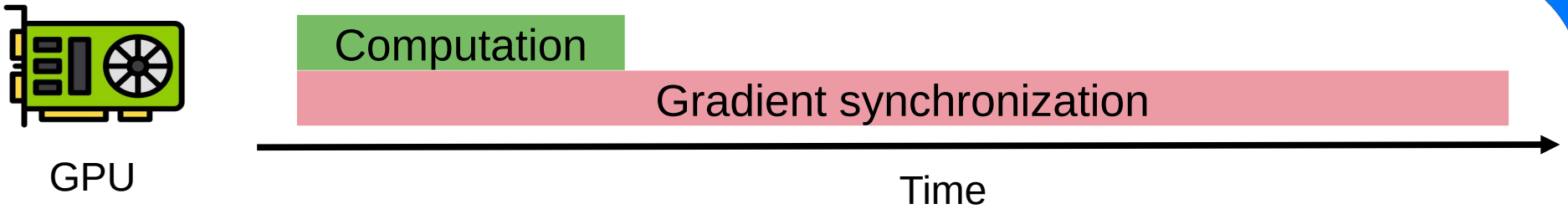
TAGC: Optimizing Gradient Communication in Distributed Transformer Training

Igor Polyakov
VK, ITMO University

Alexey Dukhanov
ITMO University

Egor Spirin
VK Lab

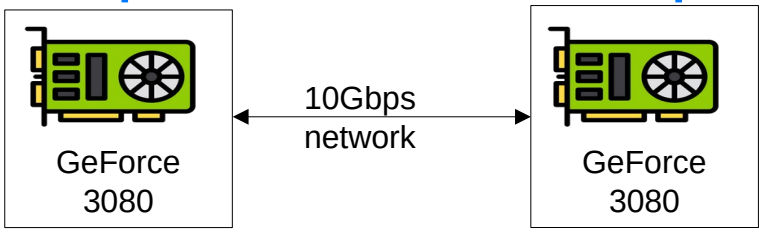


TAGC is an algorithm for distributed transformer communication:

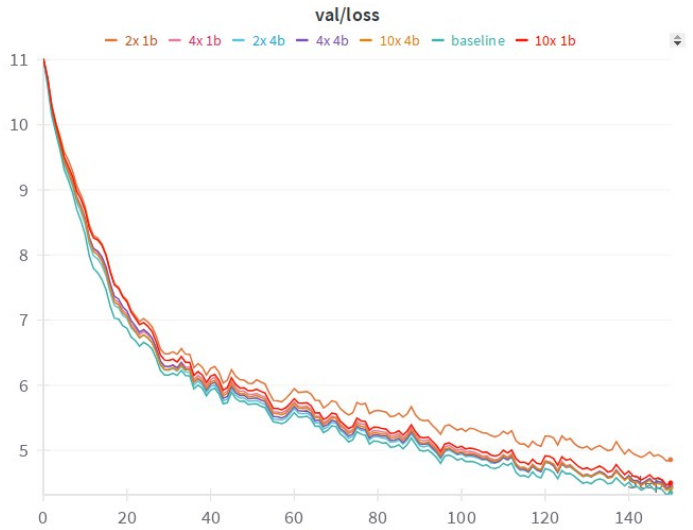
- Sparsification - we selectivley sparsify transformer gradients for large layers
- Compression - we compress those gradients for optmized communication
- Synchronization - we integrate into the popular FSDP framework

TAGC: Optimizing Gradient Communication in Distributed Transformer Training

Experimental setup



- GPT-2 with 162 million parameters
- OpenWebText dataset



TAGC speeds up training up to **15%** with loss degradation of only **3.6%**.