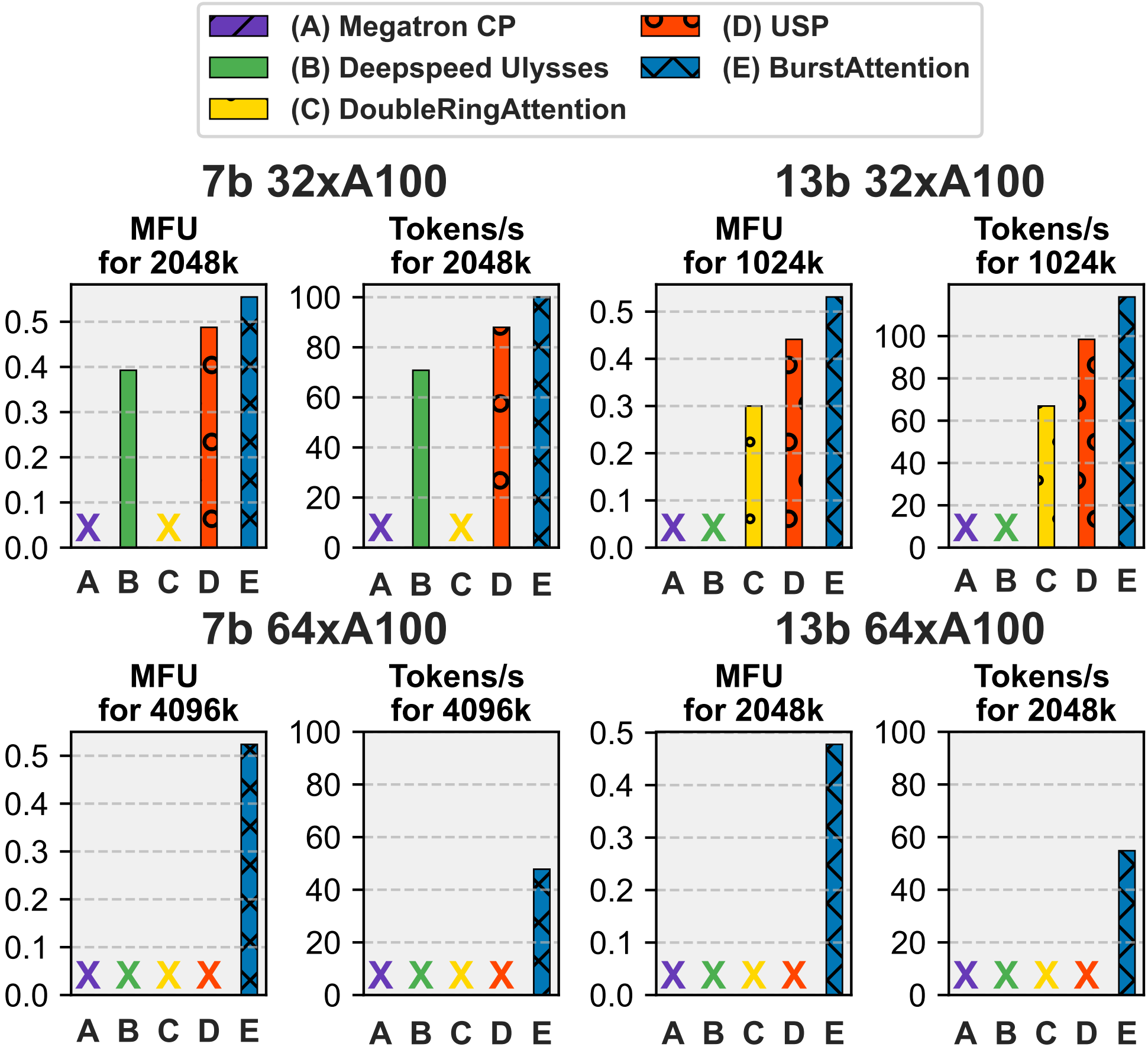
(A) Megatron CP
(B) Deepspeed Ulysses
(C) DoubleRingAttention
(D) USP
(E) BurstAttention
7b 32xA100
MFU for 2048k
Tokens/s for 2048k
13b 32xA100
MFU for 1024k
Tokens/s for 1024k
7b 64xA100
MFU for 4096k
Tokens/s for 4096k
13b 64xA100
MFU for 2048k
Tokens/s for 2048k
A B C D E