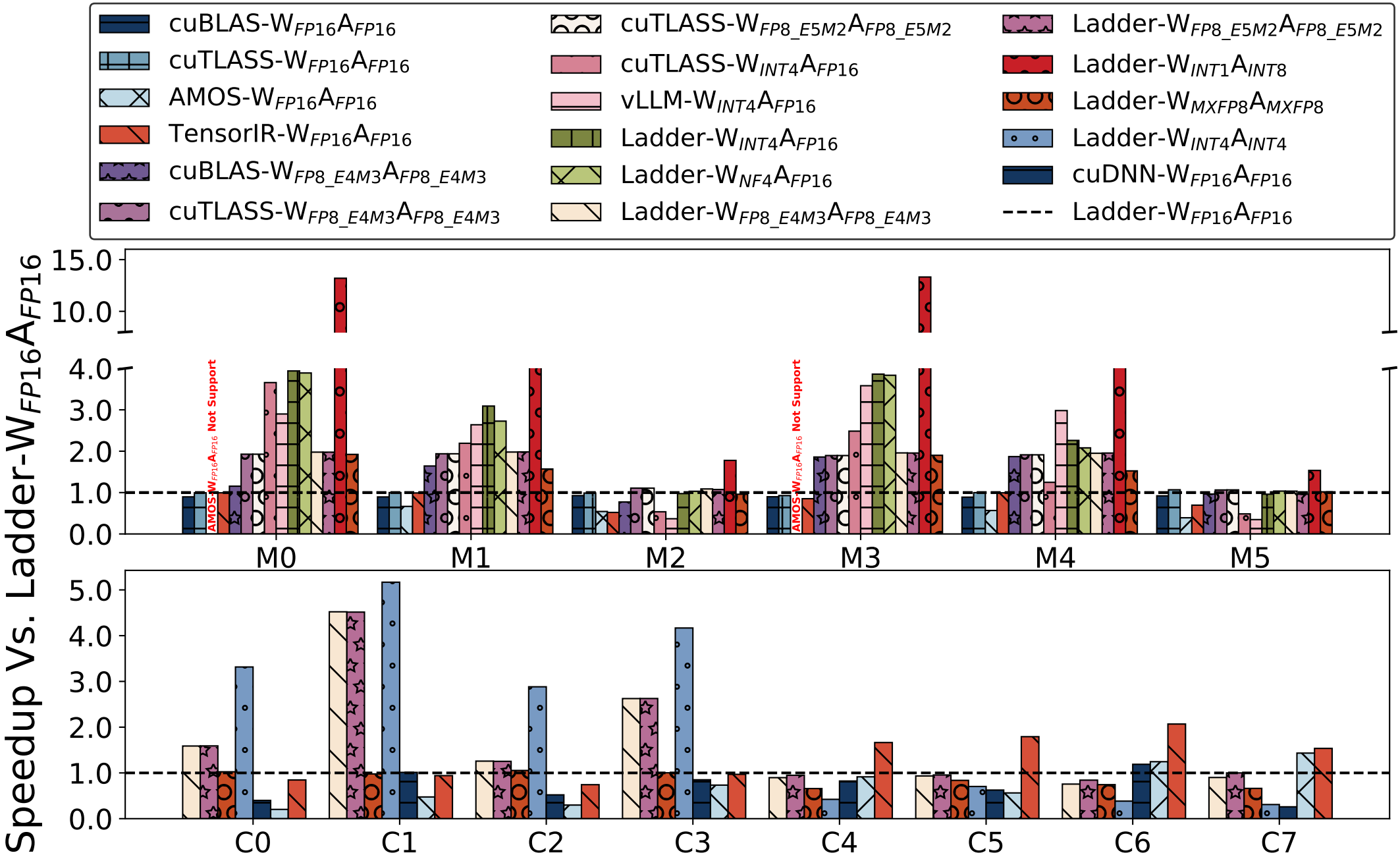
cuBLAS-W$_{FP16}$A$_{FP16}$
cuTLASS-W$_{FP16}$A$_{FP16}$
AMOS-W$_{FP16}$A$_{FP16}$
TensorIR-W$_{FP16}$A$_{FP16}$
cuBLAS-W$_{FP8_E4M3}$A$_{FP8_E4M3}$
cuTLASS-W$_{FP8_E4M3}$A$_{FP8_E4M3}$
cuTLASS-W$_{FP8_E5M2}$A$_{FP8_E5M2}$
cuTLASS-W$_{INT4}$A$_{FP16}$
vLLM-W$_{INT4}$A$_{FP16}$
Ladder-W$_{INT4}$A$_{FP16}$
Ladder-W$_{NF4}$A$_{FP16}$
Ladder-W$_{FP8_E4M3}$A$_{FP8_E4M3}$
Ladder-W$_{FP8_E5M2}$A$_{FP8_E5M2}$
Ladder-W$_{INT1}$A$_{INT8}$
Ladder-W$_{MXFP8}$A$_{MXFP8}$
Ladder-W$_{INT4}$A$_{INT4}$
cuDNN-W$_{FP16}$A$_{FP16}$
Ladder-W$_{FP16}$A$_{FP16}$
Speedup Vs. Ladder-W$_{FP16}$A$_{FP16}$
AMOS-W$_{FP16}$A$_{FP16}$ Not Support
M0
M1
M2
M3
M4
M5
C0
C1
C2
C3
C4
C5
C6
C7