CSP 07-12

- 07-distributed computing & RDMA
  - 07-distributed training & RDMA
    - 分布式训练
      - Data Parallelism
        - 为什么要分布式训练？——Sync data parallel
          - 数据、模型、硬件扩展，ML 分布式训练的三个方向
        - 如何做？分布式数据并行——Sync data parallel
          - 分布式 SGD（parallel gradient）
            - 参数同步，收集梯度，更新模型
            - 同步 SGD vs Async SGD 的区别
            - 同步 SGD 的 straggler 问题
            - 如果不等，可以用 backup worker 来解决
            - 通信量大，可以用 Sufficient Factor Broadcasting 来减少通信量
            - 通信量大，可以用 Gradient Compression 来减少通信量
            - 通信量大，可以用 Parameter Server 来减少通信量
            - 通信量大，可以用 Ring-All-Reduce 来减少通信量
            - 通信量大，可以用 Hierarchical Ring-All-Reduce 来减少通信量
            - 通信量大，可以用 2D-Torus-All-Reduce 来减少通信量
            - 通信量大，可以用 Gradient Sparsification 来减少通信量
            - 通信量大，可以用 Deep Gradient Compression 来减少通信量
            - 通信量大，可以用 Gradient Quantization 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction + Local Gradient Clipping 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction + Local Gradient Clipping + Warm-up Training 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction + Local Gradient Clipping + Warm-up Training + Momentum Factor Masking 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction + Local Gradient Clipping + Warm-up Training + Momentum Factor Masking + Layer-wise Adaptive Rate Scaling 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction + Local Gradient Clipping + Warm-up Training + Momentum Factor Masking + Layer-wise Adaptive Rate Scaling + Decoupled Weight Decay 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction + Local Gradient Clipping + Warm-up Training + Momentum Factor Masking + Layer-wise Adaptive Rate Scaling + Decoupled Weight Decay + Momentum Correction 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction + Local Gradient Clipping + Warm-up Training + Momentum Factor Masking + Layer-wise Adaptive Rate Scaling + Decoupled Weight Decay + Momentum Correction + Local Gradient Clipping 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction + Local Gradient Clipping + Warm-up Training + Momentum Factor Masking + Layer-wise Adaptive Rate Scaling + Decoupled Weight Decay + Momentum Correction + Local Gradient Clipping + Warm-up Training 来减少通信量
            - 通信量大，可以用 Gradient Sparsification + Gradient Quantization + Error Feedback + Momentum Correction + Local Gradient Clipping + Warm-up Training + Momentum Factor Masking + Layer-wise Adaptive Rate Scaling + Decoupled Weight Decay