Training 4-layer GPT models on 8xA100 80GB SXM, 900GB RAM PyTorch(FSDP,bf16-true) PyTorch(FSDP,tf32) NNTile(tf32) Tflops/s 53.2 43.8 46.9 38.1 Model size, billions