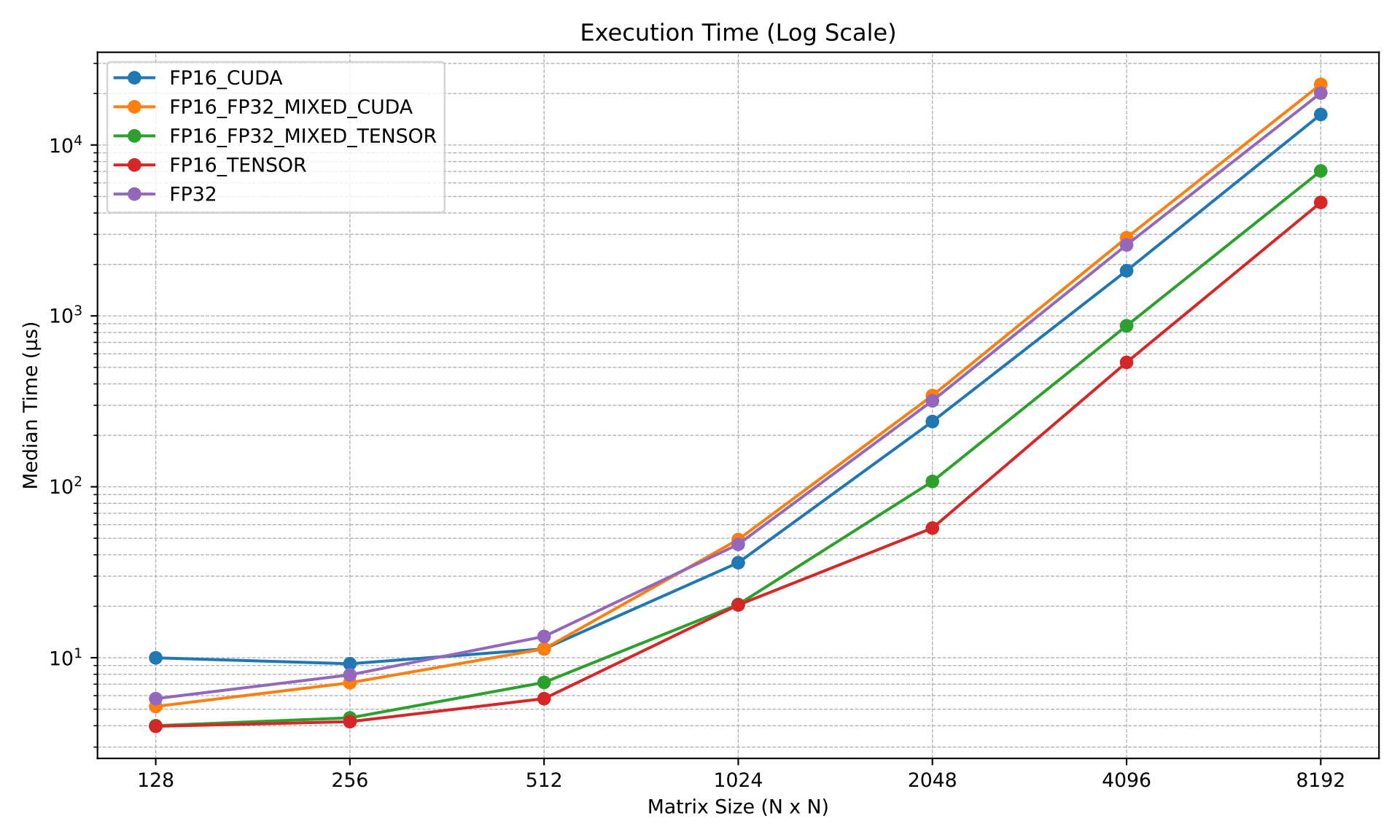
Execution Time (Log Scale)
FP16_CUDA
FP16_FP32_MIXED_CUDA
FP16_FP32_MIXED_TENSOR
FP16_TENSOR
FP32
Median Time (µs)
10^4
10^3
10^2
10^1
128
256
512
1024
2048
4096
8192
Matrix Size (N x N)