第一章　工具接口说明

第1条　核心类与接口

工具提供了三个核心类：

1）TPUArch - 芯片架构定义定义芯片的硬件参数：**参数说明：**

```
arch = TPUArch(
    core=8,                      # 单芯片核心数
    flops=128*1e12,              # 算力（FLOPS）
    dram_bw=273e9*0.893,         # DRAM 有效带宽（Byte/s）
    intra_bw=64e9,               # TP Group 内互联带宽（Byte/s）
    inter_bw=16e9                # TP Group 间互联带宽（Byte/s）
)
```

参数	含义	如何获取
`core`	单芯片核心数	查阅芯片规格表
`flops`	峰值算力	规格表，注意数据类型（FP8/BF16/FP32）
`dram_bw`	DRAM 有效带宽	理论带宽 × 利用率（0.7-0.9）
`intra_bw`	TP Group 内带宽	根据互联拓扑确定（如 All2All）
`inter_bw`	TP Group 间带宽	PCIe/网络带宽

2）Config - 部署配置定义并行策略和Batch大小：**约束条件：**

```
cfg = Config(
    bs=10,           # 单芯 Batch 大小
    tp=4,            # MLA_TP（MLA 部分的张量并行度）
    dp=32,           # MLA_DP（MLA 部分的数据并行度）
    moe_tp=1,        # MOE_TP（MoE 部分的张量并行度）
    ep=128           # MOE_EP（专家并行度）
)
```

a）`tp × dp = moe_tp × ep`（总芯片数必须一致）

b）总Batch =`bs × tp × dp`

3）DeepSeekModel - 性能评估创建模型并评估性能：**返回值：**

```
# 创建模型（Decode 阶段）
model = DeepSeekModel(arch, cfg, is_prefill=False)

# 运行评估
tps, flops, elaps, mfu, dram_occupy = model.eval(tbo=False)
```

变量	含义	单位
`tps`	总吞吐量（所有芯片）	TPS
`flops`	总浮点运算量	FLOPS

`elaps`	单次 Decode 耗时	微秒(μs)
`mfu`	模型 FLOPs 利用率	0-1
`dram_occupy`	单芯 DRAM 占用	Bytes