DistributedTensor API

- DeviceMesh:
  - Shard: shard on tensor dimension across devices
  - Replicate: replicate across devices.
  - Partial: partition values across devices.

Distributed Tensor Types

- Shard
- Replicate
- Partial

Why DistributedTensor?

Distributed Training in research community currently research community train large scale models mainly in three ways:

1. Data Parallel partitions and train data in parallel, model is usually replicated across devices
2. Within-layer Model Parallel i.e. Tensor parallel that partition weights and computation
3. Pipeline Parallel partitions the computation graph and run pipeline across stages

Tensor Parallelism (TP) that builds on top of DistributedTensor

- Colwise Parallel:
  - Partition dim -1 of weight/bias matrix
  - Rowwise Parallel:
  - Partition 0 of weight/bias matrix
  - Pairwise Parallel:
  - Cascade a colwise with a rowwise as a pair

Tensor Parallel API Examples

- Colwise Parallel:
  - Initialize a new device mesh for Tensor Parallel
  - parallelize_module(model, custom_model, device_mesh)

- Pairwise Parallel Style for a Transformer model
  - parallelize_module(model, custom_model, device_mesh)

Preliminary Results on ViT

- Quantitative analysis for peak memory for one layer:
  - Bsz = 1204, FSDP use 1.3 times more memory, which is 1.1 in experiment
  - Bsz = 512, the ratio increases to 2.4