ICS
CCS 点击此处添加 CCS 号

T/ISC

团　　体　　标　　准

T/XXX XXXX—XXXX

大模型分布式训练中断场景分类与核定技术规范

Technical specification for classification and verification of interruption scenarios in large model distributed training

（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

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：

大模型分布式训练中断场景分类与核定技术规范

1 范围

本文件规定了大模型分布式训练中断场景下的分类、核定方法、证据记录以及争议处理等内容。

本文件适用于算力服务提供方与算力服务使用方之间，在大模型分布式训练中断场景下的溯源、核定、影响范围确定及相关处理。

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 32916 信息安全技术 信息安全控制评估指南

GB/T 36344 信息技术 云计算 云服务运营通用要求

GB/T 42581 信息技术服务 数据中心业务连续性等级评价准则

GB/T 43331 互联网数据中心（IDC）技术和分级要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

大模型分布式训练 large model distributed training

将大模型（参数量达到亿级及以上）训练任务拆分为多个子任务，在多台计算设备（GPU/NPU）组成的集群环境中，采用数据并行、模型并行、流水线并行或其组合方式进行并行计算的过程。

3.2

训练中断 training interruption

大模型分布式训练任务在预期完成之前非正常终止的状态，包括训练进程崩溃、节点失联、通信超时、任务被强制终止等具体表现形式。

3.3

算力服务提供方 computing power service provider

提供用于大模型分布式训练的算力资源（包括但不限于GPU/NPU服务器、高速网络、存储系统）及其管理服务，并向使用方收取费用的法人或组织。

3.4

算力服务使用方 computing power service user

租用或使用提供方算力资源进行大模型训练任务，并按约定承担相应费用的法人或组织。

3.5

中断恢复时间 interruption recovery time

从训练中断被确认时刻起，到训练任务恢复到断点状态并可继续执行所需的时间间隔。

3.6

检查点 checkpoint

在训练过程中定期保存的模型参数快照及优化器状态，用于在中断后从最近一次保存状态恢复训练，避免重复计算。

4 分类

4.1 分类原则

中断分类应遵循以下原则。

a) 原因导向原则：根据中断的直接诱因进行归类，而非中断的表象或后果。
b) 客观性原则：分类基于可验证的技术事实和证据记录，避免主观臆断。
c) 时效性原则：中断发生后在约定的时间窗口内完成初步分类。
d) 一致性原则：同类中断在不同时间、不同节点上采用相同的分类标准。

4.2 按责任归属分类

4.2.1 算力服务提供方责任中断

指因提供方所提供的算力基础设施或管理服务存在缺陷而导致的训练中断，包括但不限于以下内容：

a) 硬件故障（GPU/NPU 故障、电源异常、存储失效）；
b) 网络故障（跨节点通信链路中断、交换机故障、路由异常、All-Reduce 通信失败）；
c) 软件或系统故障（操作系统崩溃、容器或虚拟化异常、调度器错误、深度学习框架 Bug）；
d) 供电或制冷异常（数据中心电力中断、温度超标导致降频或宕机）；
e) 提供方主动运维操作不当（未经提前通知的停机维护、错误配置变更、安全策略误拦截）；
f) 多租户资源隔离失效（邻居租户的异常流量或资源抢占导致本租户训练任务受到影响）。

4.2.2 算力服务使用方责任中断

指因使用方自身行为或应用层面问题导致的训练中断，包括但不限于以下内容：

a) 使用方提交的训练代码存在逻辑错误（如除零、内存溢出、死锁、梯度爆炸）；
b) 使用方数据集损坏、格式错误或访问权限异常；
c) 使用方自行调整训练参数（如学习率、批量大小）引发训练不稳定；
d) 使用方在未获得提供方许可下，对系统环境进行修改或安装非授权软件；
e) 使用方未按约定配置检查点（checkpoint）或存储策略，导致无法断点续训；
f) 使用方提交的训练任务超出合同约定的资源配额（如 GPU 显存、内存、磁盘 IOPS），导致资源超限被强制终止。

4.2.3 不可抗力或第三方责任中断

指超出提供方和使用方合理控制范围的事件导致的中断，包括但不限于以下内容：

a) 自然灾害（地震、洪水、台风、雷击等）；
b) 政府或公共机构的行为（电力管制、网络限行、政策调整）；
c) 第三方网络运营商故障（公网中断、ISP 骨干网故障、CDN 节点故障）；
d) 大规模外部攻击（DDoS 攻击、勒索病毒传播等）；
e) 上游供应商（如芯片原厂、云服务商）的底层服务中断。

4.2.4 混合型中断

指中断原因为多方因素复合（如使用方代码异常与提供方网络抖动同时发生），难以简单归因于单一责任方的中断场景。

4.3 按中断形态分类

4.3.1 进程级中断

训练进程因异常而崩溃退出，包括但不限于：

a) 进程收到 SIGKILL、SIGSEGV 等信号异常终止；
b) 进程因内存不足（OOM）被操作系统终止；
c) 进程因未捕获的异常（如 Python 异常）而退出。

4.3.2 节点级中断

一个或多个计算节点失去响应或脱离集群，包括但不限于：

a) 节点宕机或重启；

b） 节点网络失联；
c） 节点 GPU/NPU 设备故障导致无法继续计算。

4.3.3 通信级中断

分布式训练中节点间通信失败导致训练无法继续，包括但不限于：
a） All-Reduce 等集合通信超时或失败；
b） 通信链路拥塞导致的心跳超时；
c） NCCL/RCCL 等通信库报错。

4.3.4 存储级中断

与持久化存储相关的异常导致训练中断，包括但不限于：
a） 检查点写入失败（存储空间不足、权限错误）；
b） 训练数据读取失败（存储挂载点失效、网络存储不可达）；
c） 模型保存或加载过程中的 I/O 错误。

4.3.5 调度级中断

任务调度层面的异常导致训练中断，包括但不限于：
a） 任务被调度器驱逐（抢占、优先级调整）；
b） 节点维护导致任务被迁移或终止；
c） 配额或计费异常导致任务被终止。

4.4 按中断阶段分类

4.4.1 启动阶段中断

训练任务在初始化过程中发生的中断，包括环境准备、资源分配、模型加载、数据加载等环节的异常。

4.4.2 运行阶段中断

训练任务进入正常迭代循环后发生的中断，包括前向传播、反向传播、梯度同步、参数更新等环节的异常。

4.4.3 保存阶段中断

训练任务在保存检查点或输出结果时发生的中断，包括检查点写入失败、存储超时等。

5 核定方法

5.1 核定原则

5.1.1 客观性原则

中断核定应基于可验证的技术证据，包括但不限于系统日志、监控数据、时间戳记录等，避免依赖主观判断。

5.1.2 及时性原则

中断发生后应尽快启动核定流程，确保证据的有效性和完整性，防止关键证据因时间推移而丢失或被覆盖。

5.1.3 可追溯性原则

核定过程中的所有步骤、依据和结论均应记录在案，支持事后审计和复查。

5.2 核定流程

中断核定按照以下流程进行。

a) 中断发现与通知：任一方发现中断发生后，应在不超过 1 h 内通知对方，并提供中断的基本信息（发生时间、影响范围、初步现象）。
b) 证据保全：双方在中断发生后 24 h 内完成核心证据的保全，确保证据不被覆盖或修改。
c) 初步核定：提供方应在中断确认后 24 h 内完成初步核定，确定中断的可能原因和初步分类。
d) 双方确认：提供方将核定结果通知使用方，使用方在收到通知后 24 h 内反馈确认或提出异议。
e) 联合排查：若双方对核定结果存在分歧，应在 48 h 内启动联合排查。
f) 最终核定：经双方确认或第三方鉴定后，形成最终核定结论。

5.3 核定依据

5.3.1 证据来源

中断核定应基于以下证据来源。
a) 系统层证据：操作系统日志、内核日志、系统事件记录。
b) 硬件层证据：BMC/IPMI 日志、硬件告警记录、设备健康状态。
c) 网络层证据：交换机日志、网络流量统计、通信库日志。
d) 应用层证据：训练框架日志、任务输出日志、错误堆栈信息。
e) 监控数据：性能监控指标时序数据（GPU 利用率、内存使用率、网络吞吐量、存储 IOPS）。
f) 检查点状态：最近有效检查点的位置、时间及完整性信息。
g) 操作审计：集群管理操作记录、用户操作审计日志。

5.3.2 时间同步要求

所有证据记录的时间戳应采用统一的参考时钟源，各设备间的时间偏差不应超过100 ms。建议采用NTP（Network Time Protocol）或PTP（Precision Time Protocol）进行时间同步。

5.3.3 证据链完整性

证据记录应形成闭环的证据链，从中断发生前的正常状态到中断发生时刻再到中断后的恢复操作，均应有相应记录予以支撑。

5.4 核定方法

5.4.1 单一原因核定的判定方法

对于可明确归因于单一原因的中断，通过以下方法进行核定。
a) 日志分析法：通过分析中断时刻的系统日志和应用日志，定位中断的直接原因。
b) 监控比对法：通过比对中断前后的监控指标变化趋势，确定异常的起始点和根因。
c) 回放验证法：在测试环境中复现中断条件，验证原因假设。

5.4.2 混合型中断的核定方法

对于混合型中断，应按照以下步骤核定各方因素的贡献程度：
a) 确定中断发生时刻的完整事件序列；
b) 识别所有可能的原因因素；
c) 分析各因素与中断事件之间的时序关系和因果强度；
d) 评估各因素对中断的直接贡献程度（采用百分比表示）；
e) 结合行业经验和技术可行性，确认最终的责任比例分配。

5.4.3 不可抗力的核定方法

不可抗力事件的核定应依据以下材料：
a) 政府或权威机构发布的正式公告或证明文件；
b) 相关新闻报道或公开信息；
c) 第三方鉴定机构的鉴定报告；
d) 受影响方提供的损失说明和证明材料。

5.5 核定结果

5.5.1 核定结论

核定结论应包括以下内容：

a) 中断事件的基本信息（任务 ID、中断时间、中断形态）；
b) 中断原因分析；
c) 中断类型分类（按 4.2 和 4.3 的分类体系）；
d) 责任归属（提供方责任/使用方责任/不可抗力或第三方责任/混合型）；
e) 若为混合型中断，应包含各方的责任比例；
f) 核定依据（引用相关证据记录）。

5.5.2 核定异议

使用方对核定结论有异议的，应在收到核定通知后按双方约定的时限内（建议不超过7个工作日）提出，并附具相关证据材料。双方应在收到异议后10个工作日内协商处理。

5.5.3 核定纠错

核定结论作出后发现新的关键证据，或者原核定依据存在明显错误，任一方可申请重新核定。重新核定应由双方共同确认或委托第三方进行。

6 证据记录

6.1 记录的基本要求

提供方应建立自动化的证据记录系统，对算力集群的全部运行状态进行持续性监控和日志留存。记录应符合下列要求。

a) 完整性：涵盖所有计算节点、网络设备、存储设备的关键指标。
a) 准确性：时间戳统一至毫秒级，并保证不同设备间时间同步（建议采用 NTP 或 PTP 协议）。
b) 不可篡改性：采用日志签名、区块链存证或其他技术手段，确保记录在事后无法被修改或删除。
c) 可检索性：记录应支持按时间、节点、任务 ID、中断类型等条件快速检索。

6.2 记录的核心数据项

至少应包括以下核心数据项：

a) 任务标识（任务 ID、用户 ID）；
b) 中断发生时间（精确到毫秒）；
c) 中断类型（如进程崩溃、节点失联、通信超时）；
d) 中断前最后 10 min 的系统监控快照(CPU/GPU 利用率、内存占用、网络吞吐量、存储读写 IOPS)；
e) 中断相关日志（操作系统日志、训练框架日志、调度器日志）；
f) 检查点保存状态（最后有效检查点时间及路径）；
g) 中断恢复操作记录（重启、迁移、快照恢复等）及其操作时间。

6.3 记录的开放性要求

提供方应在保证数据安全和不泄露商业机密的前提下，向使用方开放中断相关证据的记录接口或查询权限。使用方有权在中断发生后7日内下载或查阅与本次中断直接相关的日志和监控数据。双方应签订数据保密协议，明确可用数据的范围和脱敏规则。

6.4 记录的存储与管理

6.4.1 存储期限

所有中断相关证据记录，应自中断确认之日起至少保存180天。超过保存期限后，提供方有权删除或归档，但若仍有争议未解决，应延长保存至争议处理完毕。

对于涉及重大金额争议的中断记录，保存期限应延长至争议处理完毕后不少于2年。

6.4.2 存储方式

中断记录应采用电子和纸质两种方式存储，确保数据的安全性和可追溯性。电子记录应存储在防篡改的介质上，并定期进行备份；纸质记录应存放在防火、防潮的专用档案柜中。

7 争议处理

7.1 协商解决

发生核定或分类争议时，双方应首先启动协商程序，在收到争议通知后的5个工作日内展开协商。协商期间，双方应提供各自持有的证据，并可就中断分类、核定方法等提出解决方案。协商结果应以书面形式确认。

7.2 第三方鉴定

若协商无法达成一致，双方可共同委托具备资质的第三方鉴定机构（如CNAS认可的实验室、行业内公认的测评机构）进行技术鉴定。鉴定内容包括中断原因分析、责任判定、有效时长核算等。鉴定报告应作为最终责任认定的依据，且双方应共同承担鉴定费用，具体承担比例可由鉴定机构建议或双方协商。

7.3 法律途径

7.3.1 仲裁

若第三方鉴定仍无法解决，或一方拒绝接受鉴定结果，且双方在合同中约定了仲裁条款的，任何一方可向约定的仲裁机构提起仲裁。仲裁裁决为终局裁决，对双方均具有约束力。

7.3.2 诉讼

若双方未约定仲裁条款，或仲裁条款无效，任何一方可向有管辖权的人民法院提起诉讼。争议处理期间，除争议费用部分外，双方应继续履行无争议部分的合同义务。

7.3.3 不可抗力通知义务

发生不可抗力事件后，受影响方应在24 h书面形式通知对方，并在7日内提供相关证明文件。未及时履行通知义务的，不得就因此扩大的损失主张免责。

附 录 A
（资料性）
典型中断场景责任判定参考表

表A.1列出了常见中断场景及其责任判定参考。

表 A.1 典型中断场景责任判定参考表

中断场景	中断类型	责任方	处理要点
GPU故障导致节点宕机	提供方责任中断	提供方	中断时间不计费； 提供方负责硬件更换和任务恢复
跨节点RDMA通信超时	提供方责任中断	提供方	检查网络设备和驱动版本； 中断时间不计费
All-Reduce梯度同步失败	混合型中断	视原因判定	检查代码是否正确实现分布式通信； 若为网络问题归提供方
训练代码内存溢出	使用方责任中断	使用方	中断时间正常计费； 使用方优化代码或调整批量大小
数据集文件损坏	使用方责任中断	使用方	中断时间正常计费； 使用方修复数据集
学习率设置过大导致梯度爆炸	使用方责任中断	使用方	中断时间正常计费； 使用方调整训练参数
数据中心断电	提供方责任中断	提供方	检查UPS和备用电源； 中断时间不计费
机房温度超标	提供方责任中断	提供方	检查制冷系统； 中断时间不计费
DDoS攻击导致网络中断	不可抗力/第三方	第三方	中断时间不计费； 双方协助恢复
邻居租户挖矿导致资源争用	提供方责任中断	提供方	加强多租户隔离； 中断时间不计费
straggler节点拖慢整体训练	混合型中断	视原因判定	检查节点健康状态； 若硬件问题归提供方
检查点保存失败	使用方责任中断	使用方	恢复时长计入使用方责任； 使用方优化检查点策略
地震导致数据中心损毁	不可抗力	不可抗力	双方均无责； 已产生费用按有效时长结算
政府网络管制导致公网中断	不可抗力	不可抗力	中断时间不计费； 超过容忍阈值可终止任务

附 录 B
（资料性）
中断证据记录模板

表B.1提供了中断证据记录的标准模板，供提供方在使用时参考。

表 B.1 中断证据记录模板

记录字段	记录内容示例	记录要求
任务ID	train_20260801_001	唯一标识，由系统自动生成
用户ID	user_alpha	关联算力服务使用方
中断发生时间	2026-08-01 14:32:15.234	精确到毫秒，统一时区
中断类型	节点失联（Node 7）	按3.2定义的分类填写
中断前监控快照	GPU0利用率98%，GPU7利用率0%	中断前10分钟的平均值
系统日志	kernel:Out of memory: Kill process 12345	截取关键错误信息
训练框架日志	RuntimeError:CUDA out of memory	框架层错误堆栈
检查点状态	最后检查点：step 50000,time=2026-08-01 14:30:00	记录最后有效检查点信息
恢复操作记录	14:35:00 重启节点7；14:40:00 恢复检查点	记录所有恢复操作步骤和时间
恢复完成时间	2026-08-01 14:42:30.100	任务可继续执行的时间点
中断恢复时间	10 min15 s	从确认中断到恢复完成
证据保全状态	日志已存证，哈希值：0x7f9a...	不可篡改存证标识
核定结论	提供方责任（节点 7 硬件故障）	核定依据和结论