

《AI 算力场景下高带宽内存芯片性能适配技术规范》

编制说明

一、工作简况

（一）任务来源

本项目根据关于发布《AI 算力场景下高带宽存储芯片性能适配技术规范》等 6 项团体标准立项的公告（计划号 185-T/ISC-26）项目名称为“AI 算力场景下高带宽存储芯片性能适配技术规范”的任务而进行制订。

（二）起草单位及主要起草人

本文件起草单位：上海光羽芯辰科技有限公司、浙江大学、浙江常淳科技有限公司、华中科技大学、浙江力积存储科技股份有限公司、深圳市德明利技术股份有限公司、中科智存(珠海)技术有限公司、深圳市舟鸿半导体科技有限公司、得一微电子股份有限公司、华兴中科标准技术（北京）有限公司。

本文件主要起草人：周强、薛晓勇、张亦舒、展国彬、杨北洪、张猛、于晓、祁广田、陈佳锋、蒋文、李发生、杨继光、李华、任国静、马聪。

（三）标准制定目的和意义

近年来，人工智能算力基础设施的竞争焦点正从峰值算力向“算力—带宽—能效”的协同水平迁移。在大模型训练与推理场景中，高带宽内存（HBM）已成为决定系统实际可用性能的关键变量：算力增速约为每两年三倍，而 HBM 带宽增速不足两倍，“内存墙”持续扩大。HBM 通过 3D 堆叠将带宽推升至 HBM4 代的 2048GB/s 量级，但堆叠层数增加使功率密度和散热问题急剧恶化，热约束反过来限制可持续带宽输出。更严峻的是可靠性风险：Meta 在 Llama3 训练中高达 17% 的意外中断由 HBM 故障引起。与此同时，HBM 核心供应商已将绝大部分先进产能锁定于 HBM4，国产 HBM 虽在加速突破，但供给能力与适配验证体系尚不成熟。在标准层面，JEDEC JESD238、JESD330 系列已对 HBM3、HBM4 及 SPHBM4 的接口、封装、电气、时序、功耗和热管理等作出规定，但其立足点是器件级规范，回答的是“HBM 本身应满足什么”，并不回答“HBM 在特定 AI 算力系统中是否适配”。其必要性在于：一方面为国产 HBM 的设计验证和 AI 芯片/服务器的选型集成提供统一的技术标尺，避免仅凭峰值带宽参数做出适配判断；另一方面为智算中心建设和运营方提供可复现的适配性验收依据，降低因 HBM 热降频、带宽波

动或可靠性不足引发的算力损失，以标准化手段支撑国产 HBM 从“样品突破”走向“系统可用”。

（四）主要工作过程

2026 年 04 月 29 日，通过腾讯会议召开了本团体标准的立项评估会。本次会议确定了本标准的框架结构，收集了与会各单位对标准草案（初稿）的修改意见或建议，明确了工作组各单位的工作分工，以及会后的具体工作安排。

2026 年 05 月 08 日，本团体标准由中国互联网协会正式立项，立项名称为：AI 算力场景下高带宽存储芯片性能适配技术规范。

2026 年 08 月 27 日，通过腾讯会议线上召开了本团体标准的第一次讨论会。编制组就本团体标准第一次讨论后的修改情况、意见采纳情况进行了综合阐述。与会专家及代表对该项团体标准的定义内容、技术内容等进行了进一步的深入讨论和修改，编制组认真听取和记录了与会代表提出的问题和调整建议，标准名称由《AI 算力场景下高带宽存储芯片性能适配技术规范》修改为《AI 算力场景下高带宽内存芯片性能适配技术规范》。

2026 年 11 月至 12 月，对《AI 算力场景下高带宽内存芯片性能适配技术规范》（征求意见稿）公开征求意见。

二、标准编制原则和依据

（一）编制原则

标准起草小组在编制标准过程中，以国家、行业现有的标准为制订基础，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》的规定及相关要求编制。

（1）协调性

标准编写符合 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，技术内容与现行国家标准、行业标准协调一致，避免重复和冲突。

（2）科学性

本文件依据 AI 算力场景下大模型训练与推理的真实负载特征、HBM 高带宽存储的技术原理以及热约束和可靠性规律进行设计，带宽、容量、功耗、热行为和可靠性等

适配指标具有明确依据，确保适配性评价的合理性和有效性。

（3）实用性

本文件立足 AI 芯片、服务器与智算中心的实际集成和运行需求，面向 HBM 厂商、AI 芯片/服务器厂商及智算中心运营方，提出可操作、可复现的适配性技术要求和测试评价方法，避免仅凭峰值带宽参数判断适配性，为国产 HBM 的设计验证和系统选型提供切实可行的标准依据。

（二）标准主要内容与确定依据

一、适用范围

本文件规定了高带宽内存芯片（HBM）在人工智能训练和推理场景下的技术要求、性能适配指标和测试方法。

本文件适用于搭载高带宽内存（HBM）的 AI 加速卡及其 HBM 存储子系统的设计、选型、测试与验收。采用上述 AI 加速卡的 AI 服务器和智算中心在集成、选型和验收时可参照使用。规范性引用文件。

二、技术要求

（1）技术要求

本文件从存算协同架构、工作环境和信号完整性三个方面规定技术要求。存算协同架构按应用类型和计算精度规定最低存算比，通用 AI 训练和大模型训练采用 BF16/FP16 稠密算力，最低存算比 $\geq 3.0 \text{ GB/s} \cdot \text{TfLOPs}^{-1}$ ，大模型推理和多模态推理采用 FP8 稠密算力，最低存算比 $\geq 1.5 \text{ GB/s} \cdot \text{TfLOPs}^{-1}$ ，不计入结构化稀疏峰值算力；工作环境要求在 JEDEC 规范允许范围内确保长期可靠性，记录所依据的 JEDEC 文件编号、版本及制造商规定的工作温度、电压和刷新条件；信号完整性要求链路级 $\text{BER} \leq 10^{-12}$ 、ECC 后 $\text{BER} \leq 10^{-15}$ ，HBM 数据接口数据率不低于 6.4 Gb/s/pin。

（2）性能适配指标

本文件以搭载被测 HBM 的单个 AI 计算模组或加速卡为测量单元，规定五类性能适配指标。带宽性能要求 Copy、Scale、Add、Triad 各访存模式的单任务持续有效带宽利用率均 $\geq 70\%$ ，并报告 P5、P50、P95、最小值、最大值和变异系数，多任务应报告各任务及合计有效带宽、相对单任务基线比例和最长无进展时间；延迟性能要求 GPU 类 AI 加速卡设备侧 HBM 全局内存基准平均访问延迟 $\leq 800 \text{ cycles}$ ，报告 P50、P95、P99

及最大值，随机和顺序访问分别测试报告；单位比特能耗按代际规定 $HBM3 \leq 5 \text{ pJ/bit}$ 、 $HBM3e \leq 4.5 \text{ pJ/bit}$ 、 $HBM4 \leq 4 \text{ pJ/bit}$ ，以动态功耗与实测有效带宽的比值计算；可靠性要求连续稳定运行 $\geq 72 \text{ h}$ 且无不可纠正错误、功能失效或测试中断，ECC 后不可纠正软错误率 $\leq 10^{-13} \text{ errors/(bit} \cdot \text{h)}$ ，按单侧 95% 置信上限判定；热与功耗约束要求结温 $\leq 95 \text{ }^{\circ}\text{C}$ 、热梯度 $\leq 20 \text{ }^{\circ}\text{C}$ ，报告功耗测量边界、总功耗、环境温度、冷却方式、工作负载、运行频率和测试持续时间。

（3）测试方法

本文件从测试负载、专项测试、检验规则和测试报告四个方面规定测试方法。测试负载应覆盖 LLM 训练（Transformer，序列长度 4096~8192，含周期性检查点保存/恢复）、LLM 推理（KVCache 反复读写）、稀疏访问模式（MoE 路由访问）和高并发访存（并发线程数 ≥ 32 ）；带宽测试采用 STREAM 类程序分别执行 Copy、Scale、Add、Triad，各模式单任务和多任务均测试不少于 3600 s（前 600 s 为预热期），按 1 s 固定采样窗口记录有效带宽；延迟测试采用基准、随机和顺序三种访问模式，并发深度依次为 1、32、128，每种配置测试不少于 600 s；能耗测试将负载状态与空闲状态在相同环境和配置下同步测量功耗与带宽，按公式计算单位比特能耗；可靠性测试包括不少于 72 h 连续稳定运行和 SER 测试，SER 按单侧 95% 置信上限判定，试验方案应事前确定观测条件、固定结束时点和接收事件数；热与功耗测试采用片上传感器或经校准换算的测温方法，环境温度 $25^{\circ}\text{C} \pm 2^{\circ}\text{C}$ ，满载运行不少于 30 min 至稳定状态；信号完整性测试采用高速示波器眼图测试和链路级 BER 统计，链路级与 ECC 后统计分别记录观测位置、错误位数和受测总比特数；检验规则规定抽样比例 $\geq 5\%$ 且同一批次不少于 3 个样品，复测不应覆盖首次结果；测试报告应区分实测结果、制造商资料和统计推算结果，测试缺项或测量失效时逐项说明。

三、国内外情况简要说明

经查，与本标准相关的现行国家标准情况如下：

GB/T 17574—1998《半导体器件 集成电路 第 2 部分：数字集成电路》：本文件在可靠性资料提供方式和检验规则设计上与该标准基本框架衔接，不重复其通用质量评定要求。

GB/T 36474—2018《半导体集成电路 第三代双倍数据速率同步动态随机存储器

（DDR3 SDRAM）测试方法》：本文件在 DRAM 类存储器的功能验证和电参数测试方法上与该标准保持技术逻辑上的一致。

GB/T 46346—2025《人工智能 计算中心 计算能力评估》：本文件在“以系统应用需求反推存储性能要求”的评价逻辑上与该标准一致，但评价层级为 HBM 在 AI 加速卡层面的性能适配。

四、与有关的现行法律、法规和强制性国家标准的关系

本标准的制定过程、技术要求的选定、试验方法的确定、检验项目设置等均符合现行法律法规和强制性国家标准的规定。

本标准在技术内容上与国家标准协调配套，无冲突存在。

五、重大分歧意见的处理经过和依据

本文件在制定过程中未出现重大分歧意见。

六、废止现行有关标准的建议

本标准不涉及对现行标准的废止。

七、其他应予说明的事项

无。

《AI 算力场景下高带宽内存芯片性能适配技术规范》团体标准编制组

2026 年 10 月