团体标准《人工智能加速卡管理接口规范》（征求意见稿）编制说明

一、工作简况

1、项目来源和工作单位

按照中国电子工业标准化技术协会2024年第一批团体标准制修订项目的通知（中电标通〔2024〕001号），中国电子工业标准化技术协会开放计算工作委员会启动了《人工智能加速卡管理接口规范》（立项号：CESA-2024-006）的制订工作。任务下发后，由浪潮电子信息产业股份有限公司牵头，参与单位包括上海壁仞科技股份有限公司、中科寒武纪科技股份有限公司、上海燧原科技股份有限公司、上海天数智芯半导体有限公司、新华三技术有限公司、中国质量认证中心、昆仑太科（北京）技术股份有限公司、中国民航信息网络股份有限公司等。归口单位为中国电子工业标准化技术协会。

2、主要工作过程

（一）标准预研

随着人工智能加速卡算力需求持续提升，人工智能加速卡管理接口不统一导致的AI算力集群运维难度增加，以及各人工智能加速卡厂商获取信息内容和方式差异性大，导致加速卡关键管理信息支持不全面，与服务器适配难度增加、适配周期变长等问题，2023年8月浪潮电子信息产业股份有限公司与上海壁仞科技股份有限公司、中科寒武纪科技股份有限公司、上海燧原科技股份有限公司、上海天数智芯半导体有限公司等芯片厂商对管理接口规范需求进行了初步分析，确定了人工智能加速卡管理接口规范的标准草案结构和主要内容。2023年8月至2023年10月，标准工作组组织2次标准讨论会，对标准制定的必要性，可行性，目的意义，拟解决的问题，标准范围进行了充分讨论，最终取得技术共识。

（二）标准立项

2023年10月，浪潮电子信息产业股份有限公司，联合上海壁仞科技股份有限公司、中科寒武纪科技股份有限公司、上海燧原科技股份有限公司、上海天数智

芯半导体有限公司作为共同发起方申请立项，并通过评审，成为协会正式标准制定项目。

（三）标准编制

2024年4月11日，标准工作组完成了面向工作组内成员单位的标准意见征集，并组织召开了标准启动会，对标准制定的背景、参编单位构成，标准推动计划进行了介绍。同时对标准适用性和拟要增加的管理接口进行了讨论。

2024年5月至6月，就编制组内专家所提意见，包含管理接口命令集的完整性等进行了针对性讨论。

3、主要起草人及其所做的工作

本标准由浪潮电子信息产业股份有限公司牵头组织编制、参与标准编制的单位有上海壁仞科技股份有限公司、中科寒武纪科技股份有限公司、上海燧原科技股份有限公司、上海天数智芯半导体有限公司、新华三技术有限公司、中国质量认证中心、昆仑太科（北京）技术股份有限公司、中国民航信息网络股份有限公司等。其他成员单位提供了标准所涉及的技术内容的材料，并参与了技术细节的讨论。

二、标准编制原则和确定主要内容的论据及解决的主要问题

1、编制原则

在标准编制过程中，遵循了以下五方面的原则。

a） 符合性。一是遵循国家法律、法规等相关规定，制定过程严格按照程序执行。

b） 先进性。本标准制定过程中充分考虑了管理接口的技术现状，并在方面保持了一定的前瞻性。

c） 适用性。本标准结合实际人工智能加速卡管理接口与服务器实际对加速卡信息获取的应用需求进行接口定义。

d） 中立性。本标准制定过程中编制组成员单位对标准文本进行了充分讨论。

2、确定主要内容的依据

近年来随着人工智能产业的高速发展，传统芯片的算力和性能越来越无法满足产业的发展需求，测算数据显示，到2025年，中国人工智能芯片市场规模预计

将达到1740亿元，人工智能加速卡的生产与制造已成为行业竞争的关键。当前人工智能加速卡管理接口面临着多方挑战。首先，随着算力需求提升，AI 计算集群规范急剧扩大，导致AI加速卡的管理运维变得更加困难；其次，当前各厂商对人工智能加速卡的管理并没有统一的管理协议，采用的是私有化的方式，具体表现为不同厂商所支持获取的信息内容不一致，以及即使获取同一信息，对应的协议命令格式也不同，这种管理接口上的多样性要求服务器系统需要针对不同加速卡进行单独适配，导致服务器系统整个适配难度增加，适配周期延长；再者，由于AI加速卡缺少统一的管理接口，导致有些加速卡对于终端客户希望获取的关键信息支持不全面，比如各种故障相关信息的支持，最终导致难以满足客户运维需求的尴尬局面。

本标准对管理接口进行统一定义，包括接口的物理层、命令集、命令格式等，能够统一各AI加速卡厂商的信息获取方式和获取的信息内容。本标准也将给出相应的测试方法，对管理接口实现的程度进行统一评价。

3、编制过程中解决的主要问题

编制过程主要聚焦AI加速卡的管理接口命令集进行了详细交流：

（1）关于动态信息类命令集，增加了对芯片，内存和光模块温度的接口命令，增加了获取板卡启动状态的接口命令。

（2）关于诊断信息类命令集，增加了健康状态，RMA状态，PCIe错误数，PCIe UCE状态寄存器，PCIe UCE掩码寄存器，PCIe UCE等级寄存器，PCIe CE状态寄存器，PCIe CE掩码寄存器，PCIe AER控制寄存器，PCIe AER HDRLOG寄存器和PCIe AER TLPLOG寄存器的接口命令。

三、主要试验[或验证]情况分析

在标准起草过程中，标准编制组充分考虑了 AI 加速卡管理上的需求，对 AI 加速卡管理接口命令集的实现进行了充分讨论。

四、知识产权情况说明

本标准不涉及知识产权问题。

五、产业化情况、推广应用论证和预期达到的经济效果

标准发布后，主要包含两个主要方面：一是对于 AI 芯片设计厂商而言，能够基于本标准进行 AI 加速卡管理接口开发；二是对于系统集成厂商而言，本标

准能够指导他们进行 AI 加速卡的选型适配以及基于本标准进行带外管理模块的开发。

六、转化国际标准和国外先进标准情况

本标准未采用国际标准和国外先进标准。

七、与现行相关法律、法规、规章及相关标准的协调性

本标准编制文本格式按照GB/T 1.1-2020的规定起草。

八、重大分歧意见的处理经过和依据

无重大分歧。

九、贯彻标准的要求和措施建议

建议列为推荐性标准，在标准发布后尽快组织标准宣贯、试验验证。

十、替代或废止现行相关标准的建议

无需要替代或废止的现行相关标准。

十一、其它应予说明的事项

无。

《人工智能加速卡管理接口规范》团体标准编制起草组

2024-6-17