

T/ISC

团 体 标 准

T/ISC XXXX—XXXX

异构算力度量与当量折算规范

Specification for heterogeneous computing power measurement and
equivalent conversion

（征求意见稿）

XXXX – XX – XX 发布

XXXX – XX – XX 实施

中国互联网协会 发 布

目 次

前 言 II

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 算力度量 2

6 算力当量折算与折算系数确定 5

7 度量与折算流程 8

8 报告与记录 10

附 录 A （资料性） 各类算力基准折算系数 12

附 录 B （资料性） 精度类型折算系数 13

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。

请注意本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。

本文件由中国互联网协会归口。

本文件起草单位：

本文件主要起草人：

异构算力度量与算力当量折算规范

1 范围

本文件规定了异构算力的算力度量、算力当量折算与折算系数确定、度量与折算流程、报告与记录。本文件适用于算力网络中CPU、GPU、FPGA、ASIC等异构算力资源的度量与当量折算，具体适用场景包括但不限于：

- a) 算力资源性能评估与对标比较；
- b) 算力并网与跨域算力资源池化管理；
- c) 算力调度与资源优化配置；
- d) 算力基础设施建设验收与绩效考核；

本文件不适用于量子计算、光计算等非经典计算架构的算力度量与折算。

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 46346 人工智能计算中心计算能力评估
- GB/T 47514 分布式计算 算力度量方法
- GB/T 47518 分布式计算 算力资料描述
- YD/T 6044 算力网络 算力度量与算力建模技术要求
- YD/T 6048 数据中心算力技术要求和测评方法
- YD/T 6693 算力网络 异构算力资源计算能力度量指标

3 术语和定义

下列术语和定义适用于本文件。

3.1

异构算力 heterogeneous computing power

由两种或以上不同类型计算单元提供的计算能力。

3.2

算力当量 computing power equivalent

将不同架构、不同精度下的算力，通过统一的折算方法换算为基准算力后的等效数值。用于实现异构算力之间的可比较、可计量。

3.3

算力当量值 computing power equivalent value; CEV

某一算力资源经过精度折算和架构修正后，以标准算力单位（CU）表示的等效算力数值。

3.4

标准算力单位 computing unit; CU

在标准测试条件下，基准架构以FP64执行标准基准测试程序所达到的实测算力值。

3.5

峰值算力 peak computing power

算力单元在理想条件下，依据其硬件规格计算得出的理论最大运算能力，通常以FLOPS表示。

3.6

实测算力 measured computing power

算力单元在规定的标准测试条件下，通过执行标准化基准测试程序实际测得的运算能力，通常以FLOPS或OPS表示。

3.7

算力利用率 `computing power utilization`

实测算力与峰值算力的比值，反映算力资源在实际运行中的有效利用程度，以百分比表示。

3.8

精度折算系数 `precision conversion factor`

用于将不同计算精度下的算力值归一化到基准精度的换算系数。

3.9

架构修正系数 `architecture correction factor`

用于修正不同计算架构因体系结构差异导致的实际计算效率不同的修正系数。

3.10

通用算力 `general-purpose computing power`

由以CPU为主的通用处理器提供的、适用于广泛通用计算任务的计算能力。

3.11

智能算力 `intelligent computing power`

由面向人工智能计算优化的处理器提供的、主要适用于深度学习训练与推理等AI任务的计算能力。

3.12

超算算力 `high-performance computing power`

由高性能计算集群提供的、主要适用于大规模科学计算、工程仿真等HPC任务的计算能力。

4 缩略语

以下缩略语适用于本文件。

ASIC: 专用集成电路 (Application Specific Integrated Circuit)

BF16: 16位脑浮点 (Brain Float 16)

CPU: 中央处理器 (Central Processing Unit)

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

FPGA: 现场可编程门阵列 (Field Programmable Gate Array)

FP8: 8位浮点 (Floating Point 8)

FP16: 16位半精度浮点 (Floating Point 16)

FP32: 32位单精度浮点 (Floating Point 32)

FP64: 64位双精度浮点 (Floating Point 64)

GPU: 图形处理器 (Graphics Processing Unit)

HPC: 高性能计算 (High Performance Computing)

HPCG: 高性能共轭梯度基准程序 (High Performance Conjugate Gradients)

HPL: 高性能线性包基基准程序 (High Performance Linpack)

I/O: 输入输出 (Input/Output)

INT4: 4位整数 (Integer 4)

INT8: 8位整数 (Integer 8)

MLPerf Inference: 机器学习推理性能基准套件 (Machine Learning Performance Benchmark Inference)

NPU: 神经网络处理器 (Neural Processing Unit)

OPS: 每秒操作次数 (Operations Per Second)

RISC-V: 第五代精简指令集架构 (Reduced Instruction Set Computer-V)

SPECint: 整数运算能力 (Standard Performance Evaluation Corporation Integer)

TF32: 张量浮点 (Tensor Float 32)

TPU: 张量处理器 (Tensor Processing Unit)

5 算力度量

5.1 度量基准

5.1.1 本文件以 FLOPS 作为统一算力度量的基础基准，确保不同架构算力可在同一维度上进行量化比较。

5.1.2 对于以整数运算为主的算力资源，以 OPS 作为辅助度量单位，并通过整数-浮点换算系数归一化到 FLOPS 基准。

5.1.3 度量基准应满足以下要求：

- 唯一性：同一算力资源在相同测试条件下，度量结果应唯一确定；
- 可复现性：不同测试机构在相同条件下对同一算力资源进行度量，结果偏差应不超过±5%；
- 可追溯性：度量结果应可追溯至基准测试程序和标准测试条件。

5.1.4 各类异构算力资源的度量按照表 1 进行。涉及分布式部署的算力资源，其分布式度量参数和度量方法按照 GB/T 47514 执行。算力资源的形式化描述（计算单元、通信单元、存储单元）按照 GB/T 47518 执行。

表 1 各类算力资源的度量依据

算力类型	主要度量内容	度量依据
通用算力	浮点运算能力、整数运算能力、内存带宽等	YD/T 6693、YD/T 6048
智能算力	混合精度训练性能、推理吞吐量与延迟、矩阵运算能力等	YD/T 6693、GB/T 46346
超算算力	集群FP64性能、通信带宽、并行扩展效率等	YD/T 6044、YD/T 6048

5.2 度量维度

本文件将异构算力划分为以下三个度量维度，与YD/T 6048的分类保持一致，全面反映异构算力资源的综合处理能力。

- 通用算力维度：
 - 适用对象：以 CPU 为主要计算单元的系统；
 - 典型应用：操作系统运行、数据库处理、Web 服务、通用科学计算等；
 - 度量重点：单核/多核浮点运算能力、整数运算能力、内存带宽、I/O 吞吐能力。
- 智能算力维度：
 - 适用对象：以 GPU、NPU、TPU、ASIC 等 AI 加速芯片为主要计算单元的系统；
 - 典型应用：深度学习模型训练、推理服务、计算机视觉、自然语言处理等；
 - 度量重点：矩阵运算能力、显存带宽、卡间互联带宽、混合精度训练性能、推理吞吐量与延迟。
- 超算算力维度：
 - 适用对象：由大规模计算节点集群构成的高性能计算系统；
 - 典型应用：气候模拟、分子动力学、流体力学仿真、基因组分析等；
 - 度量重点：集群整体 FP64 浮点运算能力、通信带宽、并行扩展效率。

5.3 度量指标

5.3.1 算力度量的核心指标体系见表 2。

表 2 算力度量指标体系

指标名称	符号	单位	指标说明
峰值算力	P _{peak}	TFLOPS/TOPS	硬件理论最大运算能力
实测算力	P _{measured}	TFLOPS/TOPS	标准测试条件下实测运算能力
算力利用率	η	%	$\eta = P_{\text{measured}} / P_{\text{peak}} \times 100\%$
能效比	EER	TFLOPS/W	单位功耗下的实测算力
内存带宽	BW _{mem}	GB/s	处理器与内存之间的数据传输速率
互联带宽	BW _{inter}	GB/s	计算节点之间的通信带宽
显存带宽		GB/s	计算单元与显存的数据传输速率
卡间互联带宽		GB/s	计算单元间数据传输的速率

并行扩展效率	E_scale	%	多节点/多卡并行时的线性加速比
推理吞吐量	T_infer	samples/s	单位时间内完成的推理样本数
推理延迟	L_infer	ms	单样本推理的平均响应时间
首Token延迟	TTFT	ms	从请求发出到收到首个输出Token的时间
平均Token时间	TPOT	ms/token	输出阶段平均每个输出Token的耗时
最大上下文长度	Max Context Length	tokens	单次推理支持的最大输入与输出Token总量
推理并发数	Concurrency	个	系统同时处理的推理请求数

5.3.2 峰值算力按照公式（1）计算：

$$P_{peak} = N_{cores} \times f_{clock} \times OPC \times N_{units} \quad (1)$$

式中：

P_{peak} ——计算核心数；

f_{clock} ——核心工作频率（GHz）；

OPC ——每时钟周期运算次数；

N_{units} ——每核心运算单元数。

5.3.3 能效比按照公式（2）计算：

$$EER = P_{measured}/P_{power} \quad (2)$$

式中：

P_{power} ——算力单元在满载运行时的实际功耗（单位：W）。

5.4 度量方法

5.4.1 浮点运算能力测试包括但不限于以下内容。

a) 通用算力浮点测试：

- 1) 采用 HPL 基准测试程序，测试 FP64 运算能力；
- 2) 采用 HPCG 基准测试程序，测试内存受限条件下的浮点运算能力；
- 3) 测试规模应覆盖单节点和多节点配置，多节点测试应不少于 4 个节点。

b) 智能算力浮点测试：

- 1) 采用 MLPerf Training 基准测试程序，测试混合精度（FP16/BF16）训练性能；
- 2) 采用 MLPerf Inference 基准测试程序，测试推理性能；
- 3) 对于支持 FP64 的 GPU，应同时采用 HPL 进行 FP64 浮点测试；
- 4) 测试应覆盖至少两种典型 AI 模型。

c) 超算算力浮点测试：

- 1) 采用 HPL 基准测试程序测试集群整体 FP64 性能；
- 2) 采用 HPCG 基准测试程序测试实际应用性能；
- 3) 测试应使用集群全部计算节点或不少于总节点数 80% 的节点。

5.4.2 整数运算能力测试包括但不限于以下内容：

- a) 采用 SPECint 基准测试程序测试 CPU 整数运算能力；
- b) 对于 AI 推理芯片，采用 INT8 精度的 MLPerf Inference 基准测试程序测试整数推理性能；
- c) 整数运算能力应通过整数-浮点换算系数（见附录 B）归一化到 FLOPS 基准。

5.4.3 标准测试条件包括但不限于以下内容。

a) 环境条件：

- 1) 测试环境温度：23 °C ± 5 °C；
- 2) 相对湿度：30%～70%；
- 3) 供电稳定，电压波动不超过 ±5%；
- 4) 测试期间不应有其他负载干扰。

b) 软件条件：

- 1) 操作系统版本、内核版本应明确记录；

- 2) 编译器版本及优化选项应明确记录;
 - 3) 驱动程序版本、固件版本应明确记录;
 - 4) 基准测试程序版本应明确记录;
 - 5) 所有测试应使用相同的软件配置。
 - c) 硬件条件:
 - 1) 处理器应运行在标称频率(非超频状态);
 - 2) 内存配置应达到厂商推荐的最低配置;
 - 3) 散热系统应正常工作,测试期间不应出现热降频;
 - 4) 多节点测试时,网络互联配置应明确记录。
 - d) 测试执行:
 - 1) 每项测试应至少执行 5 次,取中位数作为最终结果;
 - 2) 每次测试之间应间隔不少于 60 s 确保系统恢复稳态;
 - 3) 测试前应进行不少于 5 min 的预热运行;
 - 4) 测试过程中应监控系统资源占用,确保无异常进程干扰。
- 5.4.4 度量结果有效性判定应满足以下要求:
- a) 同一算力资源 3 次测试结果的最大偏差不应超过 $\pm 3\%$;
 - b) 若偏差超过 $\pm 3\%$,应排查原因并重新测试;
 - c) 测试结果应附带完整的测试环境信息和原始数据。

6 算力当量折算与折算系数确定

6.1 折算基准

- 6.1.1 以通用 CPU(x86 架构)在 FP64 精度下的实测算力作为 1 CU 的基准。
- 6.1.2 基准架构的选取应满足以下条件:
- a) 市场主流通用服务器 CPU 架构;
 - b) 具备完整的 FP64 浮点运算单元;
 - c) 在科学计算、通用计算领域具有广泛代表性;
 - d) 基准架构的具体型号应在附录 A 中明确列出,并定期更新。
- 6.1.3 基准测试程序为 HPL,测试精度为 FP64,测试结果以 TFLOPS 为单位。

6.2 折算原则

算力当量折算应遵循以下原则:

- a) 客观性:折算系数基于实测数据确定,不依赖厂商标称值或理论推算值;
- b) 可复现性:任何机构在相同条件下,依据本文件规定的折算方法,可得到一致的折算结果;
- c) 保守性:当实测数据不足时,折算系数取保守值,避免高估算力当量;
- d) 透明性:折算过程中使用的所有参数、系数和计算方法记录完整,可追溯、可审计;
- e) 动态更新:折算系数随硬件技术进步和应用场景变化定期评估和更新。

6.3 精度折算

- 6.3.1 不同计算精度下的算力值应通过精度折算系数归一化到基准精度(FP64)。
- 6.3.2 精度折算的基本逻辑:
- a) 低精度运算在相同硬件上的吞吐量通常高于高精度运算,但每次运算的有效信息量较低;
 - b) 精度折算系数反映了不同精度运算在计算复杂度和应用价值上的差异;
 - c) 精度折算系数应基于典型应用场景的实测性能对比确定。
- 6.3.3 精度折算的适用规则:
- a) 当算力资源仅支持单一精度时,直接使用该精度对应的折算系数;
 - b) 当算力资源支持混合精度运算时,应按各精度运算的实际占比进行加权折算;
 - c) 混合精度加权按照公式(3)计算:

$$K_{precision_mixed} = \sum(w_i \times K_{precision_i}) \dots\dots\dots (3)$$

式中：

w_i ——第*i*种精度运算的占比， $\sum w_i=1$ ；

$K_{precision}$ ——对应精度的折算系数。

6.4 架构修正

6.4.1 不同计算架构因体系结构差异（指令级并行度、内存访问模式、通信开销、专用化程度等），在相同任务负载下的实际效能存在显著差异，应通过架构修正系数予以修正。

6.4.2 架构修正系数反映的是该架构相对于基准架构（通用 CPU x86）的综合等效算力。

6.5 算力当量值计算

6.5.1 单一算力资源的算力当量值（CEV）按照公式（4）计算：

$$CEV = P_{measured} \times K_{precision} \times K_{architecture} \quad (4)$$

式中：

CEV ——算力当量值，单位为CU；

$P_{measured}$ ——实测算力值，单位为TFLOPS或经换算后的等效TFLOPS；

$K_{precision}$ ——精度折算系数（见附录B）；

$K_{architecture}$ ——架构修正系数（见附录A）。

6.5.2 算力集群的总算力当量值按照公式（5）计算：

$$CEV_{total} = \sum (CEV_i \times N_i \times \eta_i) \quad (5)$$

式中：

CEV_i ——第*i*类算力资源的单节点/单卡算力当量值；

N_i ——第*i*类算力资源的数量；

η_i ——第*i*类算力资源在实际部署条件下的算力利用率。

6.5.3 跨精度、跨架构混合算力系统的总算力当量值按照公式（6）计算：

$$CEV_{system} = \sum_j [\sum_i (P_{measured_{ij}} \times K_{precision_i} \times K_{architecture_j})] \quad (6)$$

式中：

i ——精度索引；

j ——架构索引。

6.5.4 当算力利用率无法实测时，可参照以下默认值：

表3 算力利用率默认值

部署方式	默认算力利用率
单机/单卡独立运行	70%~85%
集群并行运行（≤16节点）	60%~80%
大规模集群并行运行（>16节点）	50%~70%
云化/虚拟化部署	50%~70%

6.6 折算结果表达

6.6.1 算力当量值应保留小数点后两位有效数字。

6.6.2 折算结果应同时标注：

- 算力当量值；
- 对应的实测算力值；
- 所使用的折算系数组合。

6.6.3 折算结果应附加不确定度说明，不确定度一般不超过±10%。

6.7 折算系数构成

6.7.1 折算系数体系由以下两类系数构成：

- a) 精度折算系数：用于将不同计算精度下的算力归一化到基准精度（FP64）；
- b) 架构修正系数：用于修正不同计算架构因体系结构差异导致的实际计算效率差异。

6.7.2 最终折算系数为精度折算系数、架构修正系数的乘积。

6.8 精度折算系数确定

6.8.1 精度折算系数的确定依据：

- a) 不同精度运算在相同硬件上的实际吞吐量比值，度量数据应依据 YD/T 6693 获取；
- b) 不同精度运算在典型应用中的有效计算贡献度；
- c) 不同精度运算的能耗差异；
- d) 行业共识与专家评议结果。

6.8.2 精度折算系数的确定方法：

- a) 选取同一硬件平台，依据 YD/T 6693 规定的度量方法，分别以 FP64、FP32、FP16、BF16、INT8 等精度执行标准基准测试；
- b) 记录各精度下的实测算力值；
- c) 以 FP64 实测算力为基准，计算各精度的原始吞吐量比值；
- d) 结合应用价值权重进行修正，得到最终精度折算系数；
- e) 精度折算系数应经不少于 3 家独立机构验证。

6.8.3 精度折算系数的参考范围见附录 B。

6.9 架构修正系数确定

6.9.1 架构修正系数的确定依据：

- a) 不同架构在相同任务负载下的实际完成时间对比；
- b) 不同架构的指令级并行度、内存访问效率、通信开销等体系结构特征；
- c) 不同架构在典型应用场景中的实际效能表现；
- d) 架构的通用性与专用性差异。

6.9.2 架构修正系数的确定方法：

- a) 选取具有代表性的基准任务集，覆盖科学计算、AI 训练、AI 推理、数据处理等典型场景；
- b) 依据 YD/T 6693 和 YD/T 6048 规定的度量方法，在基准架构和待评估架构上分别执行相同任务集；
- c) 记录各架构的任务完成时间、资源占用、能耗等数据；
- d) 以基准架构的效能为 1.0，计算各架构的相对效能比；
- e) 对多个基准任务的结果进行加权平均，得到综合架构修正系数；
- f) 架构修正系数应经不少于 3 家独立机构验证。

6.9.3 架构修正系数的参考范围见附录 A。

6.10 折算系数测试方法

6.10.1 测试选取原则如下。

- a) 基准测试程序选取原则：
 - 1) 选取 YD/T 6693、YD/T 6048、GB/T 47514 等标准规定的标准基准测试程序；
 - 2) 覆盖科学计算、AI 训练、AI 推理、通用计算等主要场景；
 - 3) 基准测试程序版本统一，且为发布时间不超过 2 年的稳定版本。
- b) 测试硬件选取原则：
 - 1) 选取市场主流的、具有代表性的硬件产品；
 - 2) 每种架构类型至少选取 3 款不同厂商或不同代际的产品；
 - 3) 硬件配置为厂商推荐的标准配置。
- c) 测试规模选取原则：
 - 1) 单节点/单卡测试为基本要求；
 - 2) 对于集群级算力，补充多节点/多卡测试；
 - 3) 多节点测试规模不少于 4 个节点或 4 张加速卡。

6.10.2 系数更新要求如下。

- a) 折算系数的常规更新周期为 2 年。
- b) 当出现以下情况时，应启动临时更新：
 - 1) 出现全新计算架构（如新型 AI 芯片架构）；
 - 2) 主流硬件产品发生代际更替，性能提升超过 50%；
 - 3) 出现新的主流应用场景，现有系数无法合理覆盖；
 - 4) 行业反馈现有系数导致折算结果偏差超过 20%。
- c) 系数更新流程：
 - 1) 由标准归口单位发起系数评估；
 - 2) 组织不少于 5 家机构进行联合测试；
 - 3) 汇总测试数据，计算新系数；
 - 4) 公开征求意见，征求意见期不少于 30 天；
 - 5) 专家评审通过后发布更新版本。

6.10.3 新确定的折算系数应通过以下方式验证：

- a) 交叉验证：由独立机构使用相同方法复现测试，结果偏差应不超过 $\pm 10\%$ ；
- b) 应用验证：将折算结果应用于实际算力调度场景，验证其合理性；
- c) 专家评议：组织行业专家对系数进行评议，确认其技术合理性。

注：验证不通过的系数应退回修改，重新测试确定。

6.11 折算系数表

6.11.1 本文件提供以下折算系数参考表：

- a) 附录 A：各类算力基准折算系数表（架构修正系数）；
- b) 附录 B：精度类型折算系数表（精度折算系数）。

6.11.2 折算系数表为资料性附录，提供推荐值供参考使用。具体项目中可根据实际测试数据进行调整，但调整幅度超过 $\pm 20\%$ 时应说明理由。

6.11.3 折算系数表中的数值为无量纲系数，基准值为 1.0。

7 度量与折算流程

7.1 流程总则

度量与折算流程应遵循“识别→测试→折算→审核→出具”的五阶段流程，确保全过程可追溯、可审计。

7.2 第一阶段：算力资源识别

按照以下步骤进行。

- a) 识别待度量的算力资源基本信息：
 - 1) 算力资源类型；
 - 2) 具体型号、厂商、代际；
 - 3) 数量及部署方式；
 - 4) 硬件配置详情；
 - 5) 软件环境。
- b) 确定度量目的和适用场景，确定所需的度量维度和指标。
- c) 确定度量依据：
 - 1) 根据算力资源类型，从表 1 中选取适用的度量依据标准；
 - 2) 制定度量方案或确认已有度量报告的适用性。

7.3 第二阶段：算力度量

按照以下步骤进行。

- a) 测试环境准备：
 - 1) 按照 5.4.3 的要求检查和确认测试环境条件；

- 2) 安装并配置基准测试程序;
- 3) 进行系统预热和稳定性验证。
- b) 测试执行:
 - 1) 按照 5.4.1 和 5.4.2 的方法执行基准测试;
 - 2) 每项测试至少执行 5 次;
 - 3) 记录每次测试的原始数据和系统状态信息;
 - 4) 监控测试过程中的异常情况 (如热降频、内存溢出等)。
- c) 数据整理:
 - 1) 按照 5.4.4 对 3 次测试结果进行有效性判定;
 - 2) 取中位数作为最终实测算力值;
 - 3) 整理测试环境信息和原始数据。
- d) 已有度量报告的采用:
 - 1) 若算力资源已有符合要求的度量报告 (12 个月有效期内), 可直接采用;
 - 2) 采用已有报告时, 应核实其度量依据、测试条件、数据完整性是否满足本文件要求;
 - 3) 不满足要求的, 应补充测试或重新测试。

7.3 第三阶段: 算力当量折算

按照以下步骤进行。

- a) 系数选取:
 - 1) 根据算力资源的精度类型, 从附录 B 中选取精度折算系数;
 - 2) 根据算力资源的架构类型, 从附录 A 中选取架构修正系数;
 - 3) 若使用自定义系数, 应提供确定依据和验证数据。
- b) 当量计算:
 - 1) 按照 6.6.1 的公式计算单一资源的算力当量值;
 - 2) 对于集群系统, 按照 6.6.2 的公式计算总算力当量值;
 - 3) 对于混合系统, 按照 6.6.3 的公式计算系统总算力当量值。
- c) 结果校验:
 - 1) 检查计算过程的正确性;
 - 2) 验证折算结果的合理性;
 - 3) 评估折算结果的不确定度。

7.4 第四阶段: 结果审核

按照以下步骤进行。

- a) 技术审核:
 - 1) 按照第 5 章审核测试方法的合规性;
 - 2) 按照 5.5.3 审核测试条件的达标性;
 - 3) 审核折算系数选取的合理性;
 - 4) 审核计算过程的正确性。
- b) 质量审核:
 - 1) 审核原始数据的完整性和真实性;
 - 2) 审核测试记录的规范性;
 - 3) 审核报告内容的完整性和准确性。
- c) 审核结论:
 - 1) 审核通过: 进入报告出具阶段;
 - 2) 审核不通过: 退回修改, 说明不通过原因及整改要求;
 - 3) 存在争议: 组织专家复议。

7.5 第五阶段: 报告出具

按照以下步骤进行。

- a) 按照第 8 章要求编制算力度量与折算报告;

- b) 报告应由测试负责人和审核人签字确认；
- c) 报告应加盖测试机构公章（如适用）；
- d) 报告应分配唯一编号，纳入档案管理。

7.6 第三方测试机构要求

折算系数的确定及算力当量折算结果的出具，可选择由第三方测试机构执行。第三方测试机构应满足以下要求。

- a) 机构资质要求：
 - 1) 应为依法设立的独立法人实体，具备独立承担法律责任的能力；
 - 2) 应具备 CNAS（中国合格评定国家认可委员会）实验室认可资质，或具备省级及以上计量认证（CMA）资质；
 - 3) 应具备与算力度量相关的检测能力范围，且认可/认证证书在有效期内；
 - 4) 应建立并运行符合 GB/T 27025（ISO/IEC 17025）要求的管理体系。
- b) 技术能力要求：
 - 1) 应具备本文件第 5 章规定的各类算力度量所需的测试环境和设备条件；
 - 2) 应配备熟悉 HPL、HPCG、MLPerf、SPECint 等基准测试程序的专业技术人员，且相关人员应具备不少于 2 年的算力测试经验；
 - 3) 应具备对测试数据进行统计分析和不确定度评估的能力；
 - 4) 应具备独立的测试环境，测试环境应满足 5.4.3 规定的标准测试条件。
- c) 独立性要求：
 - 1) 第三方测试机构应与被测算力资源的提供方、使用方不存在隶属关系或利益关联；
 - 2) 测试机构的人员不应参与被测算力资源的设计、开发、生产或运维工作；
 - 3) 测试机构应建立利益冲突回避机制，并在报告中声明独立性。
- d) 报告出具要求：
 - 1) 第三方测试机构出具的算力度量与折算报告应加盖机构公章和 CMA/CNAS 标识（如适用）；
 - 2) 报告应由测试负责人、审核人和批准人三级签字；
 - 3) 报告应包含完整的测试方法、测试条件、原始数据、折算过程和最终结果；
 - 4) 报告应分配唯一编号，并纳入机构档案管理系统；
 - 5) 报告有效期为自出具之日起 12 个月。
- e) 能力验证与监督：
 - 1) 第三方测试机构应每年参加不少于 1 次由标准归口单位或行业主管部门组织的能力验证活动；
 - 2) 能力验证结果不满意的机构，应暂停出具相关报告，整改完成并通过复核后方可恢复；
 - 3) 标准归口单位有权对第三方测试机构进行不定期监督检查。

8 报告与记录

8.1 报告内容

算力度量与折算报告应至少包含以下内容：

- a) 基本信息；
- b) 被测算力资源信息；
- c) 测试方法与条件；
- d) 测试结果；
- e) 折算过程与结果；
- f) 结论与建议；
- g) 签署与附件。

8.2 记录保存

8.2.1 所有与算力度量和折算相关的资料均应纳入保存范围，包括但不限于：

- a) 测试方案及审批记录;
 - b) 原始测试数据及日志文件;
 - c) 测试环境配置记录;
 - d) 折算计算过程记录;
 - e) 审核记录及意见;
 - f) 最终报告及签发记录;
 - g) 系数选取依据及验证数据。
- 8.2.2 保存应满足以下要求:
- a) 电子记录应采用不可篡改的格式存储(如 PDF/A 格式),并附加数字签名或哈希校验;
 - b) 纸质记录应存放在防火、防潮、防虫的专用档案柜中;
 - c) 电子记录应定期进行备份,备份介质应异地存放;
 - d) 记录的访问应设置权限控制,确保数据安全。
- 8.2.3 保存期限应满足以下要求:
- a) 用于算力基建验收、绩效考核的记录,保存期限不少于 5 年;
 - b) 其他一般性度量记录,保存期限不少于 3 年。
- 8.2.4 记录销毁应满足以下要求:
- a) 超过保存期限的记录,经审批后方可销毁;
 - b) 涉及商业秘密的记录,销毁时应采用不可恢复的方式;
 - c) 销毁过程应有记录,包括销毁时间、方式、经办人等信息。

附录 A
(资料性)
各类算力基准折算系数

A.1 架构修正系数推荐值

各类算力架构修正系数推荐值见表A.1。

表 A.1 各类算力架构修正系数推荐值

算力类型	典型架构	架构修正系数	适用场景	备注
通用算力	CPU（x86架构）	1.00	通用计算（基准）	基准架构
通用算力	CPU（ARM架构）	0.85~1.05	通用计算	视具体型号而定
通用算力	CPU（RISC-V架构）	0.70~0.95	通用计算	视具体型号而定
智能算力	GPU（通用计算型）	1.20~2.50	AI训练/HPC	基于FP64效能对比
智能算力	GPU（AI加速型）	0.80~1.80	AI训练/推理	基于混合精度效能对比
智能算力	NPU/TPU	0.60~1.50	AI推理	基于INT8/FP16推理效能对比
智能算力	ASIC（AI专用）	0.50~1.30	AI推理	视专用化程度而定
专用算力	FPGA	0.40~1.20	特定算法加速	视算法匹配度而定
超算算力	超算集群（CPU+GPU）	1.50~3.00	HPC	基于集群整体效能

A.2 使用说明

使用说明如下：

- a) 表中系数为推荐范围值，具体项目应基于实测数据确定精确值；
- b) 同一架构不同代际产品的系数可能存在显著差异，应分别确定；
- c) 当算力资源用于非典型场景时，系数可能需要调整；
- d) 系数的确定应遵循第7章规定的方法和流程。

附录 B
(资料性)
精度类型折算系数

B.1 精度折算系数

不同计算精度的折算系数推荐值见表B.1。

表 B.1 精度类型折算系数推荐值

精度类型	数据类型	位宽	精度折算系数	典型应用场景	备注
双精度浮点	FP64	64-bit	1.00	科学计算、HPC	基准精度
单精度浮点	FP32	32-bit	0.50~0.65	通用计算、图形渲染	—
TF32	TF32	19-bit	0.35~0.50	AI训练 (NVIDIA)	仅特定硬件支持
半精度浮点	FP16	16-bit	0.25~0.40	AI训练/推理	—
脑浮点	BF16	16-bit	0.25~0.40	AI训练	动态范围同FP32
8位浮点	FP8	8-bit	0.12~0.25	AI推理	新一代AI芯片
8位整数	INT8	8-bit	0.10~0.20	AI推理	—
4位整数	INT4	4-bit	0.05~0.12	量化推理	大模型量化部署

B.2 混合精度折算

当算力资源采用混合精度运算时，加权精度折算系数计算示例：

示例：某GPU在AI训练任务中，FP16运算占比70%，FP32运算占比30%，则：

$$K_{precision_mixed} = 0.70 \times 0.30 + 0.30 \times 0.55 = 0.21 + 0.165 = 0.375 \cdots \cdots (B.1)$$

注：此处取FP16系数为0.30，FP32系数为0.55，为示例值。

B.3 使用说明

使用说明如下：

- a) 表中系数为推荐范围值，具体值应基于实测数据确定；
- b) 精度折算系数与具体硬件实现相关，同一精度在不同硬件上的系数可能不同；
- c) 对于 FP4、NF4 等新型精度格式，应参照本附录方法另行确定系数；
- d) 整数运算（OPS）与浮点运算（FLOPS）之间的换算，可参考 INT8 对应的系数范围。