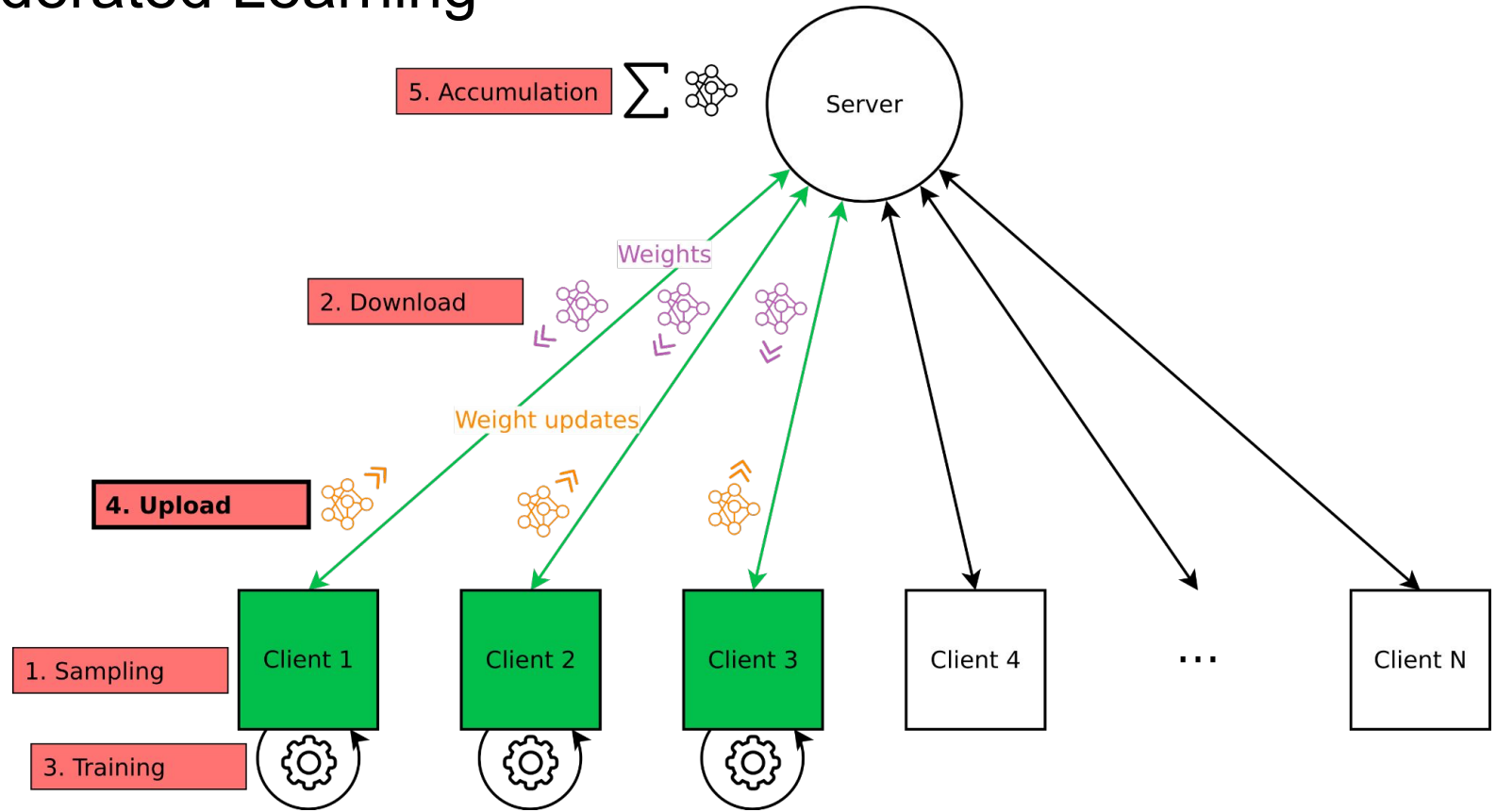
DAdaQuant

Doubly-adaptive quantization for communication-efficient Federated Learning

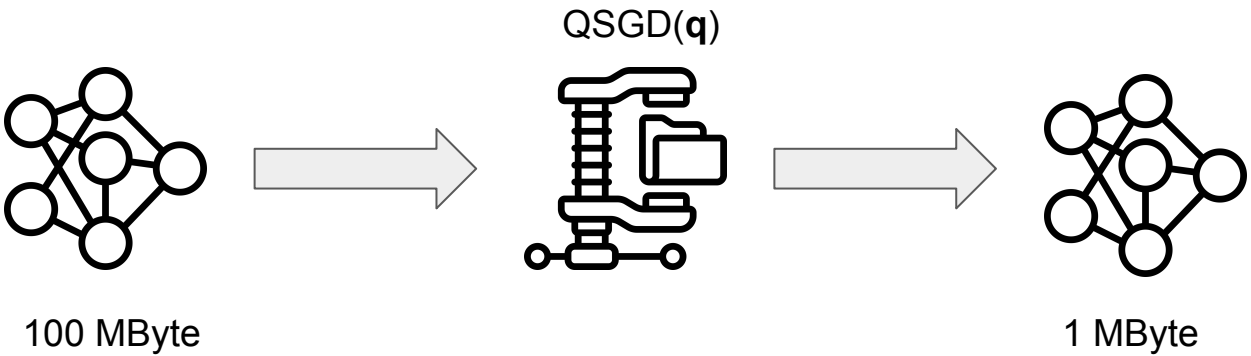
Robert Hönig (rhoenig@ethz.ch)
ETH Zurich

Yiren Zhao
Robert Mullins
University of Cambridge

Federated Learning

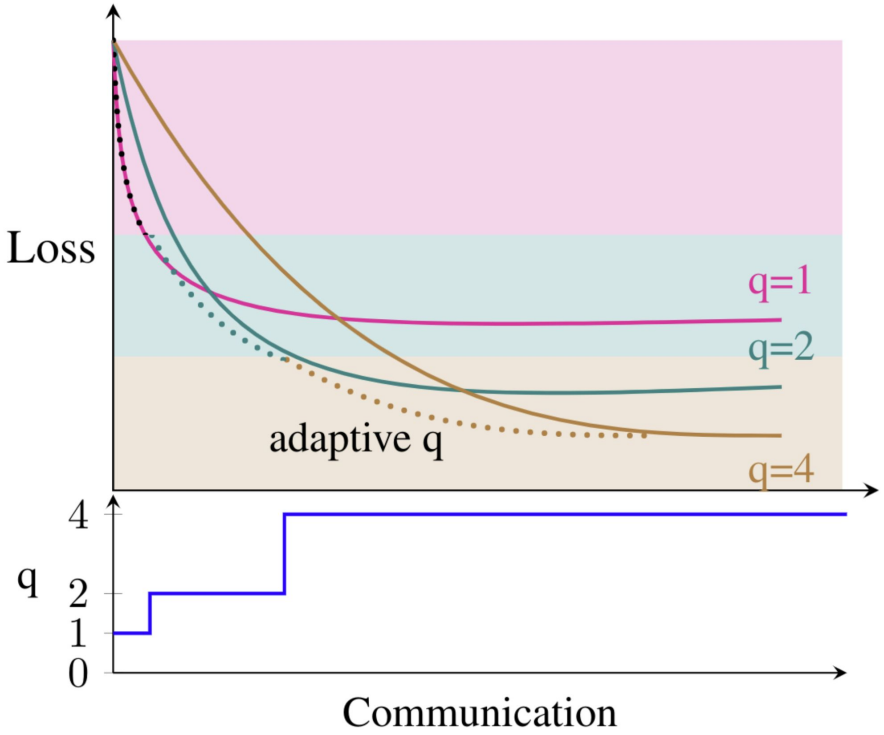


Weight update compression

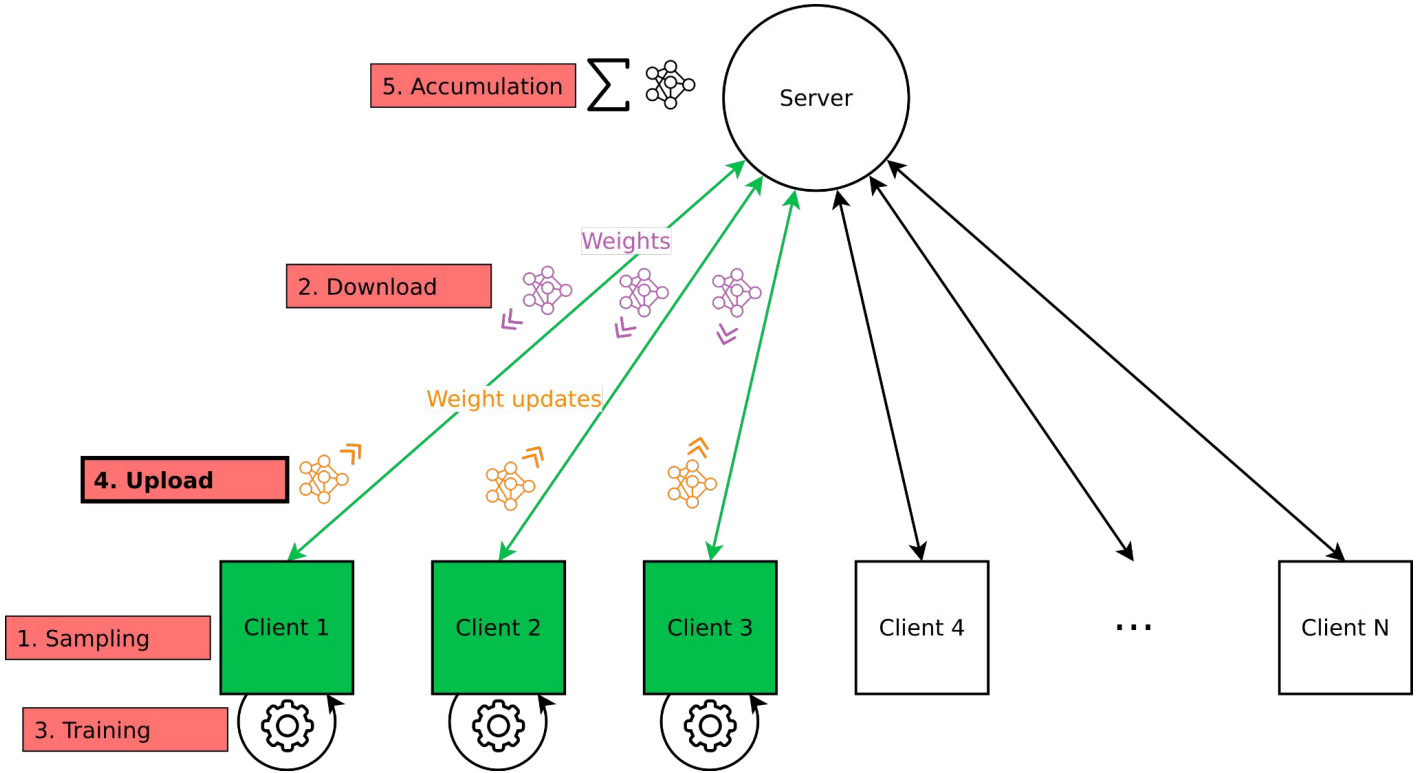


How to choose **q**?

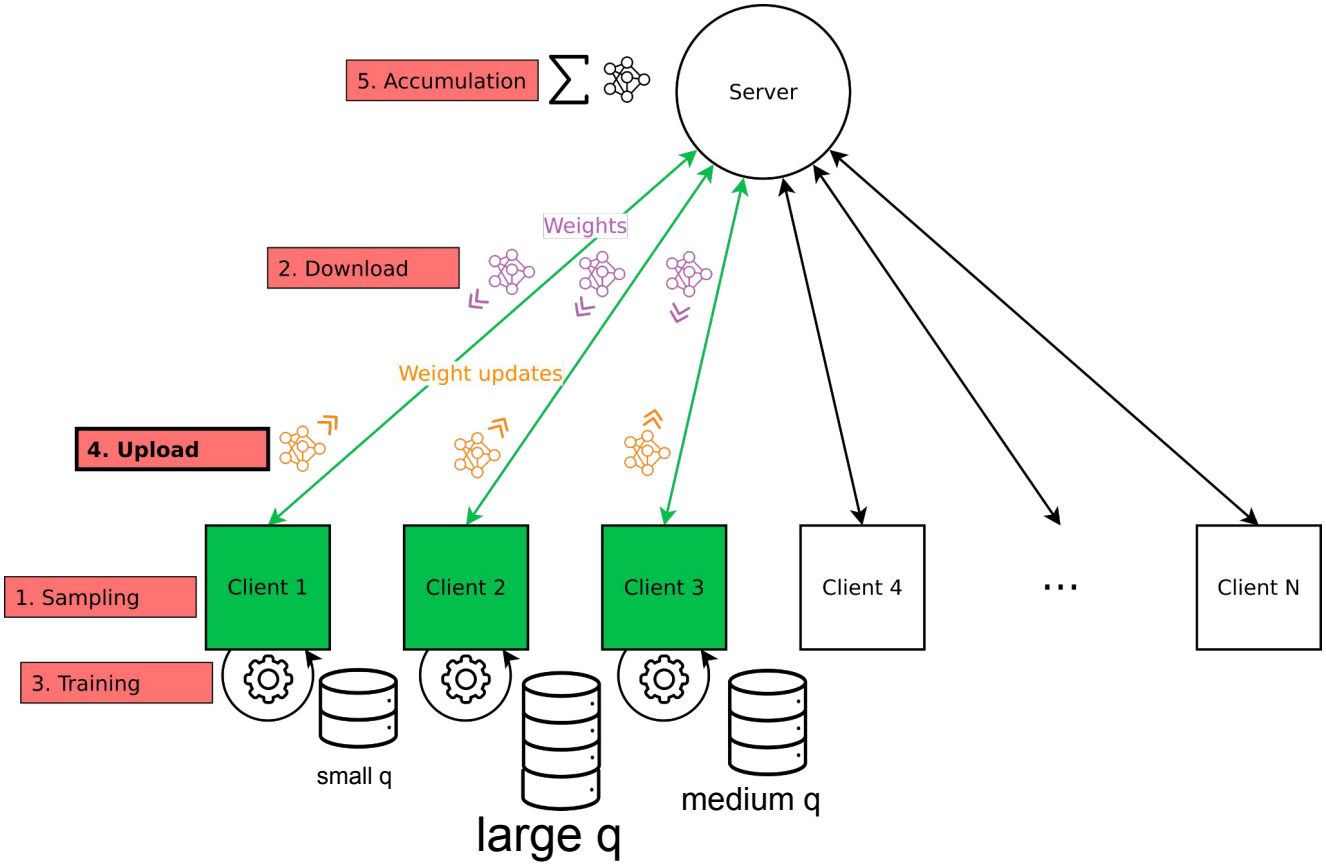
Time-adaptive quantization



Client-adaptive quantization **(new)**



Client-adaptive quantization **(new)**



DAdaQuant: Do both!

Results (selected)

Model + Dataset	**Uncompressed**	**QSGD**	**DAdaQuant**
CNN + CelebA	12.6 GB	19.4 MB	16.3 MB (1.19x)
LSTM + Shakespeare	267 MB	28.1 MB	12.7 MB (2.21x)