Decoupled Torch Network-Aware Training on Interlinked Online Nodes

DeToNATION

Communication-efficient training on - and between HPC infrastructures

Mogens Henrik From, Jacob Nielsen and Gianluca Bermina
Supervised by Peter Schneider-Kamp and Lukas Galke

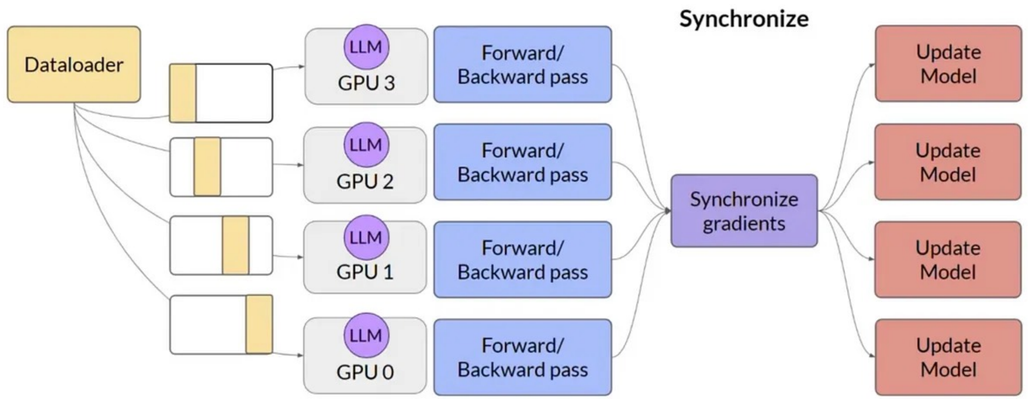
SDU

ordbogen a/s

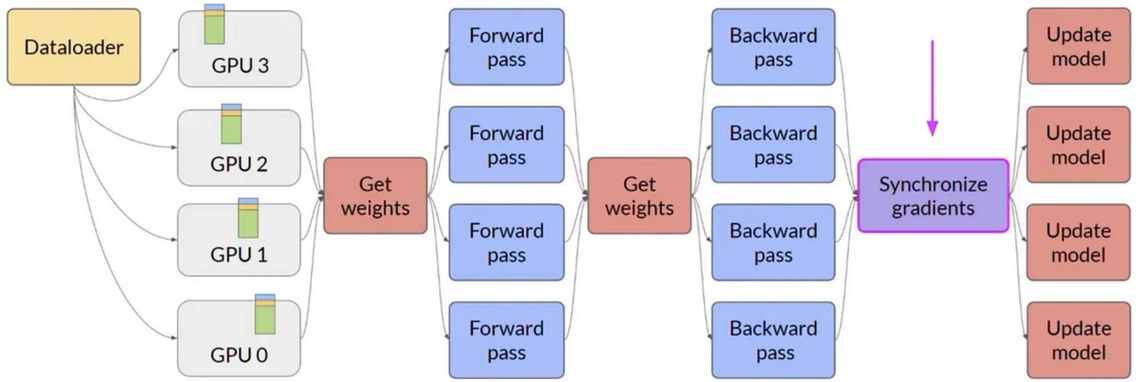
Training Neural Networks

Communication takes time

Distributed Data Parallel (DDP)



Fully Sharded Data Parallel (FSDP)



The Problem

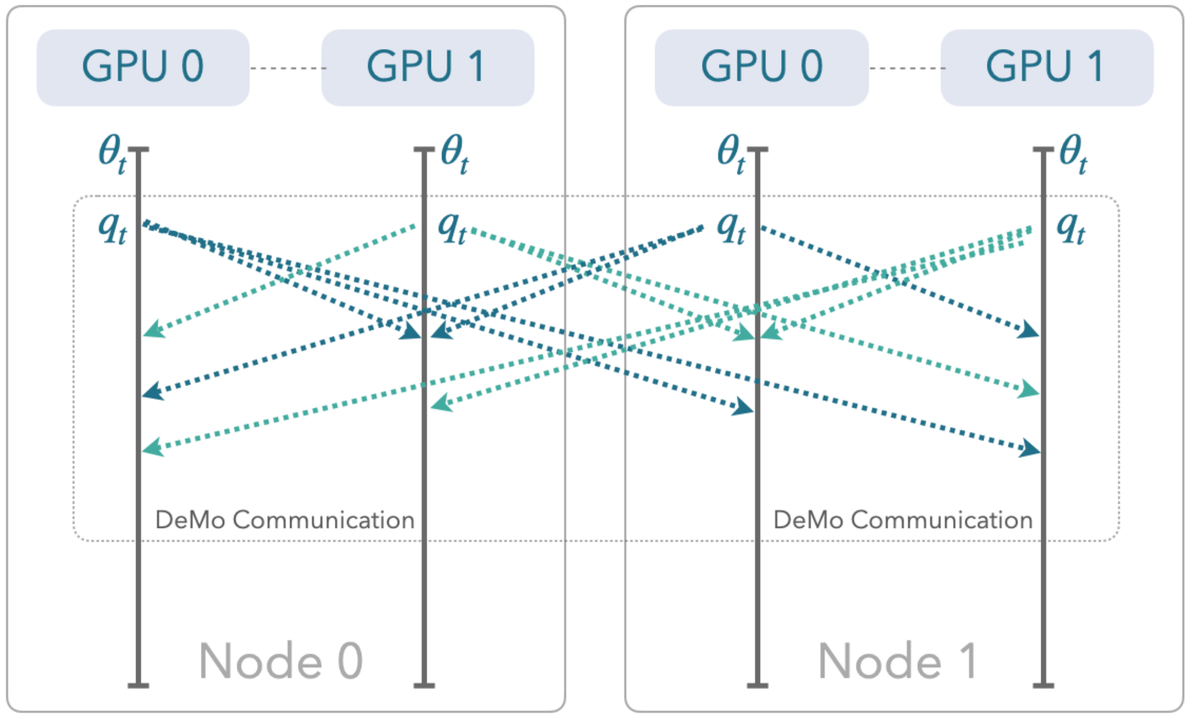
Communication takes time

- Communicating between distributed training processes is expensive
- Distributed training generally scales bad on HPCs
  - Using significant amount of time communicating *instead* of computing!
- Bottlenecks:
  - Interconnect speed
  - Network congestion both internally and externally.

A Solution - Decoupled Momentum · DeMo

Communication takes time

- Decoupled Momentum[1]
- Only exchange fast moving components in the gradients
- Only supports DDP
  - Does not scale to large models



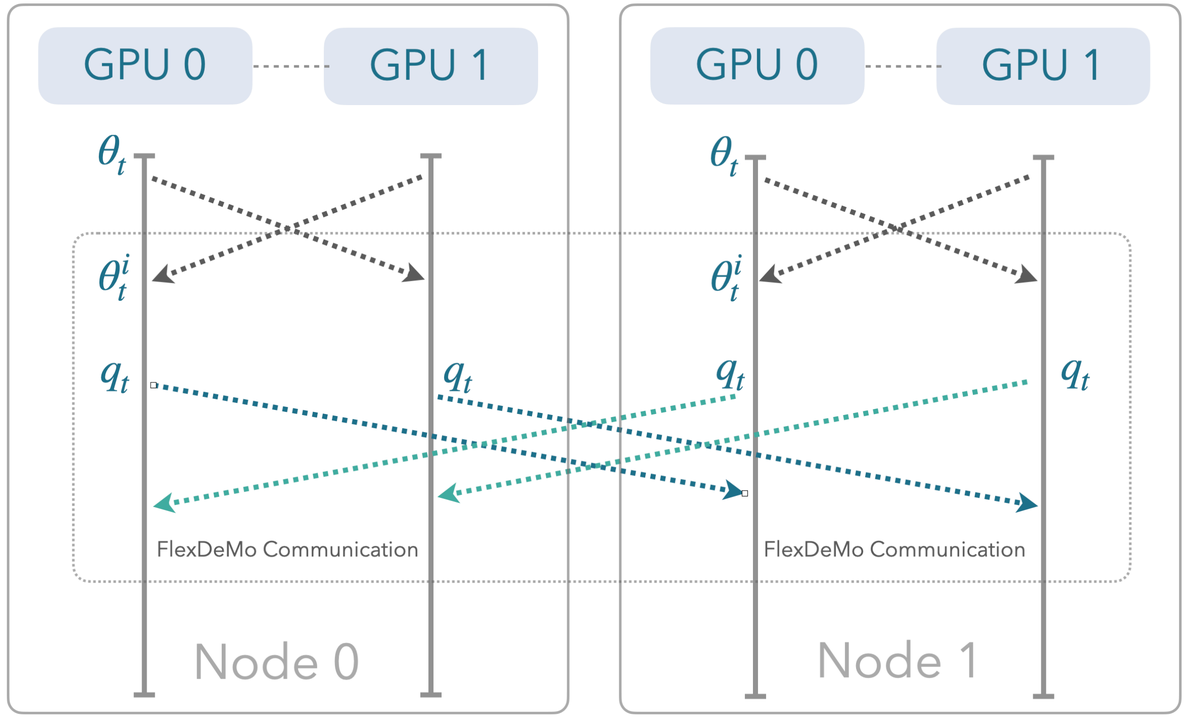
Distributed Data Parallel

1: DeMo: Decoupled Momentum Optimization, Bowen Peng, Jeffrey Quesnelle, Diederik P. Kingma, https://arxiv.org/abs/2411.19870

Our Solution

Extending from DDP to FSDP - and beyond

- Introducing FSDP into the DeMo-Scheme
  - FlexDeMo
- Introducing different optimisers
- Introducing new parameter replicator strategies

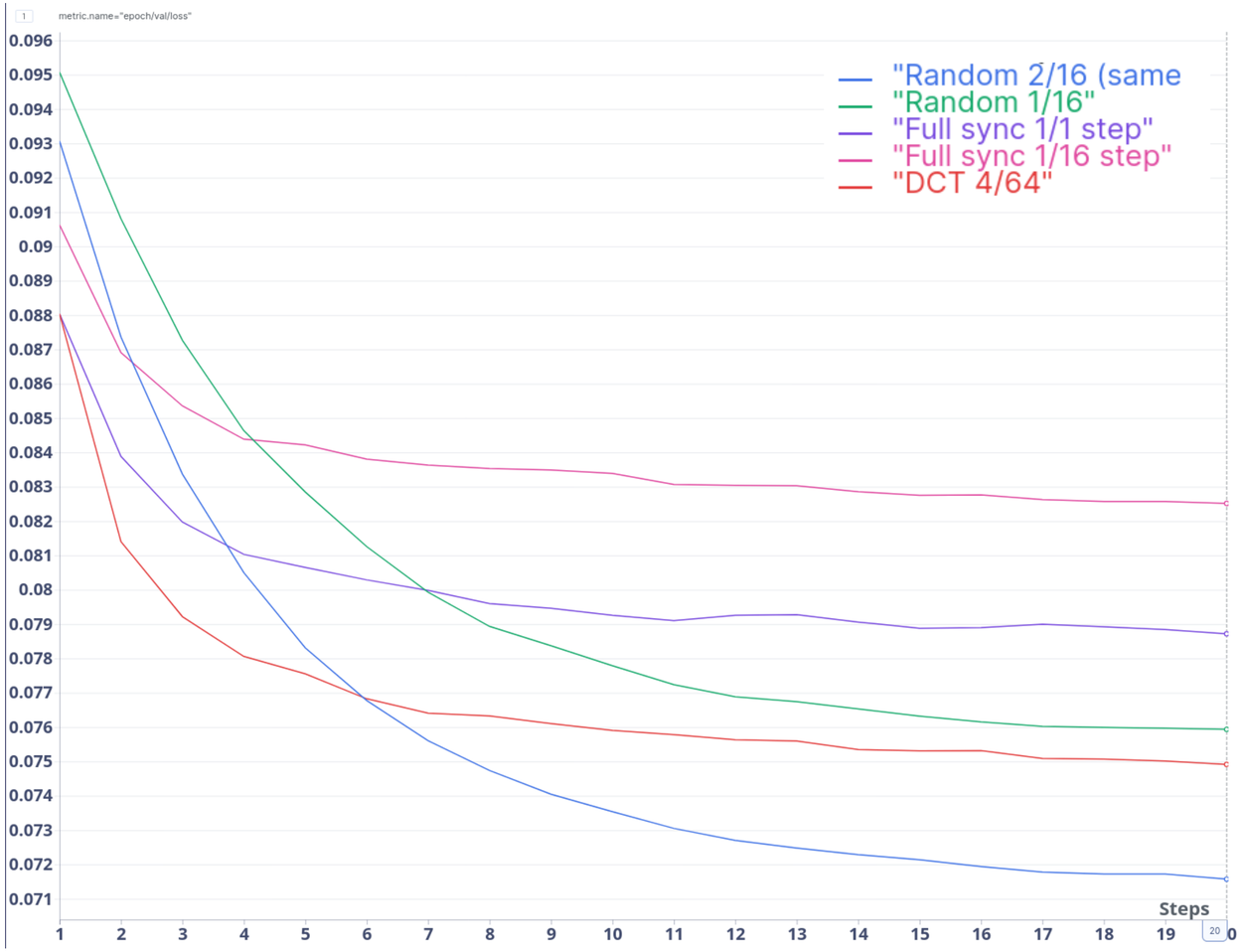


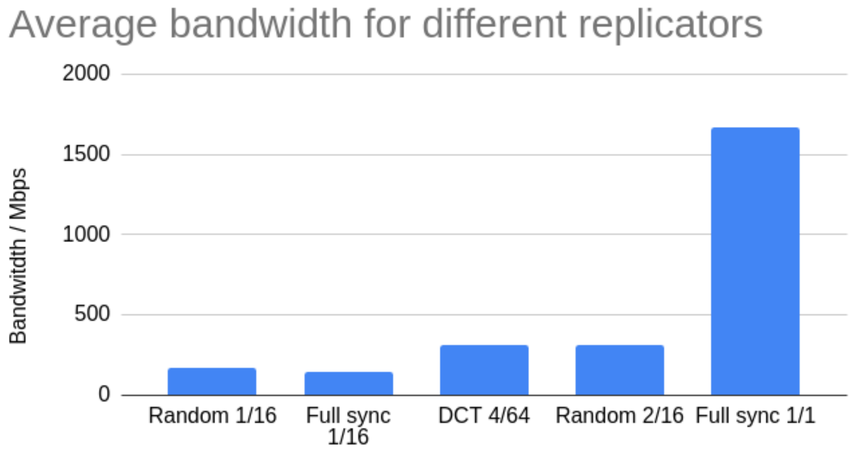
Fully Sharded Data Parallel

Preliminary Result

Communication takes time

- Introducing FSDP into the DeMo-Scheme
- FlexDeMo
- Replicators
  - DeMo
  - Random





What we are working on

Communication takes time

- Scale up experiments
  - Number of Nodes
  - Model Sizes
  - Problems and domains.
- Investigating behaviour of decoupled optimisers (SGD, AdamW)
- Different methods for selecting which data to synchronise across training-processes.
  - fast moving components are not necessarily optimal.

What we are working on

Communication takes time

- Benchmarked on NVIDIA platforms
  - Small local computer clusters
  - SDU UCloud
- Tested on LUMI (for AMD support)
  - A large scale run on LUMI remains
- Code available at: https://github.com/schneiderkamplab/DeToNATION

Hackathon goals

Our plan for the week

- Detailed benchmarks
  - Performance of Random vs. DeMo replicator (with a few accelerators)
- Scaling to many (128?) accelerators
  - How does performance of the two methods scale?
  - How is the network impacted and/or bottlenecking the training?
  - Identifying bottlenecks in the implementation / improving performance