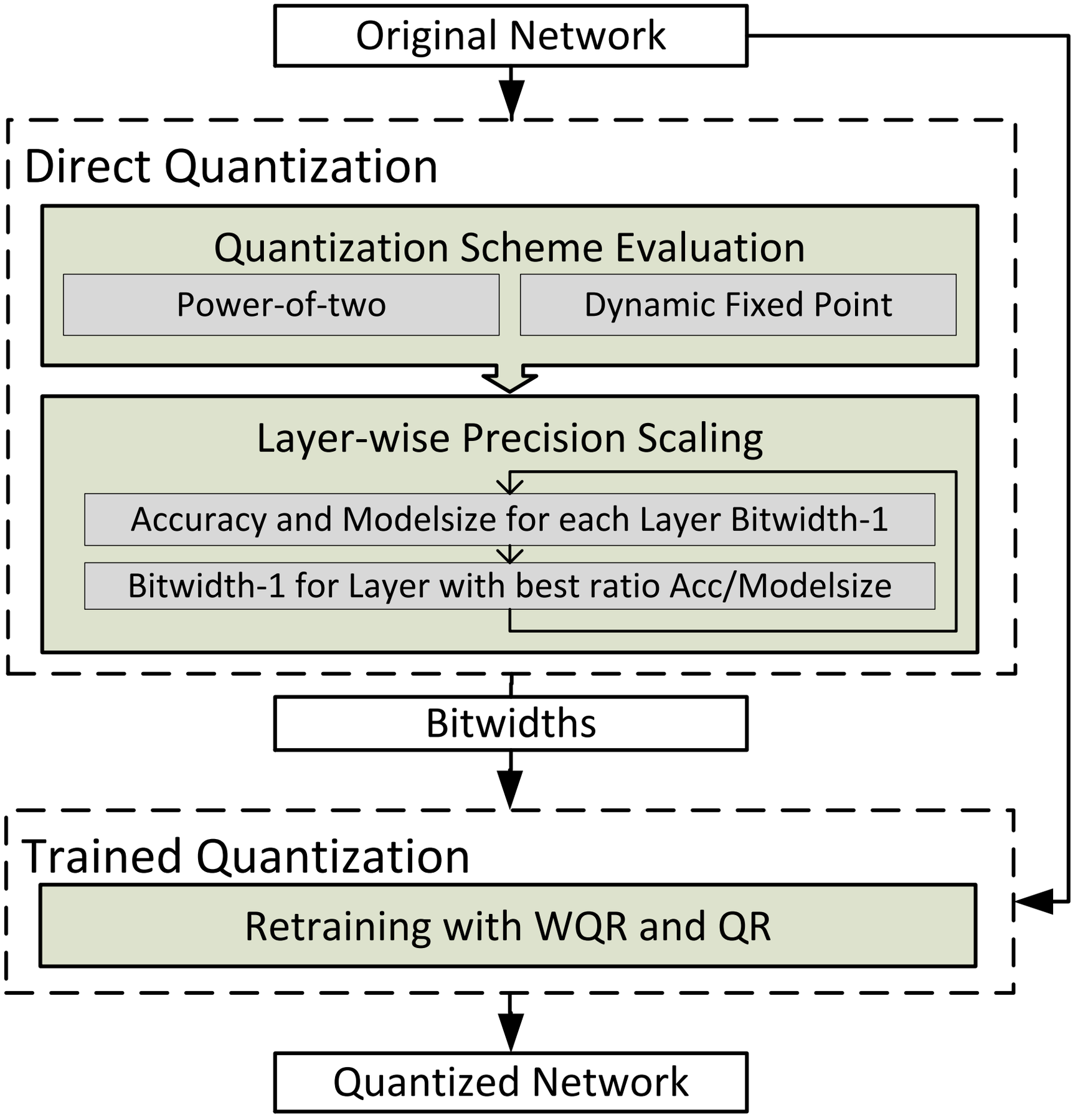
Original Network
Direct Quantization
Quantization Scheme Evaluation
Power-of-two
Dynamic Fixed Point
Layer-wise Precision Scaling
Accuracy and Modelsize for each Layer Bitwidth-1
Bitwidth-1 for Layer with best ratio Acc/Modelsize
Bitwidths
Trained Quantization
Retraining with WQR and QR
Quantized Network