算力服务器集群或各个集群组成的算力网络等大规模智能算力平台，应给出算力平台内部的算力服务器组成、型号、数量、标称算力等基本信息，并根据实际测试数据判定的综合算力效率、综合算力、算力能效等测试结果。

附件2

中国计量测试学会

《智能算力平台性能测试方法》

团体标准编制说明

中国计量测试学会

“智能算力平台性能测试方法”团体标准制定工作组

2025 年 3 月

智能算力平台性能测试方法

1 目的意义

智能算力平台是一种面向人工智能应用的，提供人工智能算法模型训练与模型运行服务的计算机系统平台，通常包括 AI 计算终端、AI 服务器、AI 算力服务器集群或各个集群组成的算力网络等。随着数字经济的快速演进，算力已逐步成为等同于“水、电、煤、汽”的基础性公共资源，形成支撑经济社会发展、赋能千行百业创新的重要新质生产力。尤其在大模型应用的持续驱动下，作为人工智能产业的创新基础，算力的需求量爆炸式增长。而作为算力资源的主要支撑主体，智能算力平台面临性能测试方法不统一，评价标准不一致等问题，无法准确、全面地表征算力供给能力和算力有效使用量，导致产业界出现"算力虚标""效能黑洞"等现象。例如：学术界提出的新型异构计算架构在实际部署中出现性能缩水，云计算厂商公布的算力利用率指标缺乏可比性。这种测评体系的缺失，使得算力资源调度存在盲目性，难以继续有效引导算力产业健康发展。

近年来，国际标准化组织（International Organization for Standardization）ISO/IEC JTC1 于 2021 年发布《人工智能系统评估方法论》，系统性构建 AI 性能评估框架；中国通信标准化协会（China Communications Standards Association , CCSA）组织开展了算力网络评测技术探索。然而，由于智能算力评价标准不一致，加之缺少资金和项目支持，目前国际国内仍然没有相应的算力量值相关校准规范，造成算力量值无据可依的局面。无论在科研领域，还是在应用领域，智能算力平台的测量标准化都是一项非常有意义并亟需完善的工作。中国计量测试学会组织本文件的起草，为智能算力平台性能评价提供依据。

2 预期的社会经济效益

标准化是一项重要的基础性工作，贯穿于产品的研发和应用的全过程。标准的制定和实施不仅可以缩短研制周期、节省研制经费，还可以提高科研成果的可靠性和通用性，推动科研成果的产业化，从而产生巨大的社会经济效益。

在本标准制定和实施前，用户一般要求研制的智能算力平台必须满足指定的技术指标，验收时提交检测报告。为了满足用户要求，在产品交付前，需要耗费大量时间编制智能算力平台性能测试大纲并进行专家评审。本标准的建立为智能

算力平台性能测试和检验提供了明确的依据和规范，提升了用户对智能算力平台性能测试的认可度，可以节约测试、检验和交付时间，进而降低研制成本。另外，本标准的建立保证了智能算力平台研制单位的产品质量，有助于提高研制单位的信誉，促进研制单位与其他科研机构和企业的合作，带来经济效益。

标准的制定、颁布和不断修订的过程是随着当前技术水平而动态变化的，标准的制定反映当前参与编制单位的科研技术水平，标准的实施在很大程度上可以加速科研成果向产业界的转化。

3 工作简况

3.1 任务来源

为高质量、高水平地快速推进我国算力基础设施的建设，统一、规范的一体化算力体系是深入推进算力基础设施建设的前提，其中尤其对需要智能算力平台的性能进行准确测试和可靠溯源，目前我国尚未建立起算力量值的统一量纲和测量标准，尤其是算力平台的算力量值溯源不准确、度量方法不规范，无法实现国内市场上的算力量值统一，无法继续有效引导算力产业健康发展。2024年3月政府工作报告提出 “适度超前建设数字基础设施，加快形成全国一体化算力体系，培育算力产业生态”，因此亟需制定《智能算力平台性能测试方法》，针对现有的智能算力平台状态，准确、全面地表征服务侧算力供给能力，客观、可靠地反映用户侧对算力有效使用量，最终解决算力平台算力量值的溯源问题，进而支撑算力基础设施的快速部署与建设。

根据中国计量测试学会公布的2025年度第一批团体标准立项的通知，中国信息通信研究院、贵州省计量测试院作为《智能算力平台性能测试方法》团体标准的申请立项单位，应尽快按要求开展“智能算力平台性能测试方法”标准的制定工作。

3.2 主要工作过程

收到《智能算力平台性能测试方法》团体标准项目立项通知后，中国信息通信研究院立即联合贵州省计量测试院共同成立了《智能算力平台性能测试方法》团体标准制定工作组（见表 1），并召开了第一次工作讨论会，会议确定了标准的制定原则、制定方案、制定工作计划及人员分工。

表1　起草人员的项目分工

姓　名	职　称	项 目 分 工
陈龙泉	高级工程师	全面规划与组织实施
胡天洋	工程师	实验方案设计与验证
毛宇博	工程师	实验方案设计与验证
张大元	高级工程师	调研和标准编写
孙小强	高级工程师	标准编写与实验
沈岸平	高级工程师	技术指导
宋茂江	正高级工程师	厂家沟通与方法确认
杨霏	正高级工程师	技术指导
廖蔚松	高级工程师	调研和标准编写

第一次会议以后，标准制定工作组查阅了国内外的相关标准和技术资料，对国内外使用单位进行了充分调研，广泛征求意见，选取了具有代表性的样品进行了测试验证工作。

随后，标准制定工作组在充分调研的基础上，按照 GB/T 1.1-2020《标准化工作导则　第 1 部分：标准的结构和编写》的要求开始起草标准文本，于 2025 年 2 月形成标准草案稿。

2025年3月23日，在严格的实验验证基础上，标准制定工作组在中国信息通信研究院召开了第二次会议，即，《智能算力平台性能测试方法》团体标准草案稿预评会，根据实验结果对团体标准的草案进行了修改和完善。2023年3月10日，标准制定工作组形成了《智能算力平台性能测试方法》团体标准初稿。

2025年4月17日，中国计量测试学会在北京组织召开了《智能算力平台性能测试方法》团体标准立项会议，会议听取了标准起草组对《智能算力平台性能测试方法》团体标准的立项情况汇报，审查专家对有关问题进行了质疑。立项会审查专家一致同意此项标准通过立项，请《智能算力平台性能测试方法》团体标准起草单位参照评审专家提出的意见和建议，抓紧开展团体标准编制工作。

2025年5月-2026年3月，标准制定工作组对照专家评审意见，结合试验验证，对团体标准进行修改完善，形成了《智能算力平台性能测试方法》（征求意见稿）。

4 标准编制的原则

本标准制定的原则是保持标准的科学性和适用性，建立一套简便、准确、可靠的智能算力平台性能测试方法，使科研和产业界同行的测试数据具有可比性，促进智能算力产业的健康发展。

5 标准编制主要技术内容的说明

面向人工智能应用，智能算力平台提供人工智能算法模型训练与模型运行服务，通常包括 AI 计算终端、AI 服务器、AI 算力服务器集群或各个集群组成的算力网络等。而算力是智能算力平台对数据进行处理并实现结果输出的一种能力，是衡量平台计算能力的一个综合指标，数值越大代表综合计算能力越强。智能算力平台不同类型的计算单元输出计算能力的综合评价指标，包括 CPU、GPU、ASIC、FPGA 等芯片的输出计算能力。多年来本标准编制工作组所在单位对不同的智能算力平台性能测试方法开展了大量的研究和验证试验，在多年的工作基础上编制此团体标准。本标准从智能算力平台性能的测试目的出发，确定了智能算力平台的性能测试方法，明确了测试过程中涉及到的测试条件、试样、设备和仪器等，并对测试步骤给出详细规定。

5.1 测量原理

智能算力平台的组成如图 1 所示，通过整合硬件、软件和算法资源，高效提供大规模计算能力，以支持人工智能训练/推理、高性能计算、大数据分析等任务。一般由以下四部分组成：

（1）硬件设备

依托 CPU、GPU、TPU、FPGA、ASIC 等芯片提供基础计算，并加速人工智能模型的训练和推理的计算单元，是智能算力的基础支撑。

（2）软件框架

提供算法开发、模型训练、推理部署等全链条的支持，通过优化算法和计算流程，提高算力利用效率，常见的软件框架包括 TensorFlow，PyTorch、Cafe 等，是智能算力的运行环境。

（3）算法优化

通过模型剪枝、量化、蒸馏等技术，减少计算复杂度，提升算力利用效率，是人工智能算力平台的关键优化途径。

（4）数据存储与传输

高效的数据存储和传输机制可减少数据访问延迟和带宽占用，从而提高整体计算效率，是人工智能算力平台的主要交互接口。

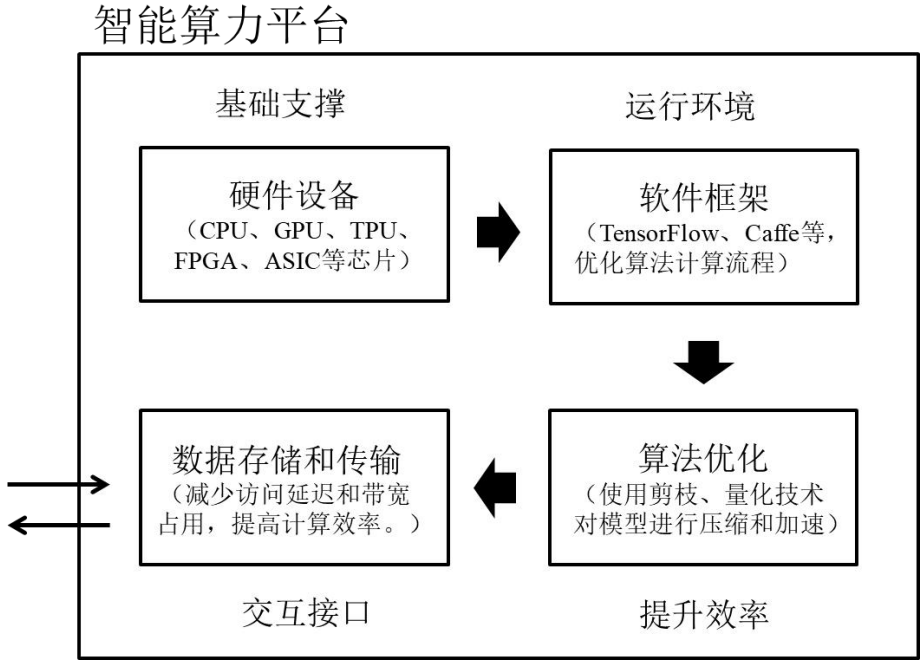

图 1 智能算力平台组成示意图

如图 1 所示为本标准采用的智能算力平台组成示意图。主要是建立统一、可控的标准算力负载环境下，对包含硬件设备、软件框架、算法优化、数据存储与传输的整体系统性能进行测试。其中算力基准测试集软件（Benchmark、CPUZ 等）是测试方法的核心，用于对被测平台的算力量值进行实时监测，该软件一般需经计量技术机构进行比对验证后方可用于基础算力的测量。另外 AI 训练框架软件（Deepspeed 等）、AI 推理框架软件（Ollama 等）用以构建标准的测试数据集用例，以应对训练或推理场景下的测试需求。

5.3 设备及仪器

5.3.1 设备

算力基准测试集软件：测量范围为（0～1.024×1018）FLOPS，最大允许误差为±1%；

AI 推理框架软件：输出范围为（0~510×1015）Tokens，最大允许误差：±1%；

存储压力测试软件：接口速率为（0~24）Gbps，最大允许误差：±1%。

5.3.2 仪器

数据网络分析仪：支持 10/100/1000 M 电口，1G/10G/40G/100G/400G/800G bps 光口等速率接口，对应数据传输速率最大允许误差为±1%，支持吞吐量、时延测试功能；

功率计：功率（0~500）W，最大允许误差：±1%。

5.4 测试程序

5.4.1 基础算力

如图 2，基于被测人工智能算力平台的操作系统，安装并适配算力基准测试集软件；运行基准测试集软件，并记录基础算力显示值。

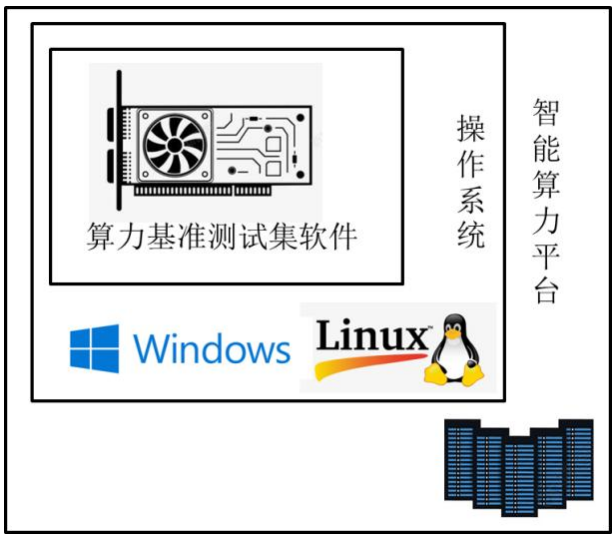

图 2 智能算力平台组成示意图

5.4.2 数据传输

如图 3，将被测人工智能算力平台的数据传输端口连接数据网络测试仪，根据被测平台的技术说明书，设置数据网络测试仪的端口速率、工作模式等，保证对应端口之间正常通信，启动吞吐量测试，测量被测平台的数据传输速率，并记录数据网络测试仪上的传输速率显示值。启动时延测试，测量被测平台的数据转发时延，并记录数据网络测试仪上的转发时延显示值。

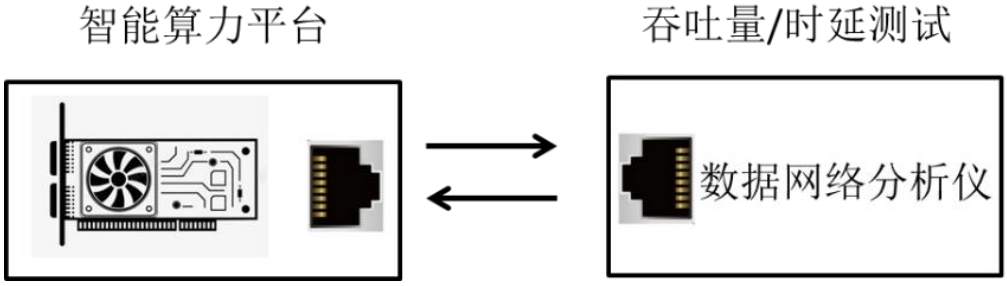

图 3 智能算力平台组成示意图

5.4.3 数据存储

如图 4，基于被测人工智能算力平台的操作系统，安装并适配存储压力测试软件；运行存储压力测试软件，并记录数据存储速率显示值。

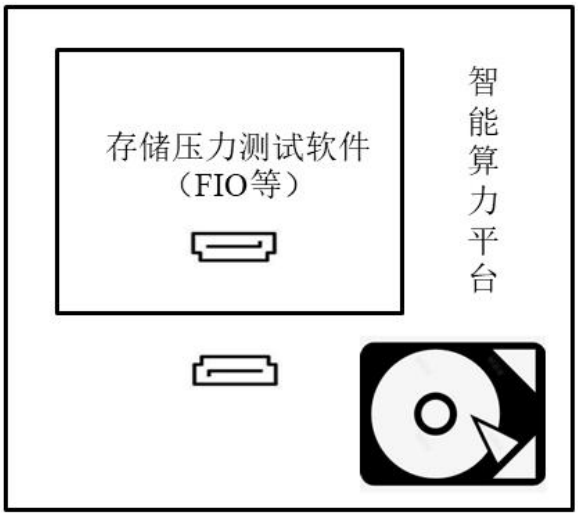

图 4 智能算力平台组成示意图

5.4.4 AI 推理场景算力

如图 5，基于被测人工智能算力平台的操作系统，部署 AI 智能模型以及推理框架；运行 AI 智能模型推理，同步启动算力基准测试集软件测试 AI 推理场景下的算力，并记录算力显示值（数据精度类型为 FP16/FP32/BF16/INT8）。

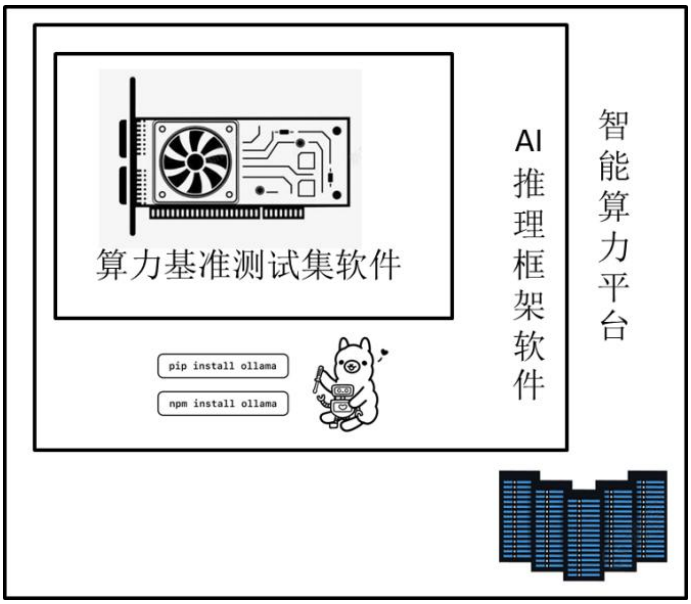

图 5 智能算力平台组成示意图

5.4.5 AI 训练场景算力

如图 6，基于被测人工智能算力平台的操作系统，部署 AI 智能模型以及训练框架；运行 AI 智能模型训练，同步启动算力基准测试集软件测试 AI 训练场景下的算力，并记录算力显示值（数据精度类型为 FP16/FP32/BF16）。

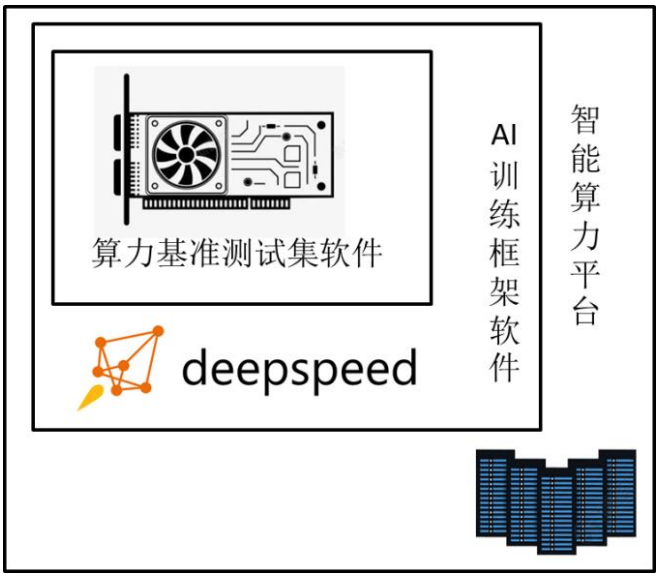

图 6 智能算力平台组成示意图

5.4.6 功耗

如图 7，在被测人工智能算力平台不同运行场景下，在电源入口处，使用功率计来测量被测平台在不同算力负载下的功耗，记录功率计显示值。

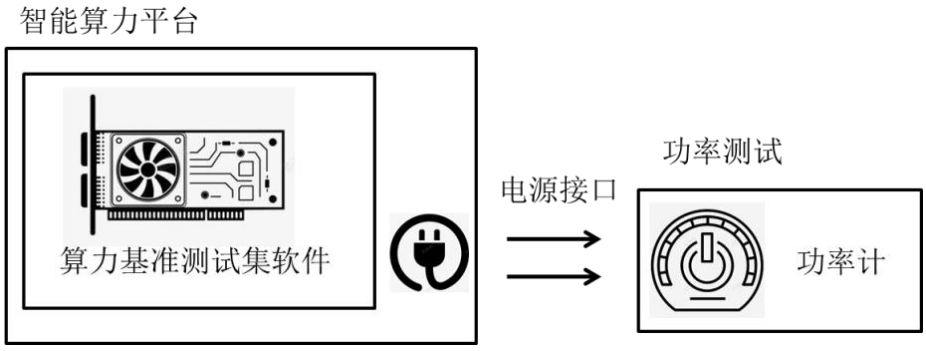

图 7 智能算力平台组成示意图

7 主要试验（或验证）结果分析

7.1 综合算力

本标准中综合算力计算涉及的“基础算力、数据存储、数据传输、AI推理场景算力、AI训练场景算力”等参数权重（如表1所示），主要依据贵州省地方标准 DB52/T 1846-2024《数据中心算力算效评估规范》中“评价指标权重”的权重框架，并进行了充分详细的行业调研，结合智能算力平台的技术特性进行适配调整，确保权重设定既符合地方成熟规范，又贴合智能算力平台的实际应用场景。

表1 参数权重列表

序号	类别	归类权重	主要指标P_i	指标权重W_i
1	基础算力	30%	有效算力比例*ECPR*	30%
2	数据传输	20%	吞吐量*Throughout*	10%
			时延*Latency*	10%
3	数据存储	20%	存力规模*Storage_size*	10%
			磁盘读写速率*IOPS*	10%
4	AI推理场景算力	15%	有效推理算力*ICE*	15%
5	AI训练场景算力	15%	有效训练算量*TCE*	15%
6	功耗	0%	功率*P*	0%

其中对单台AI计算终端、AI服务器等中小规模智能算力平台，有效算力比例根据公式（1）计算：

$$ECPR = \frac{MCP}{NCP} \cdots\cdots (1)$$

式中：

ECPR——被测智能算力平台的有效算力比例，用单精度浮点数（FP32）表示，以百分数形式表示；

NCP——被测智能算力平台的标称算力数值，单位为GFLOPS；

MCP——被测智能算力平台的测量基础算力数值，单位为GFLOPS。

存力规模、磁盘读写速率通过存储压力测试软件直接获取，根据表2分为5个等级。

表2 数据存储能力分值

指标	一级（0.2）	二级（0.4）	三级（0.6）	四级（0.8）	五级（1.0）
存力规模（TB）	[0,10)	[10,50)	[50,100)	[100,500)	[500,+∞)
磁盘读写速率（MB/s）	[0,1000)	[1000,5000)	[5000,10000)	[10000,50000)	[50000,+∞)

吞吐量、时延通过数据网络分析仪直接获取，根据表3分为5个等级。

表3 数据传输能力分值

指标	一级（0.2）	二级（0.4）	三级（0.6）	四级（0.8）	五级（1.0）
吞吐量（Gbps）	[0,1)	[1,10)	[10,40)	[40,100)	[100,+∞)
时延（us）	[1000,+∞)	[100,1000)	[10,100)	[1,10)	[0,1)

有效推理算力根据公式（2）计算：

$$\mathrm{I}CE = \frac{ICP}{NCP} \cdots\cdots (2)$$

式中：

ICE——被测智能算力平台的有效推理算力，以百分数形式表示；

NCP——被测智能算力平台的标称算力数值，单位为GFLOPS；

ICP——被测智能算力平台AI推理场景下的算力数值，单位为GFLOPS。

有效训练算力根据公式（3）计算：

$$\mathrm{T}CE = \frac{TCP}{NCP} \cdots\cdots (3)$$

式中：

TCE——被测智能算力平台的有效训练算力，以百分数形式表示；

NCP——被测智能算力平台的标称算力数值，单位为GFLOPS；

TCP——被测智能算力平台AI训练场景下的算力数值，单位为GFLOPS。

智能算力平台的综合算力效率根据公式（4）计算：

$$\mathrm{CCE} = \sum_i P_{\mathrm{i}} \times W_{\mathrm{i}} \cdots\cdots (4)$$

式中：

P_{i}——智能算力平台综合算力参考的各参数测试分值，见表2；

W_{i}——智能算力平台综合算力参考的各参数权重，见表2；

CCE——被测智能算力平台的综合算力效率，以百分数形式表示。

智能算力平台的综合算力根据公式（5）计算：

$$\mathrm{CCP} = CCE \times NCP \cdots\cdots (5)$$

式中：

CCE——被测智能算力平台的综合算力效率，以百分数形式表示；

NCP——被测智能算力平台的标称算力数值，单位为GFLOPS；

CCP——被测智能算力平台的综合算力，单位为GFLOPS。

对AI算力服务器集群或各个集群组成的算力网络等大规模智能算力平台，其有效算力比例根据公式（6）计算：

$$\mathrm{CCP} = \sum a_{\mathrm{i}}\, CPU_{\mathrm{i}} + \sum b_{\mathrm{i}}\, GPU_{\mathrm{i}} + \sum c_{\mathrm{i}}\, FPGA_{\mathrm{i}} + \sum d_{\mathrm{i}}\, ASIC_{\mathrm{i}} \cdots\cdots (6)$$

式中：

a_{i}——某型号CPU服务器数量；

b_{i}——某型号GPU服务器数量；

c_{i}——某型号FPGA服务器数量；

d_{i}——某型号ASIC服务器数量；

CPU_{i}——大规模某智能算力平台中某型号单台CPU服务器的综合算力；

GPU_{i}——大规模某智能算力平台中某型号单台GPU服务器的综合算力；

$FPGA_{\mathrm{i}}$——大规模某智能算力平台中某型号单台FPGA服务器的综合算力；

$ASIC_{\mathrm{i}}$——大规模某智能算力平台中某型号单台ASIC服务器的综合算力；

CCP——被测服务器集群或算力网络的综合算力。

7.2 算力能效

被测智能算力平台的算力能效根据公式（7）计算：

$$CPE = \frac{CCP}{P} \cdots\cdots (7)$$

式中：

CPE——被测智能算力平台的算力能效，单位为GFLOPS/W；

CCP——被测智能算力平台的综合算力数值，单位为GFLOPS；

P——被测智能算力平台的整体功耗，可在电源入口处测量获得，单位为W。

通过测试数据的整理分析表明，本标准中所确定的各项参数要求和测试方法是合理的，标准所确定的指标是适宜的，作为团体标准的依据也是可行的。

8 采用国际标准和国外先进标准情况、与国际、国外同类标准水平的对比情况

目前国内外尚无相关标准。

9与有关的现行法律、法规和强制性标准的关系

本标准在制定的过程中，已充分的查阅了相关的法律、法规及相关标准，完全符合现行相关法律、法规、强制标准的规定。

10 重大分歧意见的处理经过和依据

无。

11 贯彻标准的要求和措施建议（包括组织措施、技术措施、过度办法、实施日期等）

本标准致力于建立一套简洁、有效、准确的智能算力平台性能性能测试方法，规范智能算力平台性能的表征工作。在本标准的贯彻过程中，将依赖行业内专家的大力支持和广泛地推荐，同时也需要各同行的大力配合推广使用。最后在标准的贯彻过程中，还将广泛在行业内和用户中宣传，紧密地同用户沟通交流，保证本标准的适用性、实用性。

12 废止现行相关标准的建议：

无。

13 其他应予说明的事项：

无。

附件3

《智能算力平台性能测试方法》（征求意见稿）

征求意见反馈表

序号	条款编号	条款内容简述	修改意见和建议	原因或理由

反馈人：　　单位名称：

联系电话：　　联系邮箱：

团体标准意见反馈人：陈龙泉，电话：15001327806，邮箱：chenlongquan@caict.ac.cn