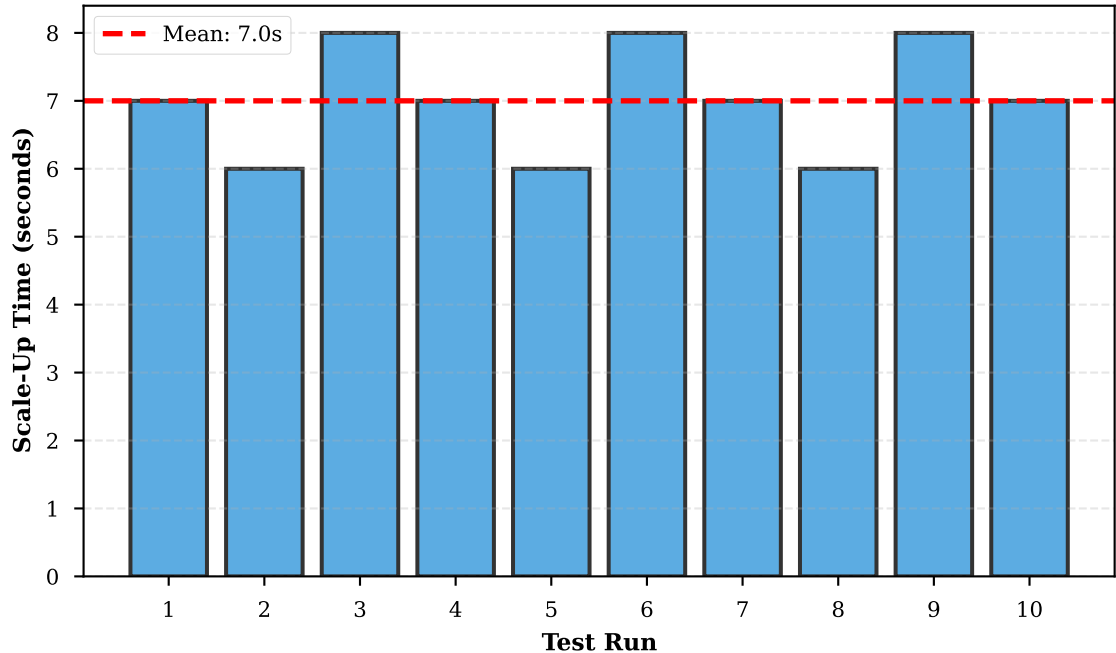
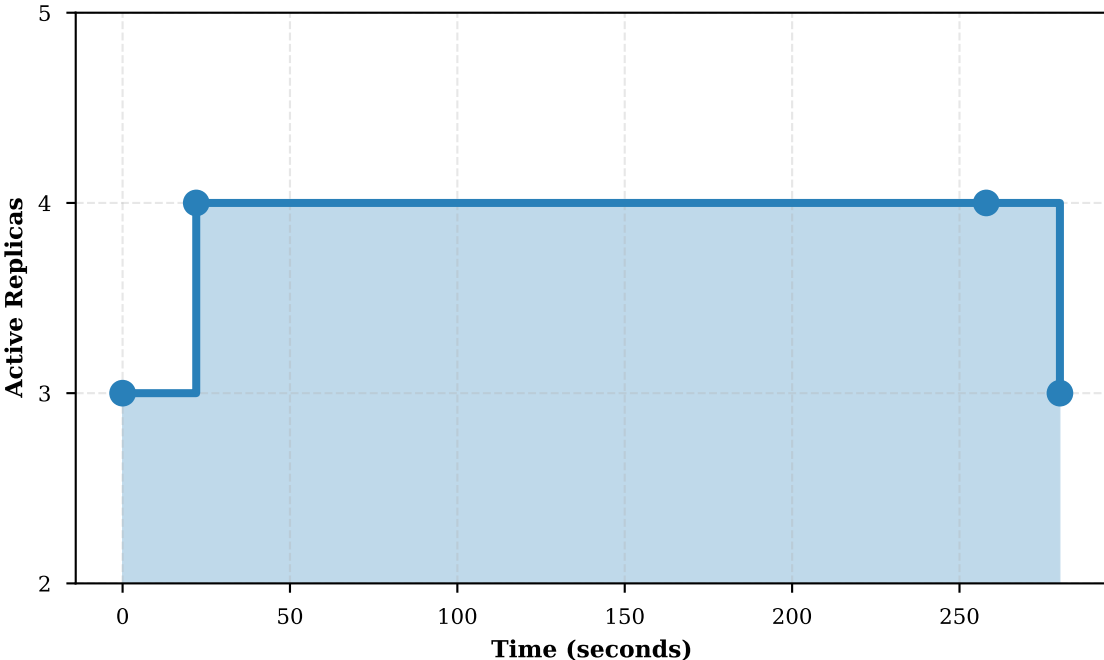


Scenario 2: Scaling Performance Analysis

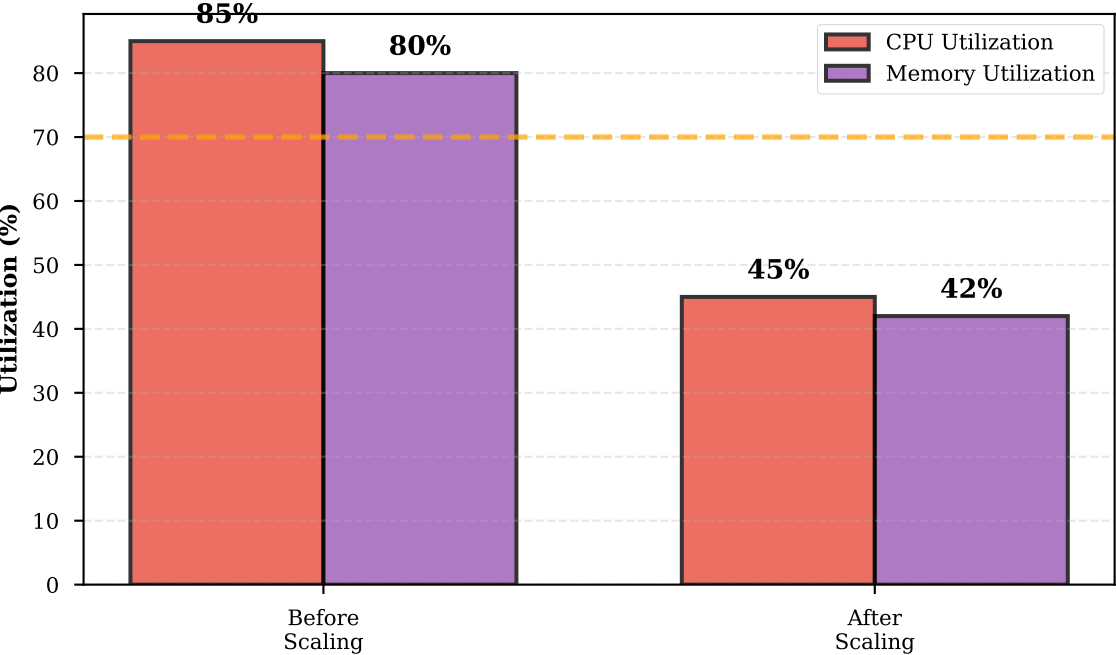
Scale-Up Latency Consistency



Replica Count Over Time



Resource Utilization: Pre vs Post Scaling



Cooldown Impact on System Stability

