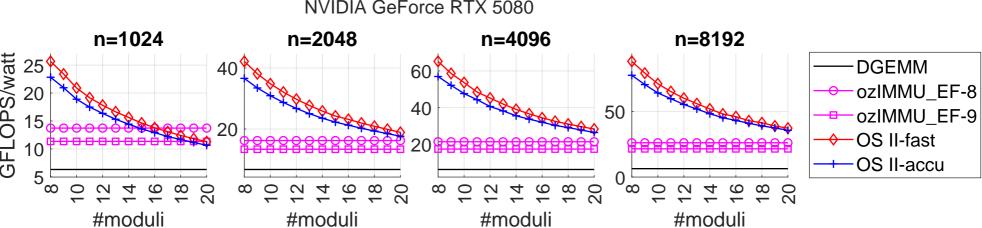
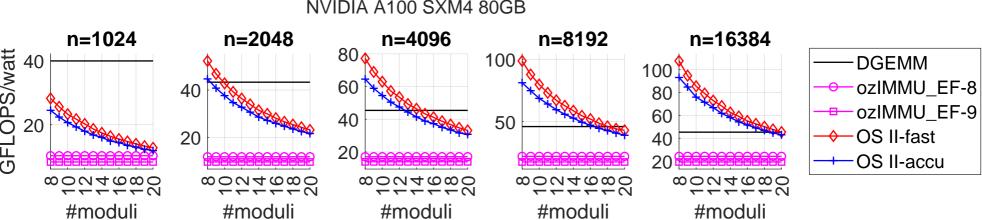
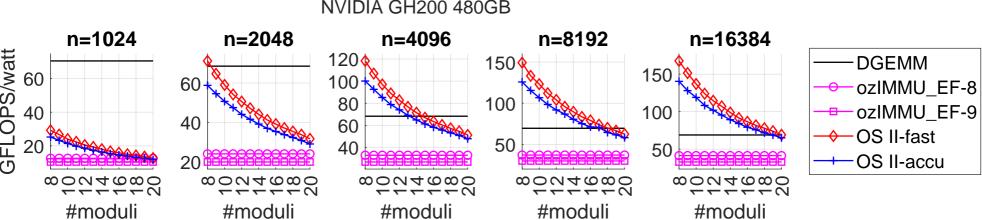


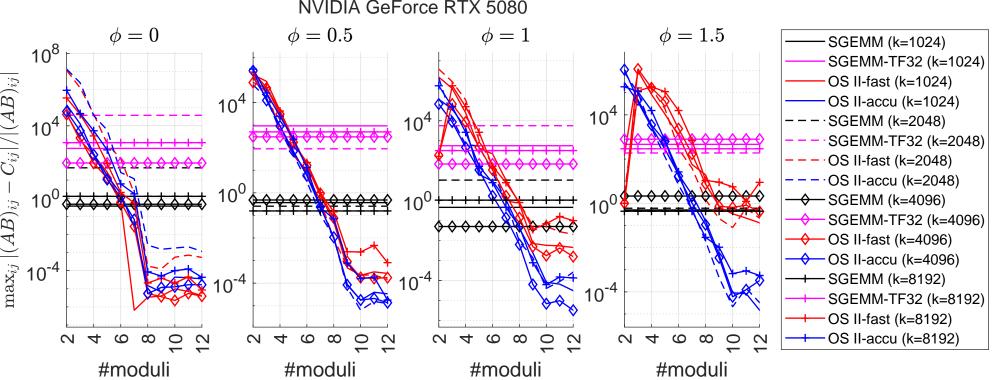
NVIDIA A100 SXM4 80GB fast n=2048 fast n=1024 fast n=4096 fast n=8192 fast n=16384 100 75 50 25 inverse_scaling conv 32i 2 8u #moduli #moduli #moduli #moduli #moduli cublasGemmEx accu n=1024 accu n=2048 accu n=4096 accu n=8192 accu n=16384 conv 64f 2 8i 100 50 #moduli #moduli #moduli #moduli #moduli

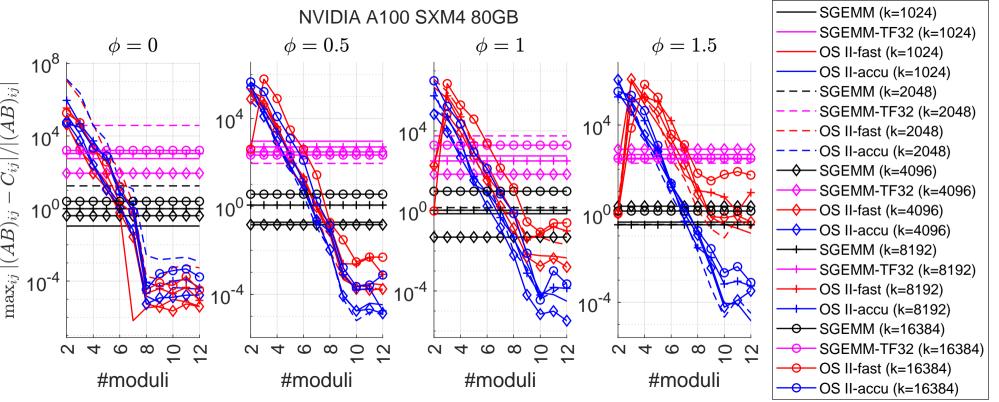
NVIDIA GH200 480GB fast n=2048 fast n=1024 fast n=4096 fast n=8192 fast n=16384 100 75 50 25 inverse_scaling conv 32i 2 8u #moduli #moduli #moduli #moduli #moduli cublasGemmEx accu n=1024 accu n=2048 accu n=4096 accu n=8192 accu n=16384 conv 64f 2 8i 100 50 25 #moduli #moduli #moduli #moduli #moduli

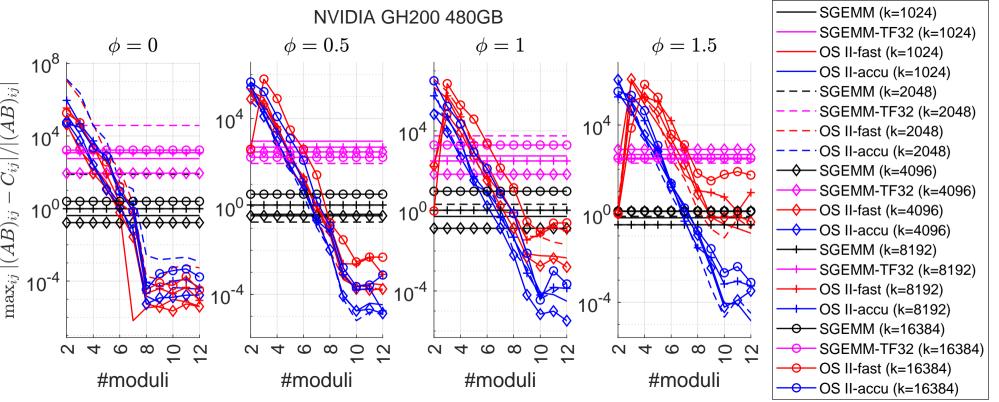


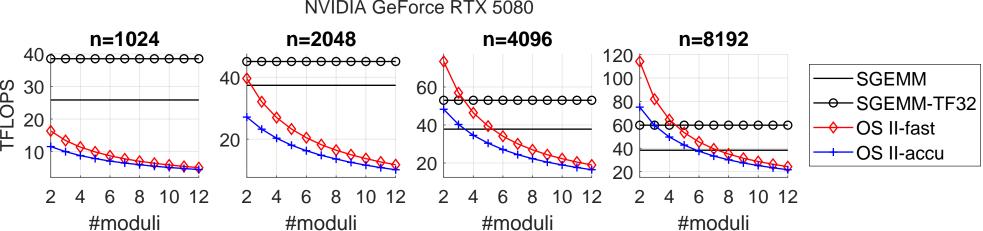


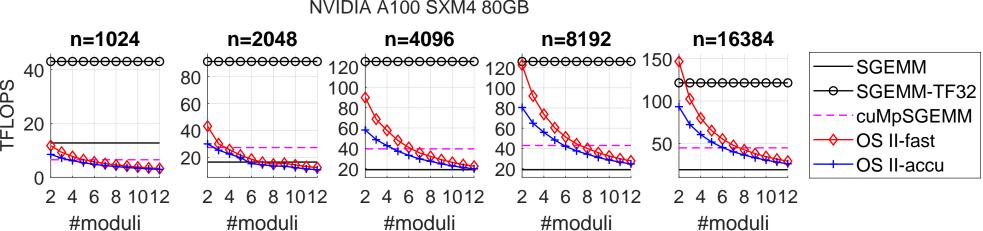


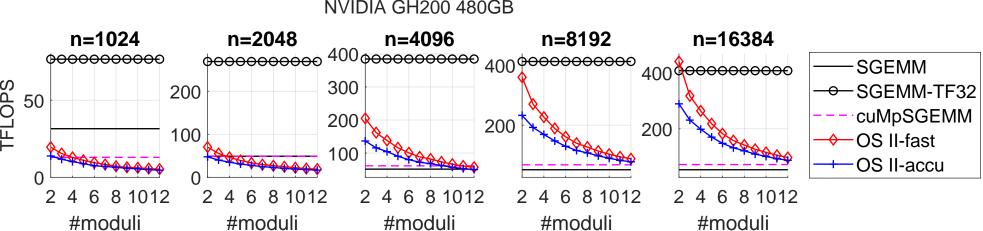


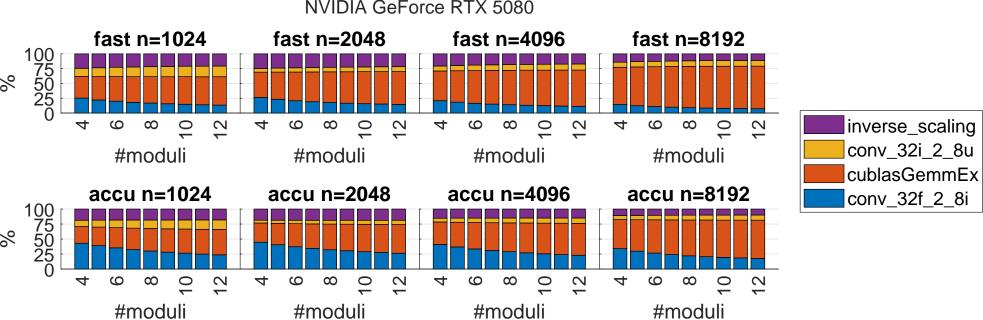


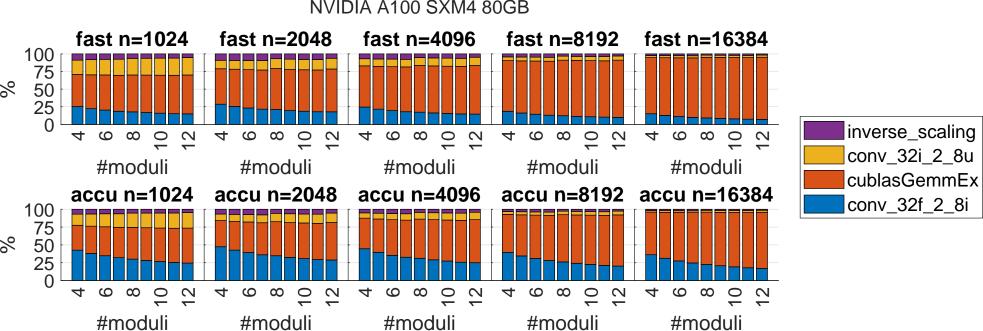


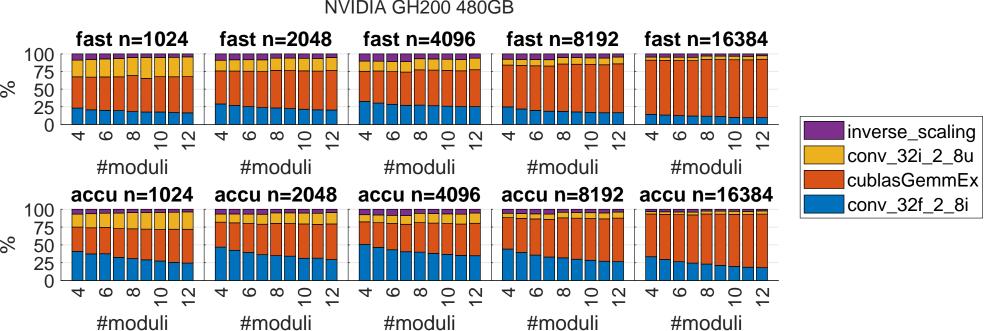


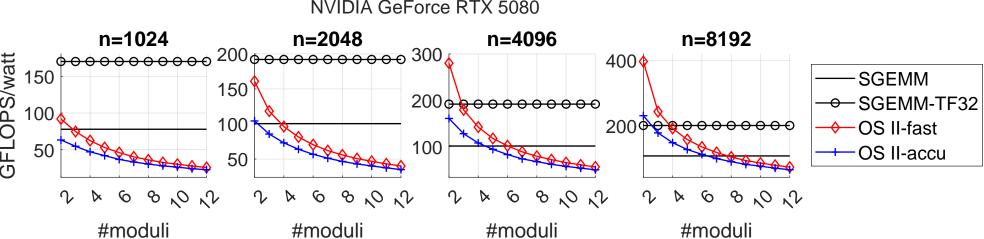


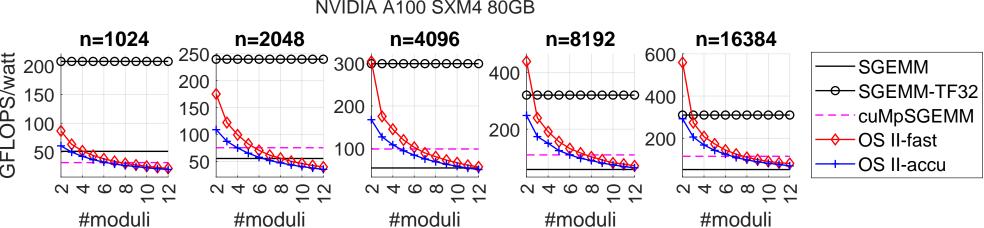












NVIDIA GH200 480GB n=1024 n=2048 n=4096n=8192 n=16384 000000000 Ф00000000 0000000000 800 **SGEMM** 600 400 0000000000 → SGEMM-TF32 200 600 400 000000000 cuMpSGEMM 200 400 100 200 OS II-fast 200 200 OS II-accu ∞ 0.49 000 **040000** 2400 0 **040000** 0 0 #moduli #moduli #moduli #moduli #moduli