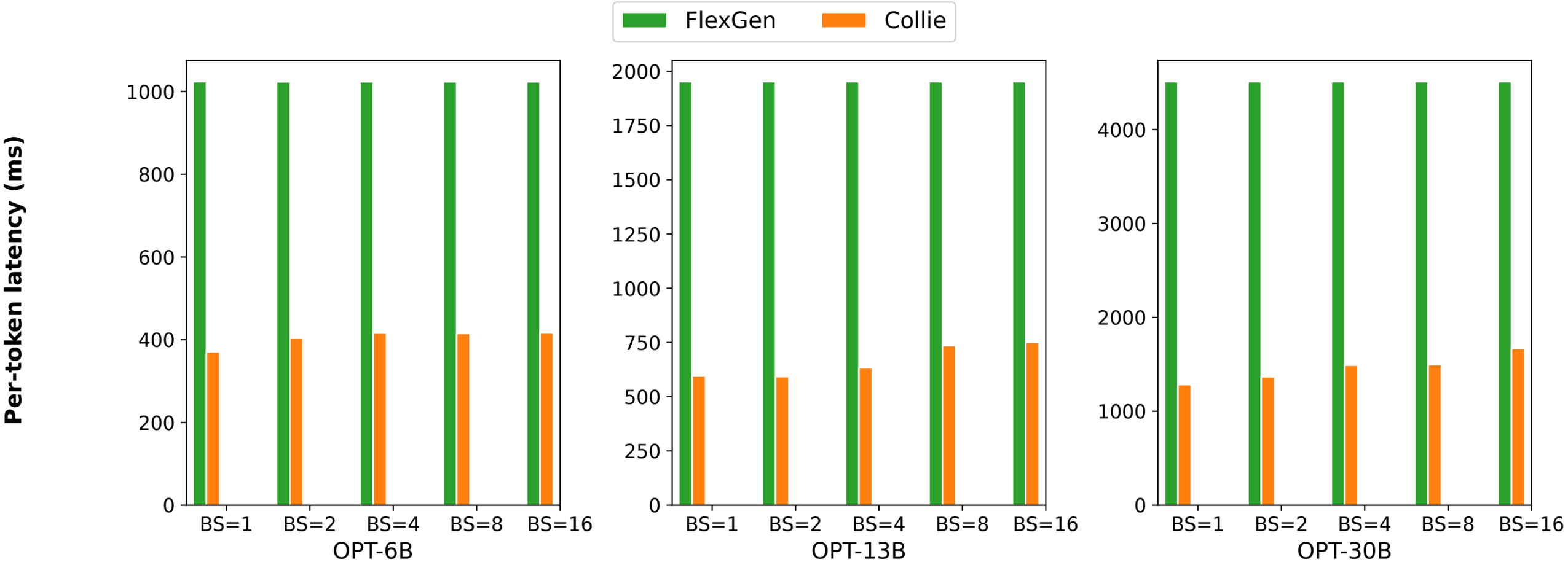
FlexGen
Collie
Per-token latency (ms)
0
200
400
600
800
1000
BS=1
BS=2
BS=4
BS=8
BS=16
OPT-6B
0
250
500
750
1000
1250
1500
1750
2000
BS=1
BS=2
BS=4
BS=8
BS=16
OPT-13B
0
1000
2000
3000
4000
BS=1
BS=2
BS=4
BS=8
BS=16
OPT-30B