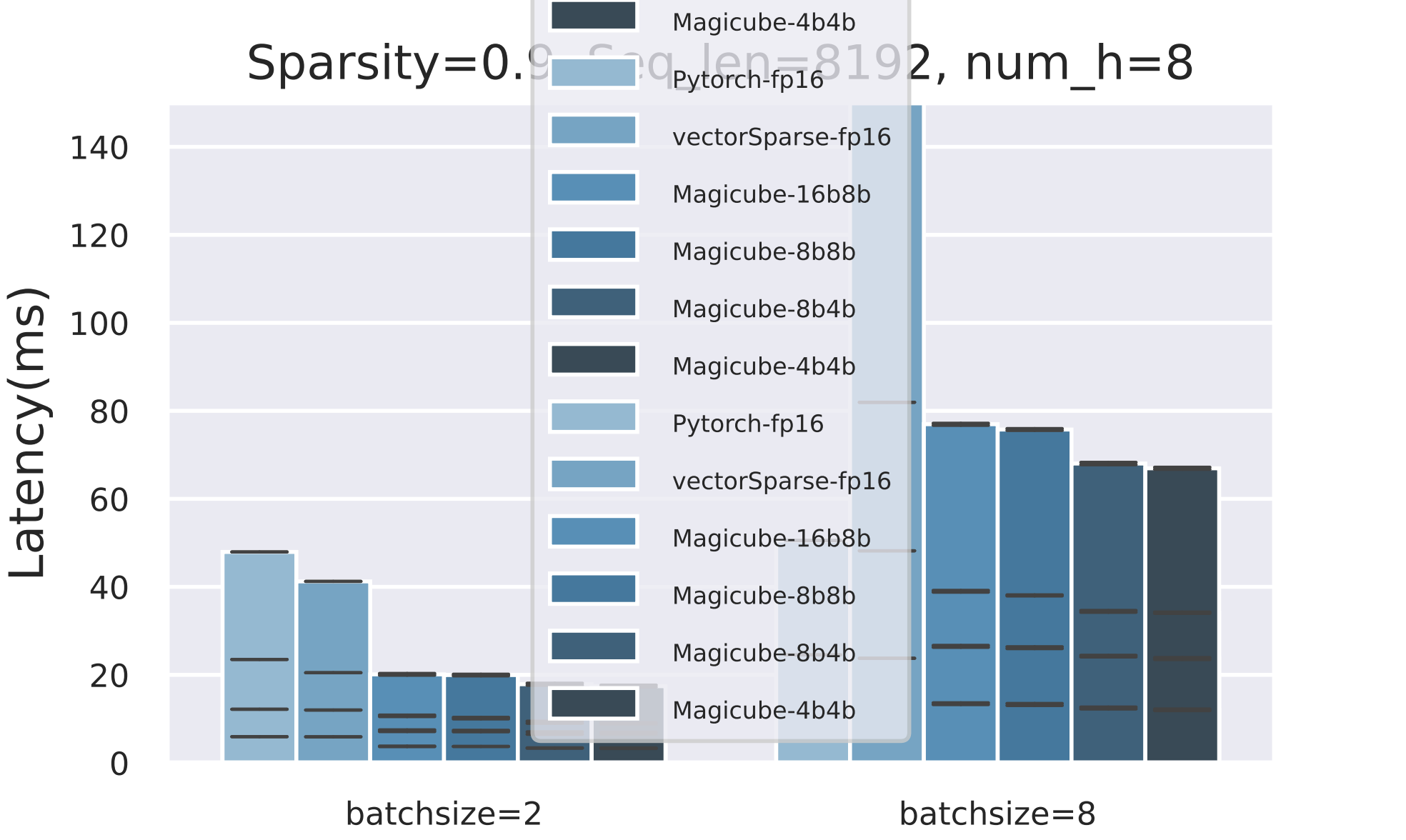
Sparsity=0.9, Seq_len=8192, num_h=8
Latency(ms)
0
20
40
60
80
100
120
140
batchsize=2
batchsize=8
Magicube-4b4b
Pytorch-fp16
vectorSparse-fp16
Magicube-16b8b
Magicube-8b8b
Magicube-8b4b
Magicube-4b4b
Pytorch-fp16
vectorSparse-fp16
Magicube-16b8b
Magicube-8b8b
Magicube-8b4b
Magicube-4b4b