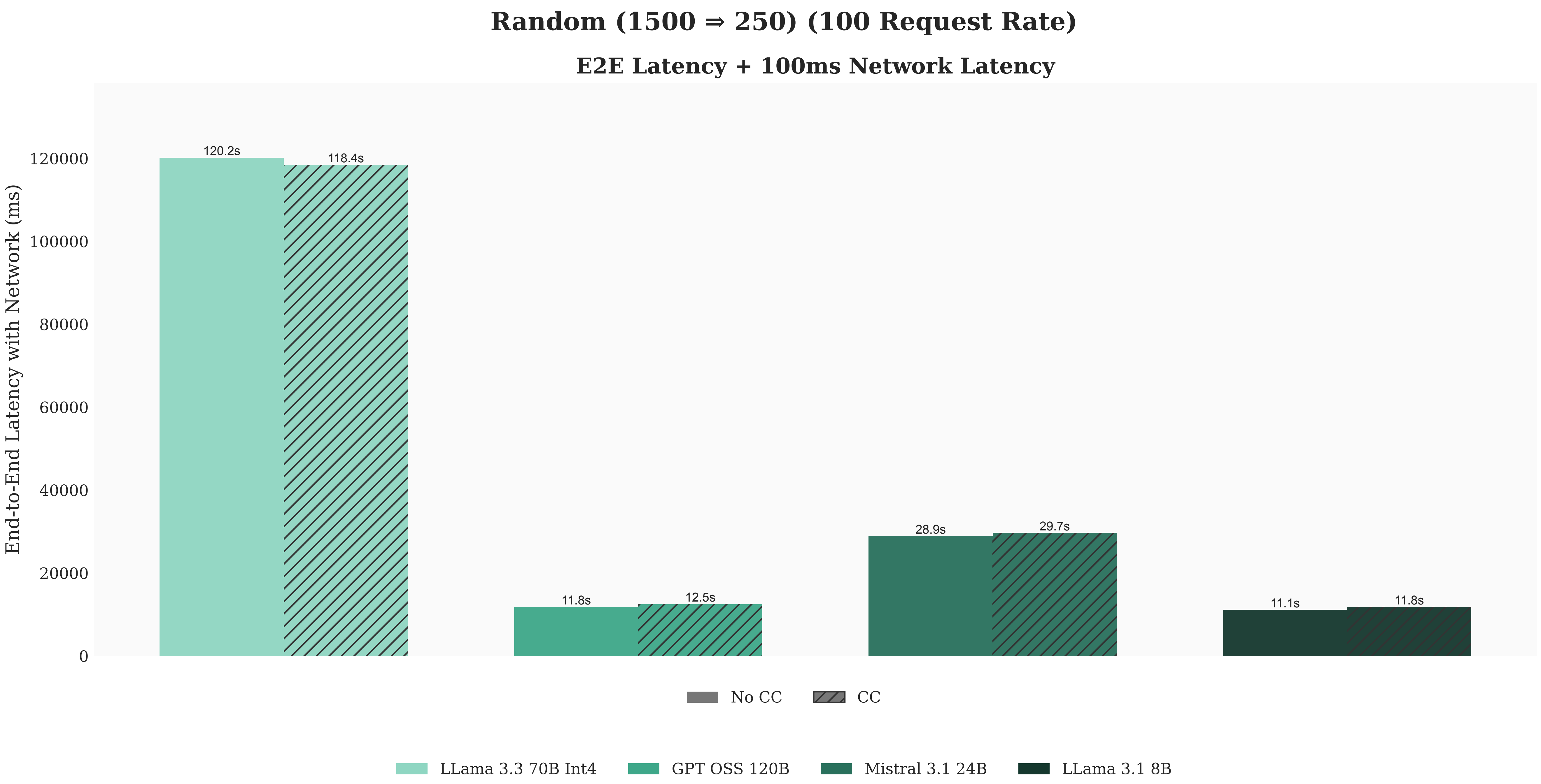
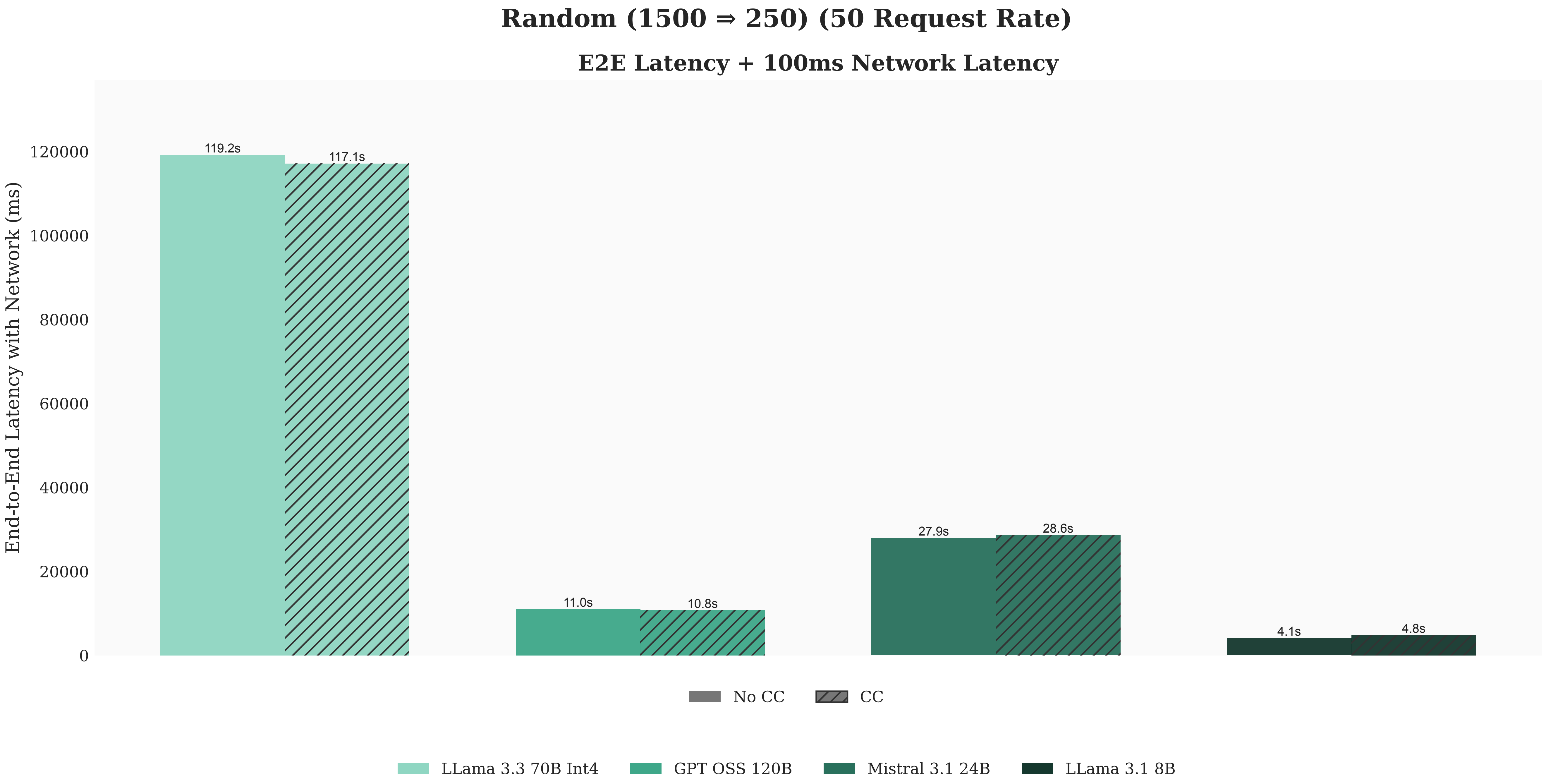


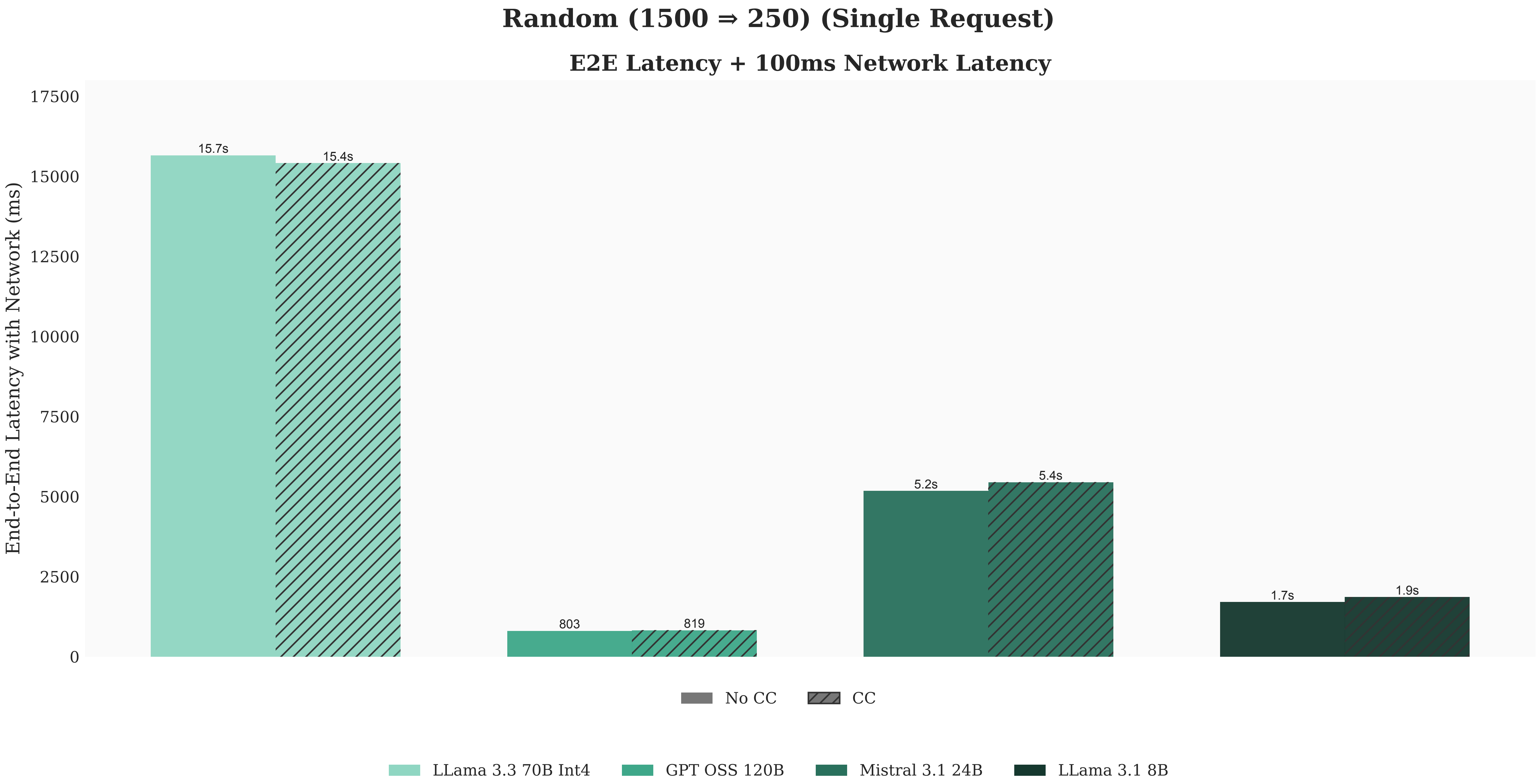
Random (1500 ⇒ 250) (100 Request Rate)

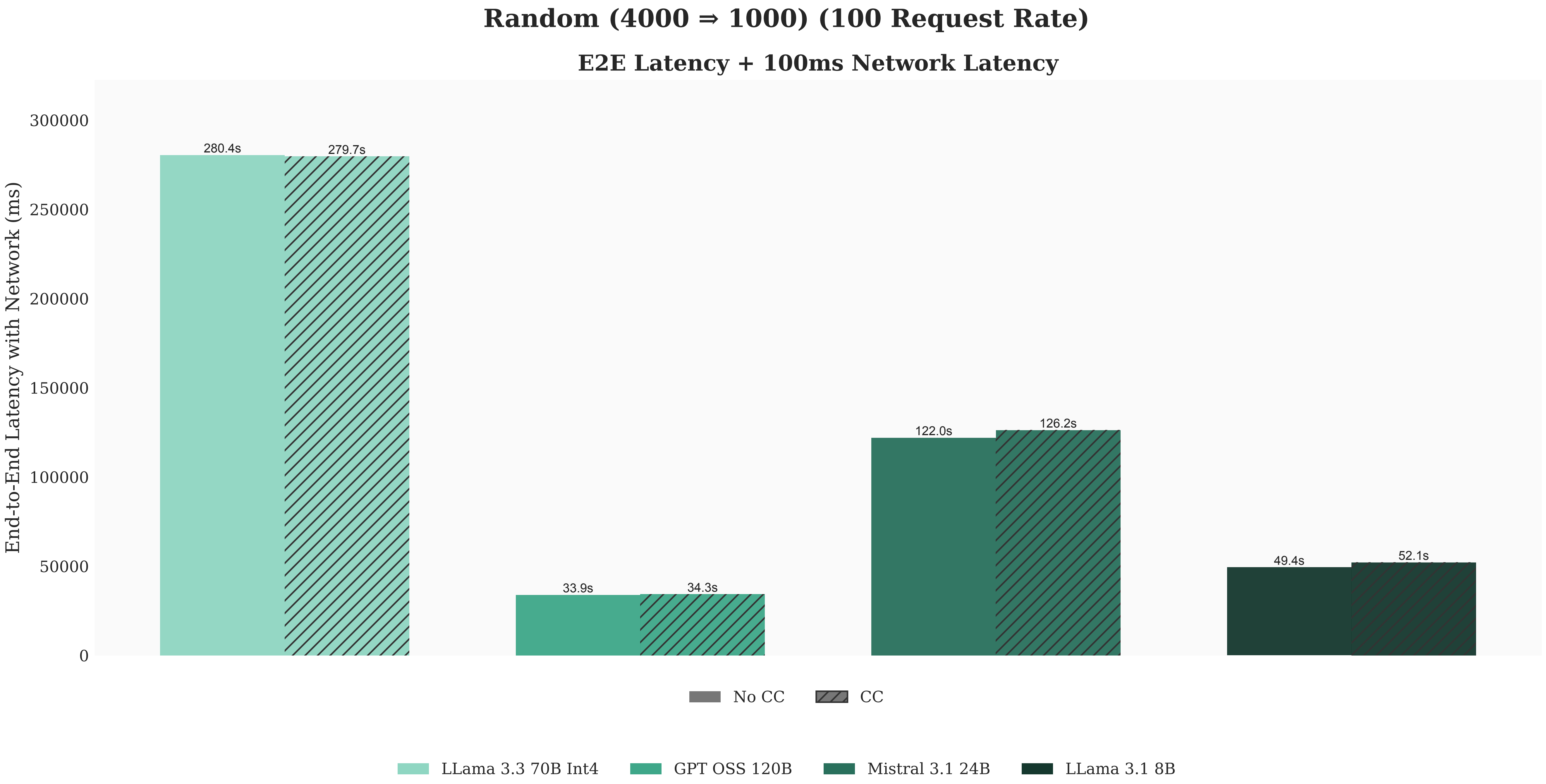
E2E Latency + 100ms Network Latency

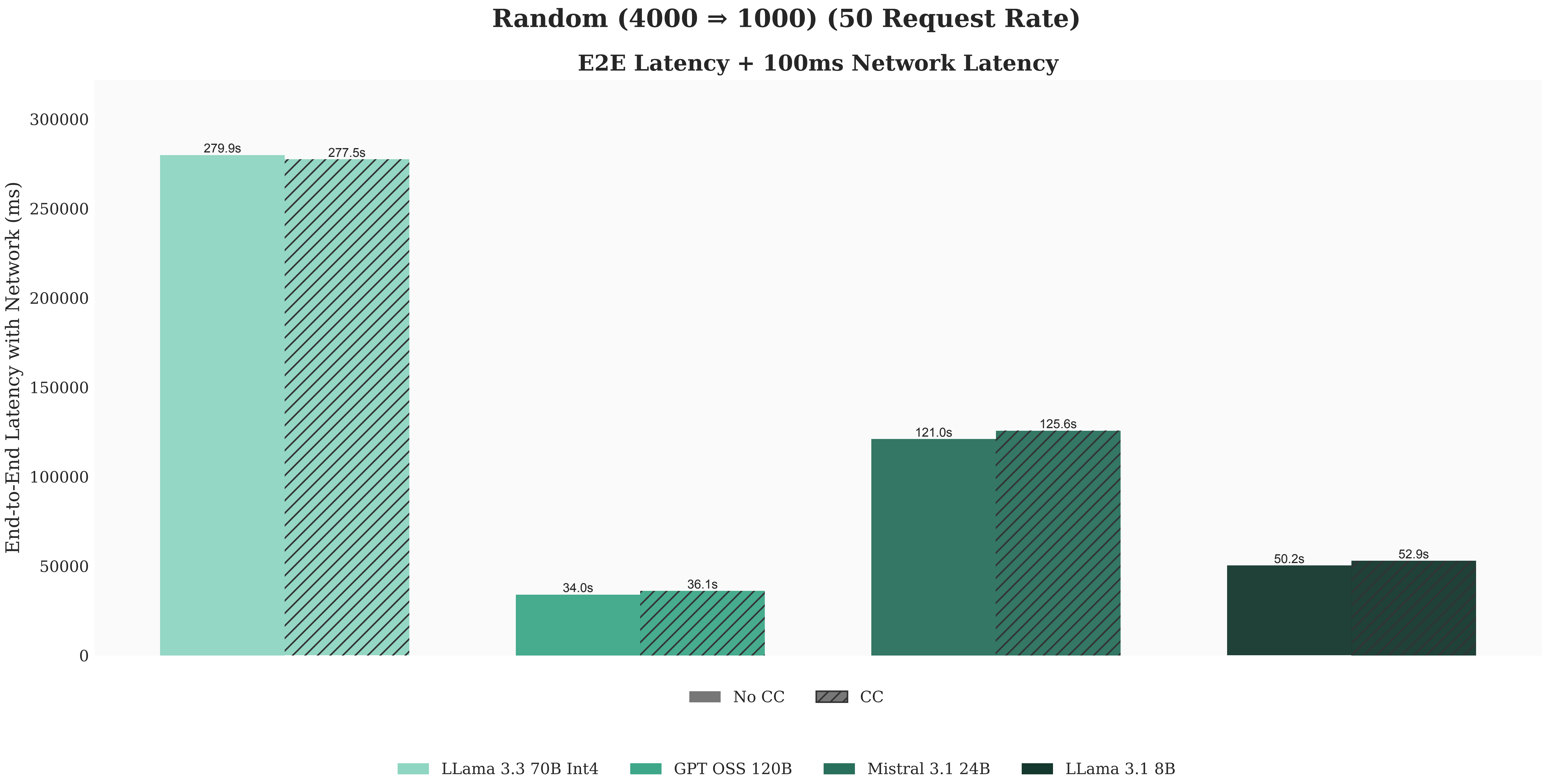


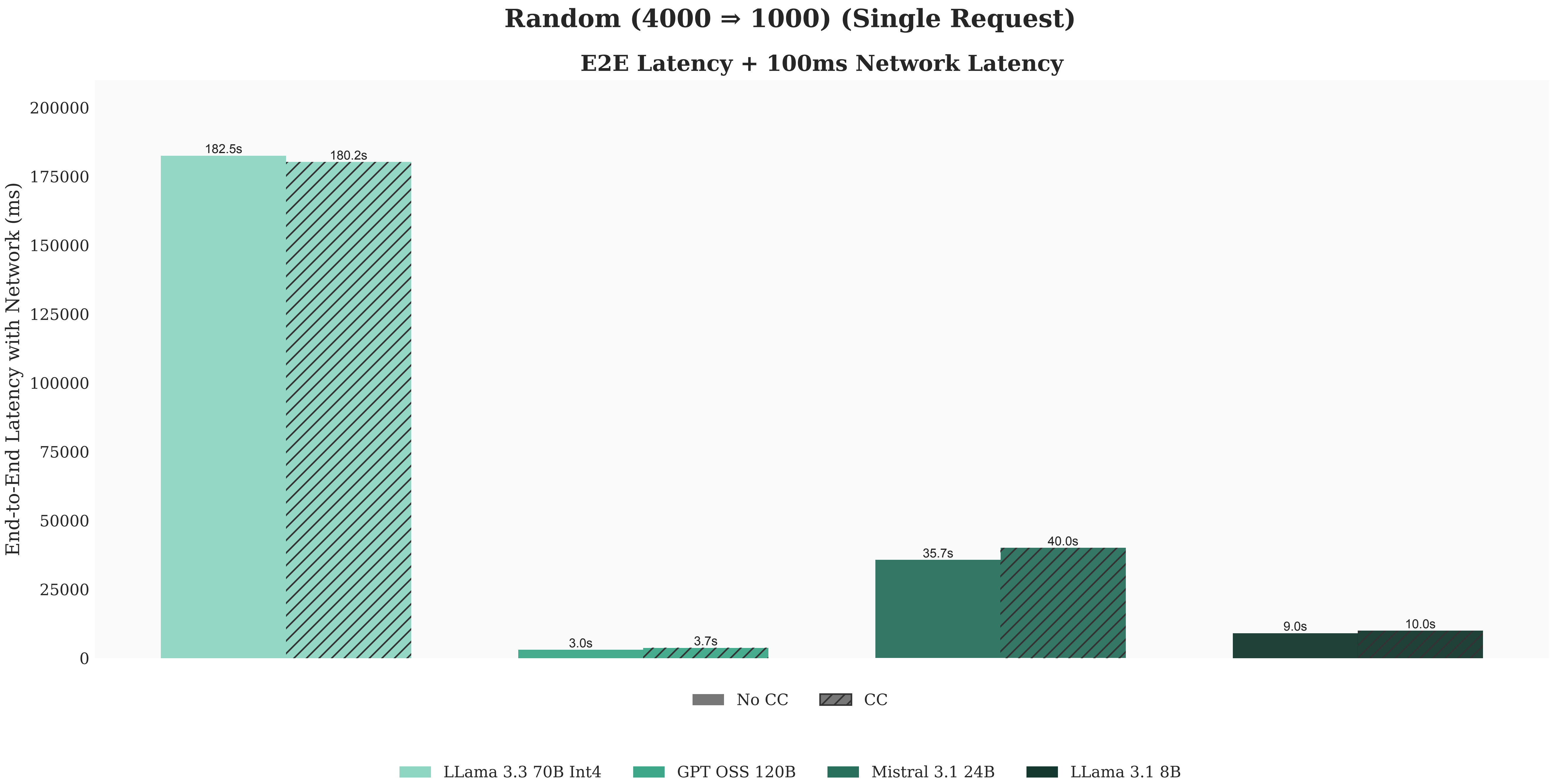
Random (1500 ⇒ 250) (50 Request Rate)

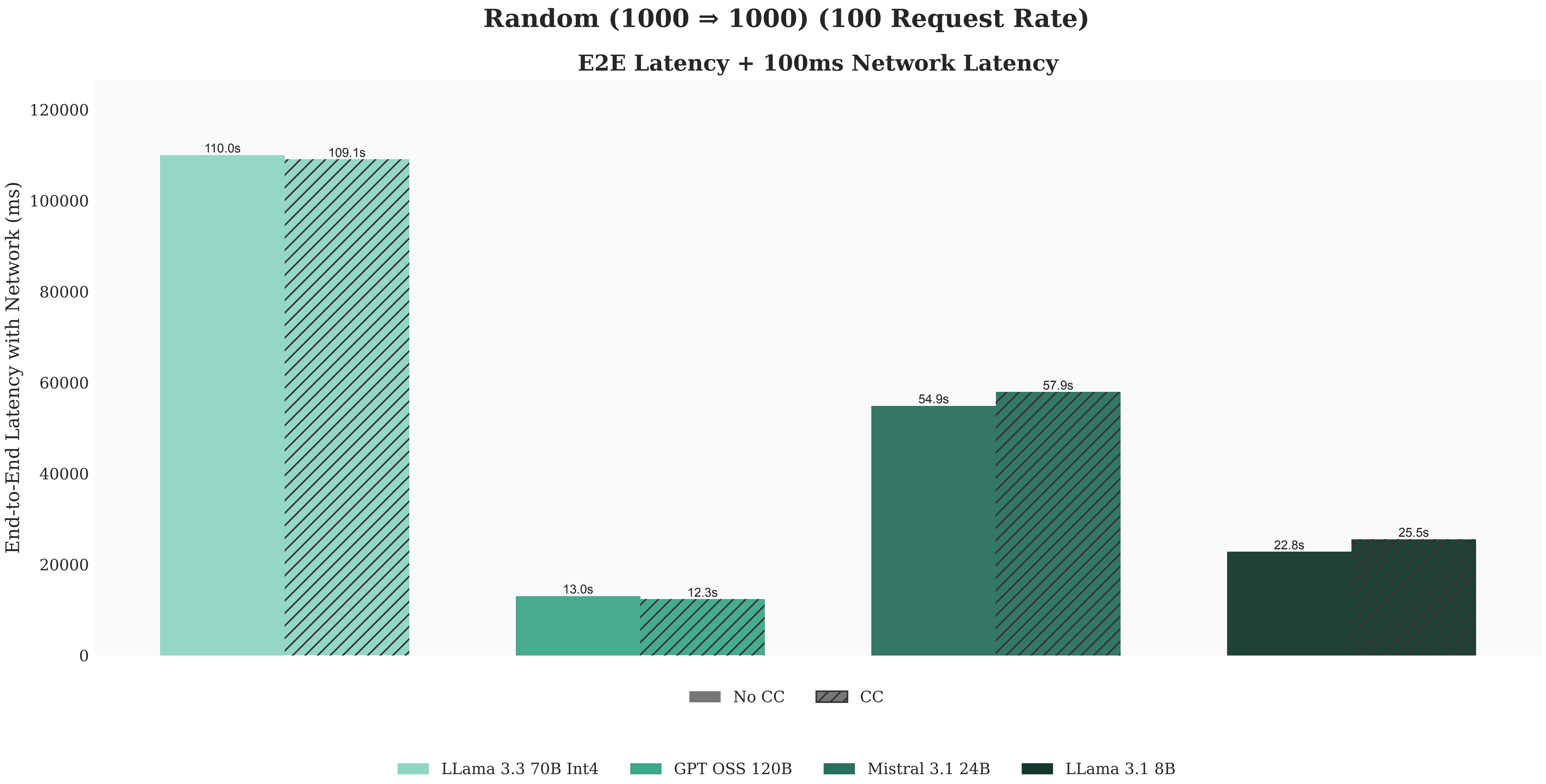
E2E Latency + 100ms Network Latency


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


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

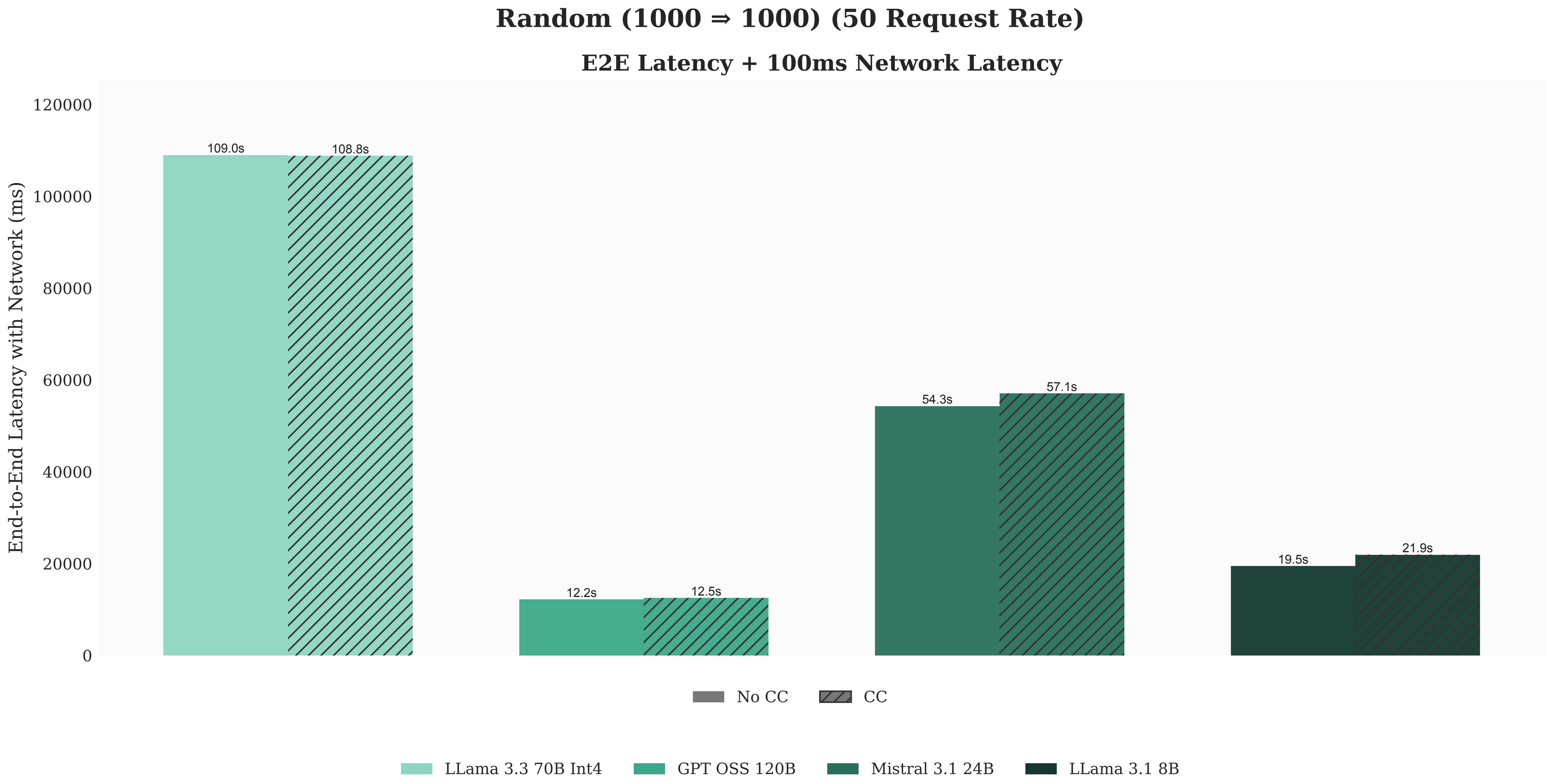

Random (4000 ⇒ 1000) (50 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
300000
250000
200000
150000
100000
50000
0
279.9s
277.5s
34.0s
36.1s
121.0s
125.6s
50.2s
52.9s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

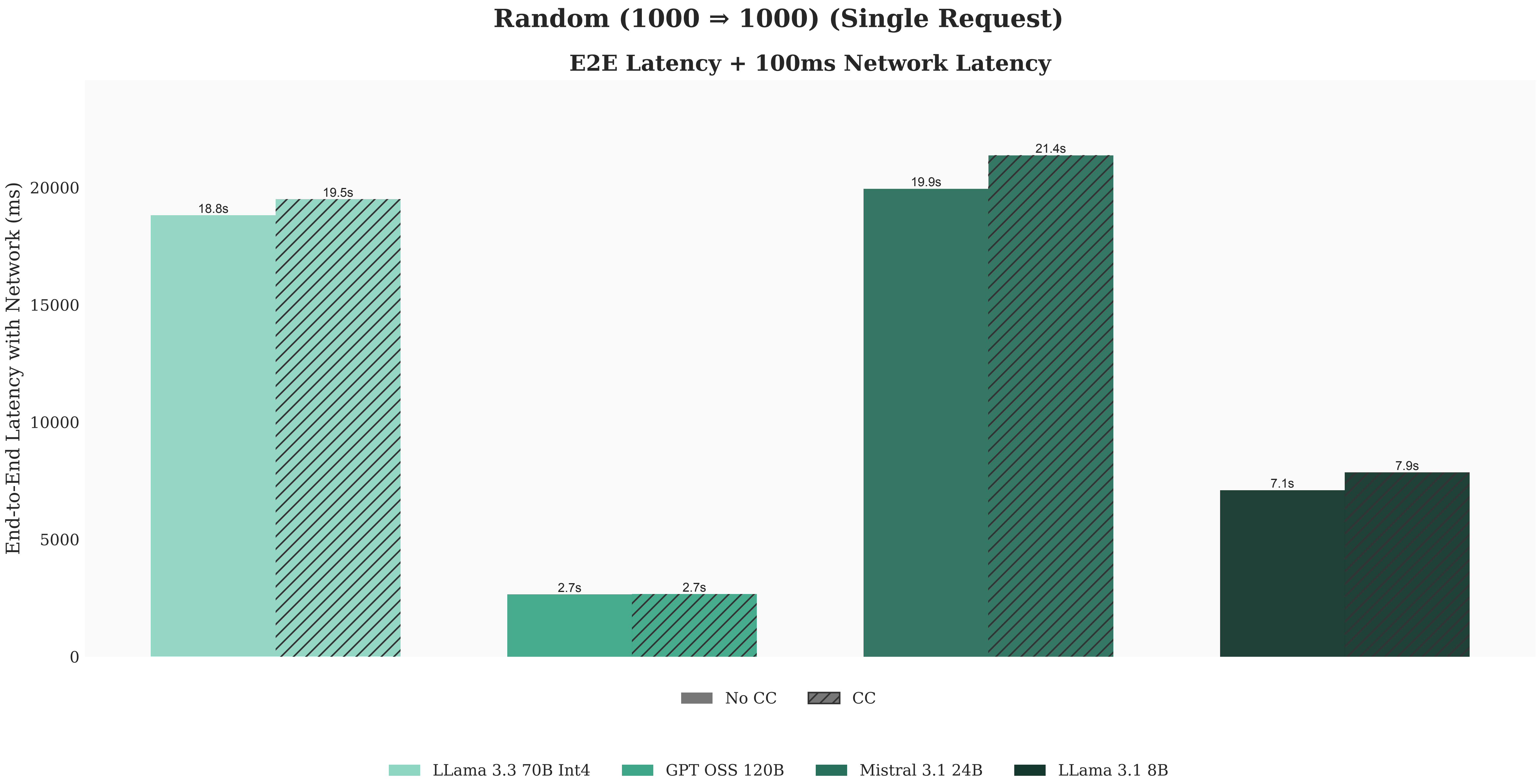

Random (4000 ⇒ 1000) (Single Request)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
200000
175000
150000
125000
100000
75000
50000
25000
0
182.5s
180.2s
3.0s
3.7s
35.7s
40.0s
9.0s
10.0s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B



Random (1000 ⇒ 1000) (100 Request Rate)

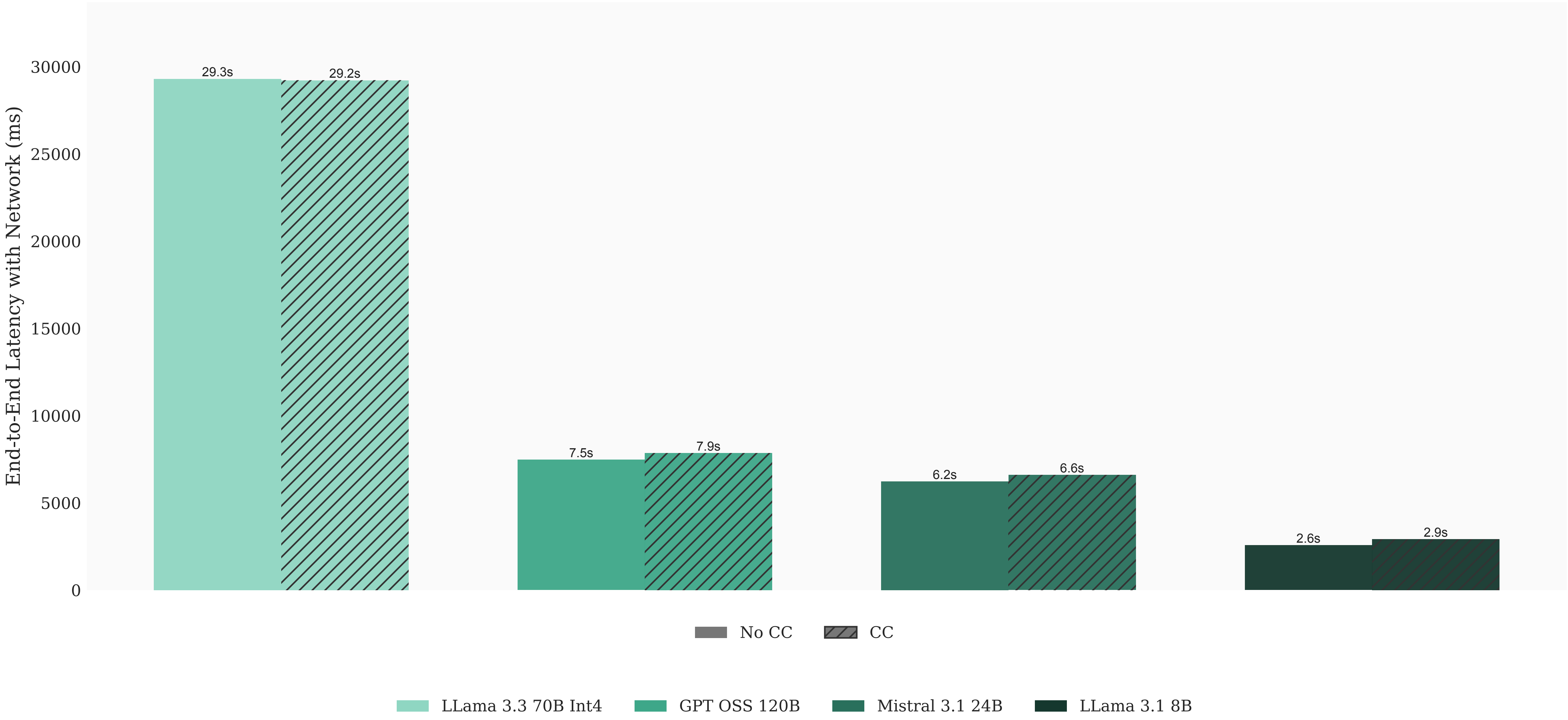
E2E Latency + 100ms Network Latency

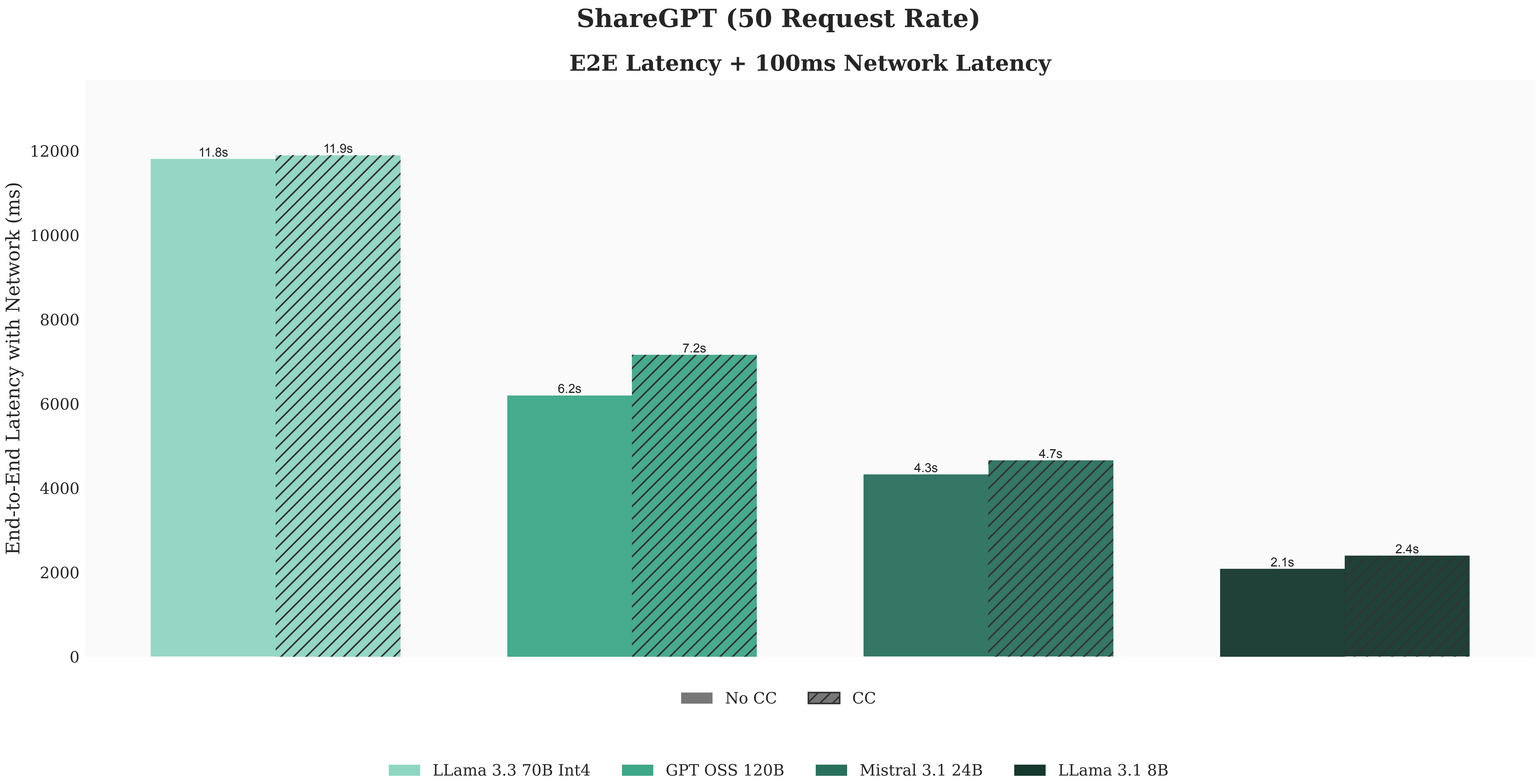

Random (1000 ⇒ 1000) (50 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
120000
100000
80000
60000
40000
20000
0
109.0s
108.8s
12.2s
12.5s
54.3s
57.1s
19.5s
21.9s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

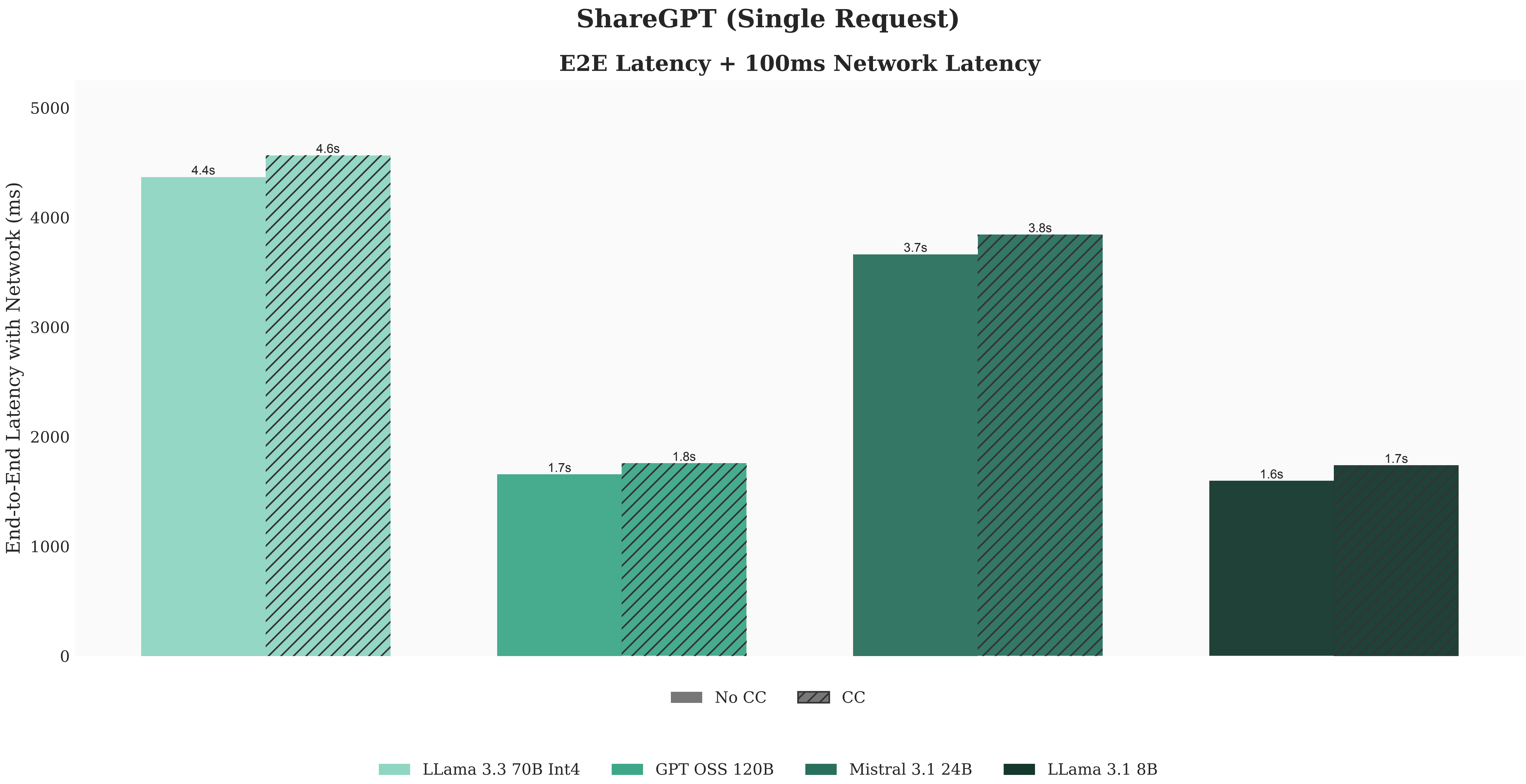

Random (1000 ⇒ 1000) (Single Request)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
5000
10000
15000
20000
18.8s
19.5s
2.7s
2.7s
19.9s
21.4s
7.1s
7.9s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

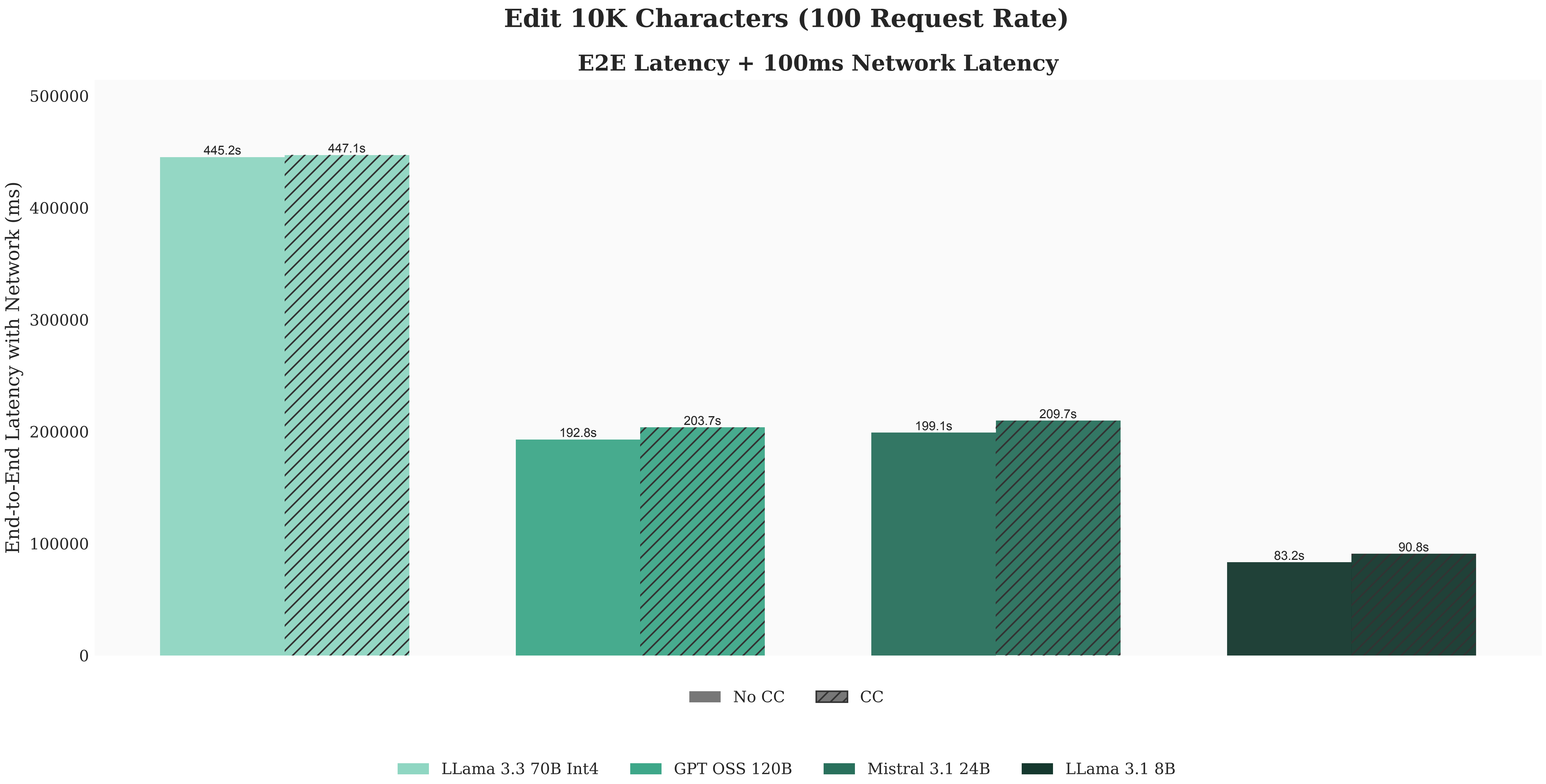
ShareGPT (100 Request Rate)

E2E Latency + 100ms Network Latency



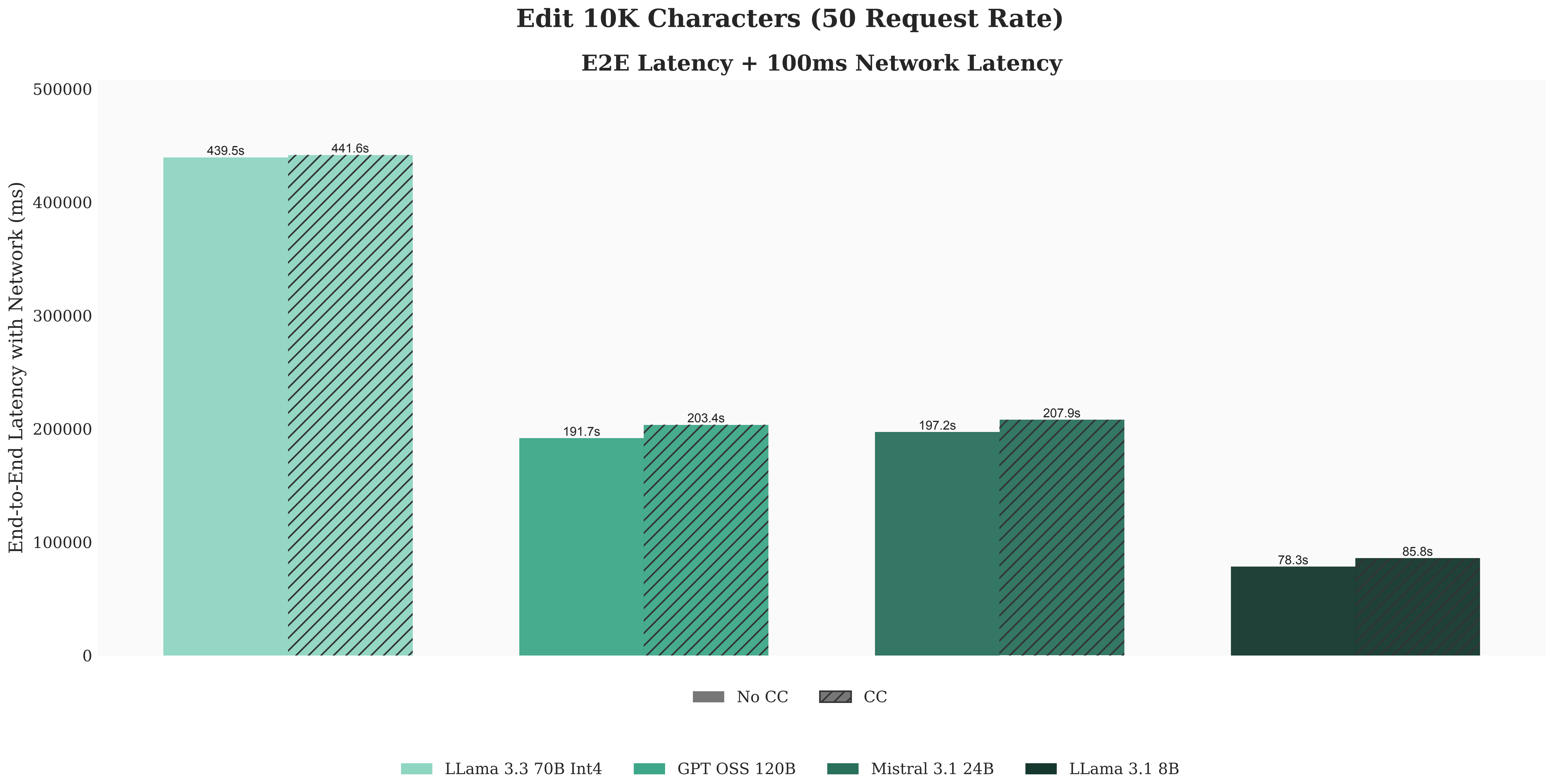

ShareGPT (50 Request Rate)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
2000
4000
6000
8000
10000
12000
11.8s
11.9s
6.2s
7.2s
4.3s
4.7s
2.1s
2.4s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


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



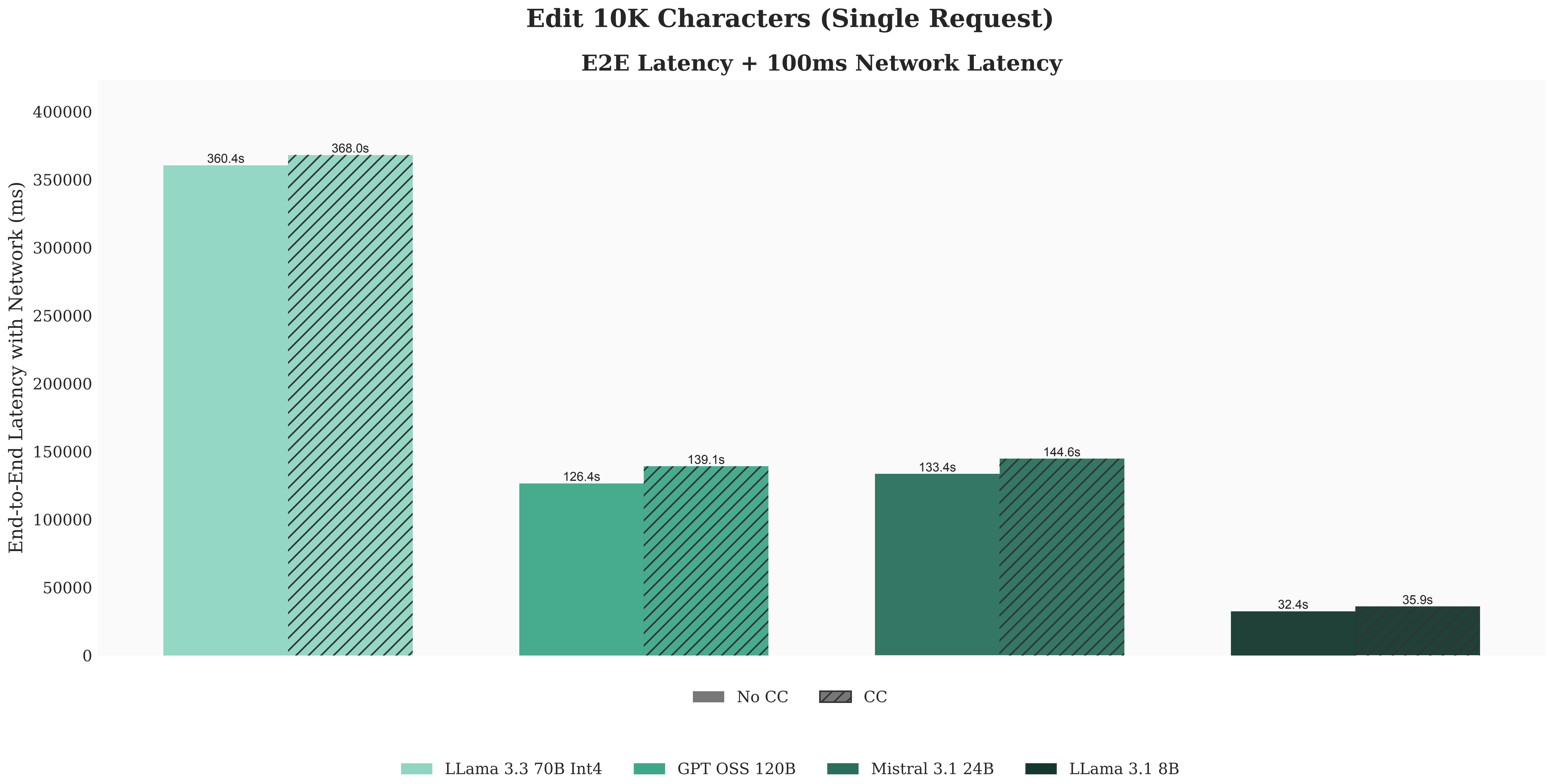
Edit 10K Characters (100 Request Rate)

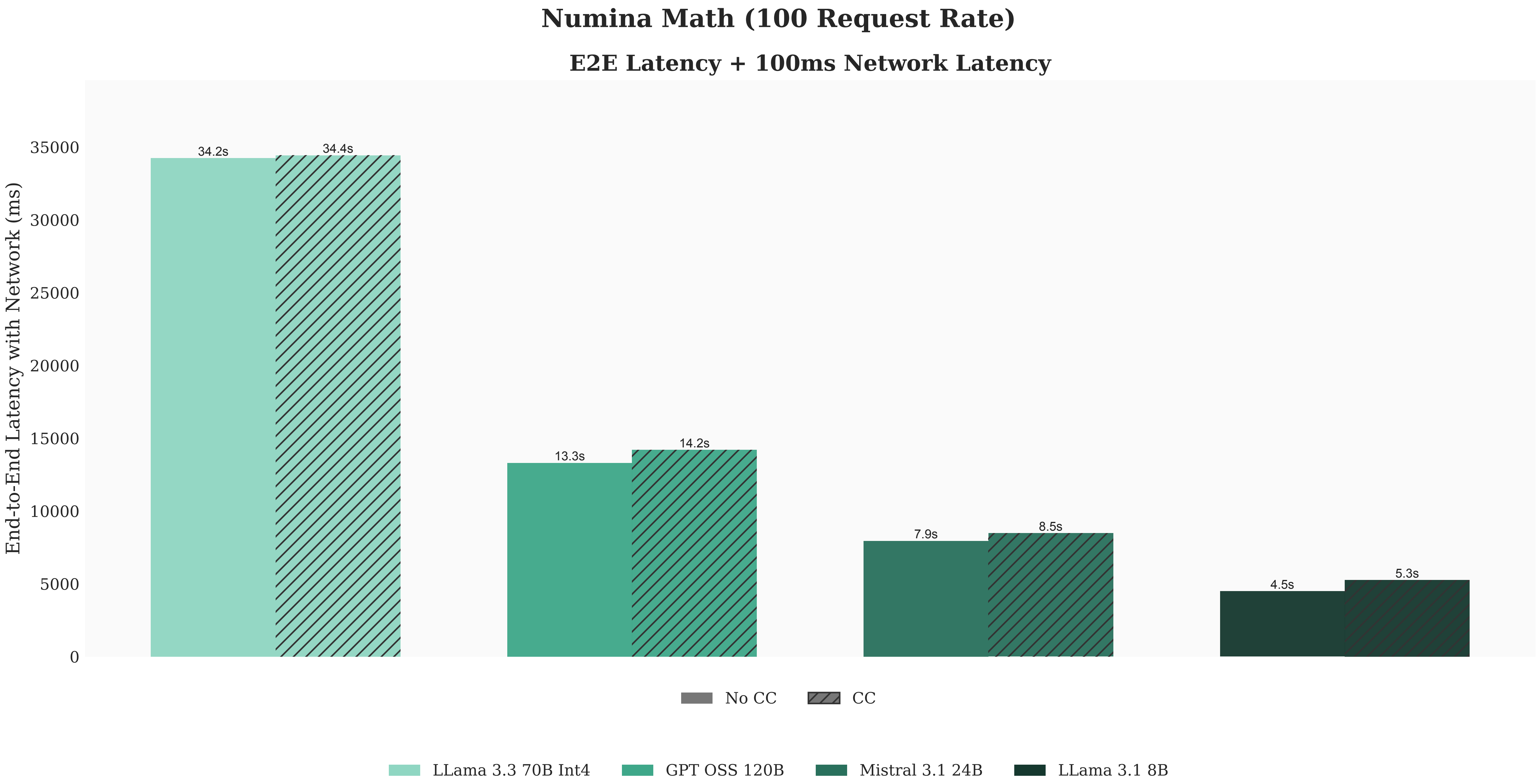
E2E Latency + 100ms Network Latency

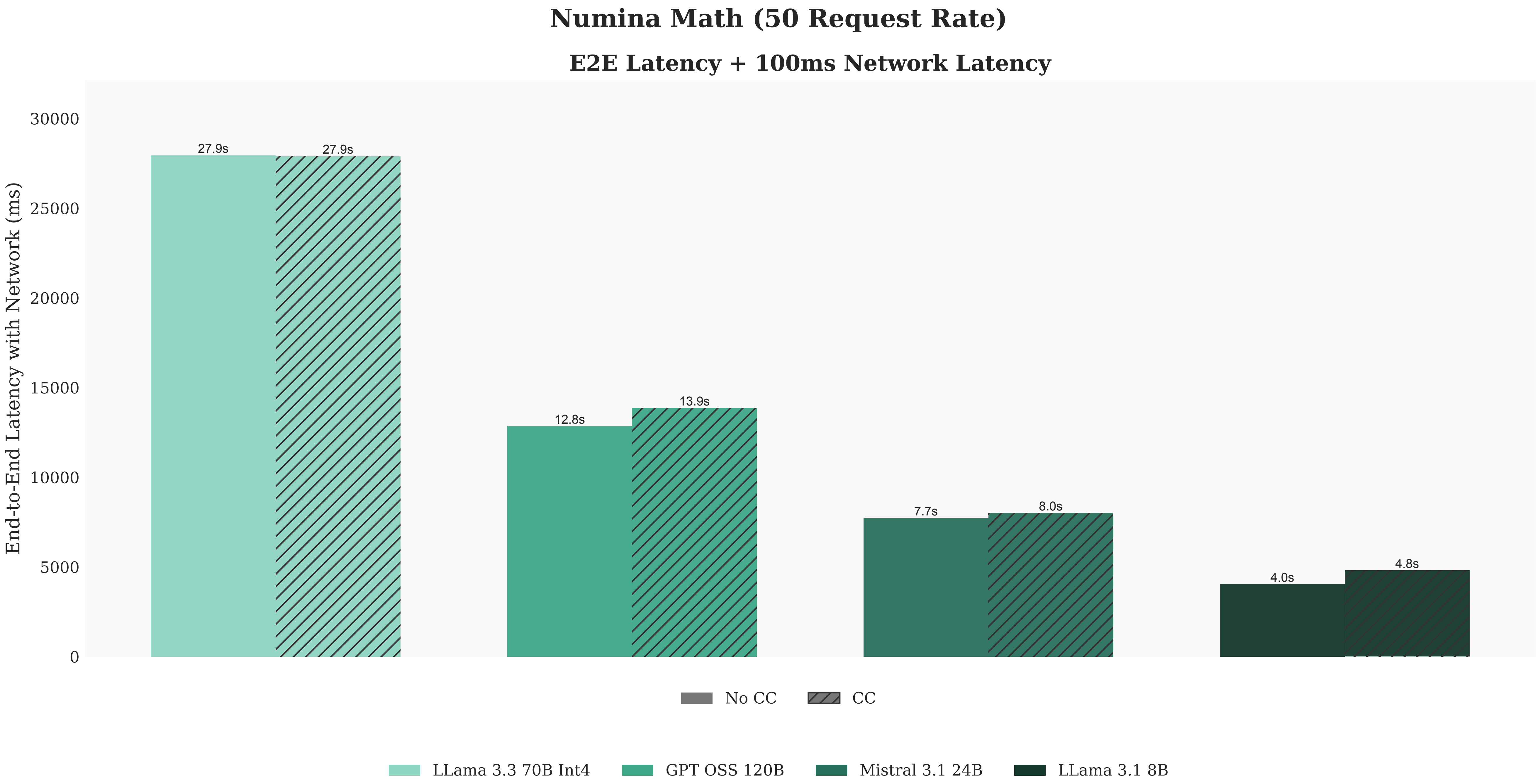


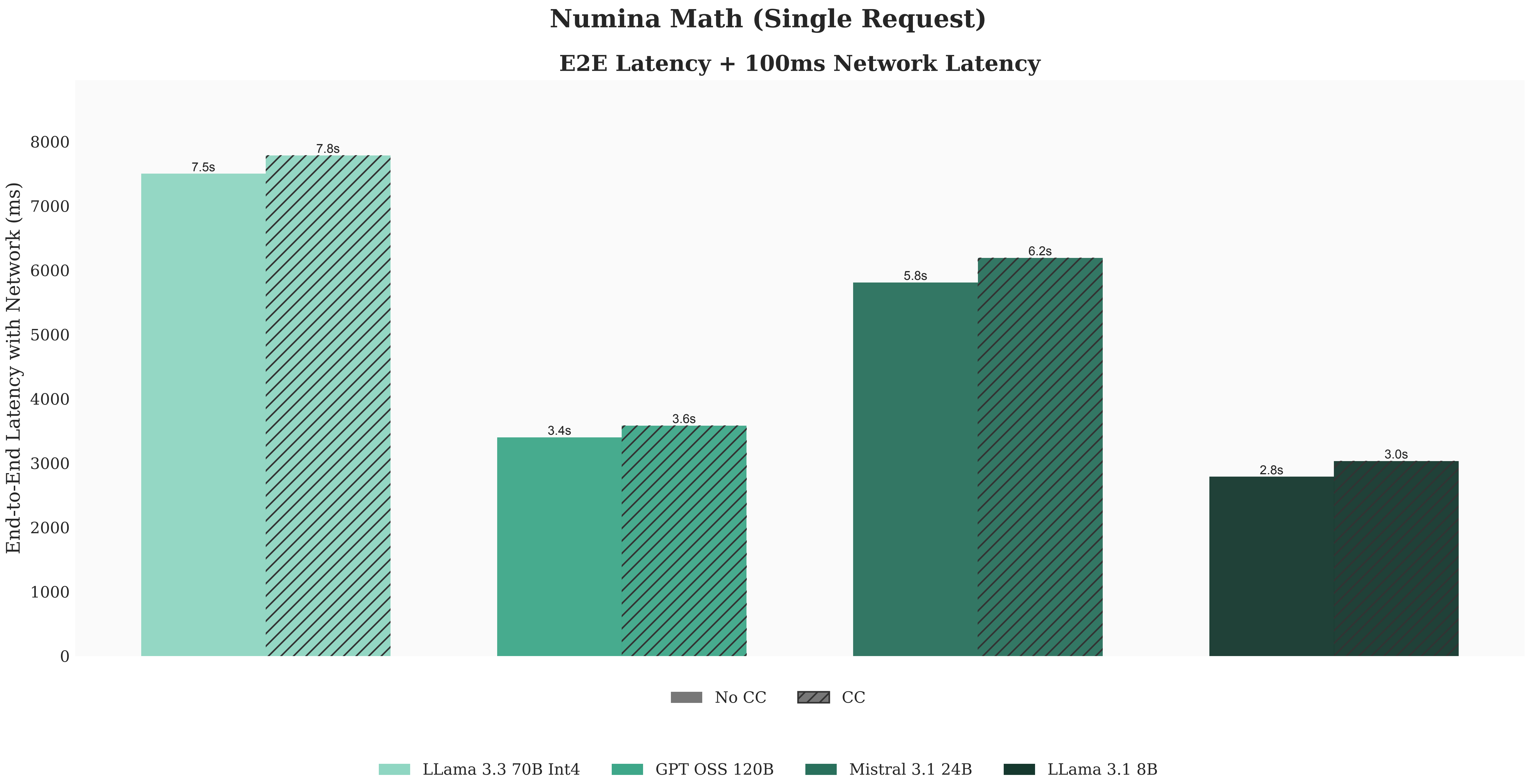
Edit 10K Characters (50 Request Rate)

E2E Latency + 100ms Network Latency


Edit 10K Characters (Single Request)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
400000
350000
300000
250000
200000
150000
100000
50000
0
360.4s
368.0s
126.4s
139.1s
133.4s
144.6s
32.4s
35.9s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


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


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


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