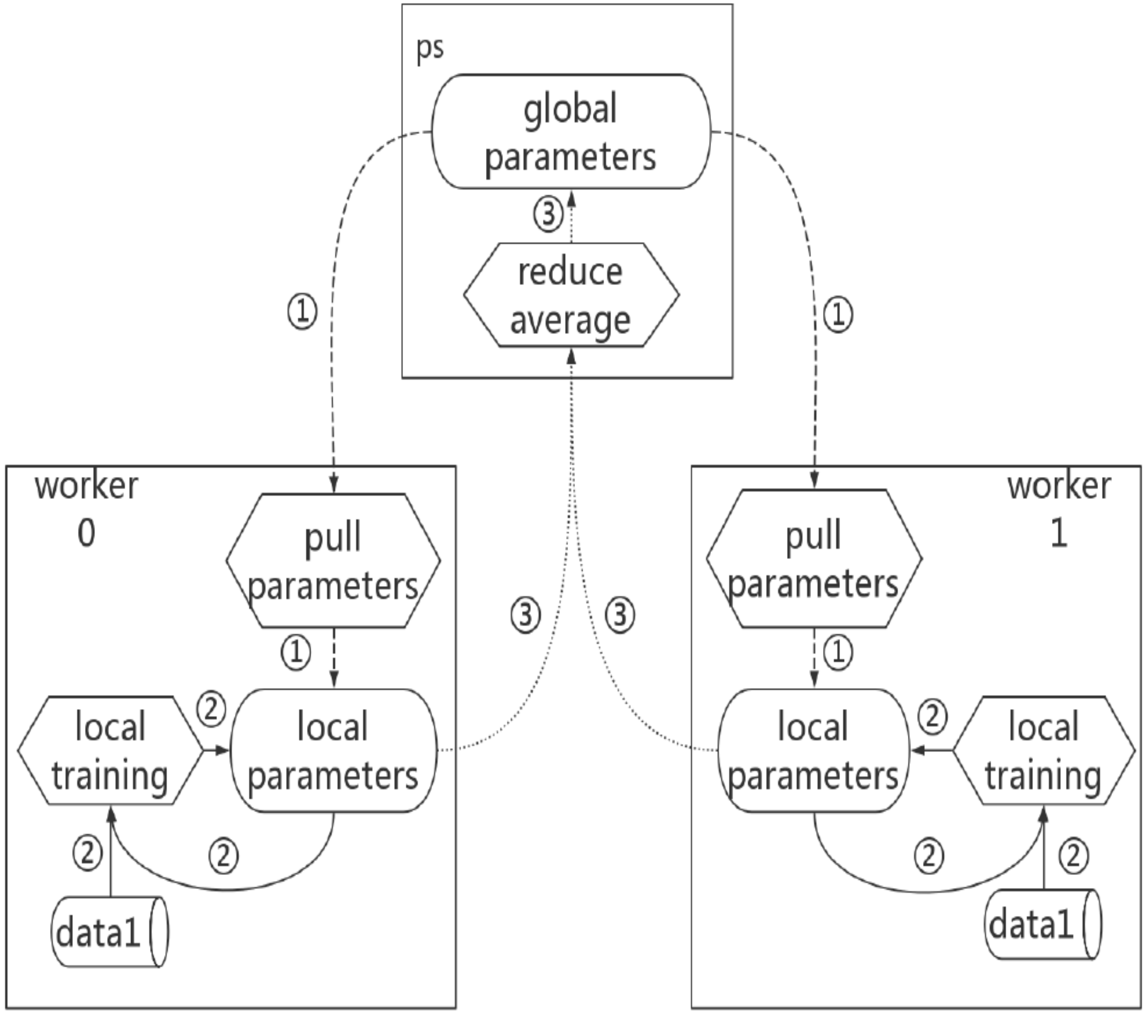
Experiences of Deep Learning Optimization on PAI

- Trends of Employ Deep Learning
- Requests for Distributed Deep Learning
  - Computation Speed
  - Memory Usage
- Memory Optimization
  - “ Swap in/out Strategy” for intermediate results
    - Utilize host memory as a bigger memory pool
  - “drop and re-computation” design for Seq2Seq Model
  - Increase training batch size 16x on deep NMT models

Alibaba Group 阿里巴巴集团

Experiences of Deep Learning Optimization on PAI

- Distributed Training
  - Simple auto placement & partitioning
  - Model Average Optimizer
    - Works well with linear learning rate rule
    - Speed up to 6.4X on 8 cards
  - Auto Parallel Optimizer
    - Extends user's graph from single mode to distributed mode
- Future work
  - Inference, Compilation, Placement, Auto tuning


ps
global parameters
③
reduce average
①
①
worker 0
pull parameters
①
local training
②
local parameters
②
②
data1
③
③
worker 1
pull parameters
①
local parameters
②
local training
②
②
data1