文件编号	FIL-C-20251029-903394-RDR
上级文件编号	RD-IC-2024-01-1-0

file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/数学建模确定最佳部署方案流程.docx

目录

目录 1

修订记录 2

第一章 总结 0

第二章 Checklist 0

第 1 条 部署方案确定流程 Checklist 0

第三章 附则 2

第 2 条 解释权 2

修订记录

版本号	修订日期	作者	修订内容
1.0	2025.10.29	王翔煜	首次制作

file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/总则.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/模型结构介绍.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/模型相关参数定义.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/分布式通信机制详解.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/需要给定的硬件资源参数与模型参数.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/性能指标计算.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/优化目标与约束条件.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/工具接口说明.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/实战案例.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/扩展案例分析.docx
file:///Users/lixiang/Documents/工作/code/CrossRing/docs/Tier6+/快速参考.docx

第一章　总结

本指南通过 128 芯 DeepSeek-V3 部署案例，详细演示了从硬件定义到最优方案输出的完整流程：

1) **定义芯片架构**：配置算力、带宽、拓扑参数
2) **确定搜索空间**：基于策略决策（TBO、MOE_TP、MLA_TP）生成候选配置
3) **批量性能评估**：使用工具快速评估所有配置
4) **筛选与优选**：应用约束条件，选择TPS per Chip最大的配置
5) **输出结果验证**：保存结果，分析瓶颈，验证精度

这套方法论可以推广到任意 Transformer 模型和任意芯片的部署评估。

第二章　Checklist

第1条　部署方案确定流程 Checklist

Table 1. LLM 部署方案确定流程

Cat.	编号	Check 项目	Check 方法与参考值	Check 结果	PR & 时间
参数准备	1	模型参数完整性	检查模型配置文件包含所有必要参数（层数、维度、专家数等） 参考值：完整	是/否	方案设计者&开始前
	2	芯片参数准确性	确认算力、DRAM、C2C 带宽等参数与芯片规格一致 参考值：一致	是/否	方案设计者&开始前
	3	性能约束明确	明确 TPS per Batch 最小值、FTL 最大值等约束	是/否	方案设计者&开始前

			参考值：明确		
	4	芯片数量确定	确定可用的芯片总数 参考值：确定	是/否	方案设计者&开始前
配置空间	1	TP-PP 组合生成	生成所有满足 TP×PP=num_chips 的组合 参考值：完整	是/否	方案设计者&配置阶段
	2	Batch 候选值合理	Batch 取值范围覆盖常用场景（1-256） 参考值：合理	是/否	方案设计者&配置阶段
	3	DRAM 预检查	排除 DRAM 容量不足的配置 参考值：完成	是/否	方案设计者&配置阶段
性能建模	1	工具配置正确	CHIPMathica 配置文件参数正确 参考值：正确	是/否	方案设计者&建模阶段
	2	批量扫描完成	所有配置均完成性能建模 参考值：100%	是/否	方案设计者&建模阶段
	3	结果数据完整	每个配置都有 TPS、DRAM 等关键指标 参考值：完整	是/否	方案设计者&建模阶段
方案筛选	1	TPS per Batch 约束	筛选 TPS per Batch >= 阈值的配置 参考值：有可行解	是/否	方案设计者&筛选阶段
	2	DRAM 容量约束	确认所有方案 DRAM 占用<95%容量 参考值：满足	是/否	方案设计者&筛选阶段
	3	最优方案选择	选择 TPS per Chip 最大的配置 参考值：明确	是/否	方案设计者&筛选阶段
	4	对比表生成	生成 Top N 配置对比表 参考值：完成	是/否	方案设计者&筛选阶段
结果分析	1	瓶颈分析	识别计算/访存/通信瓶颈 参考值：完成	是/否	方案设计者&分析阶段
	2	Roofline 分析	分析配置的计算强度与 Ridge Point 关系 参考值：完成	是/否	方案设计者&分析阶段
	3	敏感性分析	分析 TP、PP、Batch 对性能的影响	是/否	方案设计者

			参考值：完成		&分析阶段
方案输出	1	部署方案报告	生成包含配置和性能指标的完整报告 参考值：完成	是/否	方案设计者&输出阶段
	2	Excel 对比表	导出所有配置的详细对比数据 参考值：完成	是/否	方案设计者&输出阶段
	3	可视化图表	生成敏感性分析等可视化图表 参考值：完成	是/否	方案设计者&输出阶段
审核验证	1	方案合理性	专家审核方案的合理性和可行性 参考值：通过	是/否	审核人员&审核阶段
	2	实际验证	在实际硬件上验证方案性能（如可行） 参考值：误差<20%	是/否	测试人员&验证阶段

第三章　附则

第2条　解释权

TPU PMT 对本指南具有最终解释权。