

Best Config: Pipeline Parallel Dim: 1, Tensor Parallel Dim: 4, Scheduler: Sarathi-Serve, Sarathi Chunk Size: 512, Batch Size: 128, SKU: H100
QPS per Dollar: 0.03

