

《人工智能 加速卡技术要求及测试方法》

（征求意见稿）编制说明

一、工作简况

《人工智能 加速卡技术要求及测试方法》（计划号：073-T/ISC-23）于2023年8月18日由中国互联网协会下达立项计划。本标准自主制定标准，起草单位为中国移动通信集团有限公司、中国互联网协会人工智能工作委员会、中国信息通信研究院、北京智源人工智能研究院、南方电网人工智能科技有限公司、国能数智科技开发（北京）有限公司、中国石油化工集团有限公司、华为技术有限公司、上海燧原科技股份有限公司、中科寒武纪科技股份有限公司、上海天数智芯半导体有限公司、海光信息技术股份有限公司、摩尔线程智能科技（北京）有限责任公司、中讯邮电咨询设计院有限公司、北京百度网讯科技有限公司、北京智谱华章科技有限公司、新华三技术有限公司、浪潮集团有限公司、广州趣丸网络科技有限公司等行业内具有代表性的加速卡、芯片、服务器供应方、评测方、需求方等企业和科研机构，本标准提出的技术要求经主要加速卡供应商进行了实验验证，是经行业代表性企业协商一致共同认可的实际行动，具备可行性和通用性。

二、标准编制原则和确定标准主要内容的依据

当前大模型等人工智能技术的飞速发展对计算资源产生了海量需求，国内外各厂商研制的人工智能加速卡在数据精度、存储类型、通信协议、算子类型、软件栈等方面各有差异，导致加速卡在满足海量人工智能训练和推理需求时存在较大的协调难度。本标准针对该现状，以人工智能加速卡产品为标准化对象，通过对我国人工智能加速卡产品的技术现状进行充分调研，并联合相关供应方、需求方、评测方等共同研讨，以技术和产业实际为基础，结合产业需求，就人工智能加速卡产品的功能与性能指标等技术要求达成一致，并给出规范化的人工智能加速卡测试评估方法，旨在为人工智能加速卡的供需双方提供可参考的设计、选型、评测依据，并保障测试评估结果的客观性和准确性，促进人工智能加速卡产品的规模化应用和产业生态的健康发展。

本标准按照 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》要求进行编写，主要规定了人工智能加速卡的技术要求，包含

人工智能加速卡的通用技术要求，人工智能训练卡和推理卡的功能要求、性能要求、兼容性要求、可靠性要求、性能度量指标，以及人工智能加速卡的安全性要求，并给出了人工智能训练卡和推理卡的测试环境及测试方法。本标准适用于人工智能加速卡的供应方、评测方、需求方等对加速卡进行设计、测试、评估、选型和应用。

三、工作简要过程

2023年8月18日，该项标准制订计划正式下达，中国移动通信集团有限公司（以下简称中国移动）开始牵头负责标准草案大纲的编写工作。

2023年9月21日，中国移动组织中国互联网协会人工智能工作委员会、中国信息通信研究院、北京智源人工智能研究院、南方电网人工智能科技有限公司、国能数智科技开发（北京）有限公司、华为技术有限公司、上海燧原科技股份有限公司、中科寒武纪科技股份有限公司等11家首批起草单位召开标准编制启动会，正式成立标准起草组，在会上介绍了该项标准的编制背景和立项情况，就标准草案大纲进行讨论并达成一致。

2023年10月18日，标准起草组开始征集第二批标准参编单位，经过公开征集后增加了8家起草单位。

2023年11月至2024年4月，标准起草组组织召开了多次线上研讨会，围绕各起草单位的标准内容提案进行讨论和审议，共同开展标准草案的编制工作，逐步形成工作组草案。

2024年5月21日，标准起草组在北京组织召开标准编制研讨会，各起草单位共30余名代表参会，会上对标准草案内容进行了深入讨论，并提出了相关修改建议。本次会议讨论决定，标准修改完成后进入征求意见阶段。

四、主要试验（或验证）的分析

本标准起草单位中有众多业内领先的人工智能加速卡研发和生产机构，标准编制过程中，各起草单位结合各自产品在设计、研发、生产、应用过程中的经验和试验验证情况，给出了标准草案相应的技术内容和指标，因此标准草案的相关技术要求和测试方法符合行业共识，经历过实际验证。

五、采用国际标准和国外先进标准情况

本标准为自主研制标准，未采用国际标准和国外先进标准。本标准在编制过

程中参考了 ITU-T F.748.11 深度学习处理器基准测试评测指标与方法 (Metrics and evaluation methods for a deep neural network processor benchmark), 该标准主要给出了深度学习处理器的性能测试框架和评估指标, 为本标准中人工智能加速卡的训练性能指标和推理性能指标的设定提供了参考和借鉴。

六、标准涉及的专利说明

本标准暂未发现相关知识产权问题。

七、预期达到的社会效益、对产业发展的作用

当前, 以深度学习为代表的人工智能技术正在深刻的改变着人类生产和生活方式, 推动各行业提质增效, 人工智能加速卡作为一种具有适配人工智能算法运算微架构、能够完成人工智能应用运算处理的集成电路元件, 可以提供比通用计算资源更高的计算性能, 因此成为了处理大规模人工智能计算任务的重要工具, 被广泛应用于人工智能领域的训练和推理等任务中。目前由于缺乏人工智能加速卡的相关标准和规范, 一方面造成了各类 AI 加速卡产品自由发展, 各维度的功能和性能不能相互协调以满足 AI 模型训练和推理的实际需求; 另一方面也无法有效的测试和综合评估 AI 加速卡的技术能力, 应用方在选用 AI 加速卡的时候容易产生困扰。本标准对人工智能加速卡的功能、性能、兼容性、可靠性、安全性、训练性能和推理性能等技术内容进行规范, 并给出标准化的测试评估方法, 可以为人工智能加速卡的供需双方提供可参考的设计、选型、评测依据, 促进人工智能加速卡产品的规模化应用和我国人工智能产业生态的健康发展。

八、与现行相关法律、法规、规章及相关标准的协调性

本标准符合我国有关的现行法律、法规和相关强制性标准的相关规定。与本标准相关的标准有 GB/T 42018-2020 《信息技术 人工智能平台计算资源规范》, YD/T 3944-2021 《人工智能芯片基准测试评估方法》, T/CESA 1119-2020 《人工智能芯片 面向云侧的深度学习芯片测试指标与测试方法》, T/CESA 1120-2020 《人工智能芯片 面向边缘侧的深度学习芯片测试指标与测试方法》和 T/CESA 1121-2020 《人工智能芯片 面向端侧的深度学习芯片测试指标与测试方法》。

已发布的几项相关标准中, 大部分标准的标准化对象为人工智能芯片, 且主要规范的是人工智能芯片在模型训练和推理过程中有关性能方面的基准测试指标和测试方法, 缺少相关技术要求; 另一部分标准的标准化对象虽然包含人工智

能加速卡，但是只给出了内存、带宽、视频编解码等通用技术要求，规范的内容不够全面。芯片虽然是人工智能加速卡的重要组成部分，但其相关标准无法涵盖加速卡的全部技术内容，本标准不仅对人工智能加速卡本身的功能、性能、兼容性、可靠性、安全性、训练/推理性能给出全面的技术要求，同时也给出了相应的测试方法，此外还兼顾了当前大模型技术快速发展背景下对人工智能加速卡提出的新需求，国内外目前尚无相关同类标准。本标准将作为现有标准的有益补充，和现有标准互相配合，共同促进人工智能加速设备技术和产业的发展。

九、重点分歧意见的处理经过和依据

本标准在编制过程中没有重大分歧。

十、标准性质的建议说明

本标准建议作为推荐性团体标准进行发布。

十一、贯彻标准的要求和措施建议

本标准的制定可以为人工智能加速卡相关产品的设计、研发和生产提供参考和指导，在功能和性能等相关领域具有一致性和互联互通能力的人工智能加速卡也有助于构建更高效、更可靠的智算集群，为开发更高性能的人工智能算法、提供更便捷的人工智能服务提供基础设施保障，对于促进我国人工智能技术和产业的发展具有积极意义。建议在标准发布后尽快组织宣贯、推广和应用实施。

在具体贯彻实施本标准时，一是建议标准归口单位中国互联网协会在会员单位内部进行宣贯，加大标准的宣传力度，扩大影响范围；二是起草单位中各加速卡生产企业在产品设计、研发过程中，积极开展标准的应用实施工作；三是建议政府、企业等相关机构制定配套实施政策，如在招投标、认证服务等方面引导相关企业提升标准应用积极性；四是积极推动标准的符合性测试和评估工作，为标准的落地实践提供评测服务和保障。

十二、其他应予说明的事项

无。

《人工智能 加速卡技术要求及测试方法》标准起草组

2024年6月18日