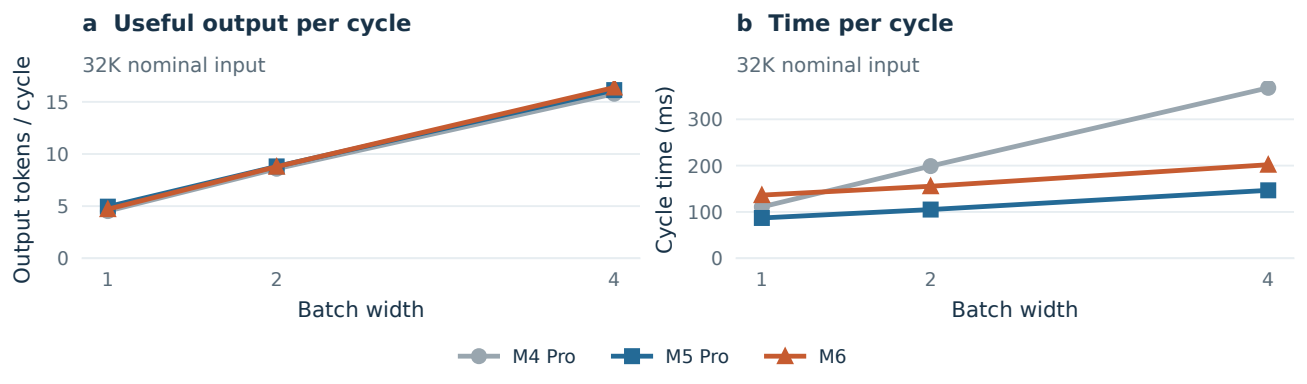

a Useful output per cycle
32K nominal input
Output tokens / cycle
15
10
5
0
1
2
4
Batch width
b Time per cycle
32K nominal input
Cycle time (ms)
300
200
100
0
1
2
4
Batch width
M4 Pro
M5 Pro
M6