第一章　需要给定的硬件资源参数与模型参数

第1条　芯片算力参数

1) **FLOPS（Floating Point Operations Per Second）**
   a) **定义**：单芯片的理论峰值算力
   b) **单位**：FLOPS（通常以TFLOPS表示，1 TFLOPS = 10^{12} FLOPS）
   c) **配置方式**：根据芯片规格确定，例如：Int8精度：128 TFLOPS；FP16精度：64 TFLOPS
   d) **在TPUArch中的参数**：`flops`
2) **核心数（core）**
   a) **定义**：单芯片的计算核心数量
   b) **配置示例**：8
   c) **在TPUArch中的参数**：`core`

第2条　内存参数

1) **DRAM容量**
   a) **定义**：单芯片的DRAM存储容量
   b) **单位**：GB
   c) **配置示例**：64GB
   d) **作用**：限制模型权重和KV Cache的存储
2) **DRAM带宽**
   a) **定义**：DRAM的访问带宽
   b) **单位**：GB/s（字节每秒）
   c) **配置示例**：273 GB/s
   d) **带宽利用率**：实际使用时需要考虑带宽利用率（通常为0.893）
   e) **在TPUArch中的参数**：`dram_bw = 理论带宽 × 利用率`
3) **L2M（L2 Memory）参数**
   a) **容量**：通常较小，如16MB
   b) **带宽**：单向带宽：512 GB/s；双向带宽：1 TB/s
   c) **作用**：作为高速缓存，减少DRAM访问

第3条　互联参数

互联拓扑目前没有参数化的方案，只能根据评估案例组织。

参数名称	符号	定义	单位	典型值
单向带宽	`intrabw`（节点内）	单方向最大传输速率	GB/s	节点内：64-900

	`inter_bw`（节点间）			节点间：12.5-50
带宽利用率	`bwurate`	实际可达带宽占理论带宽比例	无量纲	0.8-0.95
启动延迟	`startlat`	通信启动的固定开销	μs	0.5-2
同步延迟	`synclat`	多节点同步的固定开销	μs	0.5-2
链路延迟	`linkdelay`	跨节点链路传输延迟	μs	0.28-5
CPU 指令延迟	`cpufetch`	CPU 发起通信指令的延迟	μs	0-1

第4条　SG2260E硬件参数示例

以 SG2260E 为例的完整硬件参数配置：

参数类型	参数名称	数值	说明
算力	Int8 FLOPS	128 TFLOPS	单芯片峰值算力
算力	FP16 FLOPS	64 TFLOPS	单芯片峰值算力
内存	DRAM 容量	64 GB	单芯片 DRAM 容量
内存	DRAM 带宽	273 GB/s	理论带宽
内存	DRAM 带宽利用率	0.893	实际可用带宽比例
缓存	L2M 容量	16 MB	L2 缓存大小
缓存	L2M 单向带宽	512 GB/s	L2 缓存单向访问带宽
缓存	L2M 双向带宽	1 TB/s	L2 缓存双向访问带宽
互联	TP Group 内带宽	64 GB/s	直连节点间带宽
互联	TP Group 间带宽	16 GB/s	PCIe Gen5 互联带宽
其他	核心数	8	单芯片计算核心数

第5条　模型参数

从`config.json`中提取以下参数：

1) 基础维度
   a) $L\#$**(Layers)**：61（`num_hidden_layers`）
   b) $H\#$**(Hidden Size)**：7168（`hidden_size`）
   c) $V\#$**(Vocab Size)**：129,280（`vocab_size`）
   d) I_{dense}**(Dense FFN Size)**：18,432（`intermediate_size`）
   e) I_{moe}**(MoE Expert Size)**：2,048（`moe_intermediate_size`）
2) MLA（Attention）参数
   a) R_q**(Q LoRA Rank)**：1,536（`q_lora_rank`）
   b) R_{kv}**(KV LoRA Rank)**：512（`kv_lora_rank`）
   c) d_{rope}**(RoPE Dim)**：64（`qk_rope_head_dim`）
   d) d_{nope}**(NoRoPE Dim)**：128（`qk_nope_head_dim`）
   e) d_v**(Value Dim)**：128（`v_head_dim`）

f) N_{heads}: 128 (`num_attention_heads`)

3) MoE 参数

a) N_r(Routed Experts): 256 (`n_routed_experts`)

b) K_r(Active Experts per Token): 8 (`num_experts_per_tok`)

c) N_s(Shared Experts): 1 (`n_shared_experts`)

d) L_{dense}(Dense Layers): 3 (`first_k_dense_replace`)

4) 量化与精度

a) **Weight Precision**: FP8 (1 Byte)

b) **Activation Precision**: BF16 (2 Bytes)