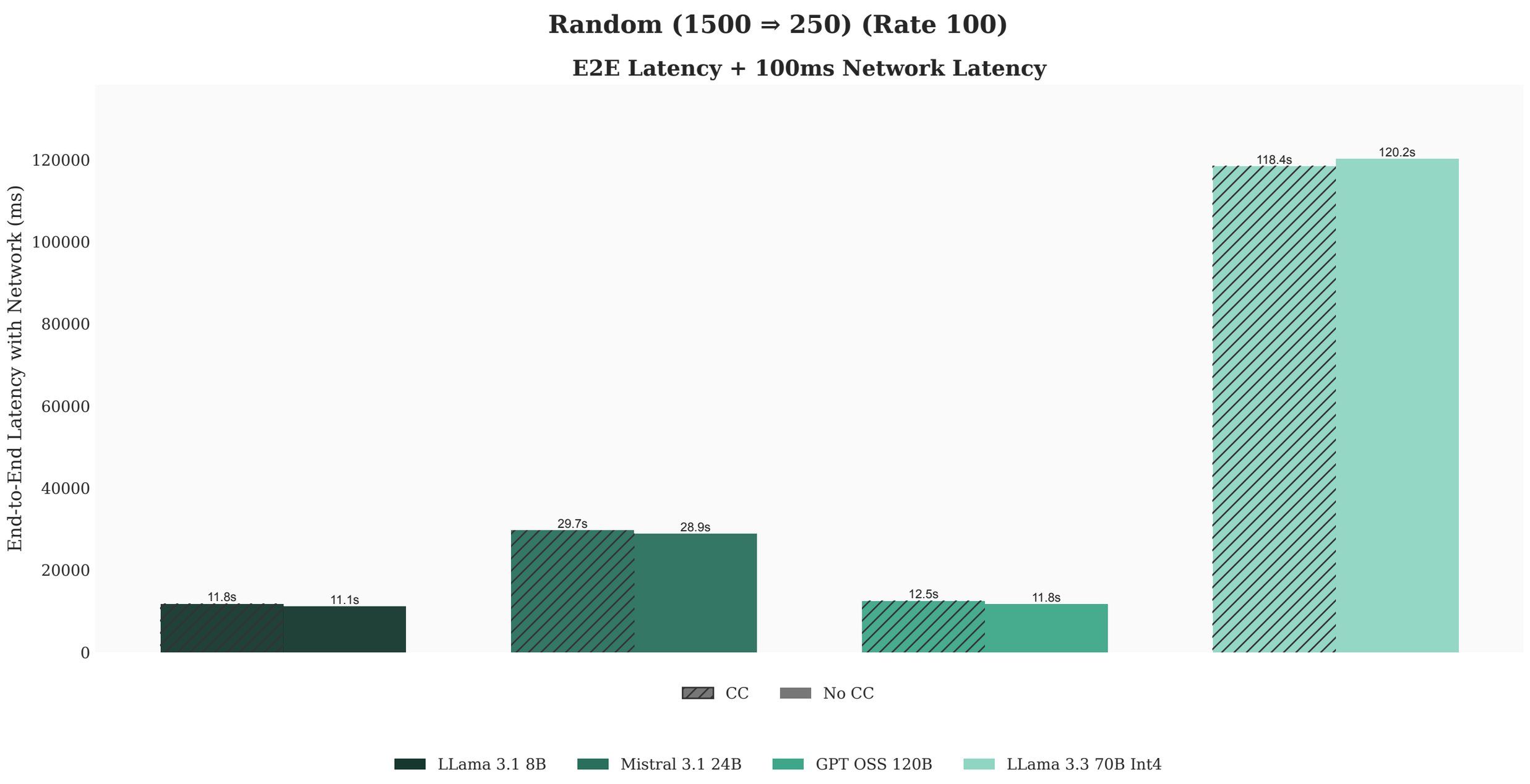
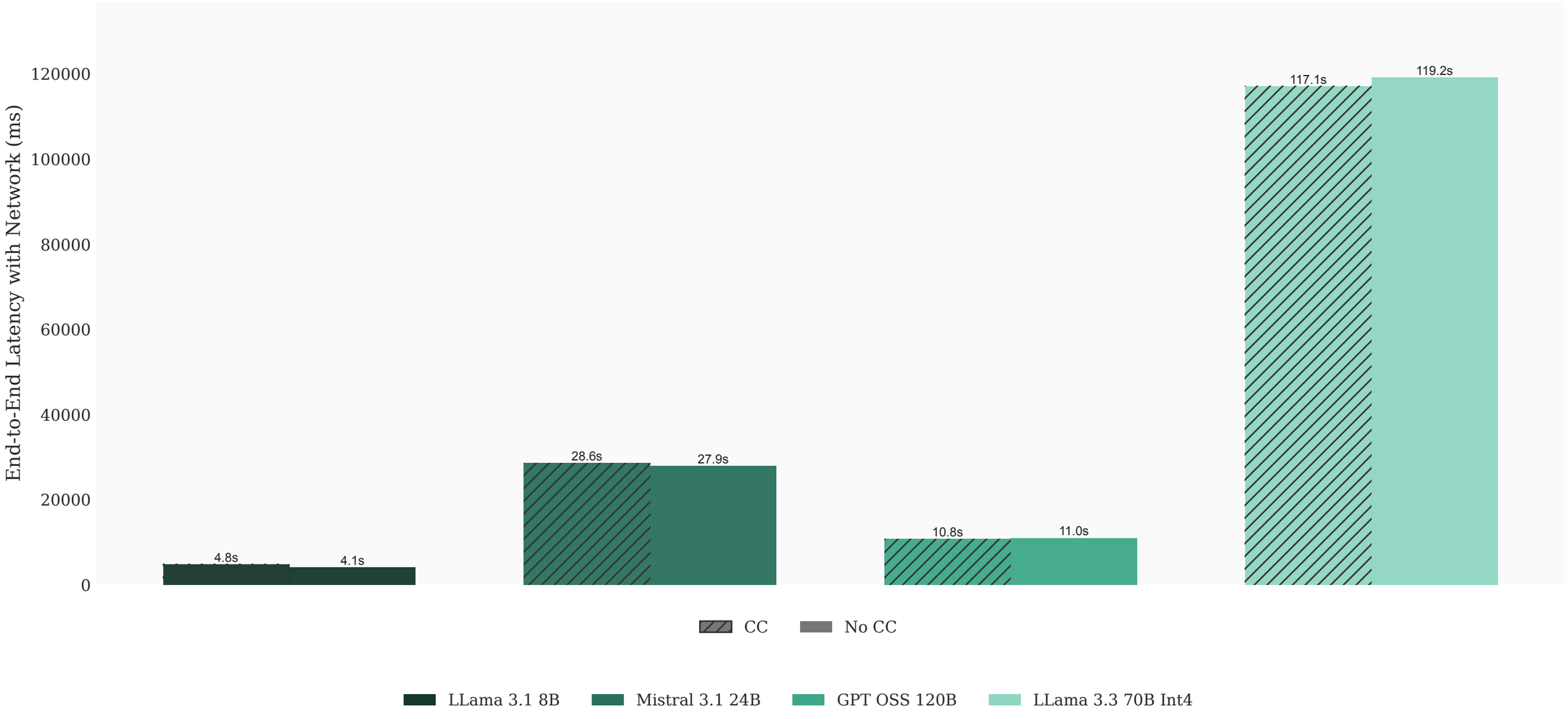
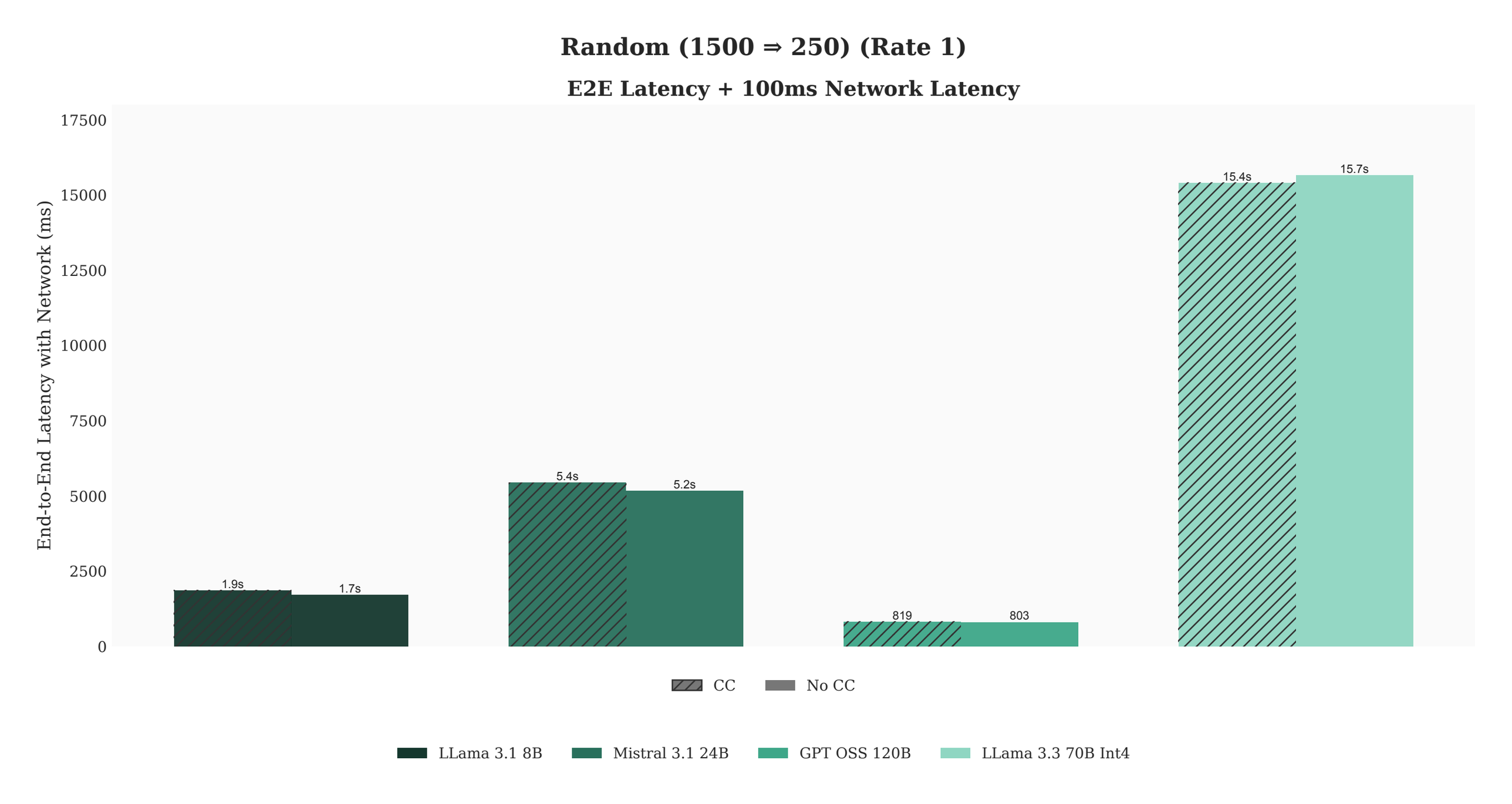
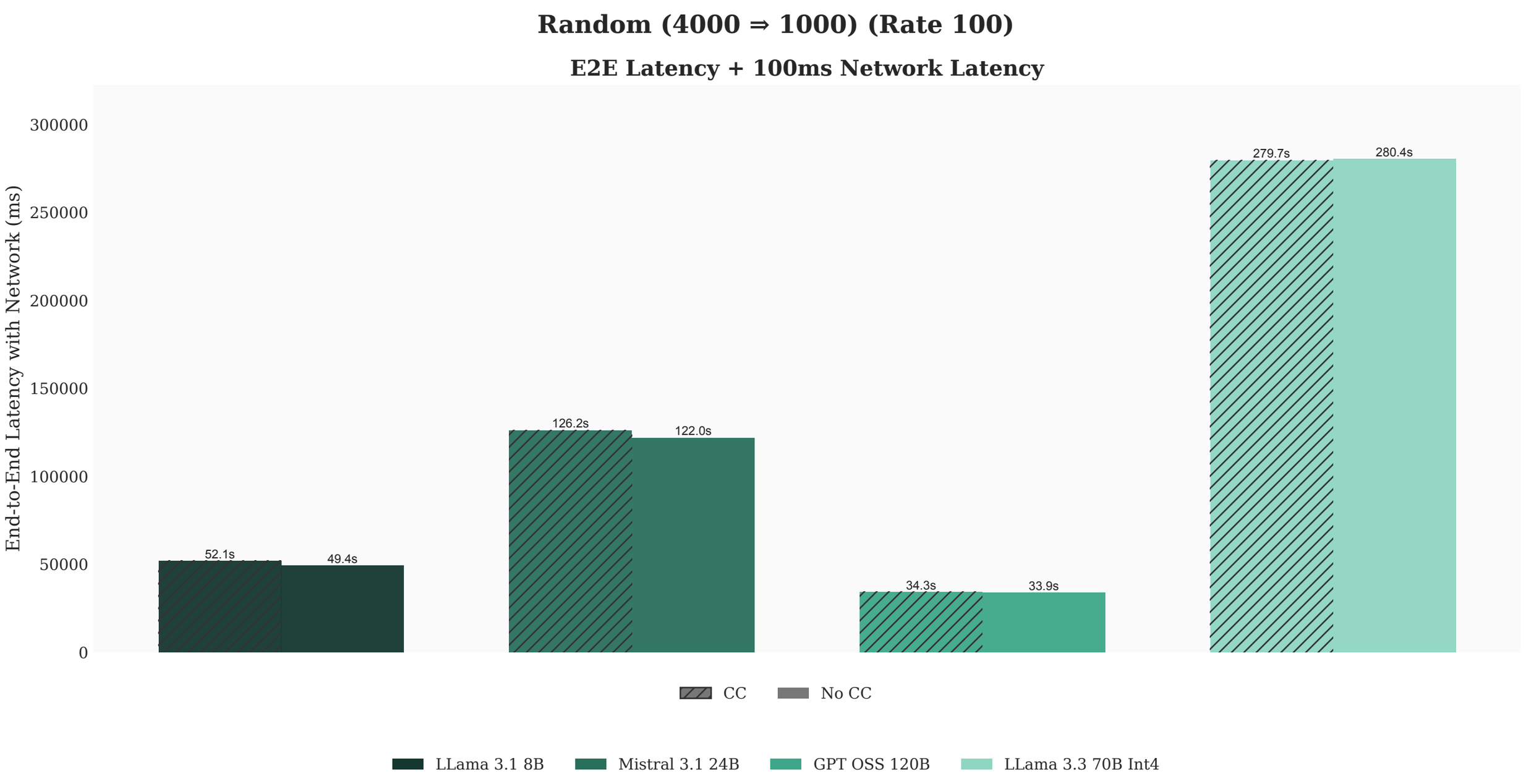

Random (1500 ⇒ 250) (Rate 100)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
120000
100000
80000
60000
40000
20000
0
11.8s
11.1s
29.7s
28.9s
12.5s
11.8s
118.4s
120.2s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

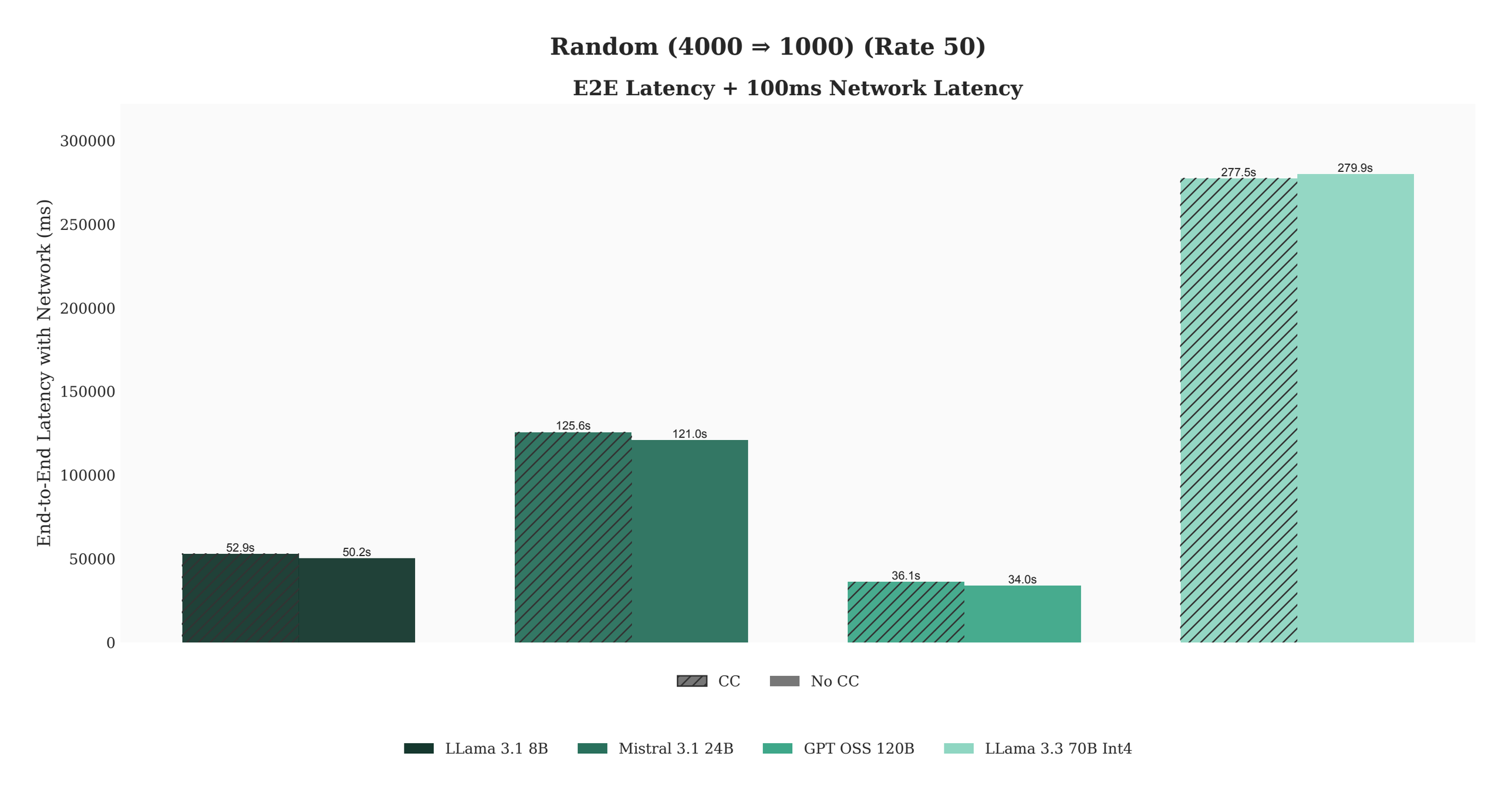
Random (1500 ⇒ 250) (Rate 50)

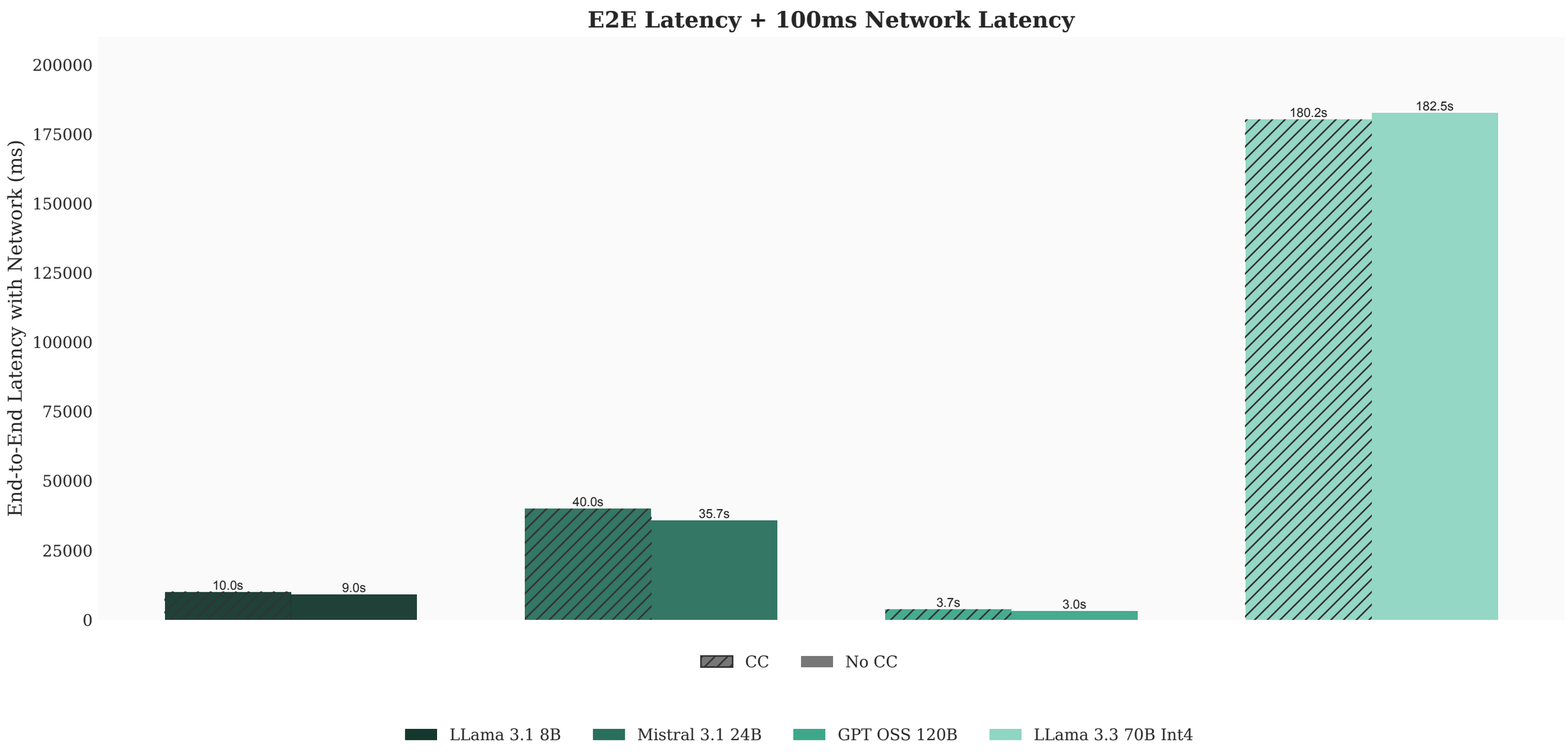
E2E Latency + 100ms Network Latency

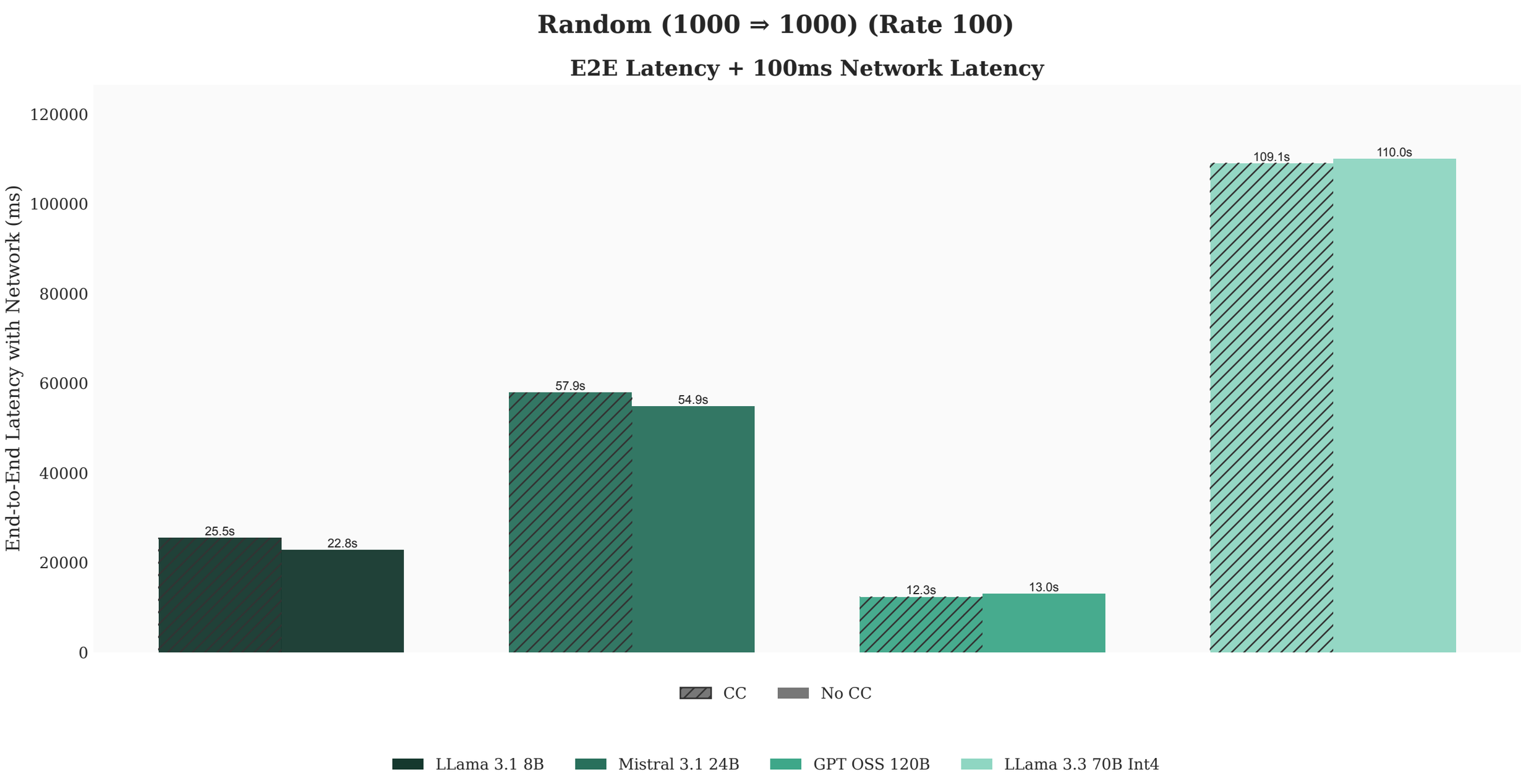


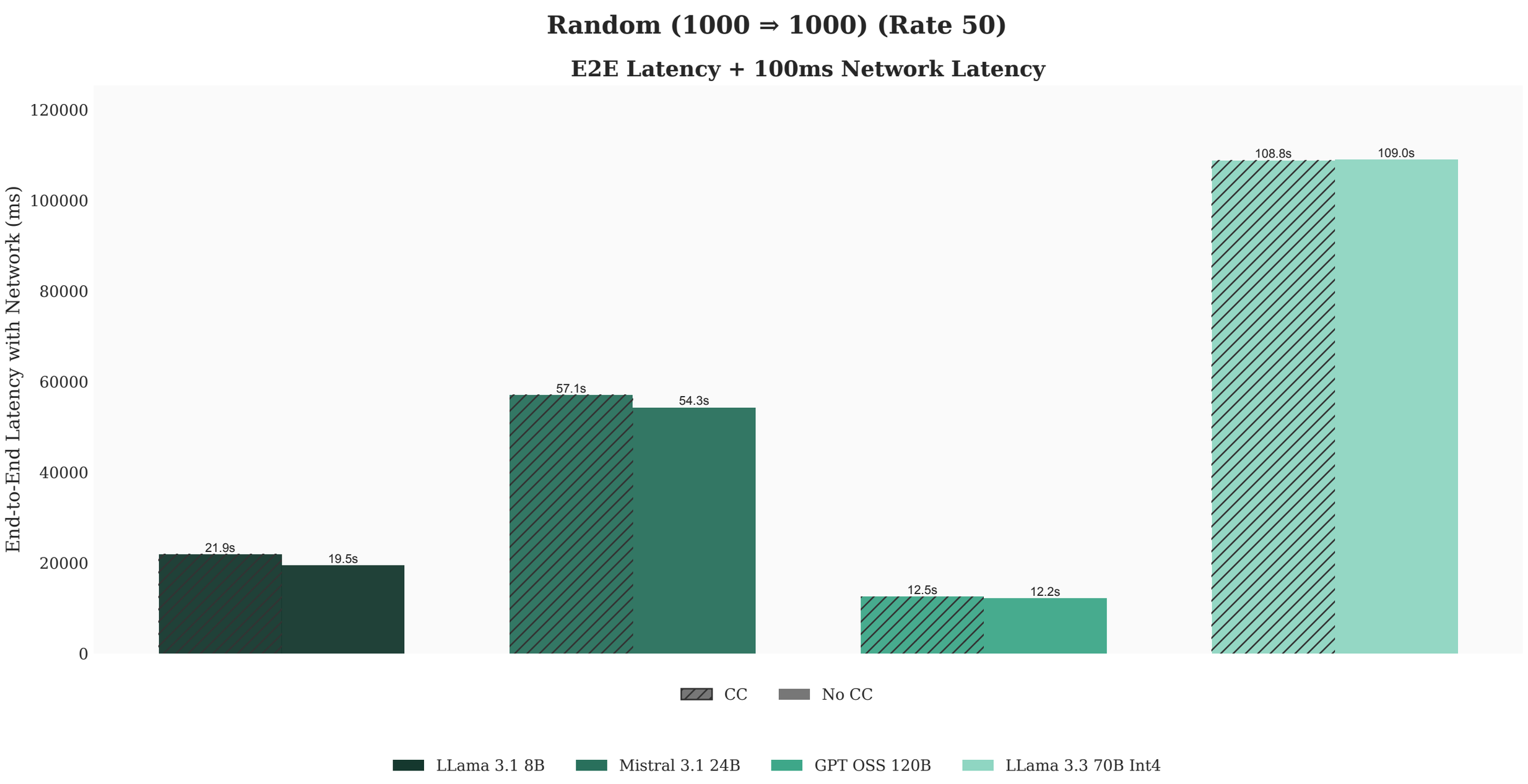

Random (1500 ⇒ 250) (Rate 1)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
17500
15000
12500
10000
7500
5000
2500
0
1.9s
1.7s
5.4s
5.2s
819
803
15.4s
15.7s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

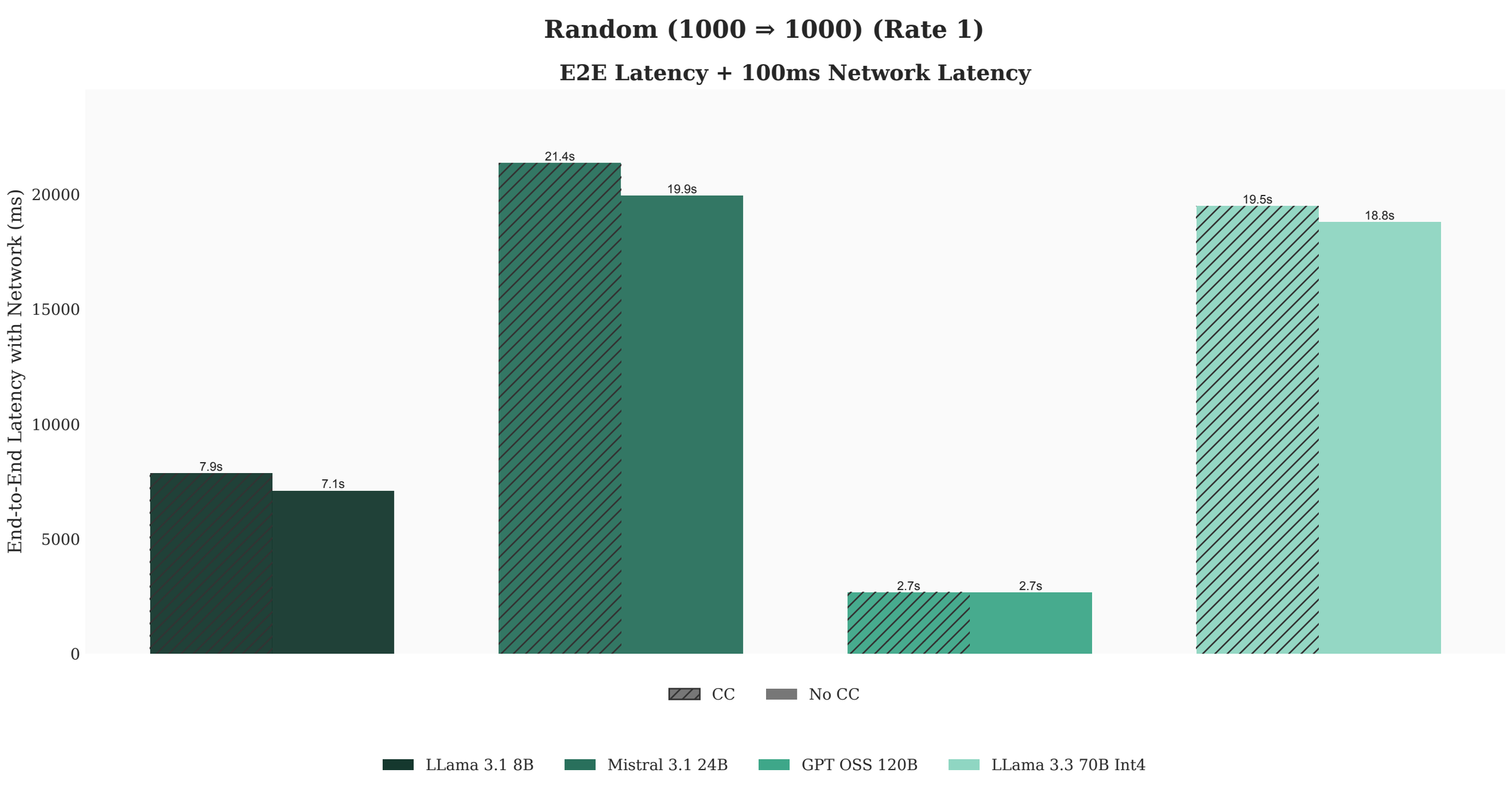

Random (4000 ⇒ 1000) (Rate 100)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
300000
250000
200000
150000
100000
50000
0
52.1s
49.4s
126.2s
122.0s
34.3s
33.9s
279.7s
280.4s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

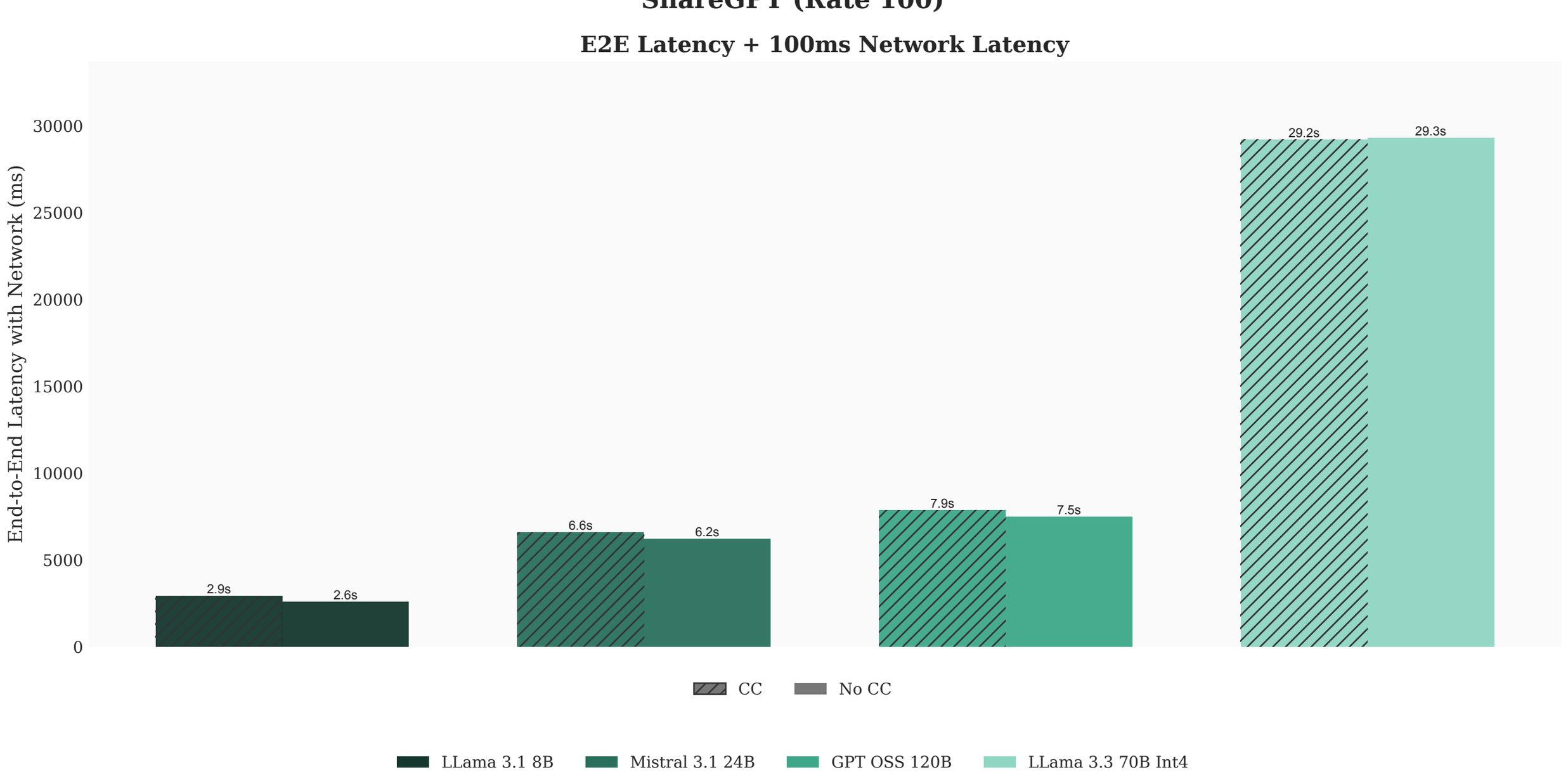

Random (4000 ⇒ 1000) (Rate 50)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
300000
250000
200000
150000
100000
50000
0
52.9s
50.2s
125.6s
121.0s
36.1s
34.0s
277.5s
279.9s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

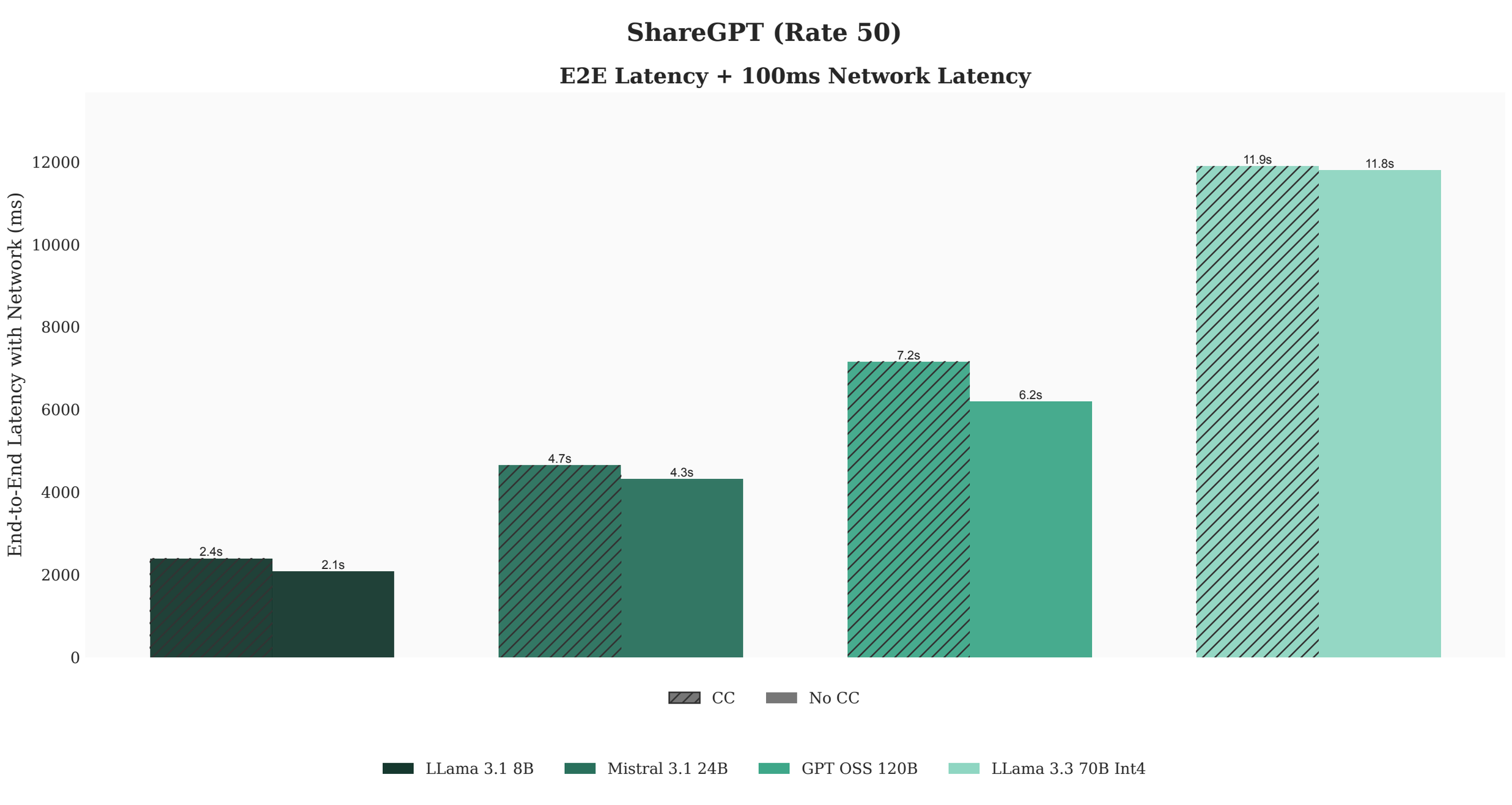

Random (4000 ⇒ 1000) (Rate 1)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
200000
175000
150000
125000
100000
75000
50000
25000
0
10.0s
9.0s
40.0s
35.7s
3.7s
3.0s
180.2s
182.5s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

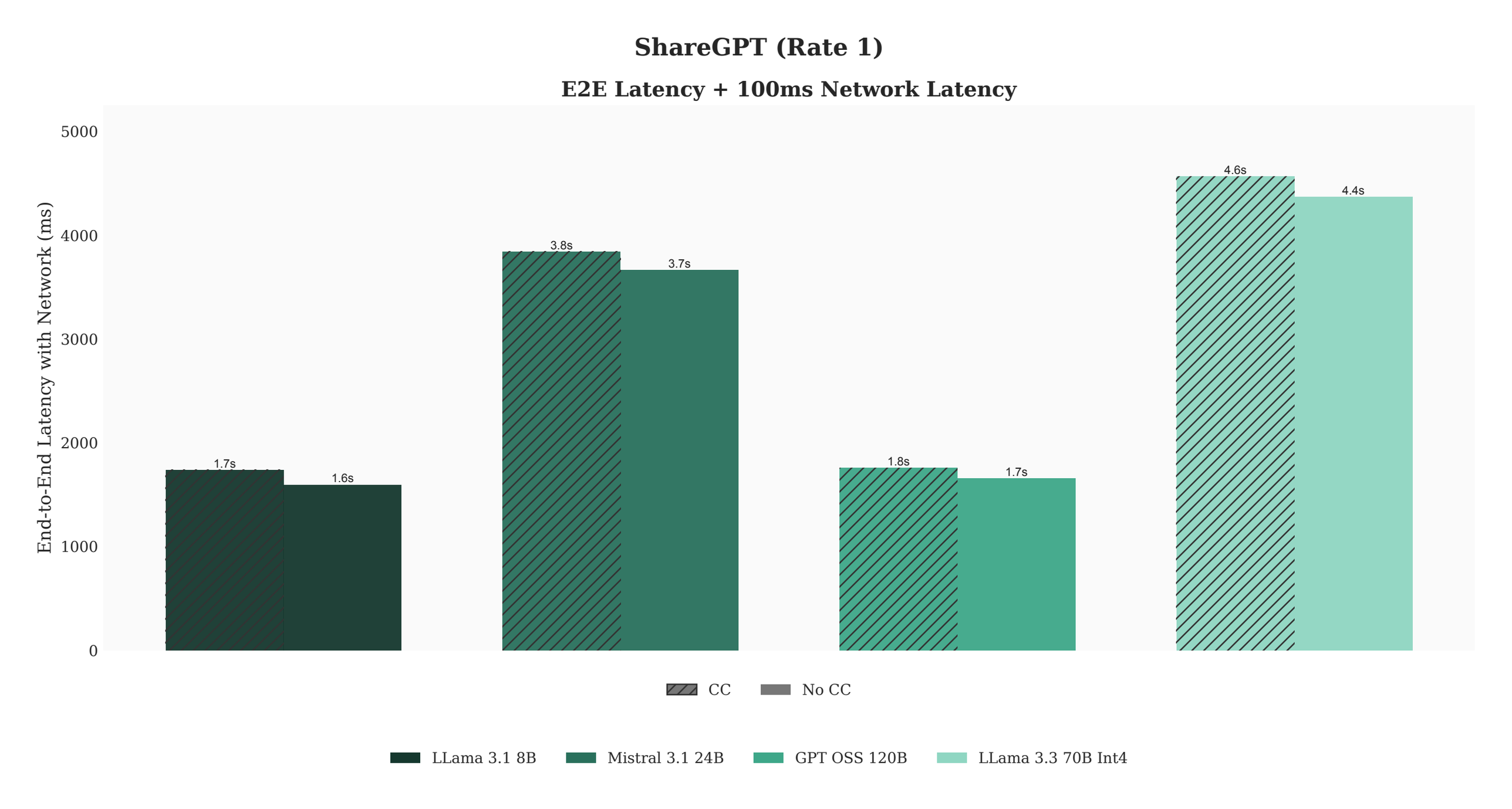

Random (1000 ⇒ 1000) (Rate 100)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
120000
100000
80000
60000
40000
20000
0
25.5s
22.8s
57.9s
54.9s
12.3s
13.0s
109.1s
110.0s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Random (1000 ⇒ 1000) (Rate 50)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
120000
100000
80000
60000
40000
20000
0
21.9s
19.5s
57.1s
54.3s
12.5s
12.2s
108.8s
109.0s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Random (1000 ⇒ 1000) (Rate 1)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
5000
10000
15000
20000
7.9s
7.1s
21.4s
19.9s
2.7s
2.7s
19.5s
18.8s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

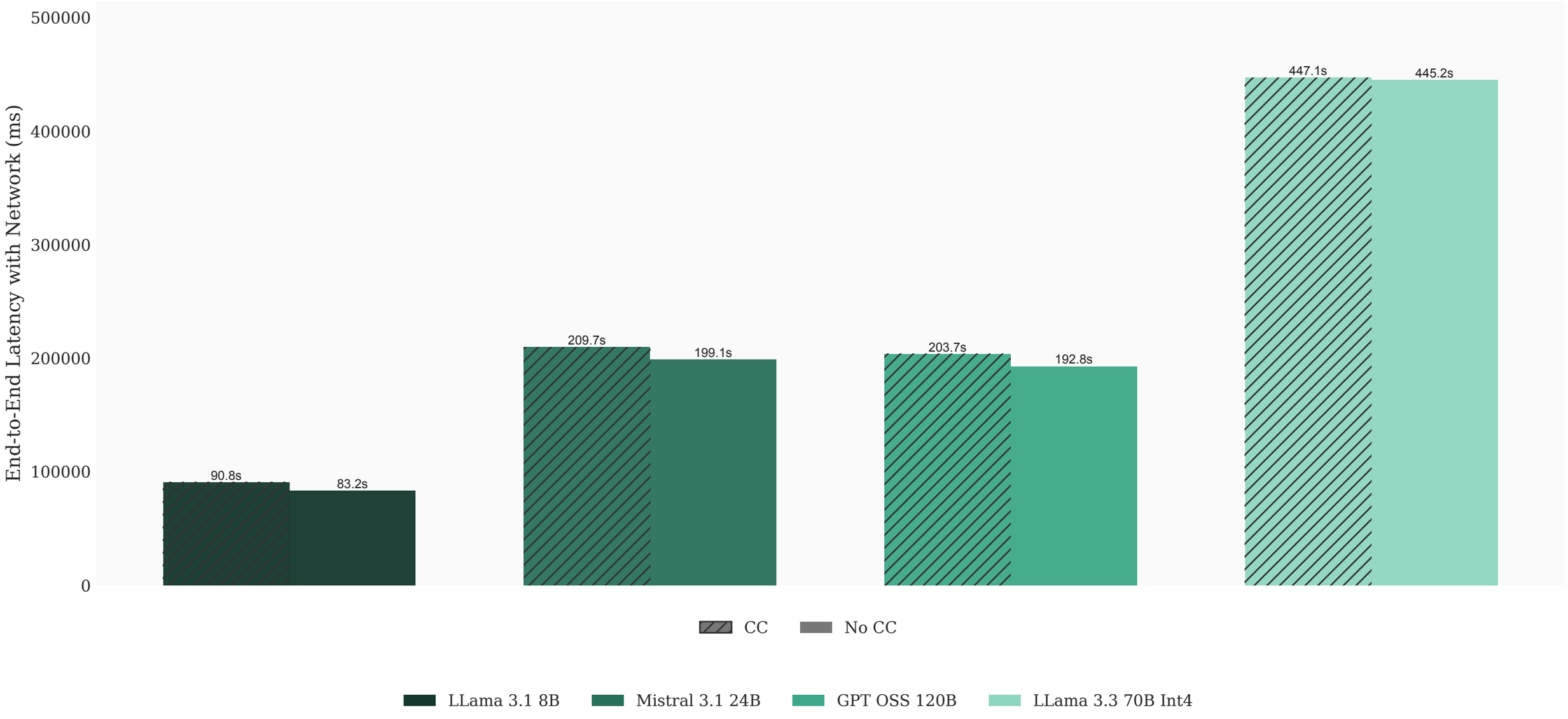

ShareGPT (Rate 100)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
30000
25000
20000
15000
10000
5000
0
2.9s
2.6s
6.6s
6.2s
7.9s
7.5s
29.2s
29.3s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

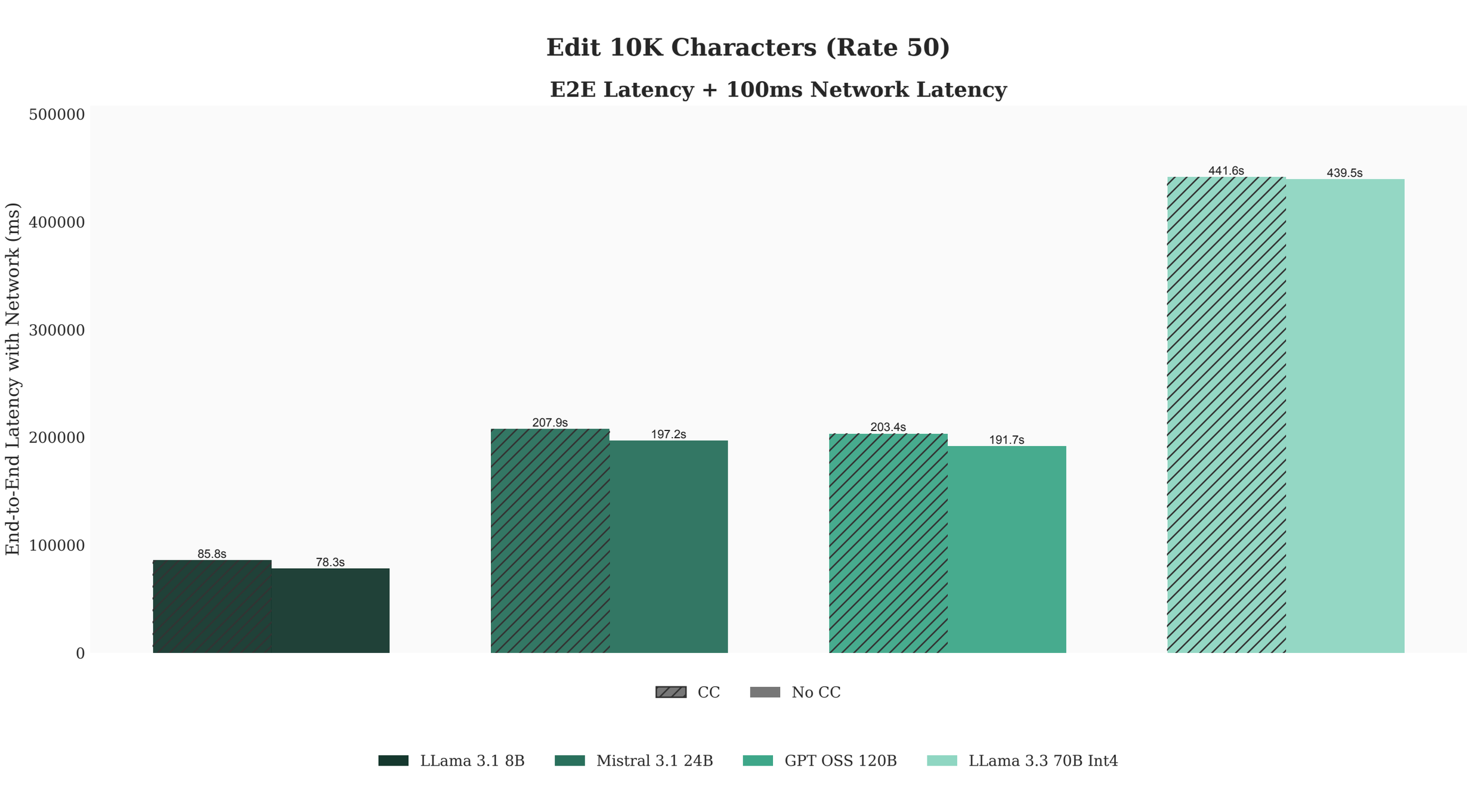

ShareGPT (Rate 50)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
2000
4000
6000
8000
10000
12000
2.4s
2.1s
4.7s
4.3s
7.2s
6.2s
11.9s
11.8s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


ShareGPT (Rate 1)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
5000
4000
3000
2000
1000
0
1.7s
1.6s
3.8s
3.7s
1.8s
1.7s
4.6s
4.4s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

Edit 10K Characters (Rate 100)

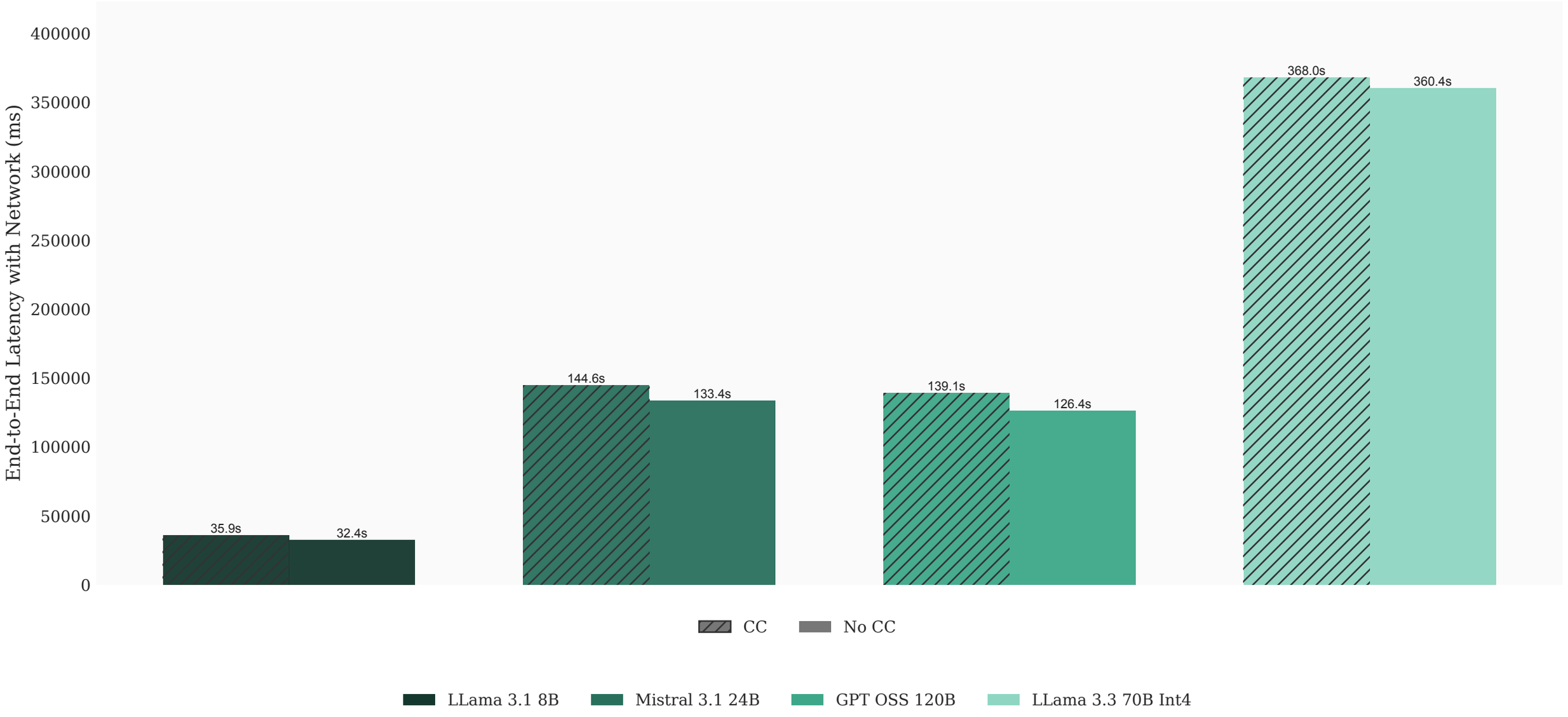
E2E Latency + 100ms Network Latency

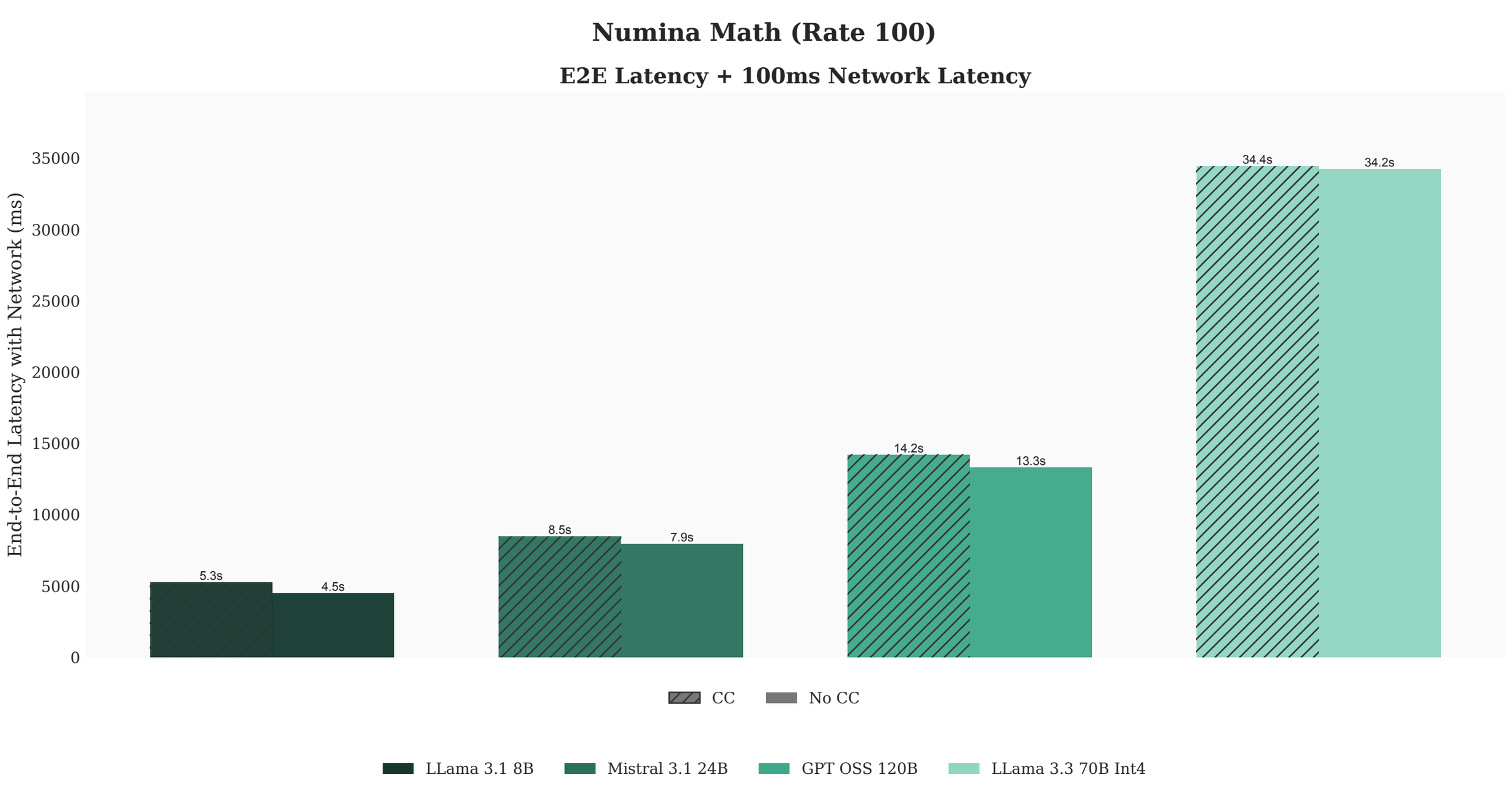


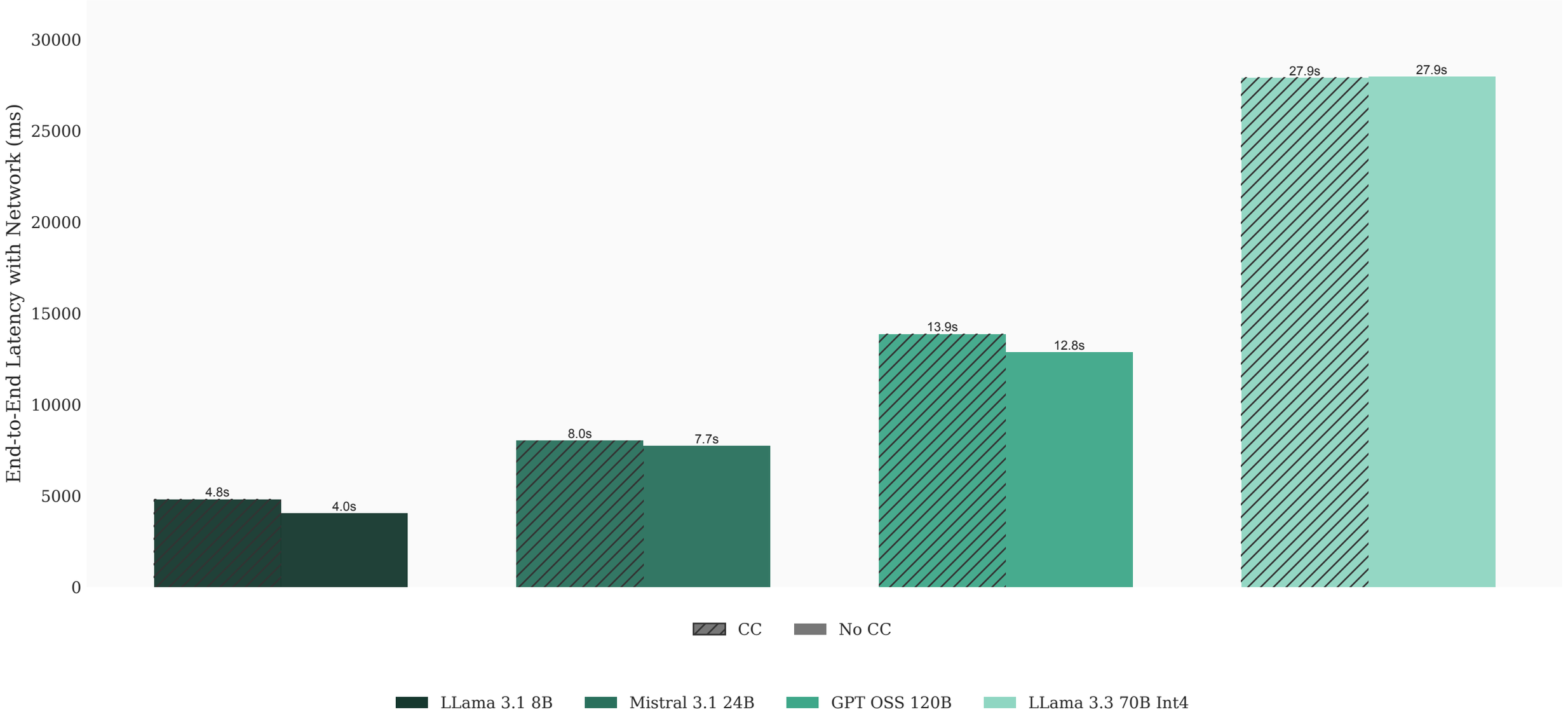

Edit 10K Characters (Rate 50)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
500000
400000
300000
200000
100000
0
85.8s
78.3s
207.9s
197.2s
203.4s
191.7s
441.6s
439.5s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

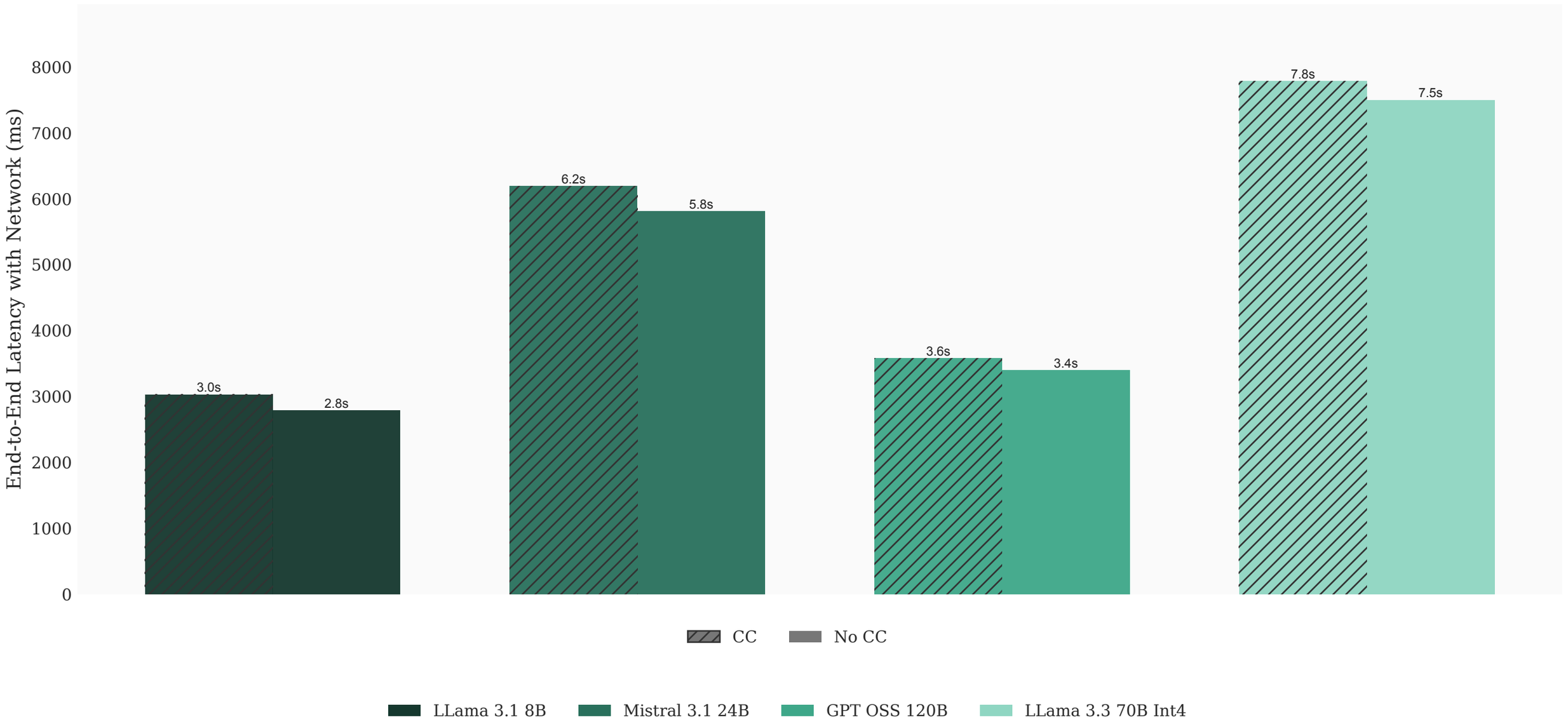
Edit 10K Characters (Rate 1)

E2E Latency + 100ms Network Latency




Numina Math (Rate 100)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
35000
30000
25000
20000
15000
10000
5000
0
5.3s
4.5s
8.5s
7.9s
14.2s
13.3s
34.4s
34.2s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (Rate 50)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
30000
25000
20000
15000
10000
5000
0
4.8s
4.0s
8.0s
7.7s
13.9s
12.8s
27.9s
27.9s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (Rate 1)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
8000
7000
6000
5000
4000
3000
2000
1000
0
3.0s
2.8s
6.2s
5.8s
3.6s
3.4s
7.8s
7.5s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4