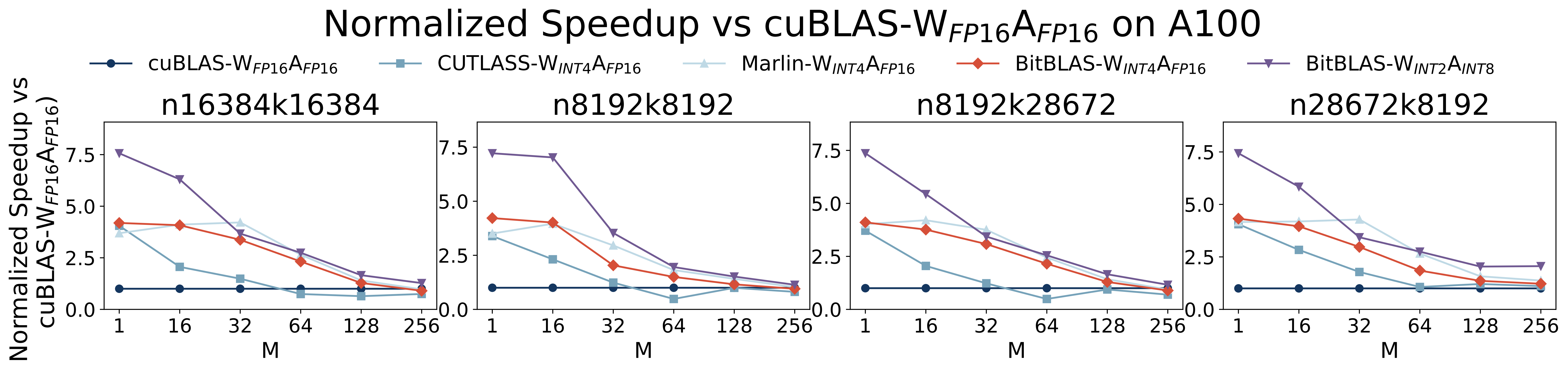
Normalized Speedup vs cuBLAS-$W_{FP16}A_{FP16}$ on A100
cuBLAS-$W_{FP16}A_{FP16}$
CUTLASS-$W_{INT4}A_{FP16}$
Marlin-$W_{INT4}A_{FP16}$
BitBLAS-$W_{INT4}A_{FP16}$
BitBLAS-$W_{INT2}A_{INT8}$
Normalized Speedup vs cuBLAS-$W_{FP16}A_{FP16}$)
n16384k16384
n8192k8192
n8192k28672
n28672k8192
0.0
2.5
5.0
7.5
1
16
32
64
128
256
M