

=====

AIramp-Accel v12.12B (Batched BLAS AG+GEMM + Baseline Fencing)

U.S. Patent No. 11_308_384 B1 - Licensed to Tensor Networks, Inc.

=====

[Checkpoint 1] Initializing HIP Runtime...

HIP Runtime Version: OK (60550421)

HIP Driver Version: OK (60550421)

Hints: pass /dev/kfd and /dev/dri, add groups video/render, set HIP_VISIBLE_DEVICES=0-7

=====

[Checkpoint 2] Found 8 AMD ROCm GPU(s).

[Checkpoint 3] Initializing RCCL for 8 GPUs...

[Checkpoint 4] RCCL Initialized.

[AIramp-Accel] Starting one-time autotuning simulation to build cost model...

[CostModel] For 'All-to-All', crossover to RING strategy at 64.00 MB

[CostModel] For 'Allgather + GEMM', crossover to PIPELINED strategy at 16.00 MB

[CostModel] For 'GEMM + ReduceScatter', crossover to FUSED strategy at 32.00 MB

[AIramp-Accel] Autotuning complete. Cost model is built.

Starting Benchmark Suite...

[>] 0/6 (0%)

>>> Testing Kernel: All-to-All -- Workload Size: 16.00 MB <<<

- Baseline Execution Time: 0.716 ms

- [AIRamp-Accel Analyzer] Decision: Workload is nominal. Baseline strategy is optimal.
- Optimized execution skipped: Baseline was determined to be optimal or in single-GPU mode.

[=====>] 1/6 (16%)

>>> Testing Kernel: Allgather + GEMM -- Workload Size: 16.00 MB <<<

- Baseline Execution Time: 217.195 ms
- [AIRamp-Accel Analyzer] Decision: Workload identified as anomalous (suboptimal for baseline). Dispatching PIPELINED strategy.
- Optimized Execution Time: 0.645 ms
- Verification: PASSED
- Speedup: 336.53x

[=====>] 2/6 (33%)

>>> Testing Kernel: GEMM + ReduceScatter -- Workload Size: 16.00 MB <<<

- Baseline Execution Time: 1.157 ms
- [AIRamp-Accel Analyzer] Decision: Workload is nominal. Baseline strategy is optimal.
- Optimized execution skipped: Baseline was determined to be optimal or in single-GPU mode.

[=====>] 3/6 (50%)

>>> Testing Kernel: All-to-All -- Workload Size: 256.00 MB <<<

- Baseline Execution Time: 1865.699 ms
- [AIRamp-Accel Analyzer] Decision: Workload identified as anomalous (suboptimal for baseline). Dispatching RING strategy.
- Optimized Execution Time: 5.506 ms
- Verification: PASSED
- Speedup: 338.83x

[=====>] 4/6 (66%)

>>> Testing Kernel: Allgather + GEMM -- Workload Size: 256.00 MB <<<

- Baseline Execution Time: 11.250 ms
- [AIRamp-Accel Analyzer] Decision: Workload identified as anomalous (suboptimal for baseline). Dispatching PIPELINED strategy.
- Optimized Execution Time: 8.922 ms
- Verification: PASSED
- Speedup: 1.26x

[=====>] 5/6 (83%)

>>> Testing Kernel: GEMM + ReduceScatter -- Workload Size: 256.00 MB <<<

- Baseline Execution Time: 10.359 ms
- [AIRamp-Accel Analyzer] Decision: Workload identified as anomalous (suboptimal for baseline). Dispatching FUSED strategy.
- Optimized Execution Time: 9.148 ms
- Verification: PASSED
- Speedup: 1.13x

[=====] 6/6 (100%)