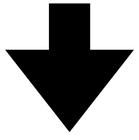
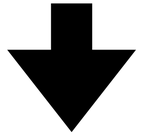
Pure Task Parallelism

- High Throughput Training
- High Processor Utilization
- Limited to training models that fit in GPU memory
- Suboptimal for heterogeneous training workloads

Pure Model Parallelism

- Enables training of larger-than-memory-models
- Low processor utilization
- Inefficient training for multi-model workloads
- Requires multiple devices to train larger-than-memory models





Hydra Shard Alternator Parallelism (SHARP)

- Enables training of larger-than-memory-models with a single device
- High Throughput Training
- High Processor Utilization
- Near-optimal for heterogeneous training workloads