ICS 35.020
CCS L 60

中华人民共和国国家标准化指导性技术文件

GB/Z 275—2026

智能计算　GPU 虚拟化与性能优化技术要求

Intelligent computing—Technical requirements for GPU virtualization and performance optimization

2026-09-24 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　GPU 虚拟化 …… 2
5.1　概述 …… 2
5.2　硬件层 …… 2
5.3　内核层 …… 3
5.4　基础支持库层 …… 3
5.5　应用层 …… 5
6　性能优化 …… 6
6.1　通则 …… 6
6.2　单应用性能优化 …… 6
6.3　多应用性能优化 …… 6
6.4　负载变化的多应用性能优化 …… 7
附录 A（规范性）　性能评估指标 …… 8
A.1　概述 …… 8
A.2　弹性度 …… 8
A.3　隔离性 …… 8
A.4　算力折损率 …… 9

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：中国信息通信研究院、之江实验室、北京趋动智能科技有限公司、天翼云科技有限公司、中国电信集团有限公司、中电信数智科技有限公司、中兴通讯股份有限公司、中移（杭州）信息技术有限公司、中国铁塔股份有限公司、京东科技信息技术有限公司、智己汽车科技有限公司、山东省计算中心（国家超级计算济南中心）、浪潮云信息技术股份公司、中国科学院计算机网络信息中心、北京思谋智能科技有限公司、江苏博云科技股份有限公司、中国移动通信集团有限公司、蚂蚁科技集团股份有限公司、中国移动通信集团辽宁有限公司、海南智算创新发展研究院、南方电网储能股份有限公司信息通信分公司、图为信息科技（深圳）有限公司、侨远科技有限公司。

本文件主要起草人：庄祎、郑立、任银垠、韩思齐、栗蔚、陈屹力、李振廷、王鲲、陈飞、江政雄、崔娜妮、杨经纬、李响、张子浪、韩冬、张博、方敏伟、王春晓、郑佳佳、陆腾、张驰、张振华、蔡旭辉、郑卿、龚明亮、张宇轩、周小淇、王波、朱俊威、苏世鹏、张磊。

智能计算　GPU虚拟化与性能优化技术要求

1　范围

本文件规定了GPU虚拟化与性能优化的技术要求。

本文件适用于GPU虚拟化相关产品的设计、研发和测试。

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 35293　信息技术　云计算　虚拟机管理通用要求

3　术语和定义

GB/T 35293界定的以及下列术语和定义适用于本文件。

3.1

GPU虚拟化　GPU virtualization

通过软件或硬件技术将物理GPU资源进行抽象、分割与分配，使多个虚拟机或容器能够独立访问和使用GPU资源，实现物理GPU资源的共享与高效利用的技术。

3.2

单根输入/输出虚拟化　single root I/O virtualization

通过在单个物理PCIe设备上创建多个独立的“虚拟功能”，使每个虚拟功能可被虚拟机或容器直接识别为独立的PCIe设备，从而实现单个物理GPU支持多虚拟机/容器并行访问的技术。

3.3

性能优化　performance optimization

通过采用内存管理优化、任务调度算法改进、通信机制升级等技术手段，提升GPU在虚拟化环境下的计算效率、降低资源损耗，以满足不同应用场景性能需求的过程。

4　缩略语

下列缩略语适用于本文件。

API：应用编程接口（Application Programming Interface）

DMA：直接内存访问（Direct Memory Access）

GPU：图像处理单元（Graphics Processing Unit）

PCIe：外设组件互联标准（Peripheral Component Interconnect Express）

RDMA：远程直接数据存取（Remote Direct Memory Access）

SR-IOV：单根输入/输出虚拟化（Single Root Input Output Virtualization）