第一章　优化目标与约束条件

第1条　问题定义

核心问题：给定硬件配置和模型架构，如何选择最优的并行策略和 Batch Size，使得在满足服务质量（SLO）的前提下，单芯片吞吐量最大化。

评估对象：目前以 DeepSeek-V3 (671B) 为基准模型进行评估

第2条　优化目标

1）主要优化目标

$$\max \mathrm{TPS}_{\text{per chip}}$$

在满足所有约束条件的前提下，最大化单芯片的吞吐量（TPS per Chip）。

2）分阶段优化目标

Prefill 阶段：

$$\begin{cases} \max & \mathrm{Throughput}_{\text{per chip}} \\ \text{s.t.} & \mathrm{FTL} \le T_{\mathrm{FTL}}^{\max} \end{cases}$$

在首Token延迟（FTL）满足SLO的前提下，最大化单芯片的Prefill吞吐量

Decode 阶段：

$$\begin{cases} \max & \mathrm{TPS}_{\text{per chip}} \\ \text{s.t.} & \mathrm{TPS}_{\text{per batch}} \ge T_{\mathrm{TPS}}^{\min} \end{cases}$$

在单Batch吞吐量（TPS per Batch）满足SLO的前提下，最大化单芯片的总吞吐量

第3条　约束条件

1）性能约束（SLO）

a）Decode阶段的**单Batch吞吐量约束**（用户体验）：

$$\mathrm{TPS}_{\text{per batch}} = \frac{1}{\mathrm{DecodeTime\ (s)}} \ge 10$$

即：

$$\mathrm{DecodeTime} \le 100 \ \text{ms}$$

b）Prefill阶段的**首Token延迟约束**：公式表示：

$$\mathrm{FTL} = T_{\mathrm{prefill}} \le T_{\mathrm{FTL}}^{\max}$$

输入长度	FTL 上限
4K tokens	≤ 3s
8K tokens	≤ 5s

2）硬件约束

a) DRAM容量约束

每个芯片的显存占用不能超过芯片DRAM容量：

$$\mathrm{DRAM}_{\mathrm{required}} \leq \mathrm{DRAM}_{\mathrm{chip}}$$

其中显存需求包括：

$$\mathrm{DRAM}_{\mathrm{required}} = \frac{\mathrm{Weight}_{\mathrm{MLA}}}{\mathrm{MLA}_{\mathrm{TP}} \times \mathrm{MLA}_{\mathrm{DP}}} + \frac{\mathrm{Weight}_{\mathrm{MoE}}}{\mathrm{MOE}_{\mathrm{TP}} \times \mathrm{MOE}_{\mathrm{EP}}} + \frac{\mathrm{KVCache} \times \mathrm{Batch}}{\mathrm{MLA}_{\mathrm{TP}} \times \mathrm{MLA}_{\mathrm{DP}}}$$

参数说明：

$\mathrm{Weight}_{\mathrm{MLA}}$：MLA部分权重大小

$\mathrm{Weight}_{\mathrm{MoE}}$：MoE部分权重大小

KVCache：单个请求的KV Cache大小

Batch：总Batch大小

b) 芯片数量约束（并行度一致性）

$$\mathrm{MLA}_{\mathrm{TP}} \times \mathrm{MLA}_{\mathrm{DP}} = \mathrm{MOE}_{\mathrm{TP}} \times \mathrm{MOE}_{\mathrm{EP}} = N_{\mathrm{chips}}$$

说明：MLA和MoE部分可以采用不同的并行策略，但总芯片数必须一致。

c) 网络拓扑约束

C2C带宽（Chip-to-Chip）：影响TP通信性能

All-to-All带宽：影响EP通信性能

并行度需与拓扑匹配：如8卡机器不能设置TP=16

第4条　可优化变量

变量	可选值	说明	约束
Batch	1, 2, 4, 8, 10, 12, 16, 32, 64, 128, 256, 512, 1024, 1280	总 Batch 大小	Batch ⩾ 1
MLA_TP	1, 2, 4, 8, 16, 32	MLA 的张量并行度	2 的幂次
MLA_DP	1, 2, 4, 8, 16, 32, 64	MLA 的数据并行度	MLA_TP × MLA_DP = N_chips
MOE_TP	1, 2, 4, 8	MoE 的张量并行度	一般 ⩽ 8
MOE_EP	16, 32, 64, 128, 256	专家并行度	MOE_TP × MOE_EP = N_chips
SP	1, 2, 4	序列并行度	SP | MLA_TP

1) 变量说明

a) Tensor Parallelism (TP)

作用：将模型权重按列/行切分到多个芯片

收益：降低单芯片显存占用和计算量

代价：增加AllReduce通信开销

建议：MLA_TP = 2-8，MOE_TP = 1-4

b) Data Parallelism (DP)

作用：不同芯片处理不同的数据样本

收益：提升总吞吐量，无通信开销

代价：不降低单样本延迟

建议：在满足SLO后，优先增大DP

c) Expert Parallelism (EP)

作用：将MoE的专家分配到不同芯片

收益：降低单芯片显存占用

代价：增加All-to-All通信（Dispatch/Combine）

建议：EP = 16-256，取决于专家数量和网络带宽

d) Sequence Parallelism (SP)

作用：将序列按长度切分到多个芯片

收益：降低Prefill阶段的激活值显存占用

代价：增加AllGather/ReduceScatter通信

建议：主要用于长序列Prefill（如8K+）

第5条　评估流程

```
1. 遍历所有可行的配置组合
   ├── 选择 (Batch, MLA_TP, MLA_DP, MOE_TP, MOE_EP, SP)
   └── 检查是否满足芯片数量约束

2. 计算性能指标
   ├── 显存占用：DRAM_required
   ├── 计算时间：T_comp
   ├── 访存时间：T_mem
   ├── 通信时间：T_comm
   └── 总时间：T_total = max(T_comp, T_mem) + T_comm

3. 检查约束条件
   ├── DRAM_required ≤ DRAM_chip?
   ├── TPS_per_batch ≥ 10? （Decode）
   └── FTL ≤ T_max? （Prefill）

4. 计算优化目标
   └── TPS_per_chip = Batch / T_decode

5. 选择最优方案
   └── max TPS_per_chip（满足所有约束）
```

第6条　评估范围

1) **模型：**

a) ✅ DeepSeek-V3 (671B, MoE)

b) ⏸️ Dense模型（暂不考虑，后续扩展）

2) **阶段：**

a) ✅ Prefill阶段（4K/8K Input）

b) ✅ Decode阶段（4K Input + 生成）

3) **优化技术：**

a) ✅ TP/DP/EP/SP 并行策略

b) ✅ Batch Size 优化

c) ⏸️ TBO（Two-Batch Overlap，可选）

d) ❌ Continuous Batching（不考虑动态调度）

e) ❌ Speculative Decoding（不考虑推测解码）

第7条　建模假设

1) **PD节点独立：**

a) 不考虑Prefill与Decode节点的吞吐量匹配

b) 实际部署可通过调整PD节点数量比例实现

2) **KV Cache传输：**

a) 不考虑PD节点之间的KV Cache传输或重算

b) 假设无KV Cache专属硬件资源

3) **负载均衡：**

a) 假设MoE路由完全均衡（理想情况）

b) 不考虑专家负载不均导致的性能下降

4) **网络稳定：**

a) 假设网络带宽恒定，无拥塞

b) 不考虑网络抖动或故障

第8条　输出最优方案

1) **Decode阶段最优方案：**

参数	最优值	说明

Batch	?	使 TPS per Chip 最大
MLA_TP	?	平衡通信与计算
MLA_DP	?	满足芯片数约束
MOE_TP	?	通常为 1 或 2
MOE_EP	?	根据专家数和带宽
TPS per Batch	⩾ 10	满足 SLO
TPS per Chip	**最大值**	**优化目标**
DRAM per Chip	⩽ DRAM_chip	满足硬件约束

2) **Prefill阶段最优方案：**

参数	**最优值**	**说明**
MLA_TP	?	降低计算时间
SP	?	处理长序列
FTL	⩽ 3s (4K) / ⩽ 5s (8K)	满足 SLO
Throughput per Chip	**最大值**	**优化目标**

3) 做出性能分析**瓶颈分析：**

Decode阶段：计算受限 vs 访存受限 vs 通信受限

Prefill阶段：计算受限 vs 通信受限**敏感性分析：成本效益：**

a) Batch Size对吞吐量和延迟的影响

b) TP对通信开销的影响

c) EP对All-to-All通信的影响

d) 单芯片TPS（成本效率）

e) 总TPS（集群吞吐）

f) 性价比：TPS /（芯片数 × 芯片单价）