

Communication Efficiency

Gradient Compression

Full
Gradients
100 MB

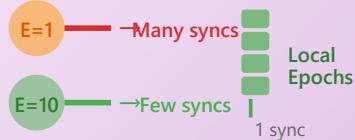


Compressed
1-10 MB

Sparsification, quantization

💾 10-100x reduction

Local Updates



More local epochs

💾 Linear in E reduction

Model Compression



Full Model

Prune
Distill



Compressed

Pruning, distillation

💾 2-10x reduction