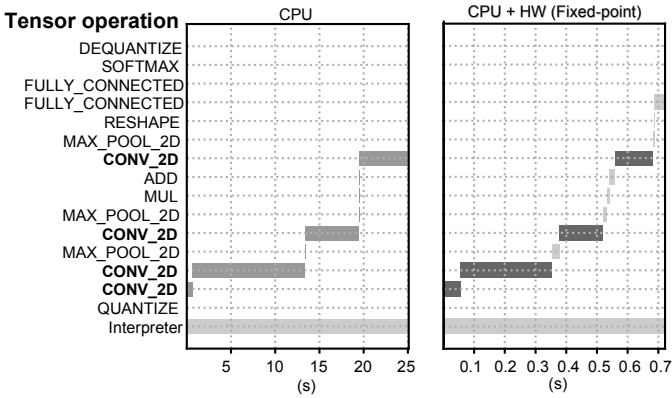
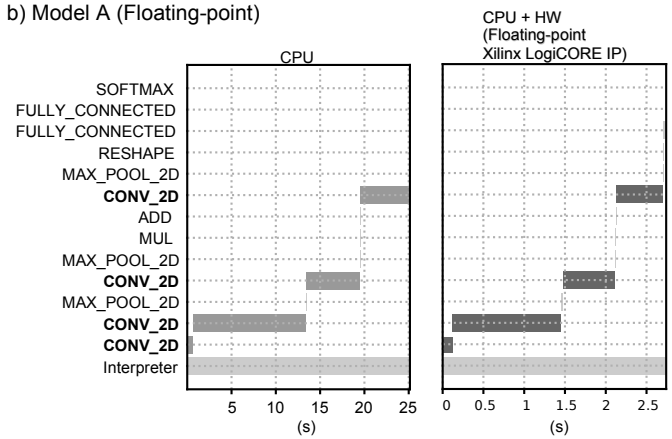
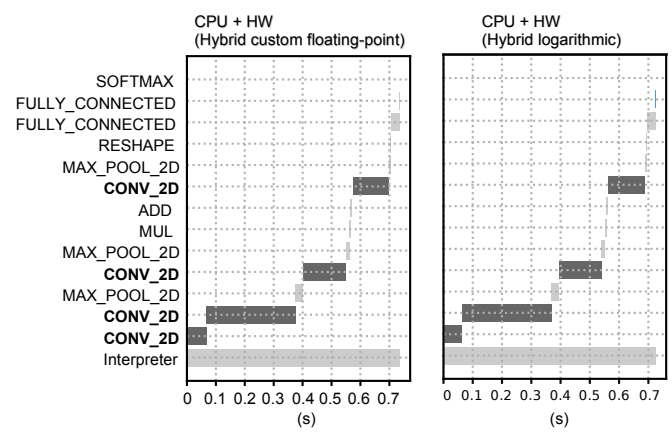
a) Model A (8-bit fixed-point quantization)



b) Model A (Floating-point)



b) Model A (Floating-point)



b) Model B (Floating-point)