LLM-para

一、 Moe

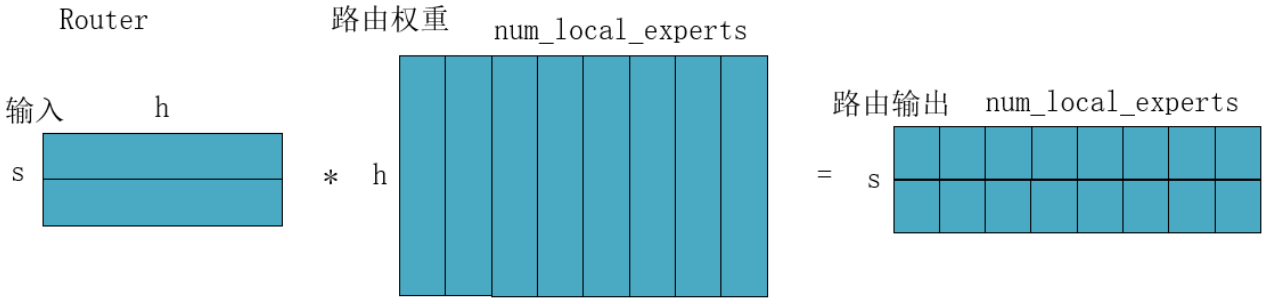
1. 启用MoE

```
use_moe = False
# 如果传入 moe 参数 则启用moe计算
if num_experts_per_tok is not None:
    use_moe = True
```

2. Prefill MoE 参数计算

分解	Input1	Input2	Output	运算量
MoE_Router	(b,s,h)	(h,num_local_experts)	(b,s,num_local_experts)	2bsh*num_local_experts
MoE_FFN1	(b, s*num_experts_per_tok , h)	(num_local_experts,h, intermediate_size*2)	(b,s*num_experts_per_tok ,intermediate_size * 2)	2bs* num_experts_per_tok * h * intermediate_size * 2 + bs * num_experts_per_tok * intermediate_size
MoE_FFN2	(b,s*num_experts_per_tok, intermediate_size)	(num_local_experts, intermediate_size, h)	(b, s*num_experts_per_tok , h)	2bs * num_experts_per_tok * intermediate_size * h

路由参数计算




```
moe_router_flops = b * seq * h * num_local_experts * 2
param_count = h * num_local_experts
add_row(phase, "Router", "(b,s,h)", f"(h,{num_local_experts})", f"(b,s,
{num_local_experts})",
        moe_router_flops, param_count,
        (b,seq,h), (h,num_local_experts), (b,seq,num_local_experts), a_bit, w_ffn)
```

FFN-1计算

假设prefill输入seq足够大，大到能够使用到所有expert。

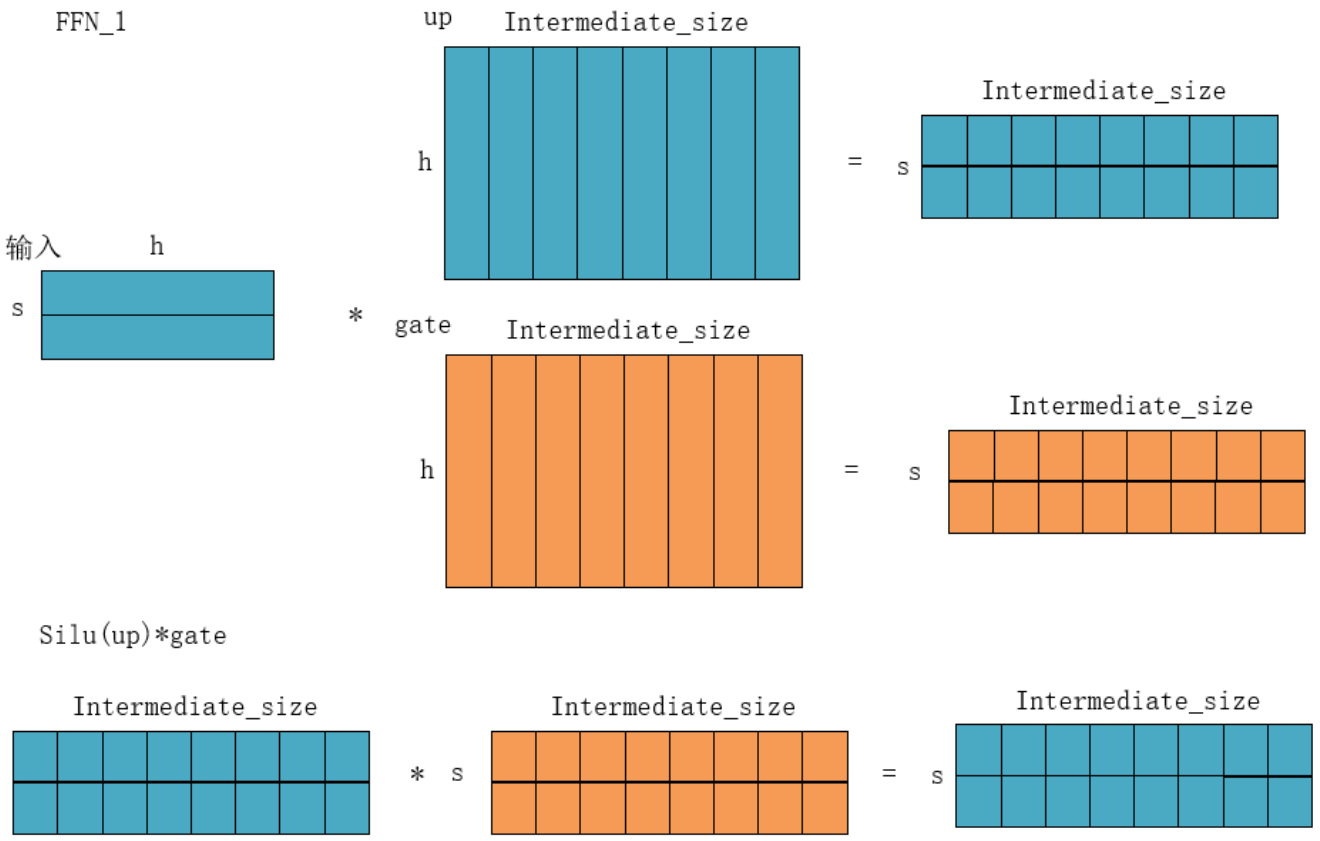
输入形状为**(b,s*num_experts_per_tok,h)**：对于每个token，需要送入到num_experts_per_tok个MLP进行计算，则相当于将每个token复制了num_experts_per_tok次，即有seq*num_experts_per_tok个token。

输入权重形状为**(num_local_experts,h,intermediate_size*2)**：使用到所有expert，且加载up和gate的权重。

所需参数量：num_experts_per_tok个 up+gate 权重矩阵

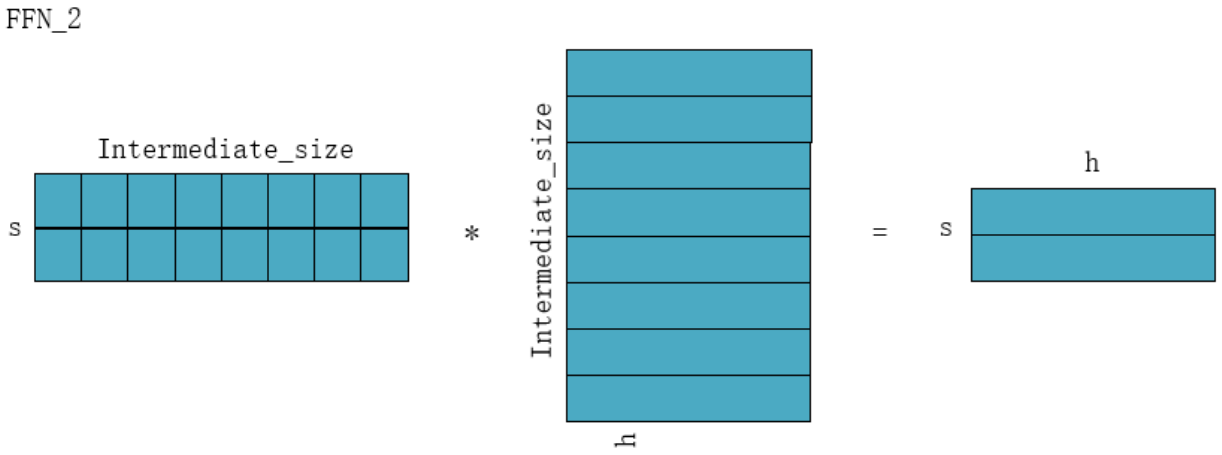
```python
# FFN-1(up + gate)(with Moe)
FFN_1_moe = b * seq * num_experts_per_tok * h * intermediate_size * 2 * 2  # * 两个矩阵乘法 * 每个token算num_experts_per_tok次
# 额外的逐元素乘法
FFN_1_moe += b * seq * num_experts_per_tok * intermediate_size  # (W₁x) ⊙ SiLU(W_gate*x)

param_count = h * intermediate_size * 2 * num_local_experts  # 两个权重矩阵 * 假设 prefill 用到了所有expert
add_row(phase, "FFN-1(with Moe)", f"(b,{s}*num_experts_per_tok,h)", f"(num_local_experts,h,{intermediate_size}*2)", f"(b,{s}*num_experts_per_tok,{intermediate_size}*2)",
        FFN_1_moe, param_count,
        (b, seq * num_experts_per_tok, h), (h, intermediate_size * 2 * num_local_experts), (b, seq * num_experts_per_tok, intermediate_size*2 ), a_bit, w_ffn)
```



FFN-2 计算

```
# FFN-2(with Moe)
FFN_2_moe = b * seq * intermediate_size * h * 2 * num_experts_per_tok
param_count = intermediate_size * h * num_local_experts
add_row(phase, "FFN-2(with Moe)", f"(b, {s}*num_experts_per_tok, {intermediate_size})",
f"(num_local_experts, {intermediate_size}, h)", f"(b,
{s}*num_experts_per_tok,h)",FFN_2_moe, param_count,
(b, seq * num_experts_per_tok, intermediate_size), (intermediate_size, h,
num_local_experts), (b, seq * num_experts_per_tok, h), a_bit, w_ffn)
```



3. Decode MoE参数计算

prefill 阶段的计算每次需要加载**num_local_experts**个MLP的参数。

decode 阶段的计算每次**只**需要加载**num_experts_per_tok**个MLP的参数。

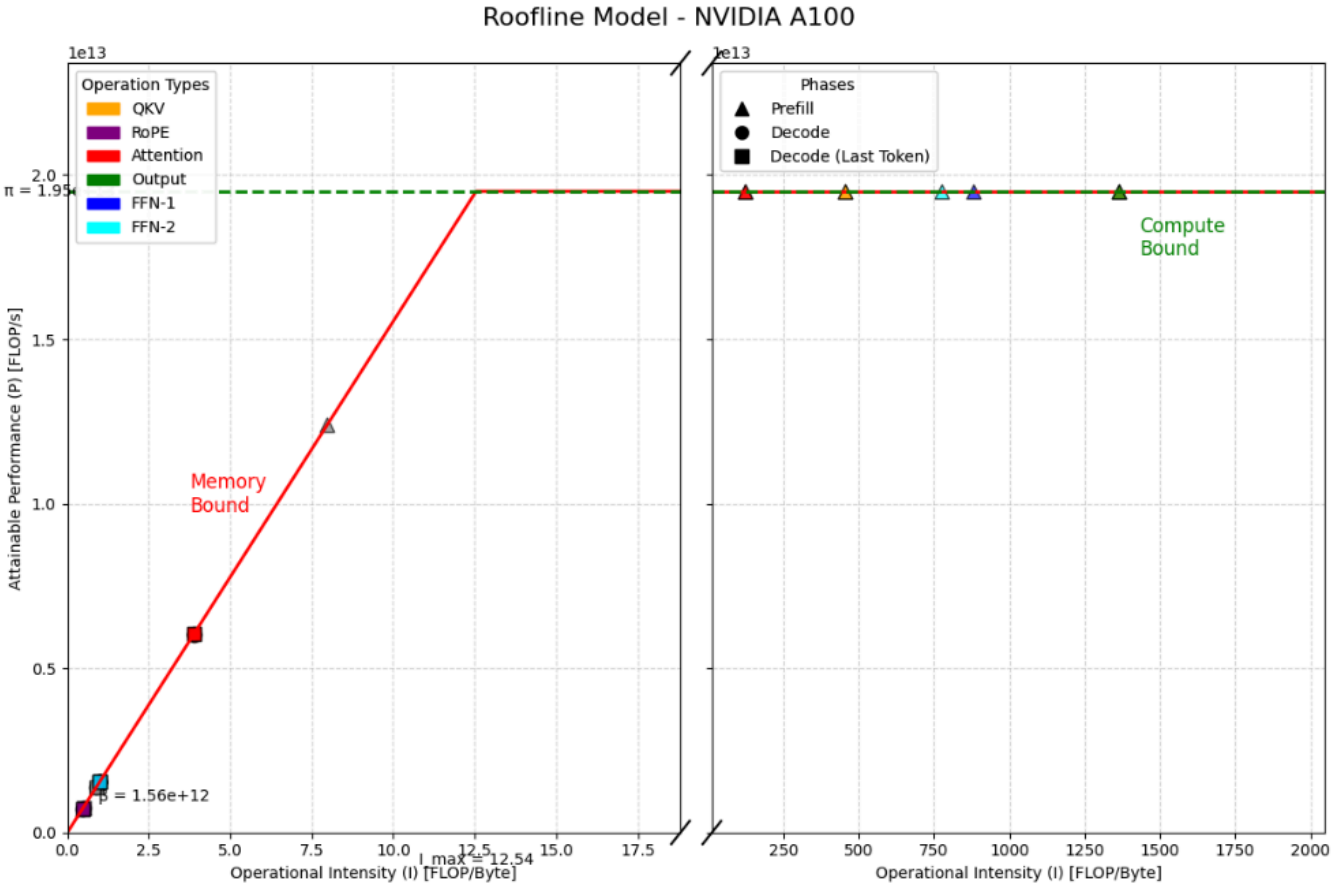
4. 计算结果对比（mixtral-8x7B）

增加Moe功能后，prefill阶段的FFN推理计算密度约为800；同一模型配置，去掉Moe功能，prefill阶段的FFN推理计算密度增加至约为2000。而decode阶段，对于FFN的推理其计算密度都为1。可得，增加Moe功能后，prefill阶段的FFN推理朝着更带宽受限的方向转变。

With_Moe仿真结果

```
num_experts_per_tok = 2
num_local_experts = 8
```

Phase	Operation	FLOPs	Param Count	Input1 Bytes	Input2 Bytes	Output Bytes	Total Bytes	Density (Op/Byte)
Prefill	xW_Q	137438953472.0	16777216.0	33554432.0	33554432.0	33554432.0	100663296.0	1365.33
Prefill	xW_K	34359738368.0	4194304.0	33554432.0	8388608.0	33554432.0	75497472.0	455.11
Prefill	xW_V	34359738368.0	4194304.0	33554432.0	8388608.0	33554432.0	75497472.0	455.11
Prefill	RoPE-Q	33554432.0	0.0	33554432.0	1048576.0	33554432.0	68157440.0	0.49
Prefill	RoPE-K	8388608.0	0.0	8388608.0	1048576.0	8388608.0	17825792.0	0.47
Prefill	Q × K^T	137438953472.0	0.0	33554432.0	8388608.0	1073741824.0	1115684864.0	123.19
Prefill	Attn × V	137438953472.0	0.0	1073741824.0	8388608.0	33554432.0	1115684864.0	123.19
Prefill	xW_O	137438953472.0	16777216.0	33554432.0	33554432.0	33554432.0	100663296.0	1365.33
Prefill	Router	268435456.0	32768.0	33554432.0	65536.0	65536.0	33685504.0	7.97
Prefill	FFN-1(with Moe)	1924262789120.0	939524096.0	67108864.0	1879048192.0	234881024.0	2181038080.0	882.27
Prefill	FFN-2(with Moe)	962072674304.0	469762048.0	234881024.0	939524096.0	67108864.0	1241513984.0	774.92
Decode	xW_Q	33554432.0	16777216.0	8192.0	33554432.0	8192.0	33570816.0	1.0
Decode	xW_K	8388608.0	4194304.0	8192.0	8388608.0	8192.0	8404992.0	1.0
Decode	xW_V	8388608.0	4194304.0	8192.0	8388608.0	8192.0	8404992.0	1.0
Decode	RoPE-Q	8192.0	0.0	8192.0	256.0	8192.0	16640.0	0.49
Decode	RoPE-K	2048.0	0.0	2048.0	256.0	2048.0	4352.0	0.47
Decode	Q × K^T	33554432.0	0.0	8192.0	8390656.0	262208.0	8661056.0	3.87
Decode	Attn × V	33562624.0	0.0	262208.0	8390656.0	8192.0	8661056.0	3.88
Decode	xW_O	33554432.0	16777216.0	8192.0	33554432.0	8192.0	33570816.0	1.0
Decode	Router	65536.0	32768.0	8192.0	65536.0	16.0	73744.0	0.89
Decode	FFN-1(with Moe)	469790720.0	234881024.0	16384.0	469762048.0	57344.0	469835776.0	1.0
Decode	FFN-2(with Moe)	234881024.0	117440512.0	57344.0	234881024.0	16384.0	234954752.0	1.0
Decode_Last	xW_Q	33554432.0	16777216.0	8192.0	33554432.0	8192.0	33570816.0	1.0
Decode_Last	xW_K	8388608.0	4194304.0	8192.0	8388608.0	8192.0	8404992.0	1.0
Decode_Last	xW_V	8388608.0	4194304.0	8192.0	8388608.0	8192.0	8404992.0	1.0
Decode_Last	RoPE-Q	8192.0	0.0	8192.0	256.0	8192.0	16640.0	0.49
Decode_Last	RoPE-K	2048.0	0.0	2048.0	256.0	2048.0	4352.0	0.47
Decode_Last	Q × K^T	67100672.0	0.0	8192.0	16777216.0	524288.0	17309696.0	3.88
Decode_Last	Attn × V	67108864.0	0.0	524288.0	16777216.0	8192.0	17309696.0	3.88
Decode_Last	xW_O	33554432.0	16777216.0	8192.0	33554432.0	8192.0	33570816.0	1.0
Decode_Last	Router	65536.0	32768.0	8192.0	65536.0	16.0	73744.0	0.89
Decode_Last	FFN-1(with Moe)	469790720.0	234881024.0	16384.0	469762048.0	57344.0	469835776.0	1.0
Decode_Last	FFN-2(with Moe)	234881024.0	117440512.0	57344.0	234881024.0	16384.0	234954752.0	1.0



Without_Moe仿真结果

```
num_experts_per_tok = None
num_local_experts = None
```

Phase	Operation	FLOPs	Param Count	Input1 Bytes	Input2 Bytes	Output Bytes	Total Bytes	Density (Op/Byte)
Prefill	xW_Q	137438953472.0	16777216.0	33554432.0	33554432.0	33554432.0	100663296.0	1365.33
Prefill	xW_K	34359738368.0	4194304.0	33554432.0	8388608.0	33554432.0	75497472.0	455.11
Prefill	xW_V	34359738368.0	4194304.0	33554432.0	8388608.0	33554432.0	75497472.0	455.11
Prefill	RoPE-Q	33554432.0	0.0	33554432.0	1048576.0	33554432.0	68157440.0	0.49
Prefill	RoPE-K	8388608.0	0.0	8388608.0	1048576.0	8388608.0	17825792.0	0.47
Prefill	Q × K^T	137438953472.0	0.0	33554432.0	8388608.0	1073741824.0	1115684864.0	123.19
Prefill	Attn × V	137438953472.0	0.0	1073741824.0	8388608.0	33554432.0	1115684864.0	123.19
Prefill	xW_O	137438953472.0	16777216.0	33554432.0	33554432.0	33554432.0	100663296.0	1365.33
Prefill	FFN-1 (with Gate)	962131394560.0	117440512.0	33554432.0	234881024.0	117440512.0	385875968.0	2493.37
Prefill	FFN-2	481036337152.0	58720256.0	117440512.0	117440512.0	33554432.0	268435456.0	1792.0
Decode	xW_Q	33554432.0	16777216.0	8192.0	33554432.0	8192.0	33570816.0	1.0
Decode	xW_K	8388608.0	4194304.0	8192.0	8388608.0	8192.0	8404992.0	1.0
Decode	xW_V	8388608.0	4194304.0	8192.0	8388608.0	8192.0	8404992.0	1.0
Decode	RoPE-Q	8192.0	0.0	8192.0	256.0	8192.0	16640.0	0.49
Decode	RoPE-K	2048.0	0.0	2048.0	256.0	2048.0	4352.0	0.47
Decode	Q × K^T	33554432.0	0.0	8192.0	8390656.0	262208.0	8661056.0	3.87
Decode	Attn × V	33562624.0	0.0	262208.0	8390656.0	8192.0	8661056.0	3.88
Decode	xW_O	33554432.0	16777216.0	8192.0	33554432.0	8192.0	33570816.0	1.0
Decode	FFN-1 (with Gate)	234895360.0	117440512.0	8192.0	234881024.0	28672.0	234917888.0	1.0
Decode	FFN-2	117440512.0	58720256.0	28672.0	117440512.0	8192.0	117477376.0	1.0
Decode_Last	xW_Q	33554432.0	16777216.0	8192.0	33554432.0	8192.0	33570816.0	1.0
Decode_Last	xW_K	8388608.0	4194304.0	8192.0	8388608.0	8192.0	8404992.0	1.0
Decode_Last	xW_V	8388608.0	4194304.0	8192.0	8388608.0	8192.0	8404992.0	1.0
Decode_Last	RoPE-Q	8192.0	0.0	8192.0	256.0	8192.0	16640.0	0.49
Decode_Last	RoPE-K	2048.0	0.0	2048.0	256.0	2048.0	4352.0	0.47
Decode_Last	Q × K^T	67100672.0	0.0	8192.0	16777216.0	524288.0	17309696.0	3.88
Decode_Last	Attn × V	67108864.0	0.0	524288.0	16777216.0	8192.0	17309696.0	3.88
Decode_Last	xW_O	33554432.0	16777216.0	8192.0	33554432.0	8192.0	33570816.0	1.0
Decode_Last	FFN-1 (with Gate)	234895360.0	117440512.0	8192.0	234881024.0	28672.0	234917888.0	1.0
Decode_Last	FFN-2	117440512.0	58720256.0	28672.0	117440512.0	8192.0	117477376.0	1.0