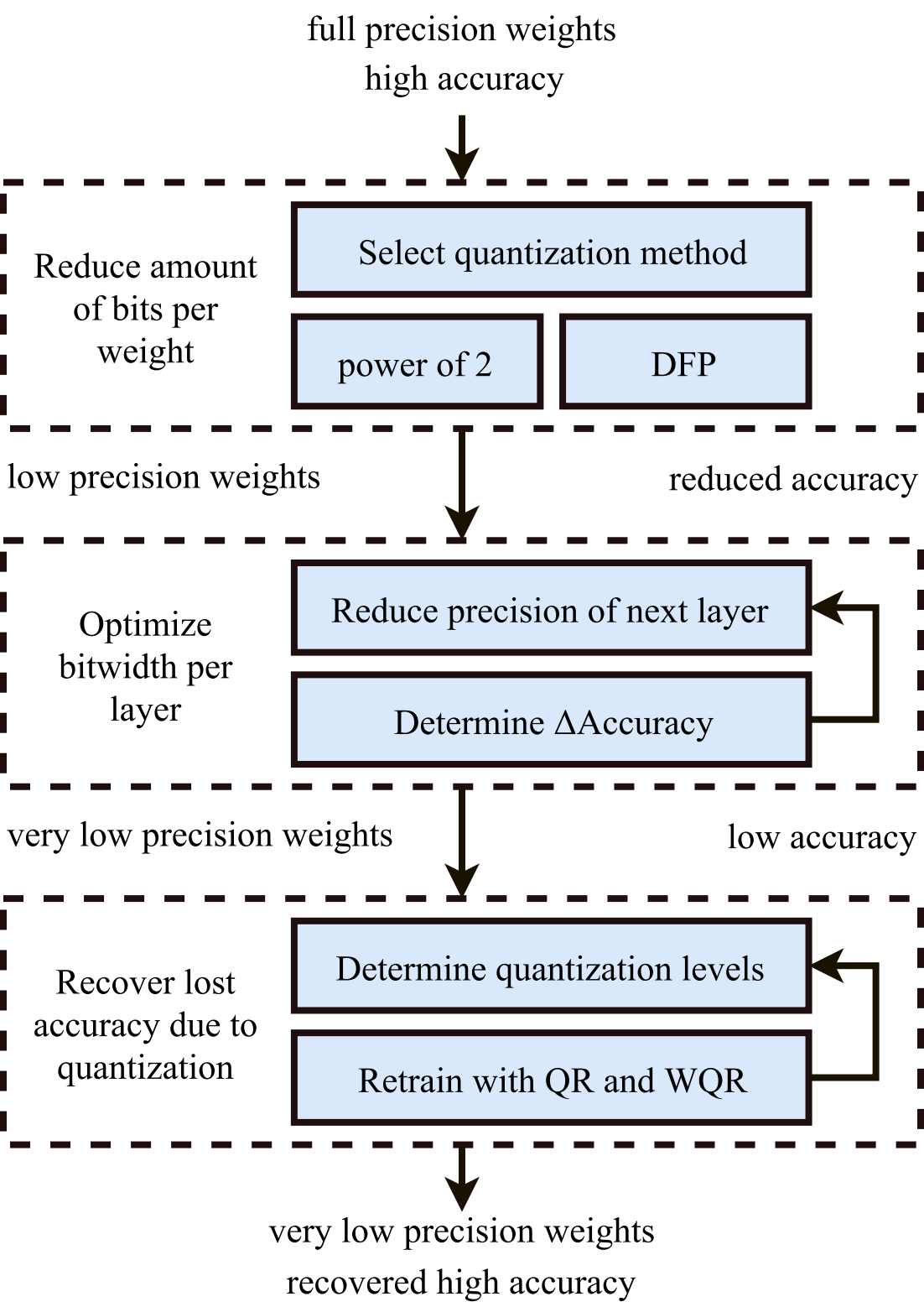
full precision weights
high accuracy
Reduce amount of bits per weight
Select quantization method
power of 2
DFP
low precision weights
reduced accuracy
Optimize bitwidth per layer
Reduce precision of next layer
Determine ΔAccuracy
very low precision weights
low accuracy
Recover lost accuracy due to quantization
Determine quantization levels
Retrain with QR and WQR
very low precision weights
recovered high accuracy