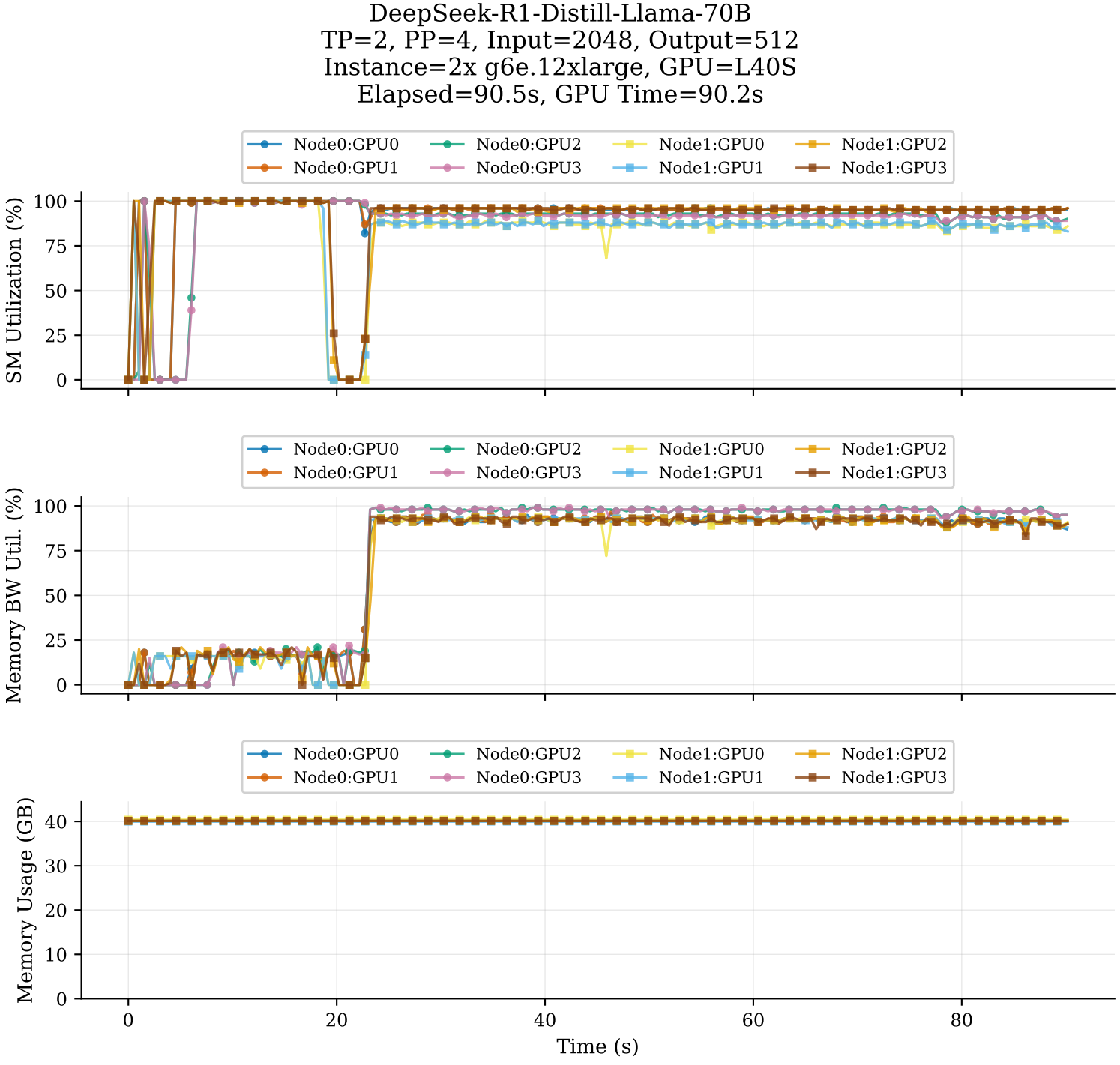
DeepSeek-R1-Distill-Llama-70B
TP=2, PP=4, Input=2048, Output=512
Instance=2x g6e.12xlarge, GPU=L40S
Elapsed=90.5s, GPU Time=90.2s
Node0:GPU0
Node0:GPU1
Node0:GPU2
Node0:GPU3
Node1:GPU0
Node1:GPU1
Node1:GPU2
Node1:GPU3
SM Utilization (%)
Memory BW Util. (%)
Memory Usage (GB)
Time (s)