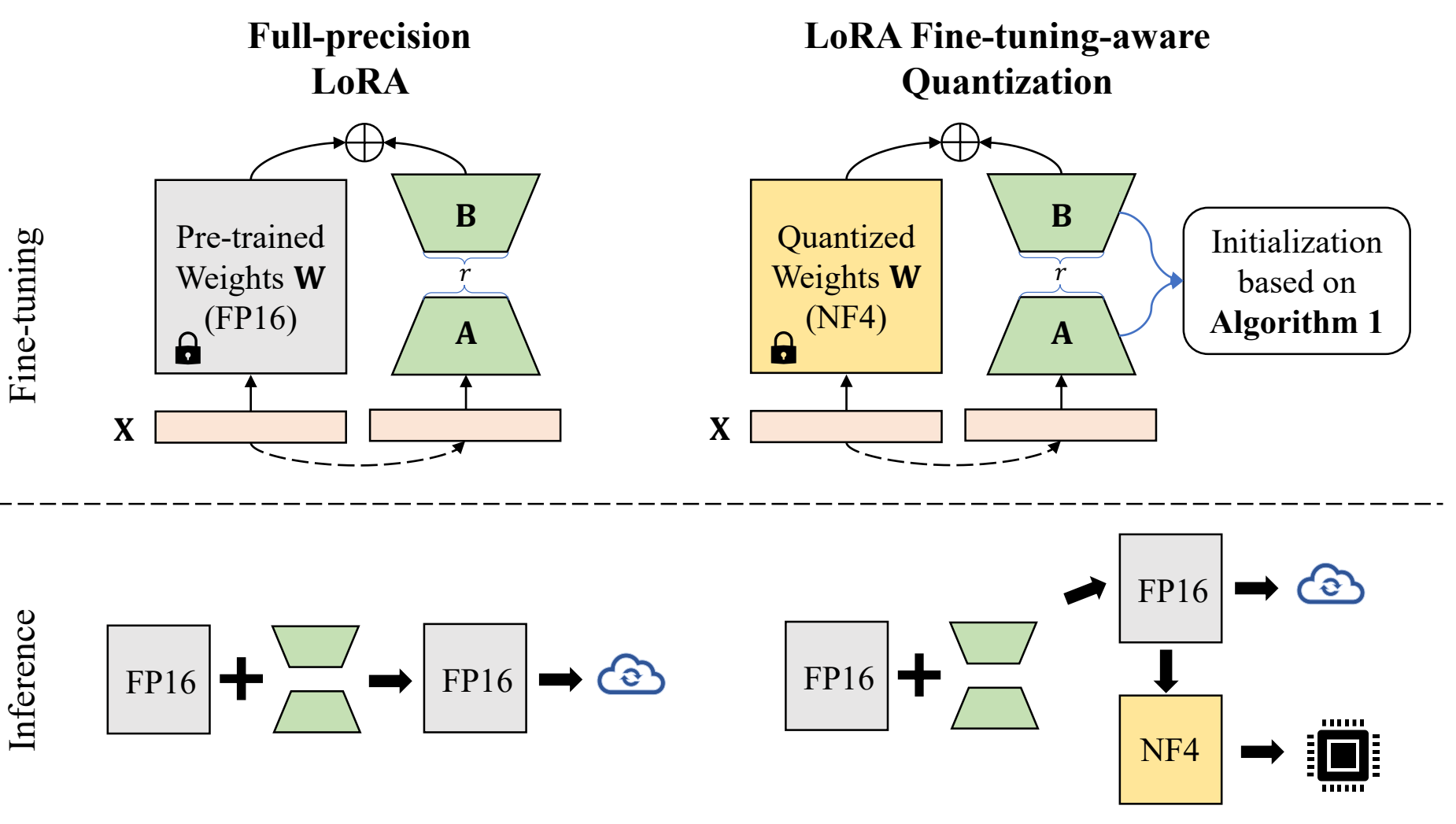
Full-precision LoRA
LoRA Fine-tuning-aware Quantization
Fine-tuning
Pre-trained Weights W (FP16)
B
r
A
X
Quantized Weights W (NF4)
Initialization based on Algorithm 1
Inference
FP16
FP16
FP16
FP16
NF4