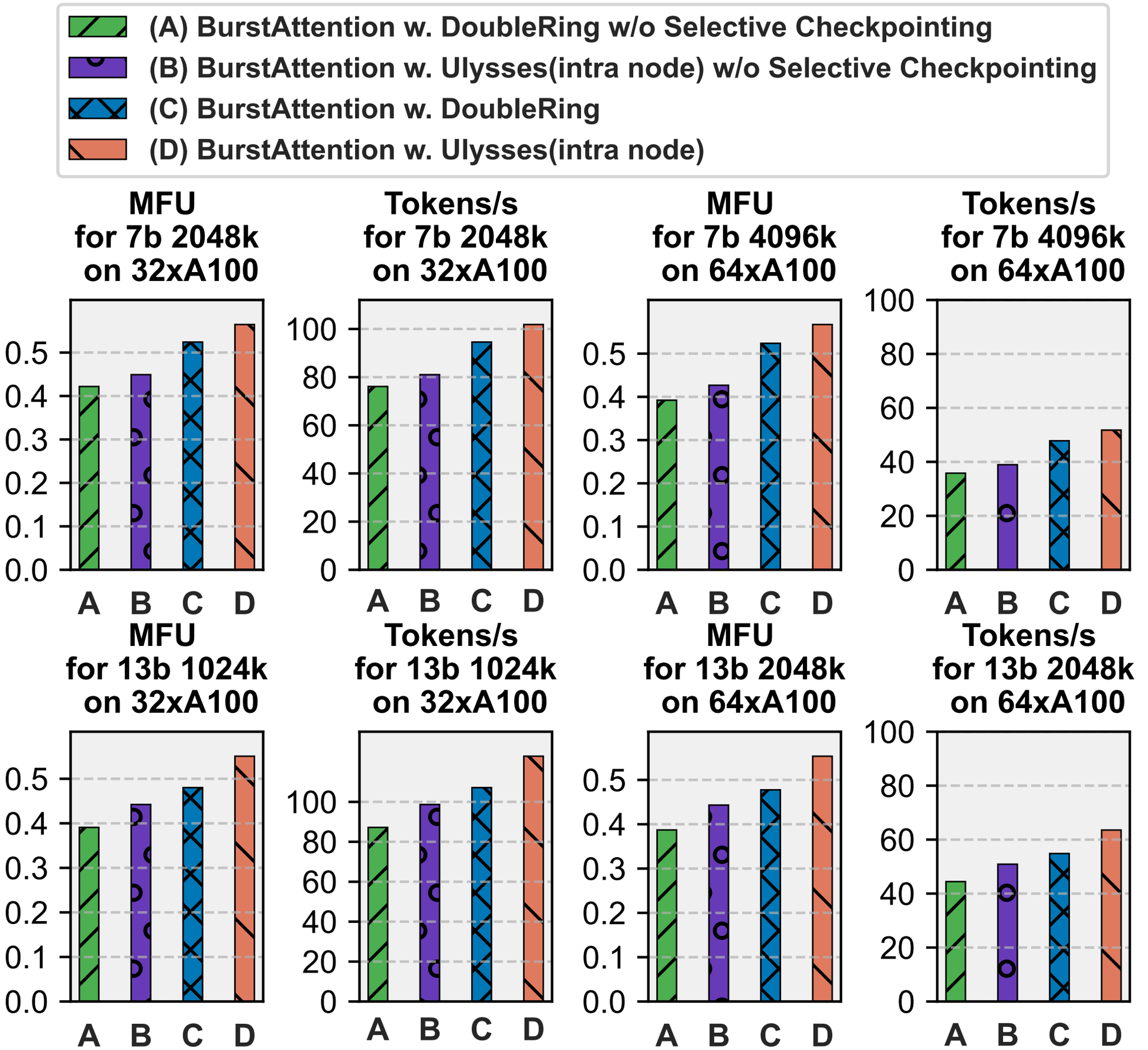
(A) BurstAttention w. DoubleRing w/o Selective Checkpointing
(B) BurstAttention w. Ulysses(intra node) w/o Selective Checkpointing
(C) BurstAttention w. DoubleRing
(D) BurstAttention w. Ulysses(intra node)
MFU for 7b 2048k on 32xA100
Tokens/s for 7b 2048k on 32xA100
MFU for 7b 4096k on 64xA100
Tokens/s for 7b 4096k on 64xA100
MFU for 13b 1024k on 32xA100
Tokens/s for 13b 1024k on 32xA100
MFU for 13b 2048k on 64xA100
Tokens/s for 13b 2048k on 64xA100
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100
A
B
C
D