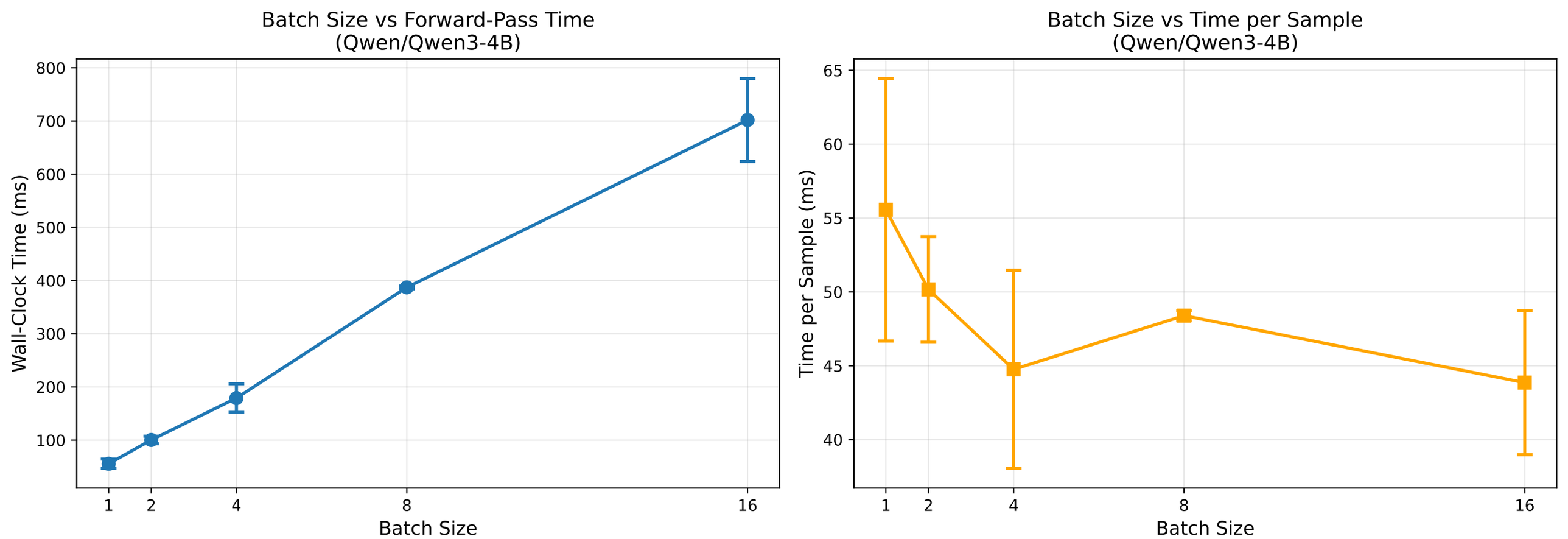
Batch Size vs Forward-Pass Time
(Qwen/Qwen3-4B)
Wall-Clock Time (ms)
800
700
600
500
400
300
200
100
1
2
4
8
16
Batch Size
Batch Size vs Time per Sample
(Qwen/Qwen3-4B)
Time per Sample (ms)
65
60
55
50
45
40
1
2
4
8
16
Batch Size