Table 1: Ring Attention Performance: Triton Kernel vs No Triton

Context Length	**No Triton (ms)**	**Triton (ms)**	**Speedup**
256	109.03	110.72	0.98×
512	178.76	168.98	1.06×
1,024	306.77	265.90	1.15×
2,048	801.77	508.88	1.58×
4,096	2,423.17	1,193.98	2.03×
8,192	8,306.90	3,559.70	2.33×
16,384	OOM	11,580.84	–

Configuration: 2× A16 GPUs, Llama 3.2-1B, Ring Attention, Mixed Precision (FP16 weights, FP32 accumulation)