Multi-GPU Scaling Efficiency (LLaMA-2-70B)