摩尔线程 MOORE THREADS

硅基流动 SiliconFlow

PD 分离异构算力混部实践技术白皮书

基于 MTT S5000 与国际高端 GPU 的异构算力协同推理方案

一、引言

1.1 背景：从“训练大模型”到“生产 Token”

大语言模型（LLM）的快速迭代与规模化落地，正在推动 AI 基础设施的根本性范式转变。行业焦点正从“训练更大模型”转向“以更低成本、更高稳定性生产高品质 Token”。据工业和信息化部数据，截至 2026 年 6 月底，我国智能算力规模已达 2185 EFLOPS（FP16），同比增长 177%，中国词元市场呈现爆发式增长。

在算力需求激增与自主安全多元化的双重背景下，AI 基础设施正面临三重核心瓶颈：

成本瓶颈：高并发推理带来持续算力消耗，Token 成本成为规模化应用的关键约束；

效率瓶颈：Prefill 与 Decode 阶段对算力资源的需求特性不同，传统一体化部署难以充分释放集群效率；

国产化瓶颈：国产 GPU/NPU/ 异构算力快速发展，但规模化部署仍需要系统级优化方案。

基础设施的价值定位正在从“供给算力”转向“组织算力、调度模型、稳定生产高品质 AI Token”。PD 分离（Prefill-Decode Disaggregation）架构与异构算力混部技术的结合，为这一转型提供了系统性解决方案。

1.2 PD 分离架构概述

在大模型推理过程中，Prefill（预填充）阶段需要并行处理整个输入序列以生成首个 Token，属于**计算密集型**操作；而 Decode（解码）阶段则逐个生成后续 Token，属于**访存密集型**操作。两类任务在计算特性、资源需求上存在显著差异。

PD 分离架构将这两个阶段拆分到不同的计算实例上独立运行，使 Prefill 和 Decode 任务互不干扰，分别针对各自特性进行专门优化。其本质是**让不同推理阶段匹配更合适的算力资源**。这种架构设计不仅消除了阶段间的资源争抢与时延干扰，还能显著提升系统的有效吞吐量。

1.3 异构算力混部的战略价值

异构算力混部是指在同一个推理集群中，将不同架构、不同厂商的 AI 芯片进行统一调度与协同工作。其核心价值在于：

因材施教：将计算密集型的 Prefill 任务分配给高算力芯片，将访存密集型的 Decode 任务分配给高带宽芯片，实现“1+1>2”的推理效能；

灵活多元：降低对单一芯片供应商的依赖，构建多元、弹性的算力供应体系；

成本优化：充分利用各类算力资源，以更优的性价比满足不同场景的推理需求；

政策响应：积极响应国家“人工智能 +”行动及全国一体化算力网建设，推动异构算力的高效调度与互联互通。

1.4 本方案的核心特色

摩尔线程针对大语言模型推出的 PD 异构分离方案，通过将高算力需求的 Prefill 计算池（MTT S5000）与高带宽的 Decode 生成池（国际高端 GPU）异构算力分离、分池部署，支持 KV Cache FP8 与超长上下文，实现高性价比词元生产。本方案在保障全链路服务质量的前提下，在 Prefill 阶段以 2:1 的部署比例实现国产 GPU 对国际高端 GPU 的算力等效替代，为大规模异构算力部署提供了可量化的实践。

二、参与方介绍

2.1 摩尔线程

摩尔线程是一家专注于全功能 GPU 芯片设计的国产领军企业，致力于为 AI 训练与推理提供自主可控的算力基础设施。

其旗舰产品 MTT S5000 是一款 AI 训推一体全功能 GPU 智算卡，核心搭载基于第四代 MUSA 架构“平湖”设计的 PH100 芯片。PH100 芯片已首批通过中国信息安全测评中心的国家《安全可靠测评》（2026 年第 2 号），安全可靠等级为 I 级。这是安全可靠测评体系建立以来，**首次将 AI 训练推理芯片纳入评测结果**。

MTT S5000 率先实现**硬件级原生 FP8 计算精度支持**，在确保模型收敛的同时显著降低显存占用。依托第四代 MUSA 全栈平台，MTT S5000 原生适配 PyTorch、Megatron-LM、vLLM 及 SGLang 等主流 AI 框架，开发者可实现代码的“零成本”迁移。

截至 2026 年 6 月 30 日，摩尔线程累计申请专利 2167 项，已获授权专利 788 项，其中发明专利 725 项，稳居国产 GPU 企业前列。

MTT S5000 在 Prefill 阶段的适配优势。Prefill 阶段的核心任务是处理整段输入，涉及大规模矩阵计算与并行算力输出，对计算密度、长序列吞吐能力要求极高。MTT S5000 凭借以下能力完美匹配 Prefill 负载：

高计算密度：具备充沛 FP8 稠密算力储备，为大规模并行矩阵计算提供强劲支撑；

长上下文吞吐优势：结合 FP8 推理与显存优化，在长上下文、多轮对话场景下保持高吞吐输出；

原生 FP8 精度支持：硬件级 FP8 计算，兼顾模型精度与推理效率；

生态迁移友好：原生适配主流 AI 框架，实现代码“零成本”迁移；

安全可控：国产化芯片保障技术栈安全可控与合规要求。

2.2 硅基流动

硅基流动是一家专注于 AI 推理基础设施的独立词元（Token）供应平台，定位为衔接算力、模型与应用的“中间层”。

公司依托自研的推理引擎与算力资源编排系统，整合上游多厂商异构算力与主流 AI 模型，将底层算力转化为标准化、可规模化交付的词元服务。其核心技术涵盖推理引擎层（融合 PD 分离、KV 缓存管理、专家并行、流水并行等技术）、异构算力调度层（统一智能调度与弹性伸缩）和模型适配层（已适配 170+ 款模型）。

硅基流动已深度适配摩尔线程等国产芯片，成为支持国产芯片数量多、使用广的 Token 生产线。公司日均 Token 调用量近万亿，服务超过 1000 万用户和 13,000 家（超 1.3 万家）企业客户，营收同比增长近 10 倍。

三、技术方案：PD 分离异构算力混部架构

3.1 总体架构与异构分工

本方案基于摩尔线程 MTT S5000 与国际高端 GPU（以 HXXX 为代表），构建 **4P2D** 异构集群（4 台 MTT S5000 负责 Prefill + 2 台 HXXX 负责 Decode），对标 **2P2D** 纯 HXXX 集群（4 台 HXXX）。选定 **Kimi K2.6** 作为部署模型，Prefill 和 Decode 阶段均采用 **FP8** 推理精度，P/D 两侧数据格式完全一致，无需额外转换，确保缓存数据的正确性与一致性。

在异构分工中，MTT S5000 依托自身充沛的 FP8 稠密算力储备，专门处理 **Prefill 重计算任务**——涵盖大规模矩阵计算、长序列编码、高并发入口请求的首段编码压力；国际高端 GPU 凭借其高显存带宽（4.8 TB/s），稳定承接 **Decode 输出任务**。同时，方案**支持 KV Cache FP8** 压缩存储与超长上下文，大幅降低显存占用，为长文本推理提供强劲支撑。

3.2 Prefill 与 Decode 的任务划分逻辑

维度	Prefill 阶段	Decode 阶段
计算特性	计算密集型	访存密集型
任务特点	并行处理整个输入序列 生成首个 Token	逐个生成后续 Token 维持会话上下文
核心需求	高算力（TFLOPS）	高显存带宽（TB/s）
推理精度	FP8	FP8

3.3 部署架构

方案采用 **4P2D** 异构集群部署架构：

Prefill 计算池：4 台 MTT S5000，承担高算力需求的预填充任务，针对长上下文场景进行极致优化；

Decode 生成池：2 台 HXXX，承担高带宽需求的解码生成任务。

部署按 **MTT S5000:HXXX = 2:1** 的比例进行算力折算与容量规划，可实现无缝扩容与平滑的国产化替代。

3.4 推理引擎与精度优化

方案基于**硅基流动自研推理引擎**实现异构算力的统一调度与优化。该引擎融合了 PD 分离、KV 缓存管理、专家并行、流水并行等多项核心技术。

在精度策略上，Prefill 和 Decode 均采用 **FP8 精度**，两侧格式完全一致，避免了跨芯片传输时的精度转换开销，同时确保了 KV Cache 数据的一致性和数值稳定性。MTT S5000 的硬件原生 FP8 支持，使该方案在保证模型质量的前提下，显著提升了推理吞吐并降低了显存占用。

针对长上下文场景，双方结合 Kimi K2.6 的模型架构进一步探索了多项优化策略：

KV Cache FP8 压缩存储：大幅减少显存占用，使长序列推理成为可能；

计算与通信重叠：在 Prefill 过程中提前传输部分 KV Cache，降低整体时延；

算子级融合与内核调优：针对长序列 Attention 进行专门优化。

3.4.1 推理加速框架总体架构

本方案的 PD 分离异构算力混部能力，依托硅基流动自研高性能推理加速框架 SiliconLLM 实现。

框架自顶向下分为四层：异构芯片层通过统一硬件抽象（HAL）适配摩尔线程、NVIDIA、AMD 等多类架构；推理引擎层融合 PD 分离、KV Cache 管理、专家并行、MTP 等核心机制；平台层基于 K8s 提供弹性调度；应用层提供兼容 OpenAI 协议的标准化接口，提供从异构芯片适配到应用接入的全栈能力。

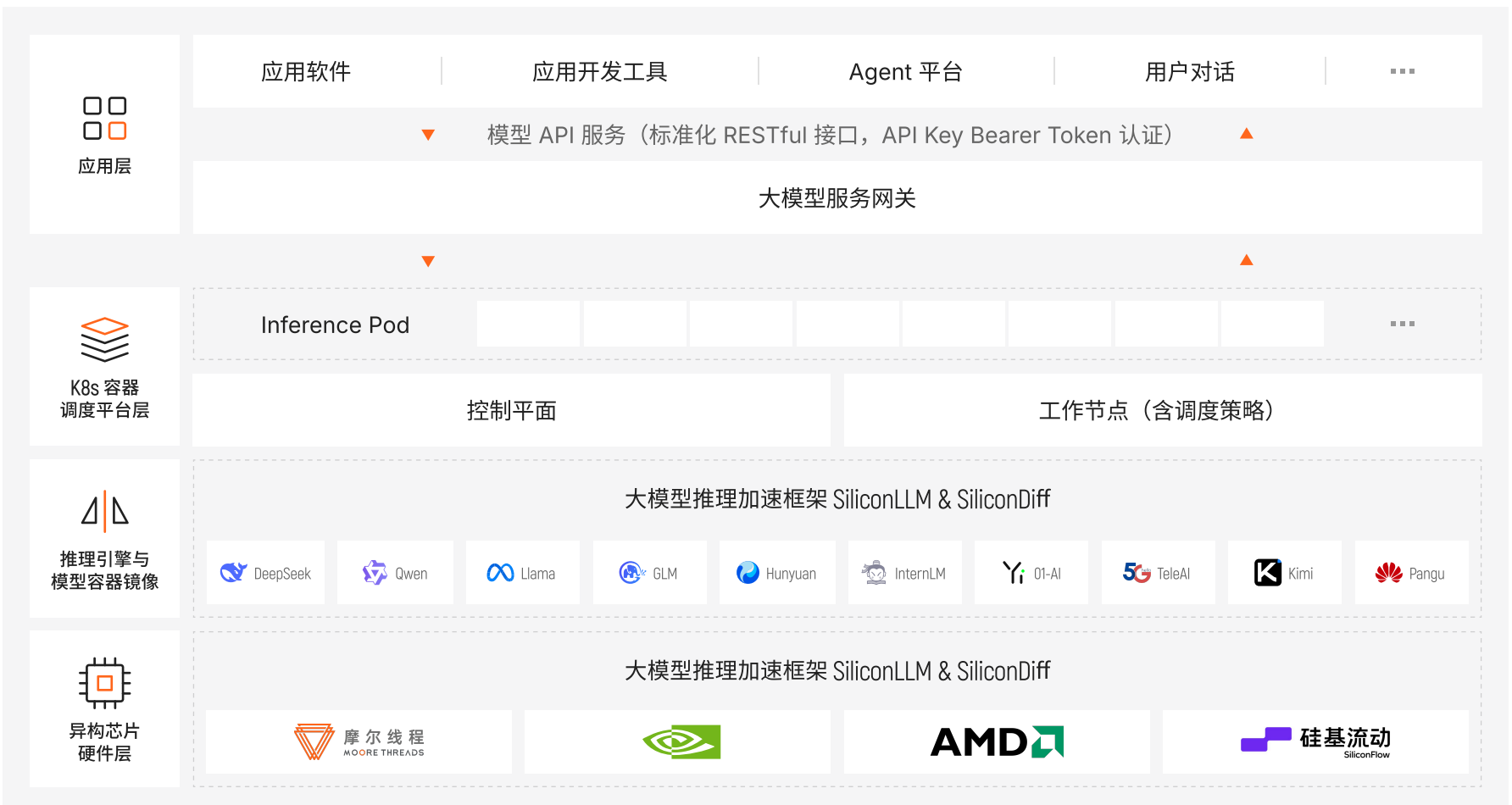

图 1 硅基流动大模型推理加速框架总体架构

3.4.2 核心技术亮点

针对异构集群的 PD 分离场景，SiliconLLM 框架在以下维度形成关键技术能力：

统一硬件抽象层（HAL）：通用 IR 兼容不同芯片指令集，标准化接入接口使 MTT S5000 与 HXXX 在同一引擎内等同调度，硬件迭代成本降低约 50%；

芯片原生高性能算子库：摩尔线程自研 Attention 算子融合 FlashAttention/MLA 与硬件指令优化，MOE 算子原生支持动态专家路由，计算效率逼近理论峰值；

自适应无损模型量化：硬件无关的 QIR 解耦量化算法与硬件实现，使 FP8 在异构芯片上推理性能与精度对齐原生 FP16。

图 2 SiliconLLM 四层联合优化框架示意

3.4.3 框架在本异构 PD 分离方案中的关键价值

针对本 MTT S5000 + HXXX 异构 PD 分离场景，SiliconLLM 框架的关键价值在于：

统一引擎消除异构壁垒：通过 HAL 抽象，P 侧 MTT S5000 与 D 侧 HXXX 在同一引擎内等同调度，模型代码与配置两侧完全一致；QIR 量化中间表示使两侧 FP8 具有等价数值语义，KV Cache 跨芯片传输不发生精度退化，从框架层面保证数据一致性；

混合并行策略灵活适配：针对 MTT S5000 Prefill 高算力与 HXXX Decode 高带宽特性，框架可为两侧分别定制并行策略，最大化异构资源利用率。

3.5 KV Cache 跨芯片通信方案

Prefill 节点（MTT S5000）与 Decode 节点（HXXX）之间通过 **Mooncake RDMA** 高速网络传输 KVCache，实现数据的高效流转：

通信机制：MTT S5000 完成 Prefill 计算后，将生成的 KV Cache 通过 Mooncake RDMA 直接发送至 HXXX 的显存中，供后续 Decode 阶段使用；

兼容性验证：双方已完成网卡层面的兼容性测试，预期可正常通信；后续将持续验证在不同网络拓扑和负载条件下的传输性能与长期可靠性；

性能预期：RDMA 零拷贝机制可大幅降低 CPU 参与和传输延迟，为端到端推理时延的稳定提供保障。

该通信方案是保证 PD 分离异构集群高效协同的关键一环，其性能与可靠性直接决定了全链路服务质量。

四、性能测试与验证

4.1 测试环境

测试模型：Kimi K2.6（FP8 精度）

推理引擎：硅基流动推理引擎（支持 FP8 KV Cache）

部署架构：4P2D 异构集群（4×MTT S5000 Prefill + 2×HXXX Decode）

对标基准：2P2D 纯 HXXX 集群（4×HXXX，同样 FP8 精度）

通信方式：Mooncake RDMA 传输 KV Cache

4.2 Prefill 单卡吞吐量化对标

基于 21 组全场景对照测试，结果如下：

场景	MTT S5000 平均吞吐	HXXX 平均吞吐	MTT S5000/HXXX 占比
24K+1K（短文本场景）	979 Tokens/s	1,809 Tokens/s	54%
64K~128K+1K（长文本场景）	1,013 Tokens/s	2,036 Tokens/s	50%
128K Input+1K Output/16 并发（极限场景）	2,047 Tokens/s	4,493 Tokens/s	46%

在 Prefill 阶段的单卡吞吐维度，MTT S5000 达到 HXXX 的 46%~54%。结合 **2:1 的部署比例**（4 台 MTT S5000 vs 2 台 HXXX），Prefill 侧总算力得到有效补足，且 KV Cache FP8 压缩使长上下文场景下显存占用大幅降低，有力支撑了超长上下文处理。

4.3 全流程服务质量（QoS）时延分析

指标	Prefill 阶段	Decode 阶段
TTFT（首 Token 时延）	1~6 秒（常规并发）；峰值 14.1 秒（16 并发 /128K 输入）	符合长文本处理预期
TPOT（每 Token 输出时延）	0.011~0.020 秒	最大并发下无剧烈抖动
TPS（系统吞吐）	90.91~40.00 Tokens/s	随并发增加呈线性平缓下降

全链路时延指标表现**稳定，无异常波动**，生成流畅度与用户体验一致，满足在线服务 SLA 要求。 Mooncake RDMA 传输 KV Cache 的时延开销控制在合理范围内，未对 TTFT/TPOT 产生可见影响。

4.4 核心结论

本方案验证了由 **4 台 MTT S5000 + 2 台 HXXX 组成的 4P2D** 异构算力集群，在处理长序列大模型推理任务时，可有效对标由 **4 台 HXXX 组成的 2P2D 算力集群使用**。

在异构算力 PD 分离部署方案下，MTT S5000 在 Prefill 阶段以 2:1 比例等效替代 HXXX，在算力总量、响应速度、生成流畅度等维度可全面对标纯 HXXX 集群，全链路时延指标表现一致。FP8 精度对齐与 KV Cache 高效通信，为跨芯片数据一致性提供了可靠保障。

五、工程化价值与生产落地

5.1 高品质 Token 工厂的体系化能力

真正可规模化的 AI 基础设施，必须同时具备技术优化能力与持续运营能力。本方案构建了完整的高品质 Token 工厂工程化能力体系：

能力维度	具体内容
模型统一接入	支持多种模型协议与格式，统一接入标准，降低集成成本
异构 PD 分离	按阶段分配最适合的算力资源，提升整体吞吐，兼顾低时延与资源利用率
KV Cache 管理	高效的 KV Cache 分配、复用与回收，显著降低延迟与显存占用
弹性伸缩	支持基于 QPS、延迟等指标的自动扩缩容，灵活应对流量波动
限流与熔断	多维度限流与熔断保护，保障系统稳定与可用性
监控告警与 SLO	全链路监控与多维告警，以 SLO 驱动稳定性持续提升
计量计费与运营分析	精确计量与用量分析，支持多维度成本与运营决策

5.2 生产落地可行性

全链路时延（TTFT/TPOT）及极限抗压能力满足**在线服务 SLA 要求**，可投入规模化生产环境；

TPOT 在最大并发下未出现剧烈抖动，生成流畅稳定；

Mooncake RDMA 通信经初步验证可正常运作，为生产级部署奠定网络基础。

六、总结与展望

6.1 核心成果

本白皮书系统阐述了摩尔线程与硅基流动基于 MTT S5000 与国际高端 GPU 构建的 PD 分离异构算力混部实践。核心成果包括：

技术验证：首次在大规模生产级环境下验证了国产 GPU（MTT S5000）与国际高端 GPU（HXXX）的异构 PD 分离协同推理，4P2D 异构集群全面对标 2P2D 纯 HXXX 集群。Prefill 和 Decode 均采用 FP8 精度，格式一致，确保数据正确性。

性能对标：在算力总量、响应速度（TTFT/TPOT）、生成流畅度等全维度指标上实现与纯 HXXX 集群的一致性表现。KV Cache FP8 与 Mooncake RDMA 通信方案为长上下文推理提供了高效支撑；

工程化落地：全链路时延满足在线服务 SLA 要求，具备规模化生产部署条件；通信兼容性验证通过，可实现大规模生产可行性；

成本优化：在 Prefill 阶段以 2:1 比例实现 MTT S5000 对 HXXX 的等效替代，显著降低推理成本及安全可控风险。

6.2 行业意义

为国产算力规模化部署提供可行路径：通过异构混部，国产算力可在不牺牲用户体验的前提下逐步替代国际高端产品；

推动异构算力标准建设：本方案为行业提供了可量化的容量规划标准与部署参考，特别是在 KV Cache 通信、精度对齐、仿真调度等方面形成了可复用的技术规范；

支撑全国一体化算力网建设：异构算力的高效协同是实现算力互联互通、资源灵活调度的关键技术基础；

构建安全可控的 AI 产业底座：形成模型、算力、平台、运营协同的新型基础设施体系，推动 AI 应用从试点走向规模化生产。

AI 基础设施的竞争，不只是算力规模的竞争，更是 Token 生产效率、系统工程能力与产业协同能力的竞争。本方案为构建安全、可控、可持续的 AI 产业底座提供了坚实的技术基础与工程实践。