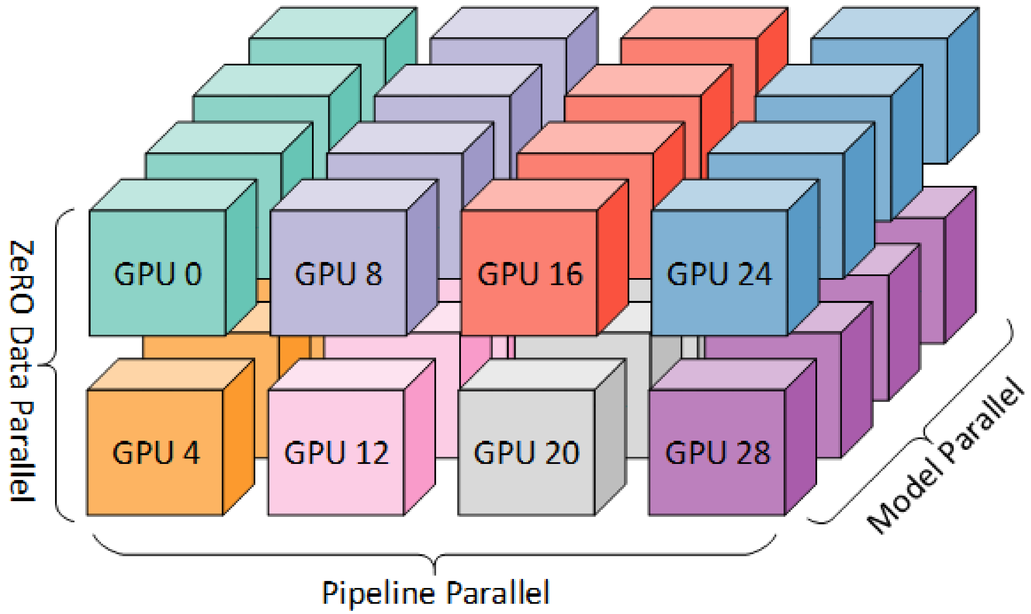
Memory Estimator for MoE LLMs

Yan Bai, Hongbin Liu, Zijie Yan

2024.12.04

Contents

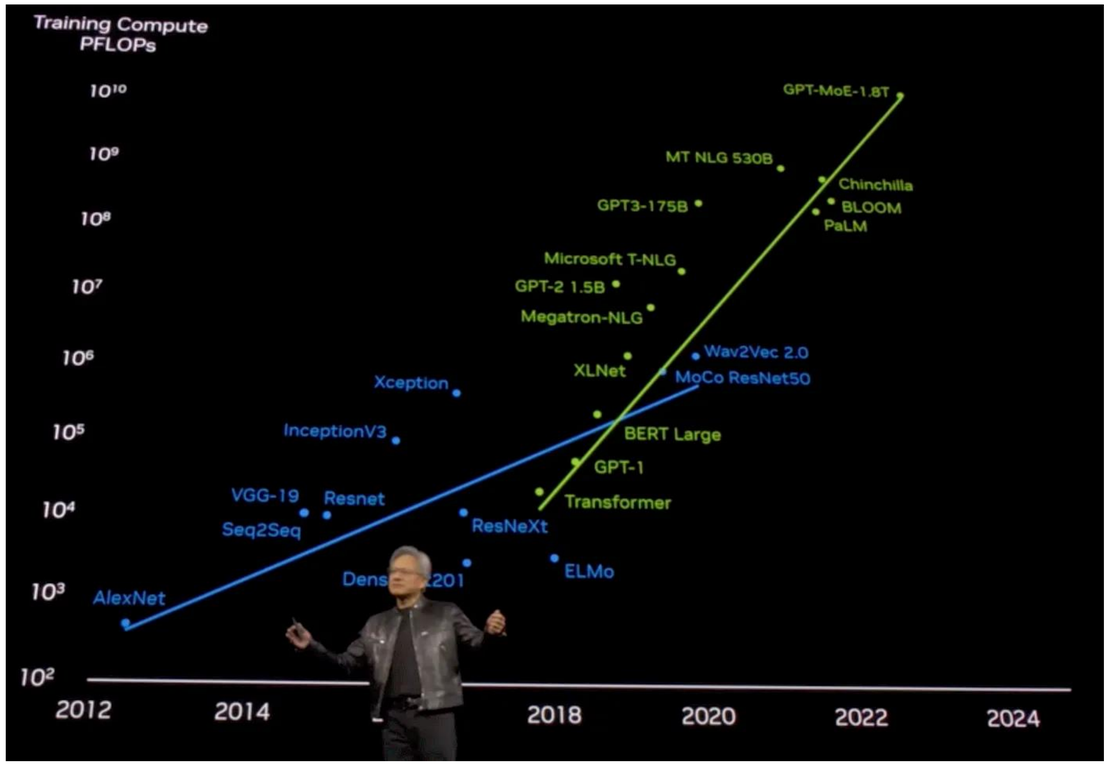
- Why we need a memory estimator for LLM?
- Introducing Memory Estimator
  - Usage
  - Design
  - Limitation
- Memory Peaks Characteristics
  - No Parallel
  - Tensor Parallel / Sequence Parallel
  - Expert Parallel
  - Pipeline Parallel / Virtual Pipeline Parallel
  - Context Parallel
  - Recompute
    - Selective (Core Attention Recompute)
    - MoE Layer Recompute
    - Full/Uniform
    - Full/Block
- Correctness
- Next Steps



Why we need a memory estimator for LLM?

Especially For MoE models

- **Memory is a critical resource for LLM Training**
  - Model is getting larger
- **Increasing complexity and diversity**
  - DP/TP/PP/SP/EP/CP/VPP/Recompute/MBS/GBS/MLA…
- **Optimizing resource allocation**
- **Reducing trial-and-error costs**
- **Enabling better scalability**



Introducing Memory Estimator

- Usage
- Design
- Limitation

Introducing Memory Estimator

Usage

- Megatron-LM Training
  - python pretrain_gpt.py ${MODEL_SPECIFIC_ARGS[@]} ${TRAINING_ARGS[@]} ${MODEL_PARALLEL_ARGS[@]} ${MOE_ARGS[@]}

- Memory Estimator
  - python estimate.py ${MODEL_SPECIFIC_ARGS[@]} ${TRAINING_ARGS[@]} ${MODEL_PARALLEL_ARGS[@]} ${MOE_ARGS[@]}

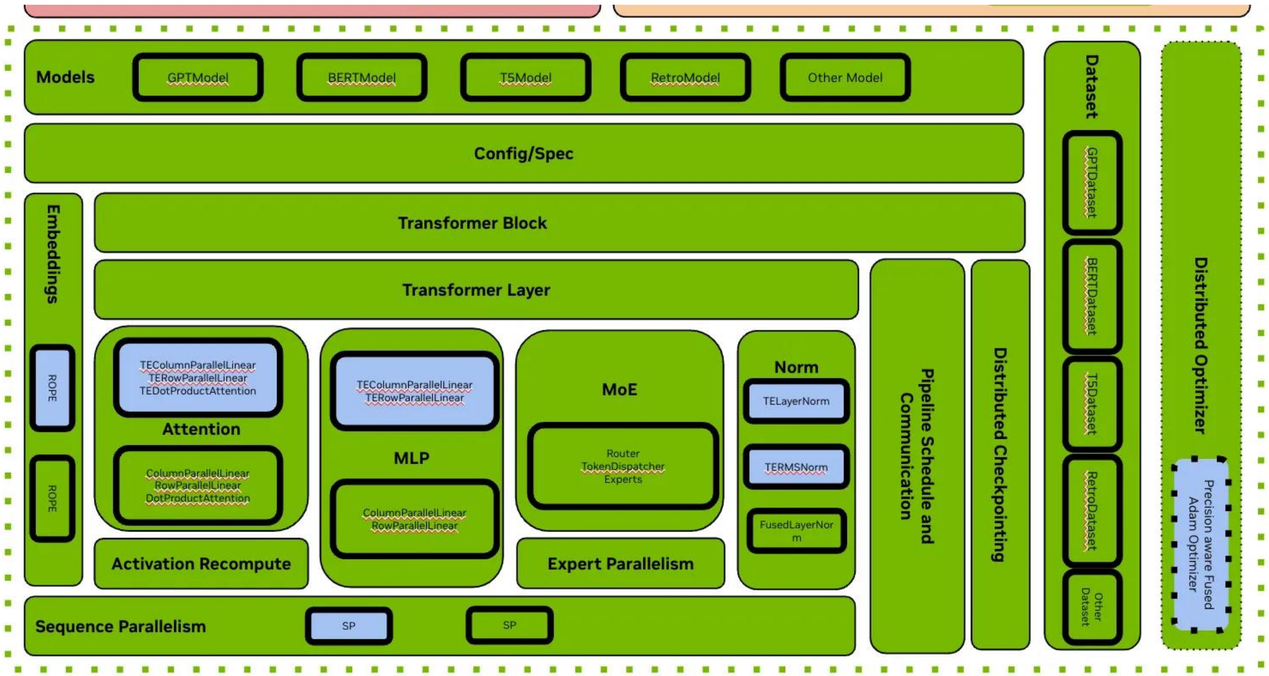
```
input_shape=[2, 4096]
GPTModel        /* n_params=1178.05M    n_act=11126.00M */ (
  (embedding): LanguageModelEmbedding   /* n_params=62.50M      n_act=16.00M */ (
    (word_embeddings): VocabParallelEmbedding   /* n_params=62.50M      n_act=16.00M */ ()
    (embedding_dropout): Dropout        /* n_params=0.00M       n_act=0.00M */ ()
  )
  (decoder): TransformerBlock   /* n_params=1053.05M    n_act=10360.00M */ (
    (layers): ModuleList        /* n_params=1053.05M    n_act=10344.00M */ (
      (0-23): 24 x TransformerLayer     /* n_params=43.88M      n_act=431.00M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=12.00M      n_act=64.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=16.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=8.00M       n_act=32.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=4.00M       n_act=16.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=16.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): RMSNorm    /* n_params=0.00M       n_act=16.00M */ ()
        (mlp): MoELayer /* n_params=31.88M      n_act=319.00M */ (
          (router): TopKRouter  /* n_params=0.00M       n_act=32.00M */ ()
          (experts): SequentialMLP      /* n_params=31.88M      n_act=255.00M */ (
            (local_experts): ModuleList /* n_params=31.88M      n_act=0.00M */ (
              (0): MLP  /* n_params=31.88M      n_act=255.00M */ (
                (linear_fc1): ColumnParallelLinear      /* n_params=21.25M      n_act=170.00M */ ()
                (linear_fc2): RowParallelLinear /* n_params=10.62M      n_act=85.00M */ ()
              )
            )
          )
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=16.00M */ ()
      )
    )
    (final_layernorm): RMSNorm  /* n_params=0.00M       n_act=16.00M */ ()
  )
  (output_layer): ColumnParallelLinear  /* n_params=62.50M      n_act=250.00M */ ()
)
Number of parameters in every GPU in billions:  1.24 where mlp part is  0.80
Number of activation in every GPU in billions:  11.67
num_bytes_per_parameter_dense=6.09375 num_bytes_per_parameter_moe=6.75
Theoretical memory footprints: weight and optimizer=7680.77 MB, activation=22252.00 MB, total=29932.77 MB
Theoretical memory footprints: weight and optimizer=7.50 GB, activation=21.73 GB, total=29.23 GB
```

Example Output: MODEL=Mixtral-8x2B TP=1 PP=1 EP=8 VPP=1 CP=1 MBS=2 GBS=256 NODES=16

Introducing Memory Estimator

Design

- Target
  - Accurate
  - Configurable, allow to load config from MCore config yaml
  - Modularized
    - easily extensible to new model structures
    - able to break down the memory usage
- Real Design
  - Split Memory to Optimizer States and Activations.
  - Define base class *MemEstimator* to mock *torch.nn.Module*
  - Implement every necessary module for MoE LLM, inheriting from *MemEstimator*, building the Model using the same way as in Megatron-LM.
  - Mock the forward, recursively calculate the number of parameters and activations layer by layer.
  - Align with [torch memory visualizer]



Introducing Memory Estimator

Limitation

- Not Measured Memory
  - NCCL states
  - Implementation specific memory pools
  - Temporary variable
- Not Accurate Enough
  - Expert parallelism
    - Assume tokens dispatching evenly
- Working in Progress
  - Only support FP16/BF16 now
  - Visual Language Model supports is WIP

Memory Peaks Characteristics

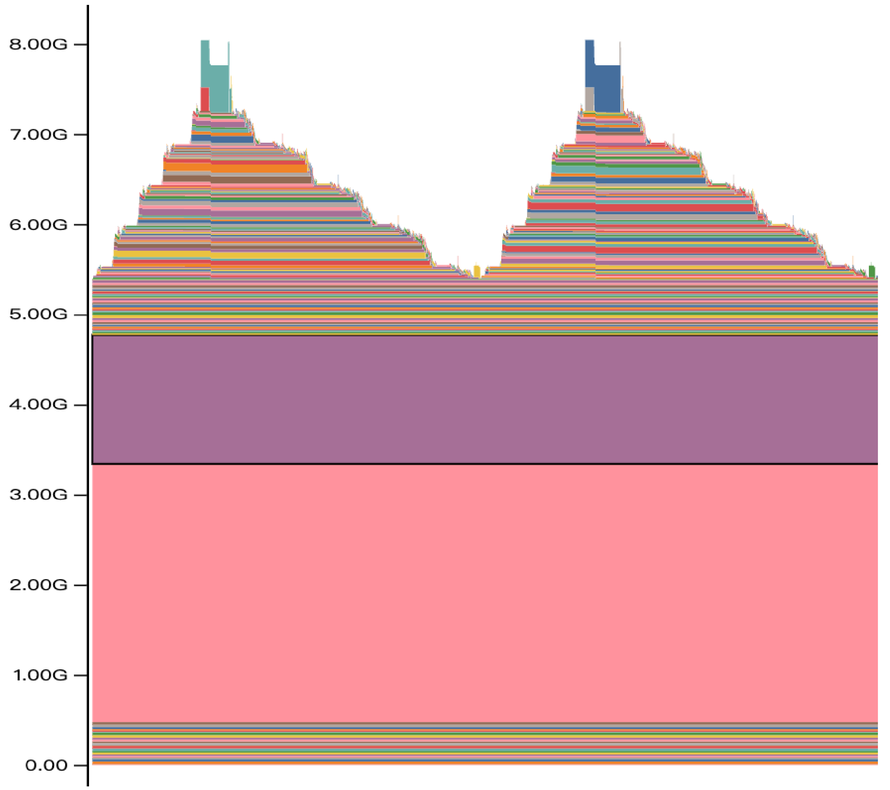
- No Parallel / Data Parallel only
- Tensor Parallel / Sequence Parallel
- Expert Parallel
- Pipeline Parallel / Virtual Pipeline Parallel
- Context Parallel
- Recompute

Memory Peaks Characteristics

No Parallel / Data Parallel only

- Optimizer States Bytes Per Parameter
  - 18 if not use distributed optimizer
    - weight(fp32)
    - weight(bf16)
    - gradient(fp32)
    - first moment(fp32)
    - second moment(fp32)
  - 6+12/dp_size otherwise
    - gradients and moments are distributed

- Achieve memory peak when calculating loss



Transformer with 4 layers

Memory Peaks Characteristics

Tensor Parallel

```
GPTModel       /* n_params=6533.05M    n_act=5563.00M */ (
  (embedding): LanguageModelEmbedding   /* n_params=62.50M      n_act=8.00M */ (
    (word_embeddings): VocabParallelEmbedding   /* n_params=62.50M      n_act=8.00M */ ()
    (embedding_dropout): Dropout        /* n_params=0.00M       n_act=0.00M */ ()
  )
  (decoder): TransformerBlock   /* n_params=6408.05M    n_act=5180.00M */ (
    (layers): ModuleList        /* n_params=6408.05M    n_act=5172.00M */ (
      (0-23): 24 x TransformerLayer     /* n_params=267.00M     n_act=215.50M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=12.00M      n_act=32.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=8.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=8.00M       n_act=16.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=4.00M       n_act=8.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=8.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): RMSNorm    /* n_params=0.00M       n_act=8.00M */ ()
        (mlp): MoELayer /* n_params=255.00M     n_act=159.50M */ (
          (router): TopKRouter  /* n_params=0.00M       n_act=16.00M */ ()
          (experts): SequentialMLP      /* n_params=255.00M     n_act=127.50M */ (
            (local_experts): ModuleList /* n_params=255.00M     n_act=0.00M */ (
              (0-7): 8 x MLP    /* n_params=31.88M      n_act=15.94M */ (
                (linear_fc1): ColumnParallelLinear      /* n_params=21.25M      n_act=10.62M */ ()
                (linear_fc2): RowParallelLinear /* n_params=10.62M      n_act=5.31M */ ()
              )
            )
          )
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=8.00M */ ()
      )
    )
    (final_layernorm): RMSNorm  /* n_params=0.00M       n_act=8.00M */ ()
  )
  (output_layer): ColumnParallelLinear  /* n_params=62.50M      n_act=125.00M */ ()
)
Number of parameters in every GPU in billions:  6.85 where mlp part is  6.42
Number of activation in every GPU in billions:  5.83
num_bytes_per_parameter=6.1875
Theoretical memory footprints: weight and optimizer=40423.24 MB, activation=11126.00 MB, total=515
49.24 MB
Theoretical memory footprints: weight and optimizer=39.48 GB, activation=10.87 GB, total=50.34 GB
```

TP=1

```
GPTModel       /* n_params=3266.55M    n_act=3461.50M */ (
  (embedding): LanguageModelEmbedding   /* n_params=31.25M      n_act=8.00M */ (
    (word_embeddings): VocabParallelEmbedding   /* n_params=31.25M      n_act=8.00M */ ()
    (embedding_dropout): Dropout        /* n_params=0.00M       n_act=0.00M */ ()
  )
  (decoder): TransformerBlock   /* n_params=3204.05M    n_act=3266.00M */ (
    (layers): ModuleList        /* n_params=3204.05M    n_act=3258.00M */ (
      (0-23): 24 x TransformerLayer     /* n_params=133.50M     n_act=135.75M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=6.00M       n_act=16.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=4.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=4.00M       n_act=8.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=2.00M       n_act=4.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=8.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): RMSNorm    /* n_params=0.00M       n_act=8.00M */ ()
        (mlp): MoELayer /* n_params=127.50M     n_act=95.75M */ (
          (router): TopKRouter  /* n_params=0.00M       n_act=16.00M */ ()
          (experts): SequentialMLP      /* n_params=127.50M     n_act=63.75M */ (
            (local_experts): ModuleList /* n_params=127.50M     n_act=0.00M */ (
              (0-7): 8 x MLP    /* n_params=15.94M      n_act=7.97M */ (
                (linear_fc1): ColumnParallelLinear      /* n_params=10.62M      n_act=5.31M */ ()
                (linear_fc2): RowParallelLinear /* n_params=5.31M       n_act=2.66M */ ()
              )
            )
          )
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=8.00M */ ()
      )
    )
    (final_layernorm): RMSNorm  /* n_params=0.00M       n_act=8.00M */ ()
  )
  (output_layer): ColumnParallelLinear  /* n_params=31.25M      n_act=62.50M */ ()
)
Number of parameters in every GPU in billions:  3.43 where mlp part is  3.21
Number of activation in every GPU in billions:  3.63
num_bytes_per_parameter=6.375
Theoretical memory footprints: weight and optimizer=20824.25 MB, activation=6923.00 MB, total=2774
7.25 MB
Theoretical memory footprints: weight and optimizer=20.34 GB, activation=6.76 GB, total=27.10 GB
```

TP=2 SP=False

- Model parameters split evenly
- Activations of sequences remain
- Activations of GEMMs split evenly

Memory Peaks Characteristics

Tensor Parallel with Sequence Parallel

```
GPTModel       /* n_params=3266.55M    n_act=3461.50M */ (
  (embedding): LanguageModelEmbedding   /* n_params=31.25M      n_act=8.00M */ (
    (word_embeddings): VocabParallelEmbedding   /* n_params=31.25M      n_act=8.00M */ ()
    (embedding_dropout): Dropout        /* n_params=0.00M       n_act=0.00M */ ()
  )
  (decoder): TransformerBlock   /* n_params=3204.05M    n_act=3266.00M */ (
    (layers): ModuleList        /* n_params=3204.05M    n_act=3258.00M */ (
      (0-23): 24 x TransformerLayer     /* n_params=133.50M     n_act=135.75M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=6.00M       n_act=16.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=4.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=4.00M       n_act=8.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=2.00M       n_act=4.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=8.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): RMSNorm    /* n_params=0.00M       n_act=8.00M */ ()
        (mlp): MoELayer /* n_params=127.50M     n_act=95.75M */ (
          (router): TopKRouter  /* n_params=0.00M       n_act=16.00M */ ()
          (experts): SequentialMLP      /* n_params=127.50M     n_act=63.75M */ (
            (local_experts): ModuleList /* n_params=127.50M     n_act=0.00M */ (
              (0-7): 8 x MLP    /* n_params=15.94M      n_act=7.97M */ (
                (linear_fc1): ColumnParallelLinear      /* n_params=10.62M      n_act=5.31M */ ()
                (linear_fc2): RowParallelLinear /* n_params=5.31M       n_act=2.66M */ ()
              )
            )
          )
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=8.00M */ ()
      )
    )
    (final_layernorm): RMSNorm  /* n_params=0.00M       n_act=8.00M */ ()
  )
  (output_layer): ColumnParallelLinear  /* n_params=31.25M      n_act=62.50M */ ()
)
Number of parameters in every GPU in billions:  3.43 where mlp part is  3.21
Number of activation in every GPU in billions:  3.63
num_bytes_per_parameter=6.375
Theoretical memory footprints: weight and optimizer=20824.25 MB, activation=6923.00 MB, total=2774
7.25 MB
Theoretical memory footprints: weight and optimizer=20.34 GB, activation=6.76 GB, total=27.10 GB
```

TP=2 SP=False

```
GPTModel       /* n_params=3266.55M    n_act=3173.50M */ (
  (embedding): LanguageModelEmbedding   /* n_params=31.25M      n_act=8.00M */ (
    (word_embeddings): VocabParallelEmbedding   /* n_params=31.25M      n_act=8.00M */ ()
    (embedding_dropout): Dropout        /* n_params=0.00M       n_act=0.00M */ ()
  )
  (decoder): TransformerBlock   /* n_params=3204.05M    n_act=2978.00M */ (
    (layers): ModuleList        /* n_params=3204.05M    n_act=2970.00M */ (
      (0-23): 24 x TransformerLayer     /* n_params=133.50M     n_act=123.75M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=6.00M       n_act=16.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=4.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=4.00M       n_act=8.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=2.00M       n_act=4.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=4.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): RMSNorm    /* n_params=0.00M       n_act=4.00M */ ()
        (mlp): MoELayer /* n_params=127.50M     n_act=95.75M */ (
          (router): TopKRouter  /* n_params=0.00M       n_act=16.00M */ ()
          (experts): SequentialMLP      /* n_params=127.50M     n_act=63.75M */ (
            (local_experts): ModuleList /* n_params=127.50M     n_act=0.00M */ (
              (0-7): 8 x MLP    /* n_params=15.94M      n_act=7.97M */ (
                (linear_fc1): ColumnParallelLinear      /* n_params=10.62M      n_act=5.31M */ ()
                (linear_fc2): RowParallelLinear /* n_params=5.31M       n_act=2.66M */ ()
              )
            )
          )
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=4.00M */ ()
      )
    )
    (final_layernorm): RMSNorm  /* n_params=0.00M       n_act=8.00M */ ()
  )
  (output_layer): ColumnParallelLinear  /* n_params=31.25M      n_act=62.50M */ ()
)
Number of parameters in every GPU in billions:  3.43 where mlp part is  3.21
Number of activation in every GPU in billions:  3.33
num_bytes_per_parameter=6.375
Theoretical memory footprints: weight and optimizer=20824.25 MB, activation=6347.00 MB, total=2717
1.25 MB
Theoretical memory footprints: weight and optimizer=20.34 GB, activation=6.20 GB, total=26.53 GB
```

TP=2 SP=True

- Activations of sequences within Transformers split
- Activations of sequences outside Transformers remain

Memory Peaks Characteristics

Expert Parallel

```
GPTModel        /* n_params=6533.05M    n_act=5563.00M */ (
  (embedding): LanguageModelEmbedding   /* n_params=62.50M      n_act=8.00M */ (
    (word_embeddings): VocabParallelEmbedding   /* n_params=62.50M      n_act=8.00M */ ()
    (embedding_dropout): Dropout        /* n_params=0.00M       n_act=0.00M */ ()
  )
  (decoder): TransformerBlock   /* n_params=6408.05M    n_act=5180.00M */ (
    (layers): ModuleList        /* n_params=6408.05M    n_act=5172.00M */ (
      (0-23): 24 x TransformerLayer     /* n_params=267.00M     n_act=215.50M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=12.00M      n_act=32.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=8.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=8.00M       n_act=16.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=4.00M       n_act=8.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=8.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): RMSNorm    /* n_params=0.00M       n_act=8.00M */ ()
        (mlp): MoELayer /* n_params=255.00M     n_act=159.50M */ (
          (router): TopKRouter  /* n_params=0.00M       n_act=16.00M */ ()
          (experts): SequentialMLP      /* n_params=255.00M     n_act=127.50M */ (
            (local_experts): ModuleList /* n_params=255.00M     n_act=0.00M */ (
              (0-7): 8 x MLP    /* n_params=31.88M      n_act=15.94M */ (
                (linear_fc1): ColumnParallelLinear      /* n_params=21.25M      n_act=10.62M */ ()
                (linear_fc2): RowParallelLinear /* n_params=10.62M      n_act=5.31M */ ()
              )
            )
          )
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=8.00M */ ()
      )
    )
    (final_layernorm): RMSNorm  /* n_params=0.00M       n_act=8.00M */ ()
  )
  (output_layer): ColumnParallelLinear  /* n_params=62.50M      n_act=125.00M */ ()
)
Number of parameters in every GPU in billions:  6.85 where mlp part is  6.42
Number of activation in every GPU in billions:  5.83
num_bytes_per_parameter=6.1875
Theoretical memory footprints: weight and optimizer=40423.24 MB, activation=11126.00 MB, total=515
49.24 MB
Theoretical memory footprints: weight and optimizer=39.48 GB, activation=10.87 GB, total=50.34 GB
```

EP=1

```
GPTModel        /* n_params=3473.05M    n_act=5563.00M */ (
  (embedding): LanguageModelEmbedding   /* n_params=62.50M      n_act=8.00M */ (
    (word_embeddings): VocabParallelEmbedding   /* n_params=62.50M      n_act=8.00M */ ()
    (embedding_dropout): Dropout        /* n_params=0.00M       n_act=0.00M */ ()
  )
  (decoder): TransformerBlock   /* n_params=3348.05M    n_act=5180.00M */ (
    (layers): ModuleList        /* n_params=3348.05M    n_act=5172.00M */ (
      (0-23): 24 x TransformerLayer     /* n_params=139.50M     n_act=215.50M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=12.00M      n_act=32.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=8.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=8.00M       n_act=16.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=4.00M       n_act=8.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=8.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): RMSNorm    /* n_params=0.00M       n_act=8.00M */ ()
        (mlp): MoELayer /* n_params=127.50M     n_act=159.50M */ (
          (router): TopKRouter  /* n_params=0.00M       n_act=16.00M */ ()
          (experts): SequentialMLP      /* n_params=127.50M     n_act=127.50M */ (
            (local_experts): ModuleList /* n_params=127.50M     n_act=0.00M */ (
              (0-3): 4 x MLP    /* n_params=31.88M      n_act=31.88M */ (
                (linear_fc1): ColumnParallelLinear      /* n_params=21.25M      n_act=21.25M */ ()
                (linear_fc2): RowParallelLinear /* n_params=10.62M      n_act=10.62M */ ()
              )
            )
          )
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=8.00M */ ()
      )
    )
    (final_layernorm): RMSNorm  /* n_params=0.00M       n_act=8.00M */ ()
  )
  (output_layer): ColumnParallelLinear  /* n_params=62.50M      n_act=125.00M */ ()
)
Number of parameters in every GPU in billions:  3.64 where mlp part is  3.21
Number of activation in every GPU in billions:  5.83
num_bytes_per_parameter_dense=6.1875 num_bytes_per_parameter_moe=6.375
Theoretical memory footprints: weight and optimizer=22063.24 MB, activation=11126.00 MB, total=331
89.24 MB
Theoretical memory footprints: weight and optimizer=21.55 GB, activation=10.87 GB, total=32.41 GB
```

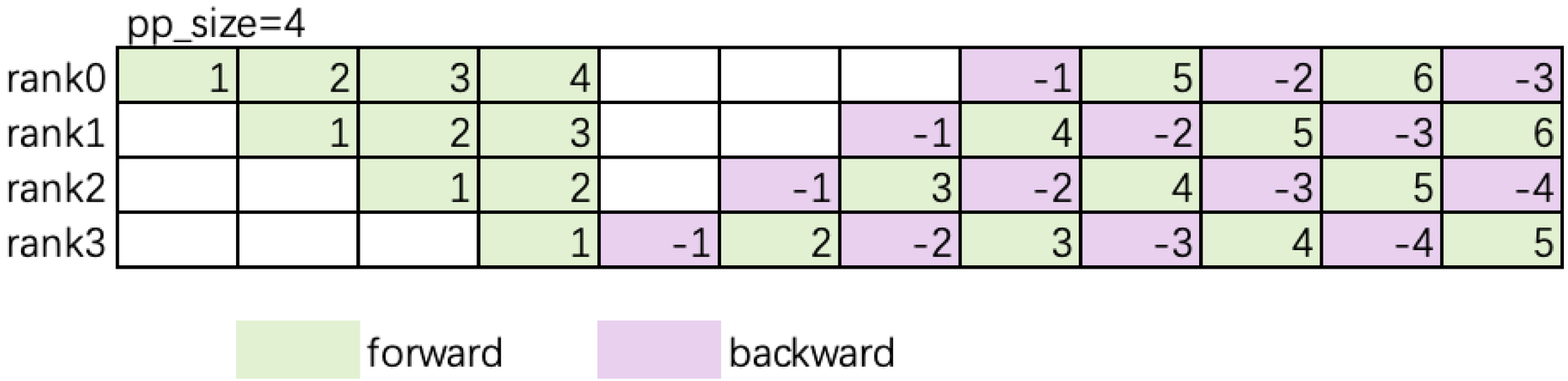
EP=2

- Model's sparse parts split evenly while dense parts remain
- Sparse and dense parts vary in dp_size
- dp_size_moe = dp_size_dense * ep_size

- Activation remains (assume tokens dispatch evenly)

Memory Peaks Characteristics

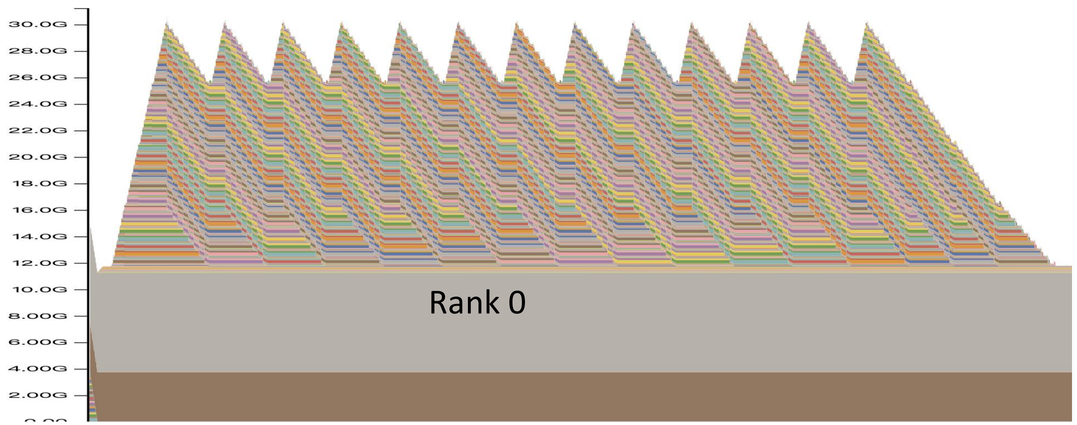
Pipeline Parallel

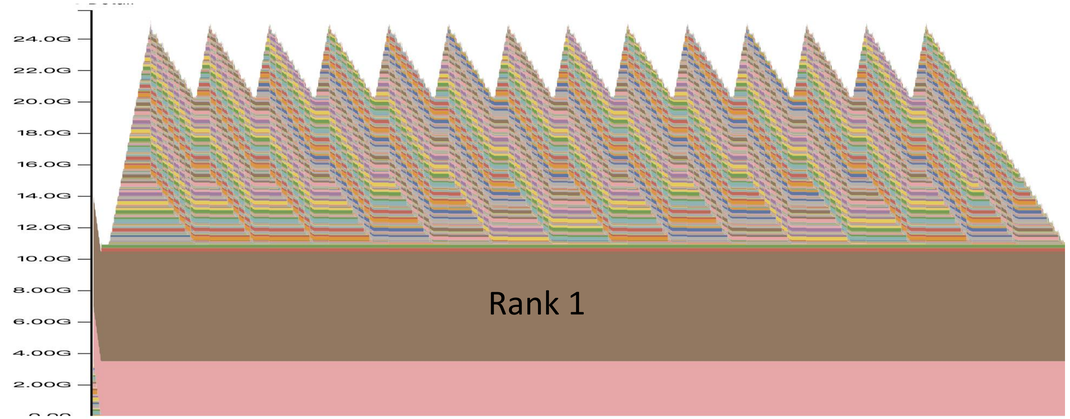


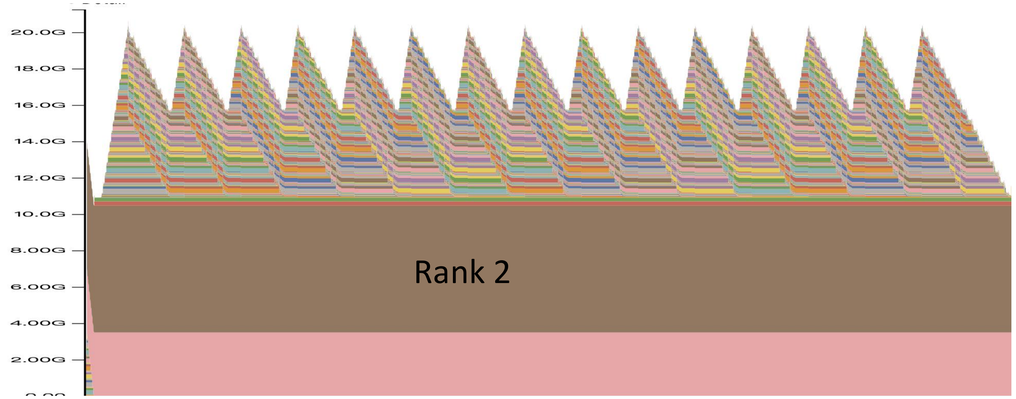
1F1B pipeline parallel with pp_size=4

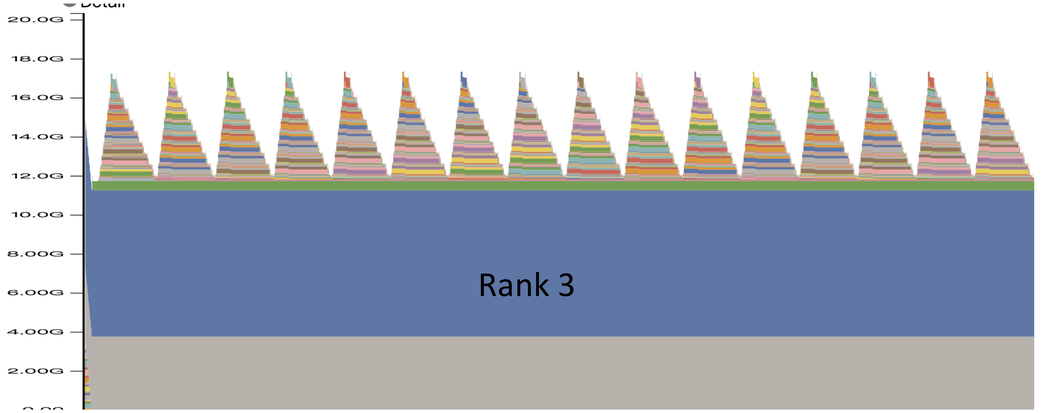
Memory Peaks Characteristics

Pipeline Parallel









Pipeline Parallel Example: Mistral7B with PP_size=4

Memory Peaks Characteristics

Pipeline Parallel

pp_size=4												
rank0	1	2	3	4				-1	5	-2	6	-3
rank1		1	2	3			-1	4	-2	5	-3	6
rank2			1	2		-1	3	-2	4	-3	5	-4
rank3				1	-1	2	-2	3	-3	4	-4	5

forward backward

1F1B pipeline parallel with pp_size=4

- Every forward will add 1x activation, while every backward sub 1x activation
- Activation multiplier = pp_size-pp_rank+1

Memory Peaks Characteristics

Pipeline Parallel

```
----------[Pipeline_Parallelism_Rank=0]---------
input_shape=[1, 4096]
GPTModel        /* n_params=1789.00M    n_act=2064.00M */ (
  (embedding): LanguageModelEmbedding   /* n_params=125.00M     n_act=16.00M */ (
    (word_embeddings): VocabParallelEmbedding   /* n_params=125.00M     n_act=16.00M */ ()
    (embedding_dropout): Dropout        /* n_params=0.00M       n_act=0.00M */ ()
  )
  (decoder): TransformerBlock   /* n_params=1664.00M    n_act=2048.00M */ (
    (layers): ModuleList        /* n_params=1664.00M    n_act=2048.00M */ (
      (0-7): 8 x TransformerLayer       /* n_params=208.00M     n_act=256.00M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=40.00M      n_act=56.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=16.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=24.00M      n_act=24.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=16.00M      n_act=16.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=16.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): IdentityOp /* n_params=0.00M       n_act=0.00M */ ()
        (mlp): MLP      /* n_params=168.00M     n_act=168.00M */ (
          (linear_fc1): ColumnParallelLinear    /* n_params=112.00M     n_act=112.00M */ ()
          (linear_fc2): RowParallelLinear       /* n_params=56.00M      n_act=56.00M */ ()
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=16.00M */ ()
      )
    )
  )
)
Number of parameters in every GPU in billions:  1.88 where mlp part is  1.41
Number of activation per microbatch in every GPU in billions:  8.66, num_microbatch_this_pp_rank=4
num_bytes_per_parameter=6.75
Theoretical memory footprints: weight and optimizer=12075.75 MB, activation=16512.00 MB, total=28587.75 MB
Theoretical memory footprints: weight and optimizer=11.79 GB, activation=16.12 GB, total=27.92 GB
```

```
----------[Pipeline_Parallelism_Rank=1]---------
input_shape=[1, 4096, 4096]
GPTModel        /* n_params=1664.00M    n_act=2048.00M */ (
  (decoder): TransformerBlock   /* n_params=1664.00M    n_act=2048.00M */ (
    (layers): ModuleList        /* n_params=1664.00M    n_act=2048.00M */ (
      (0-7): 8 x TransformerLayer       /* n_params=208.00M     n_act=256.00M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=40.00M      n_act=56.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=16.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=24.00M      n_act=24.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=16.00M      n_act=16.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=16.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): IdentityOp /* n_params=0.00M       n_act=0.00M */ ()
        (mlp): MLP      /* n_params=168.00M     n_act=168.00M */ (
          (linear_fc1): ColumnParallelLinear    /* n_params=112.00M     n_act=112.00M */ ()
          (linear_fc2): RowParallelLinear       /* n_params=56.00M      n_act=56.00M */ ()
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=16.00M */ ()
      )
    )
  )
)
Number of parameters in every GPU in billions:  1.74 where mlp part is  1.41
Number of activation per microbatch in every GPU in billions:  6.44, num_microbatch_this_pp_rank=3
num_bytes_per_parameter=6.75
Theoretical memory footprints: weight and optimizer=11232.00 MB, activation=12288.00 MB, total=23520.00 MB
Theoretical memory footprints: weight and optimizer=10.97 GB, activation=12.00 GB, total=22.97 GB
```

```
----------[Pipeline_Parallelism_Rank=2]---------
input_shape=[1, 4096, 4096]
GPTModel        /* n_params=1664.00M    n_act=2048.00M */ (
  (decoder): TransformerBlock   /* n_params=1664.00M    n_act=2048.00M */ (
    (layers): ModuleList        /* n_params=1664.00M    n_act=2048.00M */ (
      (0-7): 8 x TransformerLayer       /* n_params=208.00M     n_act=256.00M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=40.00M      n_act=56.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=16.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=24.00M      n_act=24.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=16.00M      n_act=16.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=16.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): IdentityOp /* n_params=0.00M       n_act=0.00M */ ()
        (mlp): MLP      /* n_params=168.00M     n_act=168.00M */ (
          (linear_fc1): ColumnParallelLinear    /* n_params=112.00M     n_act=112.00M */ ()
          (linear_fc2): RowParallelLinear       /* n_params=56.00M      n_act=56.00M */ ()
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=16.00M */ ()
      )
    )
  )
)
Number of parameters in every GPU in billions:  1.74 where mlp part is  1.41
Number of activation per microbatch in every GPU in billions:  4.29, num_microbatch_this_pp_rank=2
num_bytes_per_parameter=6.75
Theoretical memory footprints: weight and optimizer=11232.00 MB, activation=8192.00 MB, total=19424.00 MB
Theoretical memory footprints: weight and optimizer=10.97 GB, activation=8.00 GB, total=18.97 GB
```

```
----------[Pipeline_Parallelism_Rank=3]---------
input_shape=[1, 4096, 4096]
GPTModel        /* n_params=1789.00M    n_act=2439.00M */ (
  (decoder): TransformerBlock   /* n_params=1664.00M    n_act=2064.00M */ (
    (layers): ModuleList        /* n_params=1664.00M    n_act=2048.00M */ (
      (0-7): 8 x TransformerLayer       /* n_params=208.00M     n_act=256.00M */ (
        (input_layernorm): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (self_attention): SelfAttention /* n_params=40.00M      n_act=56.00M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=16.00M */ ()
          (linear_qkv): ColumnParallelLinear    /* n_params=24.00M      n_act=24.00M */ ()
          (q_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (k_layernorm): IdentityOp     /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=16.00M      n_act=16.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=16.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): IdentityOp /* n_params=0.00M       n_act=0.00M */ ()
        (mlp): MLP      /* n_params=168.00M     n_act=168.00M */ (
          (linear_fc1): ColumnParallelLinear    /* n_params=112.00M     n_act=112.00M */ ()
          (linear_fc2): RowParallelLinear       /* n_params=56.00M      n_act=56.00M */ ()
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=16.00M */ ()
      )
    )
    (final_layernorm): RMSNorm  /* n_params=0.00M       n_act=16.00M */ ()
  )
  (output_layer): ColumnParallelLinear  /* n_params=125.00M     n_act=125.00M */ ()
)
Number of parameters in every GPU in billions:  1.88 where mlp part is  1.41
Number of activation per microbatch in every GPU in billions:  2.56, num_microbatch_this_pp_rank=1
num_bytes_per_parameter=6.75
Theoretical memory footprints: weight and optimizer=12075.78 MB, activation=4878.00 MB, total=16953.78 MB
Theoretical memory footprints: weight and optimizer=11.79 GB, activation=4.76 GB, total=16.56 GB
```

Memory Peaks Characteristics

Virtual Pipeline Parallel

pp_size=4 vpp_size=2

rank0	1	2	3	4	1	2	3	4	5	6	7	-1	8	-2	5	-3
rank1		1	2	3	4	1	2	3	4	5	-1	6	-2	7	-3	8
rank2			1	2	3	4	1	2	3	-1	4	-2	5	-3	6	-4
rank3				1	2	3	4	1	-1	2	-2	3	-3	4	-4	5

vpp_chunk1 forward

vpp_chunk2 forward

vpp_chunk1 barckward

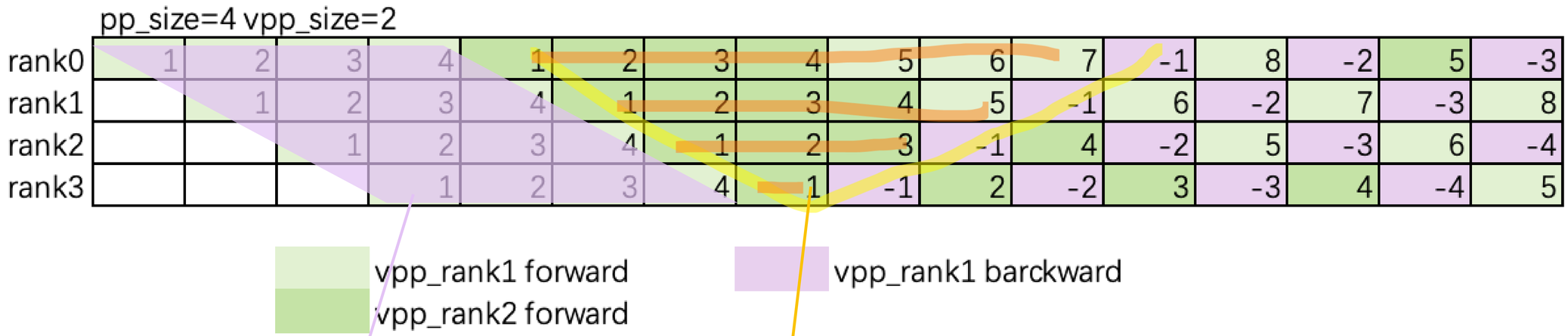
Interleave 1F1B virtual pipeline parallel with pp_size=4 vpp_size=2

- Transformers of every pp rank is splite into vpp_size chunks
- Every forward will add 1x activation of chunk, while every backward sub 1x activation of chunk
- Activation multiplier

```
num_microbatch_this_pp_rank = (
pp_size * (vpp_size - 1) + (pp_size - pp_rank) * 2 - 1
) / vpp_size
```

Memory Peaks Characteristics

Virtual Pipeline Parallel



Interleave 1F1B virtual pipeline parallel with pp_size=4 vpp_size=2

For All chunks but the last,
feed pp_size pieces of data

For the last chunk,
The gap between forward and backward passes of the first data will be filled by other data.

num_microbatch_this_pp_rank = (

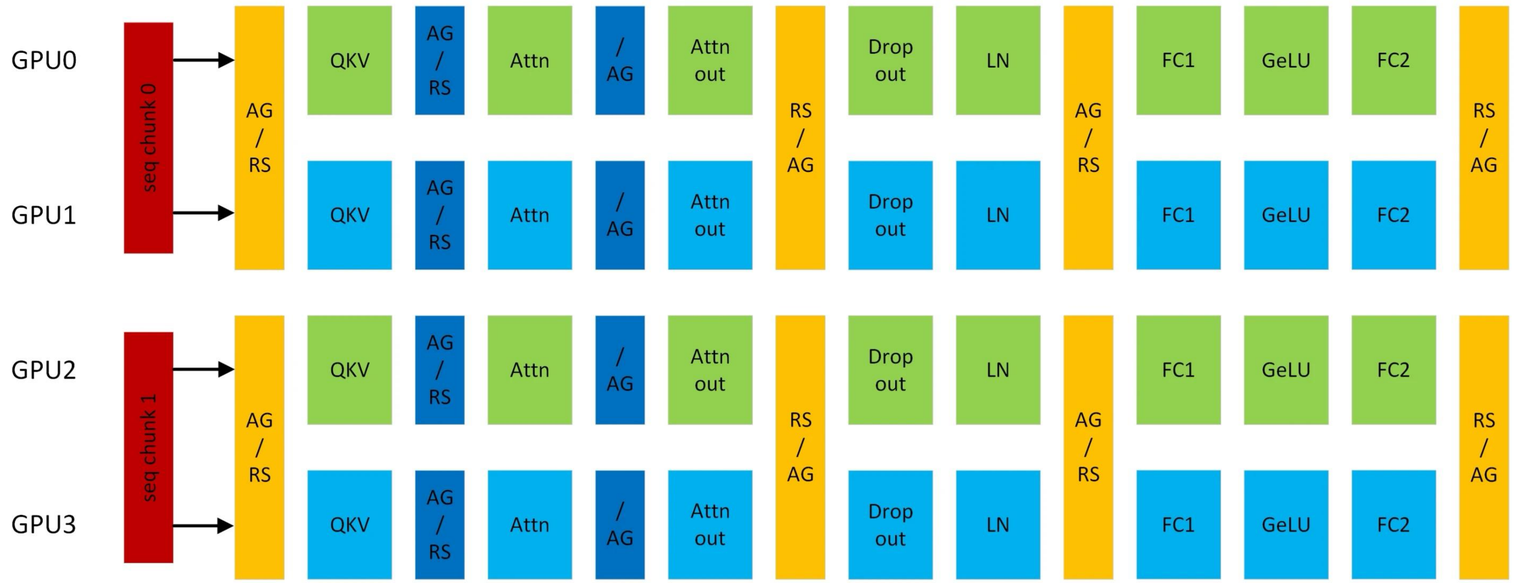
pp_size * (vpp_size - 1) + (pp_size - pp_rank) * 2 – 1

) / vpp_size

Memory Peaks Characteristics

Context Parallel

Transformer Layer with TP+CP



- In terms of memory, context parallel is a pro version of sequence parallel, which split the activation **evenly**.
- Trivial exception: core_attention's activation will be double, due to communication overlap consumption
- CP will reduce the data parallel size, so

```
num_bytes_per_parameter = (
18
if not args.use_distributed_optimizer
else 6 + (12 / args.data_parallel_size /
args.context_parallel_size)
)
```

Memory Peaks Characteristics

Recompute

- --recompute-granularity: selective
  - core_attention will be recompute, thus reducing its activation
- --moe-layer-recompute: true
  - mlp will be recompute, only mlp.router's activation leaves.

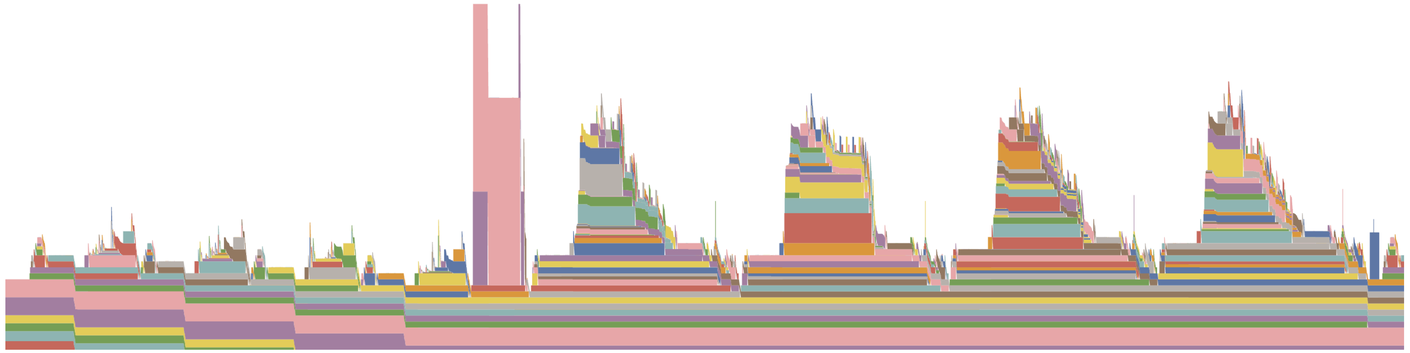
```
GPTModel        /* n_params=4852.15M    n_act=1895.34M */ (
  (embedding): LanguageModelEmbedding   /* n_params=195.56M     n_act=8.00M */ (
    (word_embeddings): VocabParallelEmbedding   /* n_params=195.56M     n_act=8.00M */ ()
    (embedding_dropout): Dropout        /* n_params=0.00M       n_act=0.00M */ ()
  )
  (decoder): TransformerBlock   /* n_params=4461.04M    n_act=714.00M */ (
    (layers): ModuleList        /* n_params=4461.04M    n_act=706.00M */ (
      (0-7): 8 x TransformerLayer       /* n_params=557.63M     n_act=88.25M */ (
        (input_layernorm): RMSNorm      /* n_params=0.00M       n_act=8.00M */ ()
        (self_attention): MLASelfAttention      /* n_params=13.13M      n_act=40.25M */ (
          (core_attention): TEDotProductAttention       /* n_params=0.00M       n_act=0.00M */ ()
          (linear_proj): RowParallelLinear      /* n_params=4.00M       n_act=8.00M */ ()
          (linear_q_proj): ColumnParallelLinear /* n_params=6.00M       n_act=12.00M */ ()
          (linear_kv_down_proj): ColumnParallelLinear   /* n_params=1.12M       n_act=2.25M */ ()
          (linear_kv_up_proj): ColumnParallelLinear     /* n_params=2.00M       n_act=16.00M */ ()
          (kv_layernorm): RMSNorm       /* n_params=0.00M       n_act=2.00M */ ()
        )
        (self_attn_bda): GetBiasDropoutAdd      /* n_params=0.00M       n_act=8.00M */ ()
        (pre_cross_attn_layernorm): IdentityOp  /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attention): IdentityOp   /* n_params=0.00M       n_act=0.00M */ ()
        (cross_attn_bda): IdentityOp    /* n_params=0.00M       n_act=0.00M */ ()
        (pre_mlp_layernorm): RMSNorm    /* n_params=0.00M       n_act=8.00M */ ()
        (mlp): MoELayer /* n_params=544.50M      n_act=16.00M */ (
          (router): TopKRouter  /* n_params=0.00M       n_act=16.00M */ ()
          (experts): TEGroupedMLP       /* n_params=528.00M     n_act=99.00M */ (
            (linear_fc1): TEColumnParallelGroupedLinear /* n_params=352.00M     n_act=66.00M */ ()
            (linear_fc2): TERowParallelGroupedLinear    /* n_params=176.00M     n_act=48.00M */ ()
          )
          (shared_experts): SharedExpertMLP     /* n_params=16.50M      n_act=33.00M */ (
            (linear_fc1): ColumnParallelLinear  /* n_params=11.00M      n_act=22.00M */ ()
            (linear_fc2): RowParallelLinear     /* n_params=5.50M       n_act=11.00M */ ()
          )
        )
        (mlp_bda): GetBiasDropoutAdd    /* n_params=0.00M       n_act=8.00M */ ()
      )
    )
    (final_layernorm): RMSNorm  /* n_params=0.00M       n_act=8.00M */ ()
  )
  (output_layer): ColumnParallelLinear  /* n_params=195.56M     n_act=391.11M */ ()
)
Number of parameters in every GPU in billions:  5.09 where mlp part is  4.57
Number of activation in every GPU in billions:  1.99
num_bytes_per_parameter=7.5
Theoretical memory footprints: weight and optimizer=36391.13 MB, activation=3790.68 MB, total=40181.81 MB
Theoretical memory footprints: weight and optimizer=35.54 GB, activation=3.70 GB, total=39.24 GB
```

Memory Peaks Characteristics

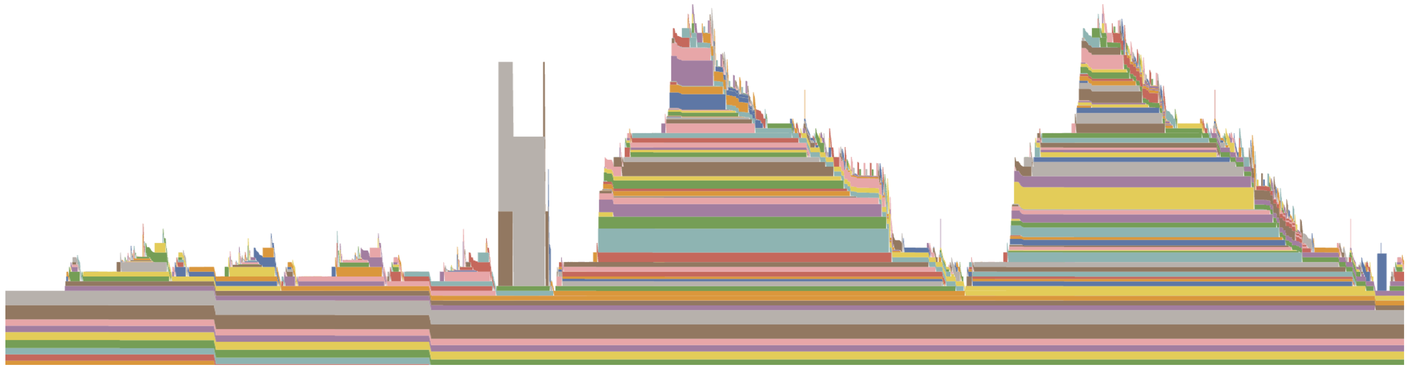
Recompute
--recompute-granularity=full

--recompute-method: uniform
--recompute-num-layers: x

- Recompute every x layer
- Memory peaks when either
  - Calculating loss, or
  - Recompute x layers



x=1 n_layer=4



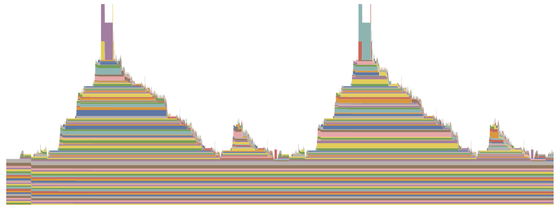
x=2 n_layer=4

Memory Peaks Characteristics

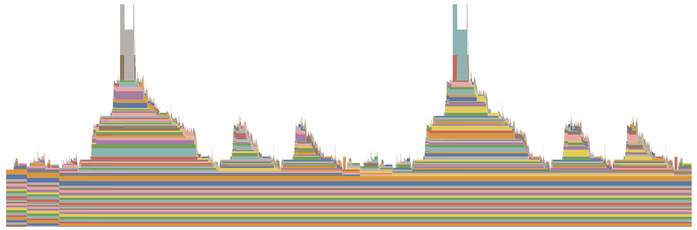
Recompute
--recompute-granularity=full

--recompute-method: block
--recompute-num-layers: x

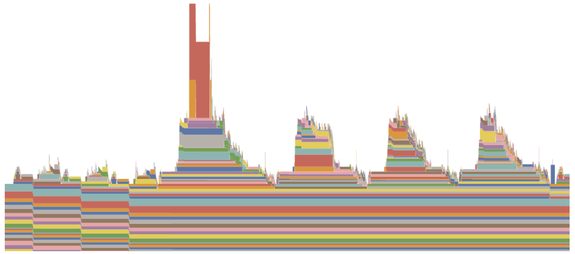
- Recompute first x layers one by one,
- The remaining are not recomputed
- Memory peaks when either
  - Calculating loss, or
  - Recompute 1 layer



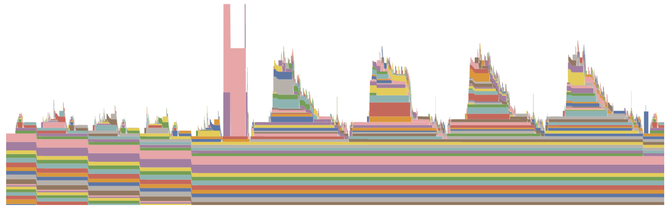
x=1 n_layer=4



x=2 n_layer=4



x=3 n_layer=4



x=4 n_layer=4

--recompute-method: block
--recompute-num-layers:
n_layer_this_pp_rank

Is equivalent to

--recompute-method: uniform
--recompute-num-layers: 1

Correctness

- Dense model
- MoE model
- Impact of in-balanced MoE tokens
- Impact of overlap-p2p-communication

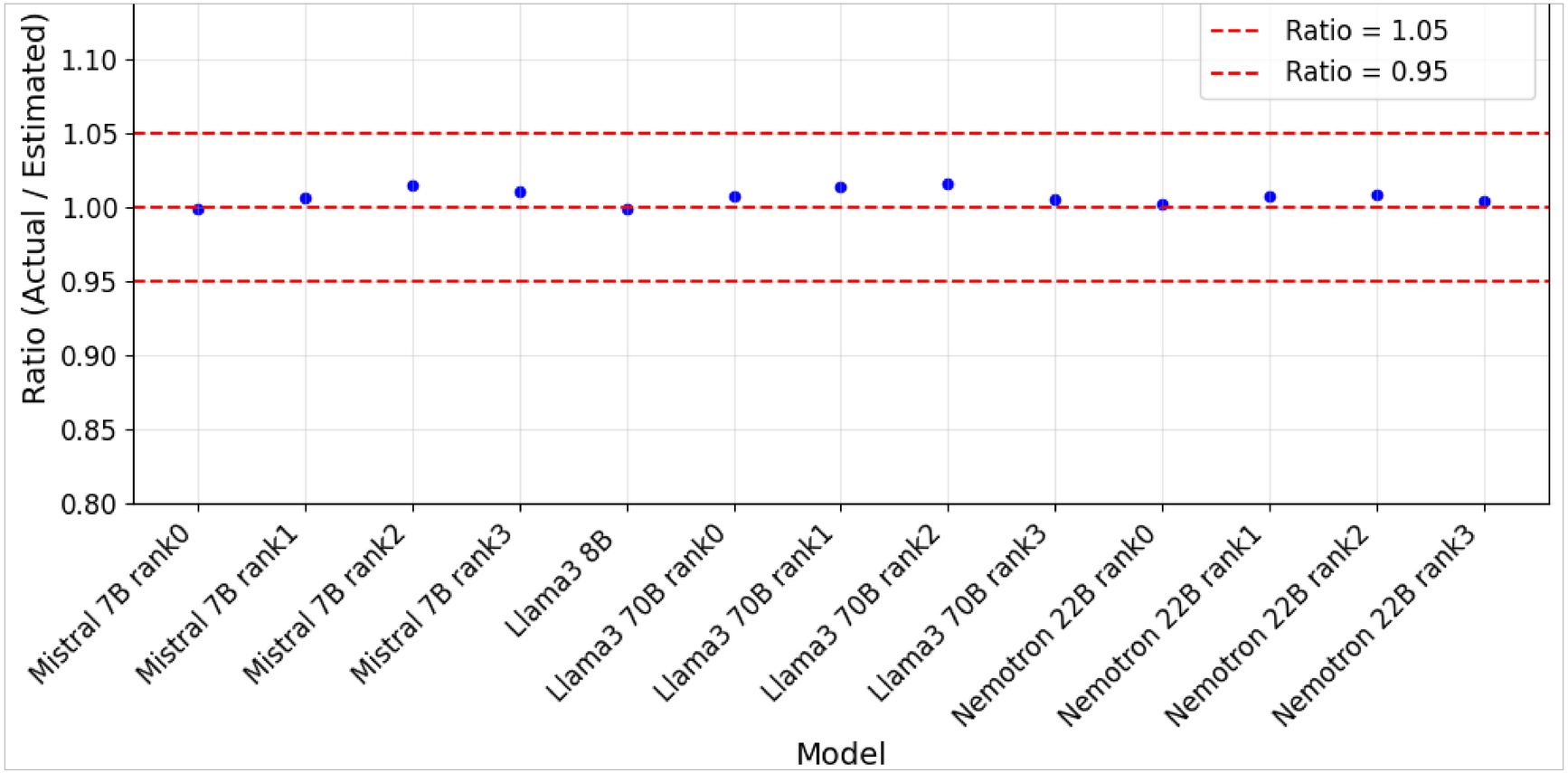
Correctness

Dense model

Model Name	TP	EP	PP	VPP	CP	GPUS	MBS	GBS	PP RANK	Estimated (GB)	Est. Model	Est. act.	Real Peak	Real inactive	Real Active Peak	diff model	diff act.
Mistral 7B	1	/	4	1	1	64	1	256	0	28.92	11.8	17.12	28.9	11.8	17.1	0	0.02
									1	23.75	11	12.75	23.9	11.1	12.8	0.1	0.05
									2	19.5	11	8.5	19.8	11.1	8.7	0.1	0.2
									3	16.81	11.8	5.01	17	12	5	0.2	0.01
Llama3 8B	2	/	1	1	2	128	1	2048	0	33.42	23.14	10.28	33.4	23.3	10.1	0.16	0.18
Llama3 70B	4	/	4	10	2	1024	1	2048	0	49.31	26.2	23.11	49.7	26.2	23.5	0	0.39
									1	46.43	24.65	21.78	47.1	24.7	22.4	0.05	0.62
									2	45.37	24.65	20.72	46.1	24.7	21.4	0.05	0.68
									3	46.82	26.2	20.62	47.1	26.3	20.8	0.1	0.18
Nemotron 22B	2	/	4	10	1	16	1	32	0	61.44	42.5	18.94	61.6	42.6	19	0.1	0.06
									1	47.81	33.75	14.06	48.2	33.8	14.4	0.05	0.34
									2	43.13	33.75	9.38	43.5	33.8	9.7	0.05	0.32
									3	50.2	42.54	7.66	50.4	42.7	7.7	0.16	0.04

Correctness

Dense model



Correctness

MoE model

Model Name	TP	EP	PP	VPP	CP	GPUS	MBS	GBS	PP RANK	Estimated (GB)	Est. Model	Est. act.	Real Peak	Real inactive	Real Active Peak	diff model	diff act
Mixtral 8x2B	1	8	1	1	1	128	2	256	/	29.98	7.5	22.48	29.4	7.6	21.8	-0.1	0.68
Mixtral 8x7B	1	8	4	1	1	64	1	256	0	42.45	14.89	27.56	37.5	15.1	22.4	-0.21	5.16
									1	34.64	14.06	20.58	34.2	14.3	19.9	-0.24	0.68
									2	27.78	14.06	13.72	27	14.3	12.7	-0.24	1.02
									3	22.51	14.89	7.62	23.5	15.2	8.3	-0.31	-0.68
Mixtral 8x7B	1	8	4	8	1	64	1	256	0	45.04	14.89	30.15	48.5	15.1	33.4	-0.21	-3.25
									1	42.35	14.06	28.29	45.5	14.3	31.2	-0.24	-2.91
									2	40.64	14.06	26.58	47.3	14.3	33	-0.24	-6.42
									3	40.6	14.89	25.71	51.8	15.2	36.6	-0.31	-10.89
Mixtral-8x22B	2	8	8	1	1	128	1	256	0	62.06	20.56	41.5	62.8	20.8	42	-0.24	-0.5
									1	55.85	19.87	35.98	56.1	20.2	35.9	-0.33	0.08
									2	50.71	19.87	30.84	48	20.2	27.8	-0.33	3.04
									3	45.57	19.87	25.7	49.4	20.2	29.2	-0.33	-3.5
									4	40.43	19.87	20.56	42.1	20.2	21.9	-0.33	-1.34
									5	35.3	19.87	15.43	35.8	20.2	15.6	-0.33	-0.17
									6	30.15	19.87	10.28	30.2	20.2	10	-0.33	0.28
									7	26.11	20.56	5.55	26.4	20.9	5.5	-0.34	0.05
DeepSeekV2	1	8	20	1	1	160	1	512	0	36.32	24.59	11.73	36.3	24.7	11.6	-0.11	0.13
									1	39.42	27.85	11.57	39.5	28.1	11.4	-0.25	0.17
									2	39.3	27.85	11.45	39.4	28.1	11.3	-0.25	0.15
									3	39.19	27.85	11.34	39.2	28.1	11.1	-0.25	0.24
									4	39.07	27.85	11.22	39.1	28.1	11	-0.25	0.22
									5	38.95	27.85	11.1	39	28.1	10.9	-0.25	0.2
									6	38.83	27.85	10.98	38.9	28.1	10.8	-0.25	0.18
									7	38.72	27.85	10.87	38.8	28.1	10.7	-0.25	0.17
									8	38.6	27.85	10.75	38.7	28.1	10.6	-0.25	0.15
									9	38.48	27.85	10.63	38.5	28.1	10.4	-0.25	0.23
									10	38.37	27.85	10.52	38.4	28.1	10.3	-0.25	0.22
									11	38.25	27.85	10.4	38.3	28.1	10.2	-0.25	0.2
									12	38.13	27.85	10.28	38.2	28.1	10.1	-0.25	0.18
									13	38.01	27.85	10.16	38.1	28.1	10	-0.25	0.16
									14	37.9	27.85	10.05	37.9	28.1	9.8	-0.25	0.25
									15	37.78	27.85	9.93	37.8	28.1	9.7	-0.25	0.23
									16	37.66	27.85	9.81	37.7	28.1	9.6	-0.25	0.21
									17	37.55	27.85	9.7	37.6	28.1	9.5	-0.25	0.2
									18	37.43	27.85	9.58	37.5	28.1	9.4	-0.25	0.18
									19	40.97	31.51	9.46	41	31.7	9.3	-0.19	0.16

Full recompute

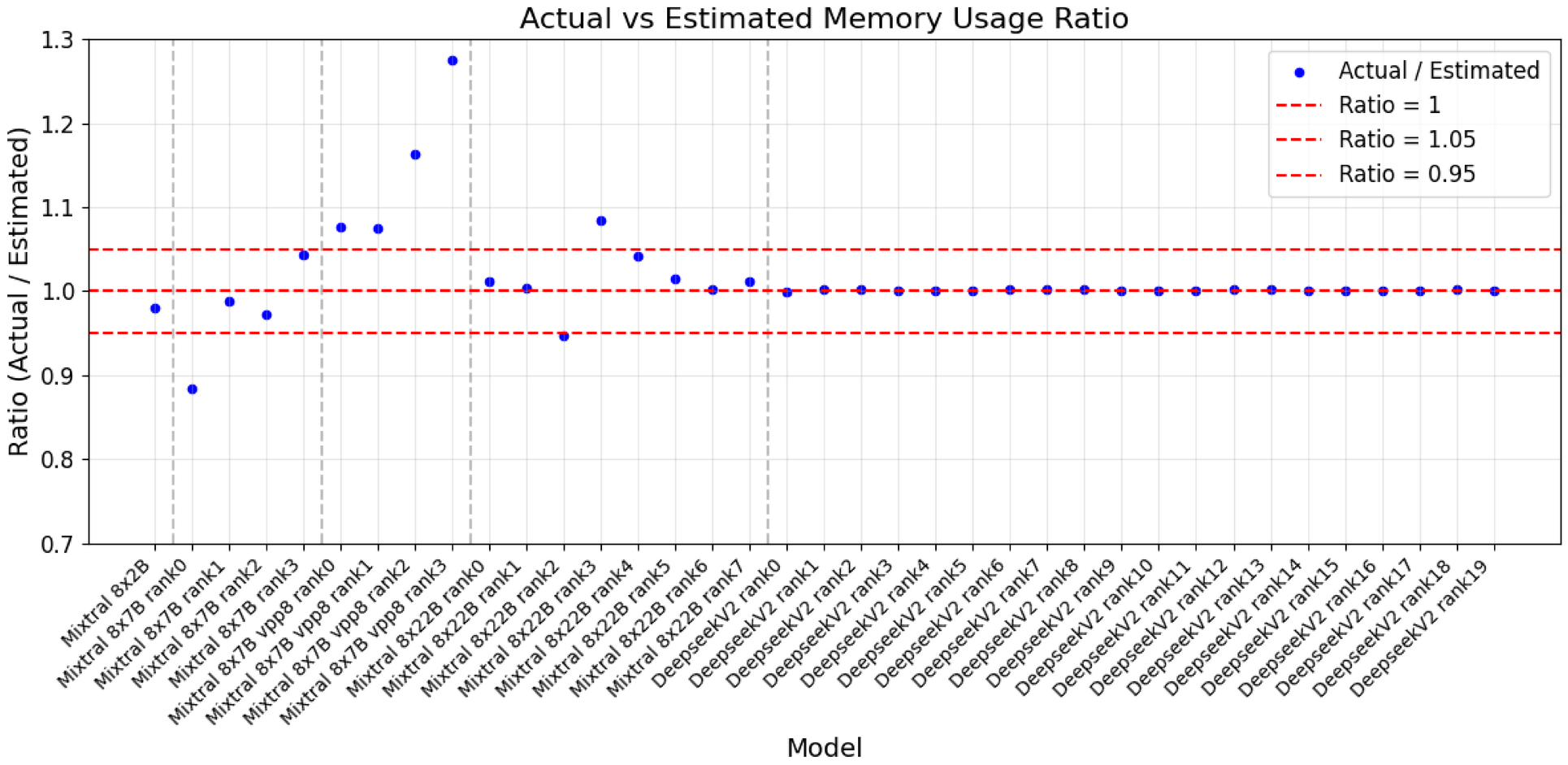
Caused by in-balanced MoE tokens

In-balanced tokens don’t matter when recomputing is on

Note: The models are measured by the first several steps from scratch, so the tokens are extremely in-balanced

Correctness

MoE model



Note: The models are measured by the first several steps from scratch, so the tokens are extremely in-balanced

Correctness

Impact of in-balanced MoE tokens

Model Name	TP	EP	PP	VPP	CP	GPUS	MBS	GBS	PP RANK	Estimated	Est. Model	Est. act.	Real Peak	Real inactive	Real Active Peak	diff model	diff act
Mixtral 8x7B	1	8	4	8	1	64	1	256	0	45.04	14.89	30.15	48.5	15.1	33.4	-0.21	-3.25
									1	42.35	14.06	28.29	45.5	14.3	31.2	-0.24	-2.91
									2	40.64	14.06	26.58	47.3	14.3	33	-0.24	-6.42
									3	40.6	14.89	25.71	51.8	15.2	36.6	-0.31	-10.89
Mixtral 8x7B	1	8	4	8	1	64	1	256	0	45.04	14.89	30.15	44.6	15.1	29.5	-0.21	0.65
									1	42.35	14.06	28.29	42.1	14.3	27.8	-0.24	0.49
									2	40.64	14.06	26.58	40.6	14.3	26.3	-0.24	0.28
									3	40.6	14.89	25.71	40	15.2	24.8	-0.31	0.91

force token balance

With balanced MoE token dispatches, the estimation is accurate

Correctness

Impact of overlap-p2p-communication

Model Name	TP	EP	PP	VPP	CP	GPUS	MBS	GBS	PP RANK	Estimated	Est. Model	Est. act.	Real Peak	Real inactive	Real Active Peak	diff model	diff act
Mistral 7B	1	/	4	8	1	64	1	256	0	30.53	11.8	18.73	32.9	11.7	21.2	0.1	2.47
									1	28.53	11	17.53	31.2	11.5	19.7	0.5	2.17
									2	27.47	11	16.47	30.2	11.5	18.7	0.5	2.23
									3	28.05	11.8	16.25	30.5	12.1	18.4	0.3	2.15
	1	/	4	8	1	64	1	256	0	30.53	11.8	18.73	31	11.7	19.3	0.1	0.57
									1	28.53	11	17.53	29.1	11.5	17.6	0.5	0.07
				no-overlap-p2p-communication					2	27.47	11	16.47	28.1	11.5	16.6	0.5	0.13
									3	28.05	11.8	16.25	28.2	12.1	16.1	0.3	0.15

When overlap-p2p-communication is enabled, there will be a relatively fixed memory overhead.

Next Steps

Next Steps

- Accuracy
  - Investigate the impact of overlap-p2p-communication
- Functional
  - Handy interface for uneven EP token distribution
  - Support VLM
  - Keep tracking Megatron-Core’s evolution.
    - Support FP8
    - Support Zero-2, Zero-3
    - ……