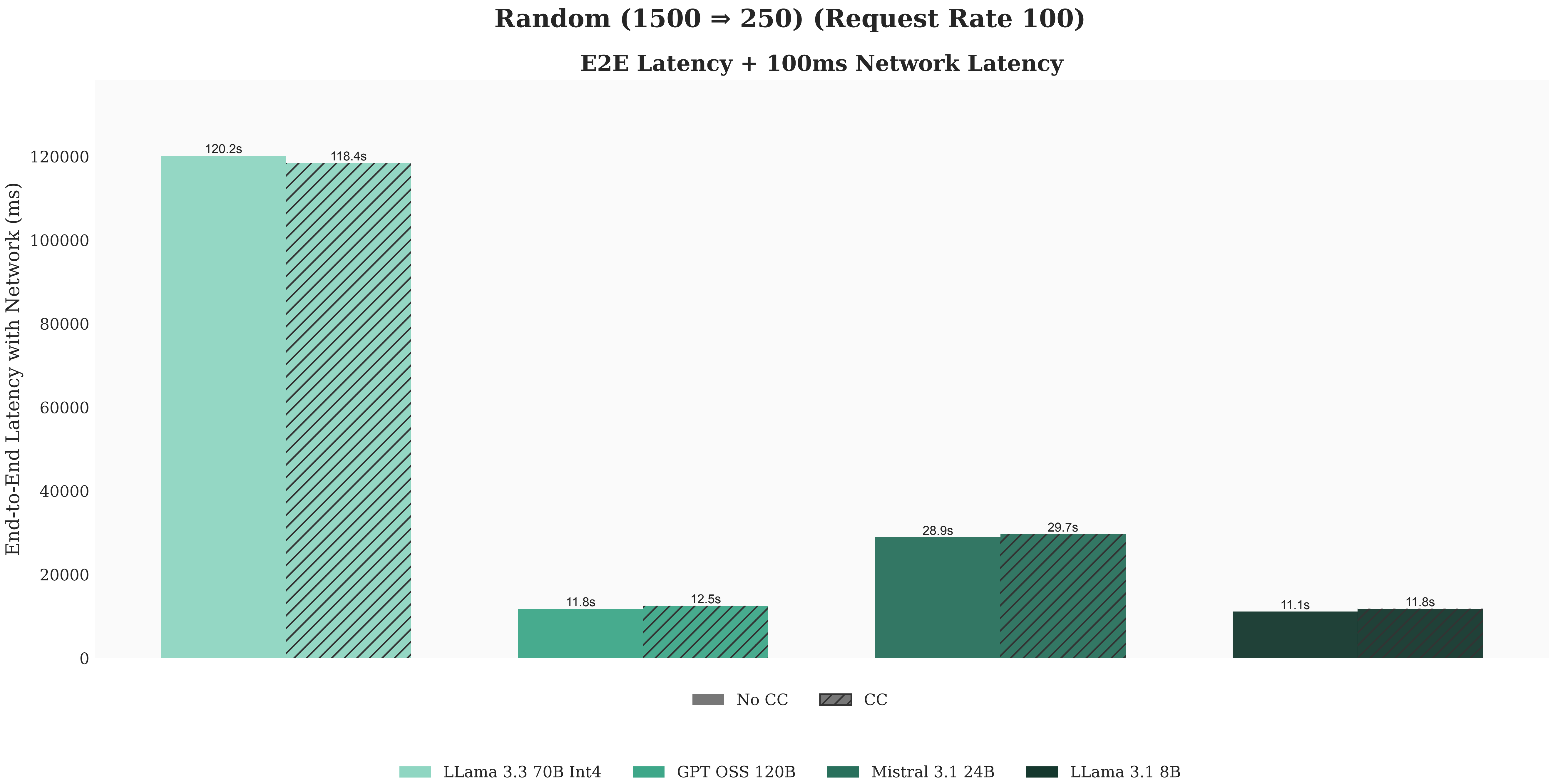
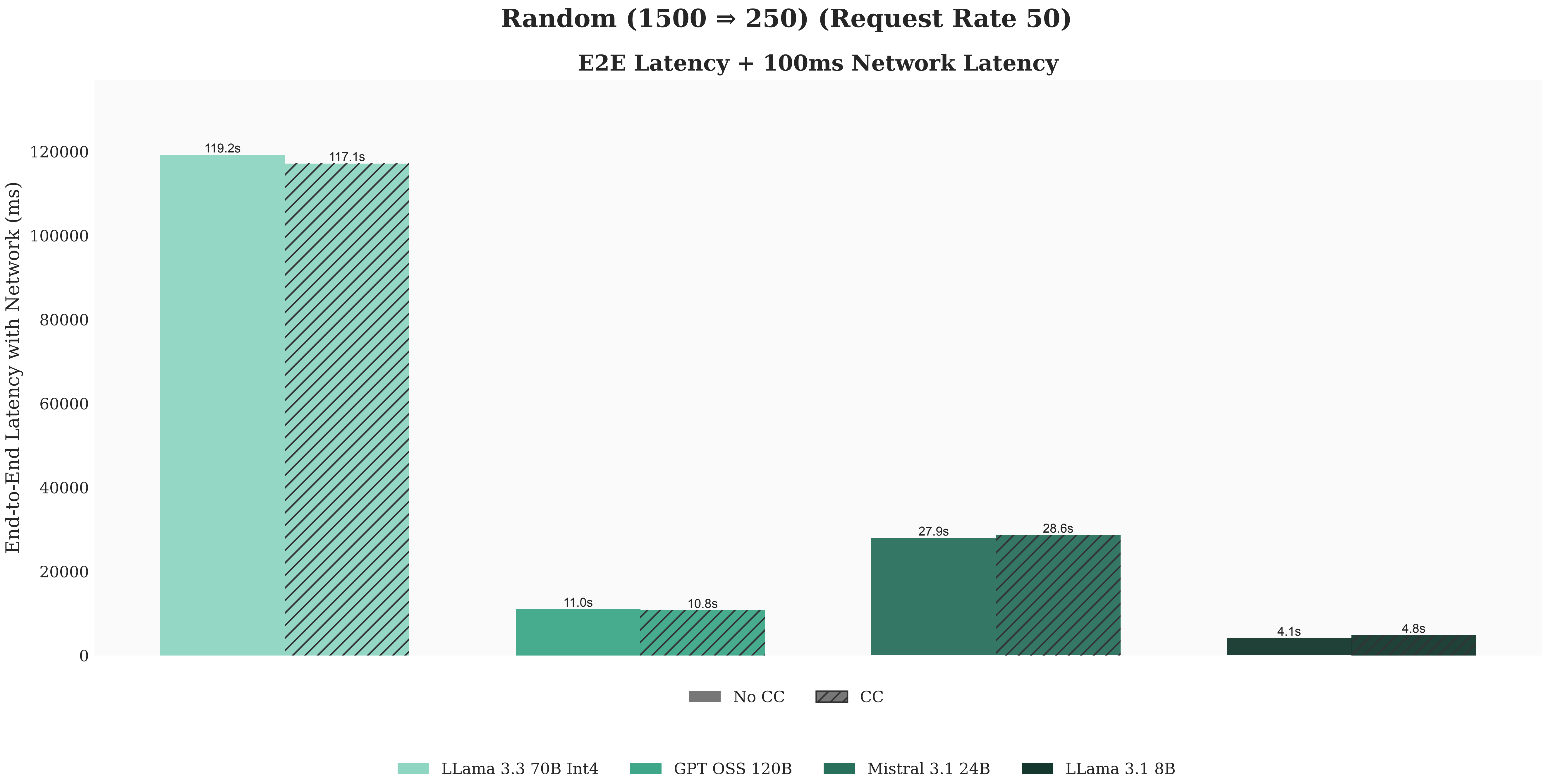
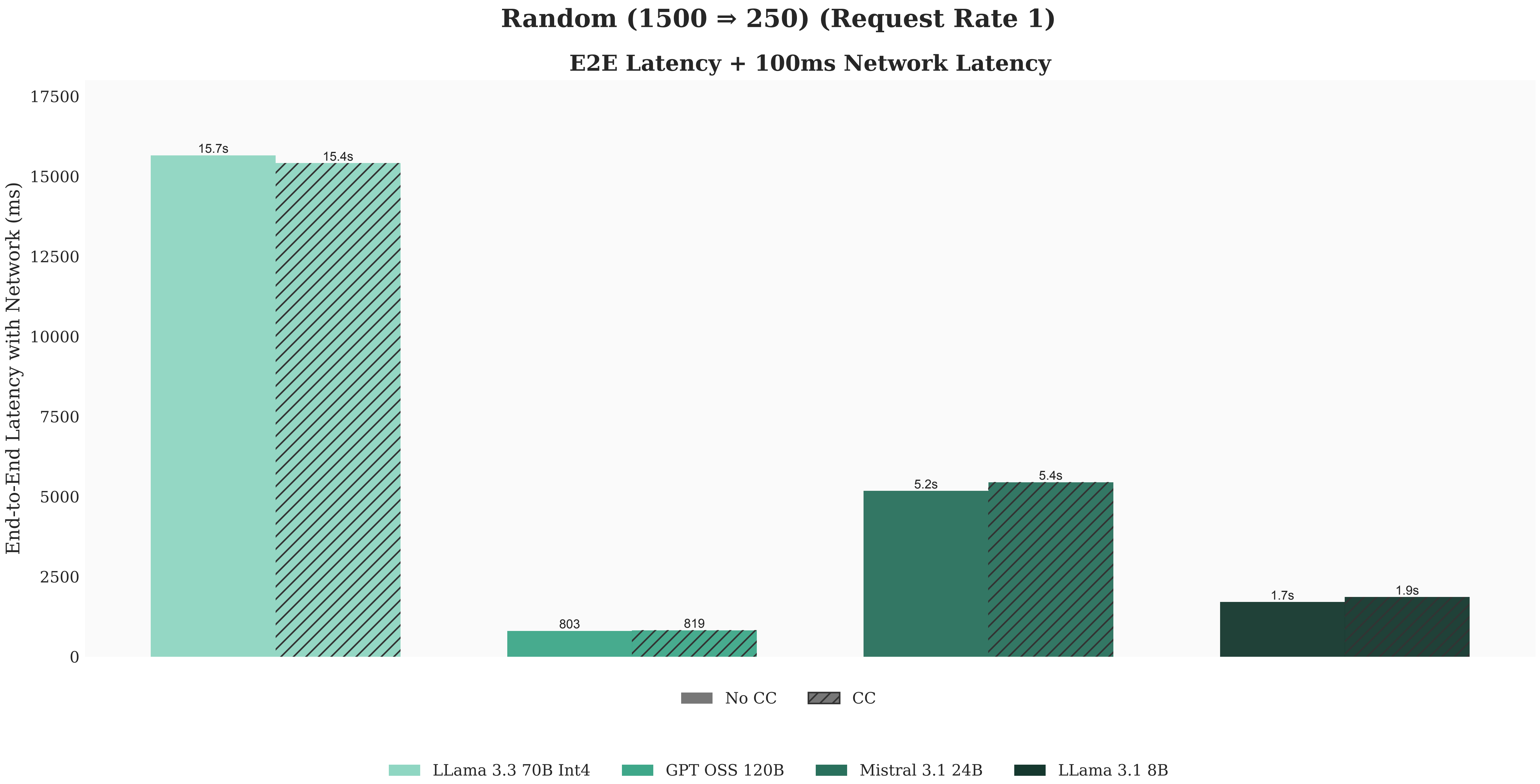
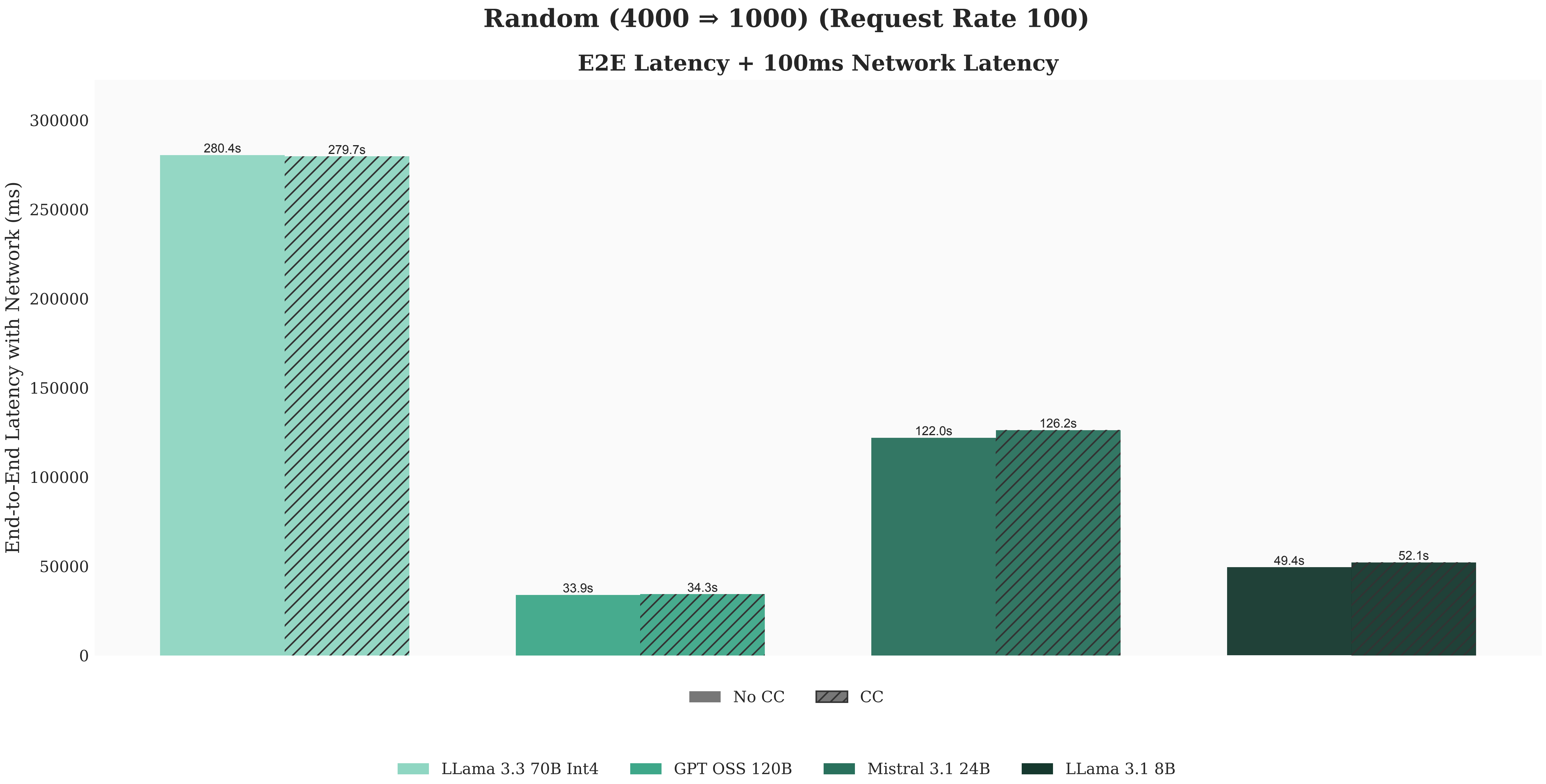

Random (1500 ⇒ 250) (Request Rate 100)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
20000
40000
60000
80000
100000
120000
120.2s
118.4s
11.8s
12.5s
28.9s
29.7s
11.1s
11.8s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


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



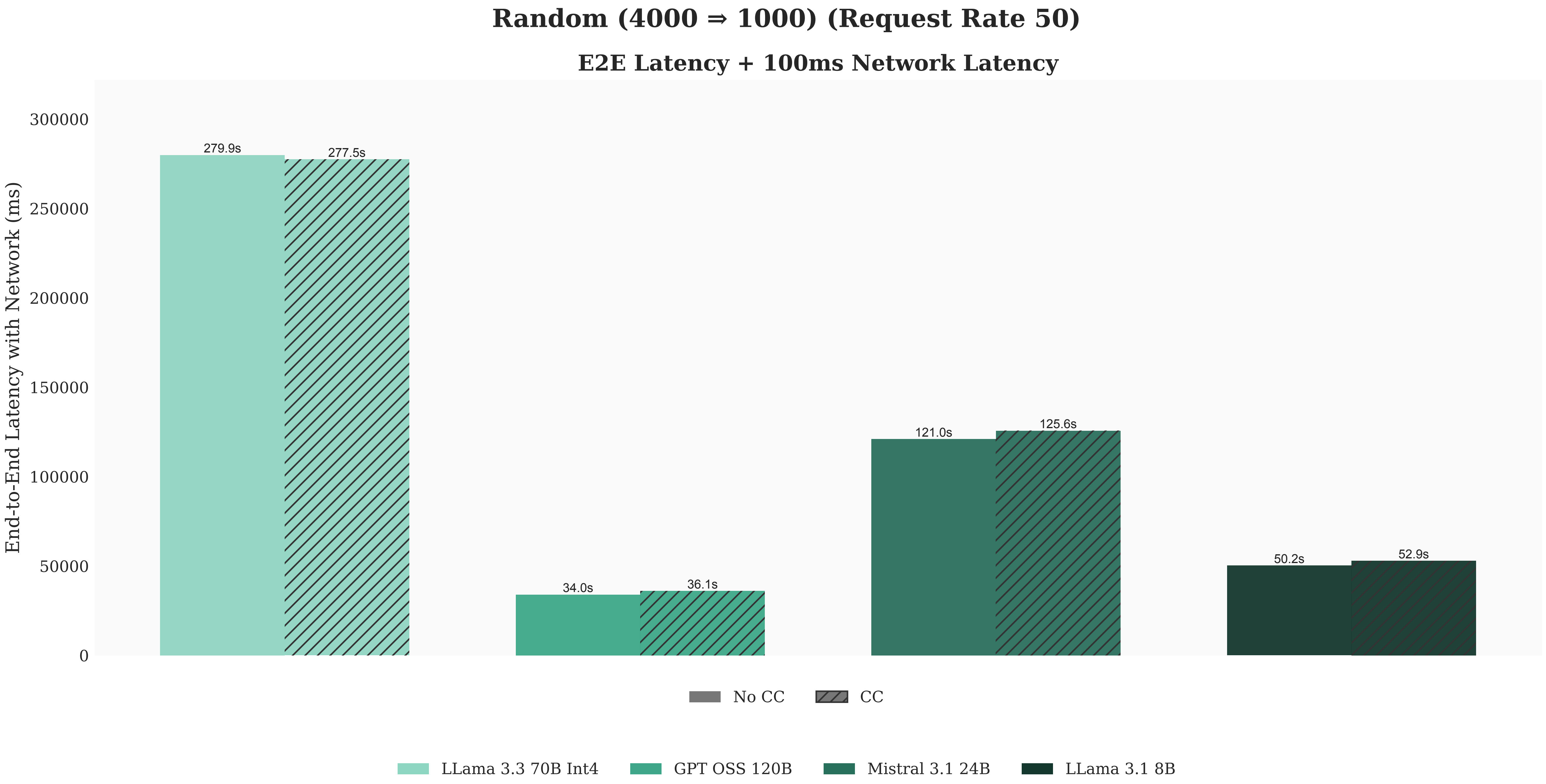
Random (1500 ⇒ 250) (Request Rate 1)

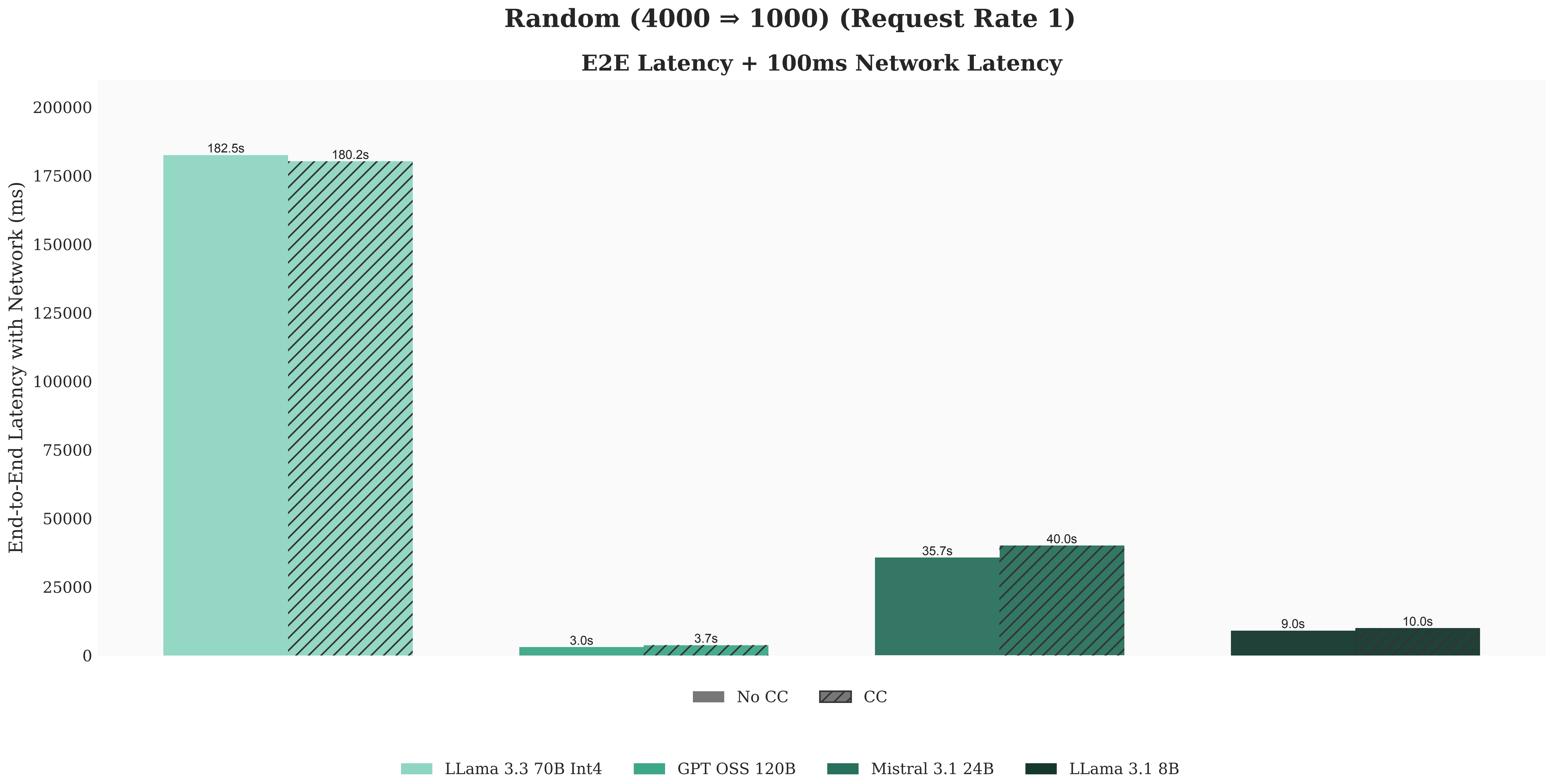
E2E Latency + 100ms Network Latency

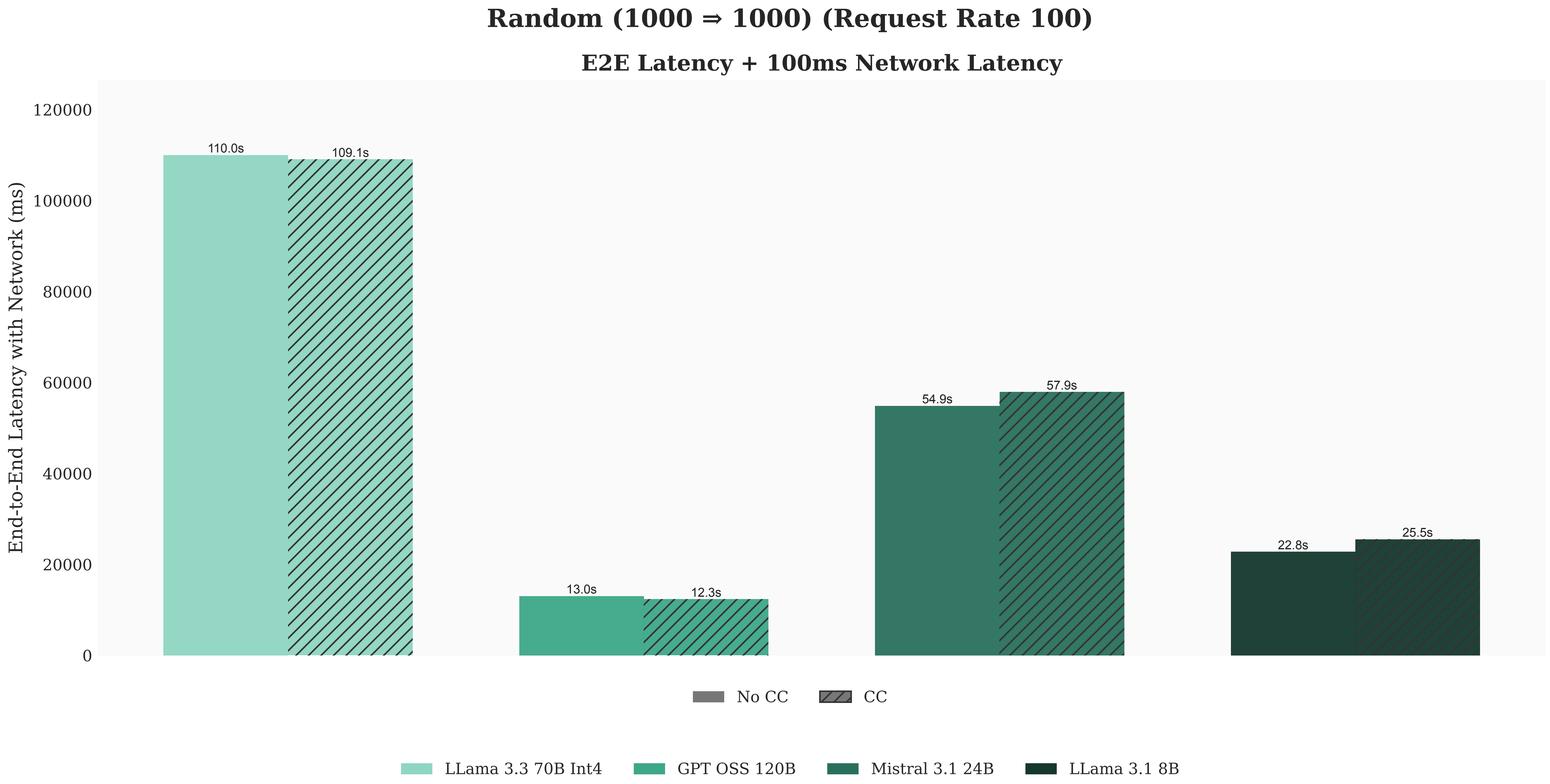


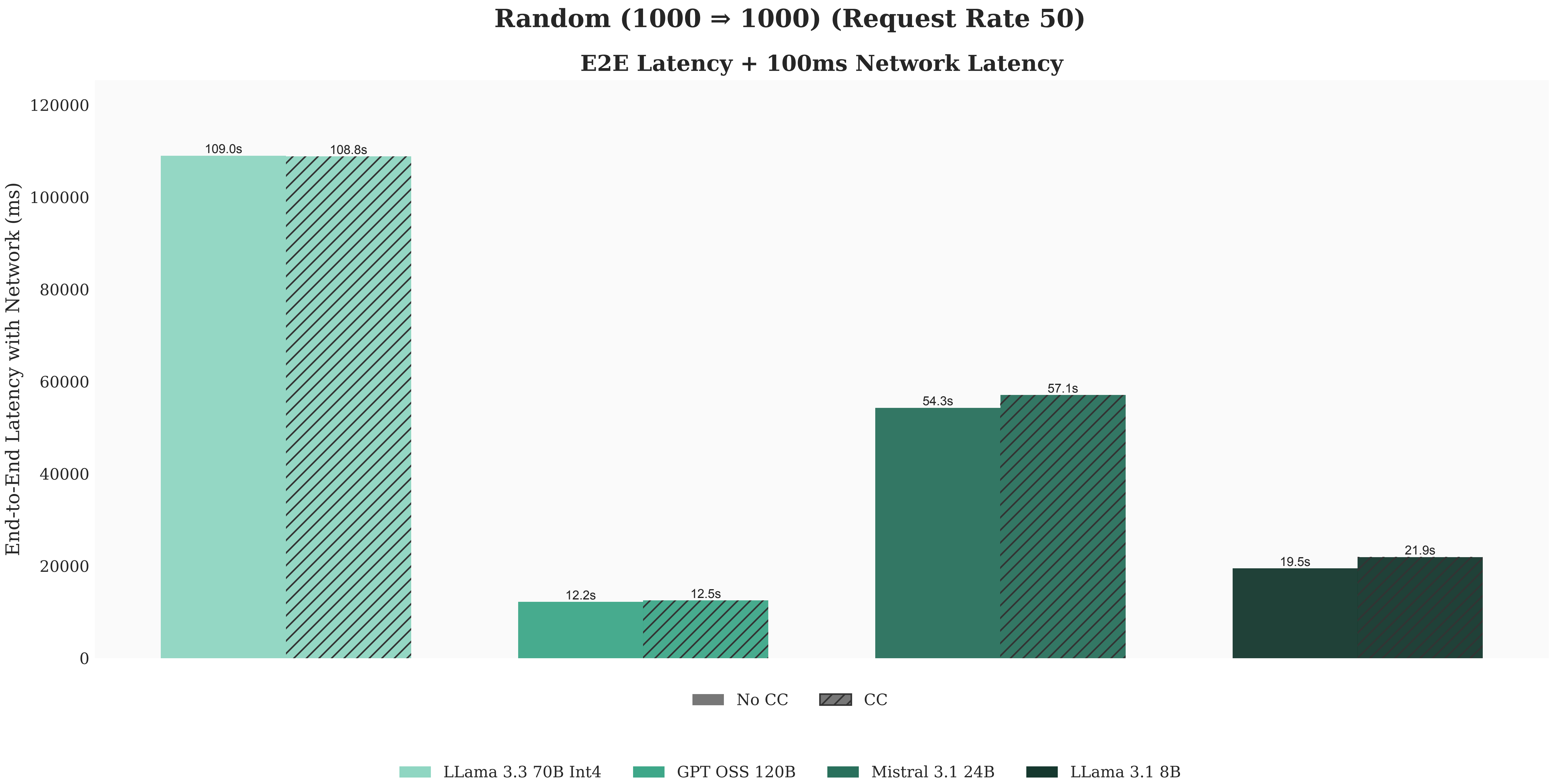
Random (4000 ⇒ 1000) (Request Rate 100)

E2E Latency + 100ms Network Latency


Random (4000 ⇒ 1000) (Request Rate 50)
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

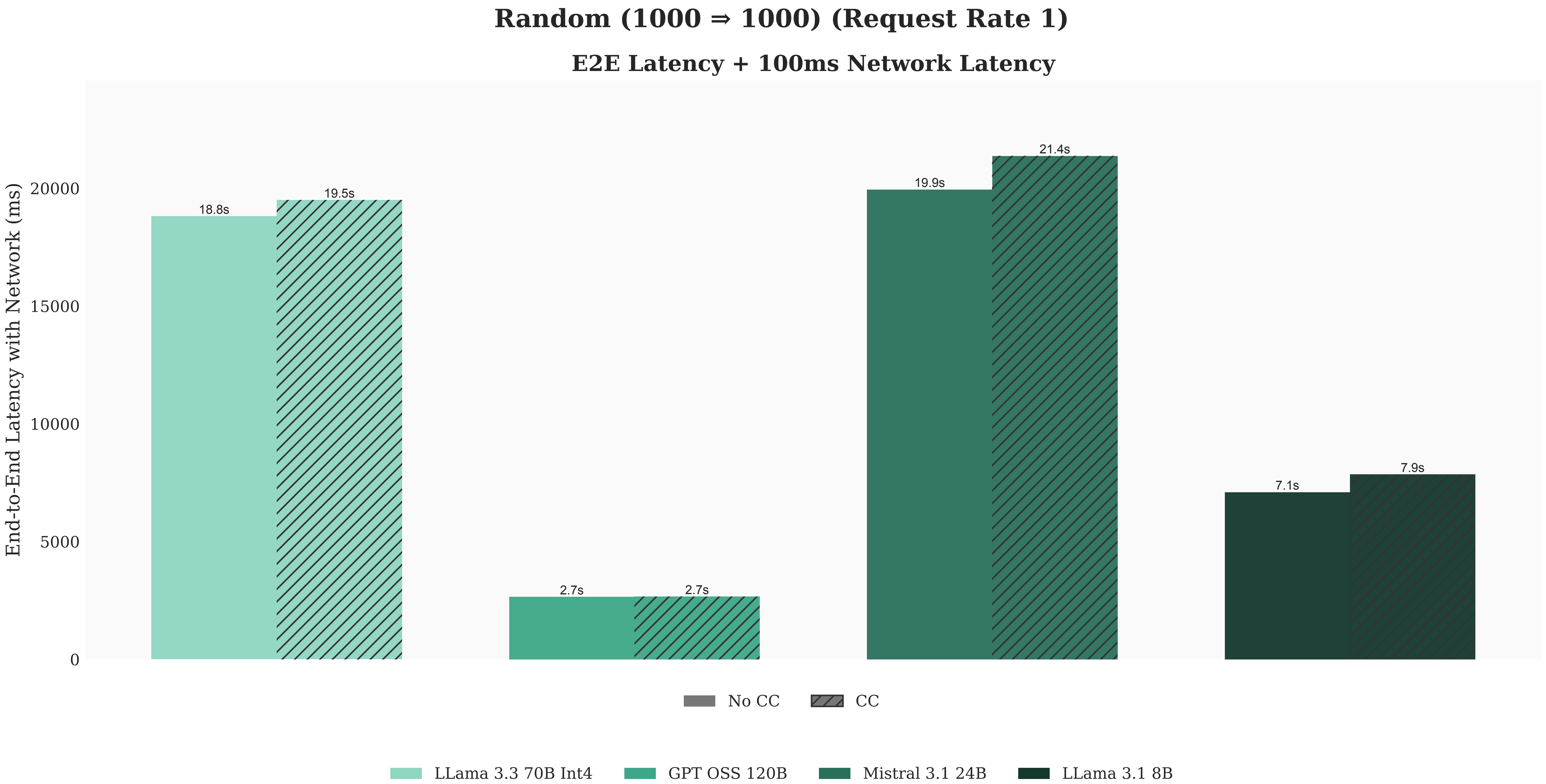

Random (4000 ⇒ 1000) (Request Rate 1)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
200000
175000
150000
125000
100000
75000
50000
25000
0
182.5s
180.2s
3.0s
3.7s
35.7s
40.0s
9.0s
10.0s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


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



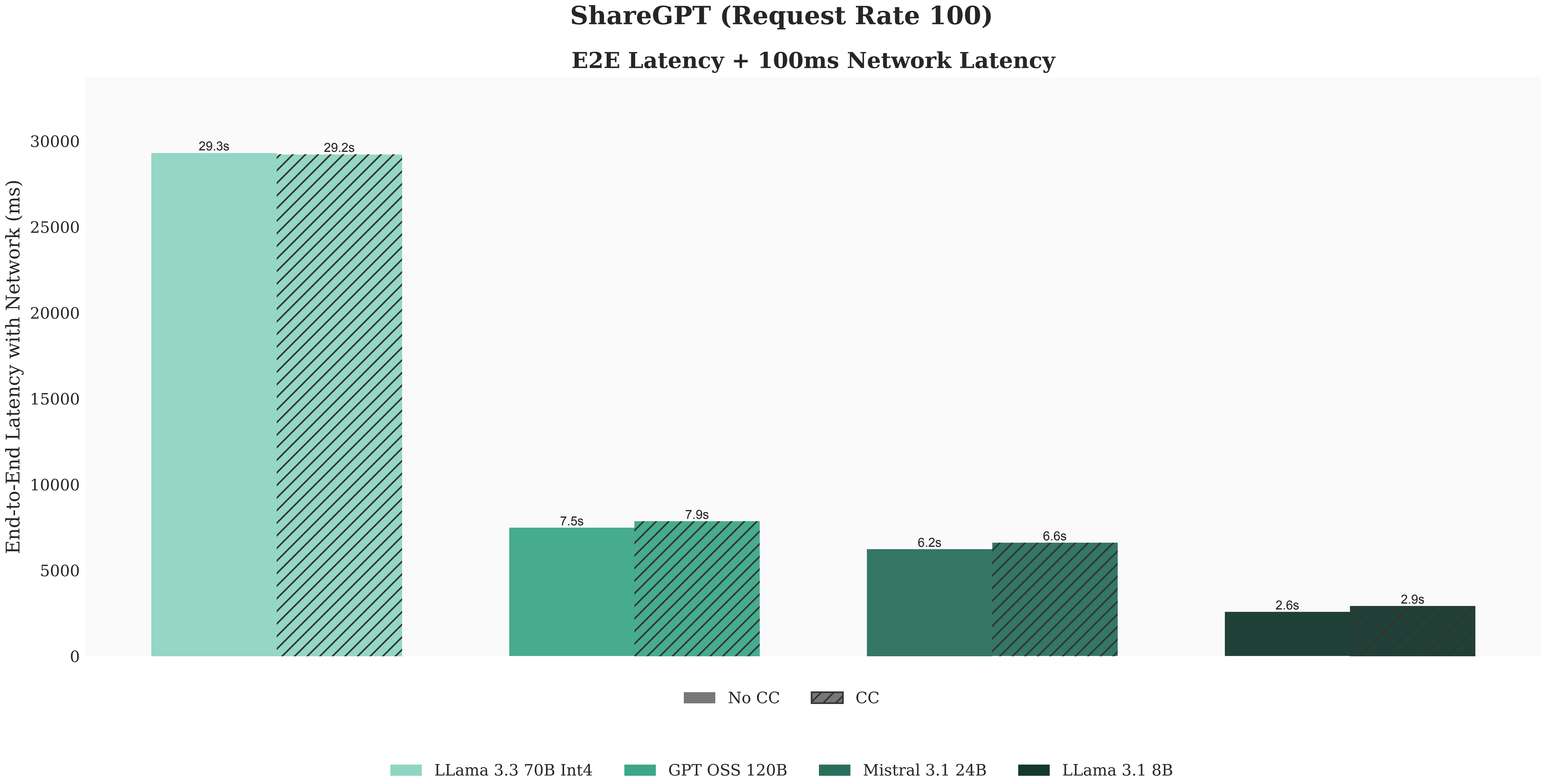
Random (1000 ⇒ 1000) (Request Rate 50)

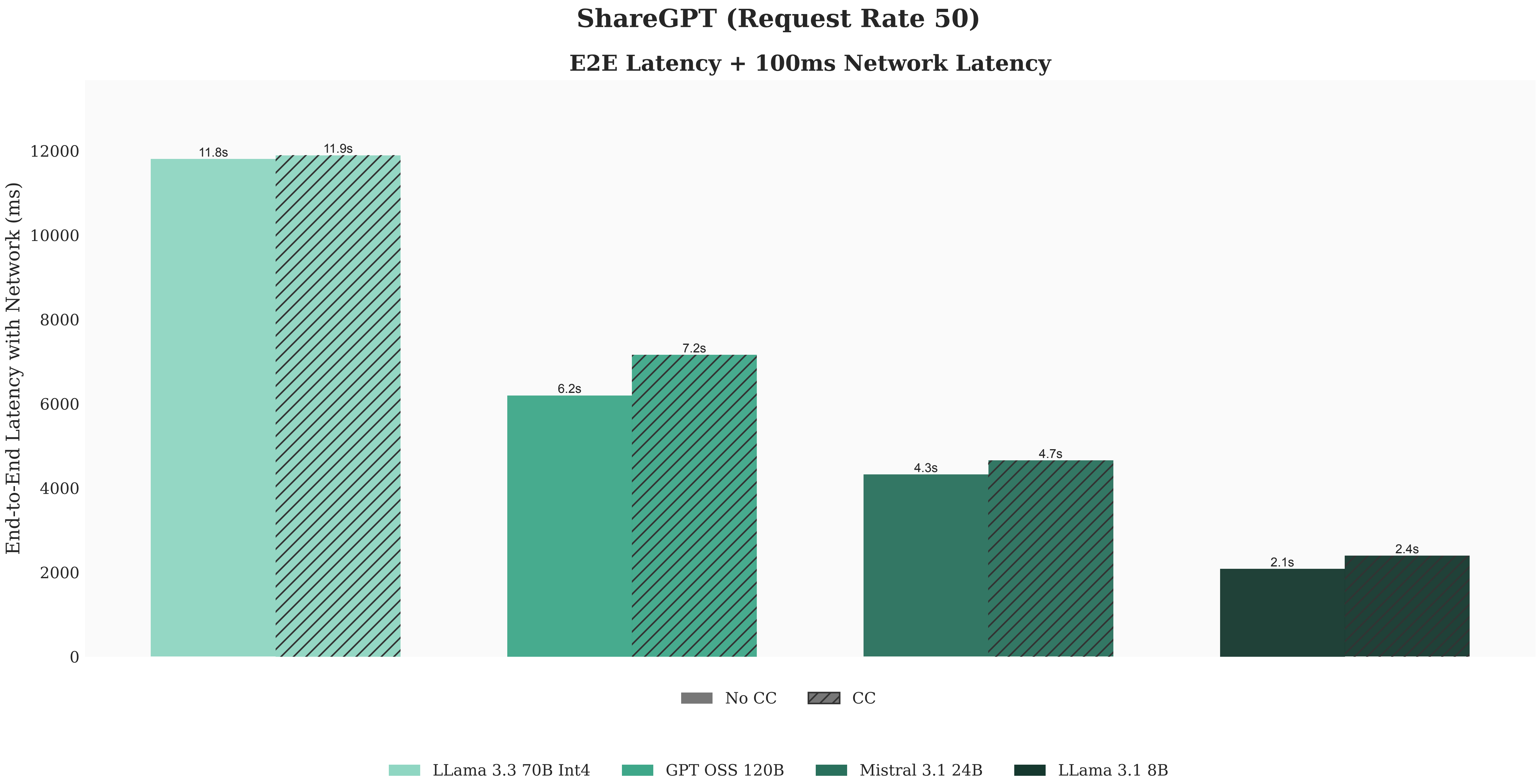
E2E Latency + 100ms Network Latency

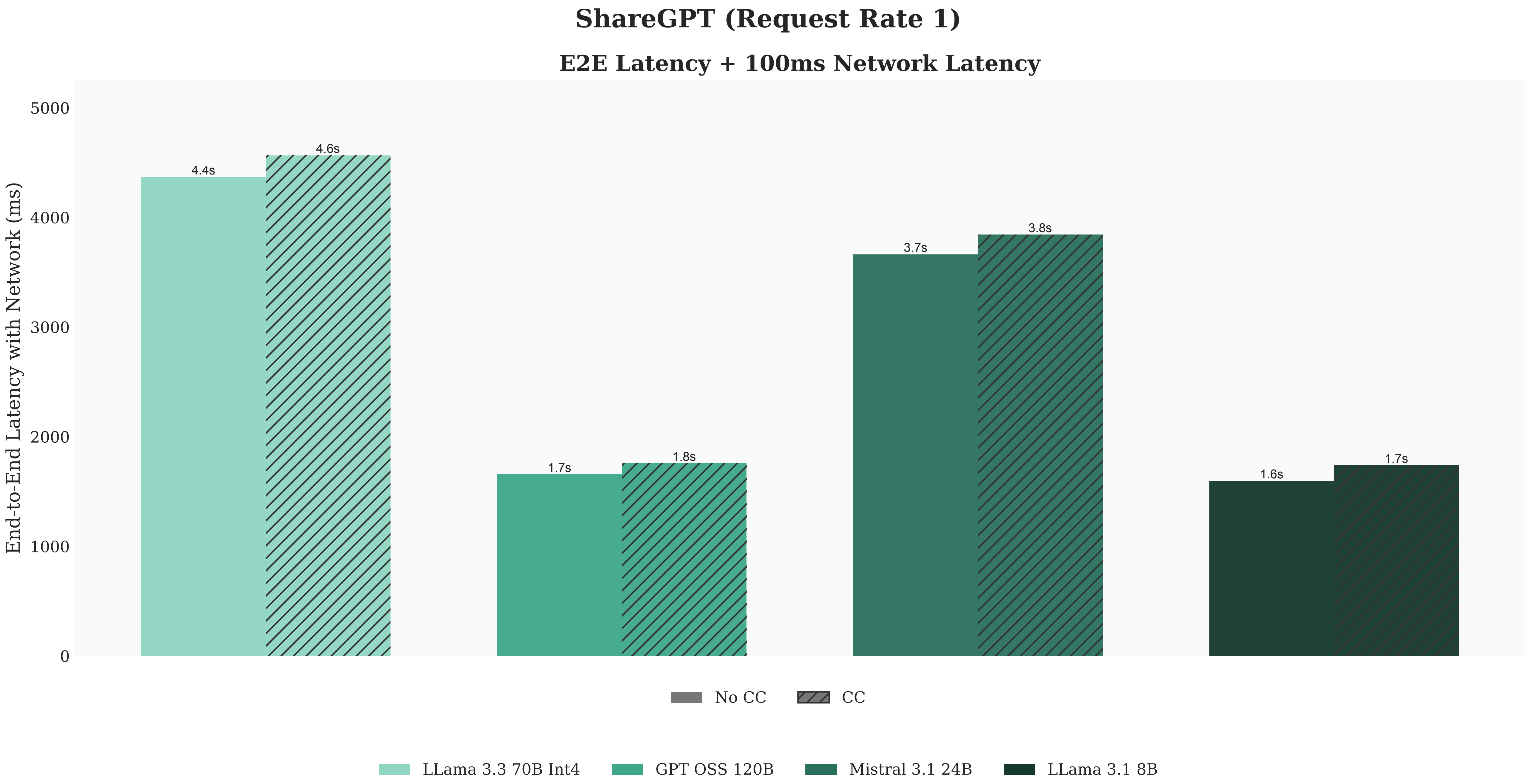


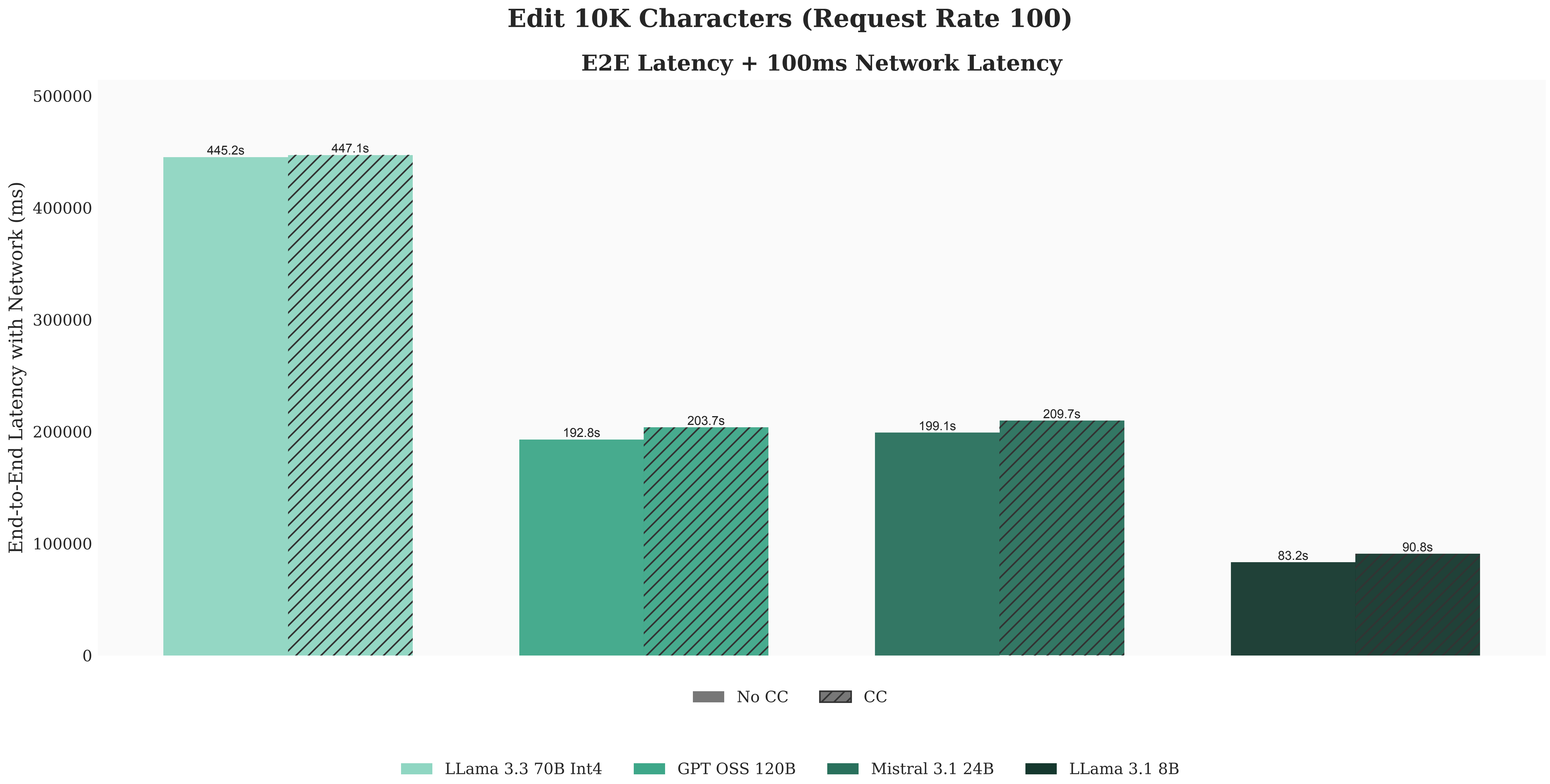
Random (1000 ⇒ 1000) (Request Rate 1)

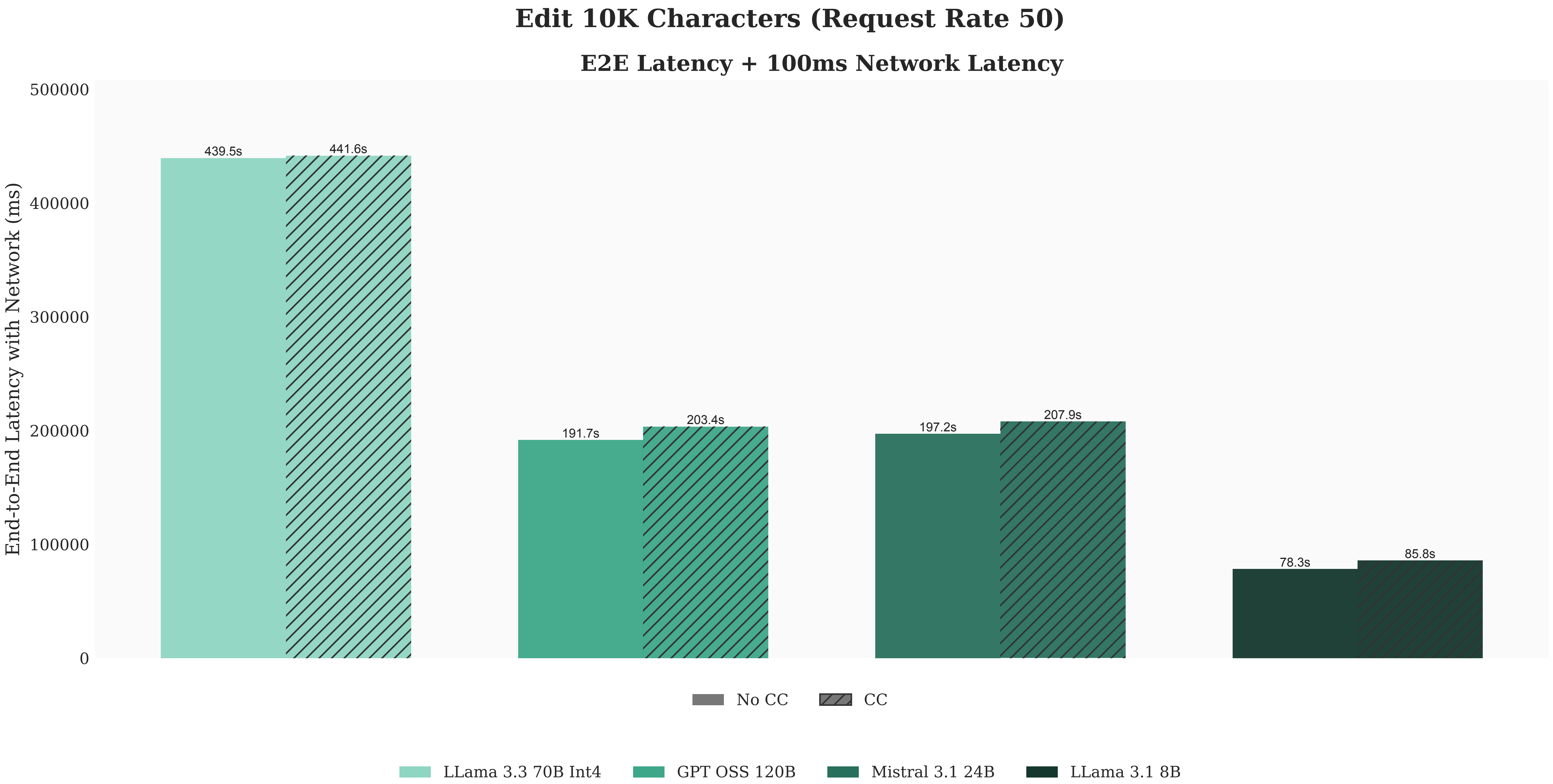
E2E Latency + 100ms Network Latency

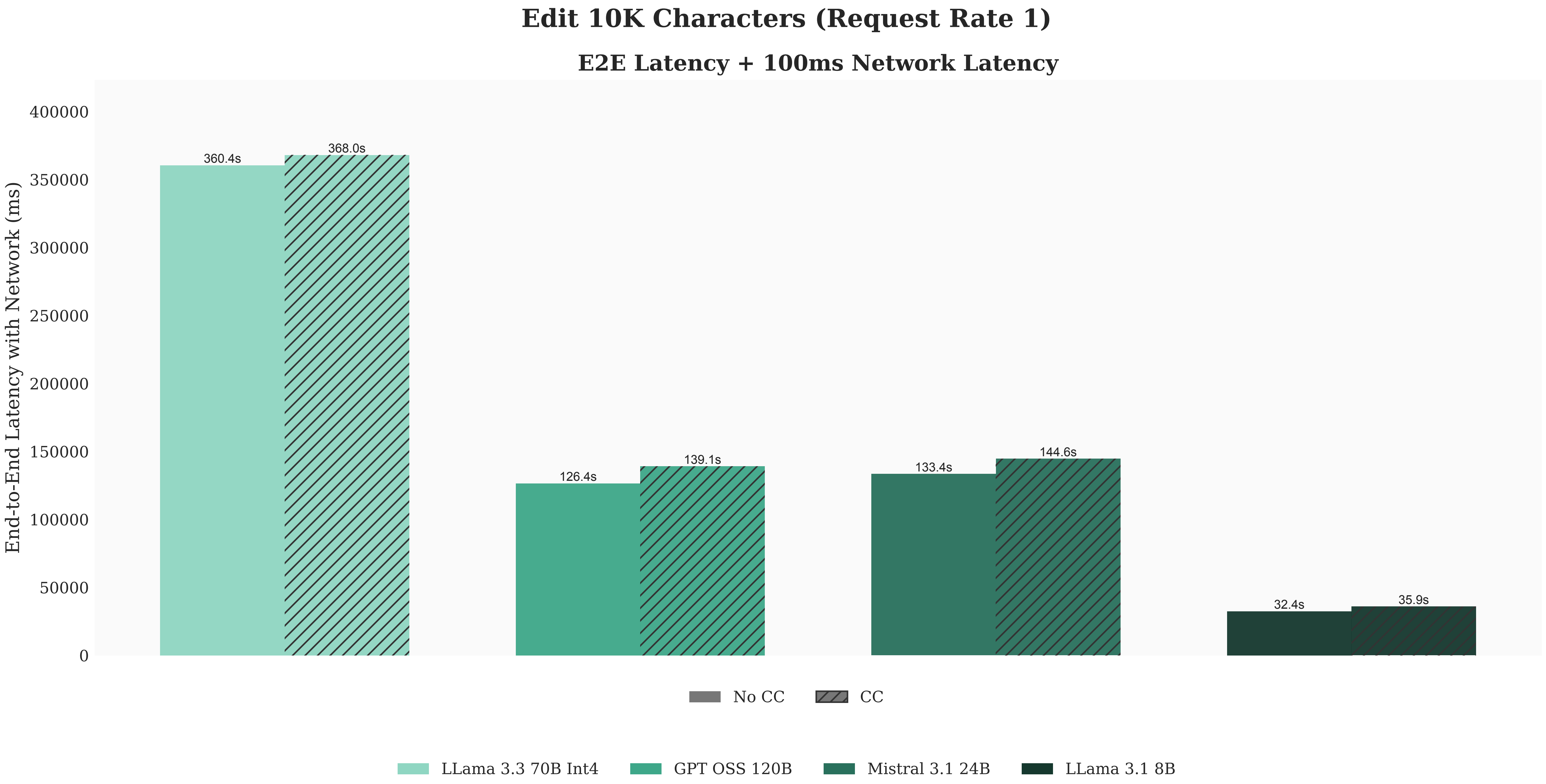

ShareGPT (Request Rate 100)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
30000
25000
20000
15000
10000
5000
0
29.3s
29.2s
7.5s
7.9s
6.2s
6.6s
2.6s
2.9s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

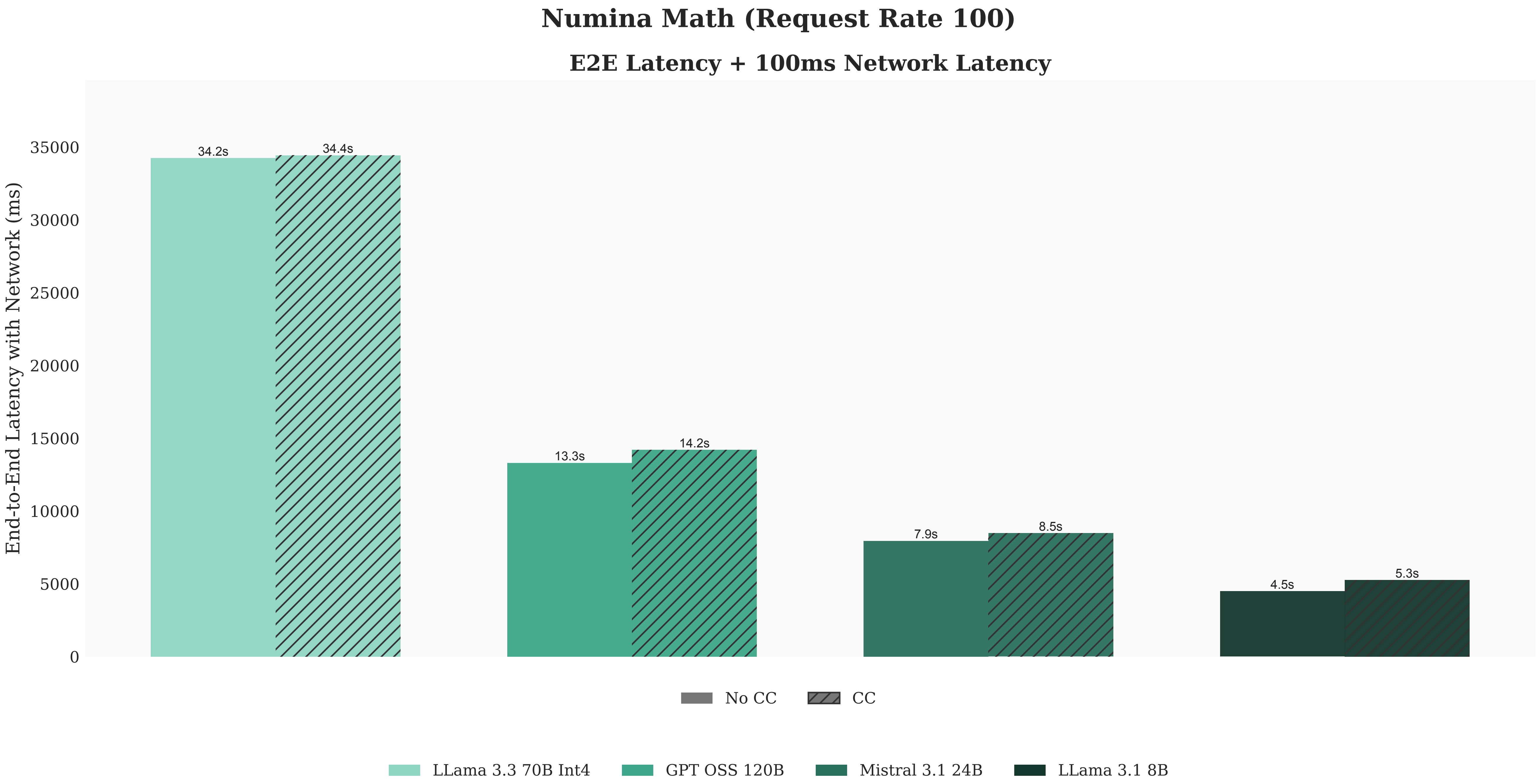

ShareGPT (Request Rate 50)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
2000
4000
6000
8000
10000
12000
11.8s
11.9s
6.2s
7.2s
4.3s
4.7s
2.1s
2.4s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

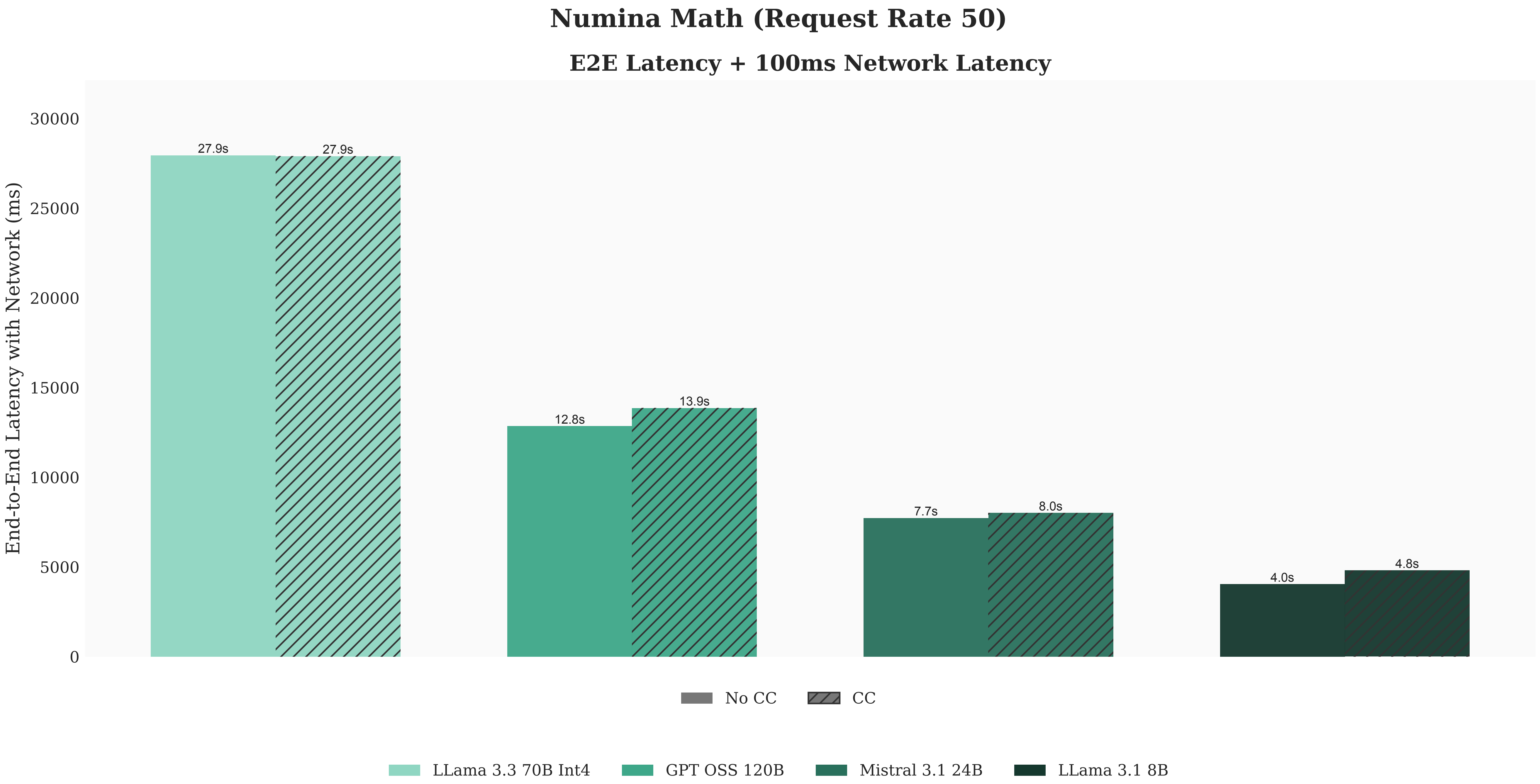

ShareGPT (Request Rate 1)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
5000
4000
3000
2000
1000
0
4.4s
4.6s
1.7s
1.8s
3.7s
3.8s
1.6s
1.7s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

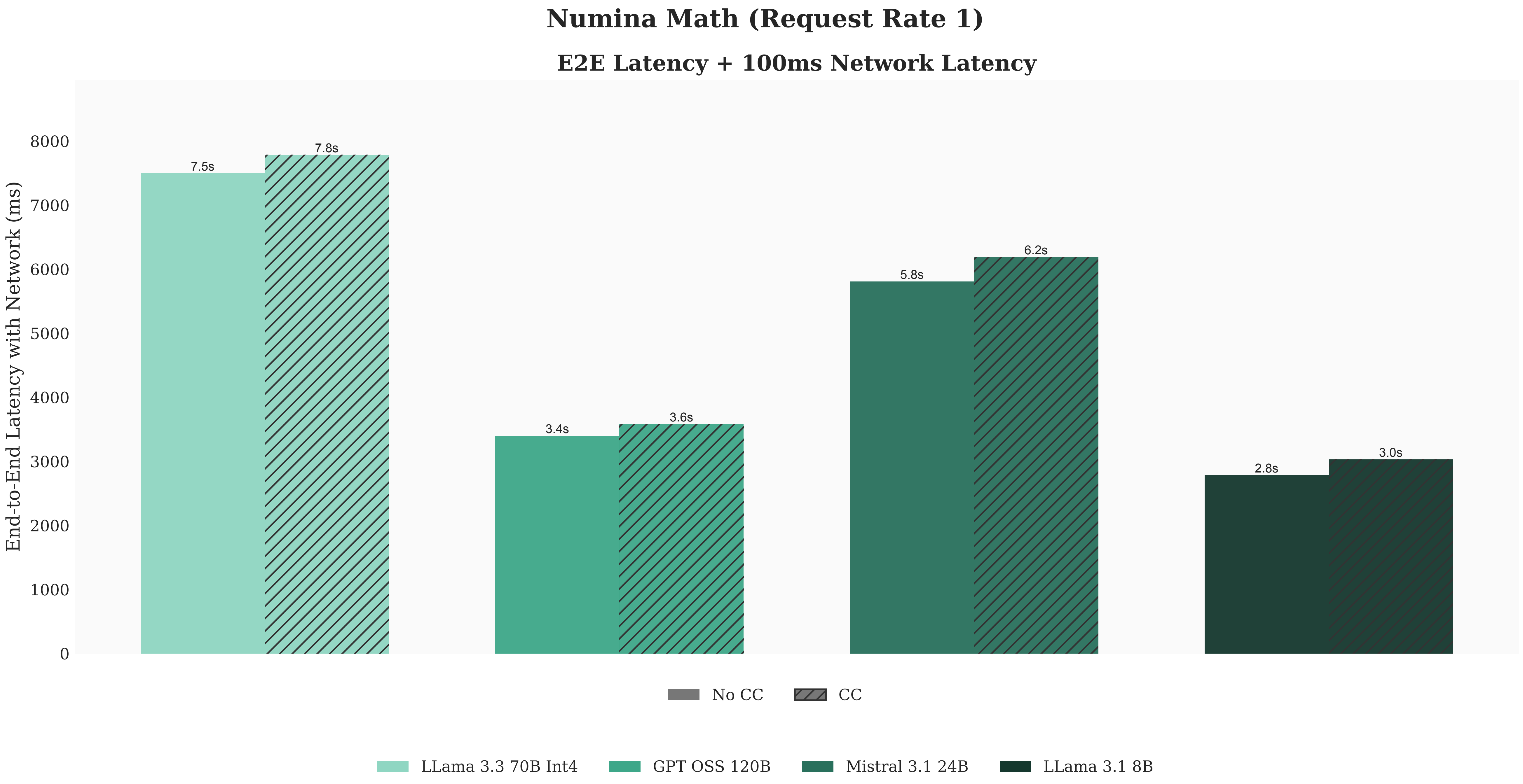

Edit 10K Characters (Request Rate 100)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
500000
400000
300000
200000
100000
0
445.2s
447.1s
192.8s
203.7s
199.1s
209.7s
83.2s
90.8s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


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


Edit 10K Characters (Request Rate 1)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
50000
100000
150000
200000
250000
300000
350000
400000
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 (Request Rate 100)
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 (Request Rate 50)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
30000
25000
20000
15000
10000
5000
0
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 (Request Rate 1)
E2E Latency + 100ms Network Latency
End-to-End Latency with Network (ms)
0
1000
2000
3000
4000
5000
6000
7000
8000
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