第一章　快速参考

第1条　关键配置速查表

场景	芯片数	推荐配置	性能预期	注意事项
Decode 标准部署	128	TP4+DP32+EP128, B10	TPS/Chip=108	性价比最佳⭐
Decode 高吞吐	256	TP4+DP64+EP256, B12	TPS/Chip=138 (+28%)	总吞吐翻倍
Prefill	16-32	TP4+SP4+EP16/32, B4-8	FTL=2.2-2.5s	单芯效率 400%+
小规模部署	64	❌ 不推荐	TPS/Chip=40 (-63%)	建议用 128 芯

1) **核心策略：**

   a) ✅ TBO：仅高带宽芯片+大Batch场景（SG2260E不需要）

   b) ✅ MOE_TP：固定为1（仅做EP）

   c) ✅ MLA_TP：优先选择4（平衡通信与计算）

   d) ✅ EP：越大越好（256 > 128 > 64）

   e) ✅ Batch：满足SLO前提下越大越好

第2条　常见问题FAQ

1) Q1: 为什么TPS per Chip随Batch增大而增大，但有SLO约束？
   a) Batch越大 → 计算密集度高 → TPS per Chip高
   b) 但Batch越大 → TPOT越长 → TPS per Batch低
   c) 需要平衡用户体验（TPS per Batch）和芯片利用率（TPS per Chip）
2) Q2: 为什么MFU只有12%？
   a) DeepSeek-V3是访存密集型模型（大量权重和KV Cache访问）
   b) DRAM带宽273GB/s相对计算能力128T是瓶颈
   c) 这是正常现象，不代表芯片性能差
3) Q3: 如何选择网络拓扑？
   a) 如果TP Group间能直连（All2All）：设置`inter_bw=64GB/s`
   b) 如果TP Group间只能PCIe：设置`inter_bw=16GB/s`
   c) 性能差异约10-15%
4) Q4: 数学建模结果可信吗？
   a) **相对比较可信**：配置排序准确（TP4优于TP2/TP8等）

b) **绝对值有误差：**实际TPS可能有5-15%的偏差

c) **趋势分析准确：**EP越大性能越好等结论可靠

d) **建议：**最终方案在实际硬件上验证

5) Q5: 如何处理Prefill和Decode节点的吞吐量不匹配？

a) 通过调整PD节点数量比例实现匹配

b) 例如：Prefill节点数 / Decode节点数 = Decode吞吐 / Prefill吞吐

c) 暂不考虑PD节点之间的KV Cache传输