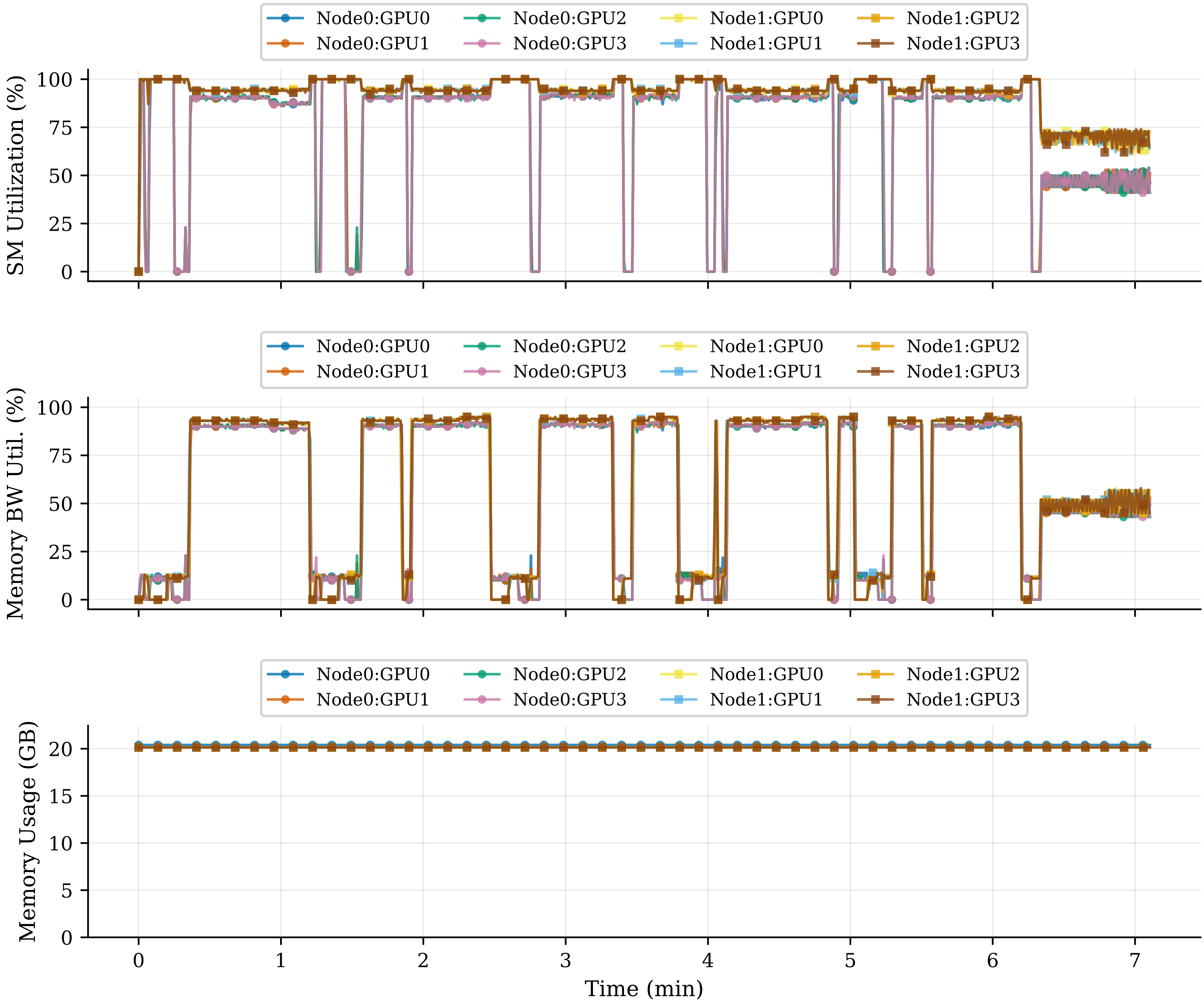
DeepSeek-R1-Distill-Llama-70B
TP=4, PP=2, Input=2048, Output=512
Instance=2x g5.12xlarge, GPU=A10G
Elapsed=426.4s, GPU Time=426.3s


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 (min)