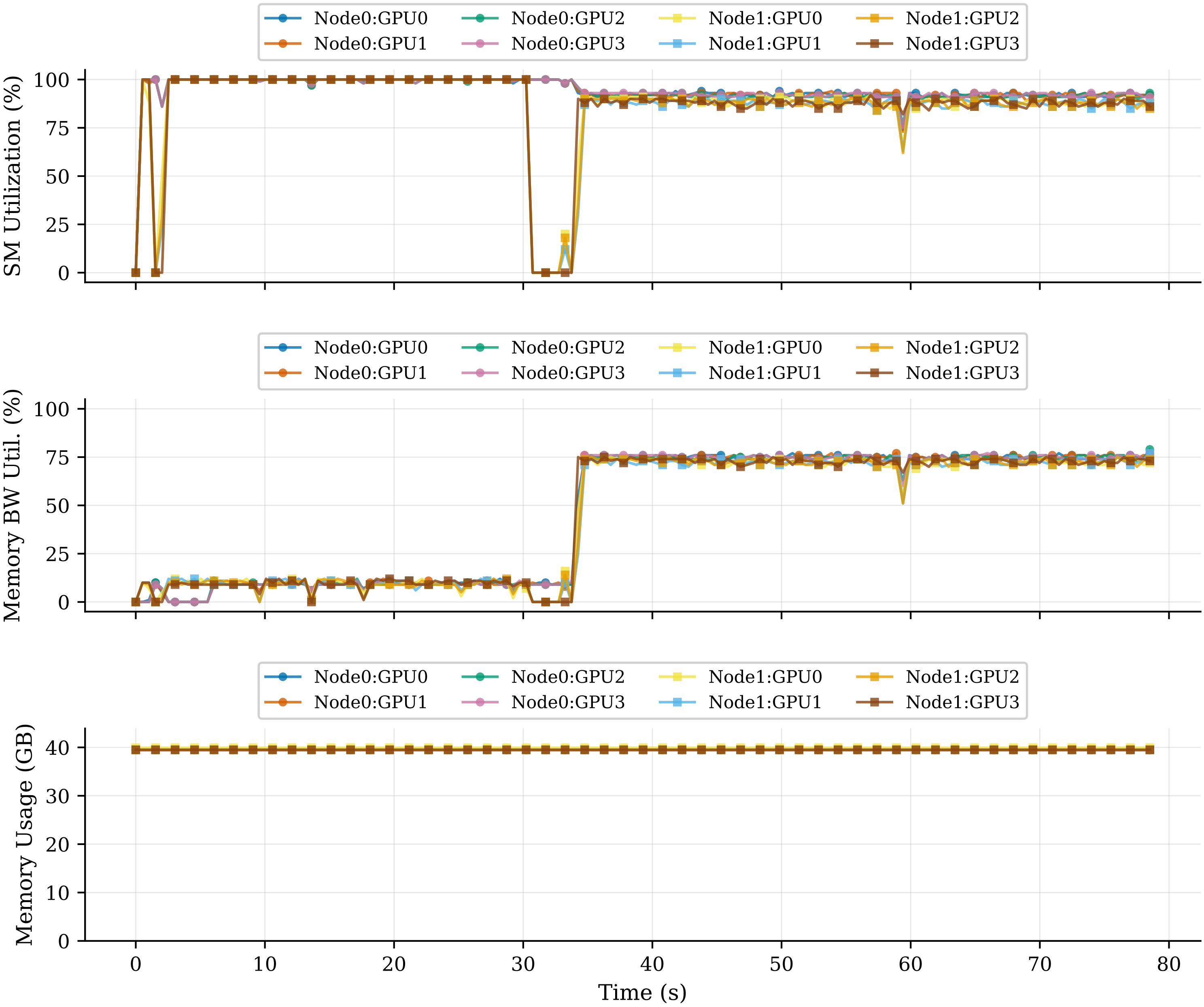

DeepSeek-R1-Distill-Llama-70B
TP=4, PP=2, Input=2048, Output=512
Instance=2x g6e.12xlarge, GPU=L40S
Elapsed=78.6s, GPU Time=78.5s