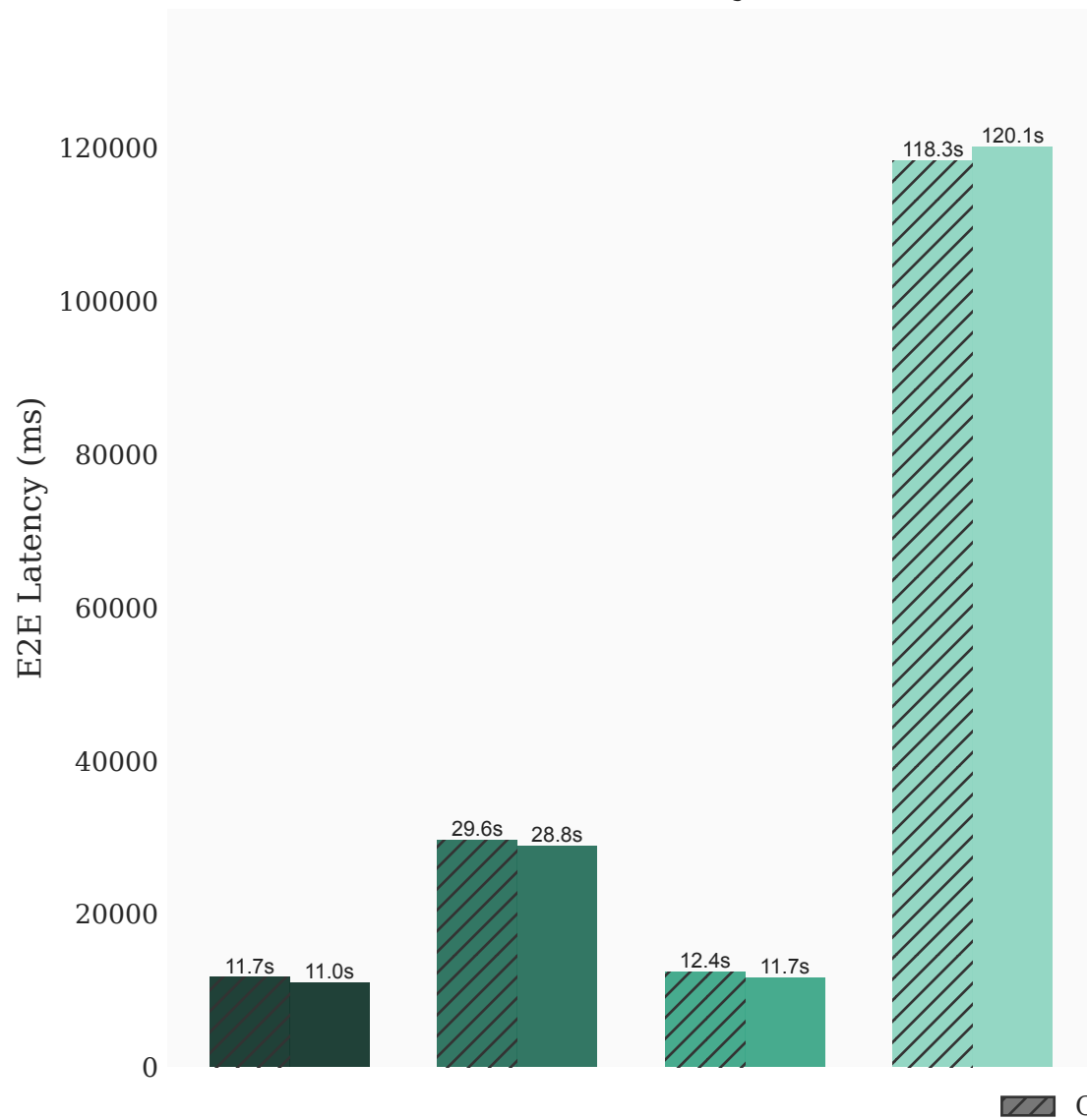# Random (1500 ⇒ 250) (100 Concurrent Requests)

End-to-End Latency (Mean)

E2E Latency (ms)

- 11.7s / 11.0s — LLama 3.1 8B
- 29.6s / 28.8s — Mistral 3.1 24B
- 12.4s / 11.7s — GPT OSS 120B
- 118.3s / 120.1s — LLama 3.3 70B Int4

End-to-End Latency (P99)

E2E Latency (ms)

- 13.0s / 12.3s — LLama 3.1 8B
- 35.9s / 34.8s — Mistral 3.1 24B
- 18.4s / 17.5s — GPT OSS 120B
- 162.6s / 164.5s — LLama 3.3 70B Int4

CC   No CC

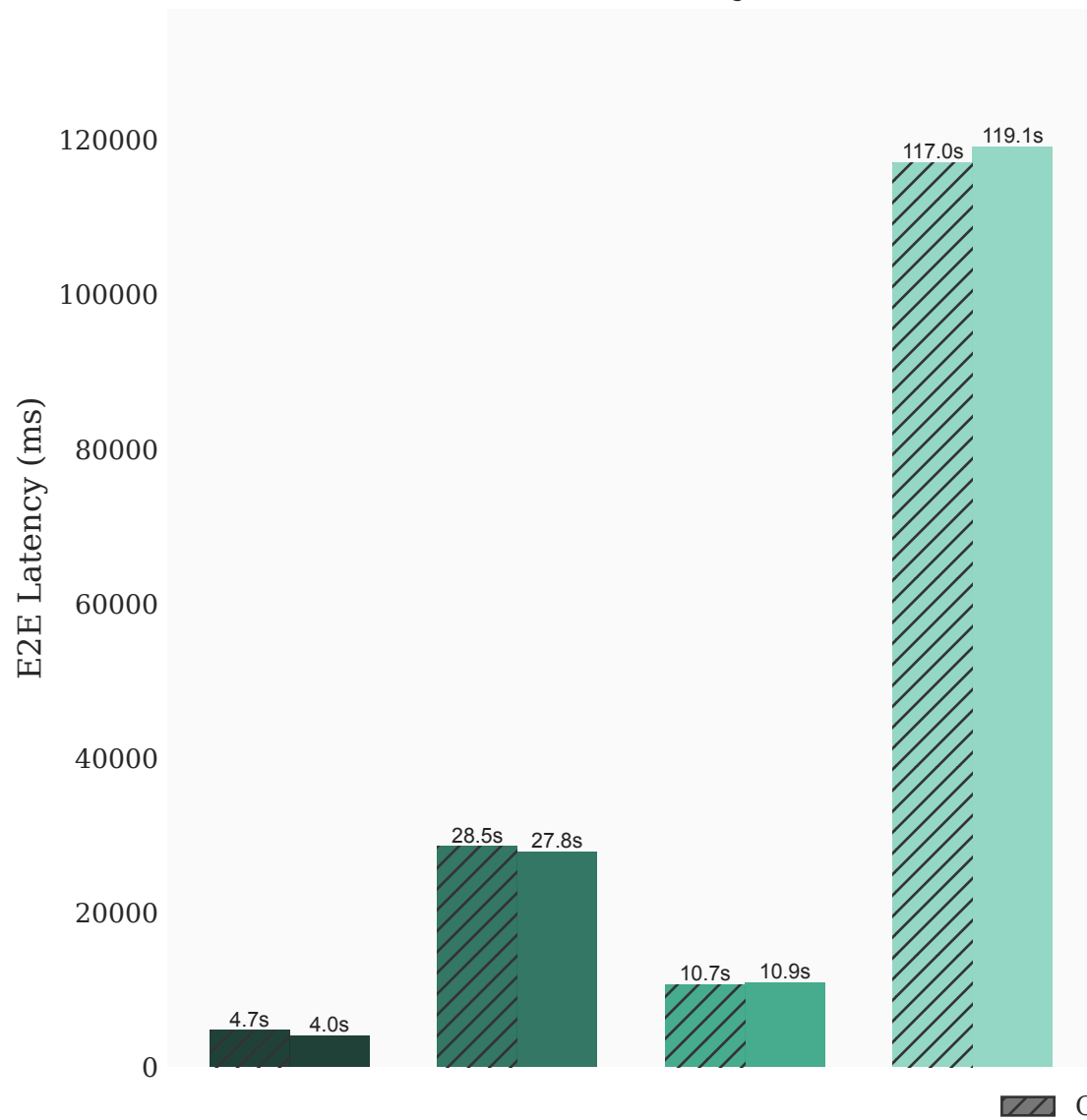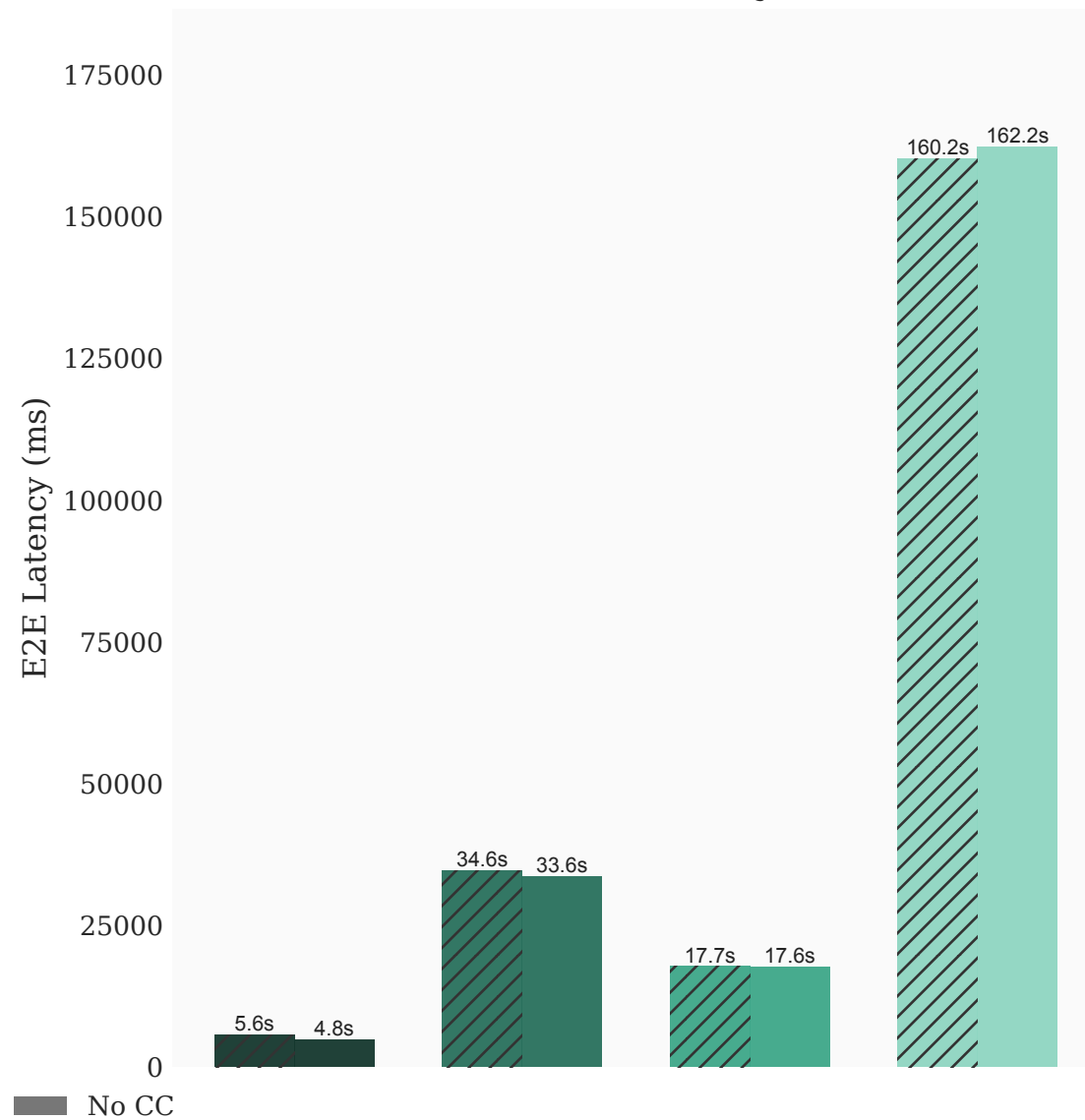LLama 3.1 8B   Mistral 3.1 24B   GPT OSS 120B   LLama 3.3 70B Int4

Random (1500 ⇒ 250) (50 Concurrent Requests)
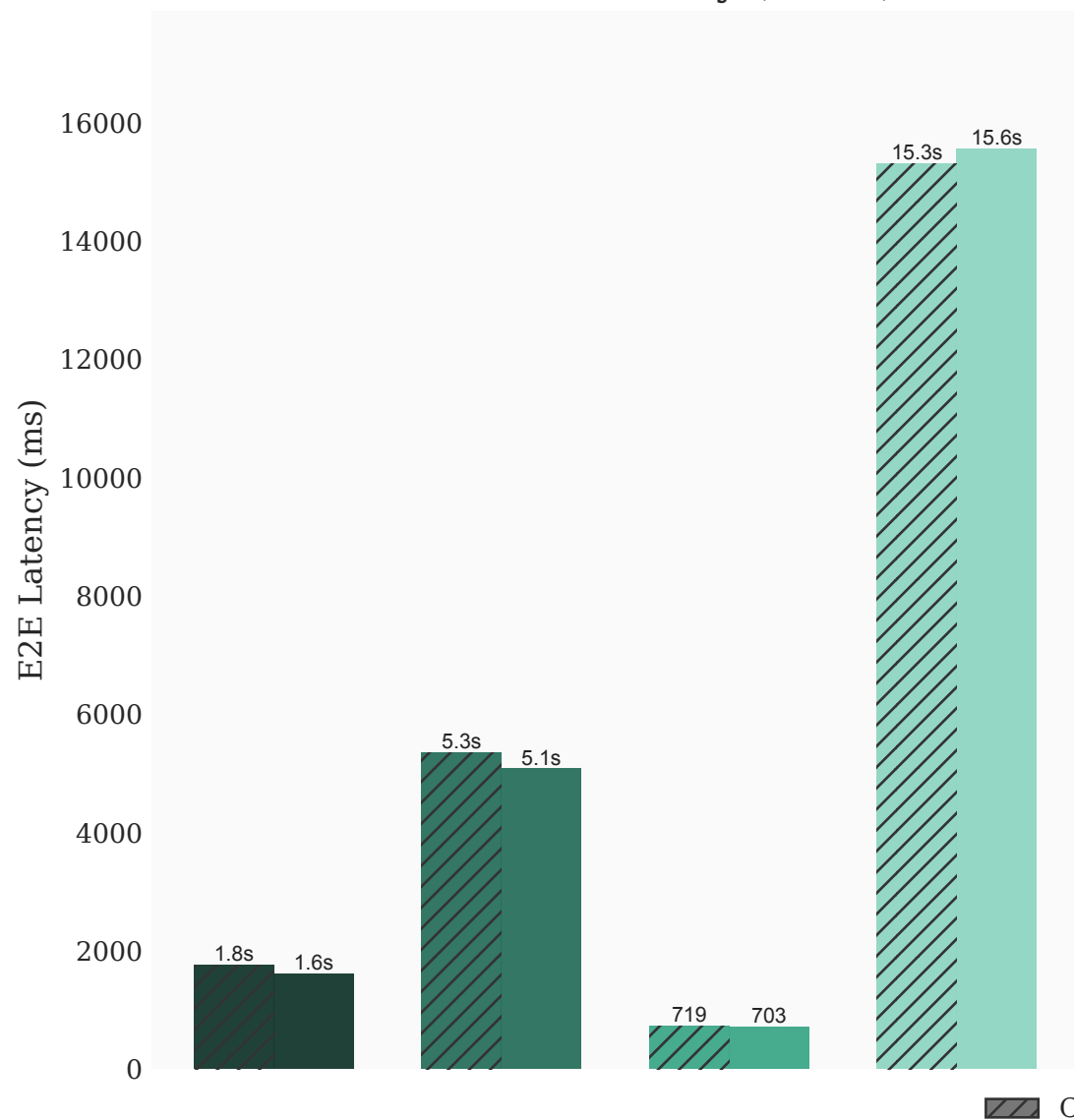
End-to-End Latency (Mean)

End-to-End Latency (P99)

Legend: CC | No CC
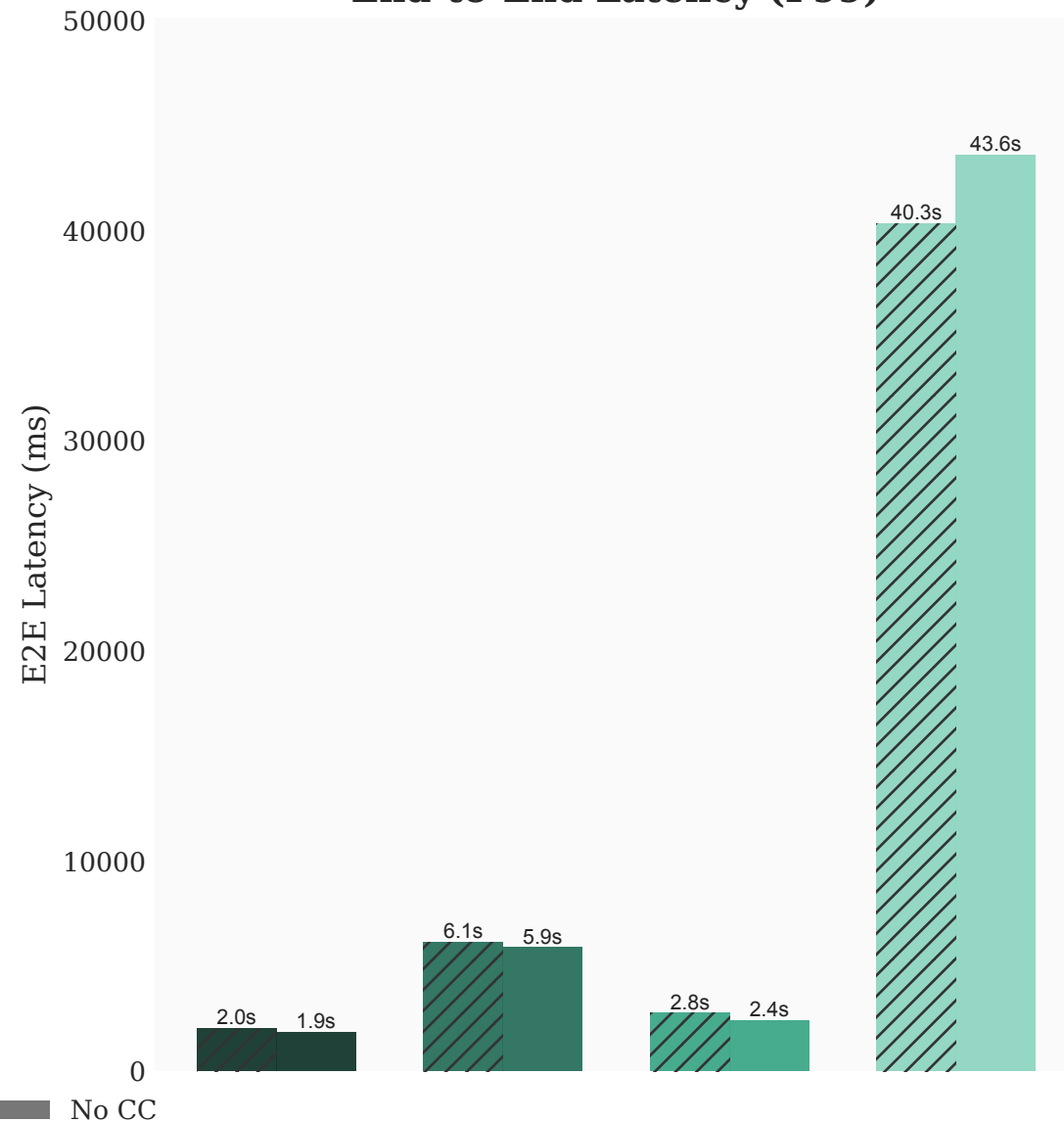
LLama 3.1 8B | Mistral 3.1 24B | GPT OSS 120B | LLama 3.3 70B Int4

Random (1500 ⇒ 250) (1 Concurrent Requests)

End-to-End Latency (Mean)

E2E Latency (ms)

- 1.8s / 1.6s
- 5.3s / 5.1s
- 719 / 703
- 15.3s / 15.6s

End-to-End Latency (P99)

E2E Latency (ms)

- 2.0s / 1.9s
- 6.1s / 5.9s
- 2.8s / 2.4s
- 40.3s / 43.6s

CC   No CC

LLama 3.1 8B   Mistral 3.1 24B   GPT OSS 120B   LLama 3.3 70B Int4

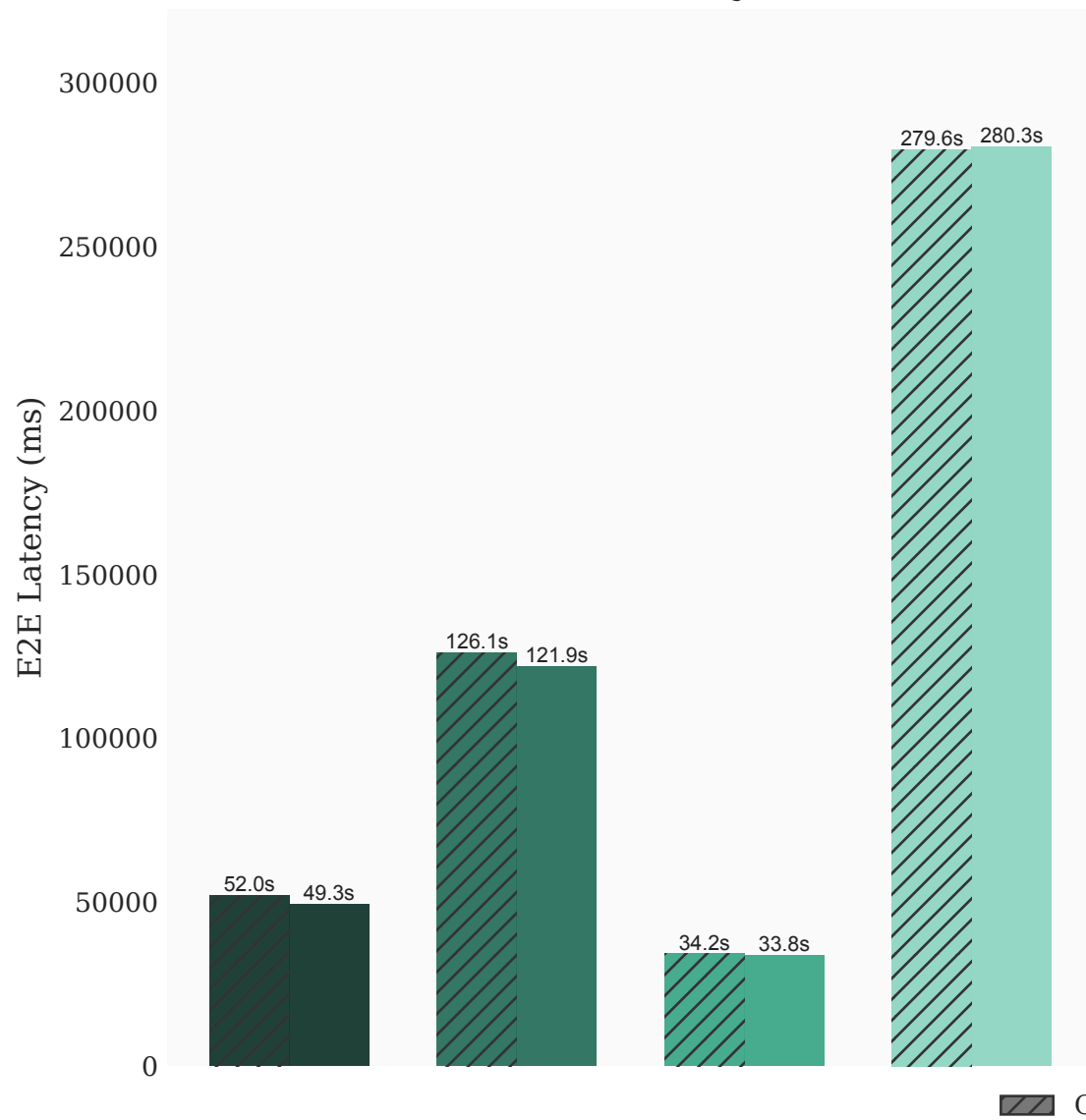# Random (4000 ⇒ 1000) (100 Concurrent Requests)

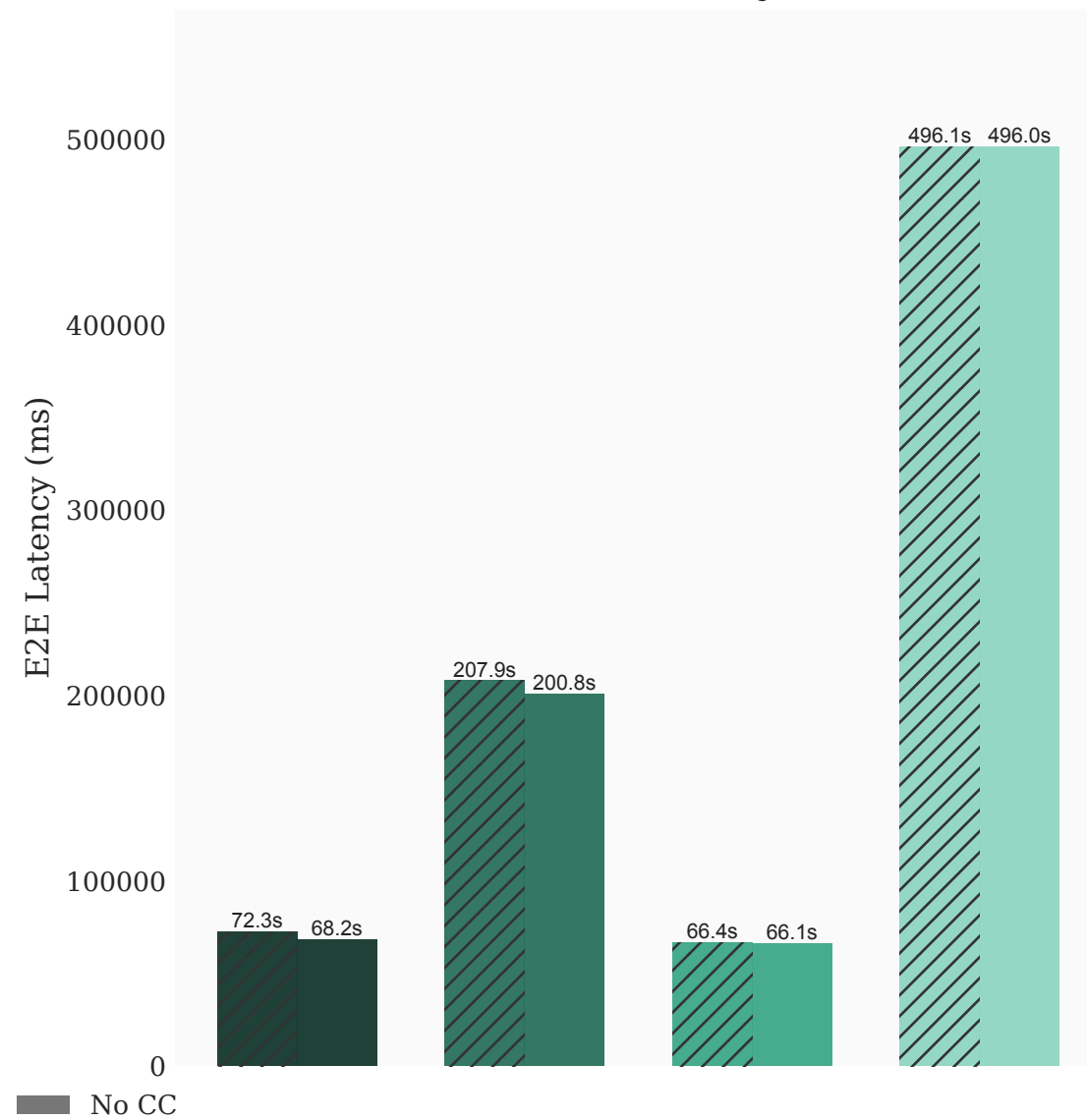## End-to-End Latency (Mean)

End-to-End Latency (P99)

E2E Latency (ms)

Mean values: 52.0s / 49.3s (LLama 3.1 8B), 126.1s / 121.9s (Mistral 3.1 24B), 34.2s / 33.8s (GPT OSS 120B), 279.6s / 280.3s (LLama 3.3 70B Int4)

P99 values: 72.3s / 68.2s (LLama 3.1 8B), 207.9s / 200.8s (Mistral 3.1 24B), 66.4s / 66.1s (GPT OSS 120B), 496.1s / 496.0s (LLama 3.3 70B Int4)

CC   No CC

LLama 3.1 8B   Mistral 3.1 24B   GPT OSS 120B   LLama 3.3 70B Int4

Random (4000 ⇒ 1000) (50 Concurrent Requests)

End-to-End Latency (Mean)

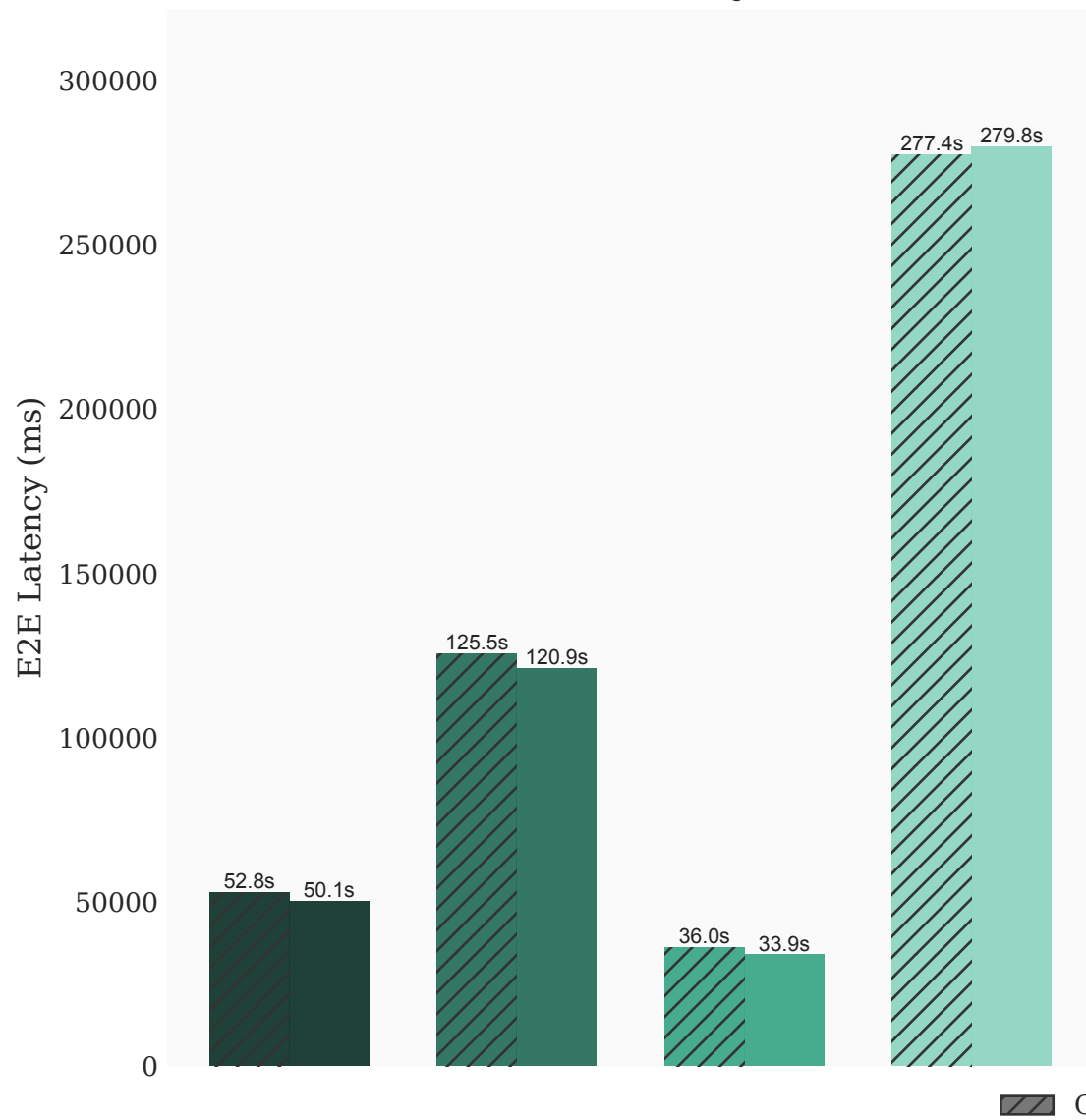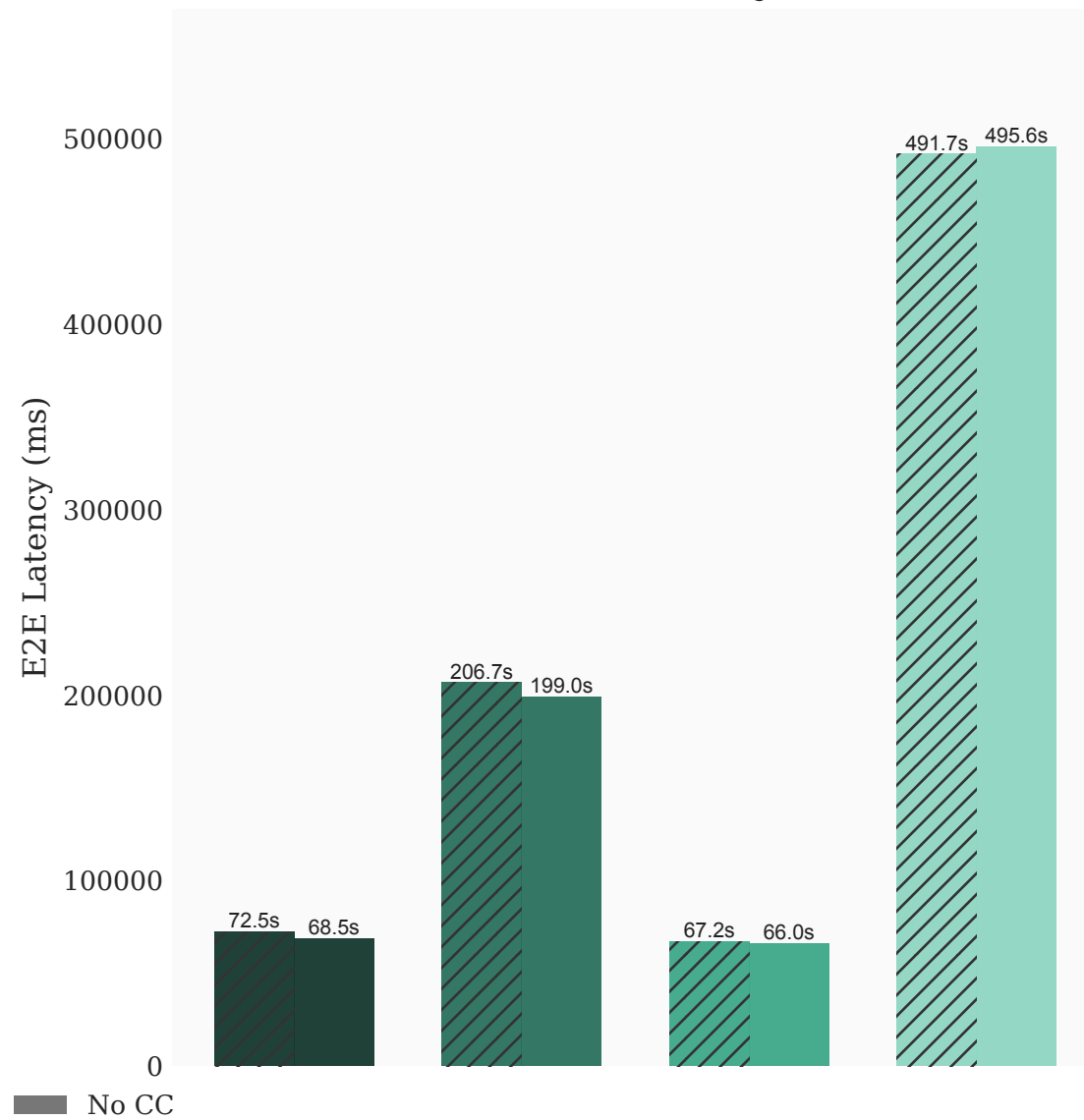## End-to-End Latency (P99)

	CC	No CC

- LLama 3.1 8B
- Mistral 3.1 24B
- GPT OSS 120B
- LLama 3.3 70B Int4

Mean values: 52.8s, 50.1s, 125.5s, 120.9s, 36.0s, 33.9s, 277.4s, 279.8s

P99 values: 72.5s, 68.5s, 206.7s, 199.0s, 67.2s, 66.0s, 491.7s, 495.6s

Random (4000 ⇒ 1000) (1 Concurrent Requests)

End-to-End Latency (Mean)

End-to-End Latency (P99)

Mean chart values:
- 9.9s / 8.9s
- 39.9s / 35.6s
- 3.6s / 2.9s
- 180.1s / 182.4s

P99 chart values:
- 12.9s / 11.6s
- 51.7s / 49.0s
- 14.7s / 12.9s
- 317.8s / 321.5s

Legend: CC, No CC

LLama 3.1 8B   Mistral 3.1 24B   GPT OSS 120B   LLama 3.3 70B Int4

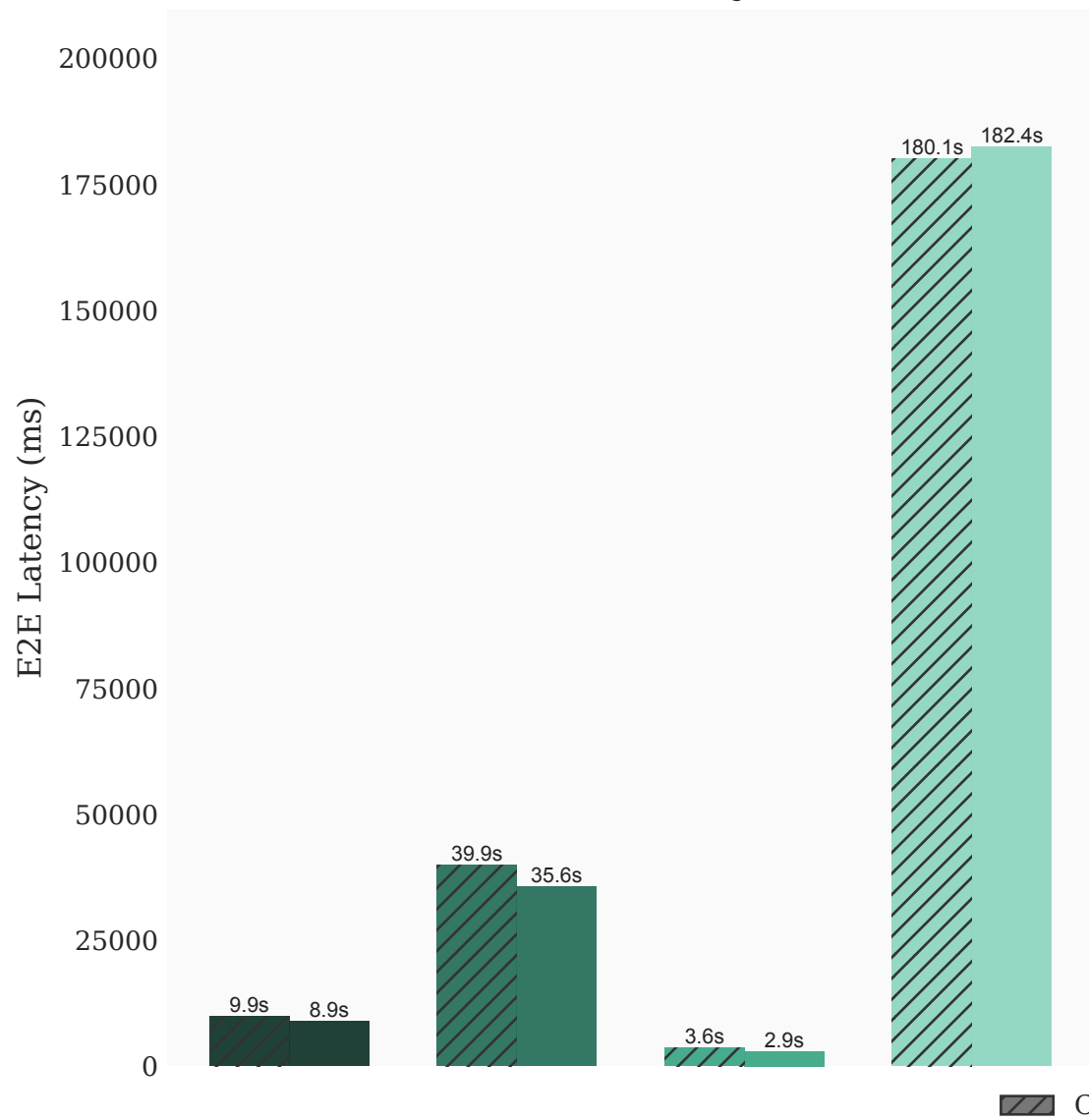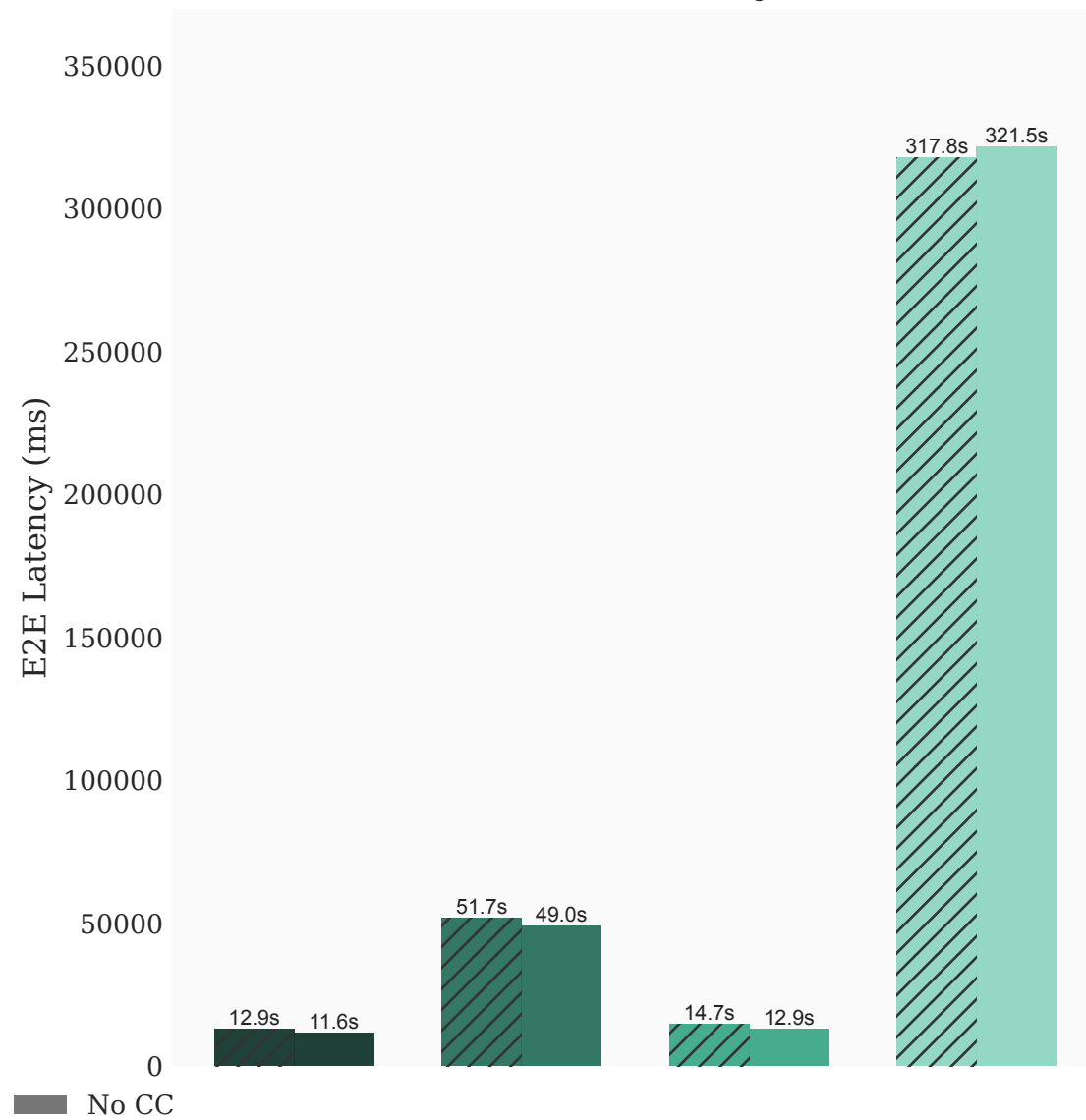# Random (1000 ⇒ 1000) (100 Concurrent Requests)

End-to-End Latency (Mean)

End-to-End Latency (P99)
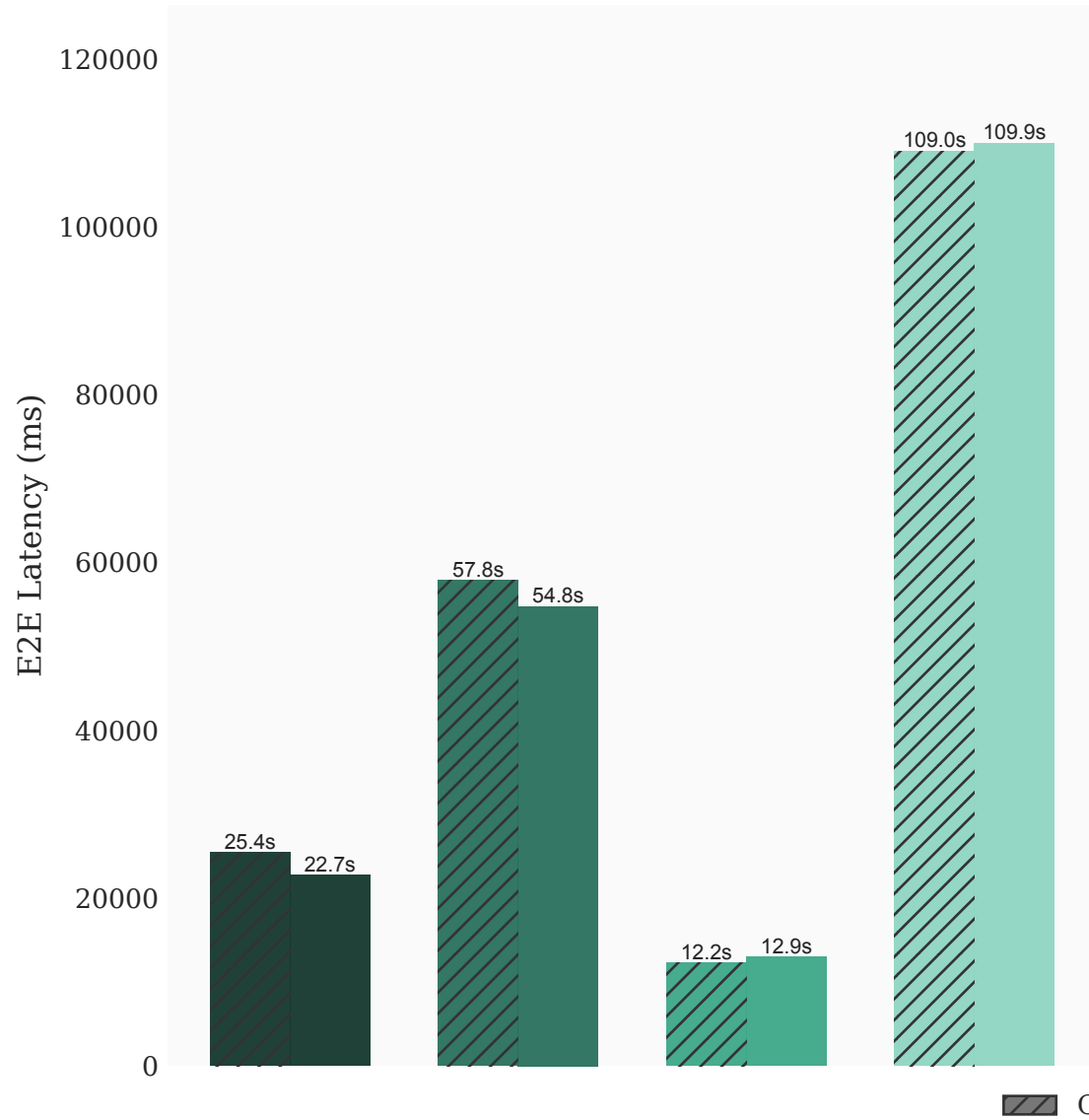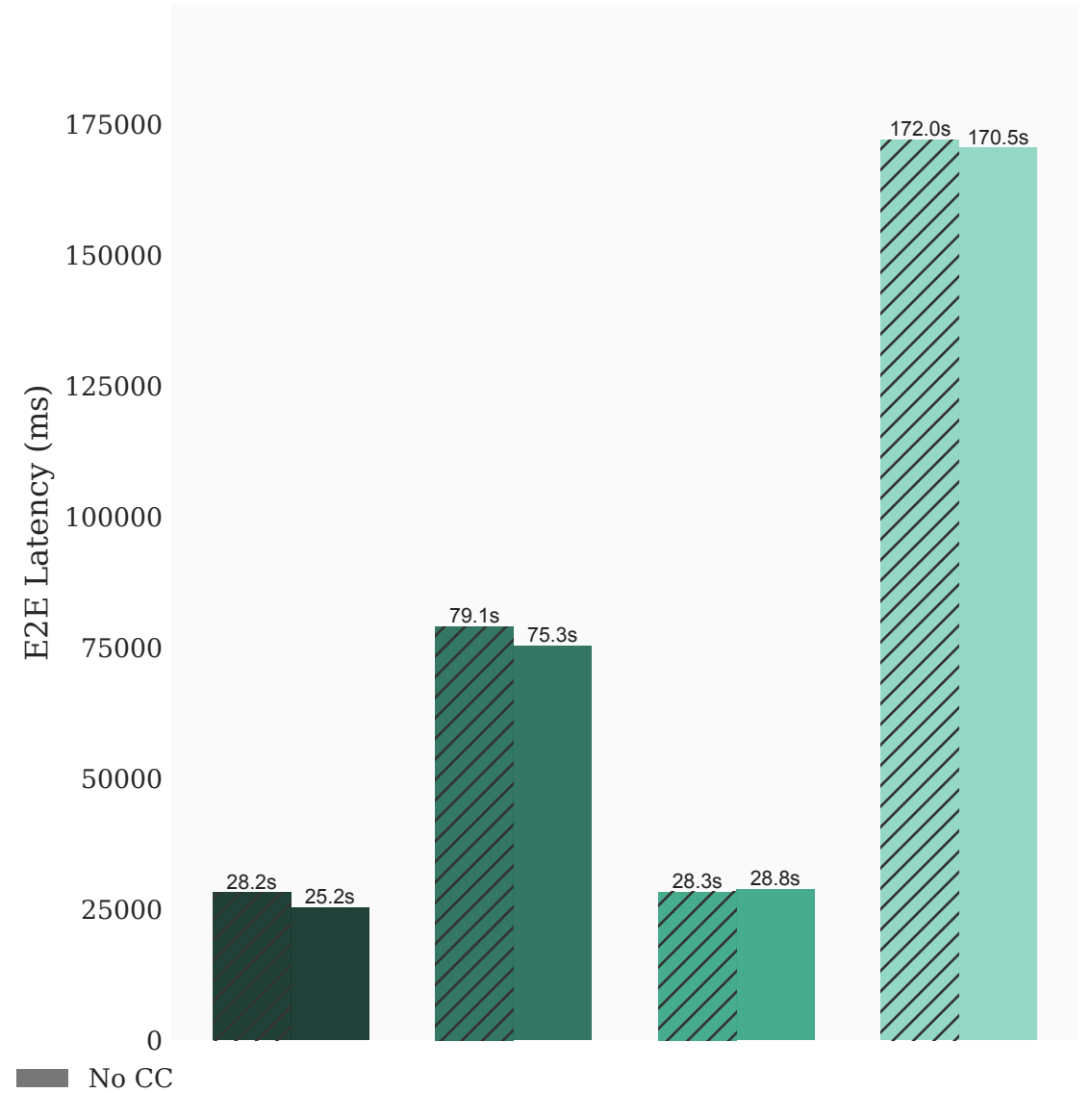

Legend: CC, No CC

LLama 3.1 8B, Mistral 3.1 24B, GPT OSS 120B, LLama 3.3 70B Int4

Random (1000 ⇒ 1000) (50 Concurrent Requests)

End-to-End Latency (Mean)

End-to-End Latency (P99)

E2E Latency (ms)

Mean values: 21.8s, 19.4s, 57.0s, 54.2s, 12.4s, 12.1s, 108.7s, 108.9s

P99 values: 24.2s, 21.4s, 77.3s, 73.6s, 30.0s, 29.5s, 172.6s, 170.1s

Legend: CC, No CC

LLama 3.1 8B   Mistral 3.1 24B   GPT OSS 120B   LLama 3.3 70B Int4

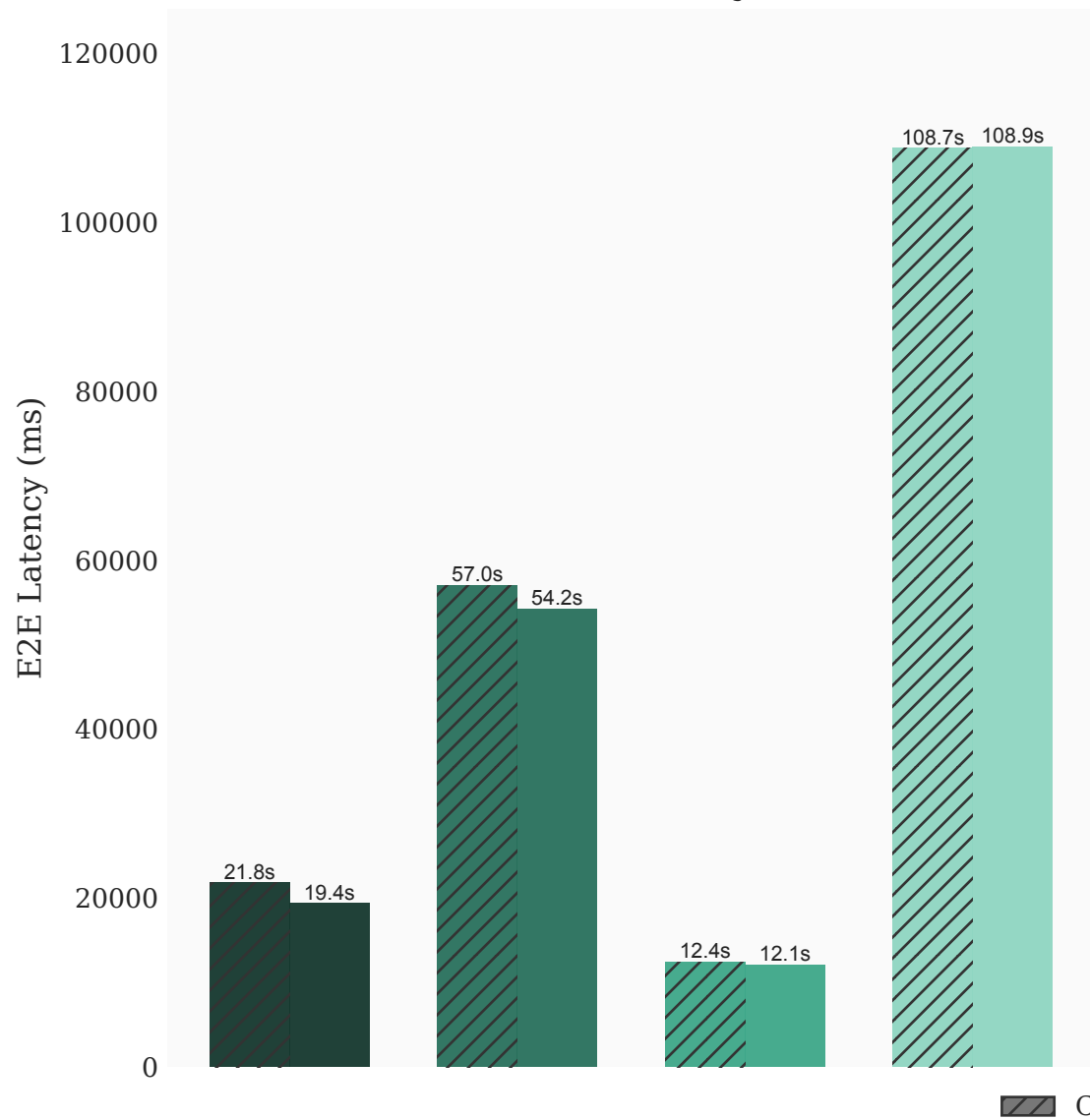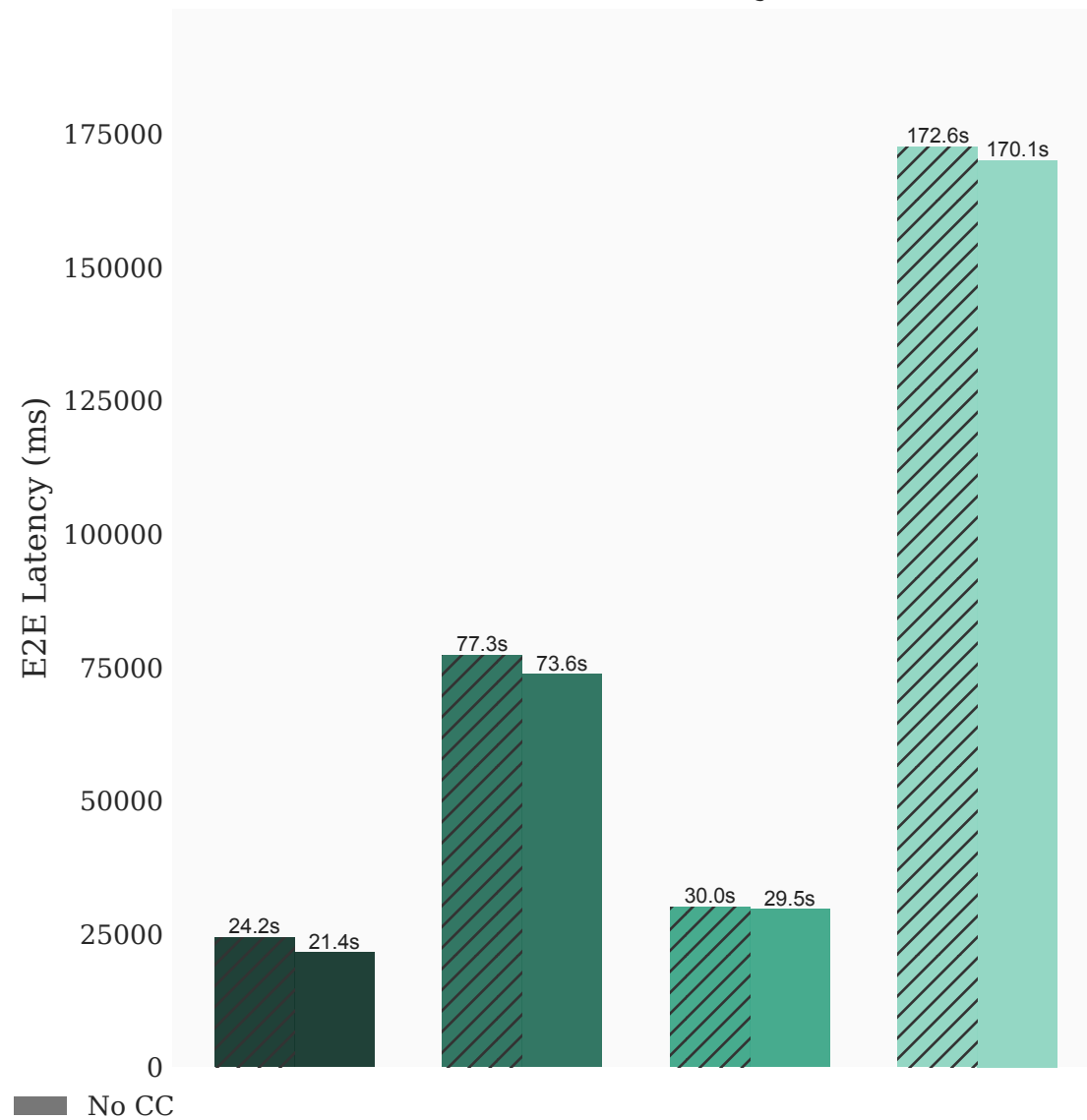# Random (1000 ⇒ 1000) (1 Concurrent Requests)

End-to-End Latency (Mean)

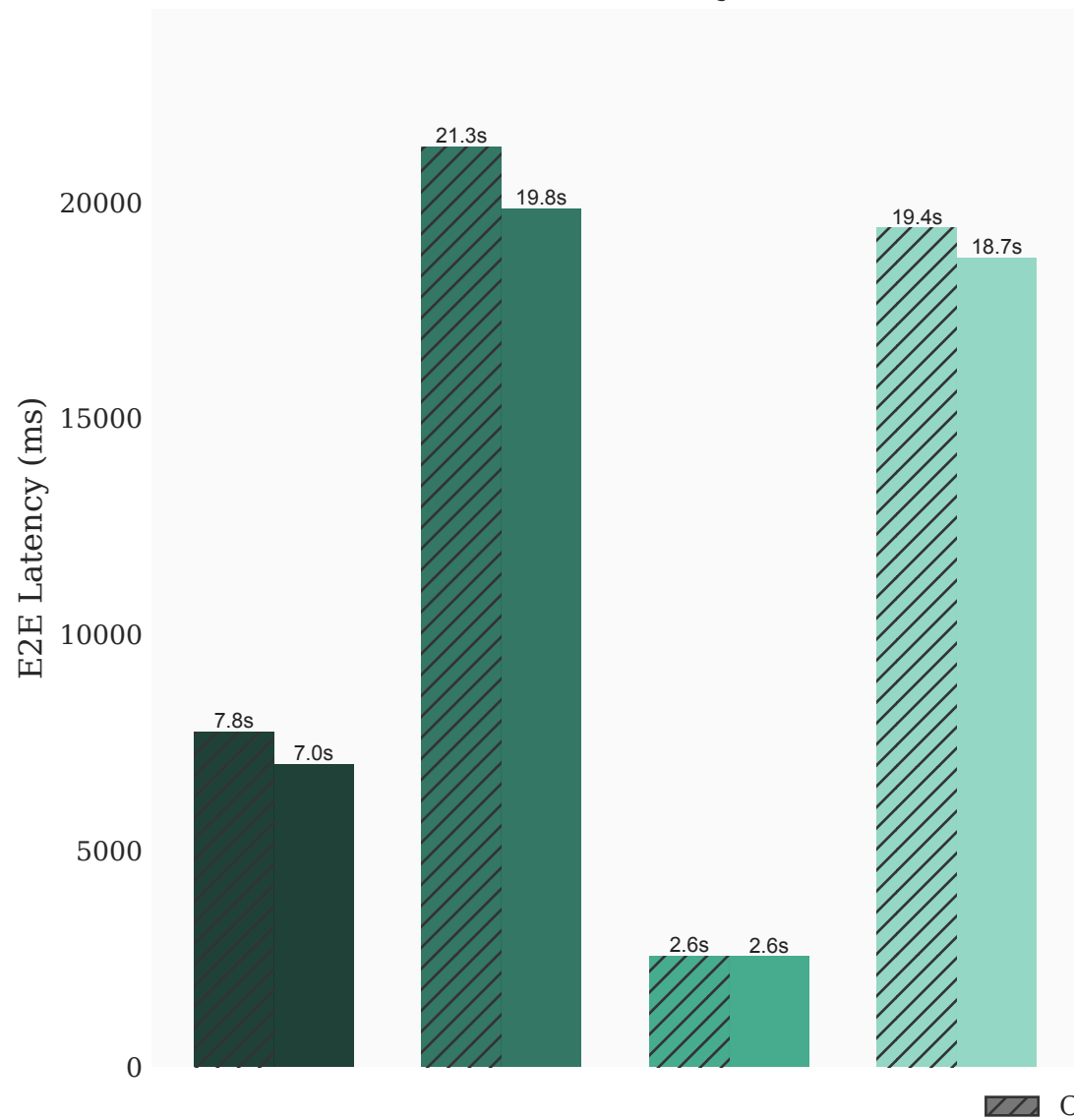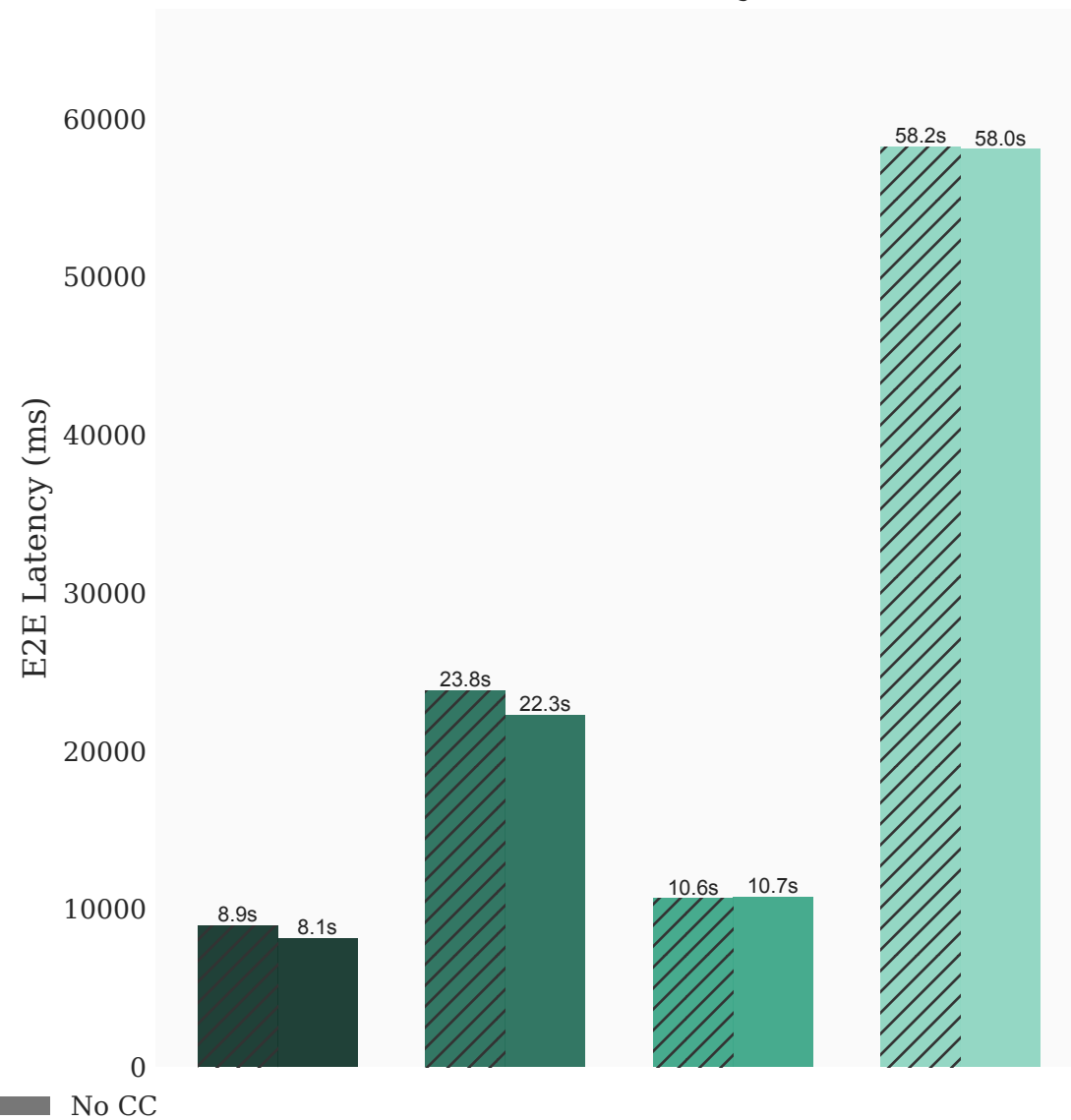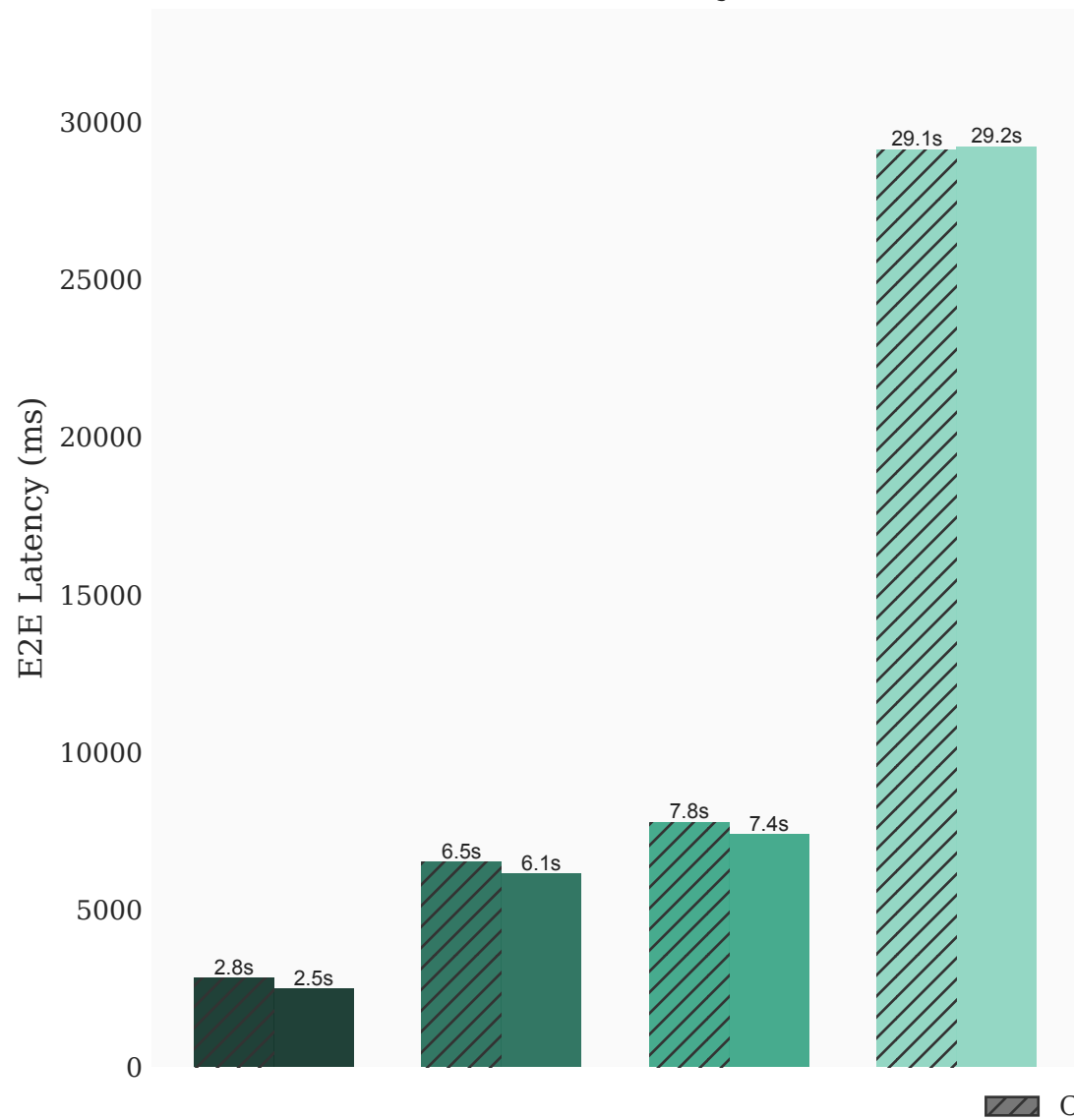E2E Latency (ms)

- LLama 3.1 8B: 7.8s (CC), 7.0s (No CC)
- Mistral 3.1 24B: 21.3s (CC), 19.8s (No CC)
- GPT OSS 120B: 2.6s (CC), 2.6s (No CC)
- LLama 3.3 70B Int4: 19.4s (CC), 18.7s (No CC)

End-to-End Latency (P99)

E2E Latency (ms)

- LLama 3.1 8B: 8.9s (CC), 8.1s (No CC)
- Mistral 3.1 24B: 23.8s (CC), 22.3s (No CC)
- GPT OSS 120B: 10.6s (CC), 10.7s (No CC)
- LLama 3.3 70B Int4: 58.2s (CC), 58.0s (No CC)

Legend: ⫽ CC | ▬ No CC

■ LLama 3.1 8B   ■ Mistral 3.1 24B   ■ GPT OSS 120B   ■ LLama 3.3 70B Int4

ShareGPT (100 Concurrent Requests)

End-to-End Latency (Mean)

End-to-End Latency (P99)

CC   No CC
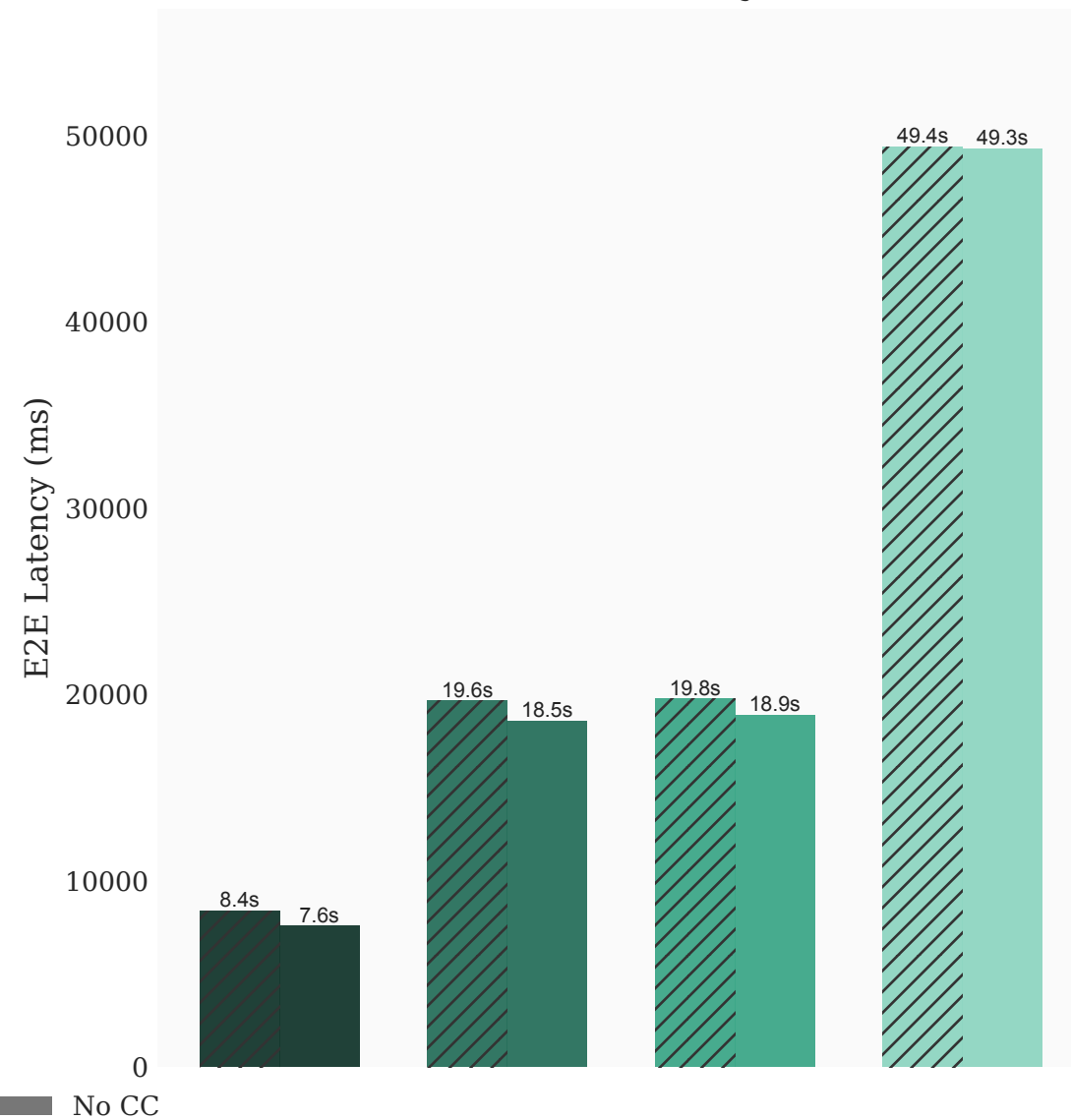
■ LLama 3.1 8B   ■ Mistral 3.1 24B   ■ GPT OSS 120B   ■ LLama 3.3 70B Int4

ShareGPT (50 Concurrent Requests)

End-to-End Latency (Mean)

End-to-End Latency (P99)

Legend: ▨ CC   ▬ No CC

LLama 3.1 8B   ■ Mistral 3.1 24B   ■ GPT OSS 120B   ■ LLama 3.3 70B Int4

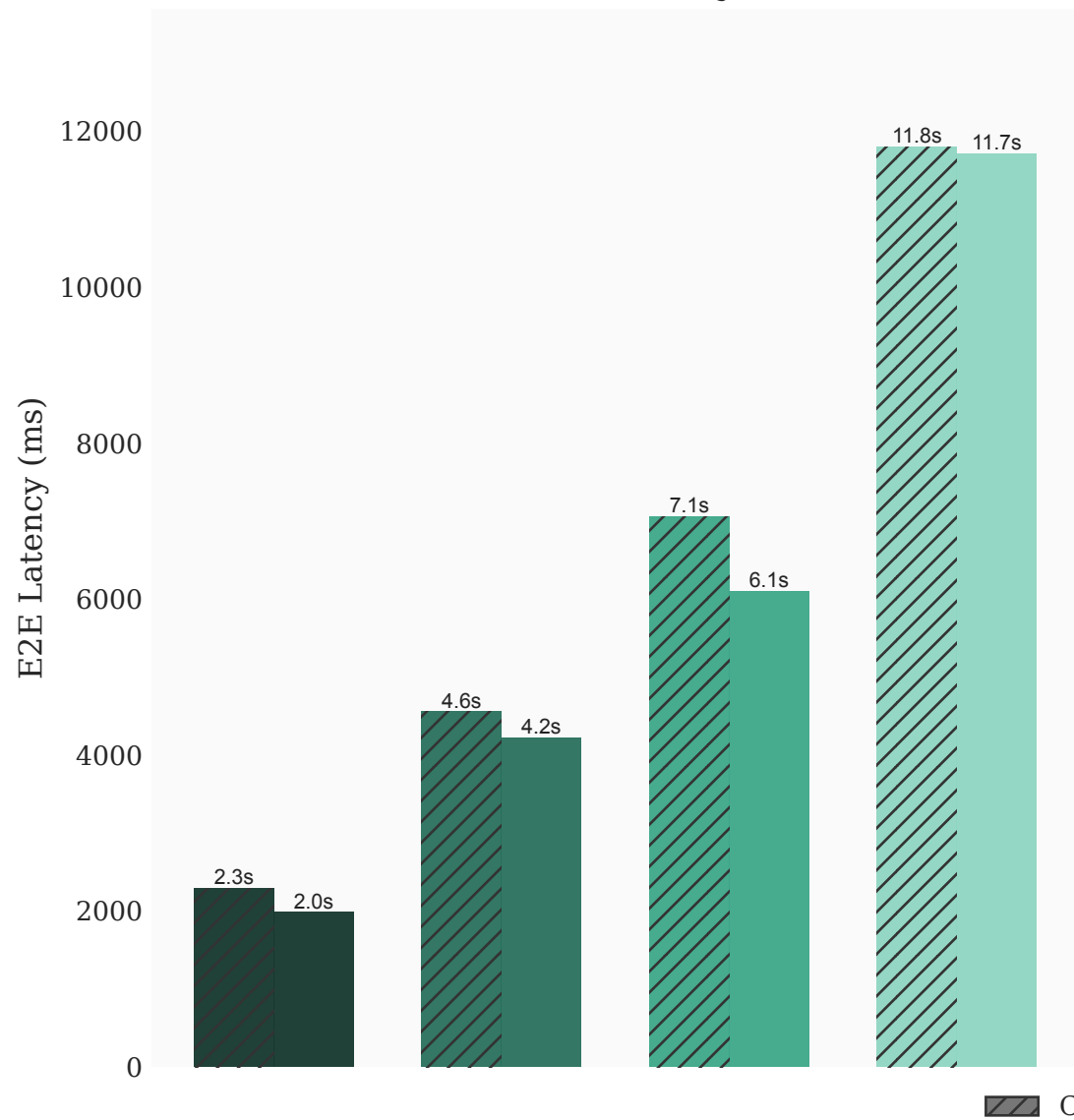# ShareGPT (1 Concurrent Requests)

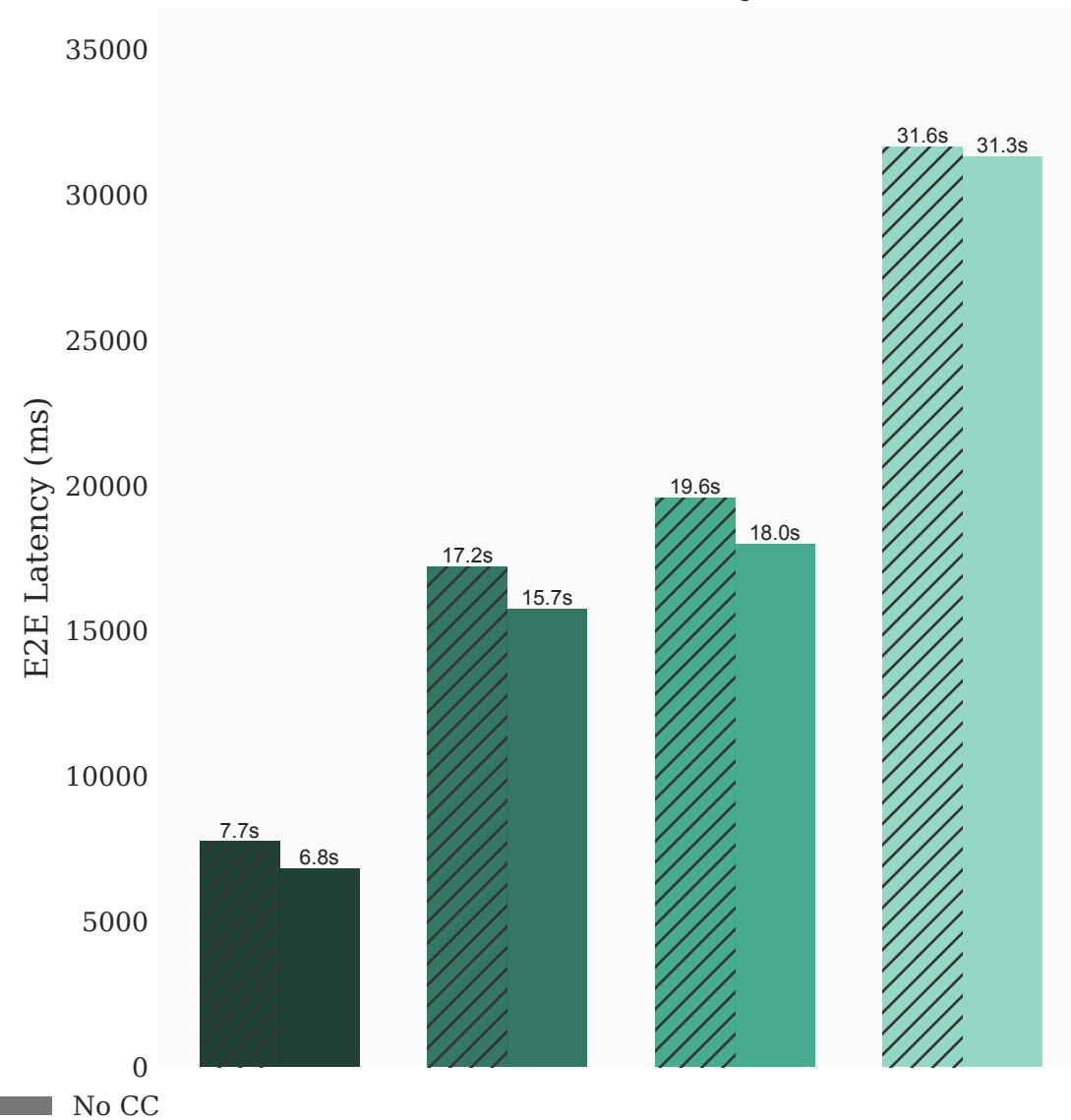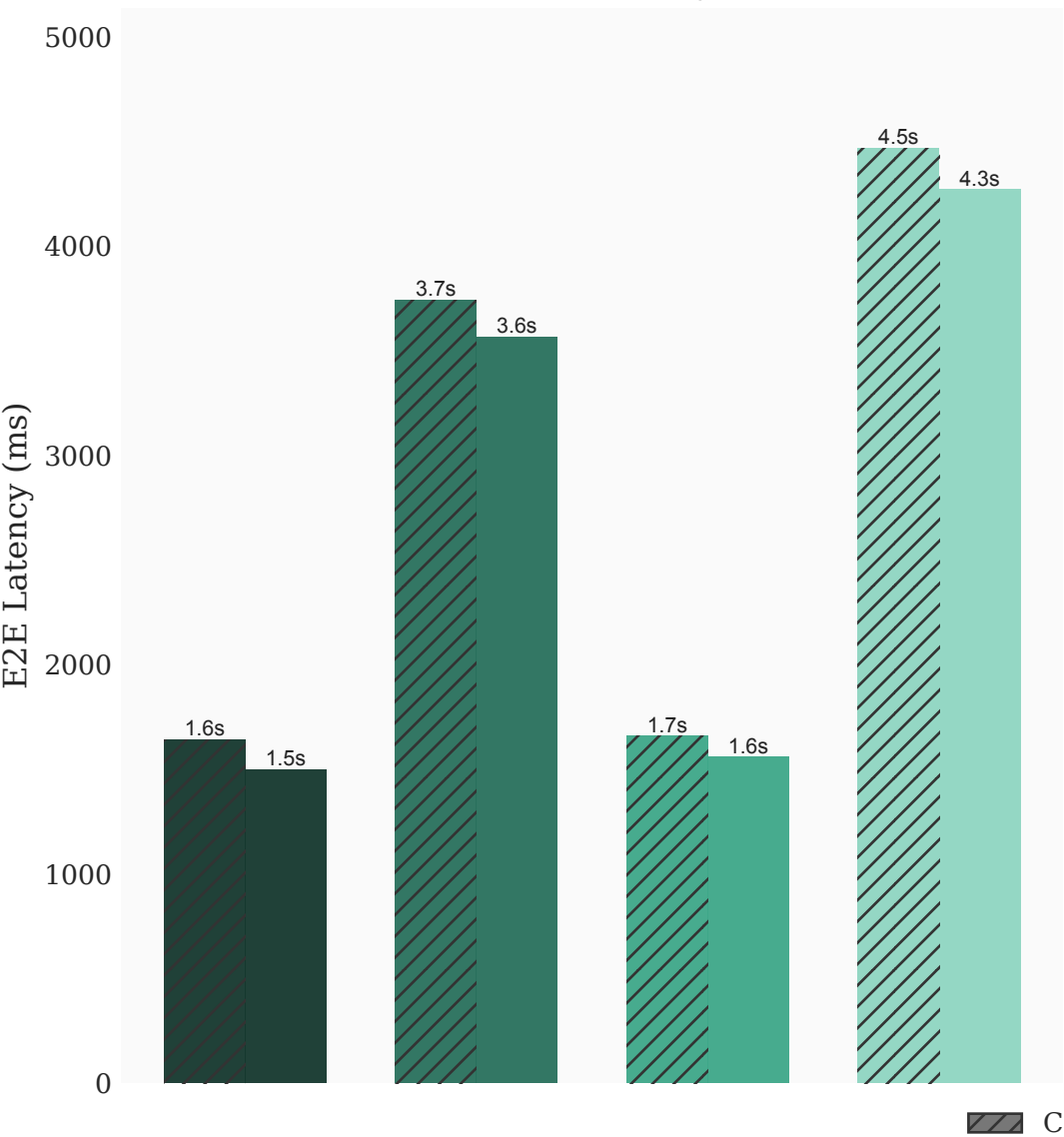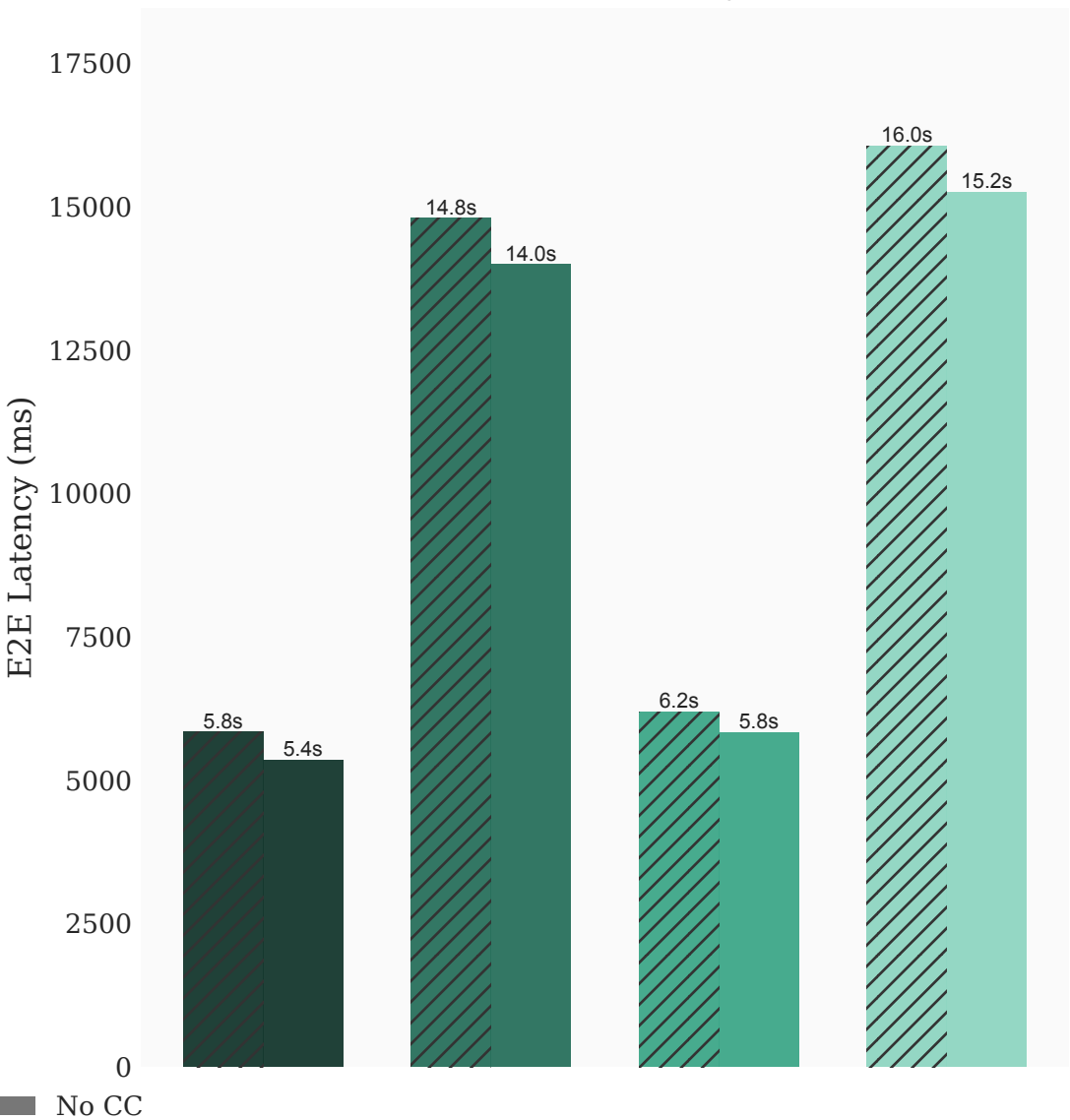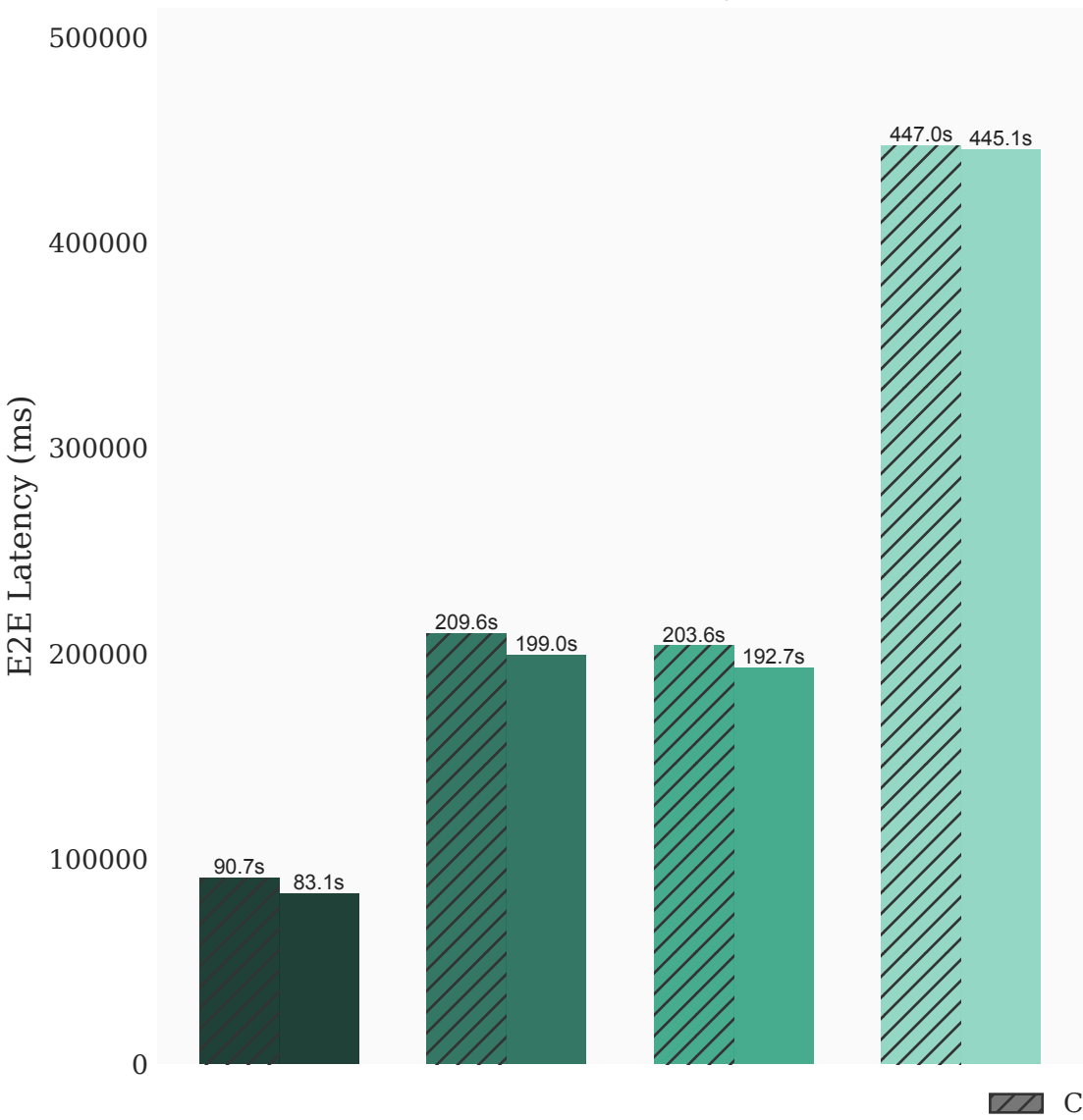## End-to-End Latency (Mean)
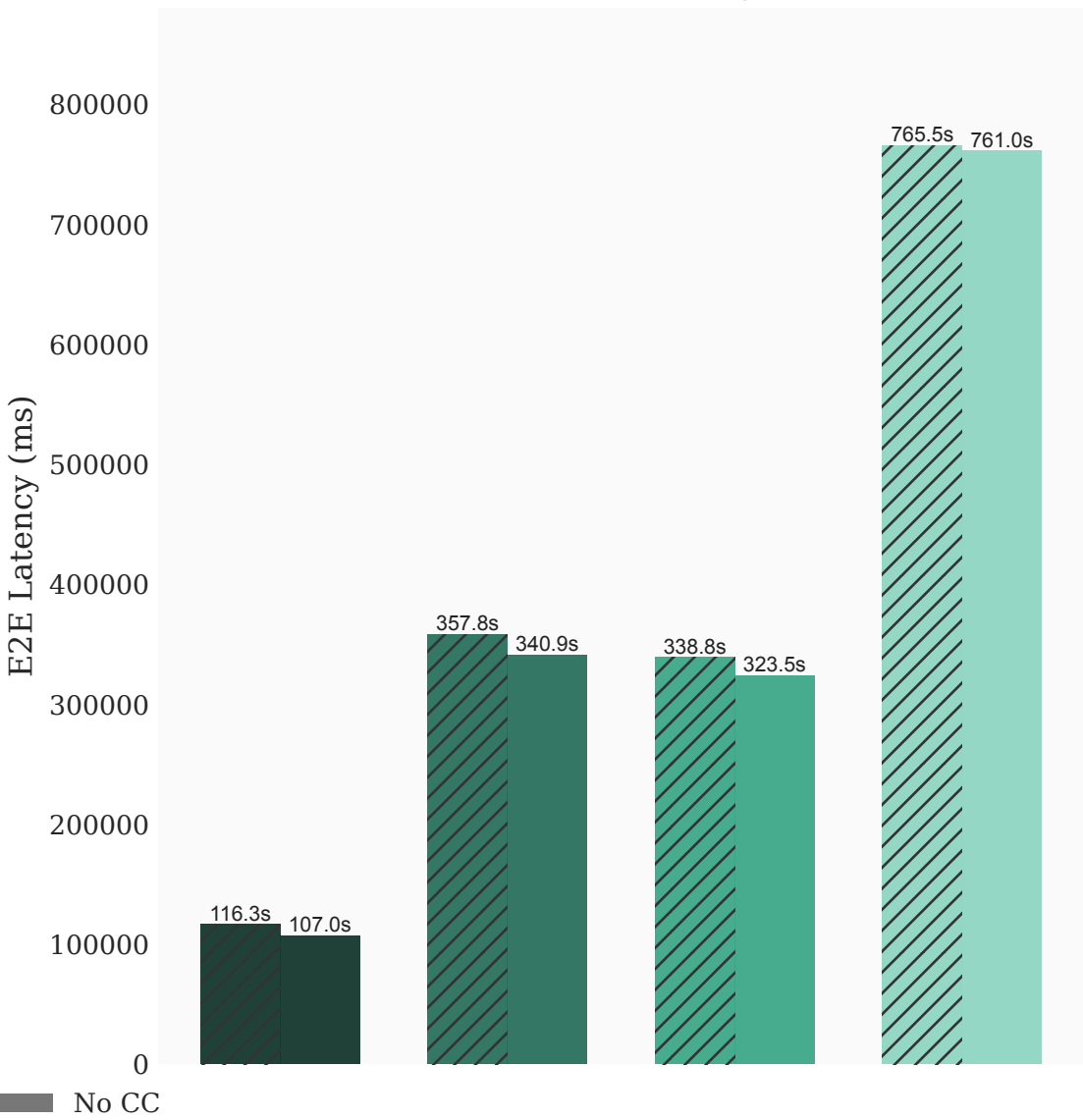

End-to-End Latency (P99)

Legend: CC, No CC

LLama 3.1 8B · Mistral 3.1 24B · GPT OSS 120B · LLama 3.3 70B Int4

Edit 10K Characters (100 Concurrent Requests)

End-to-End Latency (Mean)

End-to-End Latency (P99)

Legend: CC / No CC

LLama 3.1 8B   Mistral 3.1 24B   GPT OSS 120B   LLama 3.3 70B Int4

Mean values: LLama 3.1 8B — 90.7s (CC), 83.1s (No CC); Mistral 3.1 24B — 209.6s (CC), 199.0s (No CC); GPT OSS 120B — 203.6s (CC), 192.7s (No CC); LLama 3.3 70B Int4 — 447.0s (CC), 445.1s (No CC)

P99 values: LLama 3.1 8B — 116.3s (CC), 107.0s (No CC); Mistral 3.1 24B — 357.8s (CC), 340.9s (No CC); GPT OSS 120B — 338.8s (CC), 323.5s (No CC); LLama 3.3 70B Int4 — 765.5s (CC), 761.0s (No CC)

Edit 10K Characters (50 Concurrent Requests)

End-to-End Latency (Mean)

End-to-End Latency (P99)

Bar chart legend: CC (hatched), No CC (solid)

Models: LLama 3.1 8B, Mistral 3.1 24B, GPT OSS 120B, LLama 3.3 70B Int4

End-to-End Latency (Mean) values:
- LLama 3.1 8B: 85.7s (CC), 78.2s (No CC)
- Mistral 3.1 24B: 207.8s (CC), 197.1s (No CC)
- GPT OSS 120B: 203.3s (CC), 191.6s (No CC)
- LLama 3.3 70B Int4: 441.5s (CC), 439.4s (No CC)

End-to-End Latency (P99) values:
- LLama 3.1 8B: 112.0s (CC), 101.3s (No CC)
- Mistral 3.1 24B: 355.5s (CC), 336.1s (No CC)
- GPT OSS 120B: 338.0s (CC), 320.7s (No CC)
- LLama 3.3 70B Int4: 765.8s (CC), 753.6s (No CC)

Edit 10K Characters (1 Concurrent Requests)

End-to-End Latency (Mean)

	CC	No CC
LLama 3.1 8B	35.8s	32.3s
Mistral 3.1 24B	144.5s	133.3s
GPT OSS 120B	139.0s	126.3s
LLama 3.3 70B Int4	367.9s	360.3s

End-to-End Latency (P99)

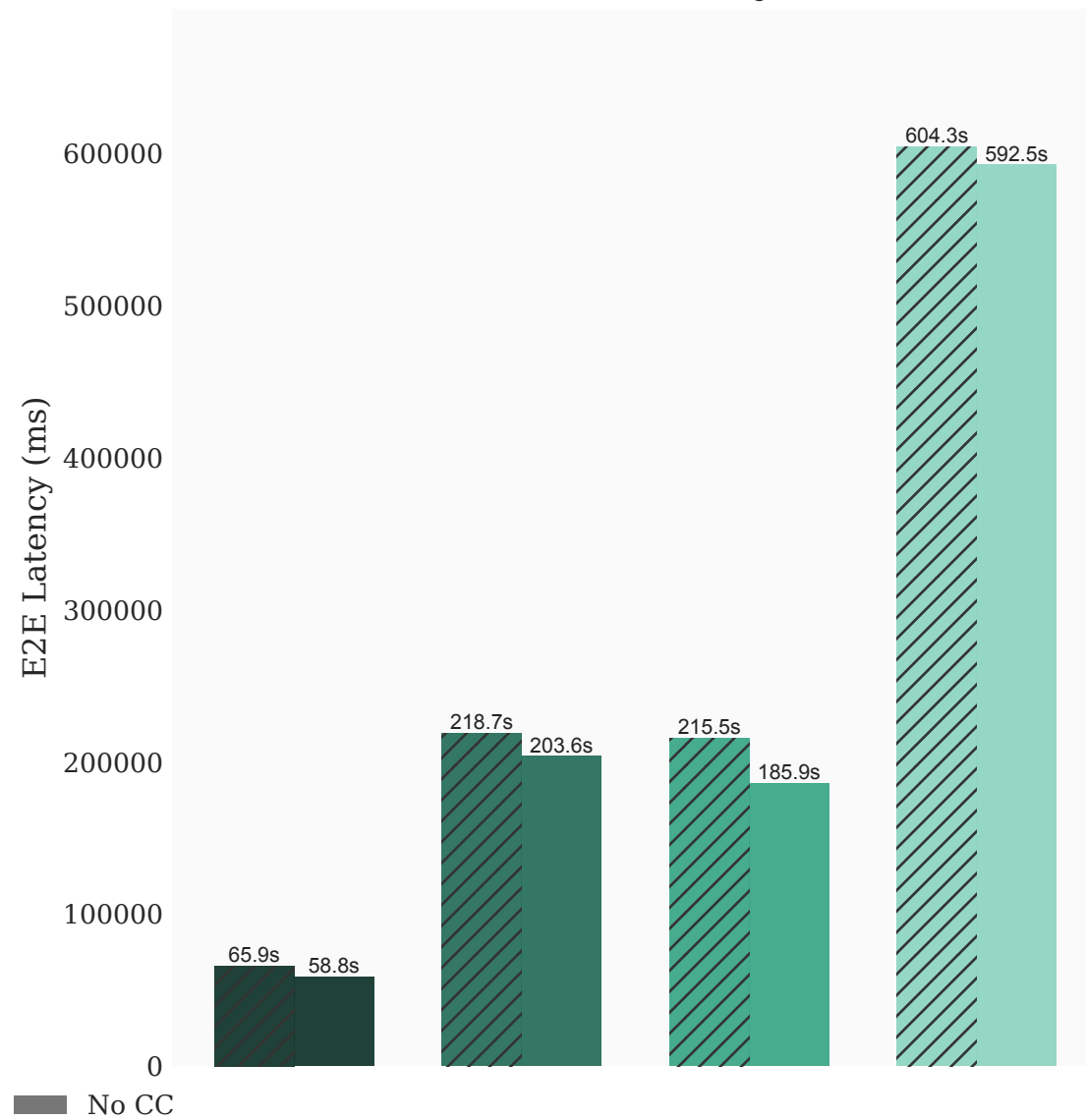|	CC	No CC
LLama 3.1 8B	65.9s	58.8s
Mistral 3.1 24B	218.7s	203.6s
GPT OSS 120B	215.5s	185.9s
LLama 3.3 70B Int4	604.3s	592.5s

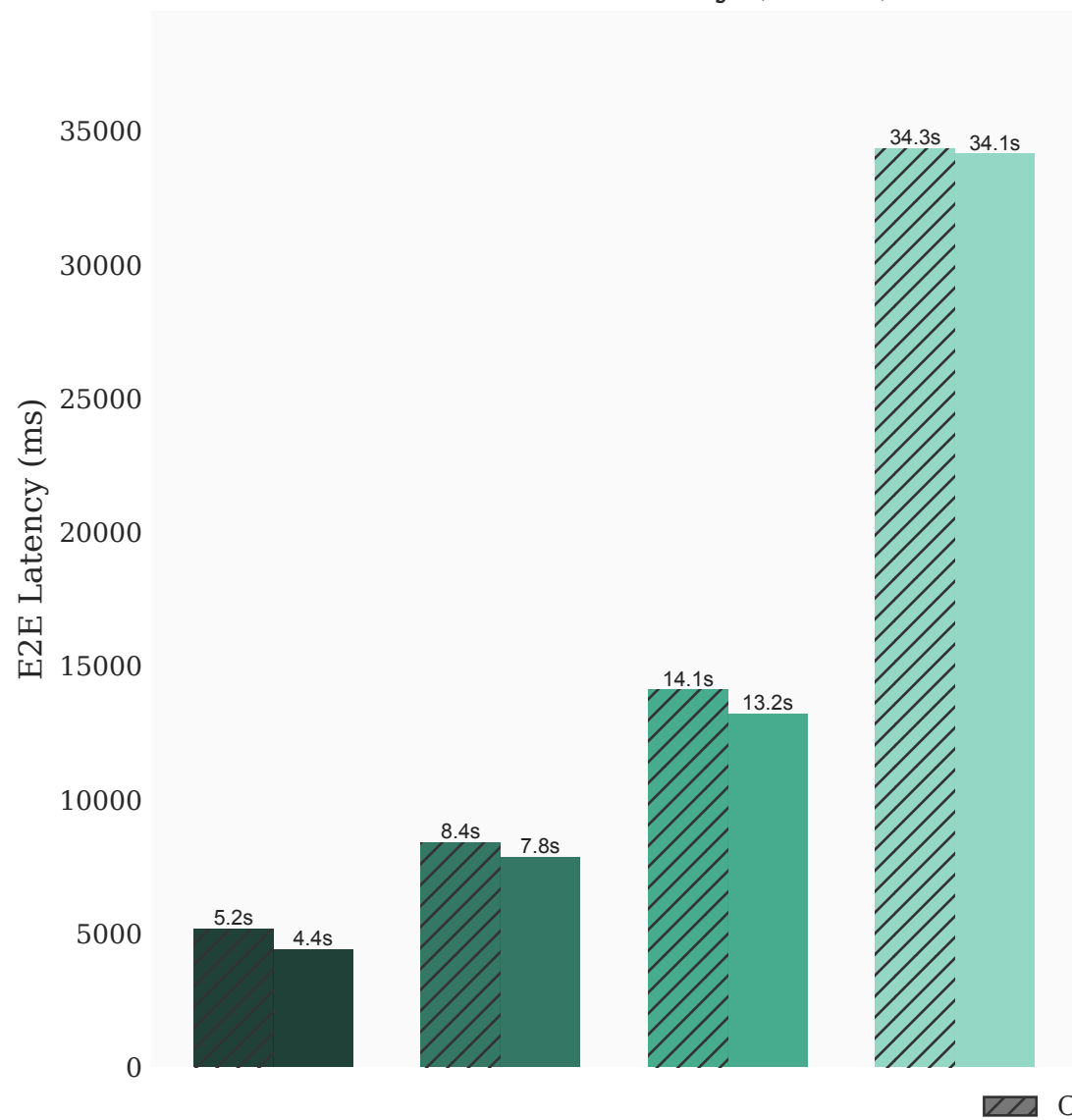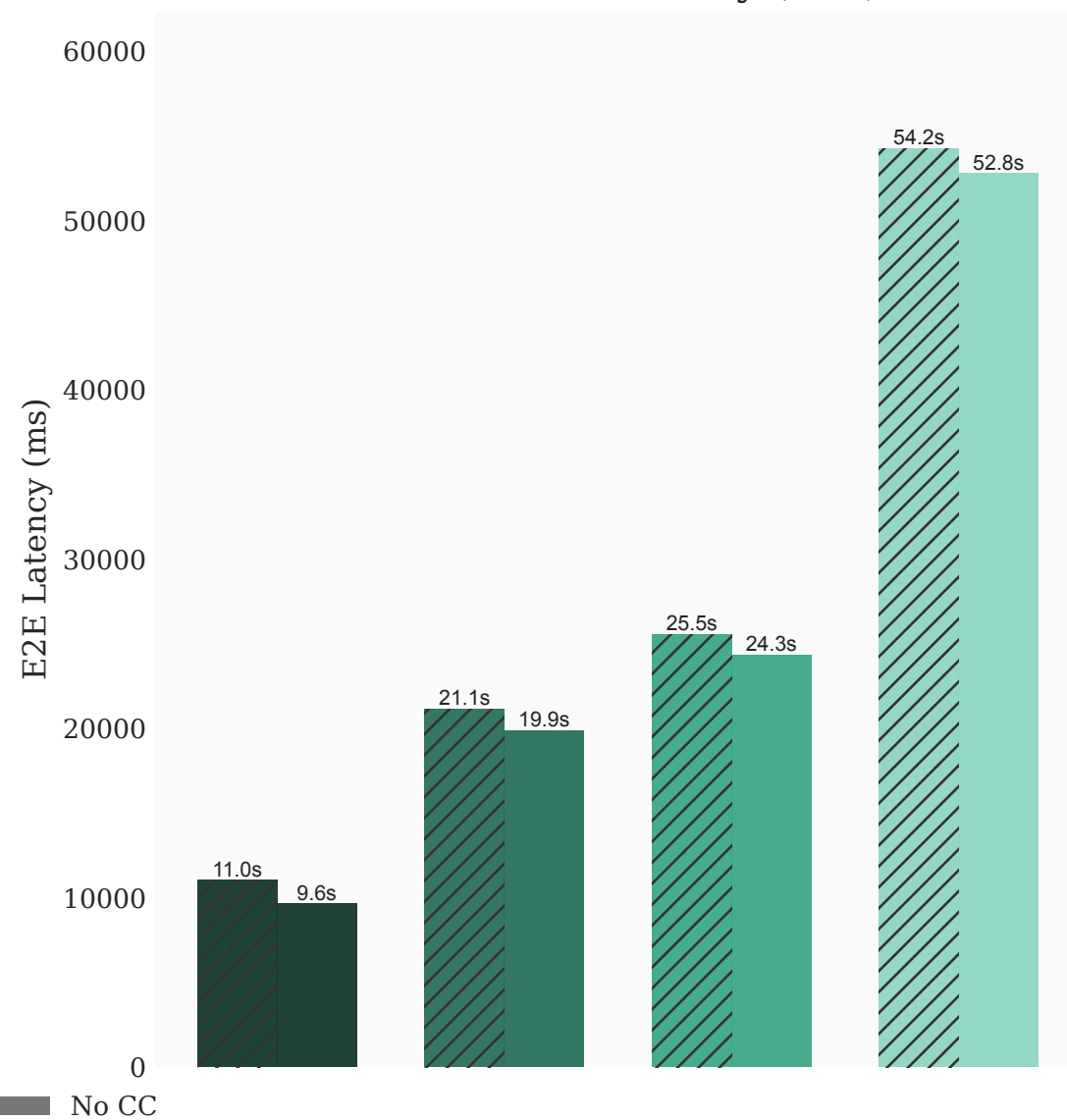Legend: CC, No CC — LLama 3.1 8B, Mistral 3.1 24B, GPT OSS 120B, LLama 3.3 70B Int4

Numina Math (100 Concurrent Requests)

End-to-End Latency (Mean)

E2E Latency (ms)
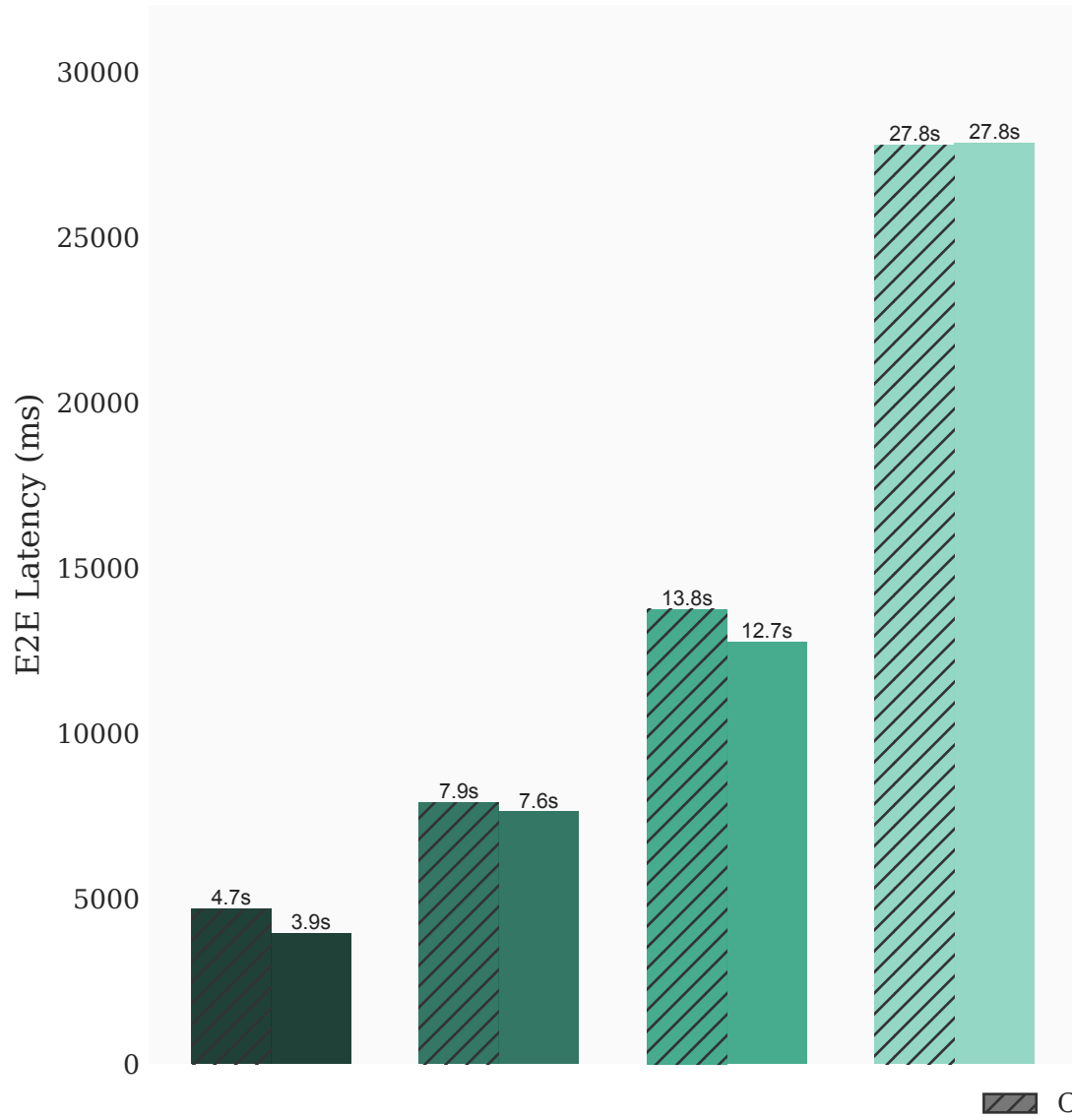
- 5.2s / 4.4s (LLama 3.1 8B)
- 8.4s / 7.8s (Mistral 3.1 24B)
- 14.1s / 13.2s (GPT OSS 120B)
- 34.3s / 34.1s (LLama 3.3 70B Int4)

End-to-End Latency (P99)

E2E Latency (ms)

- 11.0s / 9.6s (LLama 3.1 8B)
- 21.1s / 19.9s (Mistral 3.1 24B)
- 25.5s / 24.3s (GPT OSS 120B)
- 54.2s / 52.8s (LLama 3.3 70B Int4)
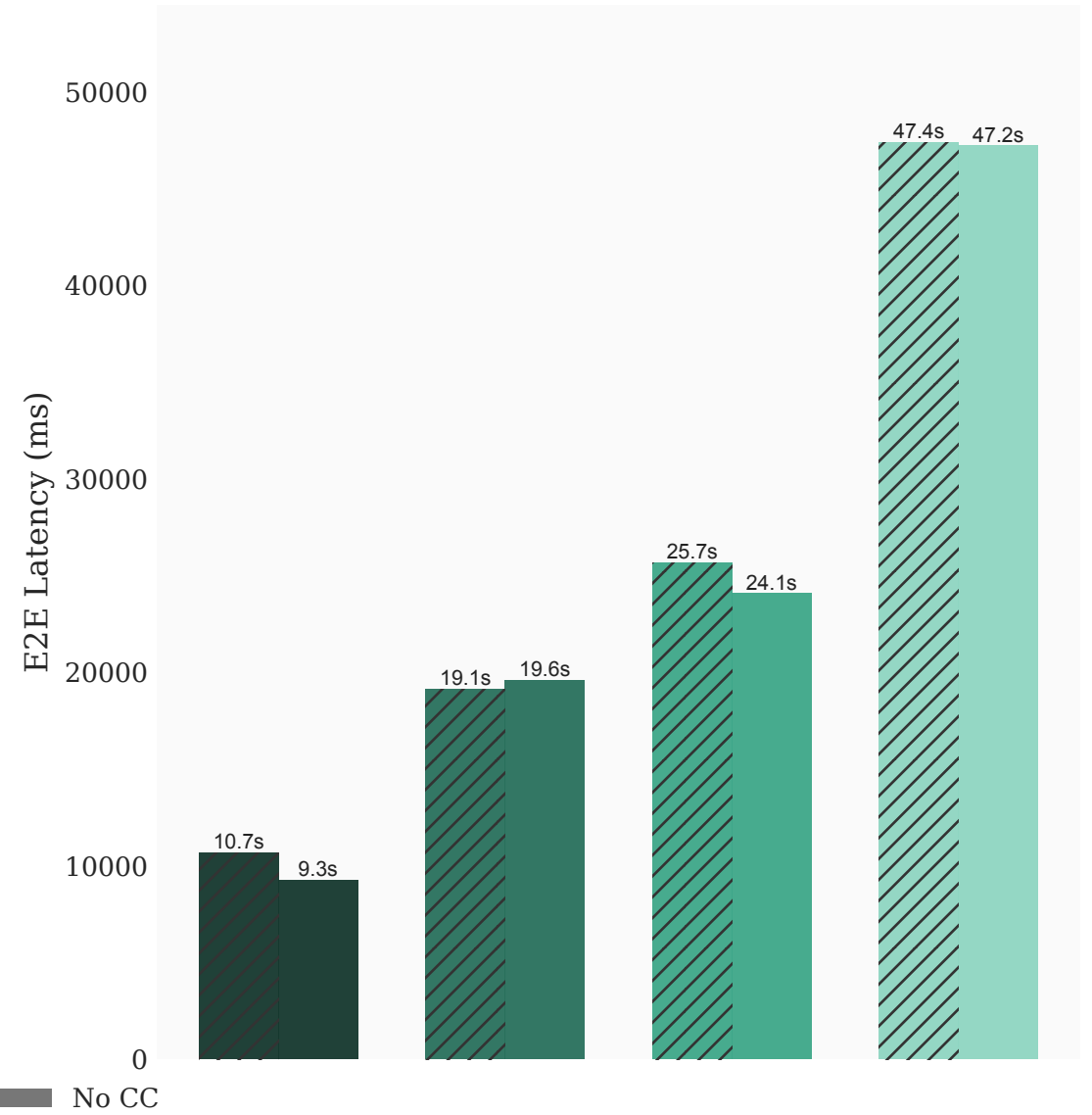
Legend: CC / No CC

LLama 3.1 8B   Mistral 3.1 24B   GPT OSS 120B   LLama 3.3 70B Int4

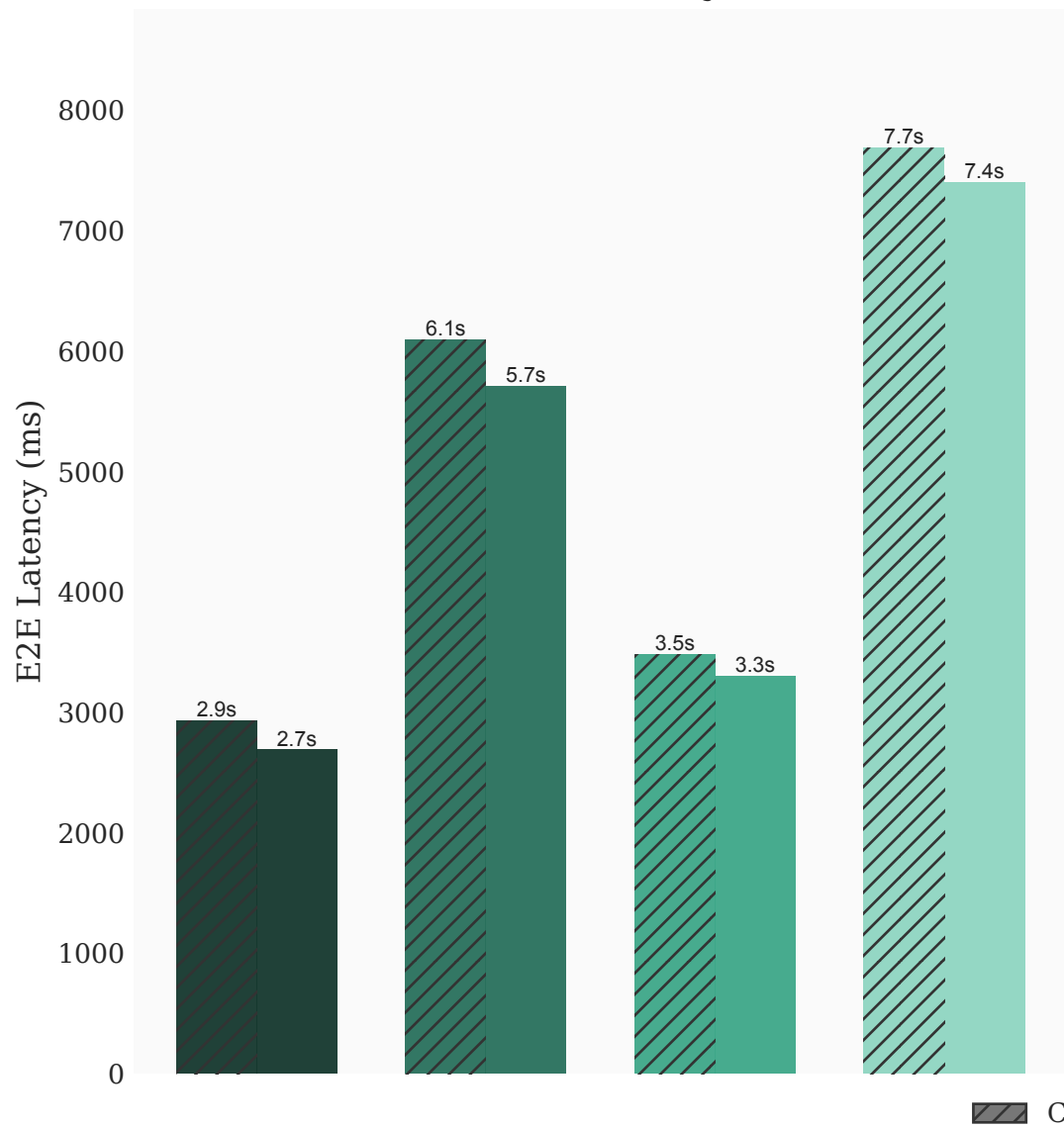# Numina Math (50 Concurrent Requests)

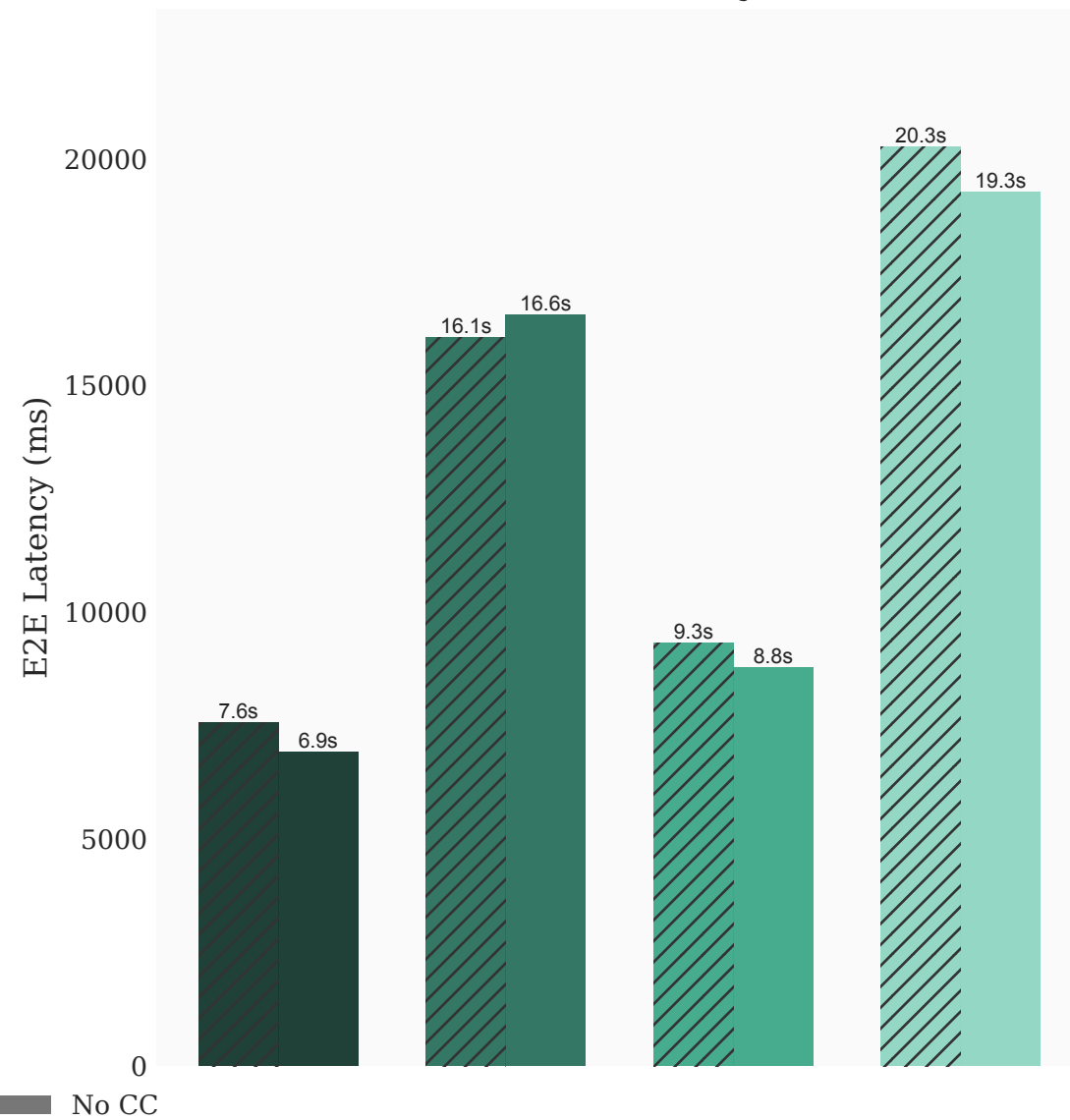## End-to-End Latency (Mean)

End-to-End Latency (P99)

Legend: ⧄ CC   ▬ No CC

■ LLama 3.1 8B   ■ Mistral 3.1 24B   ■ GPT OSS 120B   ■ LLama 3.3 70B Int4

Numina Math (1 Concurrent Requests)

End-to-End Latency (Mean)

End-to-End Latency (P99)

Legend: CC | No CC

LLama 3.1 8B | Mistral 3.1 24B | GPT OSS 120B | LLama 3.3 70B Int4