

Best Config: Pipeline Parallel Dim: 2, Tensor Parallel Dim: 2, Scheduler: Sarathi-Serve, Sarathi Chunk Size: 512, Batch Size: 256, SKU: H100
QPS per Dollar: 0.20

