[EOD REPORT] Adhitya Charan - 28-11-2025

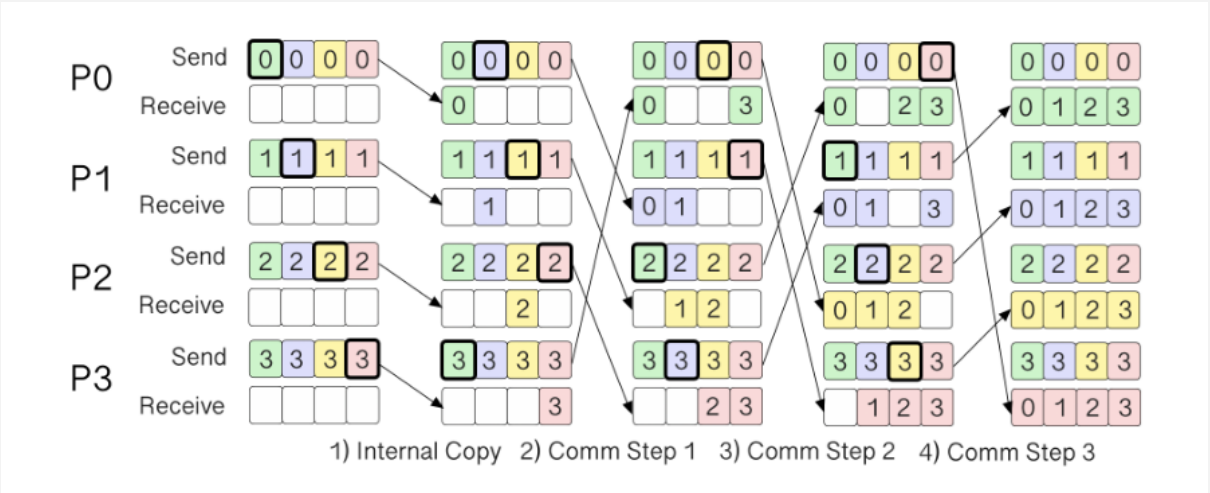
To	Praveen Kumar S. Kripa Shankar Person
Cc	Manoj Kumar Singh Navin Kumar Kiruthik Kanna Anuj Kumar
Bcc	Person
Subject	[EOD REPORT] Adhitya Charan - 20-11-2025

Subject : AlltoAll Operation in MOE and CollCommOperations Benchmarking

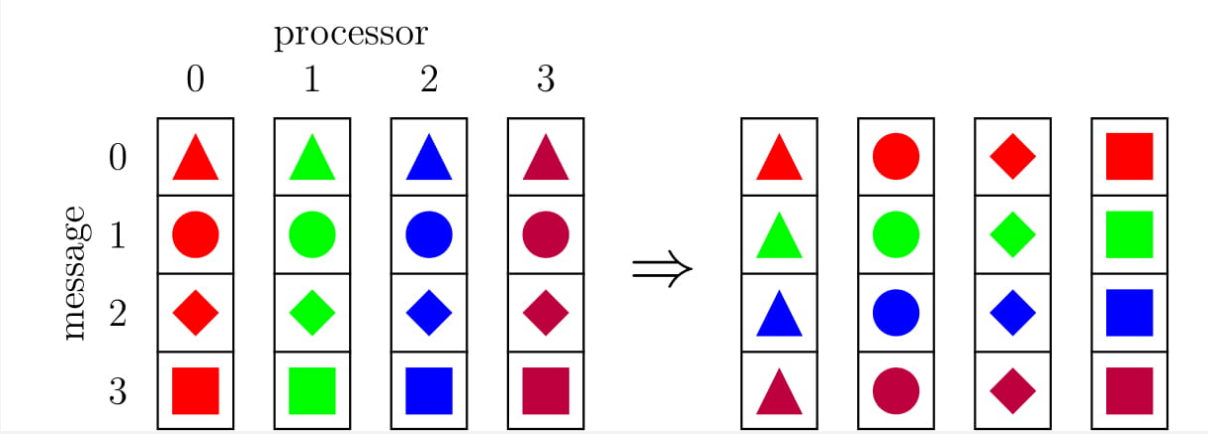
All to All Operation

All to All is another collective communication operation
In it there are a set of processing elements. Among a set of processing elements (nodes) each node has distinct (personalized) data items destined for each of the other nodes.

The all-to-all operation accomplishes this total data exchange among the set of nodes, such that each node ends up having an individual data item from each of the other nodes.



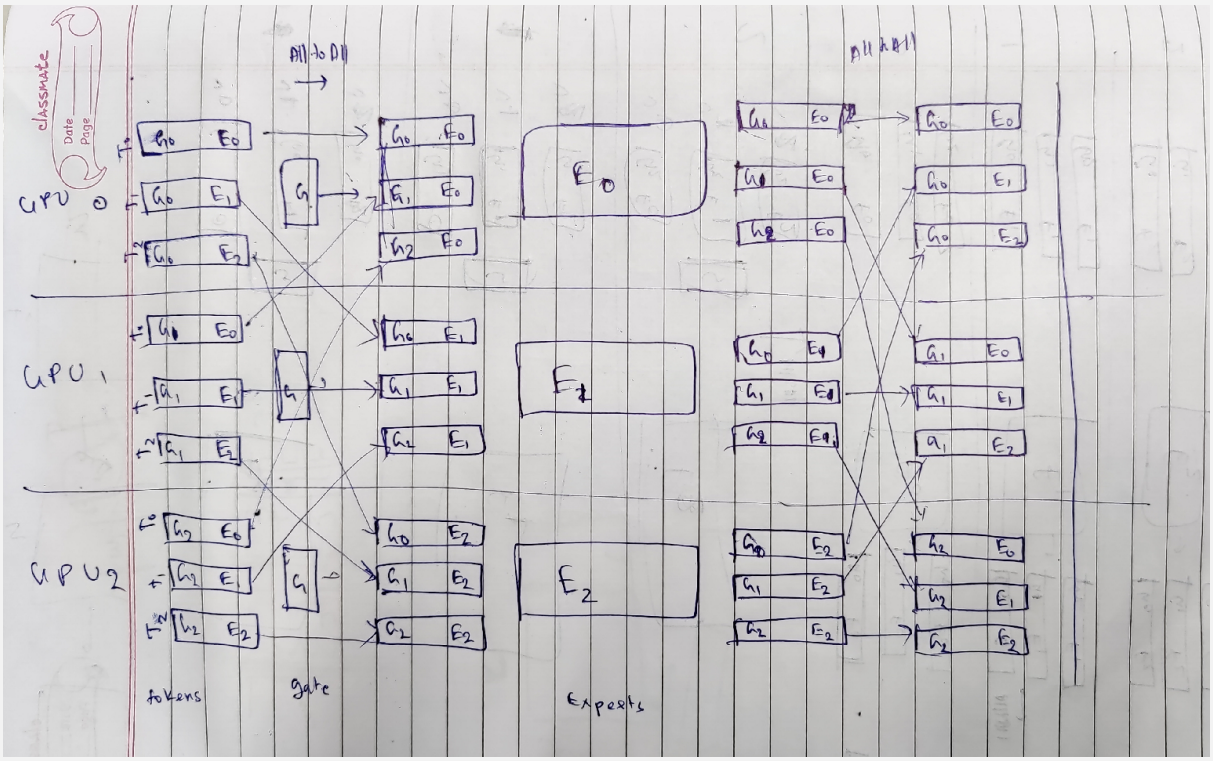
All to All is a Transpose operation on each sub element in the data in each GPU vs the GPU in which the data resides on.



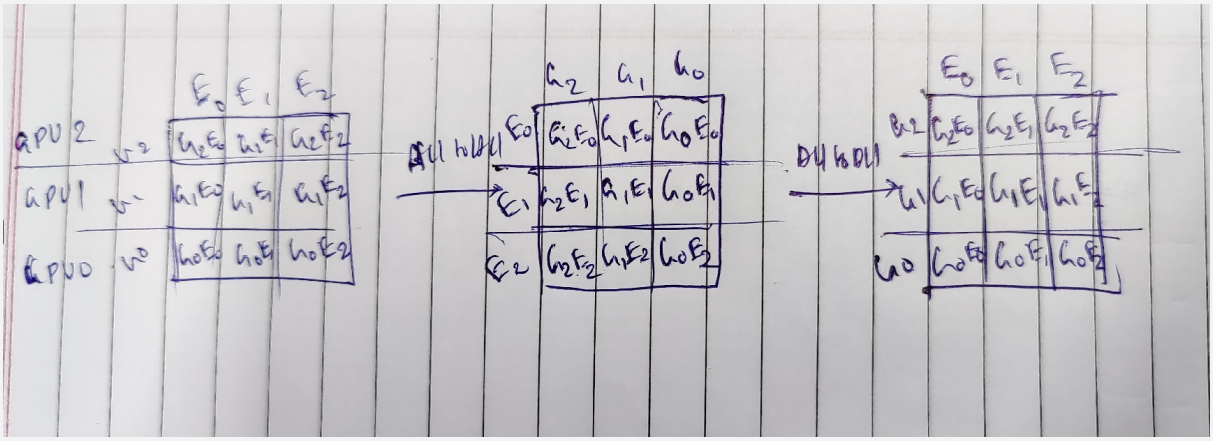
This kind of data transfer is useful while working on Mixture of experts because the Experts are located in each GPU separately but the tokens are available in all the GPUs

Once they are allocated in each GPU based on the router. Then they all need to be reallocated such that they are gathered together in each GPU based on what expert it is assigned to (whichever GPU has that particular expert) rather than the GPU in which it was found in, so they need to perform an All to All collective communication operation again.

And once the tokens are passed through the expert block, they need to be reallocated back again to be gathered together based on GPUs in which they came from again.
So there needs to be an All to All operation performed again.



This way there is a Transpose operation that is performed on the matrix denoting the GPU in which the token needs to reside vs Expert that the token needs to go to.



I also looked into refining the benchmarking of the Collective Communication Operations to stress test how much memory each operation could allocate and transfer and with how much latency.

I also made the All Reduce Operation run by combining the Reduce Scatter and All Gather operations together.

```cpp
void benchmarkRSplusAG(const std::vector<long unsigned int>& elmts) {
    if (rank_ == 0) {
        std::cout << "--- ReduceScatter + AllGather ---\n";
    }

    for (long unsigned int count : elmts) {

        long unsigned int small_count = count / world_size_;
        if (small_count == 0) continue;  // skip if too smalll

        auto op = [&](float* d_data, long unsigned int count) {

            NCCL_CHECK(ncclReduceScatter(d_data, d_data + rank_ * small_count, small_count, ncclFloat, ncclSum, nccl_comm_, stream_));

            NCCL_CHECK(ncclAllGather(d_data + rank_ * small_count, d_data, small_count, ncclFloat, nccl_comm_, stream_));

        };

        double latency_ms = benchmarkOp(op, count, 0);
        double size_mb = (count * sizeof(float)) / (1024.0 * 1024.0);
        double bandwidth_gbps = (size_mb / 1024.0) / (latency_ms / 1000.0);

        if (rank_ == 0) {
            printResult(size_mb, latency_ms, bandwidth_gbps);
        }
    }

    if (rank_ == 0) std::cout << "\n";
}
```

I was able to make all the operations run until 10.75 GB but once it crossed that the memory started to go out of bound.

```
NCCL Collectives Benchmark (2 GPUs)
--- AllReduce ---
Size: 10752.00 MB | Latency:  2889.97 ms | Bandwidth:     3.63 GB/s

--- Reduce ---
Size: 10752.00 MB | Latency:  1699.40 ms | Bandwidth:     6.18 GB/s

--- Broadcast ---
Size: 10752.00 MB | Latency:  1706.69 ms | Bandwidth:     6.15 GB/s

--- AllGather ---
Size: 10752.00 MB | Latency:  1765.81 ms | Bandwidth:     5.95 GB/s

--- Gather ---
Size: 10752.00 MB | Latency:   842.55 ms | Bandwidth:    12.46 GB/s

--- ReduceScatter ---
Size: 10752.00 MB | Latency:  1700.59 ms | Bandwidth:     6.17 GB/s

--- Scatter ---
Size: 10752.00 MB | Latency:   853.20 ms | Bandwidth:    12.31 GB/s

--- AlltoAll ---
Size: 10752.00 MB | Latency:   725.42 ms | Bandwidth:    14.47 GB/s

--- ReduceScatter + AllGather ---
Size: 10752.00 MB | Latency:  3452.59 ms | Bandwidth:     3.04 GB/s

Benchmark Complete
```

```
Every 0.1s: nvidia-smi                                  blubridge25-MS-7E06: Thu Nov 27 16:13:47 2025

Thu Nov 27 16:13:47 2025
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 580.95.05              Driver Version: 580.95.05      CUDA Version: 13.0     |
+-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GeForce RTX 3060        On  |   00000000:01:00.0  On |                  N/A |
| 34%   59C    P2             58W /  170W |   11423MiB /  12288MiB |    100%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+
|   1  NVIDIA GeForce RTX 3060        On  |   00000000:05:00.0 Off |                  N/A |
|  0%   61C    P2             51W /  170W |   10971MiB /  12288MiB |    100%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI              PID   Type   Process name                        GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|    0   N/A  N/A            3475      G   /usr/lib/xorg/Xorg                      192MiB |
|    0   N/A  N/A            3692      G   /usr/bin/gnome-shell                     73MiB |
|    0   N/A  N/A            9116      G   /proc/self/exe                           79MiB |
|    0   N/A  N/A          100517      G   ...share/antigravity/antigravity         76MiB |
|    0   N/A  N/A          358516      C   ./benchmarks/nccl_benchmark           10948MiB |
|    1   N/A  N/A            3475      G   /usr/lib/xorg/Xorg                       4MiB |
|    1   N/A  N/A          358517      C   ./benchmarks/nccl_benchmark           10948MiB |
+-----------------------------------------------------------------------------------------+
```

AllReduce

```
NCCL Collectives Benchmark (2 GPUs)
--- AllReduce ---
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.09 GB/s
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.31 GB/s
Size:     0.02 MB | Latency:     0.10 ms | Bandwidth:     0.15 GB/s
Size:     0.06 MB | Latency:     0.10 ms | Bandwidth:     0.63 GB/s
Size:     0.25 MB | Latency:     0.10 ms | Bandwidth:     2.56 GB/s
Size:     1.00 MB | Latency:     0.30 ms | Bandwidth:     3.22 GB/s
Size:     4.00 MB | Latency:     1.12 ms | Bandwidth:     3.48 GB/s
Size:    16.00 MB | Latency:     4.35 ms | Bandwidth:     3.59 GB/s
Size:    64.00 MB | Latency:    17.27 ms | Bandwidth:     3.62 GB/s
Size:   256.00 MB | Latency:    68.33 ms | Bandwidth:     3.66 GB/s
Size:  1024.00 MB | Latency:   274.09 ms | Bandwidth:     3.65 GB/s
Size:  2048.00 MB | Latency:   549.74 ms | Bandwidth:     3.64 GB/s
Size:  3072.00 MB | Latency:   823.88 ms | Bandwidth:     3.64 GB/s
Size:  4096.00 MB | Latency:  1099.95 ms | Bandwidth:     3.64 GB/s
Size:  8192.00 MB | Latency:  2206.40 ms | Bandwidth:     3.63 GB/s
Size: 10752.00 MB | Latency:  2896.98 ms | Bandwidth:     3.62 GB/s
```

Reduce

```
--- Reduce ---
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.15 GB/s
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.55 GB/s
Size:     0.02 MB | Latency:     0.01 ms | Bandwidth:     1.35 GB/s
Size:     0.06 MB | Latency:     0.04 ms | Bandwidth:     1.67 GB/s
Size:     0.25 MB | Latency:     0.05 ms | Bandwidth:     4.55 GB/s
Size:     1.00 MB | Latency:     0.18 ms | Bandwidth:     5.55 GB/s
Size:     4.00 MB | Latency:     0.68 ms | Bandwidth:     5.77 GB/s
Size:    16.00 MB | Latency:     2.57 ms | Bandwidth:     6.07 GB/s
Size:    64.00 MB | Latency:    10.17 ms | Bandwidth:     6.15 GB/s
Size:   256.00 MB | Latency:    40.48 ms | Bandwidth:     6.18 GB/s
Size:  1024.00 MB | Latency:   162.27 ms | Bandwidth:     6.16 GB/s
Size:  2048.00 MB | Latency:   324.74 ms | Bandwidth:     6.16 GB/s
Size:  3072.00 MB | Latency:   486.57 ms | Bandwidth:     6.17 GB/s
Size:  4096.00 MB | Latency:   648.48 ms | Bandwidth:     6.17 GB/s
Size:  8192.00 MB | Latency:  1297.59 ms | Bandwidth:     6.17 GB/s
Size: 10752.00 MB | Latency:  1703.17 ms | Bandwidth:     6.16 GB/s
```

Broadcast

```
--- Broadcast ---
Size:     0.00 MB | Latency:     0.05 ms | Bandwidth:     0.02 GB/s
Size:     0.00 MB | Latency:     0.05 ms | Bandwidth:     0.07 GB/s
Size:     0.02 MB | Latency:     0.06 ms | Bandwidth:     0.27 GB/s
Size:     0.06 MB | Latency:     0.06 ms | Bandwidth:     0.98 GB/s
Size:     0.25 MB | Latency:     0.09 ms | Bandwidth:     2.79 GB/s
Size:     1.00 MB | Latency:     0.13 ms | Bandwidth:     7.34 GB/s
Size:     4.00 MB | Latency:     0.68 ms | Bandwidth:     5.73 GB/s
Size:    16.00 MB | Latency:     2.58 ms | Bandwidth:     6.06 GB/s
Size:    64.00 MB | Latency:    10.24 ms | Bandwidth:     6.11 GB/s
Size:   256.00 MB | Latency:    41.07 ms | Bandwidth:     6.09 GB/s
Size:  1024.00 MB | Latency:   163.12 ms | Bandwidth:     6.13 GB/s
Size:  2048.00 MB | Latency:   326.68 ms | Bandwidth:     6.12 GB/s
Size:  3072.00 MB | Latency:   488.17 ms | Bandwidth:     6.15 GB/s
Size:  4096.00 MB | Latency:   652.41 ms | Bandwidth:     6.13 GB/s
Size:  8192.00 MB | Latency:  1305.75 ms | Bandwidth:     6.13 GB/s
Size: 10752.00 MB | Latency:  1712.74 ms | Bandwidth:     6.13 GB/s
```

AllGather

```
--- AllGather ---
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.12 GB/s
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.41 GB/s
Size:     0.02 MB | Latency:     0.01 ms | Bandwidth:     1.18 GB/s
Size:     0.06 MB | Latency:     0.03 ms | Bandwidth:     2.06 GB/s
Size:     0.25 MB | Latency:     0.06 ms | Bandwidth:     4.15 GB/s
Size:     1.00 MB | Latency:     0.18 ms | Bandwidth:     5.37 GB/s
Size:     4.00 MB | Latency:     0.71 ms | Bandwidth:     5.51 GB/s
Size:    16.00 MB | Latency:     2.62 ms | Bandwidth:     5.97 GB/s
Size:    64.00 MB | Latency:    10.55 ms | Bandwidth:     5.92 GB/s
Size:   256.00 MB | Latency:    42.08 ms | Bandwidth:     5.94 GB/s
Size:  1024.00 MB | Latency:   168.26 ms | Bandwidth:     5.94 GB/s
Size:  2048.00 MB | Latency:   337.02 ms | Bandwidth:     5.93 GB/s
Size:  3072.00 MB | Latency:   505.04 ms | Bandwidth:     5.94 GB/s
Size:  4096.00 MB | Latency:   673.89 ms | Bandwidth:     5.94 GB/s
Size:  8192.00 MB | Latency:  1350.22 ms | Bandwidth:     5.92 GB/s
Size: 10752.00 MB | Latency:  1774.85 ms | Bandwidth:     5.92 GB/s
```

Gather

```
--- Gather ---
Size:     0.00 MB | Latency:     0.03 ms | Bandwidth:     0.04 GB/s
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.56 GB/s
Size:     0.02 MB | Latency:     0.01 ms | Bandwidth:     1.92 GB/s
Size:     0.06 MB | Latency:     0.01 ms | Bandwidth:     4.85 GB/s
Size:     0.25 MB | Latency:     0.02 ms | Bandwidth:    15.02 GB/s
Size:     1.00 MB | Latency:     0.09 ms | Bandwidth:    10.80 GB/s
Size:     4.00 MB | Latency:     0.38 ms | Bandwidth:    10.18 GB/s
Size:    16.00 MB | Latency:     1.31 ms | Bandwidth:    11.89 GB/s
Size:    64.00 MB | Latency:     5.01 ms | Bandwidth:    12.46 GB/s
Size:   256.00 MB | Latency:    20.23 ms | Bandwidth:    12.36 GB/s
Size:  1024.00 MB | Latency:    81.14 ms | Bandwidth:    12.32 GB/s
Size:  2048.00 MB | Latency:   165.35 ms | Bandwidth:    12.10 GB/s
Size:  3072.00 MB | Latency:   250.76 ms | Bandwidth:    11.96 GB/s
Size:  4096.00 MB | Latency:   327.81 ms | Bandwidth:    12.20 GB/s
Size:  8192.00 MB | Latency:   649.96 ms | Bandwidth:    12.31 GB/s
Size: 10752.00 MB | Latency:   860.09 ms | Bandwidth:    12.21 GB/s
```

ReduceScatter

```
--- ReduceScatter ---
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.12 GB/s
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.41 GB/s
Size:     0.02 MB | Latency:     0.01 ms | Bandwidth:     1.21 GB/s
Size:     0.06 MB | Latency:     0.03 ms | Bandwidth:     2.13 GB/s
Size:     0.25 MB | Latency:     0.11 ms | Bandwidth:     2.30 GB/s
Size:     1.00 MB | Latency:     0.18 ms | Bandwidth:     5.55 GB/s
Size:     4.00 MB | Latency:     0.73 ms | Bandwidth:     5.39 GB/s
Size:    16.00 MB | Latency:     2.56 ms | Bandwidth:     6.09 GB/s
Size:    64.00 MB | Latency:    10.33 ms | Bandwidth:     6.05 GB/s
Size:   256.00 MB | Latency:    40.77 ms | Bandwidth:     6.13 GB/s
Size:  1024.00 MB | Latency:   164.07 ms | Bandwidth:     6.10 GB/s
Size:  2048.00 MB | Latency:   326.52 ms | Bandwidth:     6.13 GB/s
Size:  3072.00 MB | Latency:   488.61 ms | Bandwidth:     6.14 GB/s
Size:  4096.00 MB | Latency:   651.38 ms | Bandwidth:     6.14 GB/s
Size:  8192.00 MB | Latency:  1302.49 ms | Bandwidth:     6.14 GB/s
Size: 10752.00 MB | Latency:  1706.56 ms | Bandwidth:     6.15 GB/s
```

Scatter

```
--- Scatter ---
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.11 GB/s
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.42 GB/s
Size:     0.02 MB | Latency:     0.01 ms | Bandwidth:     2.13 GB/s
Size:     0.06 MB | Latency:     0.01 ms | Bandwidth:     6.17 GB/s
Size:     0.25 MB | Latency:     0.02 ms | Bandwidth:    14.54 GB/s
Size:     1.00 MB | Latency:     0.10 ms | Bandwidth:     9.34 GB/s
Size:     4.00 MB | Latency:     0.45 ms | Bandwidth:     8.65 GB/s
Size:    16.00 MB | Latency:     1.35 ms | Bandwidth:    11.60 GB/s
Size:    64.00 MB | Latency:     5.13 ms | Bandwidth:    12.19 GB/s
Size:   256.00 MB | Latency:    20.32 ms | Bandwidth:    12.30 GB/s
Size:  1024.00 MB | Latency:    81.67 ms | Bandwidth:    12.24 GB/s
Size:  2048.00 MB | Latency:   162.80 ms | Bandwidth:    12.29 GB/s
Size:  3072.00 MB | Latency:   244.33 ms | Bandwidth:    12.28 GB/s
Size:  4096.00 MB | Latency:   325.54 ms | Bandwidth:    12.29 GB/s
Size:  8192.00 MB | Latency:   651.59 ms | Bandwidth:    12.28 GB/s
Size: 10752.00 MB | Latency:   854.41 ms | Bandwidth:    12.29 GB/s
```

AlltoAll

```
--- AlltoAll ---
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.10 GB/s
Size:     0.00 MB | Latency:     0.01 ms | Bandwidth:     0.41 GB/s
Size:     0.02 MB | Latency:     0.01 ms | Bandwidth:     1.29 GB/s
Size:     0.06 MB | Latency:     0.04 ms | Bandwidth:     1.73 GB/s
Size:     0.25 MB | Latency:     0.03 ms | Bandwidth:     7.38 GB/s
Size:     1.00 MB | Latency:     0.08 ms | Bandwidth:    11.60 GB/s
Size:     4.00 MB | Latency:     0.33 ms | Bandwidth:    11.84 GB/s
Size:    16.00 MB | Latency:     1.30 ms | Bandwidth:    12.03 GB/s
Size:    64.00 MB | Latency:     4.36 ms | Bandwidth:    14.35 GB/s
Size:   256.00 MB | Latency:    17.34 ms | Bandwidth:    14.41 GB/s
Size:  1024.00 MB | Latency:    69.95 ms | Bandwidth:    14.30 GB/s
Size:  2048.00 MB | Latency:   138.59 ms | Bandwidth:    14.43 GB/s
Size:  3072.00 MB | Latency:   208.72 ms | Bandwidth:    14.37 GB/s
Size:  4096.00 MB | Latency:   277.81 ms | Bandwidth:    14.40 GB/s
Size:  8192.00 MB | Latency:   559.27 ms | Bandwidth:    14.30 GB/s
Size: 10752.00 MB | Latency:   733.37 ms | Bandwidth:    14.32 GB/s
```

ReduceScatter + AllGather (AllReduce)

```
--- ReduceScatter + AllGather ---
Size:     0.00 MB | Latency:     0.02 ms | Bandwidth:     0.06 GB/s
Size:     0.00 MB | Latency:     0.02 ms | Bandwidth:     0.21 GB/s
Size:     0.02 MB | Latency:     0.03 ms | Bandwidth:     0.57 GB/s
Size:     0.06 MB | Latency:     0.06 ms | Bandwidth:     1.05 GB/s
Size:     0.25 MB | Latency:     0.11 ms | Bandwidth:     2.15 GB/s
Size:     1.00 MB | Latency:     0.35 ms | Bandwidth:     2.83 GB/s
Size:     4.00 MB | Latency:     1.35 ms | Bandwidth:     2.89 GB/s
Size:    16.00 MB | Latency:     5.31 ms | Bandwidth:     2.94 GB/s
Size:    64.00 MB | Latency:    20.67 ms | Bandwidth:     3.02 GB/s
Size:   256.00 MB | Latency:    82.52 ms | Bandwidth:     3.03 GB/s
Size:  1024.00 MB | Latency:   329.58 ms | Bandwidth:     3.03 GB/s
Size:  2048.00 MB | Latency:   655.75 ms | Bandwidth:     3.05 GB/s
Size:  3072.00 MB | Latency:   981.92 ms | Bandwidth:     3.06 GB/s
Size:  4096.00 MB | Latency:  1313.44 ms | Bandwidth:     3.05 GB/s
Size:  8192.00 MB | Latency:  2627.62 ms | Bandwidth:     3.04 GB/s
Size: 10752.00 MB | Latency:  3448.99 ms | Bandwidth:     3.04 GB/s
```

I am planning to look into integrating autograd into DTensor tomorrow.

Adhitya Charan

Regards