Fuse-all:

C: B\*4096\*4096\*2\*4+B\*4096\*S\*2\*+B\*S\*4096\*2+B\*4096\*11008\*2\*3

M: B\*4096+4096\*4096\*4+B\*S\*4096\*2+B\*4096+B\*4096+4096\*11008\*3+ B\*4096

B=1, S=1024: C/M=

B=4096, S=1024: C/M=

Baseline:

C: B\*4096\*4096\*2\*4+B\*4096\*S\*2\*+B\*S\*4096\*2+B\*4096\*11008\*2\*3

M: B\*4096+B\*4096+B\*4096+4096\*4096\*3+B\*4096\*3+B\*4096+B\*S\*4096\*2+B\*4096+B\*4096+4096\*4096+B\*4096+B\*4096+4096\*11008\*2+B\*11008\*2+B\*11008\*2+B\*11008+ B\*11008+11008\*4096+B\*4096

B=1, S=1024: C/M=

B=4096, S=1024: C/M=

Activation占比:

Activation: B\*4096+B\*4096+B\*4096+B\*4096\*3+B\*4096+B\*4096+B\*4096+B\*4096+B\*4096+B\*11008\*2+B\*11008\*2+B\*11008+ B\*11008+B\*4096

Weight:

4096\*4096\*3+B\*S\*4096\*2+4096\*4096+4096\*11008\*2+11008\*4096

Total:

B\*4096+B\*4096+B\*4096+4096\*4096\*3+B\*4096\*3+B\*4096+B\*S\*4096\*2+B\*4096+B\*4096+4096\*4096+B\*4096+B\*4096+4096\*11008\*2+B\*11008\*2+B\*11008\*2+B\*11008+ B\*11008+11008\*4096+B\*4096

B=4096, S=1024: Activation/ Total=, Activation=, Weight=