Workload Generator in AICB

The **Workload Generator** simulates the computational and communication workload of large language models (LLMs) like GPT, LLaMA, and T5 *without requiring actual GPUs*.

It produces workload trace files describing how training steps would use compute resources and network bandwidth in a distributed setup.

Types of Workload We Can Generate:

The tool can create workloads for different frameworks:

- **Megatron** (standard GPT-style transformers) (This is what I am trying to use)
- **DeepSpeed** (ZeRO and MoE)
- **Custom (AIOB)** workloads with specific compute time

Each workload describes:

- **Layers:** embedding, attention, MLP, normalization, optimizer steps.
- **Forward & backward compute times**.
- **Collective communication operations:** all-reduce, reduce-scatter, all-gather.
- **Batching and parallelism configurations.**

Main Parameters

`--frame=Megatron`
Tells the generator to emulate the **Megatron-LM**, a framework built for large-scale, multi-GPU training.

`--gpu_type`
Specify the GPU hardware (e.g., A100, H100) so the simulator can match its speed and memory limits.

`--model_name`
Model size

`--world_size`
Total number of GPUs/nodes in the simulation.

`--tensor_model_parallel_size=8`
Splits every transformer layer across 8 GPUs, giving “tensor parallelism.”

`--pipeline_model_parallel=1`
Number of pipeline stages, 1 means no pipelining—every layer runs as a single block.

`--global_batch`
Total batch size across all GPUs.

`--micro_batch=1`
Per-device micro-batch size.
This forces maximum gradient accumulation steps, letting the model have more reasonable workload behavior.

`--num_layers`
Number of transformer layers.

`--seq_length`
Sequence length of each input (Maximum tokens fed into the model at once).

`--hidden_size`
The transformer's hidden dimension depends on the size of the model we chose

---

`--epoch_num`
Number of training epochs.

`--num_attention_heads`
Number of attention heads.

`--max_position_embeddings`
Max sequence length supported.

`--vocab_size`
Tokenizer vocabulary size (standard for GPT-2/3).
Ensures the embedding table size is realistic.

`--use-distributed-optimizer`
Simulates the optimizer running across multiple GPUs.

`--aiob_enable`
Enables **AIOB** (Automatic Inference of Operator Behavior).
Captures how long each operation takes.

`--use_flash_attn`
Swaps standard attention for FlashAttention kernels, faster and lighter on memory

---

`--swiglu`
Enables SwiGLU activation in the MLP. This activation improves performance.

`--num_experts`
Specifies how many separate mini-MLPs “experts” live inside every Mixture-of-Experts (MoE) layer. / when training GPT models, we set it to 1 because there is no need for MoE.

`--comp_filepath=workload/aiob_inputs/Example.txt`
Load a precomputed **computation time description file**, because we don’t have access to physical GPUs.

Generation

We must use the custom workload generator (AIOB) since the Megatron and Deepspeed workload generators do not work for us. Unfortunately, these workload generators rely on having access to at least 1 NVIDIA GPU and therefore generate errors while running since neither of our computers have the required equipment.

After running the workload generator command, a .txt file is generated with outputs of the workload, and the file name uses the parameters that were passed to the Python script. The info in the output file is provided by each layer in the model. The main layer types are as follows:

- **Embedding layer**: Reduces dimensionality of data
- **Attention layer**: Weighted mean reduction

- **MLP layer**: Neural network layer
- **Grad param compute**: Computes gradients
- **Grad param comm**: Communicates gradients
- **Layernorm**: Layer normalization

The first line of output contains the following information:

- **HYBRID_TRANSFORMER_FWD_IN_BCKWD**: type of model pass being simulated
- **model_parallel_NPU_group**: Represents the size of Tensor Parallelism
- **ep**: Represents the size of the Expert model parallelism
- **pp**: Represents the size of pipeline model parallelism
- **vpp**: Virtual Pipeline Parallelism
- **ga**: Gradient-accumulation steps per full update.
- **all_gpus:** GPU count
- **Checkpoints**: How many activation checkpoints the model uses.
- **pp_comm**: Pipeline-parallel communication ratio.

Starting with the embedding, attention, and MLP layer output information, we start to have a consistent data pattern. The following data pattern is:

- **Column 1**: Layer type (MLP, attention, embedding…)
- **Column 2**: Not explained
- **Column 3**: Forward compute value
- **Column 4**: Forward communication type (tensor parallel gradient)
- **Column 5**: Forward communication value (tensor parallel gradient)
- **Column 6**: Backward compute value
- **Column 7**: Backward communication value (tensor parallel gradient)
- **Column 8**: Backward communication type (tensor parallel gradient)
- **Column 9**: Backward compute value (not a typo)
- **Column 10**: Backward communication value (data parallel gradient)
- **Column 11**: Backward communication type (data parallel gradient)
- **Column 12**: Not explained

Execution 1
Command
--frame=Megatron
--gpu_type=None
--model_name=gpt_7B
--world_size=512
--tensor_model_parallel_size=2
--pipeline_model_parallel=1
--global_batch=2048
--micro_batch=1
--num_layers=36

--seq_length=1024
--hidden_size=4096
--epoch_num=1
--num_attention_heads=32
--max_position_embeddings=1024
--vocab_size=50257
--use-distributed-optimizer
--aiob_enable
--use_flash_attn
--swiglu
--num_experts=1
--comp_filepath=workload/aiob_inputs/Example.txt

Output

This will act as a baseline for the rest of the tests to be compared to. We plan to just change one parameter at a time in order to see how the parameter affects the rest of the output. The first line contains the following information:

```
HYBRID_TRANSFORMER_FWD_IN_BCKWD model_parallel_NPU_group: 2 ep: 1 pp: 1 vpp: 36 ga: 8 all_gpus: 512 checkpoints: 0 checkpoint_initiates: 0 pp_comm: 0
598
grad_gather     -1      1       NONE    0       1       NONE    0       1       ALLGATHER       6459228160      100
grad_param_comm -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER   12918456320     100
grad_param_compute      -1      1       NONE    0       34021000        NONE    0       1       NONE    0       100
layernorm       -1      1       NONE    0       1       ALLREDUCE       6459228160      1       NONE    0       100
embedding_grads -1      1       NONE    0       1       ALLREDUCE       8388608 1       NONE    0       100
moe_grad_norm1  -1      1       NONE    0       1       NONE    0       1       ALLGATHER_DP_EP 0       100
moe_grad_norm2  -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER_DP_EP     0       100
embedding_layer -1      799000  ALLREDUCE       8388608 1       NONE    0       17374000        NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
```

—---

Execution 2: Changed the sequence length and max position embeddings to 2048 from 1024. Kept everything else the same as test 1

Command

--frame=Megatron
--gpu_type=A100
--model_name=gpt_7B

--world_size=512
--tensor_model_parallel_size=2
--pipeline_model_parallel=1
--global_batch=2048
--micro_batch=1
--num_layers=36
--seq_length=2048
--hidden_size=4096
--epoch_num=1
--num_attention_heads=32
--max_position_embeddings=2048
--vocab_size=50257
--use-distributed-optimizer
--aiob_enable
--use_flash_attn
--swiglu
--num_experts=1
--comp_filepath=workload/aiob_inputs/Example.txt

Output

```
HYBRID_TRANSFORMER_FWD_IN_BCKWD model_parallel_NPU_group: 2 ep: 1 pp: 1 vpp: 36 ga: 8 all_gpus: 512 checkpoints: 0 checkpoint_initiates: 0 pp_comm: 0
598
grad_gather     -1      1       NONE    0       1       NONE    0       1       ALLGATHER       6467616768      100
grad_param_comm -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER   12935233536     100
grad_param_compute      -1      1       NONE    0       34021000        NONE    0       1       NONE    0       100
layernorm       -1      1       NONE    0       1       ALLREDUCE       6467616768      1       NONE    0       100
embedding_grads -1      1       NONE    0       1       ALLREDUCE       16777216        1       NONE    0       100
moe_grad_norm1  -1      1       NONE    0       1       NONE    0       1       ALLGATHER_DP_EP 0       100
moe_grad_norm2  -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER_DP_EP     0       100
embedding_layer -1      799000  ALLREDUCE       16777216        1       NONE    0       17374000        NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       16777216        2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       16777216        2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       16777216        2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       16777216        2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       16777216        2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       16777216        2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       16777216        2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       16777216        2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       16777216        2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       16777216        1820000 NONE    0       1820000 NONE    0       100
```

Lines 3 and 4 ALLGATHER AND REDUCESCATTER numbers slightly increased
Lines 6 and 7 ALLREDUCE numbers slightly increased

Value for forward communication has increased from 8mb to 16mb due to the increased sequence length. ALLGATHER, REDUCESCATTER, and ALLREDUCE have increased due to the increase in communication volume

Execution 3: Change the hidden size and number of heads to 2048 and 16 from 4096 and 32. 2048/16 = 128 as required. Kept everything else the same as test 1

Command

```
--frame=Megatron
--gpu_type=None
--model_name=gpt_7B
--world_size=512
--tensor_model_parallel_size=2
--pipeline_model_parallel=1
--global_batch=2048
--micro_batch=1
--num_layers=36
--seq_length=1024
--hidden_size=2048
--epoch_num=1
--num_attention_heads=16
--max_position_embeddings=1024
--vocab_size=50257
--use-distributed-optimizer
--aiob_enable
--use_flash_attn
--swiglu
--num_experts=1
--comp_filepath=workload/aiob_inputs/Example.txt
```

Output

```
HYBRID_TRANSFORMER_FWD_IN_BCKWD model_parallel_NPU_group: 2 ep: 1 pp: 1 vpp: 36 ga: 8 all_gpus: 512 checkpoints: 0 checkpoint_initiates: 0 pp_comm: 0
598
grad_gather     -1      1       NONE    0       1       NONE    0       1       ALLGATHER       1823473664      100
grad_param_comm -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER   3646947328      100
grad_param_compute      -1      1       NONE    0       34021000        NONE    0       1       NONE    0       100
layernorm       -1      1       NONE    0       1       ALLREDUCE       1823473664      1       NONE    0       100
embedding_grads -1      1       NONE    0       1       ALLREDUCE       4194304 1       NONE    0       100
moe_grad_norm1  -1      1       NONE    0       1       NONE    0       1       ALLGATHER_DP_EP 0       100
moe_grad_norm2  -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER_DP_EP     0       100
embedding_layer -1      799000  ALLREDUCE       4194304 1       NONE    0       17374000        NONE    0       100
attention_layer -1      1820000 ALLREDUCE       4194304 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       4194304 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       4194304 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       4194304 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       4194304 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       4194304 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       4194304 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       4194304 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       4194304 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       4194304 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       4194304 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       4194304 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       4194304 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       4194304 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       4194304 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       4194304 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       4194304 1820000 NONE    0       1820000 NONE    0       100
```

Lines 3 and 4 ALLGATHER AND REDUCESCATTER numbers have significantly decreased by 3.54
Lines 6 and 7 ALLREDUCE numbers have significantly decreased by 3.54

Value for forward communication has decreased from 8mb to 4mb due to the decreased hidden size and number of heads. ALLGATHER, REDUCESCATTER, and ALLREDUCE have increased due to the increase in communication volume

Execution 4: Changed the global batch to 1024 from 2048. Kept everything else the same as test 1

Command

```
--frame=Megatron
--gpu_type=None
--model_name=gpt_7B
--world_size=512
--tensor_model_parallel_size=2
--pipeline_model_parallel=1
--global_batch=1024
--micro_batch=1
--num_layers=36
--seq_length=1024
--hidden_size=4096
--epoch_num=1
--num_attention_heads=32
--max_position_embeddings=1024
--vocab_size=50257
--use-distributed-optimizer
--aiob_enable
--use_flash_attn
--swiglu
--num_experts=1
--comp_filepath=workload/aiob_inputs/Example.txt
```

Output

```
HYBRID_TRANSFORMER_FWD_IN_BCKWD model_parallel_NPU_group: 2 ep: 1 pp: 1 vpp: 36 ga: 4 all_gpus: 512 checkpoints: 0 checkpoint_initiates: 0 pp_comm: 0
306
grad_gather     -1      1       NONE    0       1       NONE    0       1       ALLGATHER       6459228160      100
grad_param_comm -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER   12918456320     100
grad_param_compute      -1      1       NONE    0       34021000        NONE    0       1       NONE    0       100
layernorm       -1      1       NONE    0       1       ALLREDUCE       6459228160      1       NONE    0       100
embedding_grads -1      1       NONE    0       1       ALLREDUCE       8388608 1       NONE    0       100
moe_grad_norm1  -1      1       NONE    0       1       NONE    0       1       ALLGATHER_DP_EP 0       100
moe_grad_norm2  -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER_DP_EP     0       100
embedding_layer -1      799000  ALLREDUCE       8388608 1       NONE    0       17374000        NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
mlp_layer       -1      2478000 ALLREDUCE       8388608 2478000 NONE    0       2478000 NONE    0       100
attention_layer -1      1820000 ALLREDUCE       8388608 1820000 NONE    0       1820000 NONE    0       100
```

Line 1 ga changed from 8 to 4
Line 2 changed from 508 to 306

All other values stayed the same for this workload. The number of gradient accumulation steps dropped from 8 to 4. Global Batch Size = Micro Batch Size × Gradient Accumulation Steps × Data Parallel Size, so it makes sense that as we decrease the global batch size while keeping all other parameters the same, the gradient accumulation steps will have to decrease as well.

Line 2 is the number of rows in the output file. Since we're doing fewer gradient accumulation operations, we have fewer events happening and therefore, fewer rows of output.

Execution 5
Command

--frame=Megatron
--gpu_type=None
--model_name=gpt_22B
--world_size=512
--tensor_model_parallel_size=2
--pipeline_model_parallel=1
--global_batch=2048
--micro_batch=1
--num_layers=48
--seq_length=4096
--hidden_size=6144
--epoch_num=1
--num_attention_heads=48
--max_position_embeddings=4096

--vocab_size=50257
--use-distributed-optimizer
--aiob_enable
--use_flash_attn
--swiglu
--num_experts=1
--comp_filepath=workload/aiob_inputs/Example.txt

Output:

```
HYBRID_TRANSFORMER_FWD_IN_BCKWD model_parallel_NPU_group: 2 ep: 1 pp: 1 vpp: 48 ga: 8 all_gpus: 512 checkpoints: 0
checkpoint_initiates: 0 pp_comm: 0
790
grad_gather      -1      1       NONE     0       1       NONE     0       1       ALLGATHER       18201182208     100
grad_param_comm  -1      1       NONE     0       1       NONE     0       1       REDUCESCATTER   36402364416     100
grad_param_compute      -1      1       NONE    0       34021000        NONE    0       1       NONE    0       100
layernorm        -1      1       NONE     0       1       ALLREDUCE       18201182208     1       NONE    0       100
embedding_grads  -1      1       NONE     0       1       ALLREDUCE       50331648        1       NONE    0       100
moe_grad_norm1   -1      1       NONE     0       1       NONE     0       1       ALLGATHER_DP_EP 0       100
moe_grad_norm2   -1      1       NONE     0       1       NONE     0       1       REDUCESCATTER_DP_EP     0       100
embedding_layer  -1      799000  ALLREDUCE       50331648        1       NONE    0       17374000        NONE    0       100
attention_layer  -1      1820000 ALLREDUCE       50331648        1820000 NONE    0       1820000 NONE    0       100
mlp_layer        -1      2478000 ALLREDUCE       50331648        2478000 NONE    0       2478000 NONE    0       100
attention_layer  -1      1820000 ALLREDUCE       50331648        1820000 NONE    0       1820000 NONE    0       100
mlp_layer        -1      2478000 ALLREDUCE       50331648        2478000 NONE    0       2478000 NONE    0       100
attention_layer  -1      1820000 ALLREDUCE       50331648        1820000 NONE    0       1820000 NONE    0       100
mlp_layer        -1      2478000 ALLREDUCE       50331648        2478000 NONE    0       2478000 NONE    0       100
attention_layer  -1      1820000 ALLREDUCE       50331648        1820000 NONE    0       1820000 NONE    0       100
mlp_layer        -1      2478000 ALLREDUCE       50331648        2478000 NONE    0       2478000 NONE    0       100
attention_layer  -1      1820000 ALLREDUCE       50331648        1820000 NONE    0       1820000 NONE    0       100
mlp_layer        -1      2478000 ALLREDUCE       50331648        2478000 NONE    0       2478000 NONE    0       100
attention_layer  -1      1820000 ALLREDUCE       50331648        1820000 NONE    0       1820000 NONE    0       100
mlp_layer        -1      2478000 ALLREDUCE       50331648        2478000 NONE    0       2478000 NONE    0       100
attention_layer  -1      1820000 ALLREDUCE       50331648        1820000 NONE    0       1820000 NONE    0       100
mlp_layer        -1      2478000 ALLREDUCE       50331648        2478000 NONE    0       2478000 NONE    0       100
attention_layer  -1      1820000 ALLREDUCE       50331648        1820000 NONE    0       1820000 NONE    0       100
mlp_layer        -1      2478000 ALLREDUCE       50331648        2478000 NONE    0       2478000 NONE    0       100
```

——----------

Workload Differences (Execution 1: “gpt_7B” **& 5:** “gpt_22B”**)**

1. **Model Size & Complexity: execution 5** is a much larger model than **execution 1**: more layers, higher hidden size, and more attention heads. Processes much more data and parameters per forward/backward pass.

2. **Sequence Length: gpt_22B** uses a much longer sequence (4096 tokens vs 1024). This increases memory and computation for layers like attention and MLP, and also causes a higher volume of communication (larger tensors being gathered, reduced, or all-reduced).

3. **Number of Layers:** More transformer layers (48 vs 36) in the larger model. This increases the total number of repeated `attention_layer` and `mlp_layer` workload entries.

4. **Communication Volume:** In the generated workload files, collective communication ops (like **ALLGATHER** and **REDUCESCATTER**) have much larger data volumes for the 22B model. For example:

    - `grad_gather` size jumps from **6,459,228,160** to **18,201,182,208**
    - `grad_param_comm` jumps from **12,918,456,320** to **36,402,364,416**

```
grad_gather      -1      1       NONE    0       1       NONE    0       1       ALLGATHER       34552676352     100
grad_param_comm -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER   69105352704     100
```

```
grad_gather      -1      1       NONE    0       1       NONE    0       1       ALLGATHER       18201182208     100
grad_param_comm -1      1       NONE    0       1       NONE    0       1       REDUCESCATTER   36402364416     100
```

This is because the parameter count and activation sizes grow with both the hidden size and the number of layers/heads.

5. **Layer Operations:** Every core operation (`attention_layer`, `mlp_layer`, `embedding_layer`) in the GPT-22B workload shows much larger tensor sizes for collective operations (see ALLREDUCE amounts). This reflects the scaling up of both model width (hidden size) and sequence length.

6. **Layernorm & Gradients:** The size of tensors being reduced/all-reduced for operations like `layernorm` and `embedding_grads` is also much larger for the 22B model due to increased hidden size and sequence.

7. **Workload File Length:** The number of operations listed grows in proportion to the number of layers. More layers = more repeated `attention_layer` and `mlp_layer` steps per training iteration.

—----------

With 5 different executions of the workload generator, we're only seeing a change in the value of **row 5**, the **forward communication time**. Currently, the backward communication type and time for both TP and DP are null, with a type of None and value of 0. The changes we've tested are not affecting the forward compute time either. We tried experimenting with additional arguments, but did not find any notable findings with effects on any of the other fields.