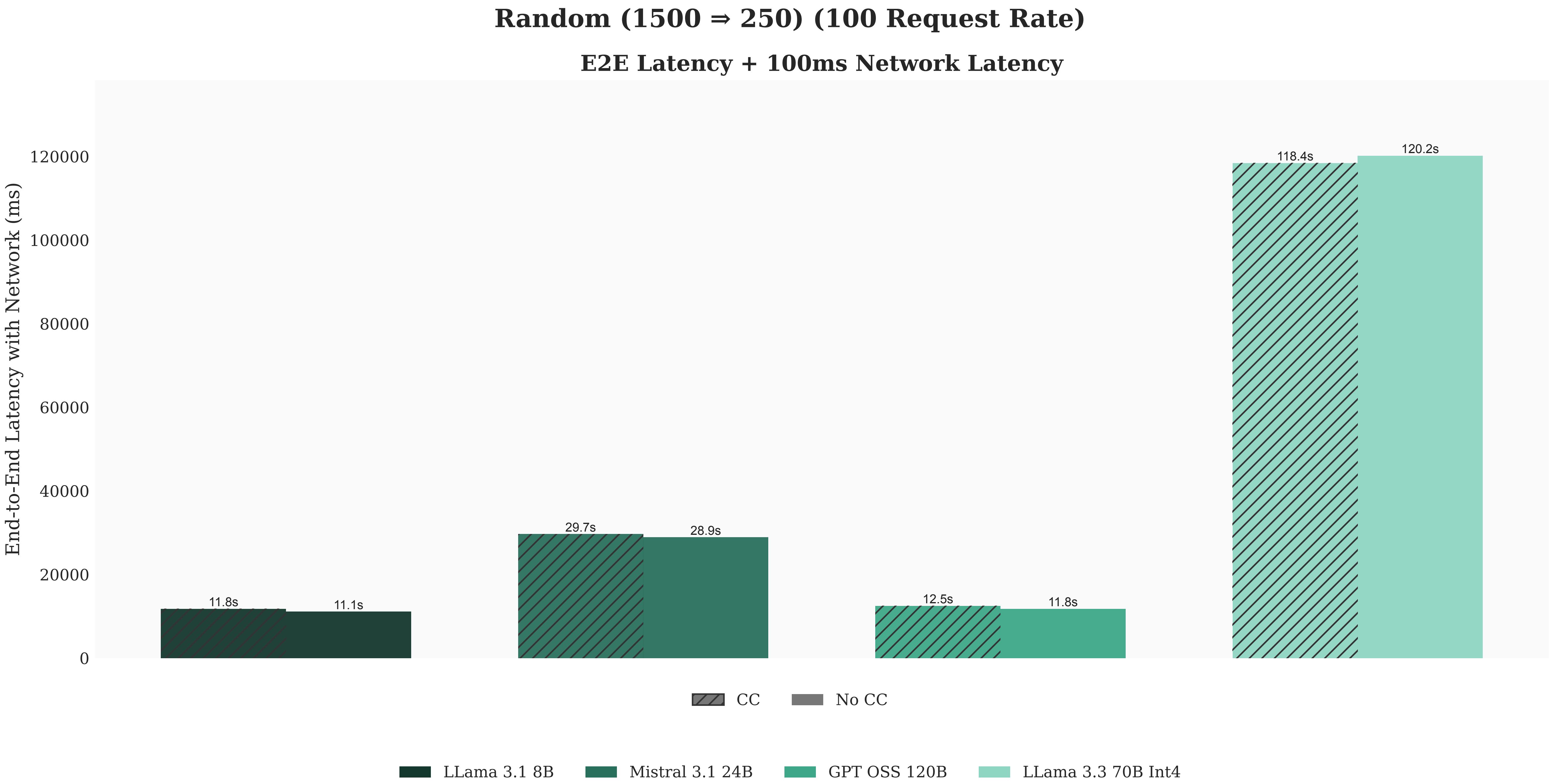
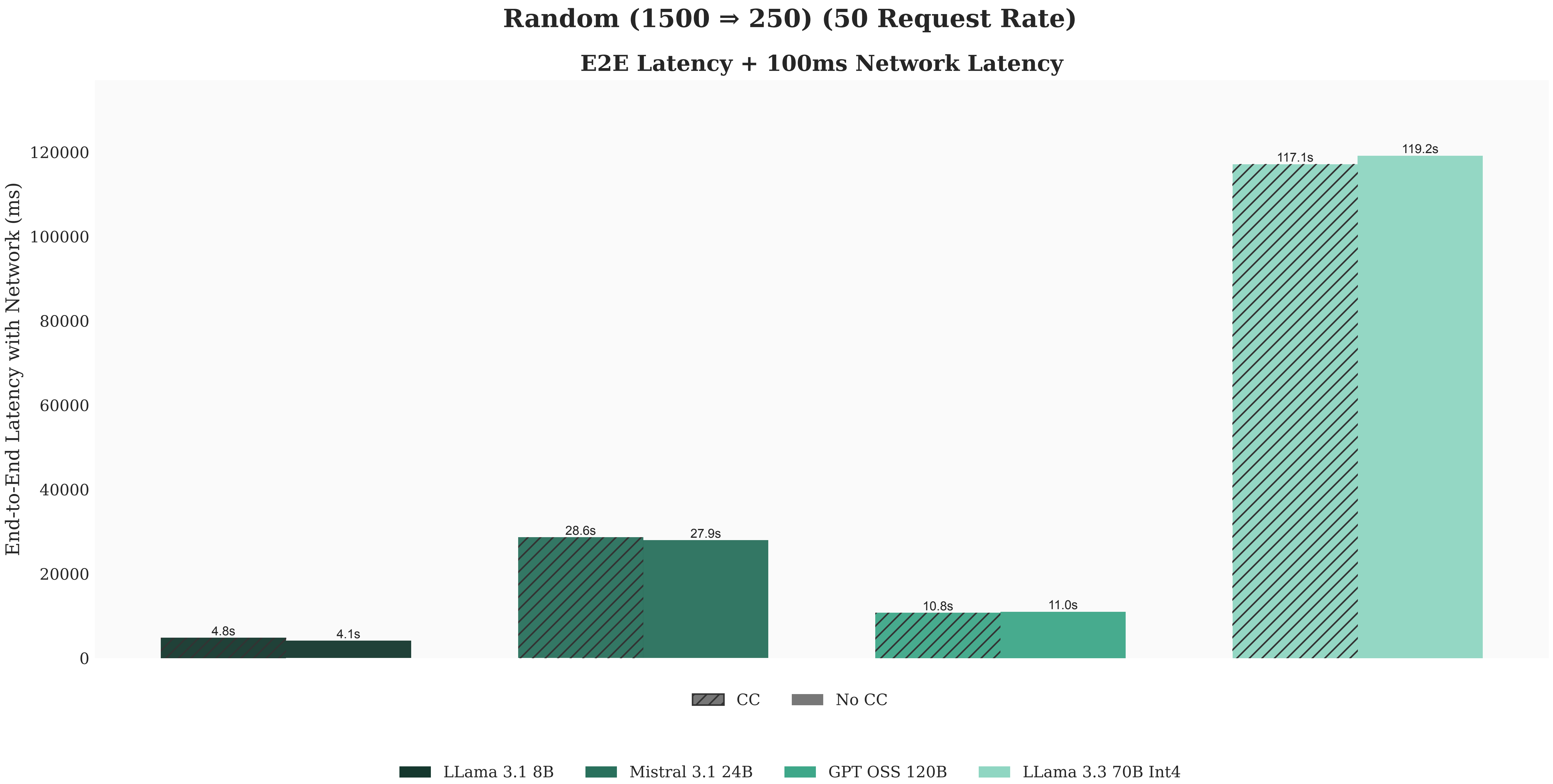
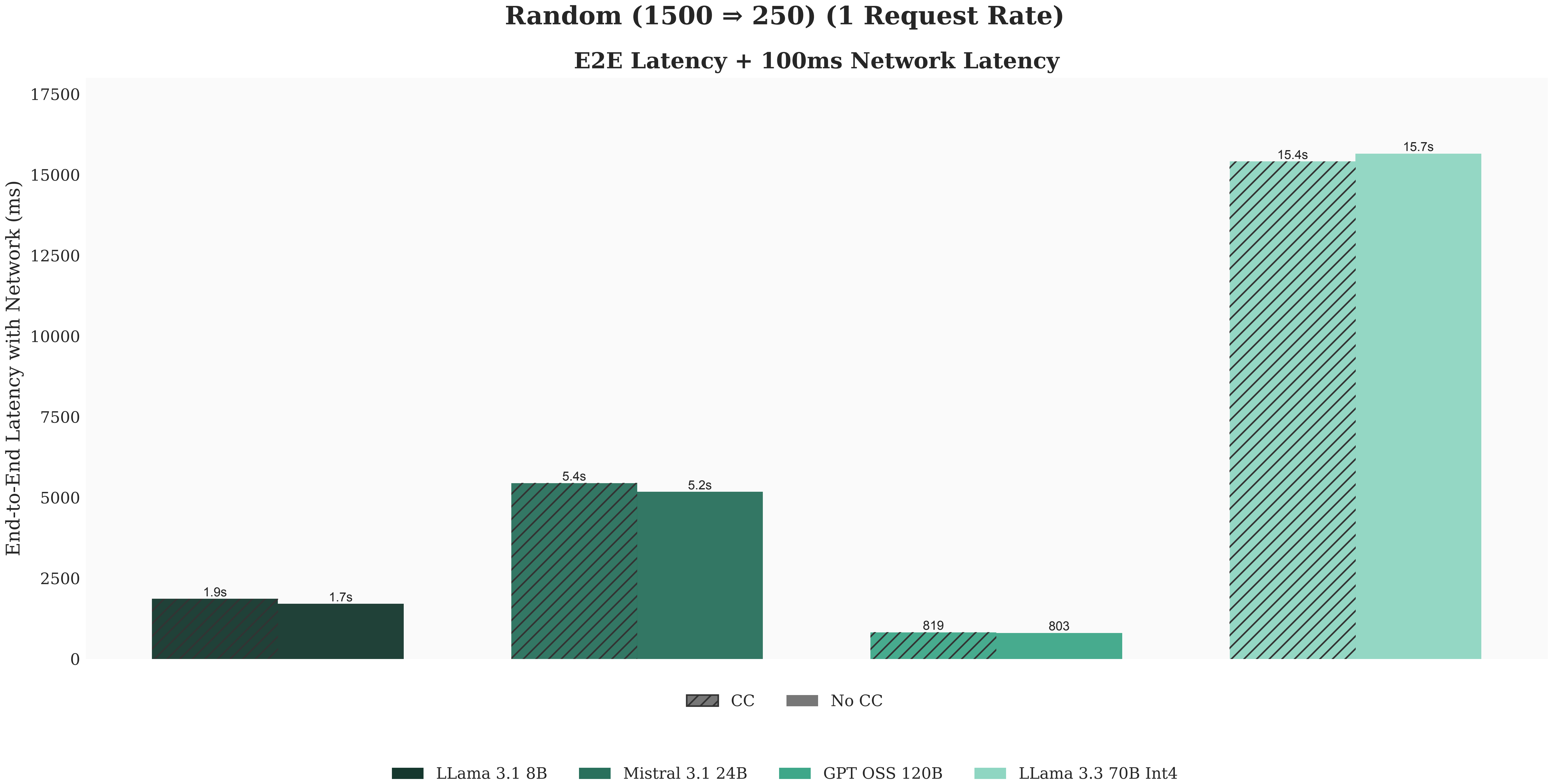
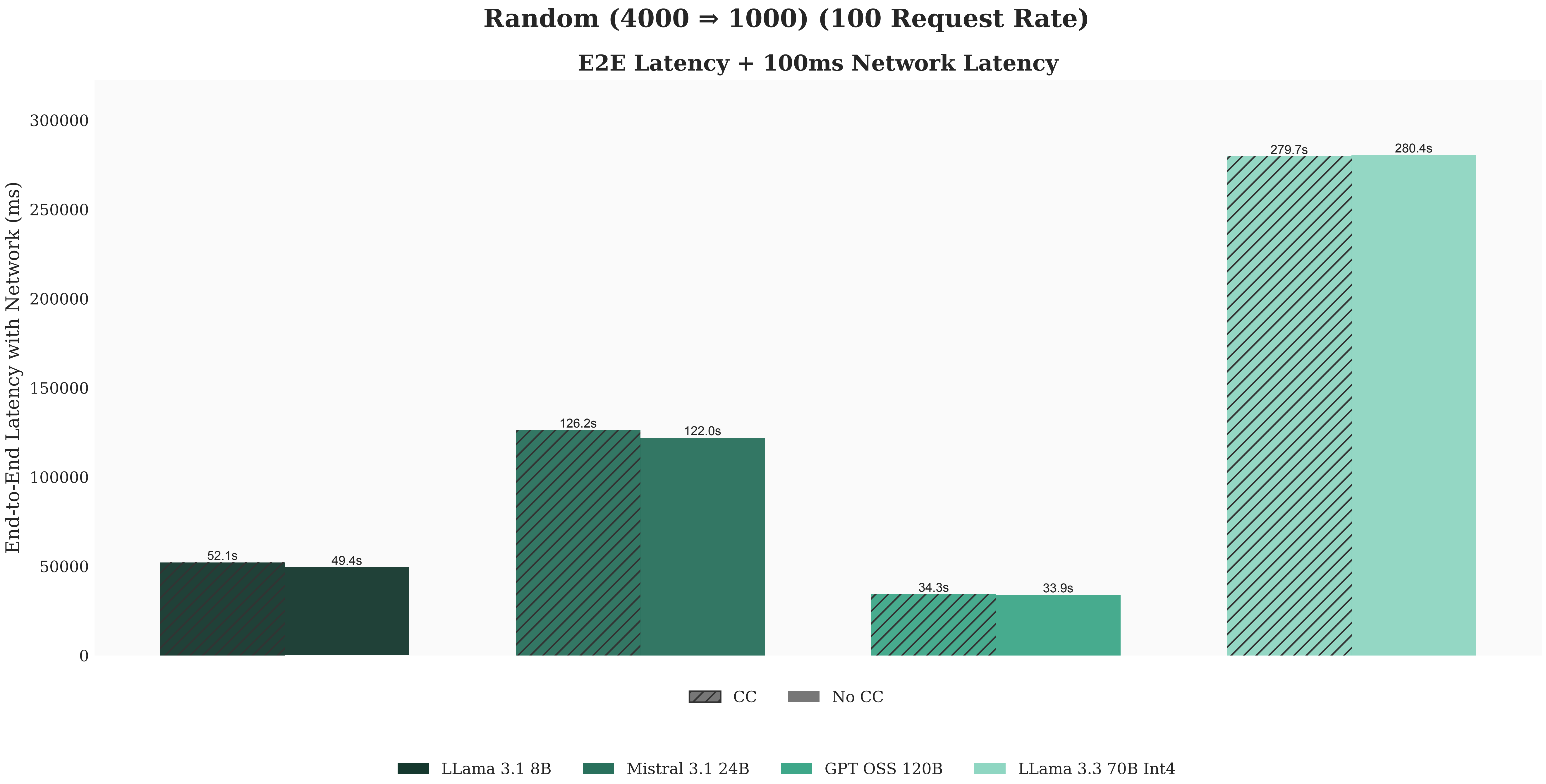

Random (1500 ⇒ 250) (100 Request Rate)

E2E Latency + 100ms Network Latency

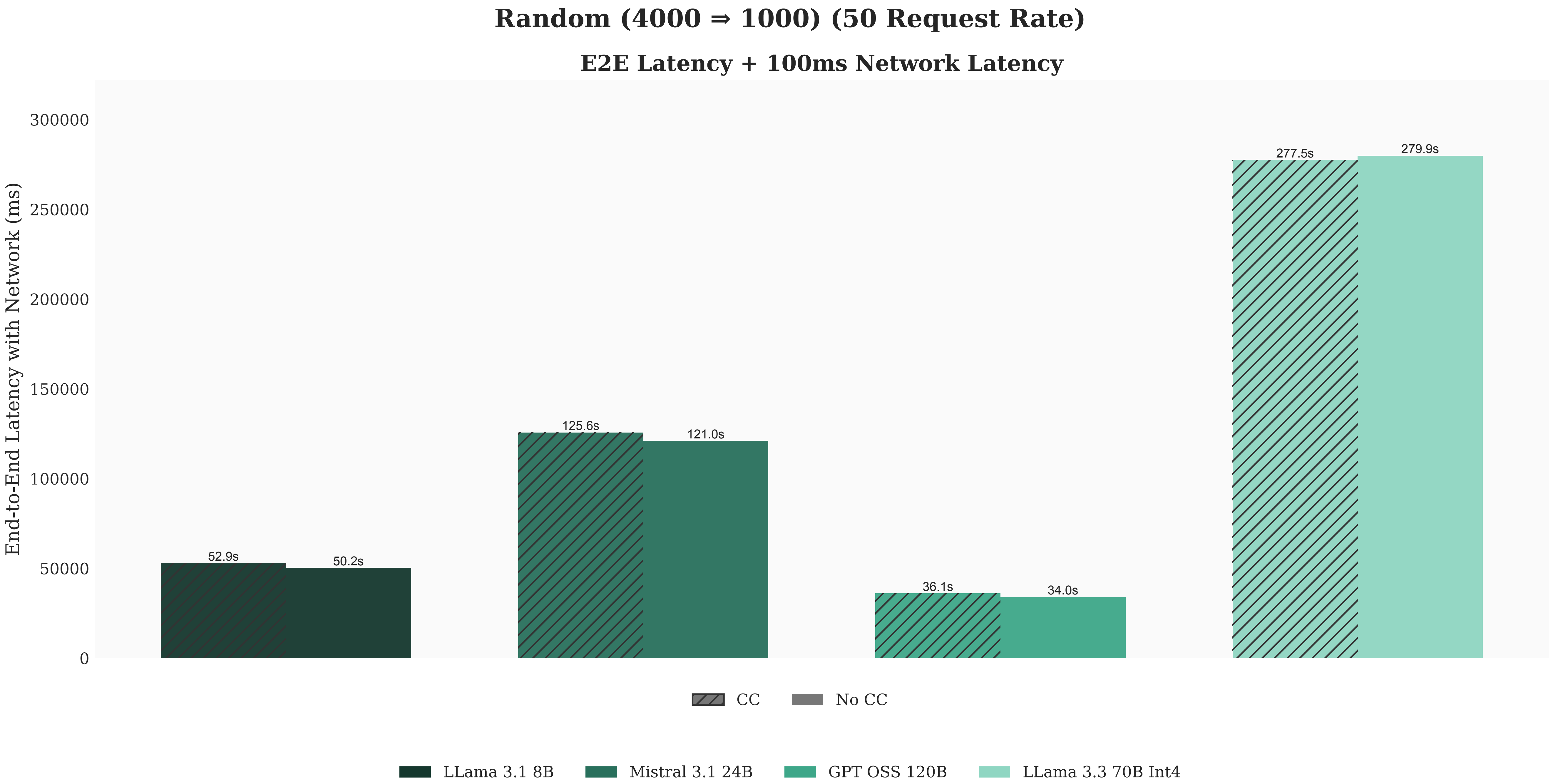

Random (1500 ⇒ 250) (50 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
20000
40000
60000
80000
100000
120000
4.8s
4.1s
28.6s
27.9s
10.8s
11.0s
117.1s
119.2s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

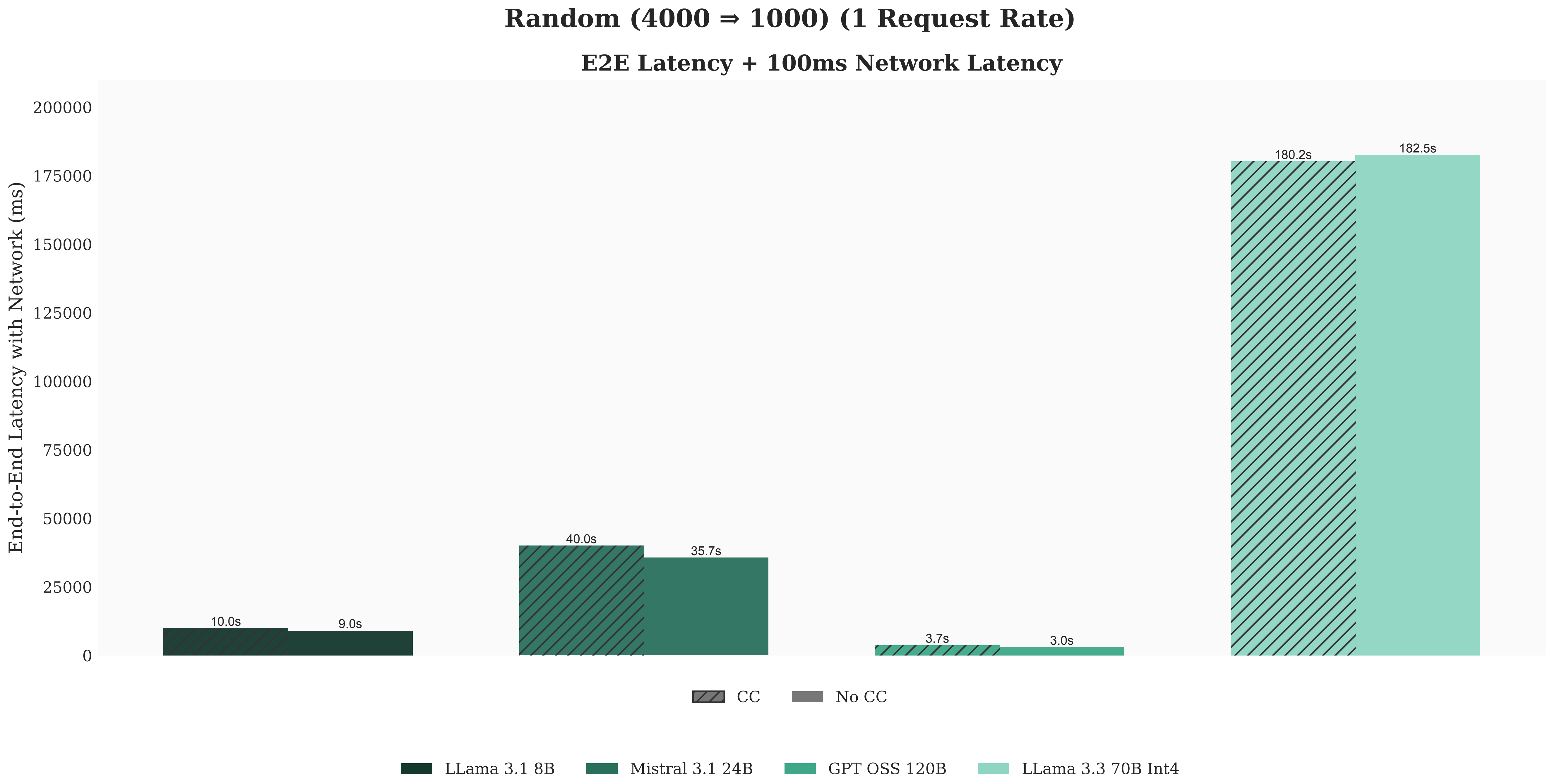

Random (1500 ⇒ 250) (1 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
17500
15000
12500
10000
7500
5000
2500
0
1.9s
1.7s
5.4s
5.2s
819
803
15.4s
15.7s
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

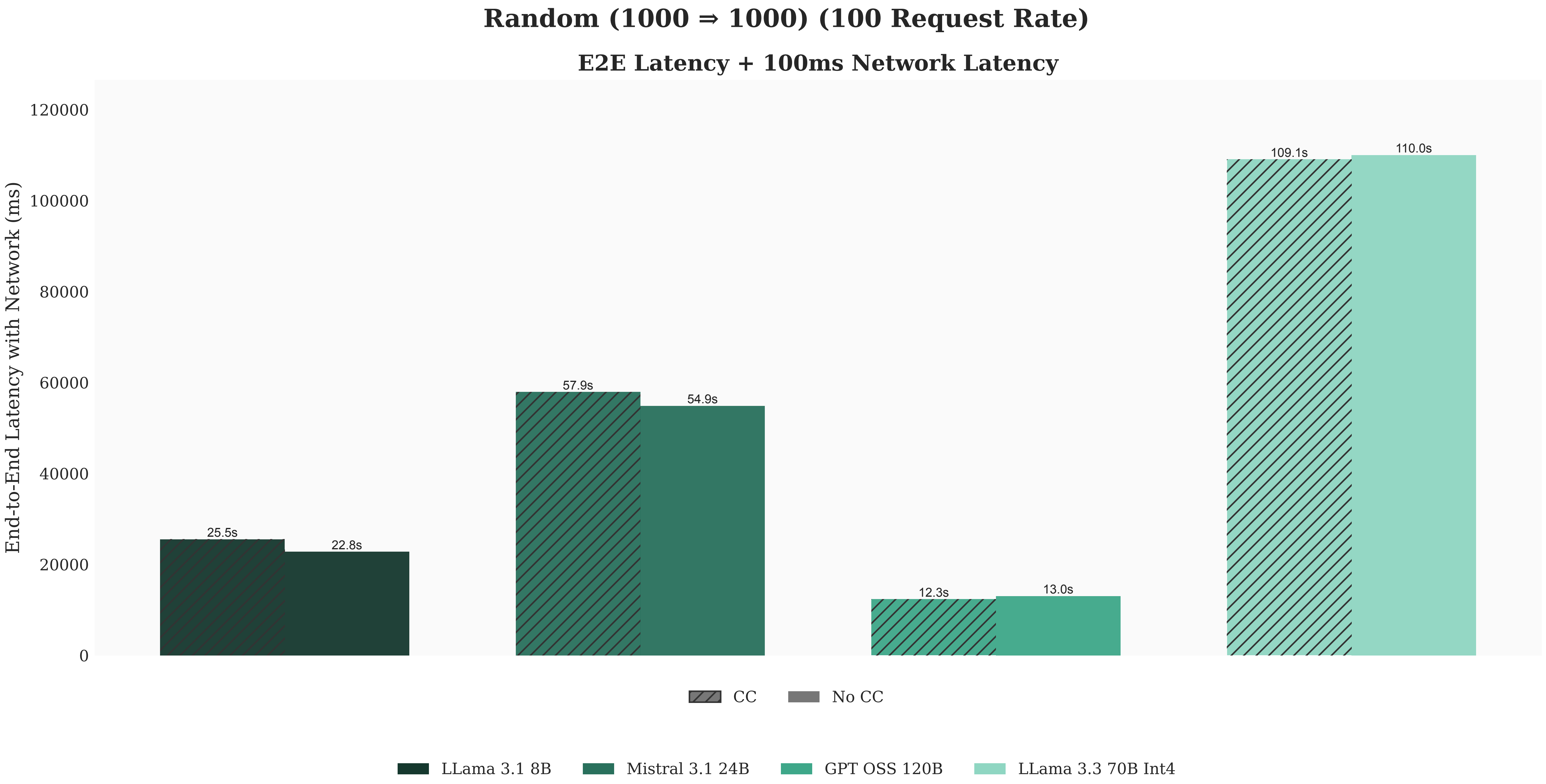


Random (4000 ⇒ 1000) (100 Request Rate)

E2E Latency + 100ms Network Latency

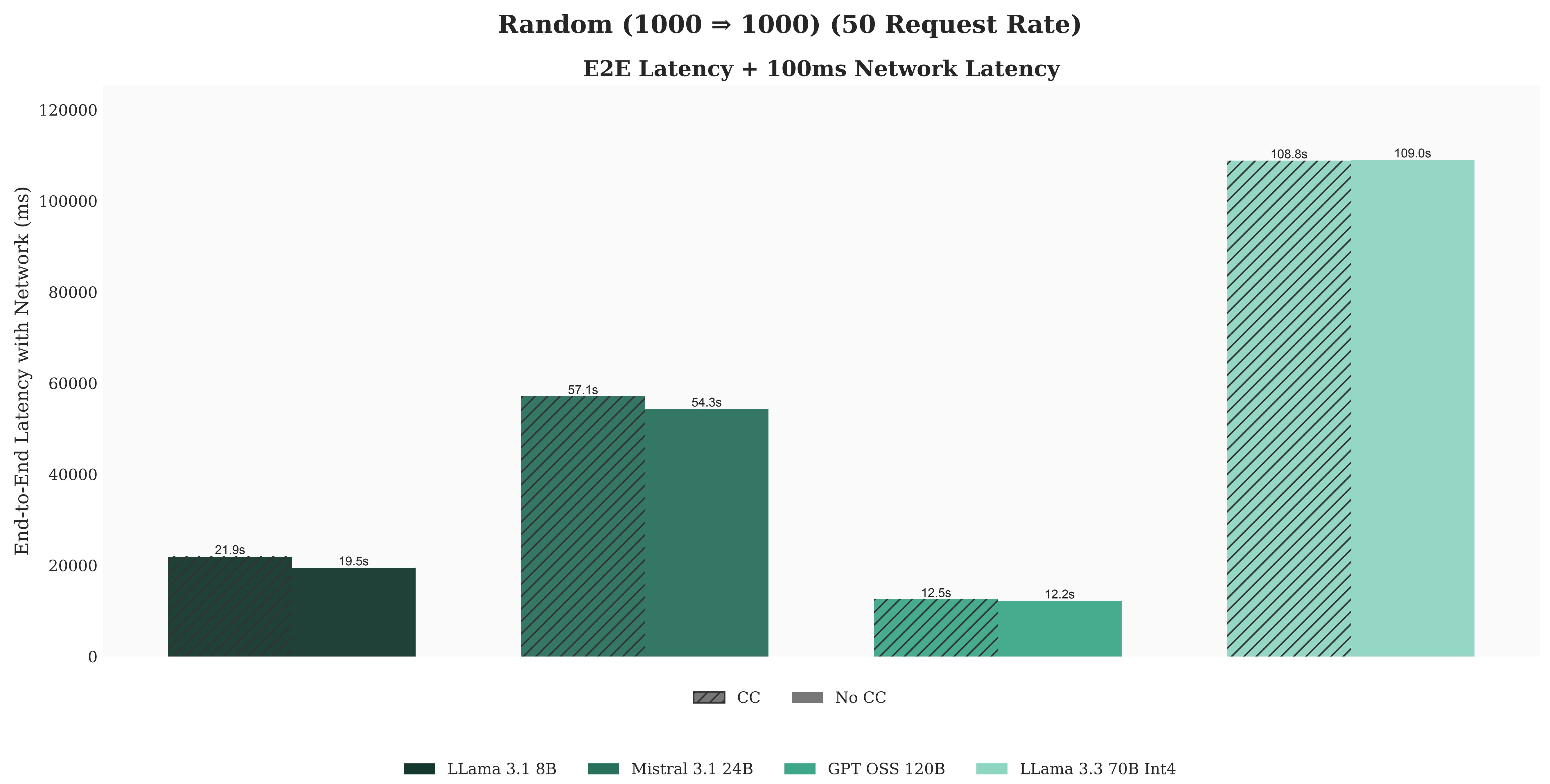

Random (4000 ⇒ 1000) (50 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
300000
250000
200000
150000
100000
50000
0
52.9s
50.2s
125.6s
121.0s
36.1s
34.0s
277.5s
279.9s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Random (4000 ⇒ 1000) (1 Request Rate)
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



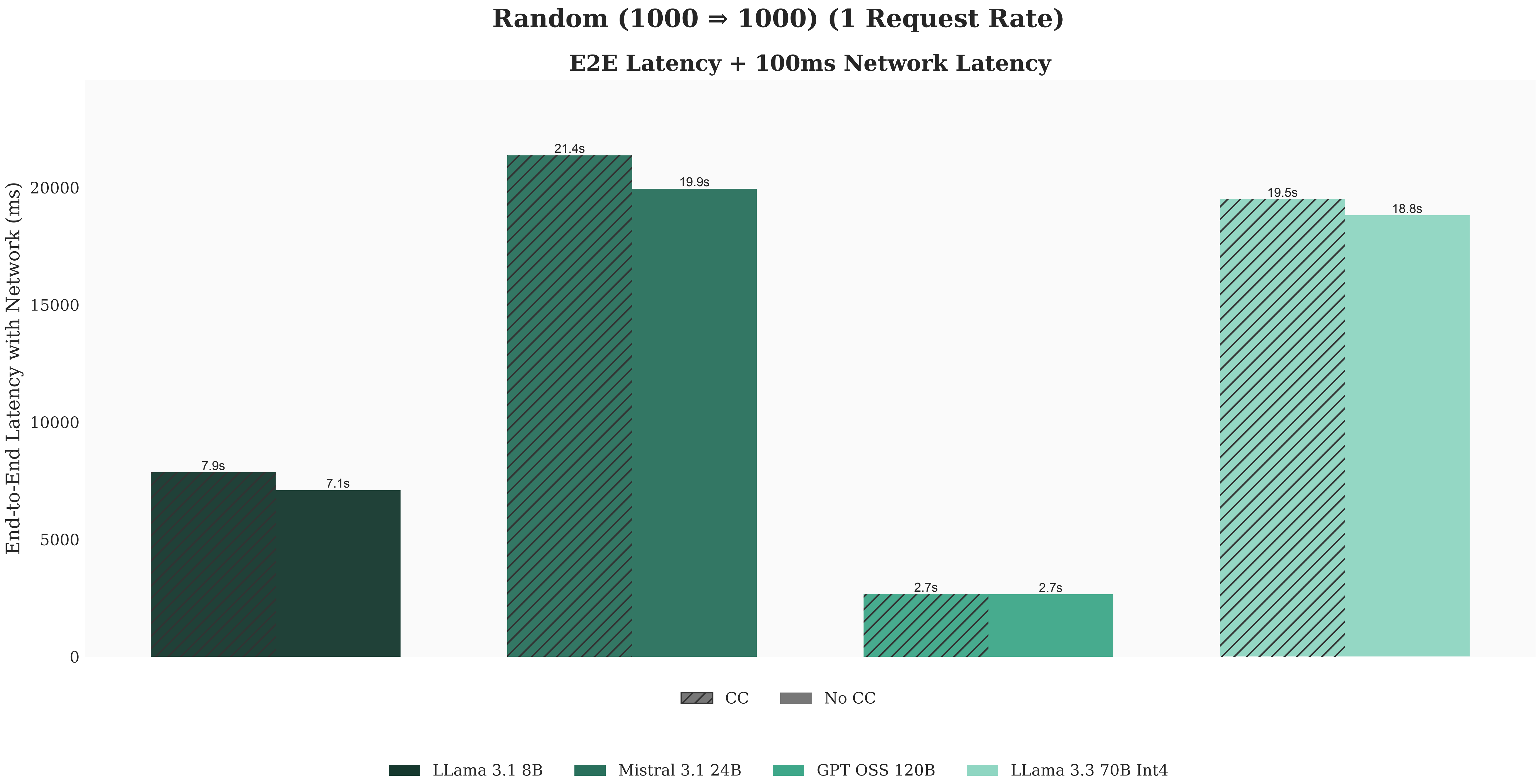
Random (1000 ⇒ 1000) (100 Request Rate)

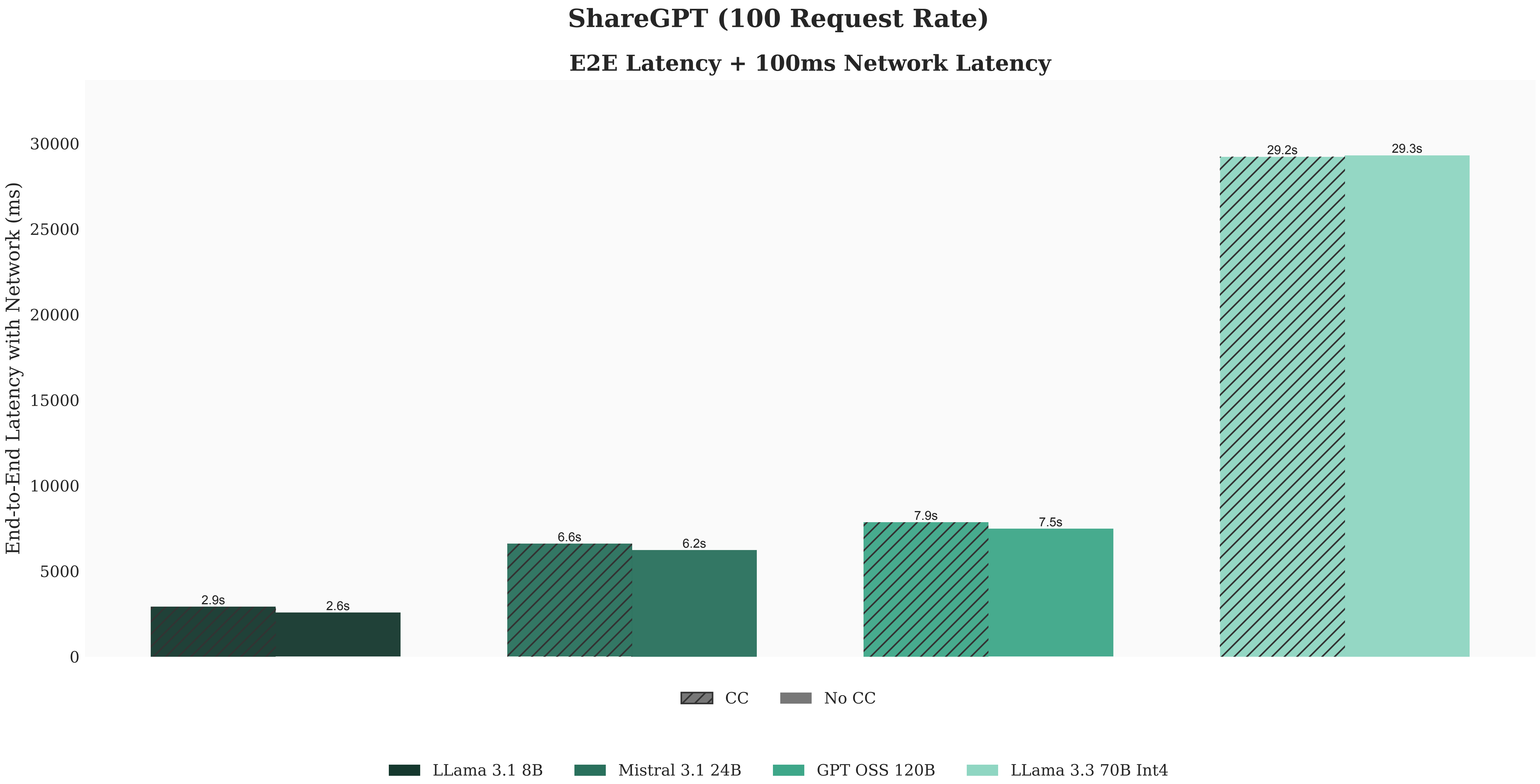
E2E Latency + 100ms Network Latency

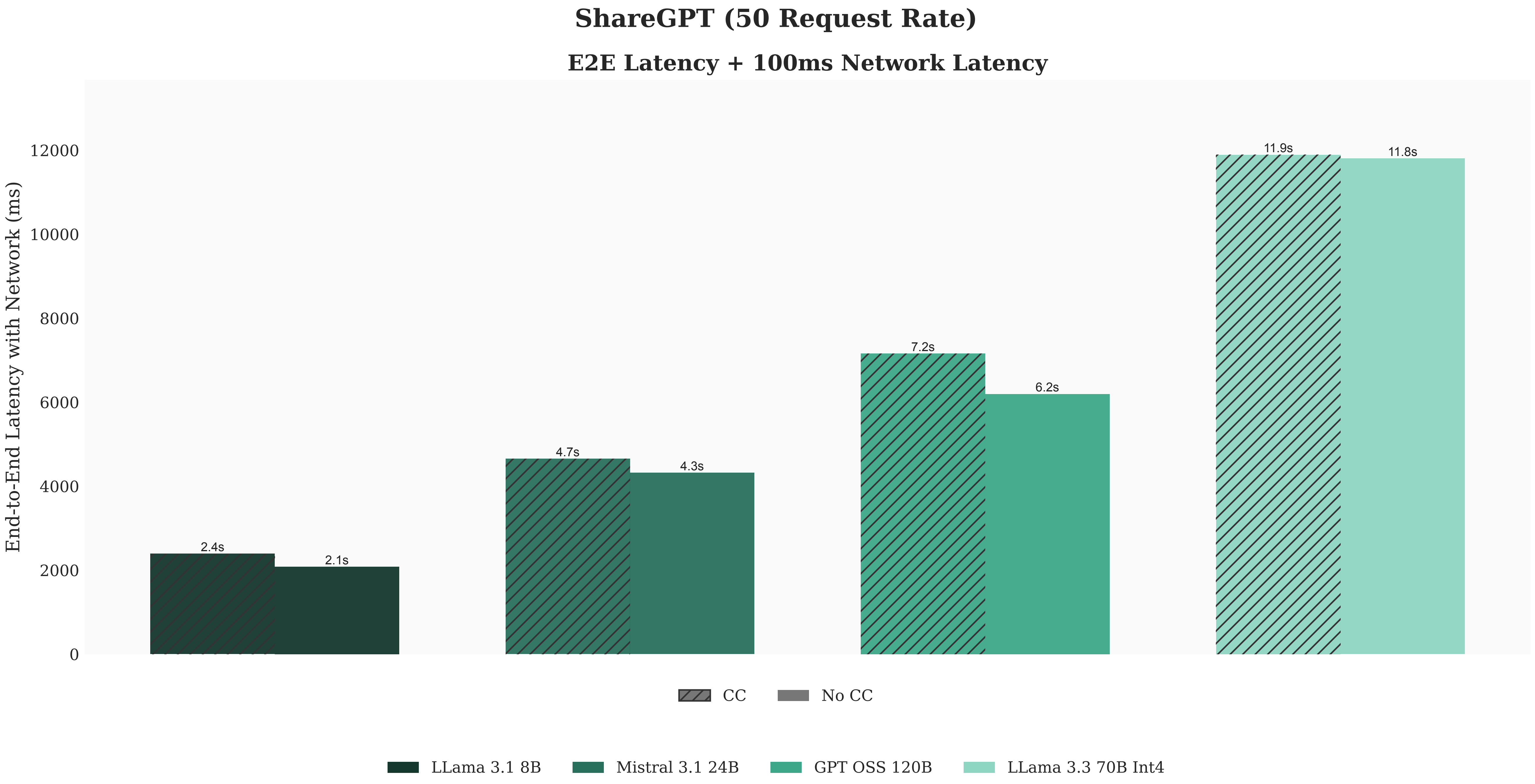


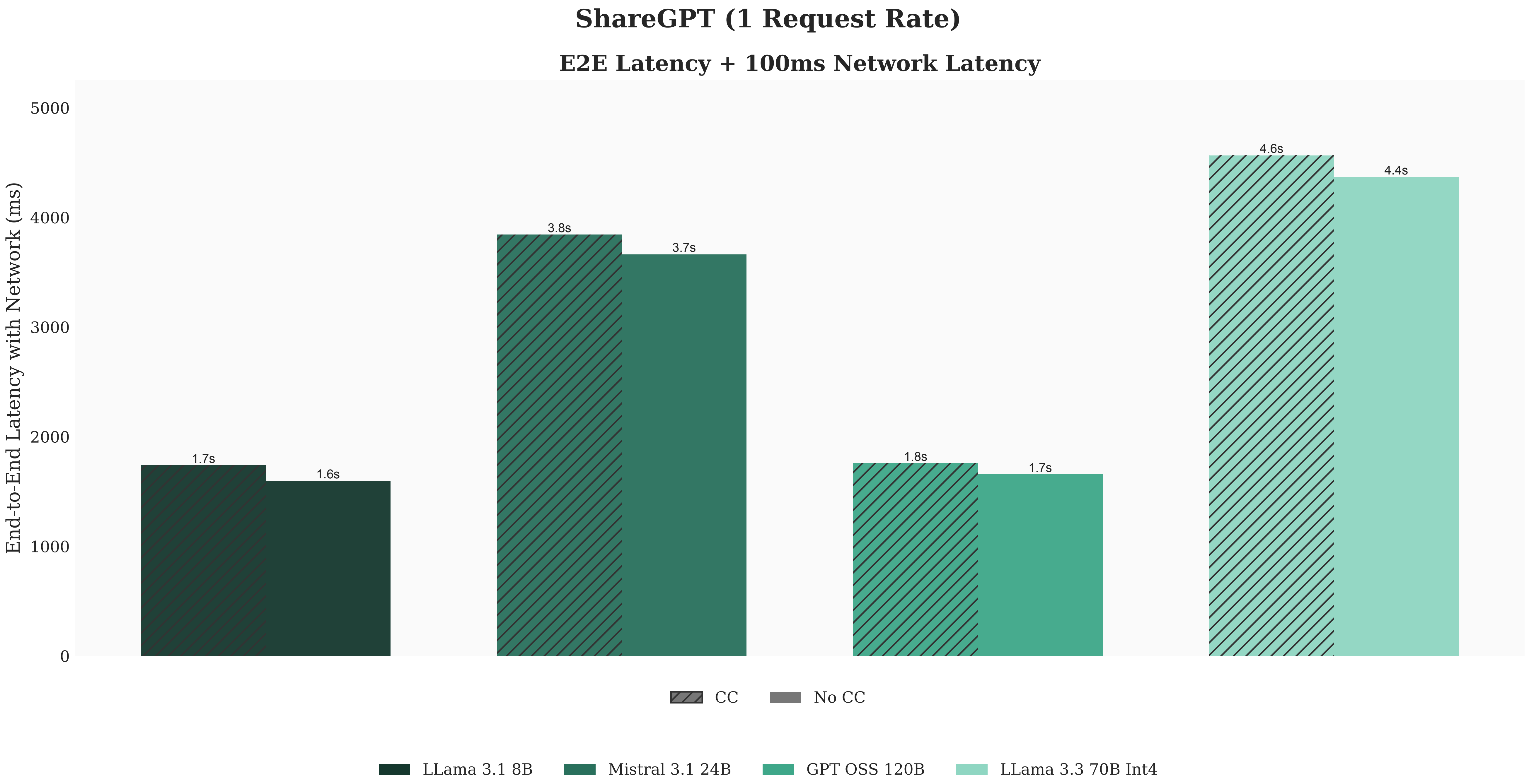
Random (1000 ⇒ 1000) (50 Request Rate)

E2E Latency + 100ms Network Latency


Random (1000 ⇒ 1000) (1 Request Rate)
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

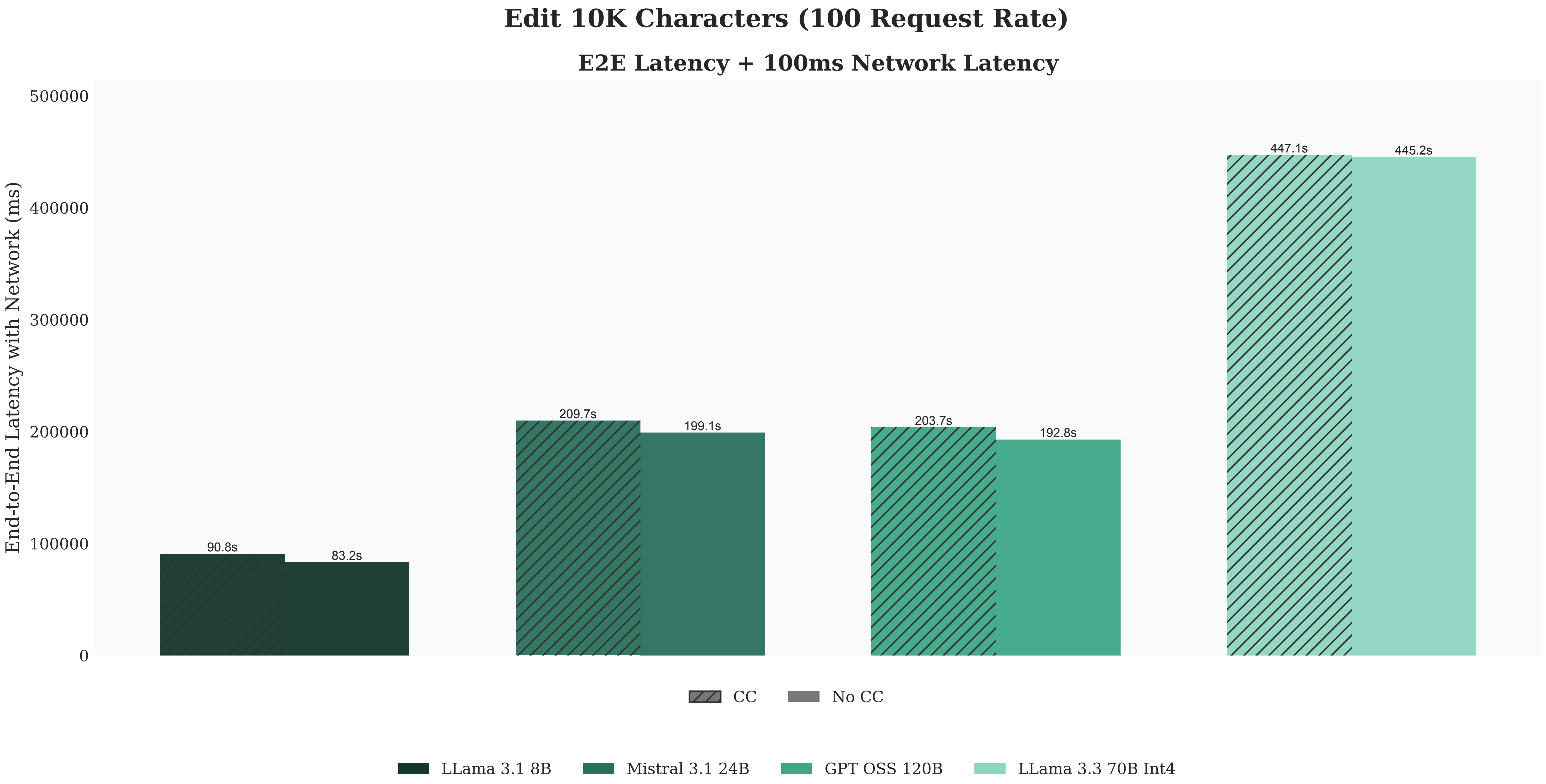

ShareGPT (100 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
5000
10000
15000
20000
25000
30000
2.9s
2.6s
6.6s
6.2s
7.9s
7.5s
29.2s
29.3s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

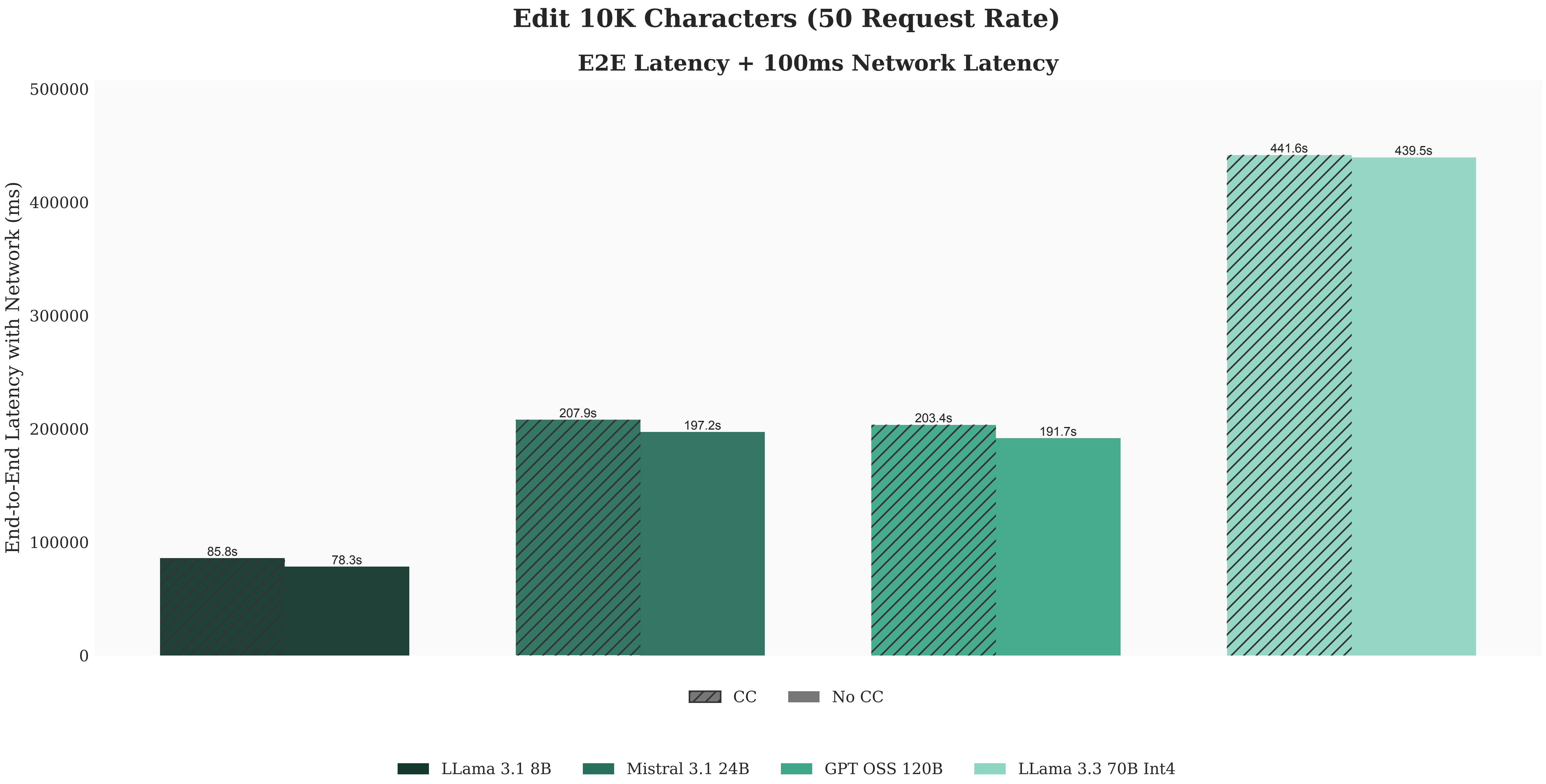

ShareGPT (50 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
2000
4000
6000
8000
10000
12000
2.4s
2.1s
4.7s
4.3s
7.2s
6.2s
11.9s
11.8s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4



ShareGPT (1 Request Rate)

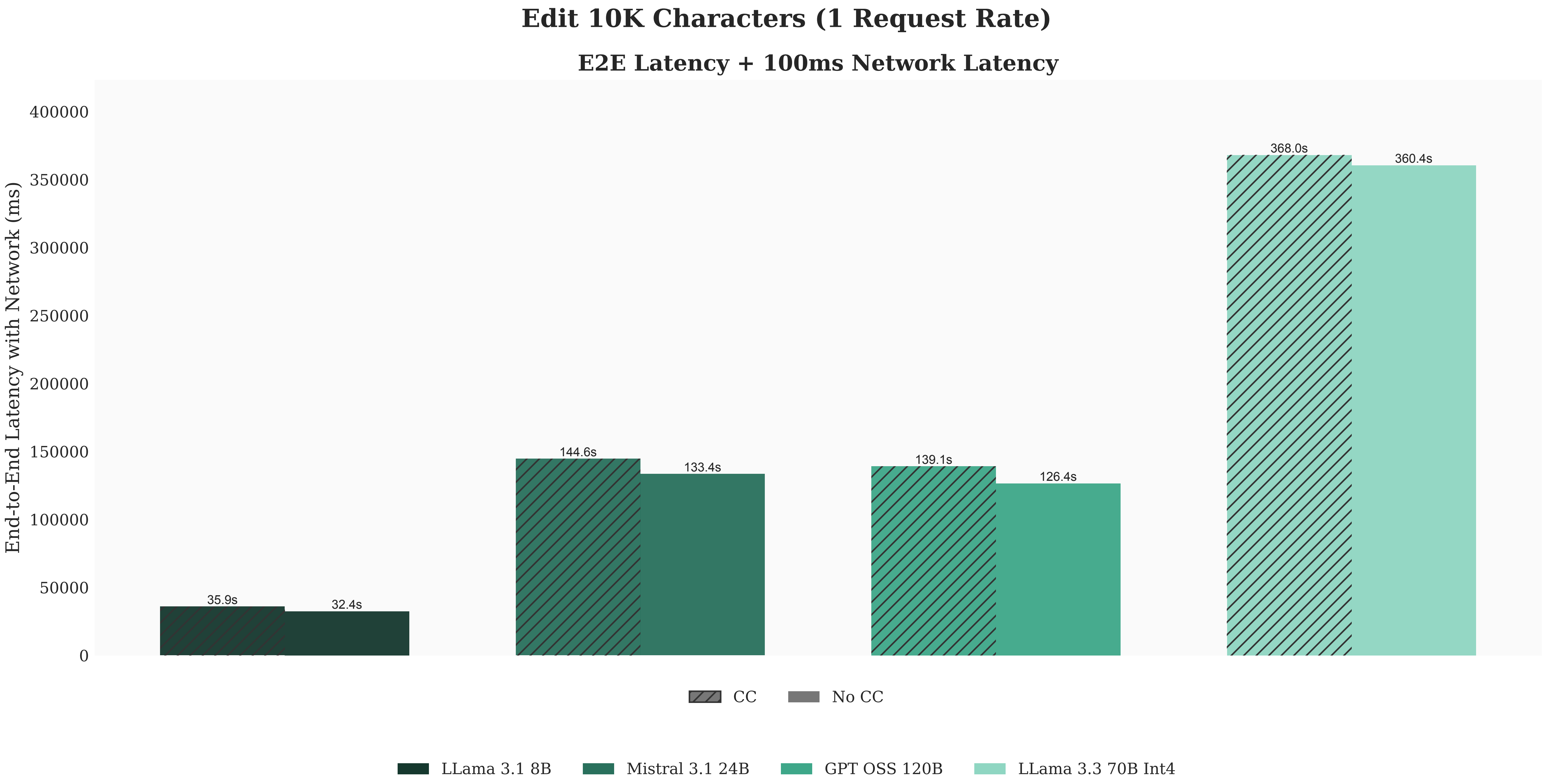
E2E Latency + 100ms Network Latency


Edit 10K Characters (100 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
500000
400000
300000
200000
100000
0
90.8s
83.2s
209.7s
199.1s
203.7s
192.8s
447.1s
445.2s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4



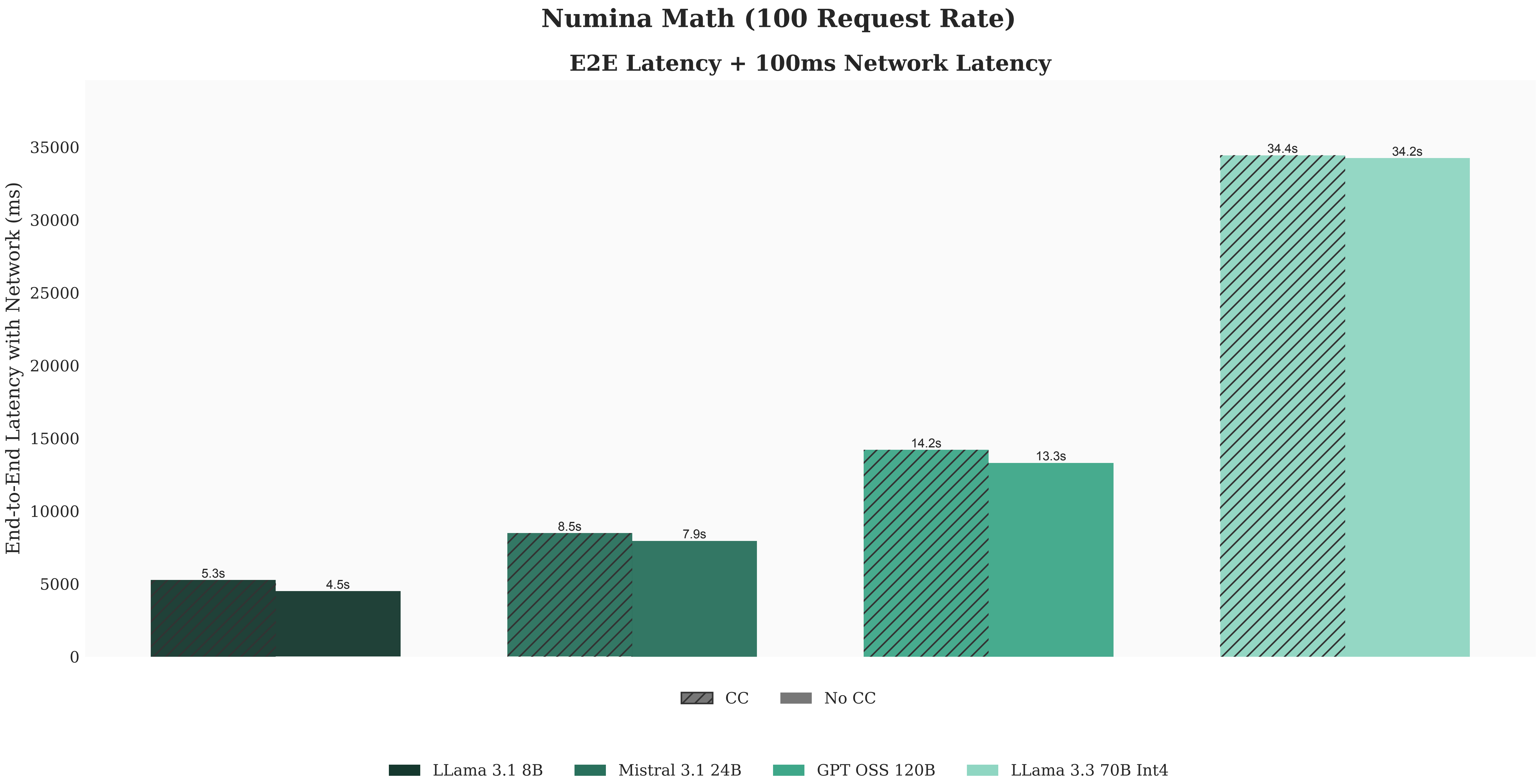
Edit 10K Characters (50 Request Rate)

E2E Latency + 100ms Network Latency



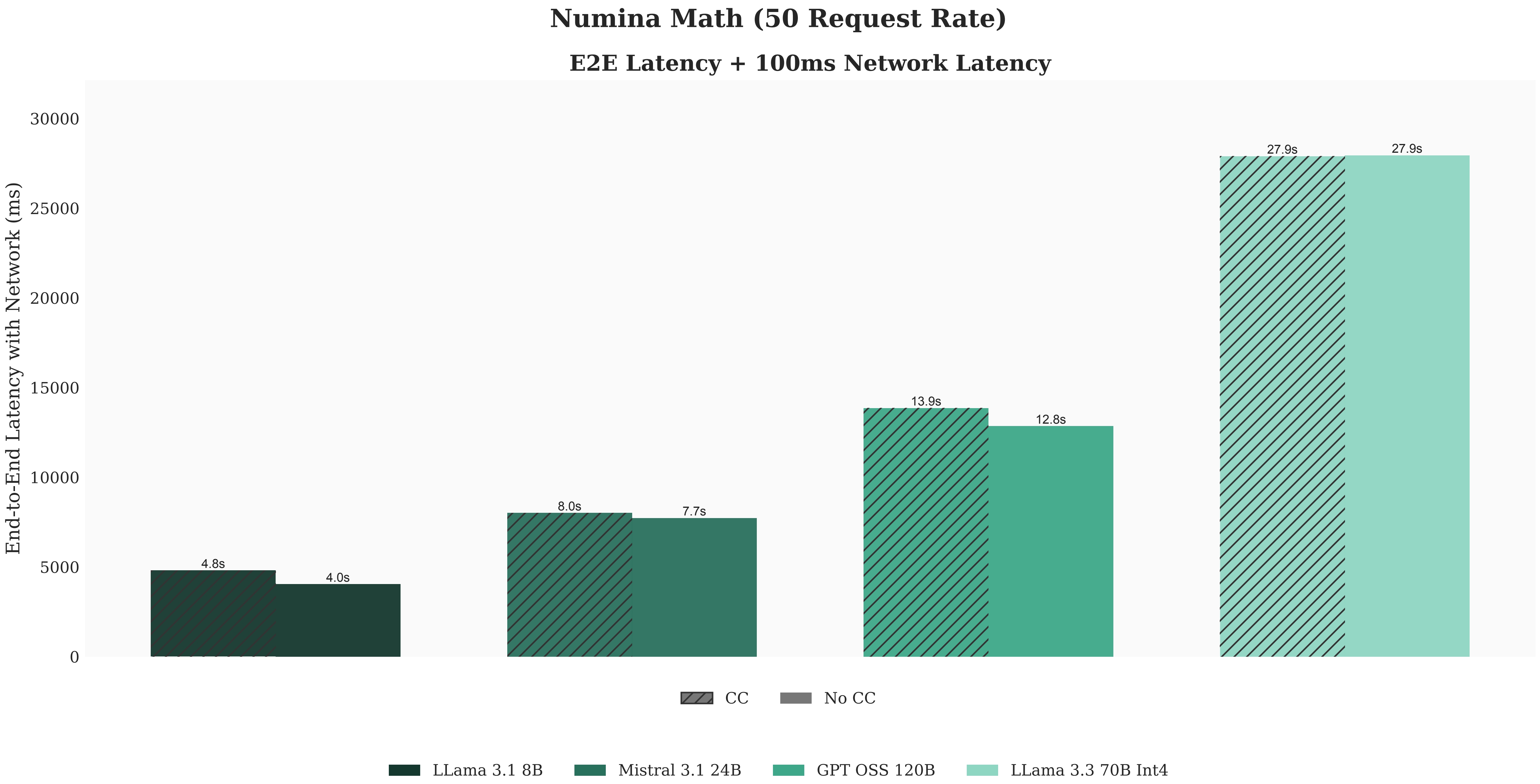
Edit 10K Characters (1 Request Rate)

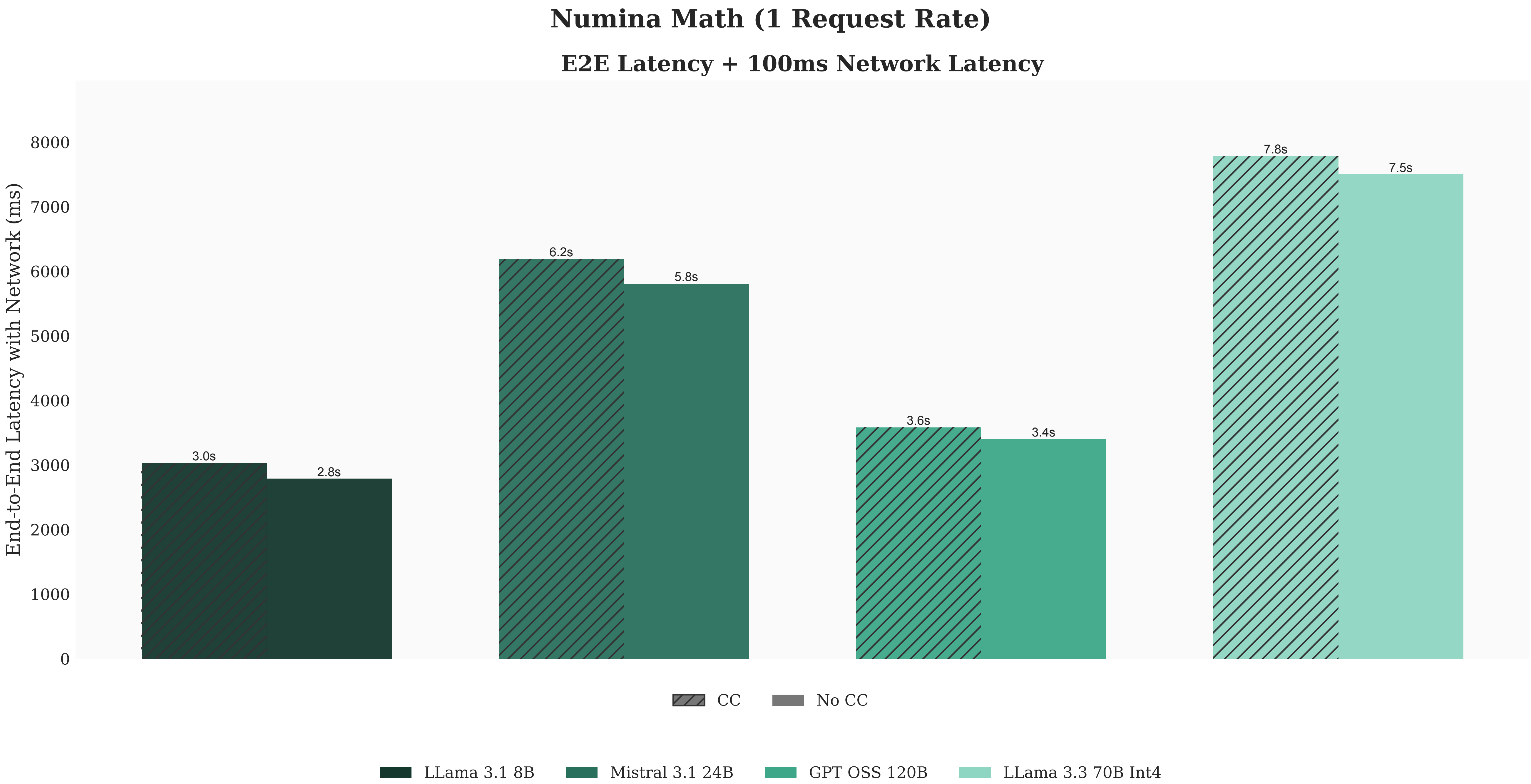
E2E Latency + 100ms Network Latency



Numina Math (100 Request Rate)

E2E Latency + 100ms Network Latency


Numina Math (50 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
5000
10000
15000
20000
25000
30000
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 (1 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
1000
2000
3000
4000
5000
6000
7000
8000
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