Sequence Length: 1M

Legend: Context Parallelism, Medha 2D Parallel

Speedup labels: 1.10x, 1.08x, 1.08x, 1.29x, 1.64x

X-axis: Number of GPUs (8, 16, 32, 64, 128)

Y-axis: Prefill Latency (s)