第一章　总则

第1条　目的

本文档旨在系统性地介绍如何利用数学建模工具，根据芯片的硬件参数（算力、DRAM 带宽、C2C 带宽等），结合通讯拓扑设计，为大语言模型确定最佳的部署方案。通过优化 Tensor Parallelism（TP）、Pipeline Parallelism（PP）、Batch Size 等部署参数，在满足性能约束（如 TPS per Batch 达标）的前提下，实现 TPS per Chip 的最大化，从而充分发挥芯片的计算能力和成本效益。

第2条　适用范围

本文档适用于使用 CHIPMathica 工具进行基于 Transformer 架构的大语言模型部署方案设计的芯片架构师、算法工程师和系统设计人员，特别适用于需要在给定芯片资源下优化模型推理性能的场景。

第3条　名词定义

序号	术语	定义说明
1.	TPS	Token Per Second，每秒生成的 Token 数量，衡量推理吞吐量
2.	TPS per Batch	单个 Batch 的 TPS，体现用户体验（值越大响应越快）
3.	TPS per Chip	单个芯片的 TPS，体现芯片利用率和成本效益
4.	TP (Tensor Parallelism)	张量并行度，将模型参数在多个芯片上切分
5.	PP (Pipeline Parallelism)	流水线并行度，将模型层在多个芯片上切分
6.	EP (Expert Parallelism)	专家并行度，用于 MoE 模型的专家分布
7.	FTL (First Token Latency)	首 Token 延迟，从输入到输出第一个 Token 的时间
8.	MLA (Multi-head Latent Attention)	多头潜在注意力，DeepSeek-V3 的注意力机制
9.	SLO (Service Level Objective)	服务质量目标，如 TPS per Batch ≥ 10
10.	Prefill 阶段	处理输入提示词的阶段，计算量大，batch 并行度高
11.	Decode 阶段	逐个生成输出 Token 的阶段，访存密集，对延迟敏感
12.	AllReduce Time	张量并行中的通信同步时间
13.	Computing Time	模型推理的计算时间
14.	Xfer Time	数据搬运时间（权重和 KV Cache 的 DRAM 访问）
15.	Min DRAM Size	部署模型所需的最小 DRAM 容量