PyTorch/XLA Data Parallel with SPMD

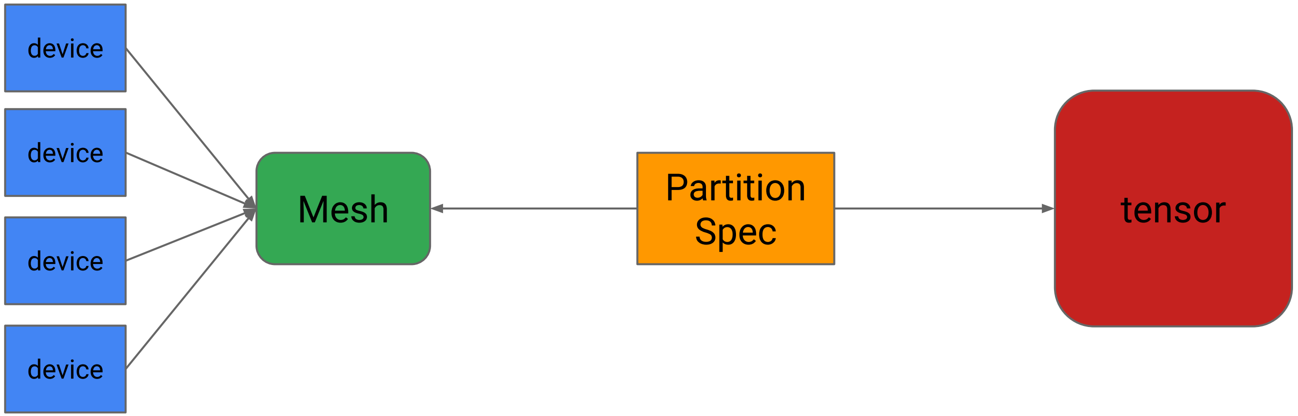
Agenda

1. Recap of SPMD
2. What’s data parallel
3. How data parallel + SPMD works

GSPMD

1. https://arxiv.org/abs/2105.04663
2. User only express sharding intention, let compiler shard the tensor for you.
3. User don’t need to shard every tensor, compiler will propagate the sharding for the user.
4. Collective ops(all_gather, reduce_scatter etc) will be added after compilation

Basic ideas


device

device

device

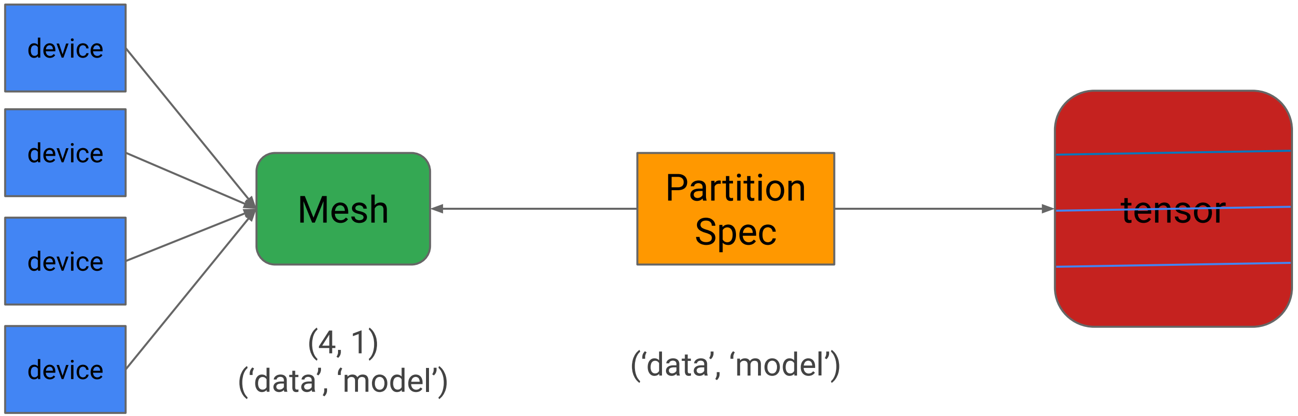
device

Mesh

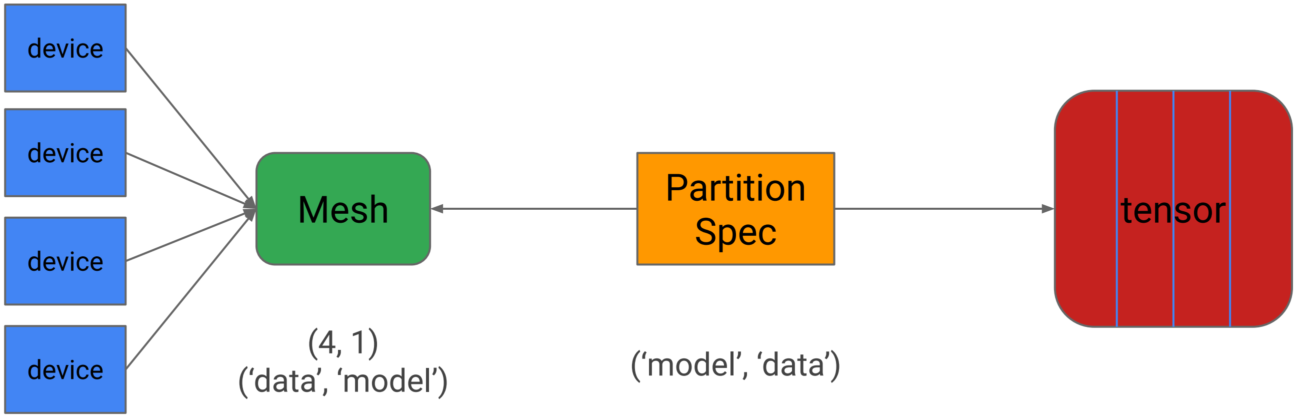
Partition Spec

tensor

Basic ideas


device
device
device
device
Mesh
Partition Spec
tensor
(4, 1)
('data', 'model')
('data', 'model')

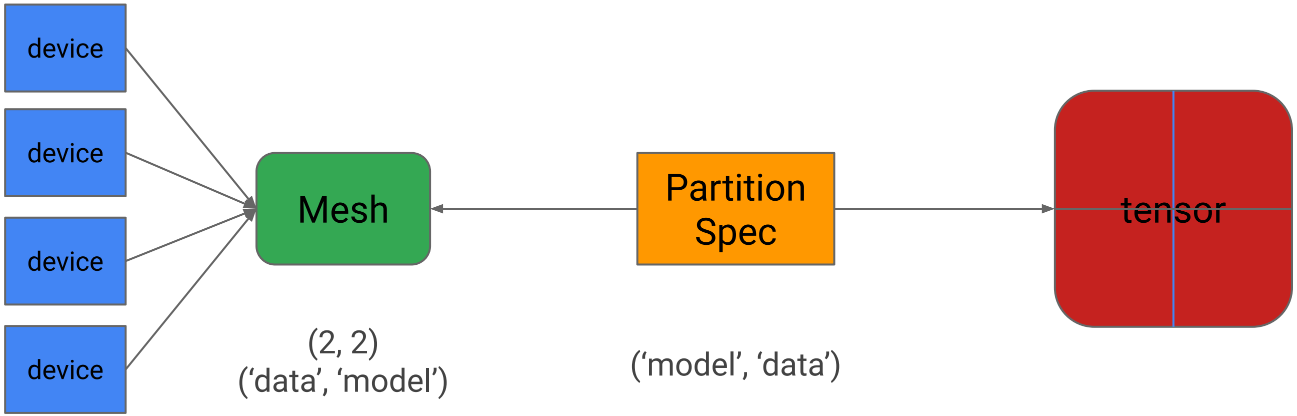
Basic ideas



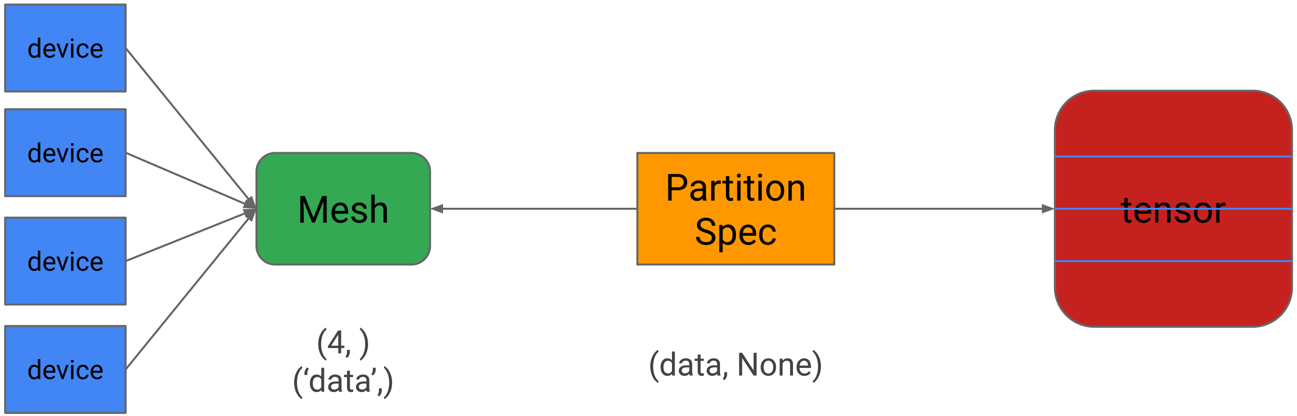
(4, 1)
('data', 'model')

('model', 'data')

Basic ideas


device
device
device
device
Mesh
Partition Spec
tensor
(2, 2)
(‘data’, ‘model’)
(‘model’, ‘data’)

Basic ideas



device

device

device

device

Mesh

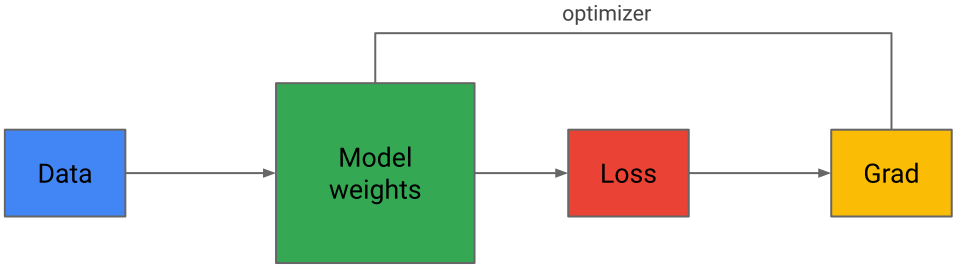
Partition Spec

tensor

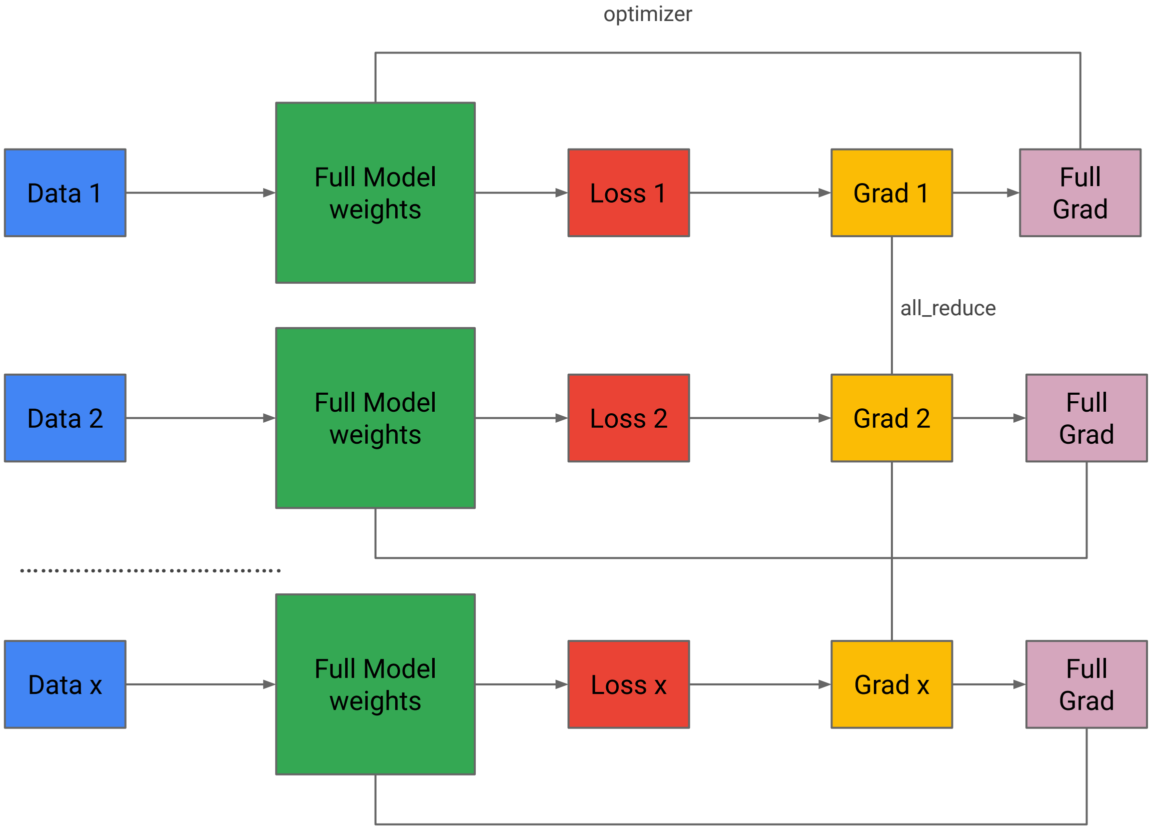
(4,)
(‘data’,)

(data, None)

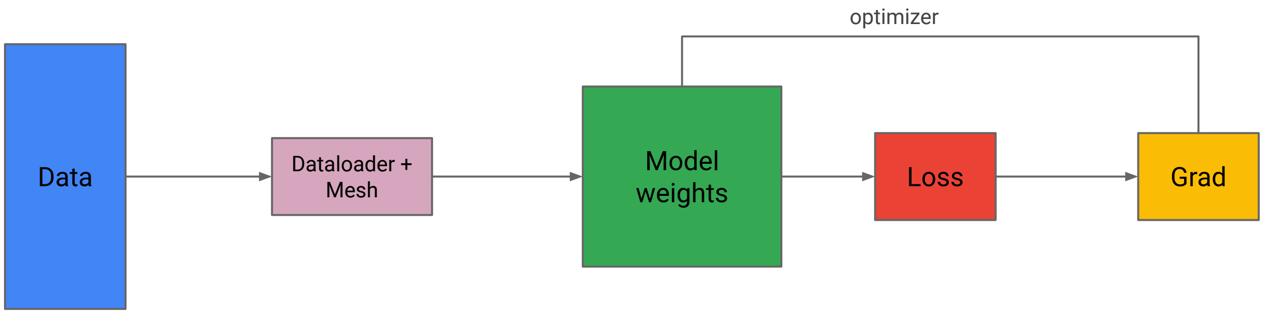
Single device



Data Parallel



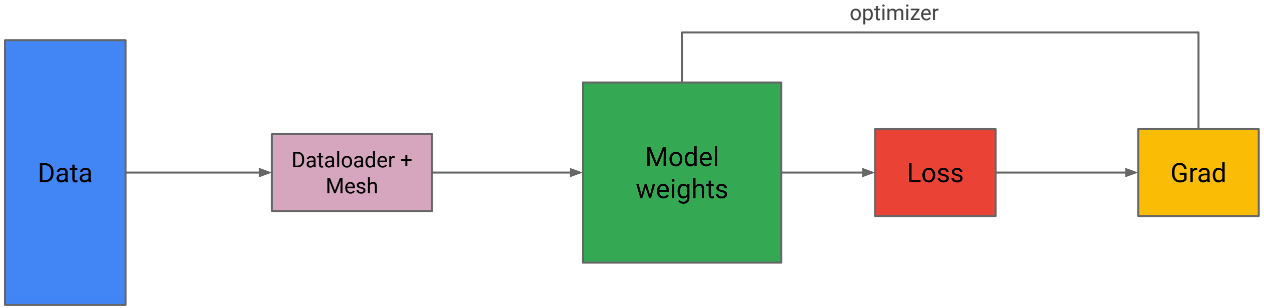
SPMD + Data Parallel



Sharded at batch

replicated

SPMD + Data Parallel



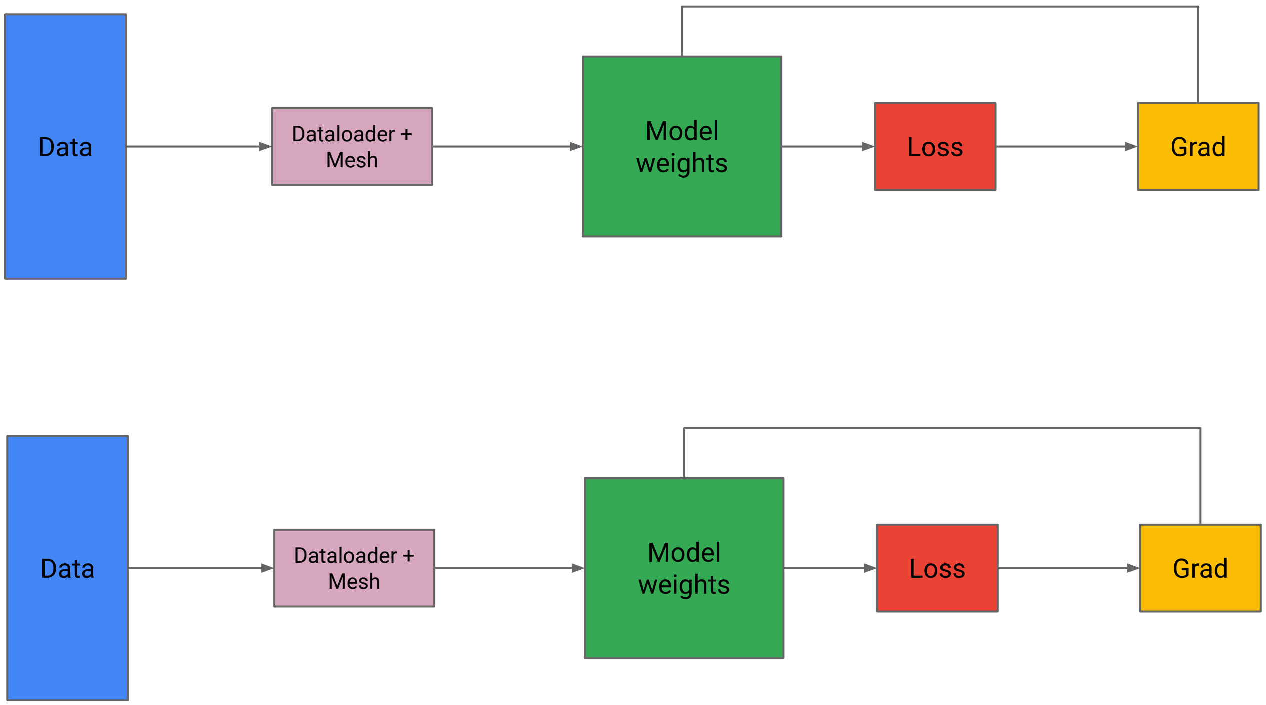
Sharded at batch

replicated

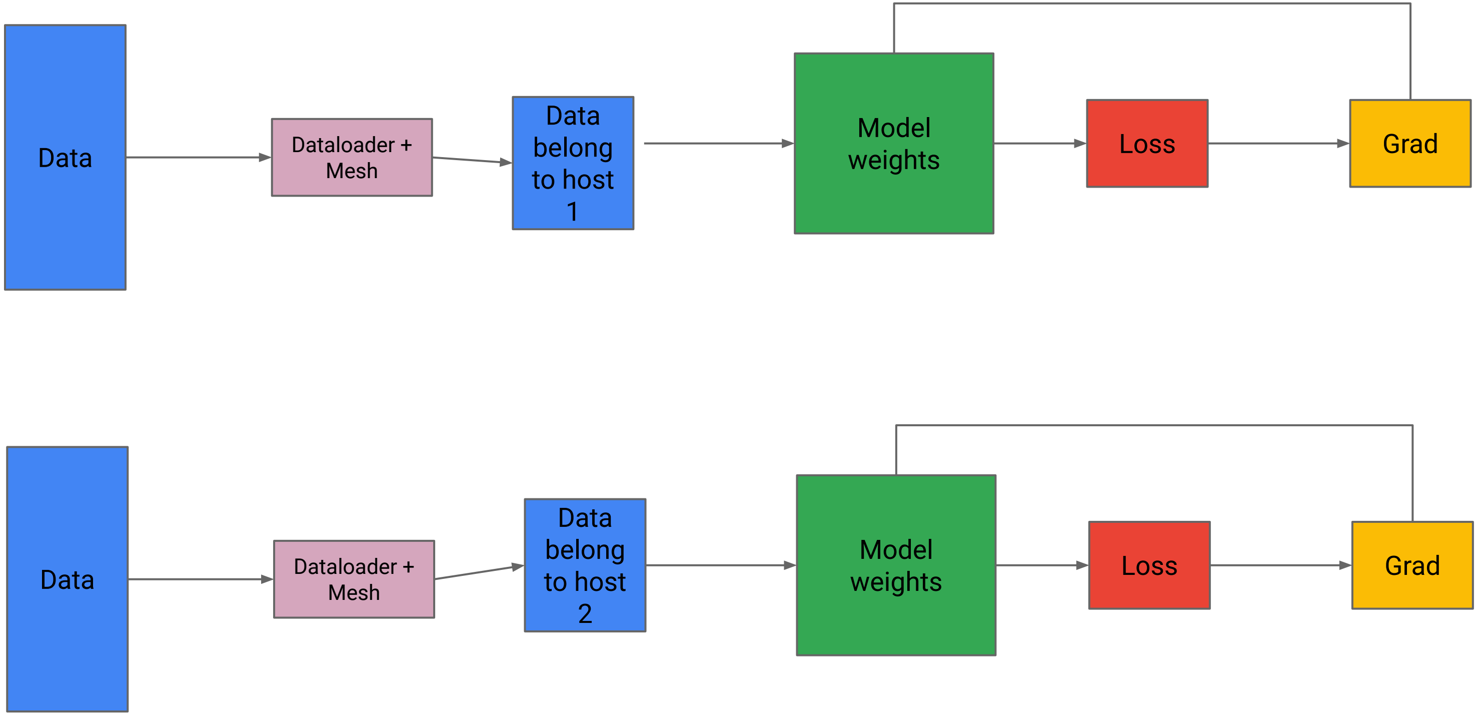
partial

Partial. Compiler will apply all_reduce

SPMD + Data Parallel + multi host



SPMD + Data Parallel + multi host


Data
Dataloader + Mesh
Data belong to host 1
Model weights
Loss
Grad
Data
Dataloader + Mesh
Data belong to host 2
Model weights
Loss
Grad