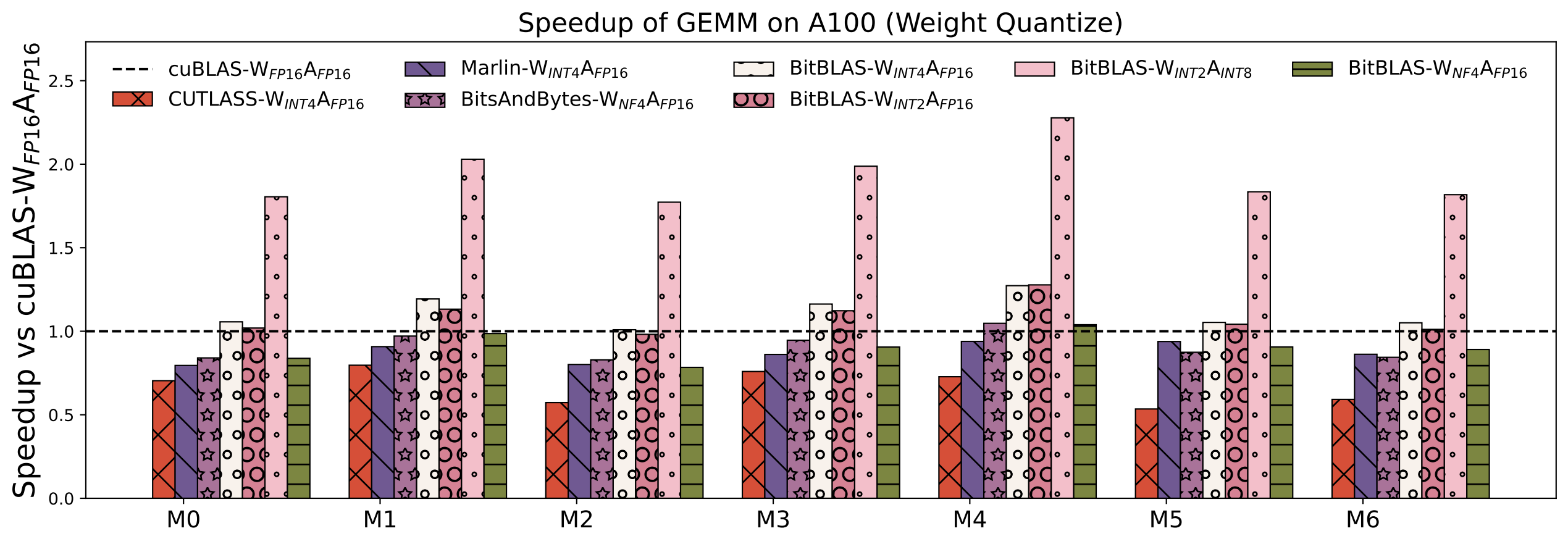
Speedup of GEMM on A100 (Weight Quantize)
Speedup vs cuBLAS-W$_{FP16}$A$_{FP16}$
cuBLAS-W$_{FP16}$A$_{FP16}$
CUTLASS-W$_{INT4}$A$_{FP16}$
Marlin-W$_{INT4}$A$_{FP16}$
BitsAndBytes-W$_{NF4}$A$_{FP16}$
BitBLAS-W$_{INT4}$A$_{FP16}$
BitBLAS-W$_{INT2}$A$_{FP16}$
BitBLAS-W$_{INT2}$A$_{INT8}$
BitBLAS-W$_{NF4}$A$_{FP16}$
0.0
0.5
1.0
1.5
2.0
2.5
M0
M1
M2
M3
M4
M5
M6