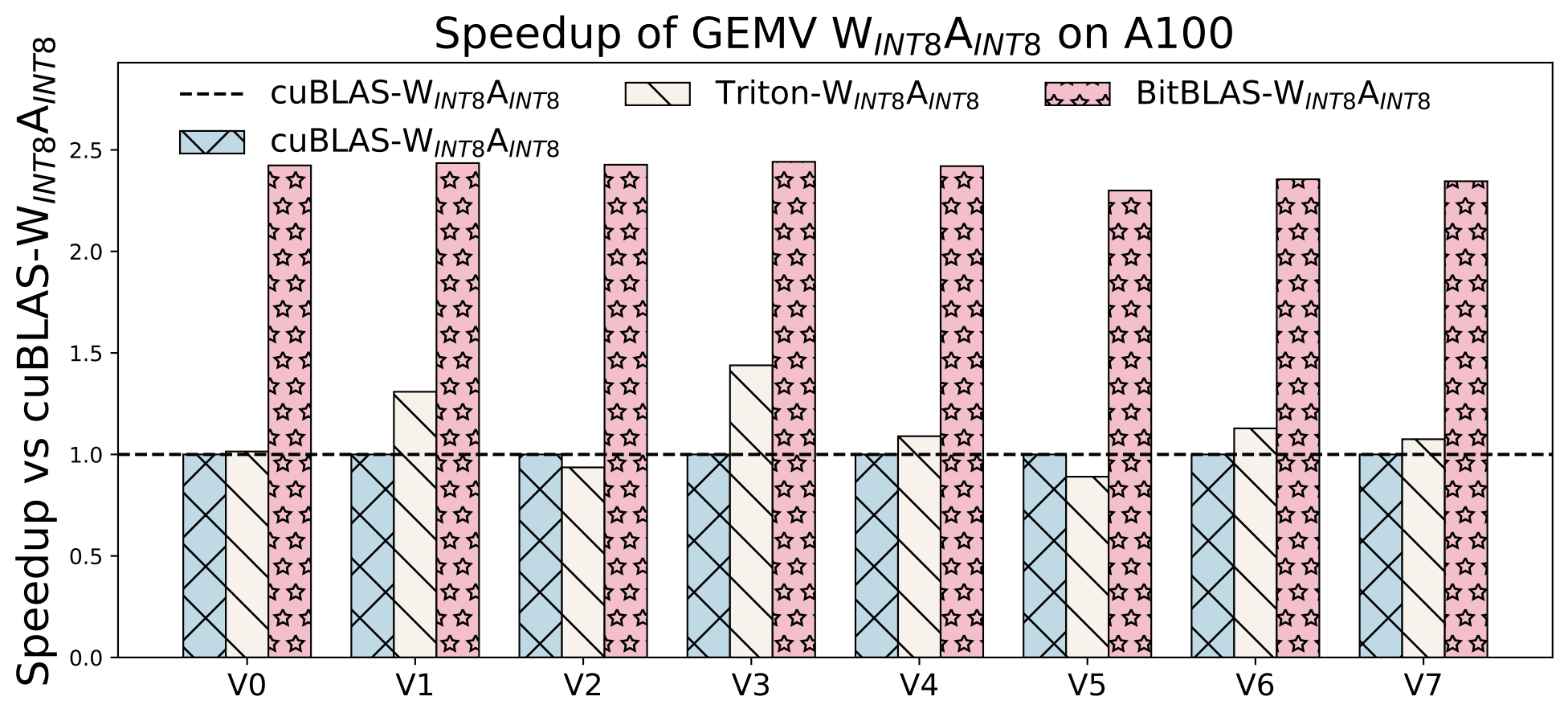
Speedup of GEMV $W_{INT8}A_{INT8}$ on A100
cuBLAS-$W_{INT8}A_{INT8}$
Triton-$W_{INT8}A_{INT8}$
BitBLAS-$W_{INT8}A_{INT8}$
cuBLAS-$W_{INT8}A_{INT8}$
Speedup vs cuBLAS-$W_{INT8}A_{INT8}$
0.0
0.5
1.0
1.5
2.0
2.5
V0
V1
V2
V3
V4
V5
V6
V7