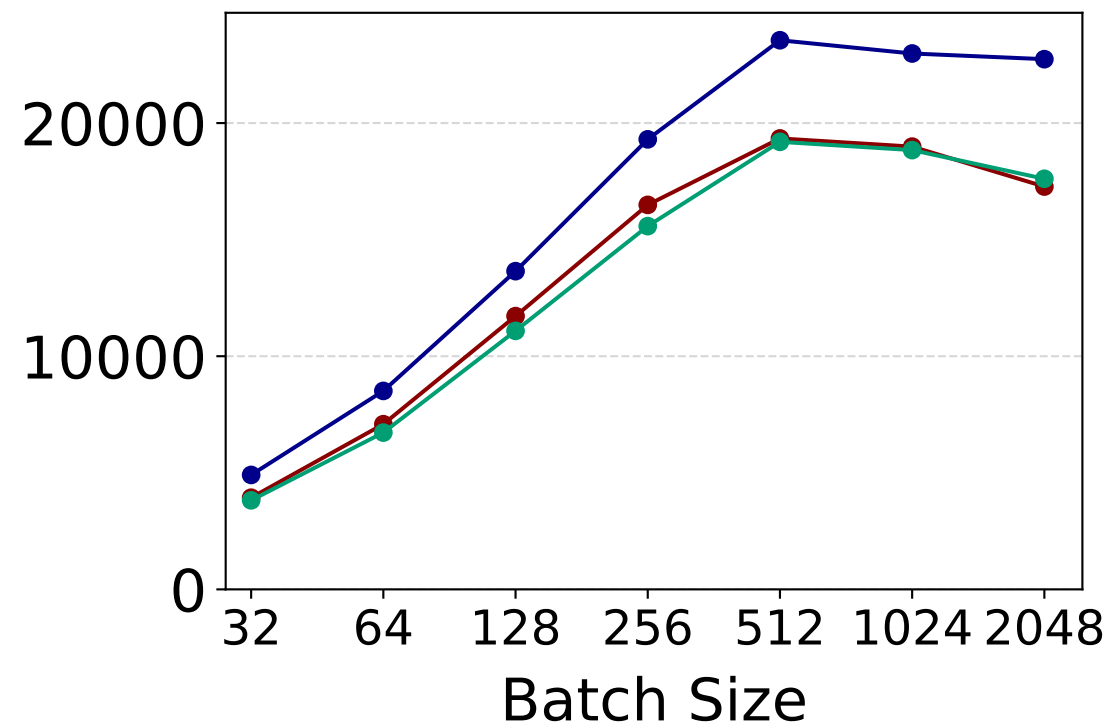


Throughput (toks/s)

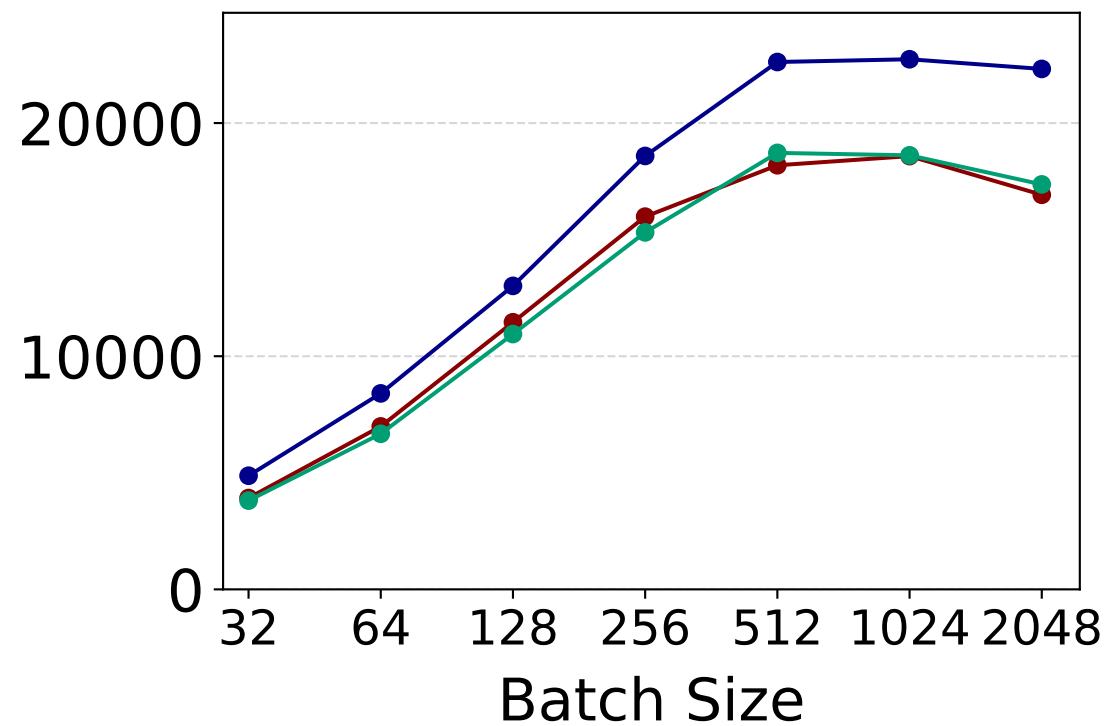
—●— Torch FP16

—●— Dual-FP FP16

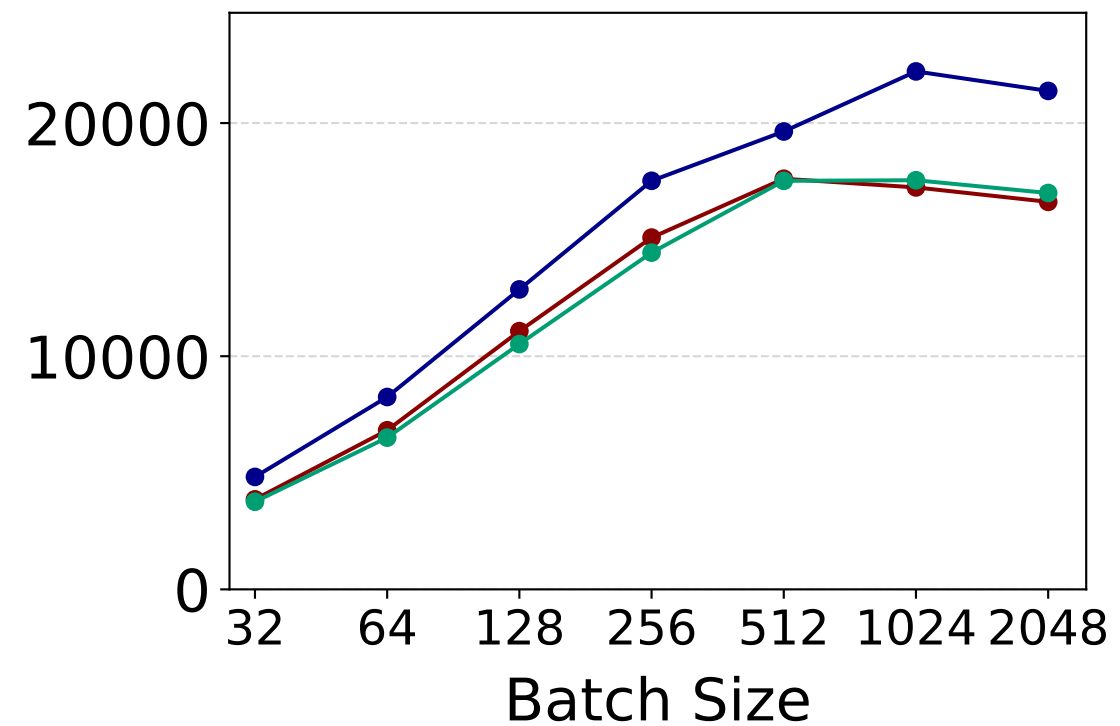
—●— Dual-FP FP8



(a) Input Token = 16



(b) Input Token = 256



(c) Input Token = 1024