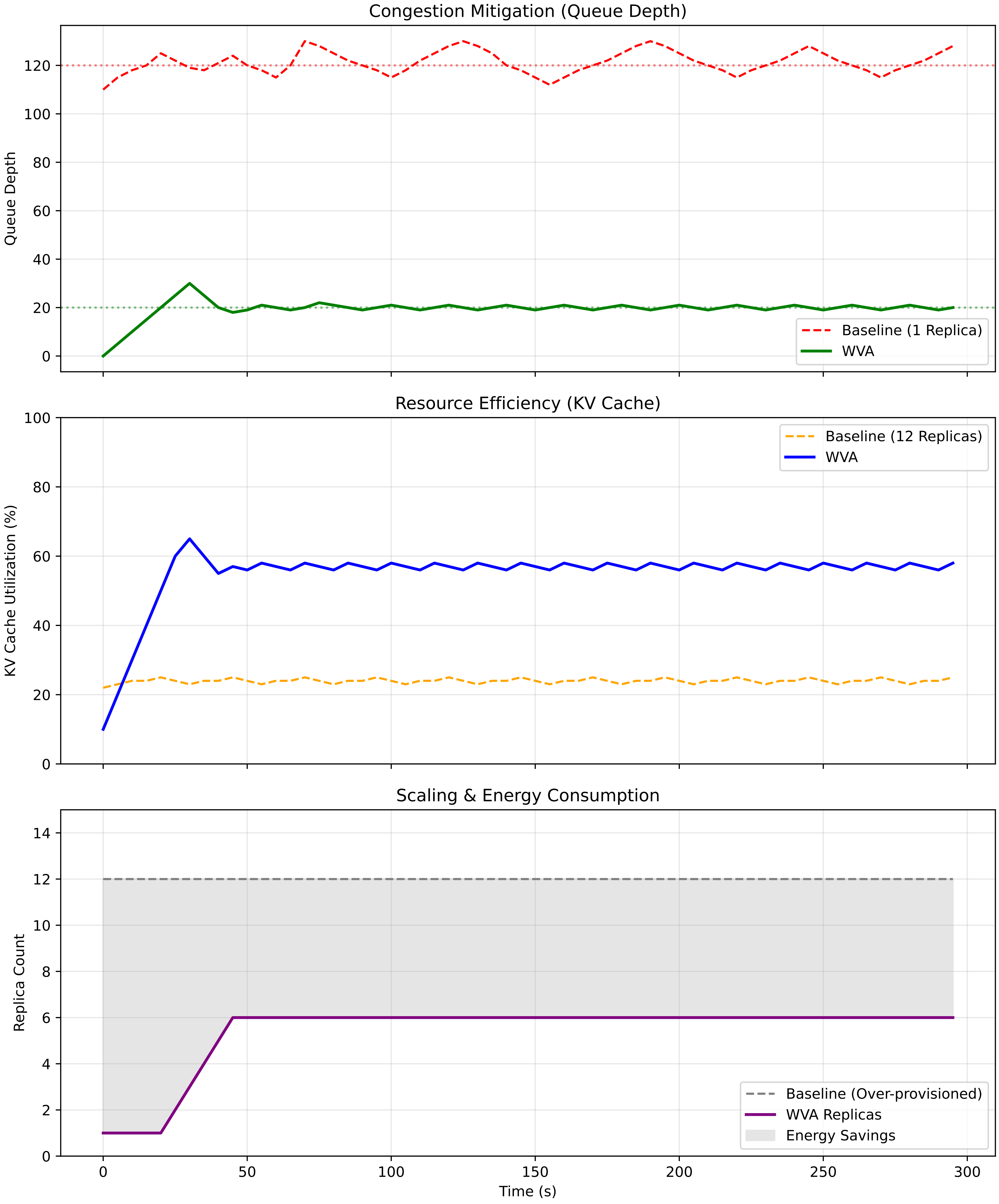
Congestion Mitigation (Queue Depth)
Queue Depth
0
20
40
60
80
100
120
Baseline (1 Replica)
WVA
Resource Efficiency (KV Cache)
KV Cache Utilization (%)
0
20
40
60
80
100
Baseline (12 Replicas)
WVA
Scaling & Energy Consumption
Replica Count
0
2
4
6
8
10
12
14
Baseline (Over-provisioned)
WVA Replicas
Energy Savings
0
50
100
150
200
250
300
Time (s)