Sparsity=0.9, Seq len=8192, num h=8 140 algs Pytorch-fp16 120 vectorSparse-fp16 Latency(ms) 100 Magicube-16b8b Magicube-8b8b 80 Magicube-8b4b 60 Magicube-4b4b 40 20 0 batchsize=2 batchsize=8