第一章　扩展案例分析

第1条　案例A：256芯方案（高吞吐场景）

1） **配置：**

```
configs_256 = generate_configs(
    num_chips=256,
    tp_candidates=[4],
    batch_candidates=[10, 12, 16]
)
```

2） **最优方案：**

参数	256 芯	128 芯	提升
配置	TP4+DP64+EP256, B12	TP4+DP32+EP128, B10	EP 翻倍
TPS/Chip	**138**	108	**+27.8%**
总吞吐量	**35,328**TPS	13,824 TPS	**+155.6%**
TPS/Batch	12.6	10.3	+22.3%
最佳 Batch	12	10	+2

3） **关键发现：**

a） ✅**EP256显著优于EP128：**MoE通信减少，性能提升28%

b） ✅**支持更大Batch：**Batch=12仍满足TPS/Batch≥10

c） ✅**总吞吐量翻倍：**从13.8K提升到35.3K TPS

d） ✅**单芯效率也提升：**不仅总量大，单芯效率也提升

4） **适用场景：**超大规模推理服务（日请求量千万级）

第2条　案例B：64芯方案（不推荐）

1） **配置扫描结果：**

Batch	MLA_TP	TPS/Batch	TPS/Chip	DRAM(GB)	满足 SLO?
4	4	10.0	39.8	14.8	✅ 刚好满足
6	4	9.3	55.5	14.8	❌ 不满足
8	4	8.6	69.1	14.8	❌ 不满足

2） **问题分析：**

a） 仅Batch=4能满足SLO，但TPS/Chip仅39.8（**128芯的37%**）

b） Batch做不大，无法充分利用硬件

c） EP=64相比EP=128，MoE通信开销显著增加

3） **建议：**

a) ❌ 不使用64芯作为单一Decode节点

b) ✅ 部署2个独立32芯实例

c) ✅ 或与其他64芯组合成128芯

第3条　案例C：16芯Prefill方案

1) **配置与结果：**

Batch	MLA_TP	MLA_DP	MOE_EP	TPS/Chip	FTL(ms)	满足FTL≤3s	评价
4	**4**	**4**	**16**	**462**	**2213**	✅	**最优⭐**
4	4	4	8	443	2621	✅	次优
5	16	1	16	329	3896	❌	不满足FTL
8	2	8	16	407	2517	✅	可选

2) **Prefill方案分析：**

a) ✅**最优配置：**TP4+SP4+EP16，Batch4

b) ✅**性能：**FTL=2.2s < 3s，单芯吞吐462 TPS/Chip

c) ✅**总吞吐量：**7,392 TPS（16芯）

d) 📌**关键特点：**

Prefill适合较小Batch（4-8），因为Input长度大
TP16虽然通信少，但单芯计算效率低，FTL反而更长
SP（Sequence Parallel）在Prefill阶段很有效