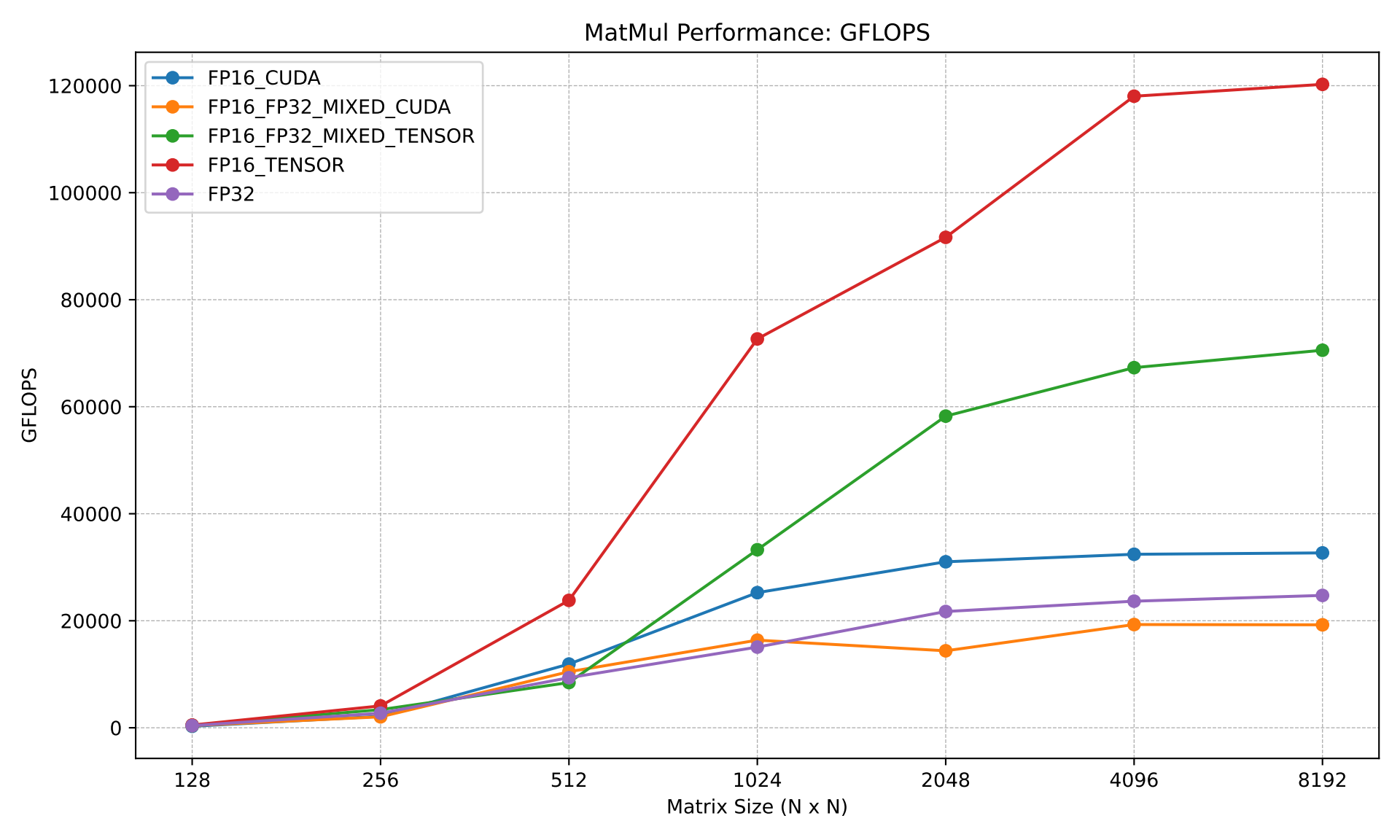
MatMul Performance: GFLOPS
FP16_CUDA
FP16_FP32_MIXED_CUDA
FP16_FP32_MIXED_TENSOR
FP16_TENSOR
FP32
GFLOPS
0
20000
40000
60000
80000
100000
120000
128
256
512
1024
2048
4096
8192
Matrix Size (N x N)