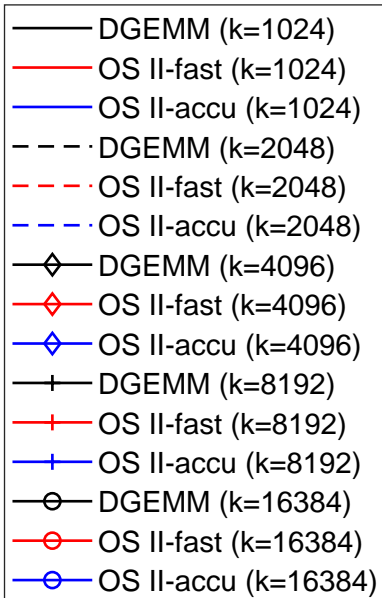
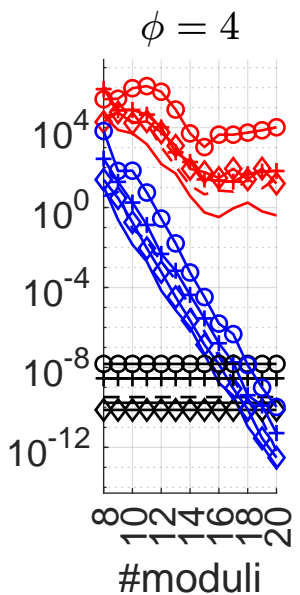
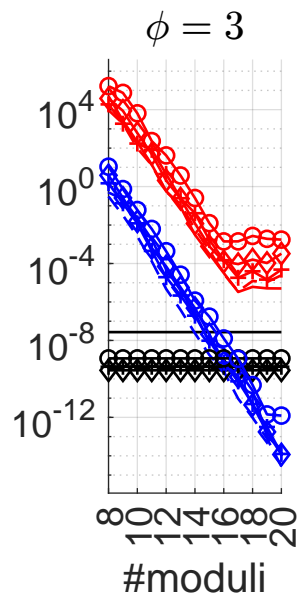
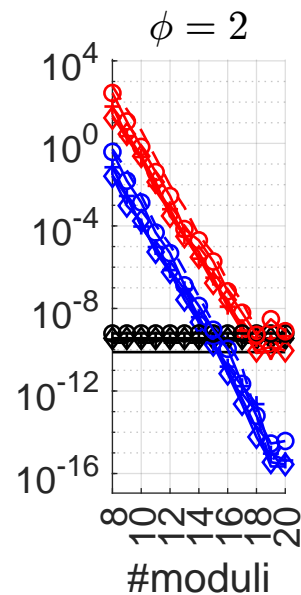
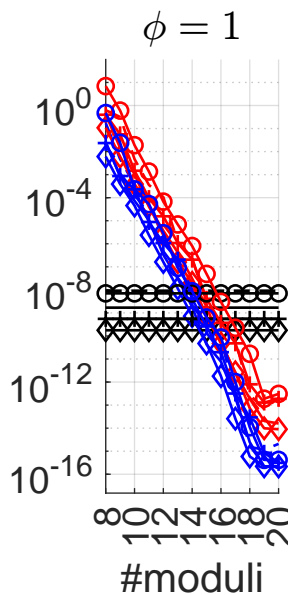
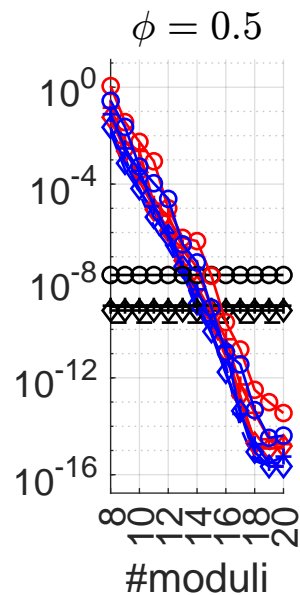


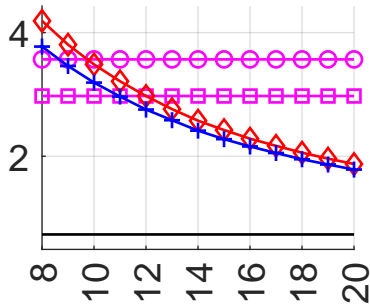
NVIDIA GH200 480GB

$$\max_{ij} |(AB)_{ij} - C_{ij}| / |(AB)_{ij}|$$



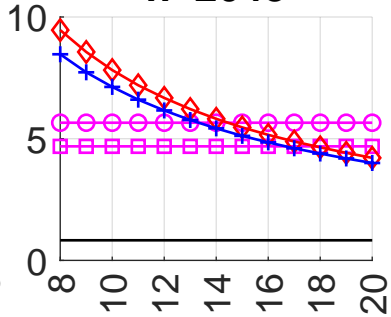
TFLOPS

n=1024



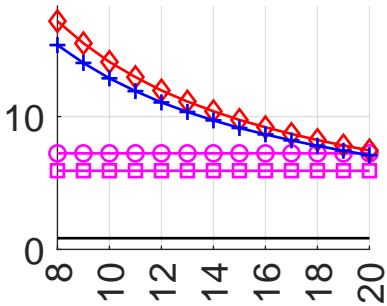
#moduli

n=2048



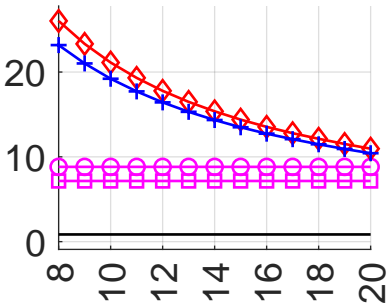
#moduli

n=4096



#moduli

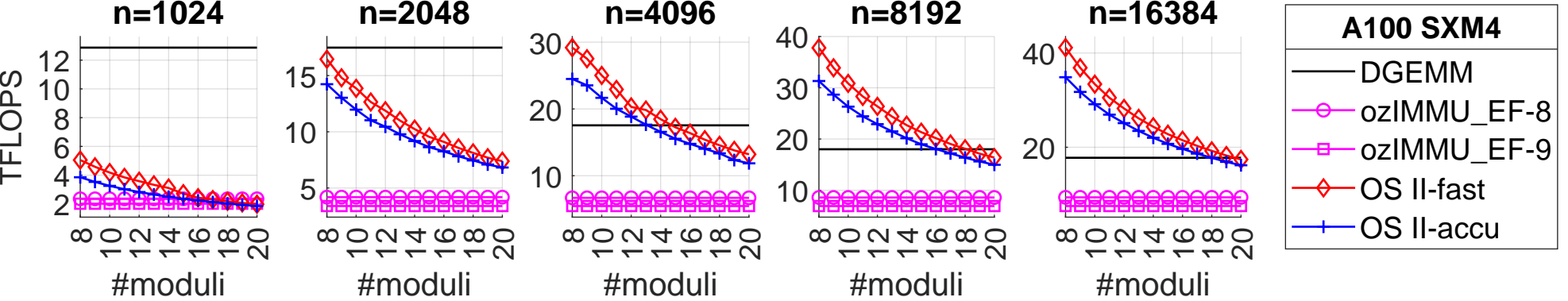
n=8192

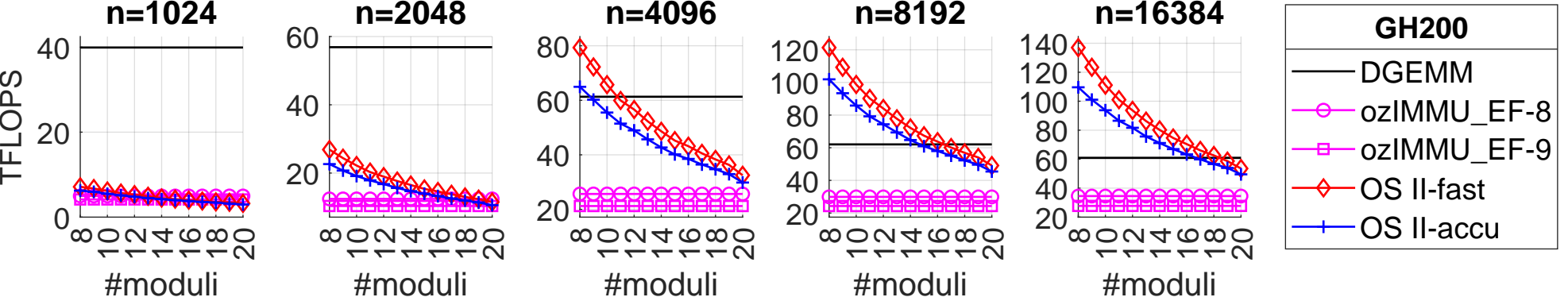


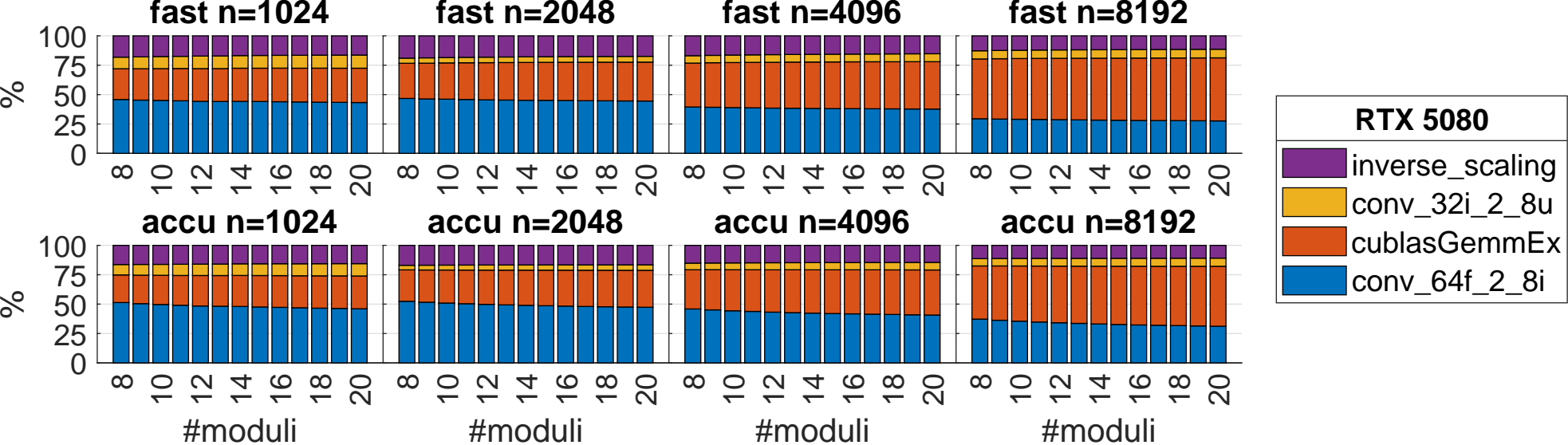
#moduli

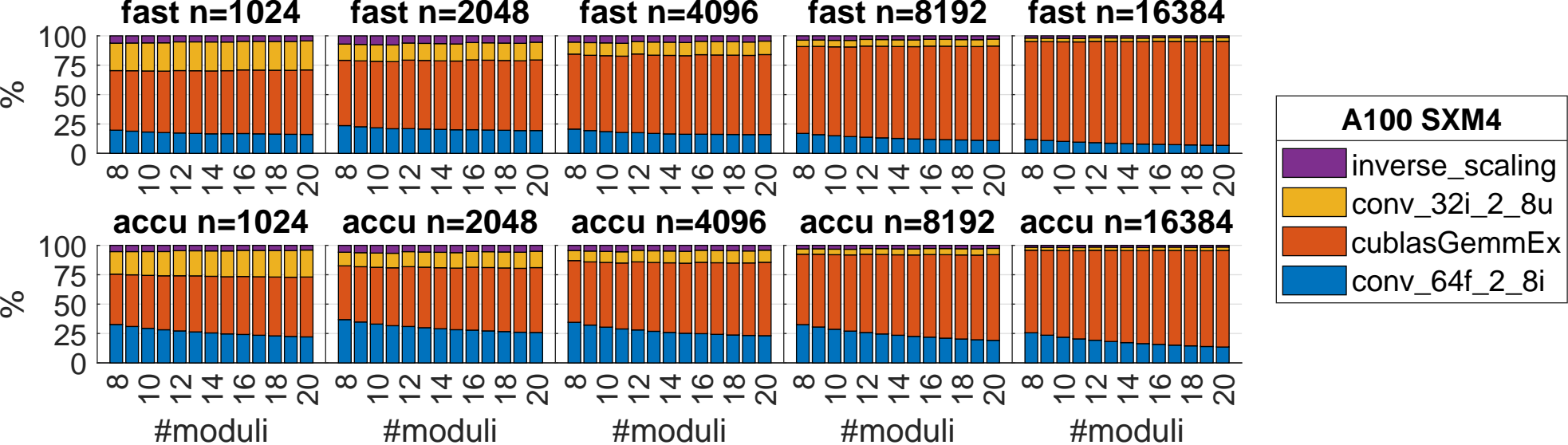
RTX 5080

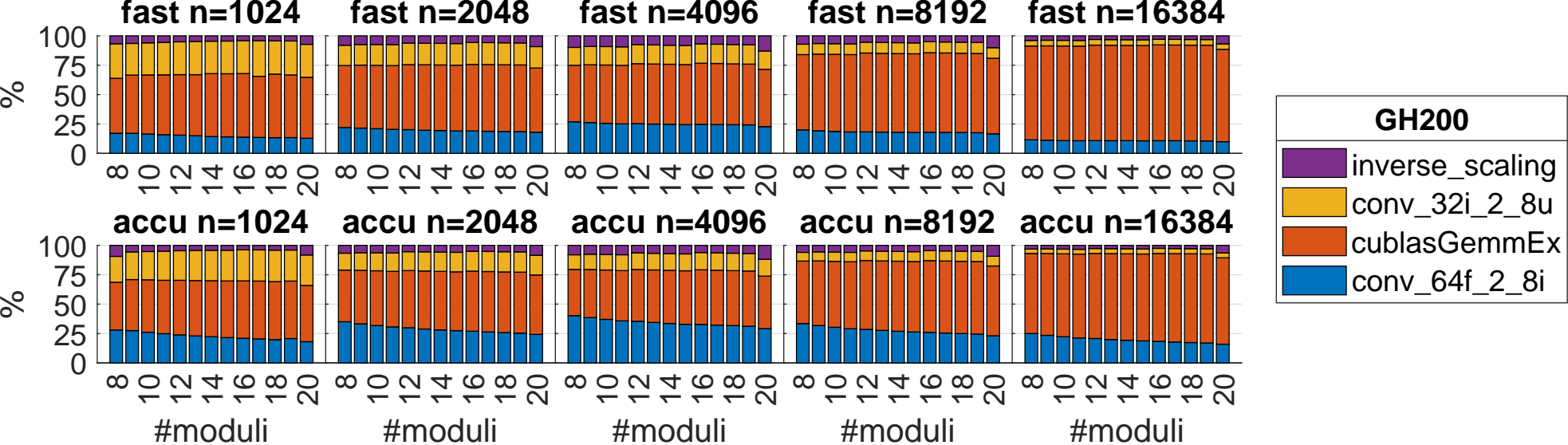
- DGEMM
- ozIMMU_EF-8
- ozIMMU_EF-9
- ◇— OS II-fast
- +— OS II-accu

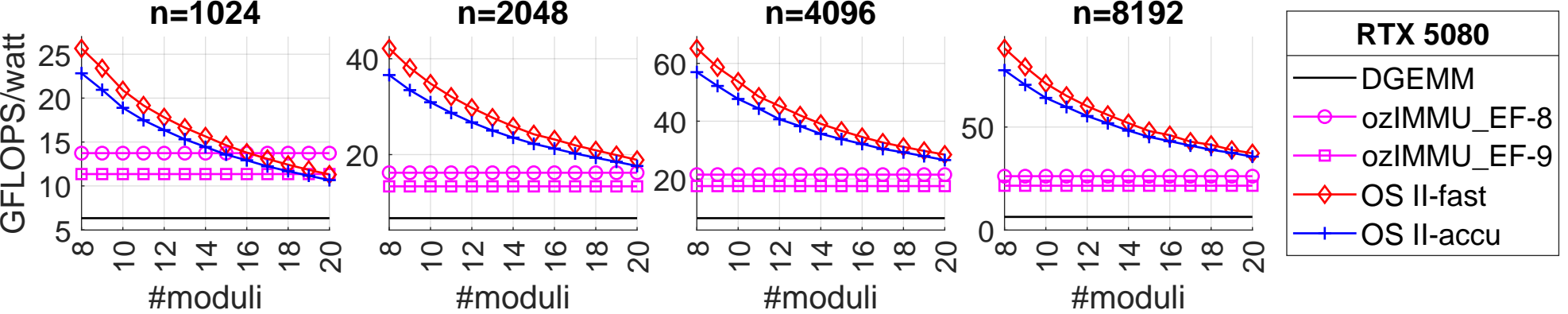






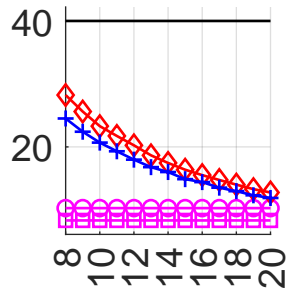




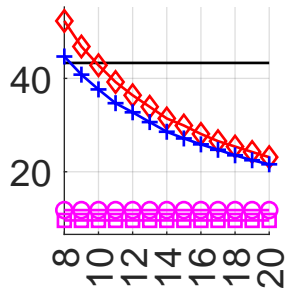


Gflops/watt

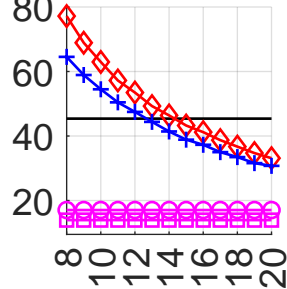
n=1024



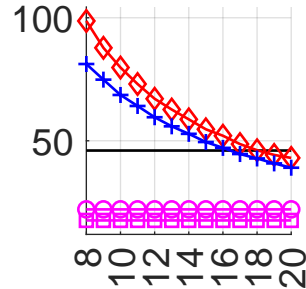
n=2048



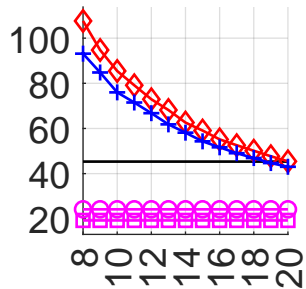
n=4096



n=8192



n=16384

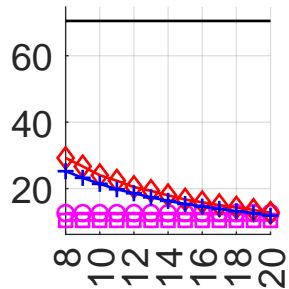


A100 SXM4

- DGEMM
- ozIMMU_EF-8
- ozIMMU_EF-9
- ◇— OS II-fast
- +— OS II-accu

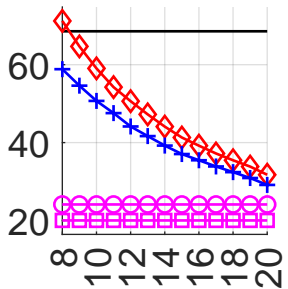
Gflops/watt

n=1024



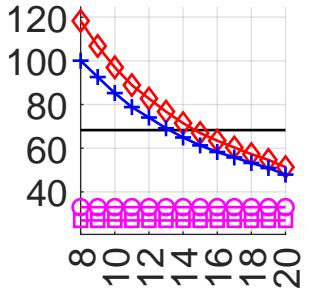
#moduli

n=2048



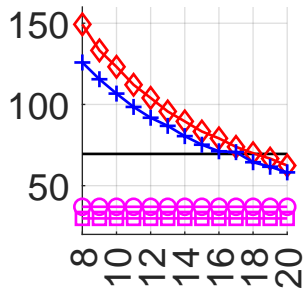
#moduli

n=4096



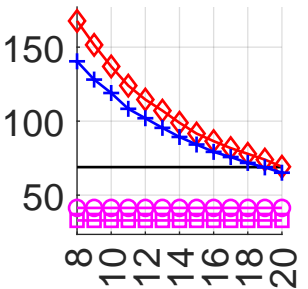
#moduli

n=8192



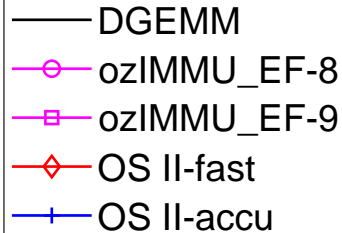
#moduli

n=16384

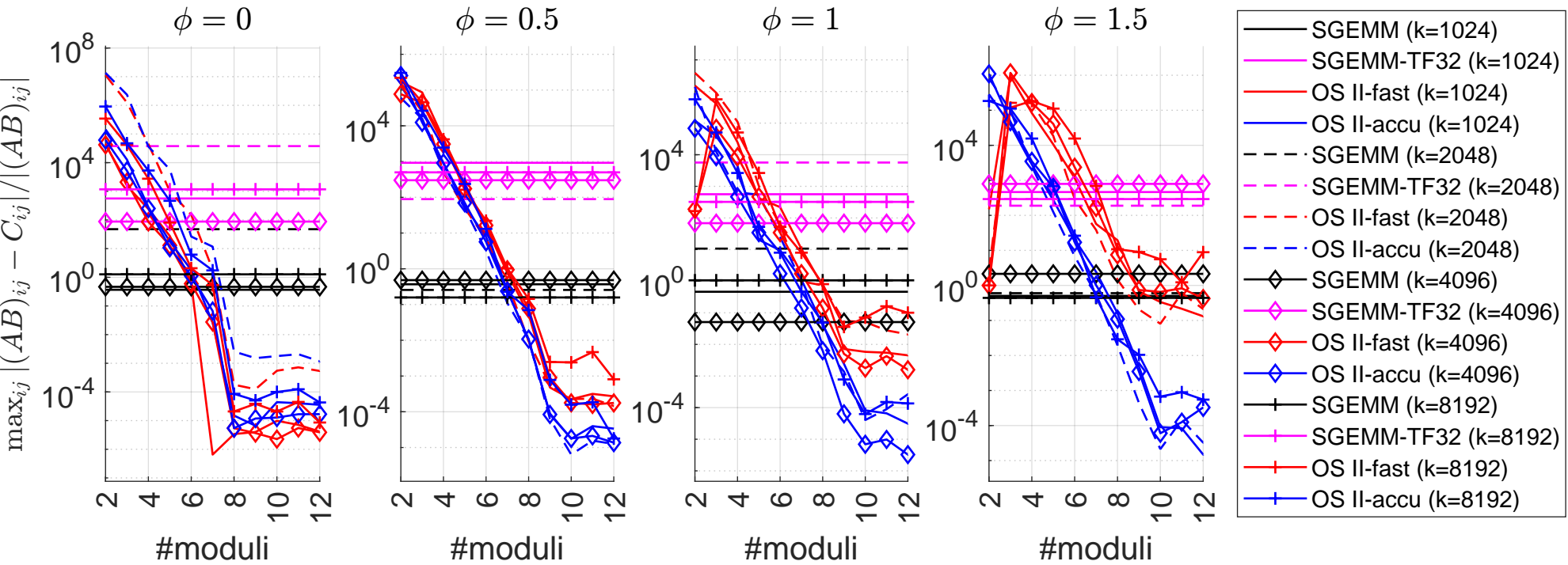


#moduli

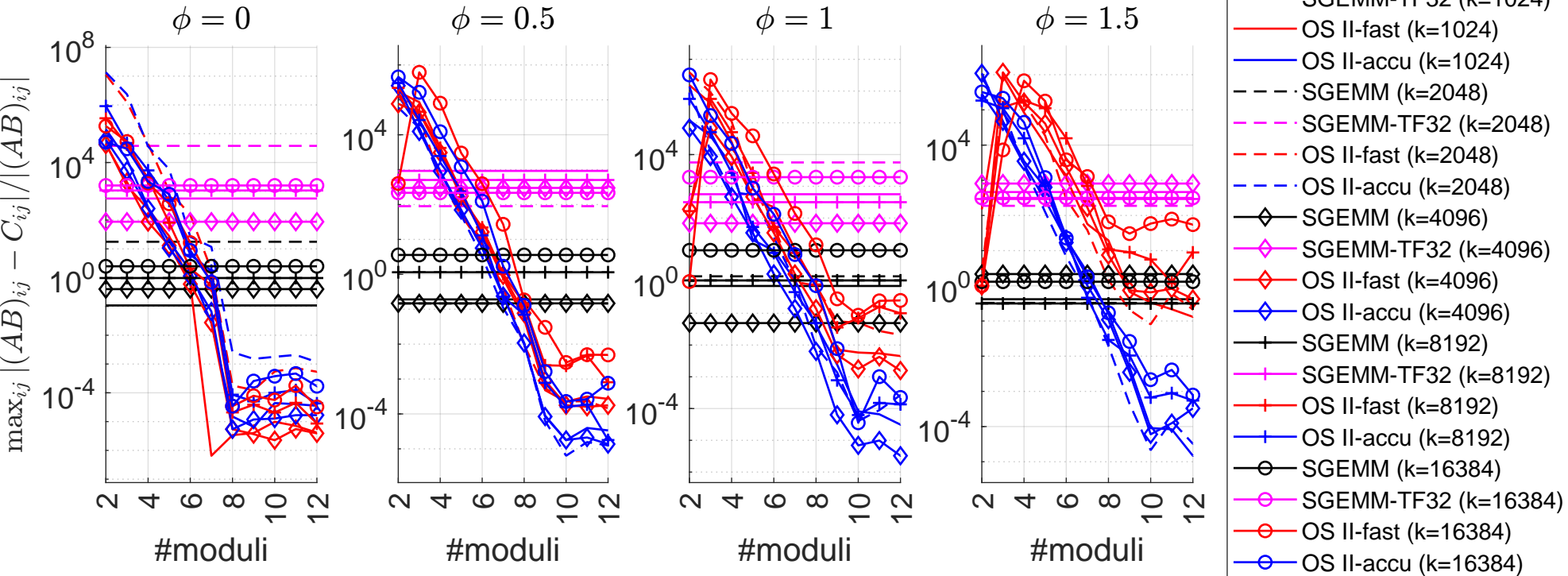
GH200



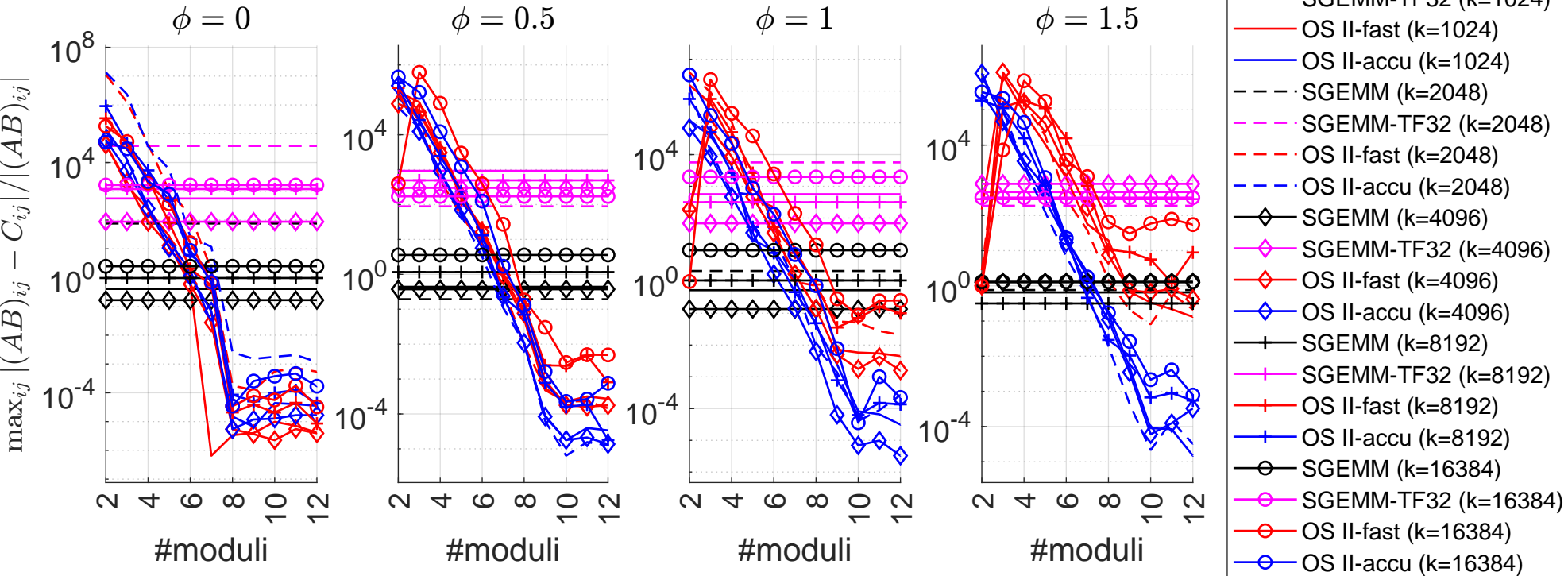
NVIDIA GeForce RTX 5080

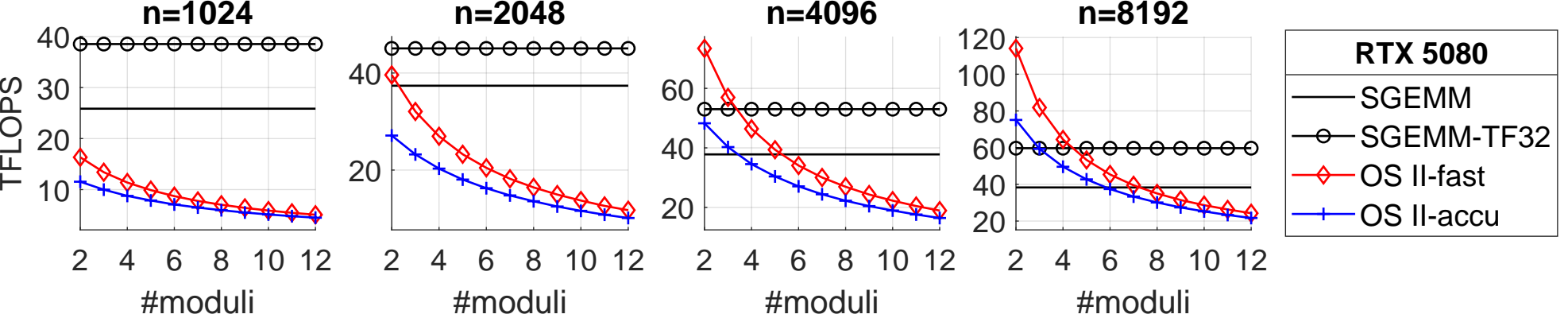


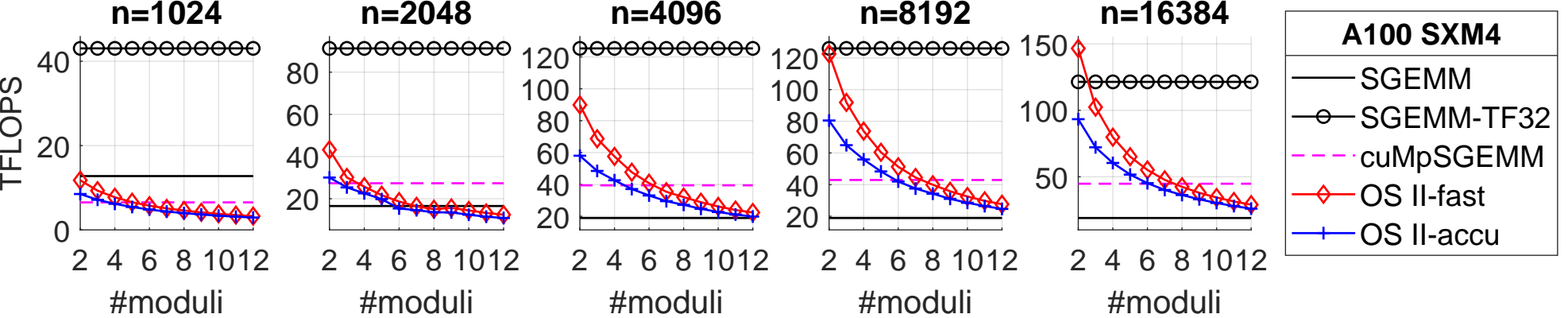
NVIDIA A100 SXM4 80GB

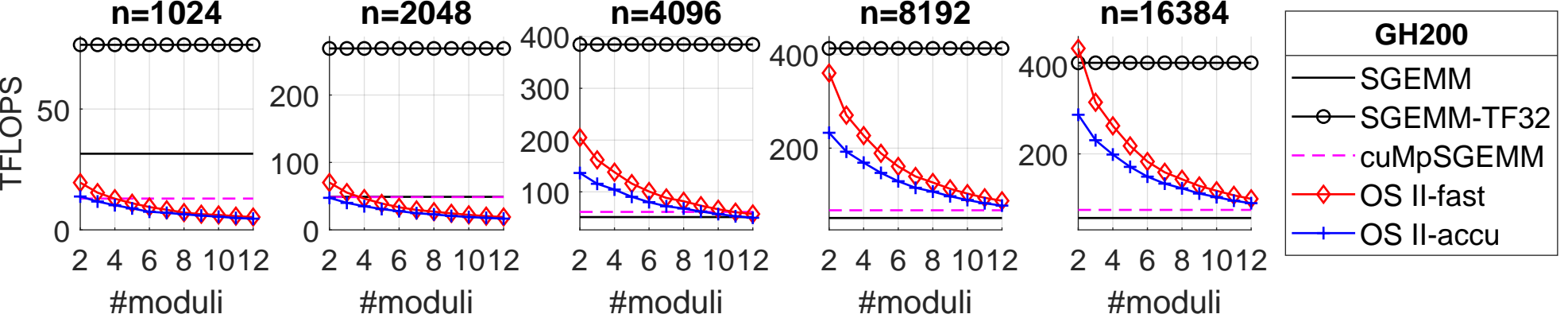


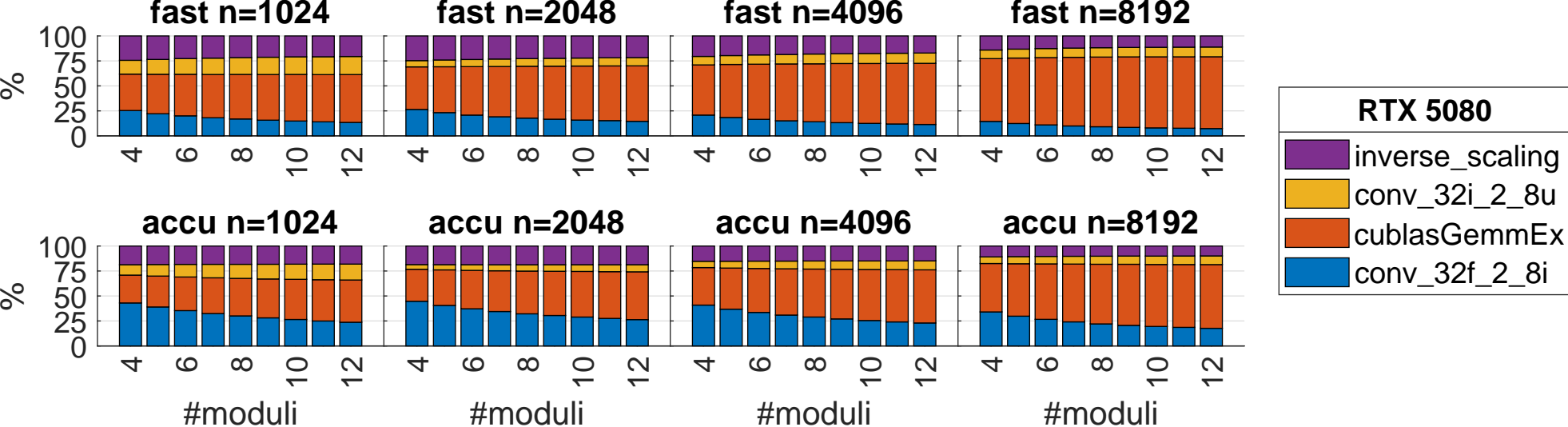
NVIDIA GH200 480GB

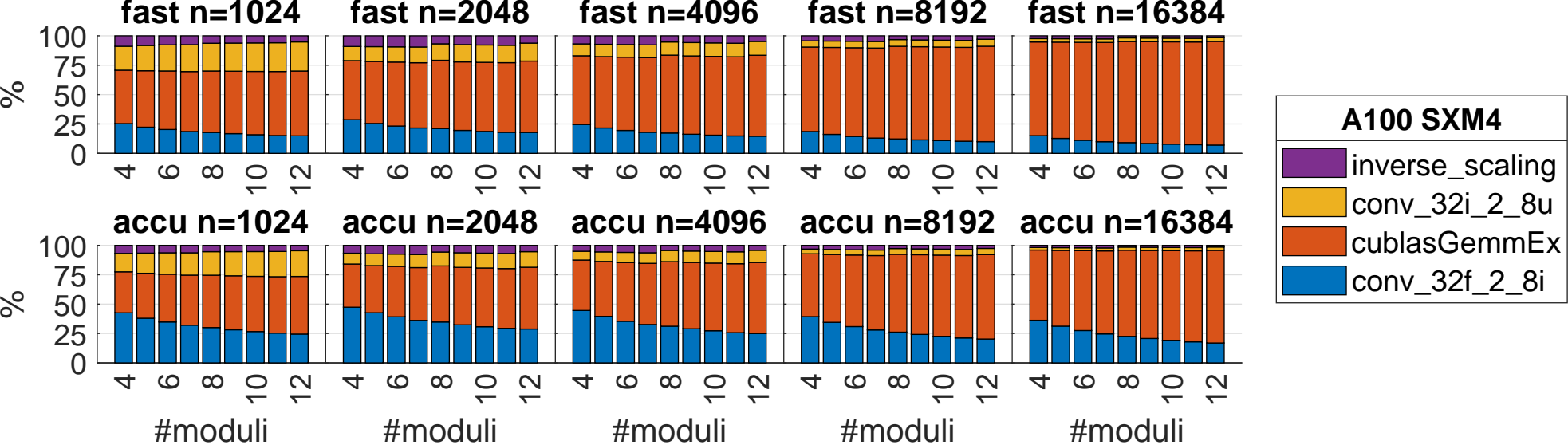


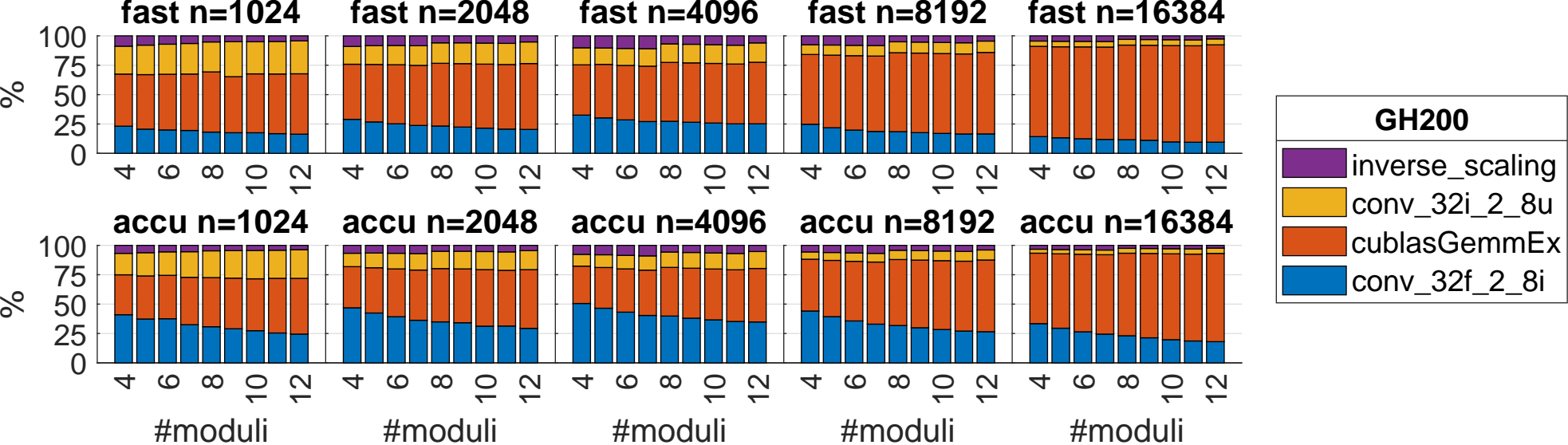






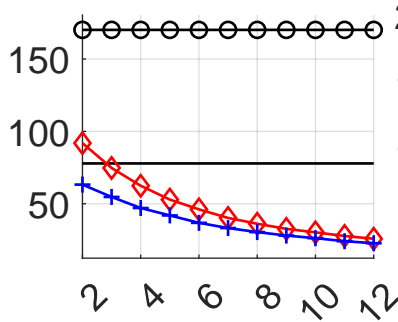






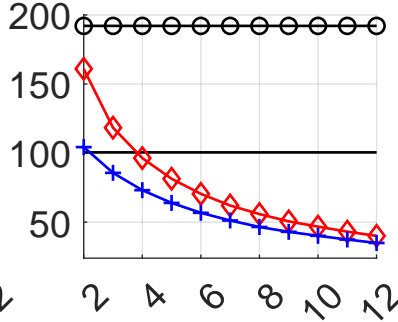
G FloPS/watt

n=1024



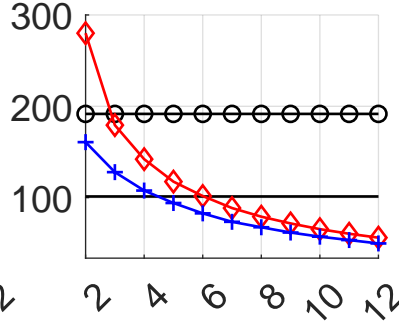
#moduli

n=2048



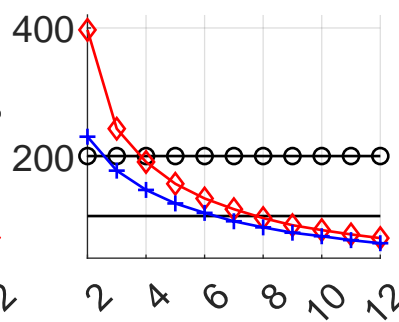
#moduli

n=4096



#moduli

n=8192



#moduli

RTX 5080

- SGEMM
- SGEMM-TF32
- ◇— OS II-fast
- +— OS II-accu

