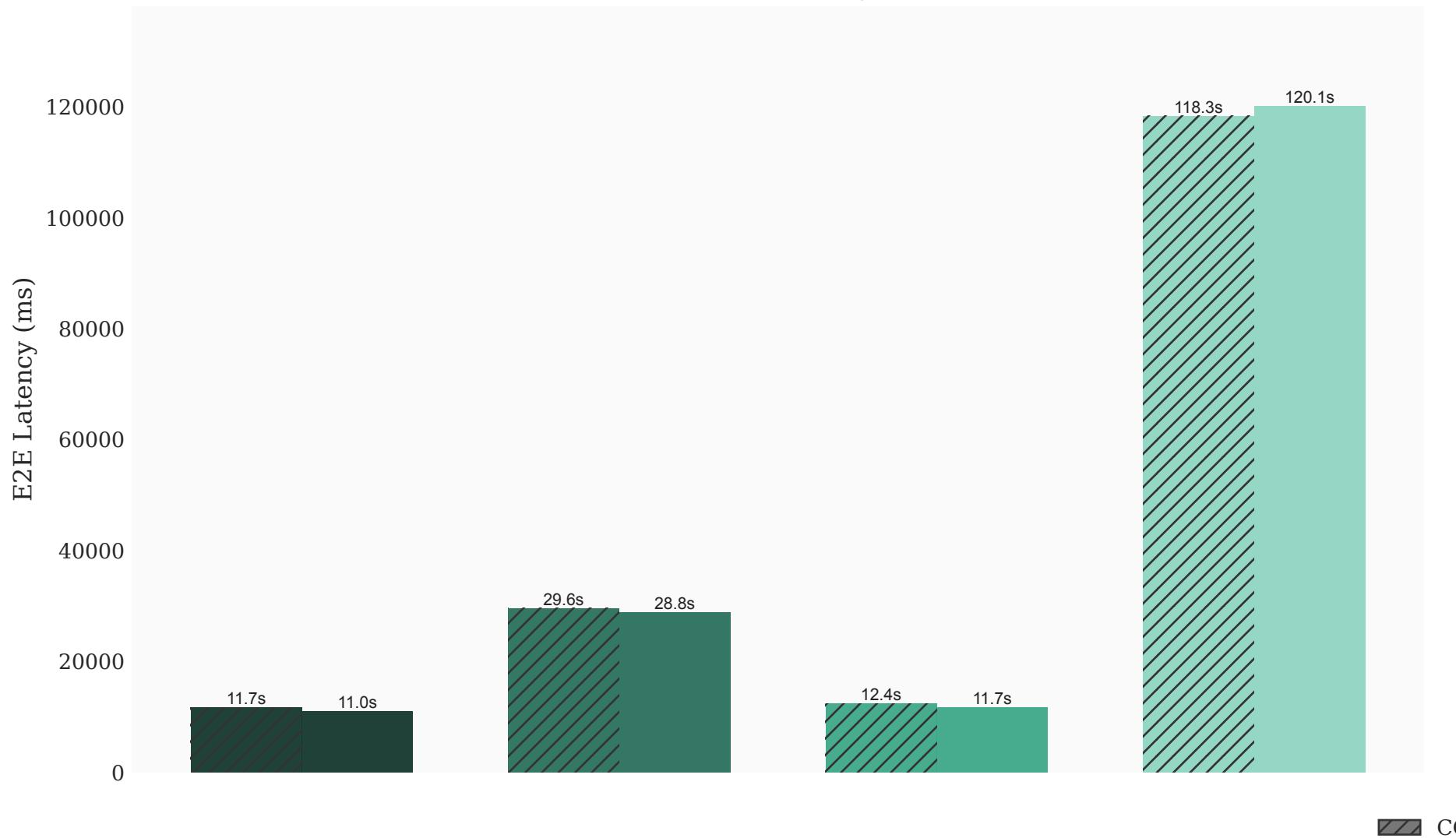
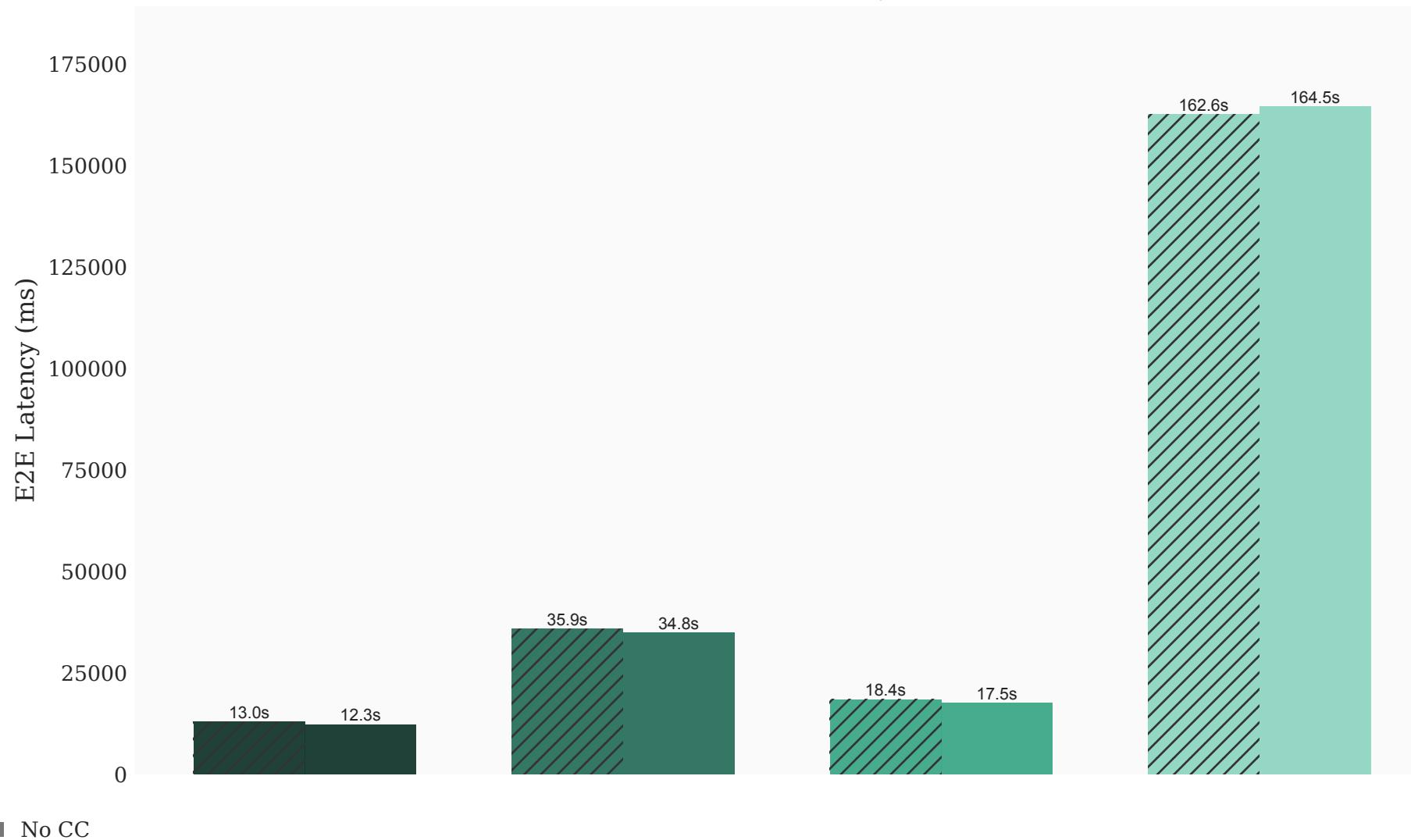


Random (1500 ⇒ 250) (Rate 100)

End-to-End Latency (Mean)



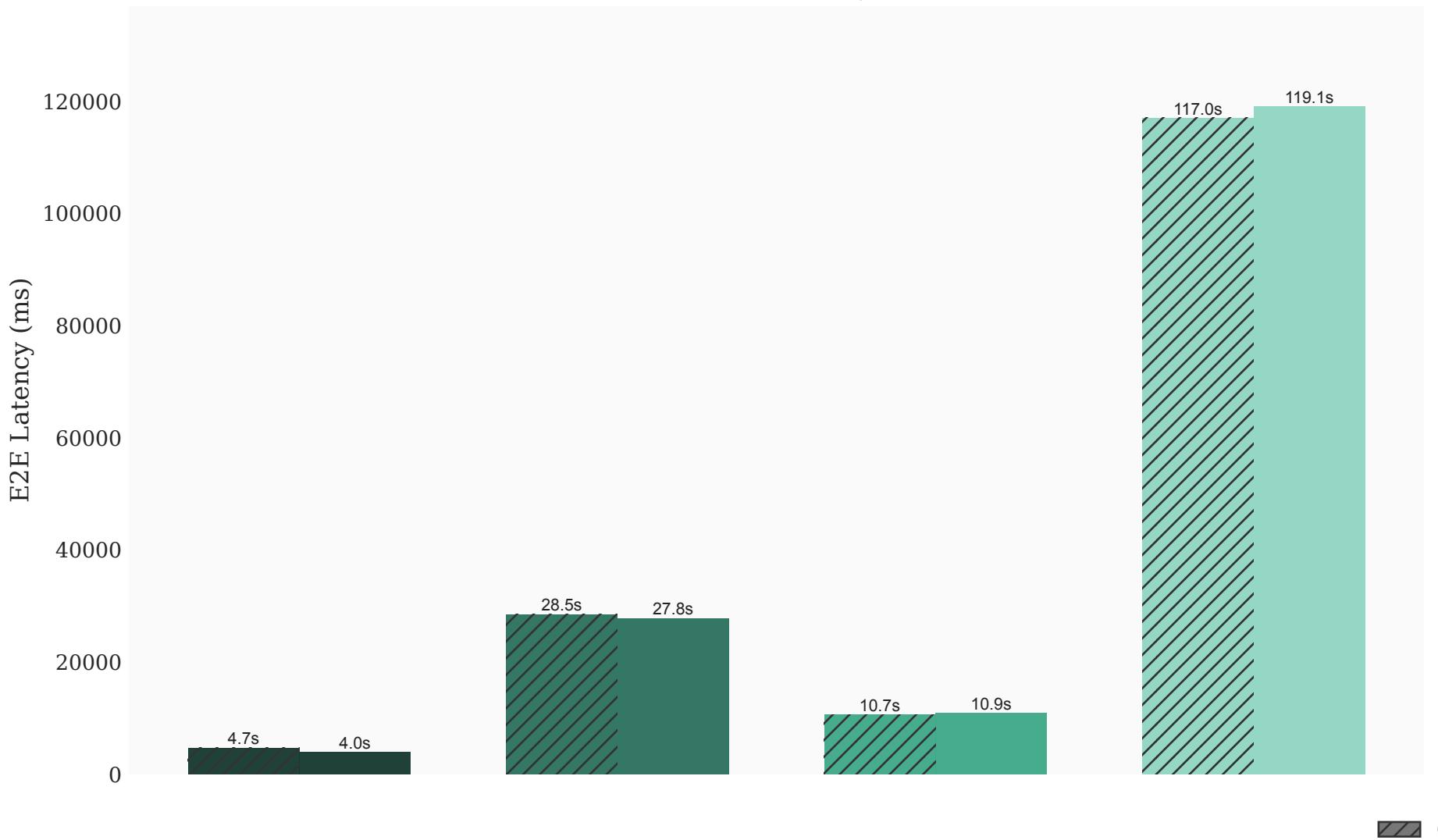
End-to-End Latency (P99)



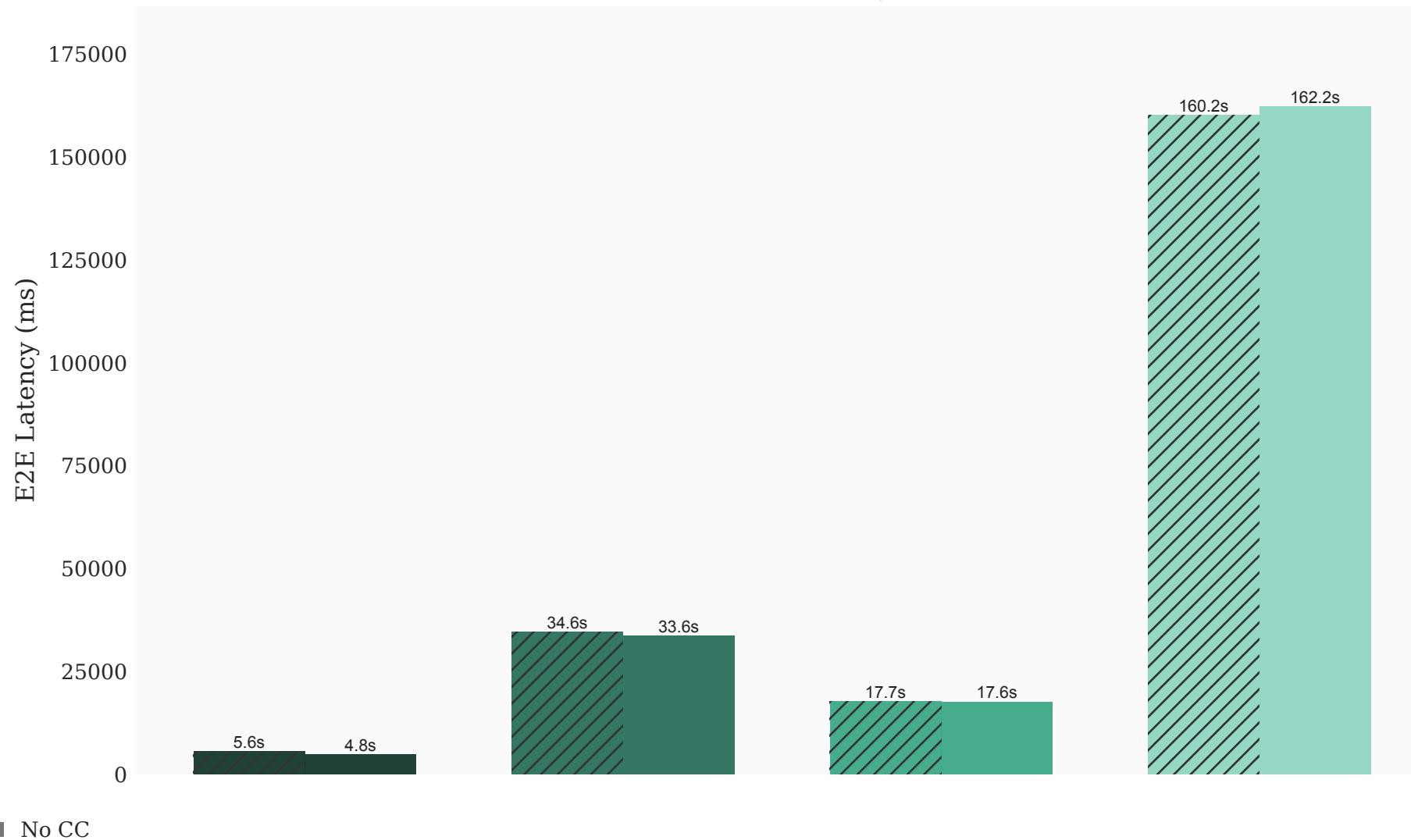
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (1500 ⇒ 250) (Rate 50)

End-to-End Latency (Mean)



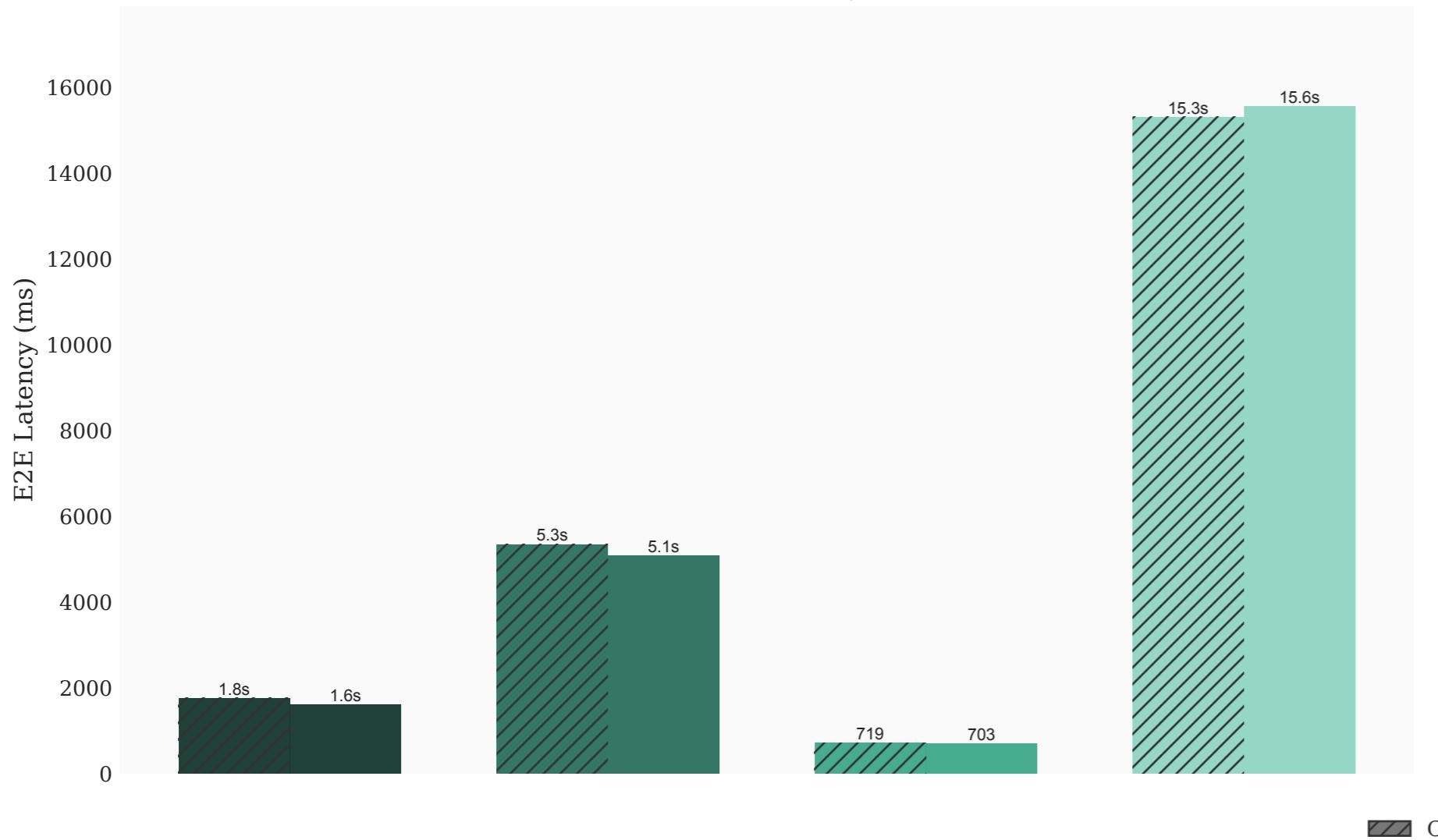
End-to-End Latency (P99)



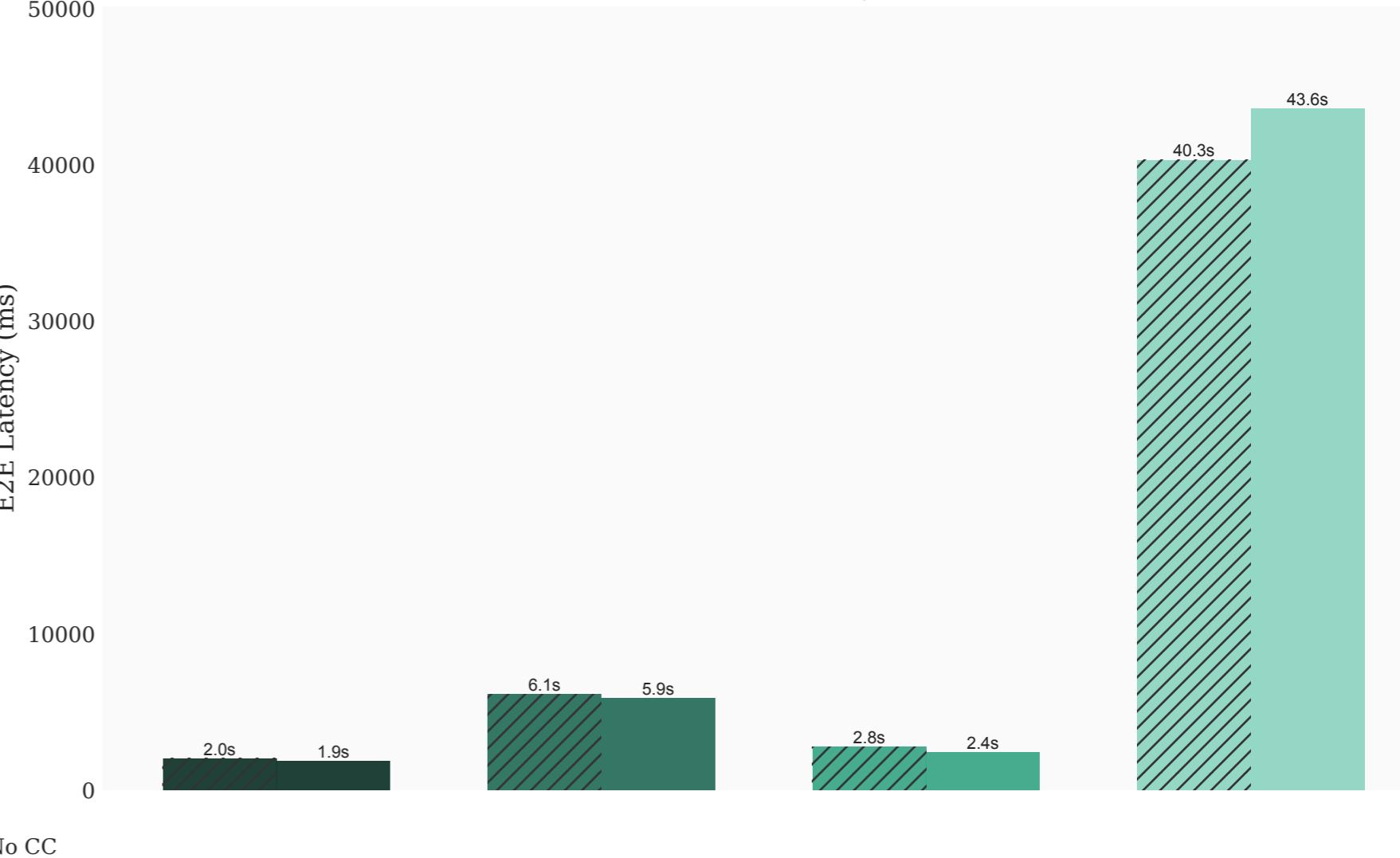
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (1500 \Rightarrow 250) (Rate 1)

End-to-End Latency (Mean)



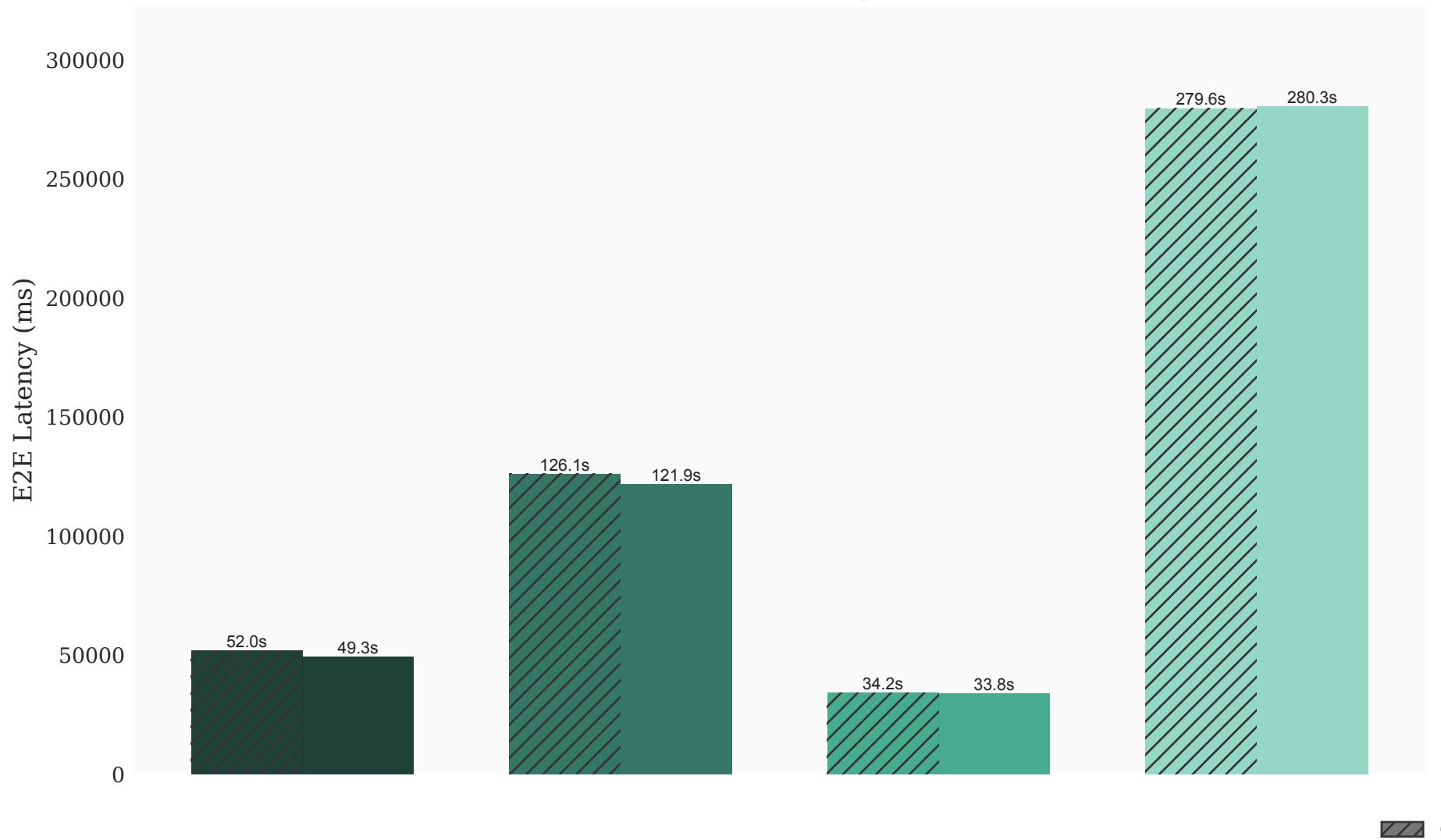
End-to-End Latency (P99)



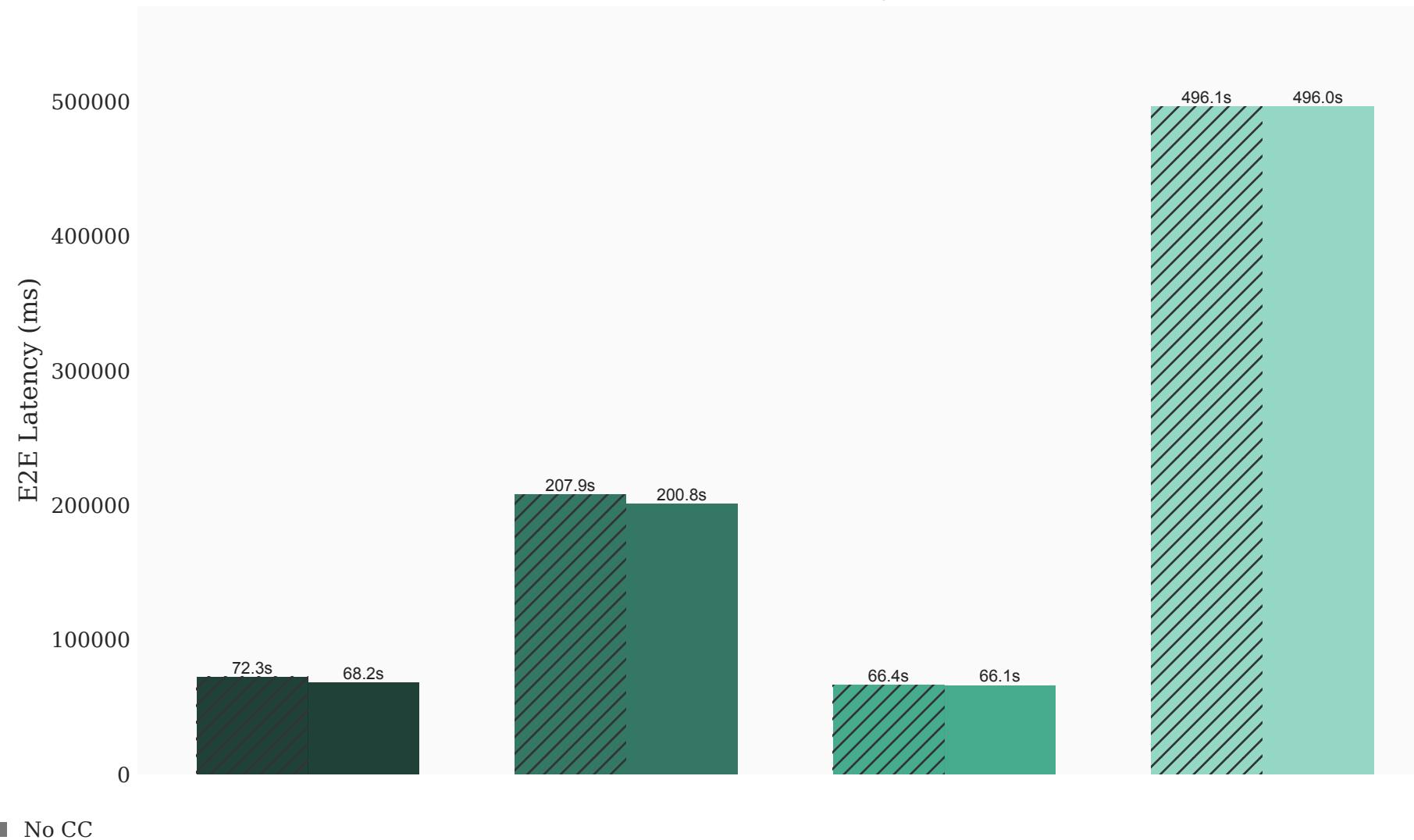
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (4000 ⇒ 1000) (Rate 100)

End-to-End Latency (Mean)



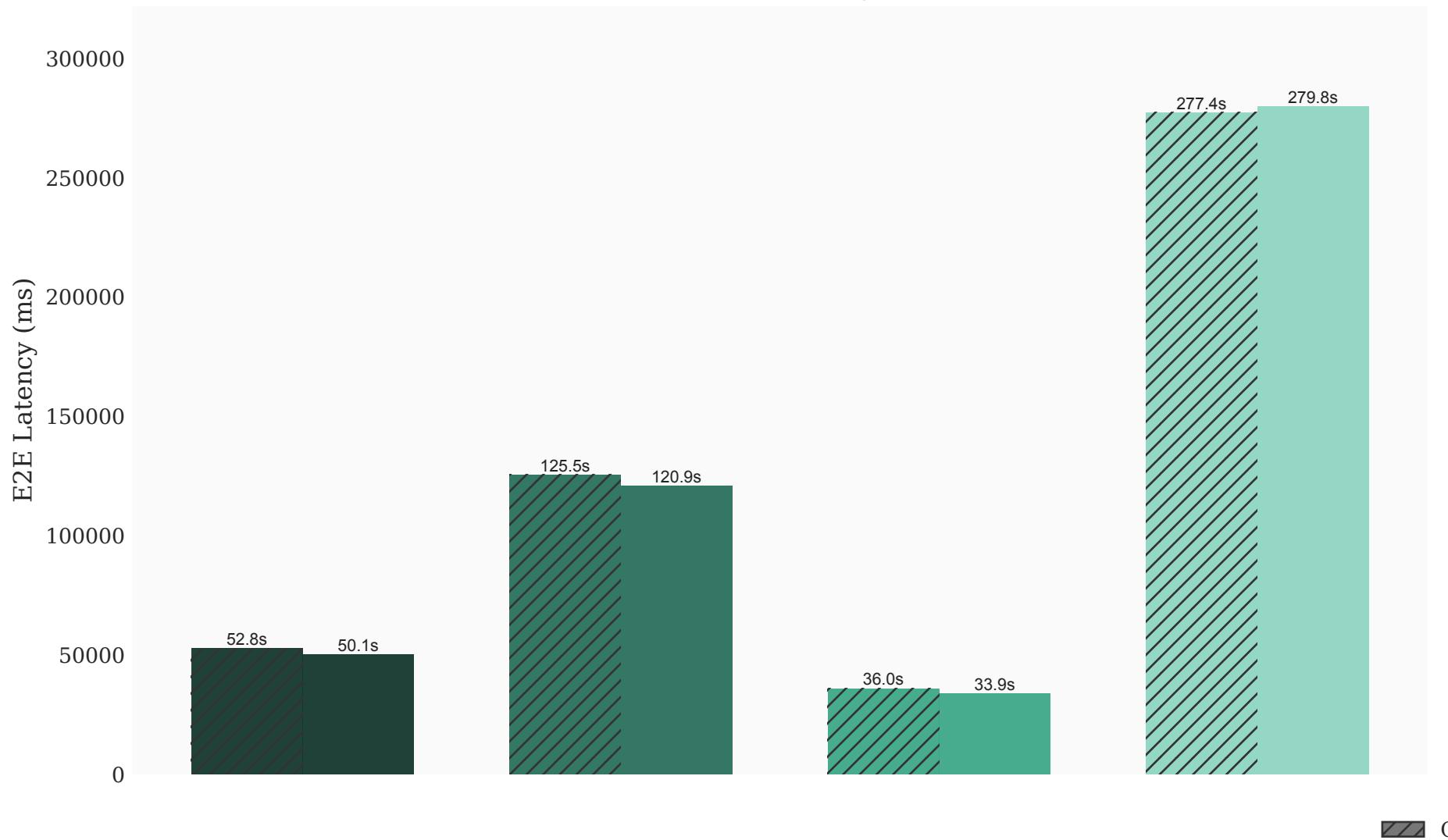
End-to-End Latency (P99)



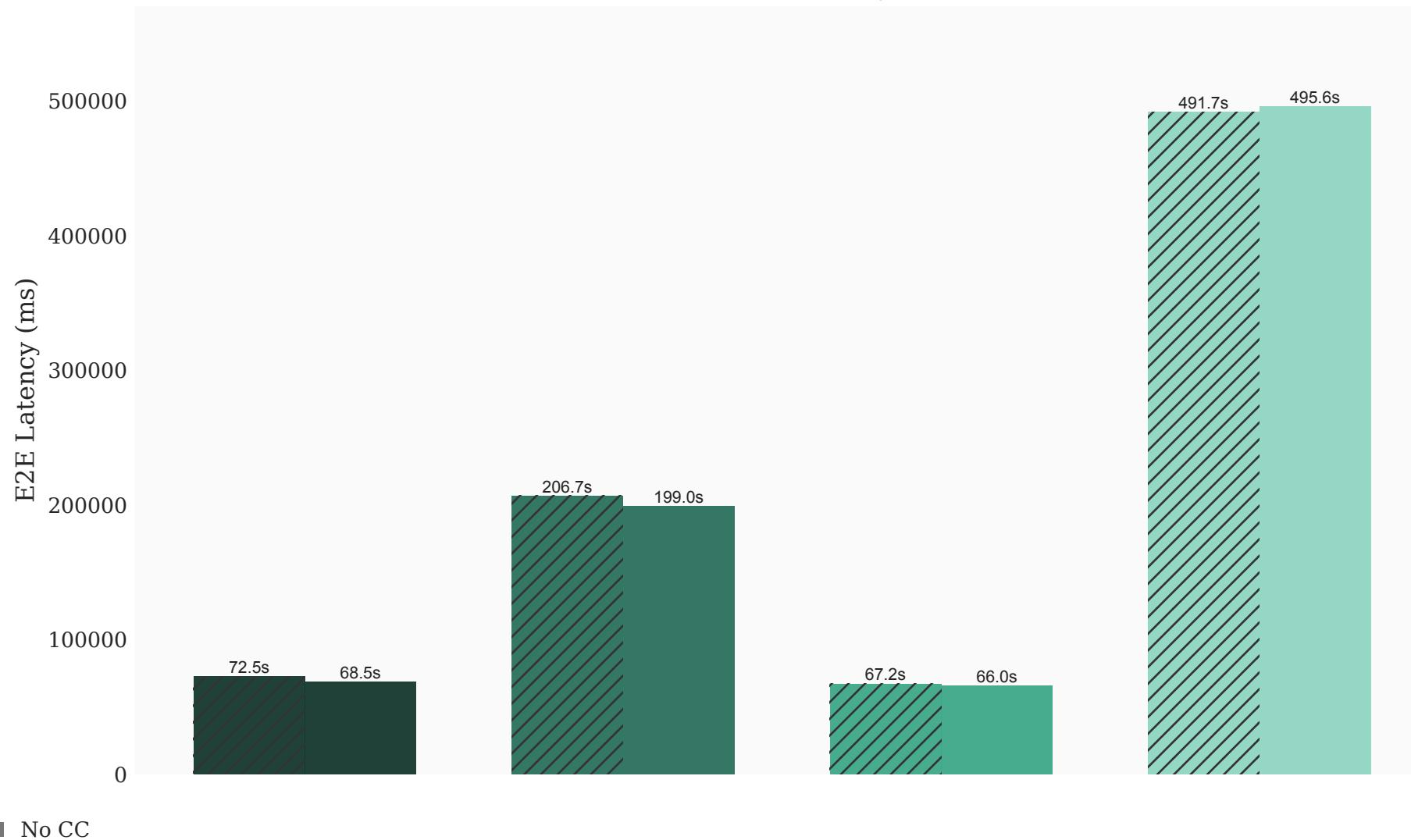
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (4000 ⇒ 1000) (Rate 50)

End-to-End Latency (Mean)



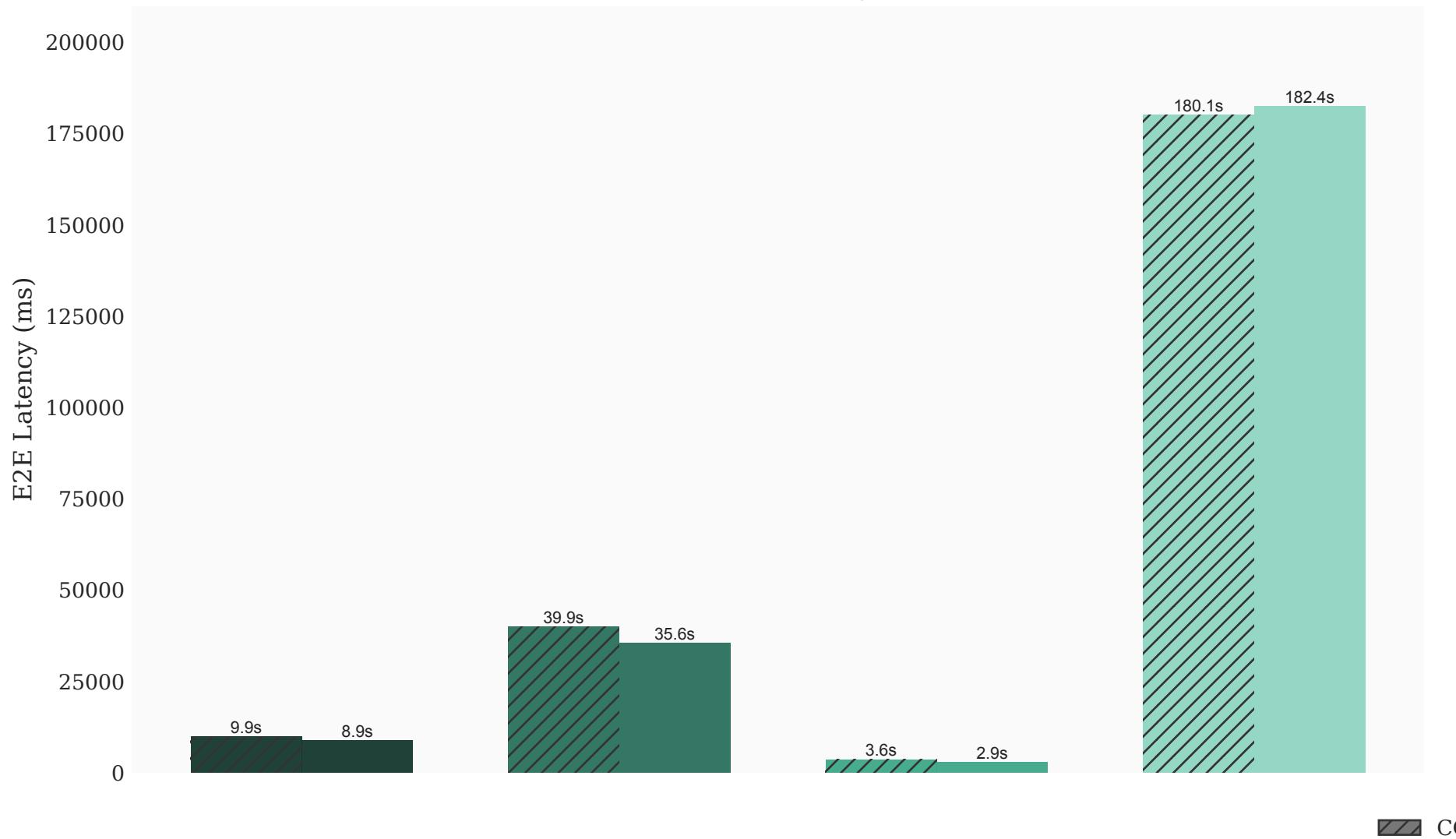
End-to-End Latency (P99)



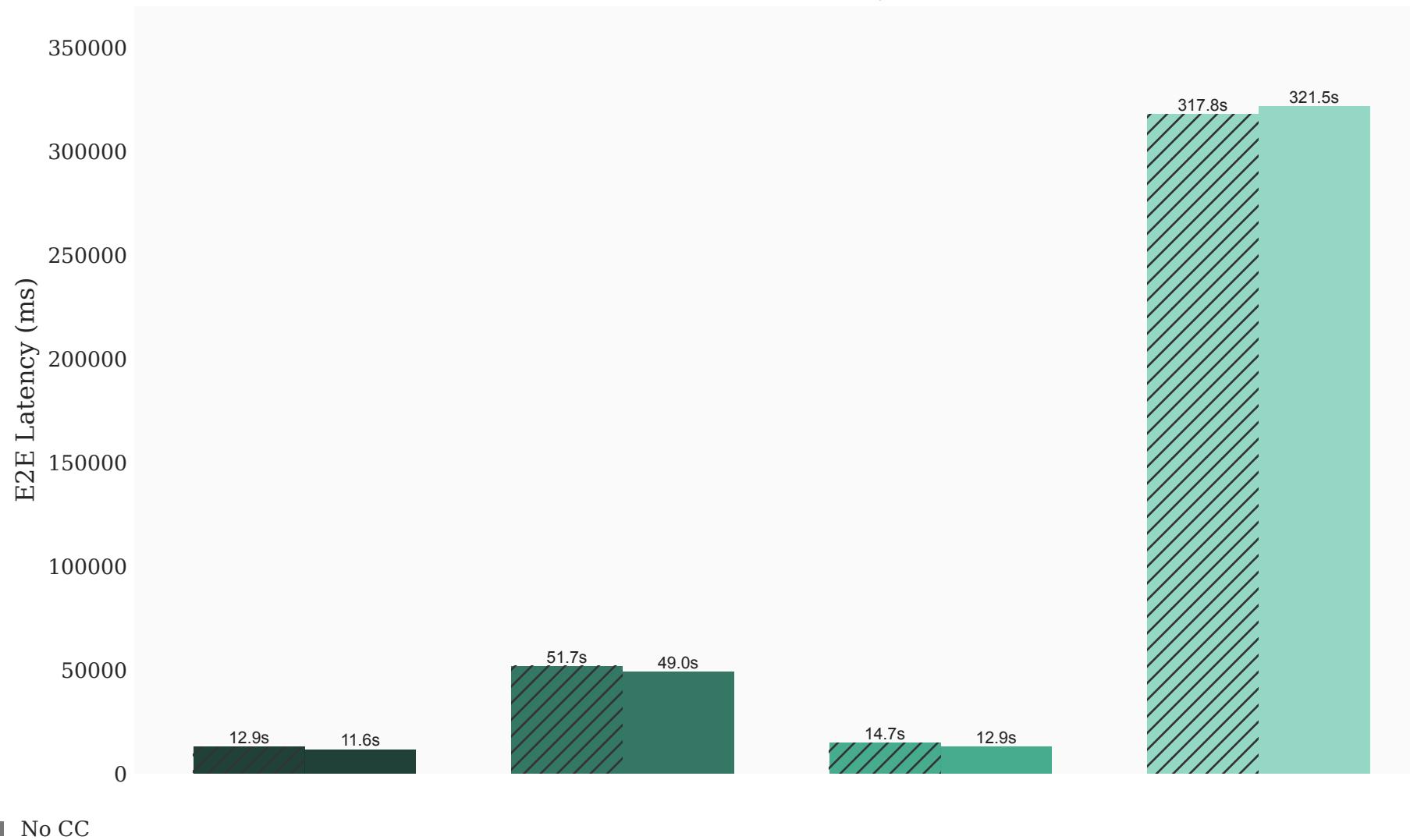
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (4000 \Rightarrow 1000) (Rate 1)

End-to-End Latency (Mean)



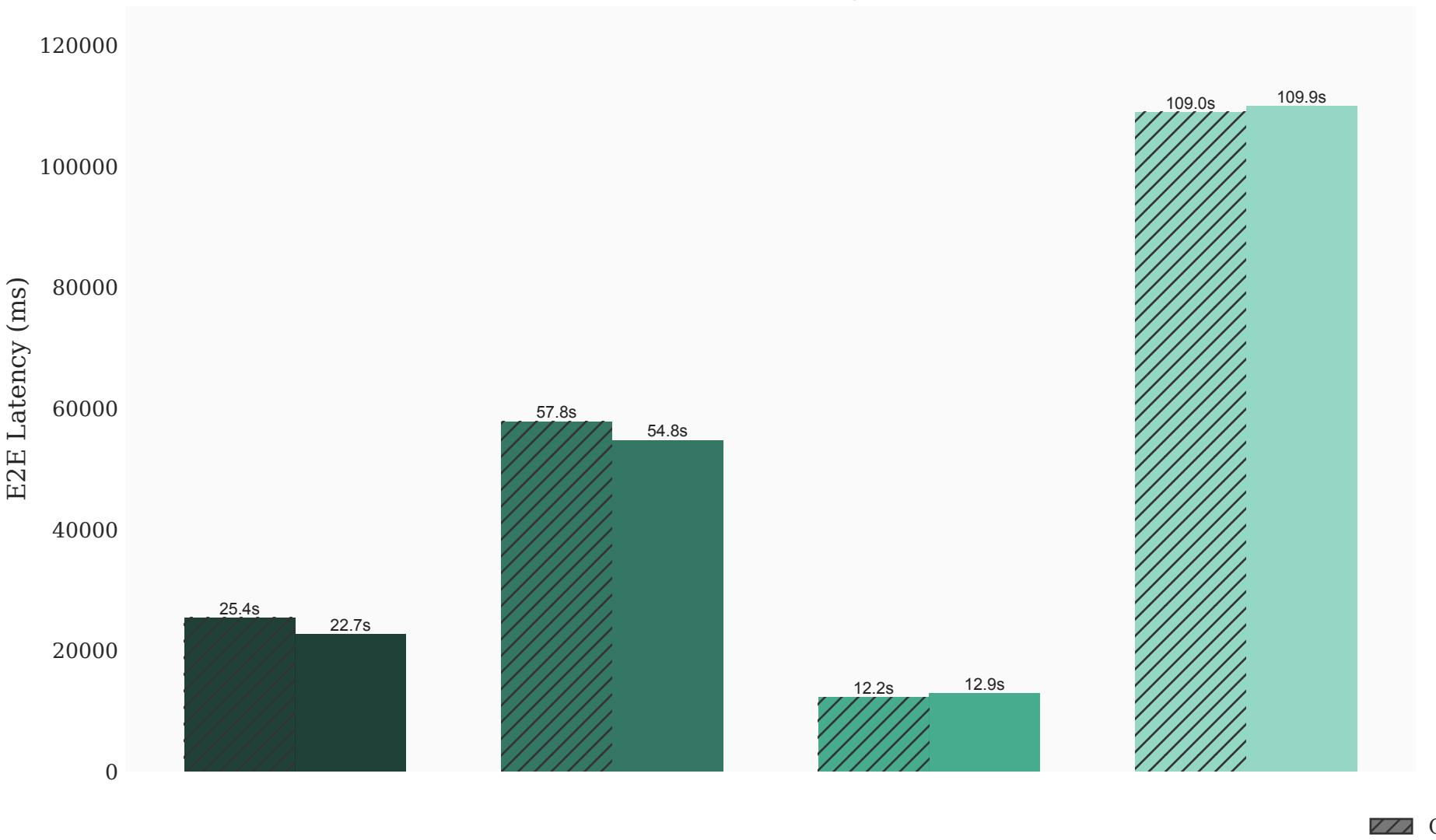
End-to-End Latency (P99)



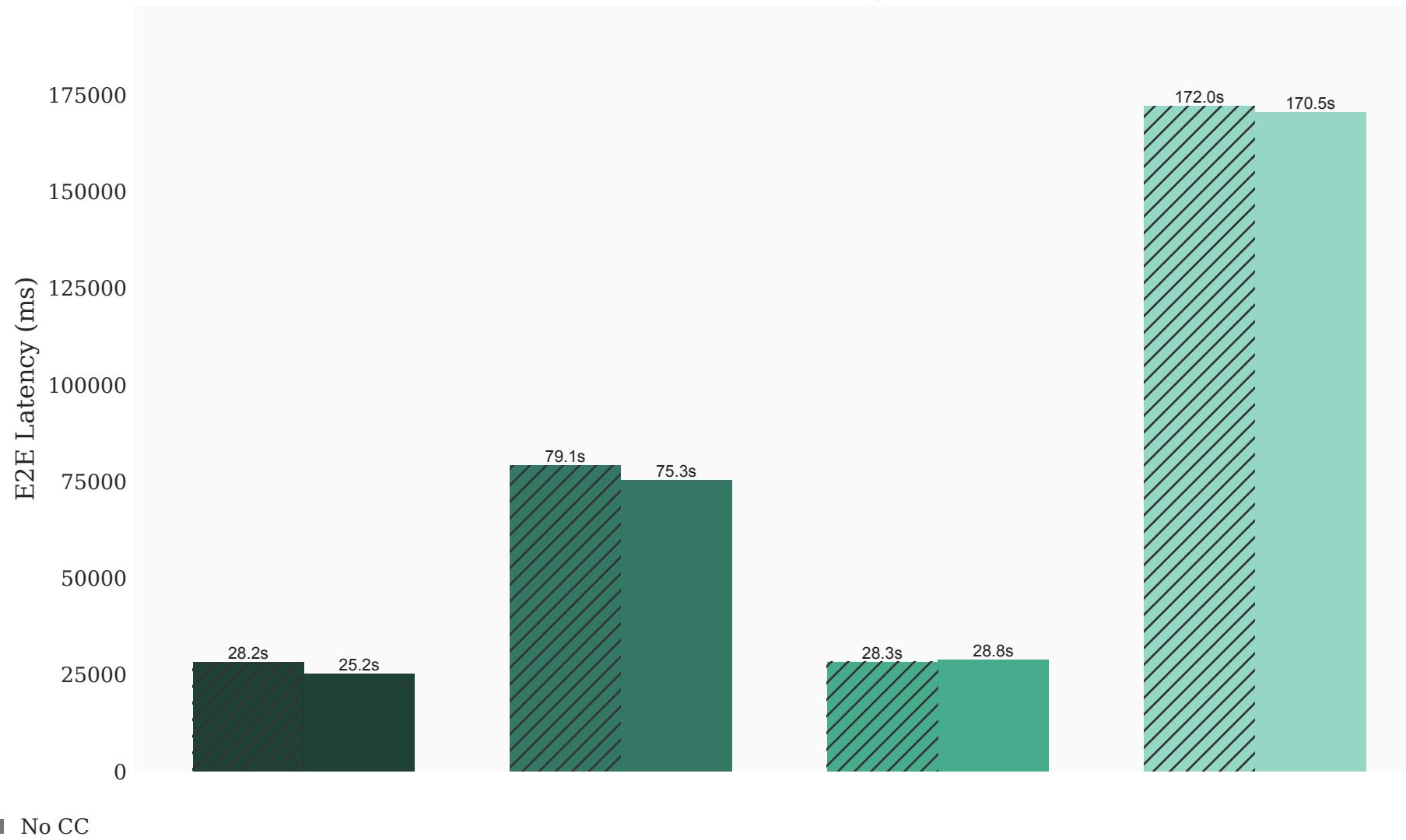
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (1000 ⇒ 1000) (Rate 100)

End-to-End Latency (Mean)



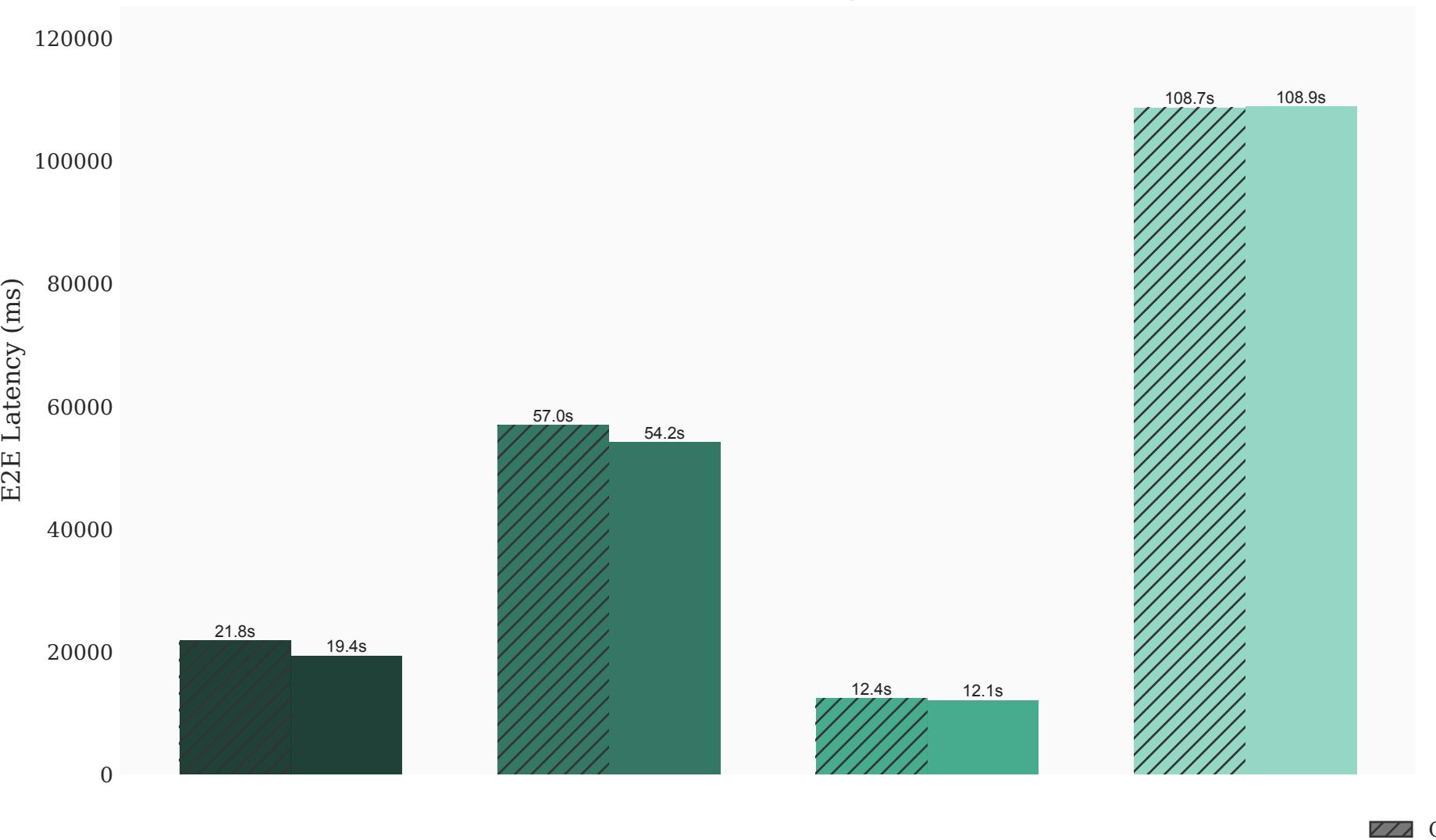
End-to-End Latency (P99)



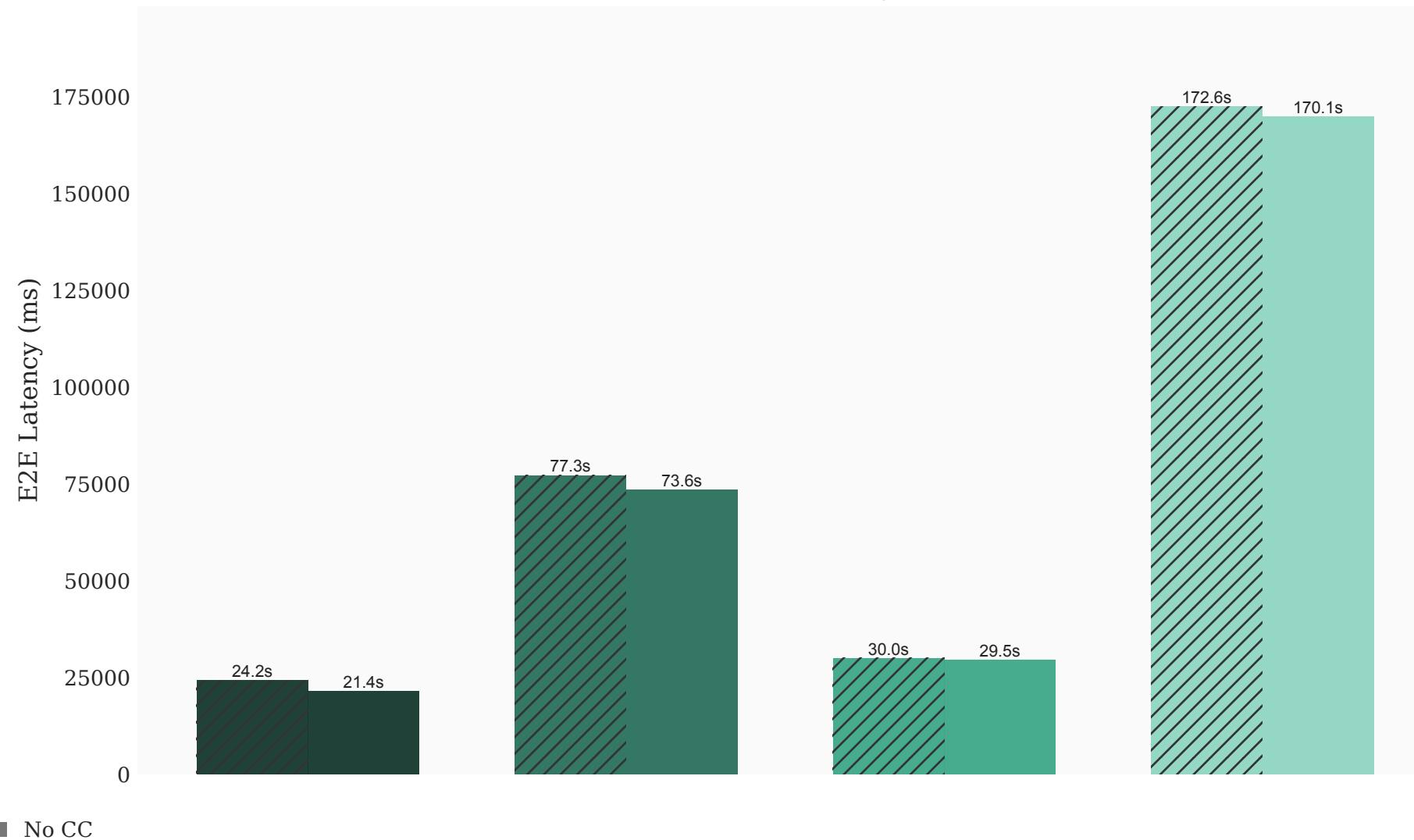
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (1000 ⇒ 1000) (Rate 50)

End-to-End Latency (Mean)

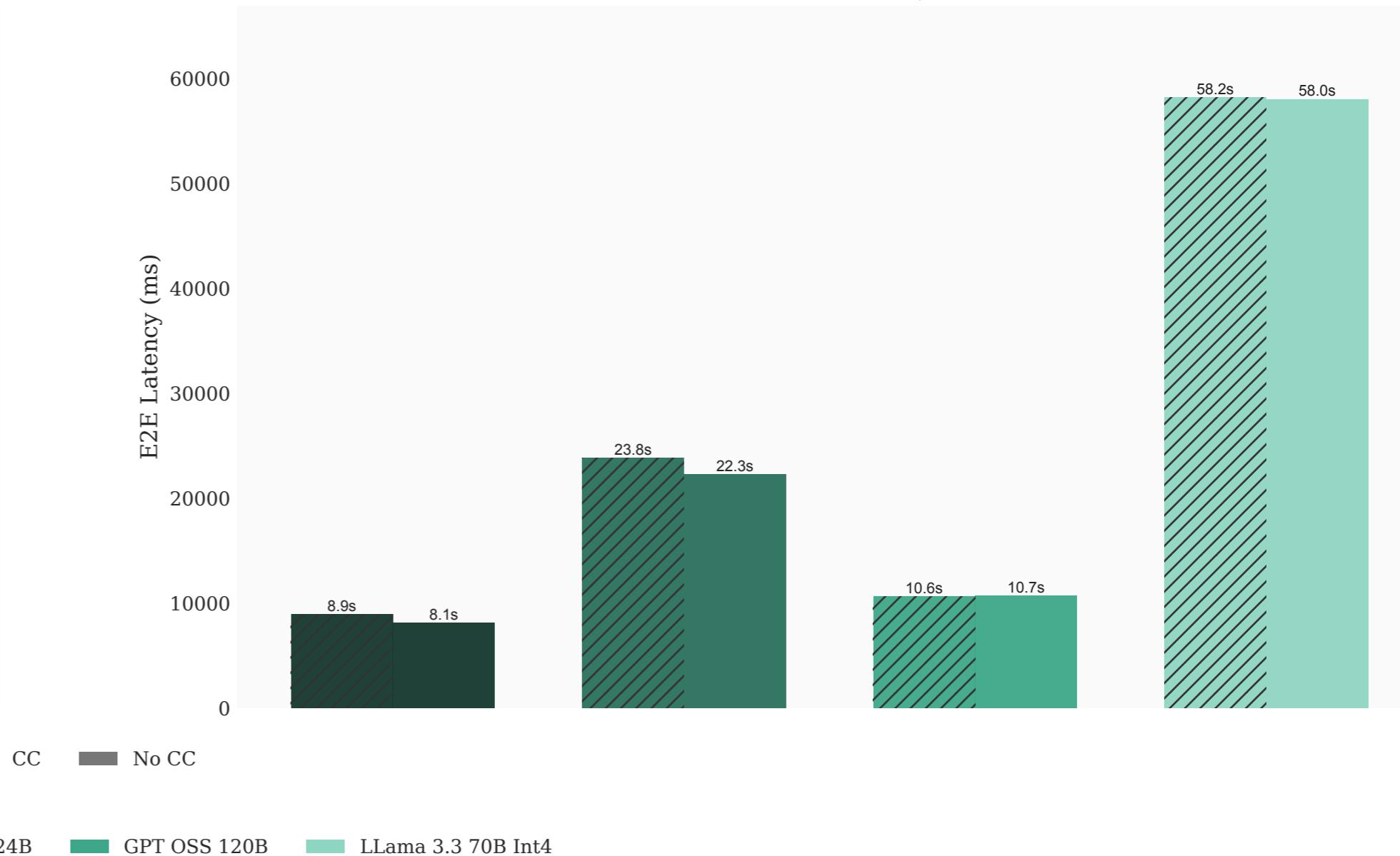
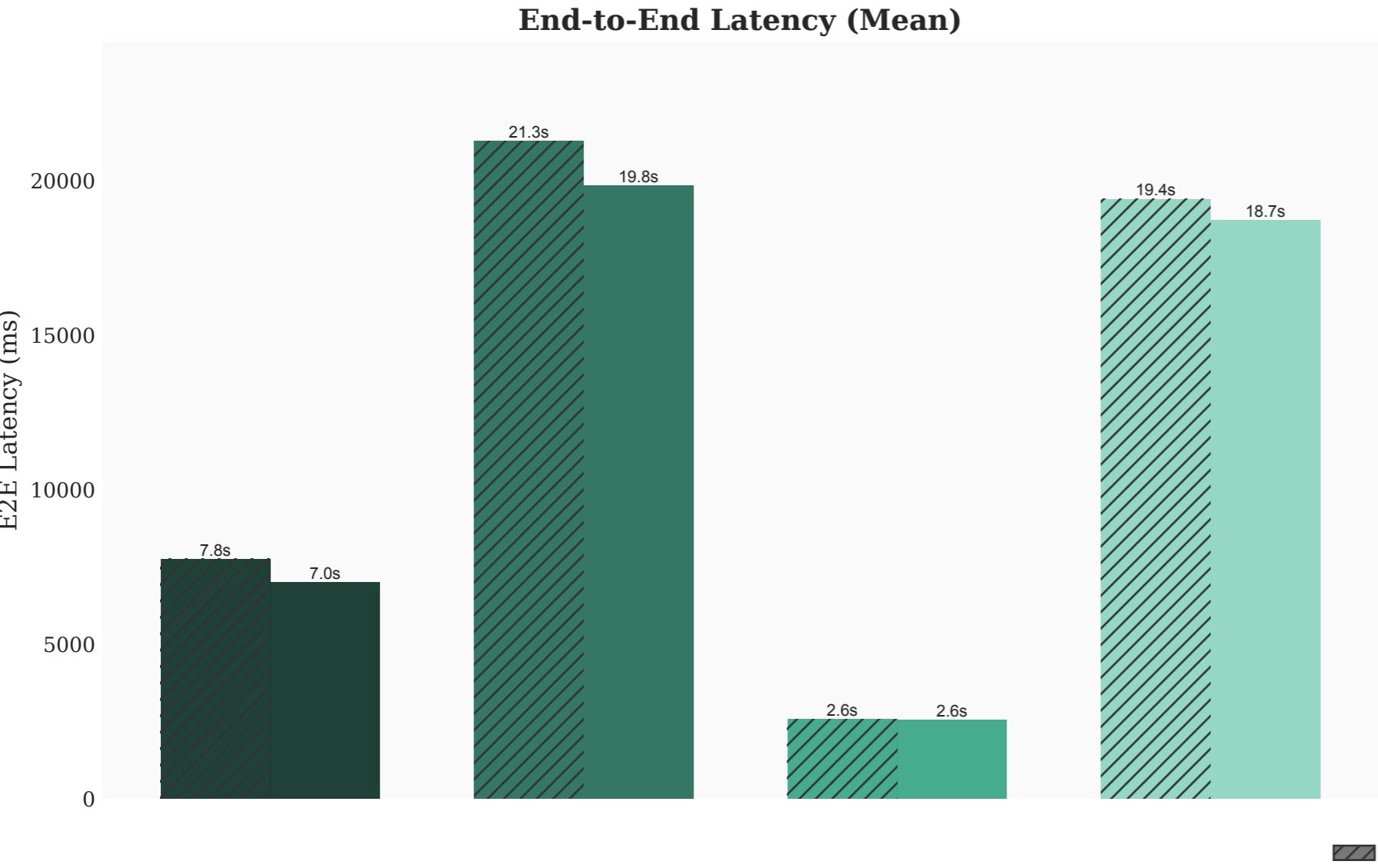


End-to-End Latency (P99)



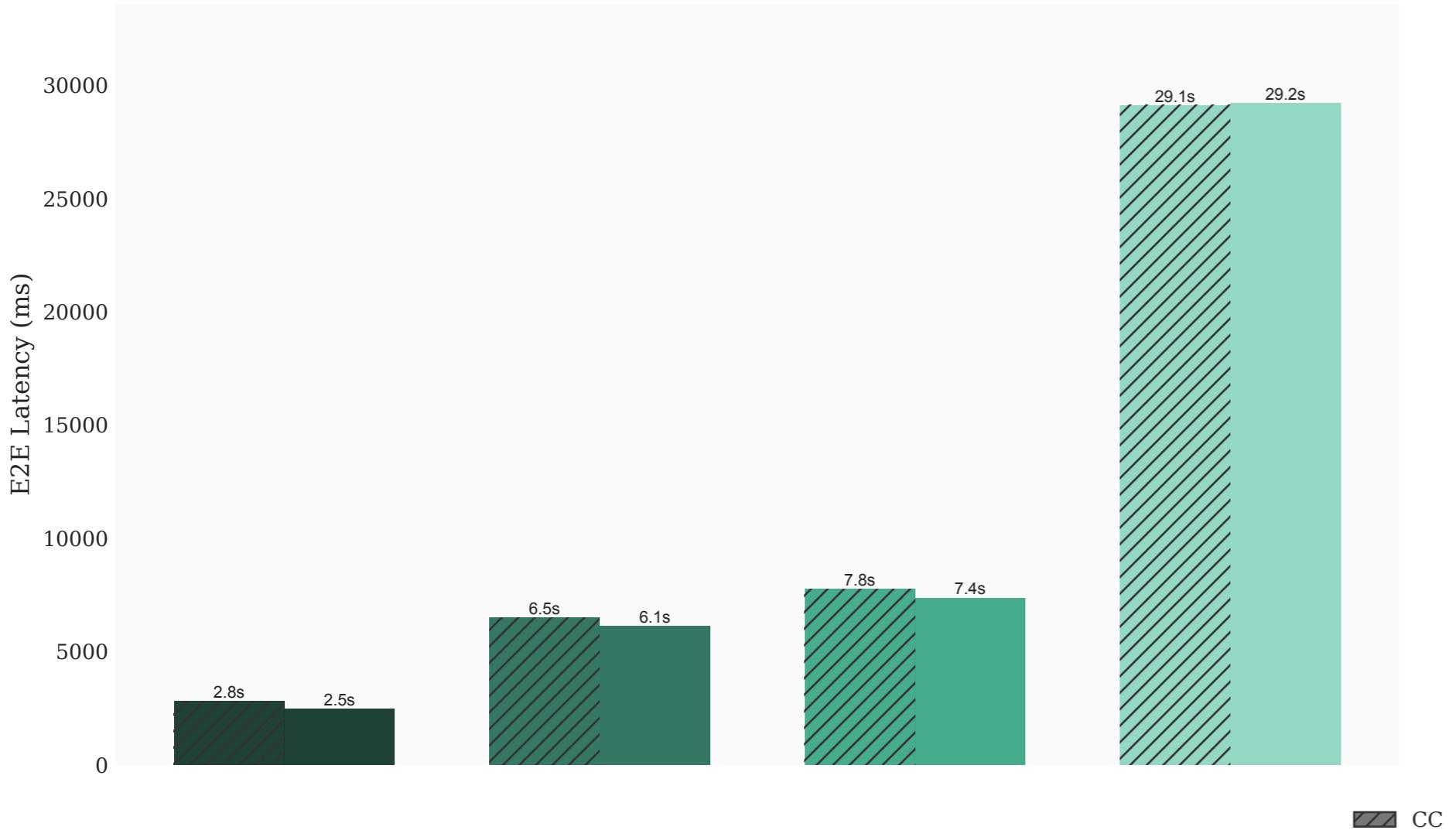
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (1000 \Rightarrow 1000) (Rate 1)

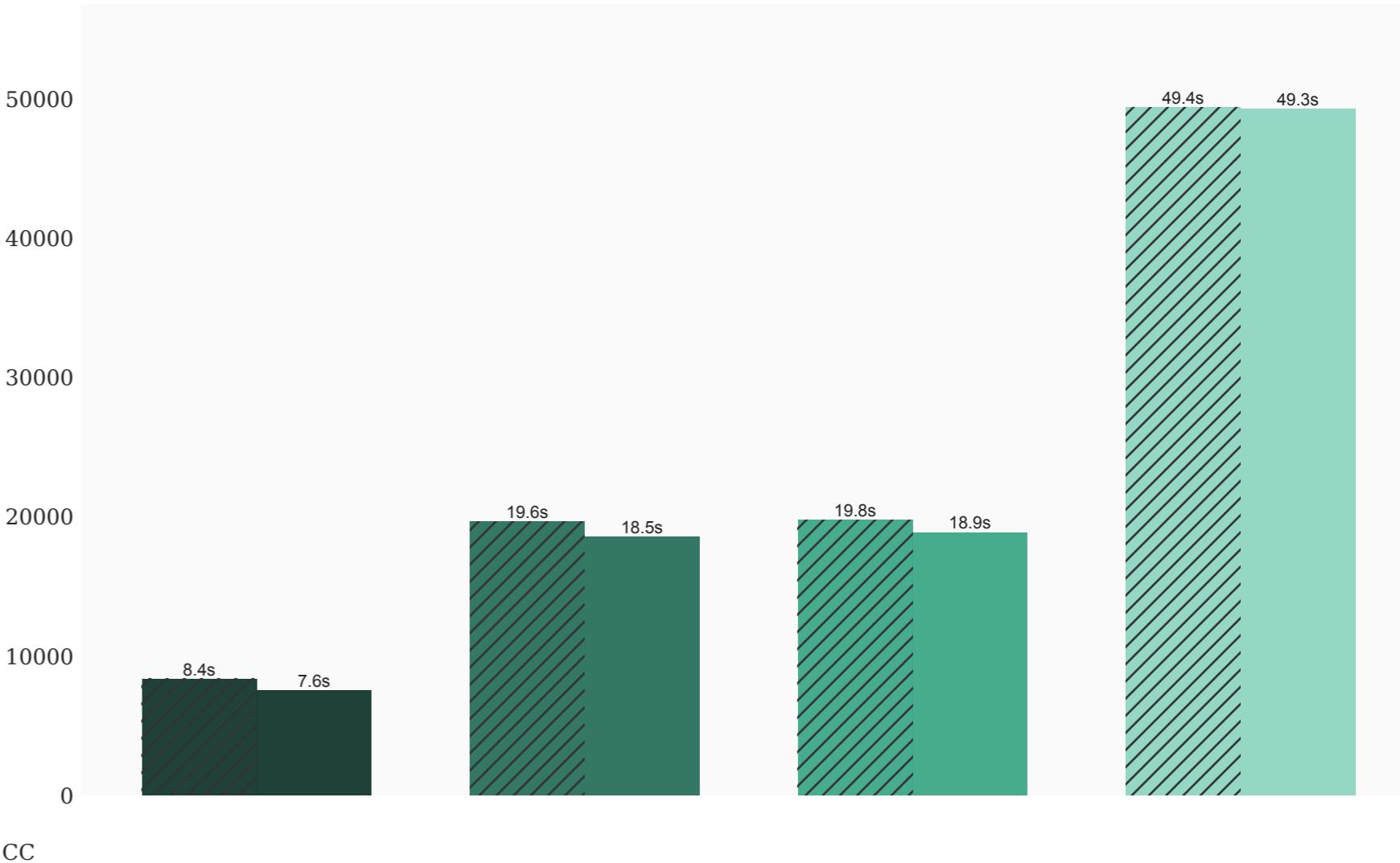


ShareGPT (Rate 100)

End-to-End Latency (Mean)



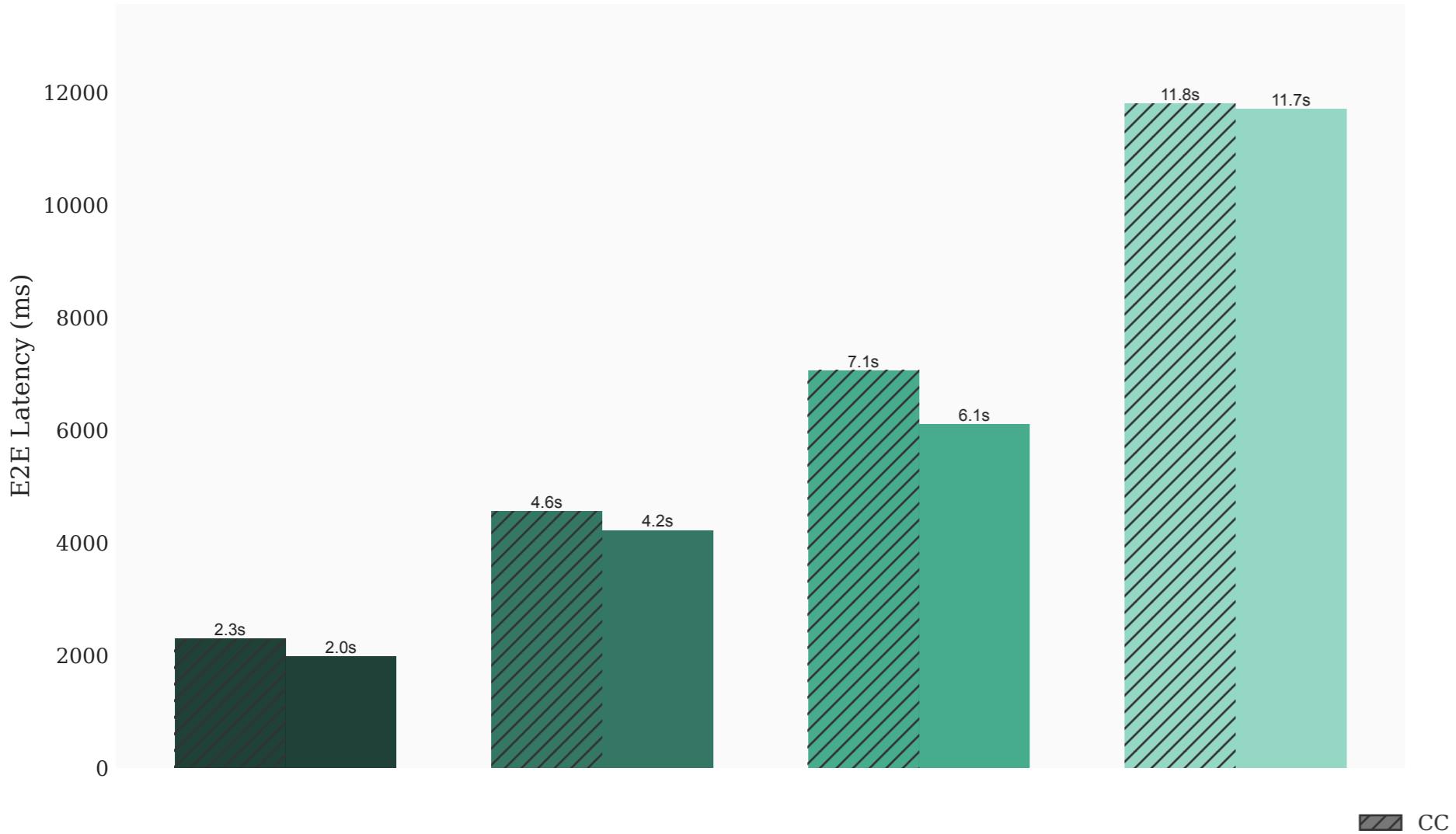
End-to-End Latency (P99)



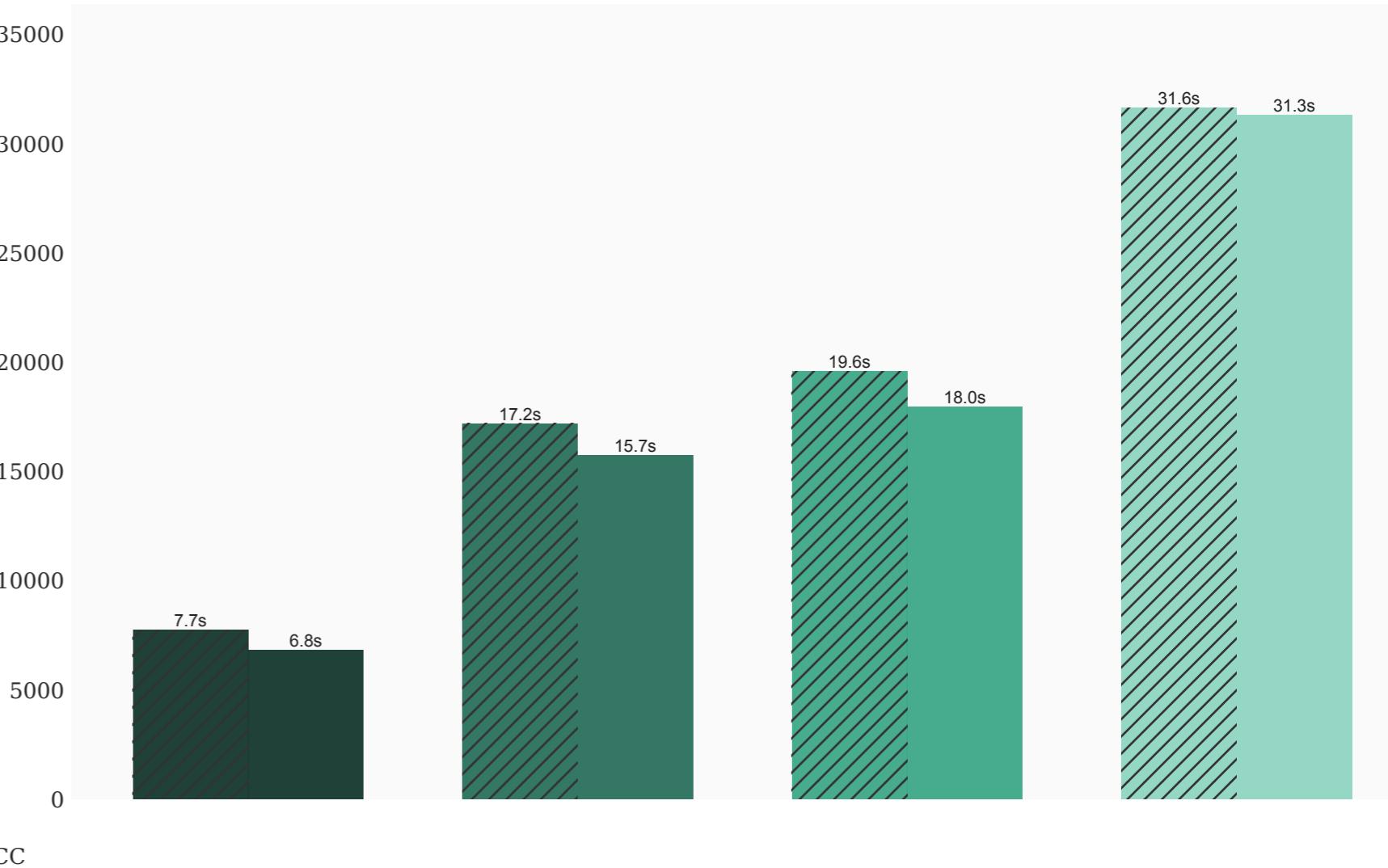
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

ShareGPT (Rate 50)

End-to-End Latency (Mean)

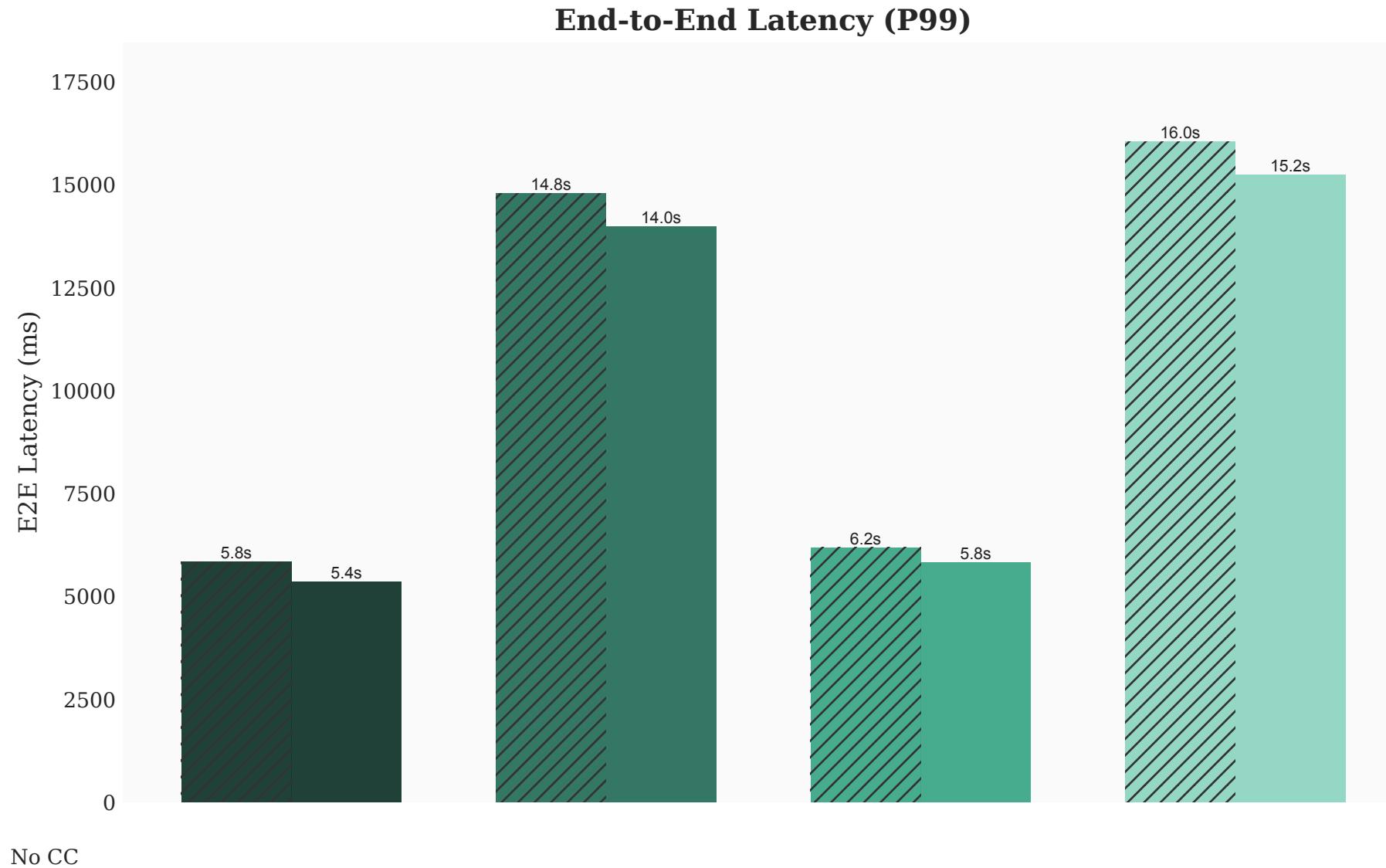
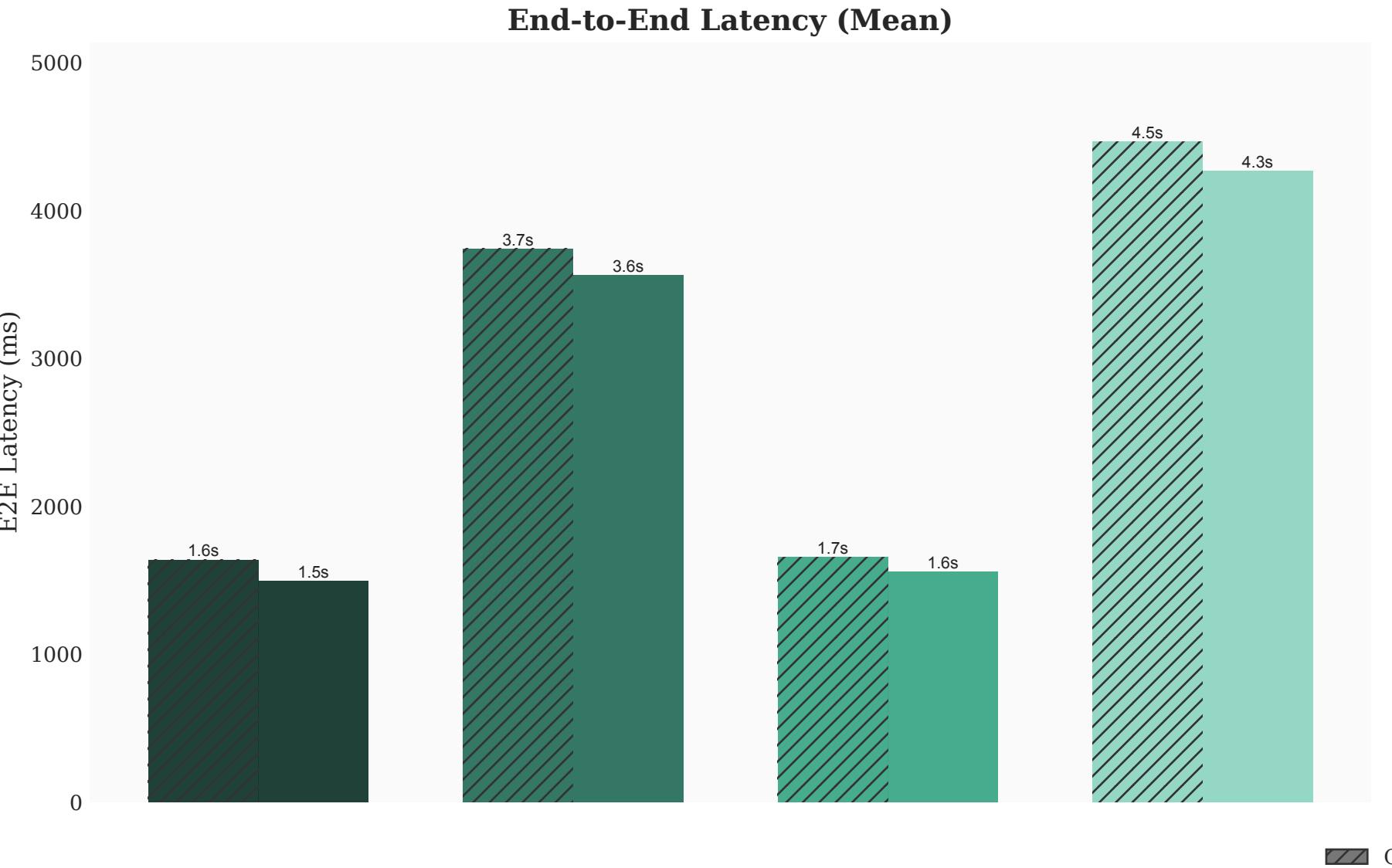


End-to-End Latency (P99)



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

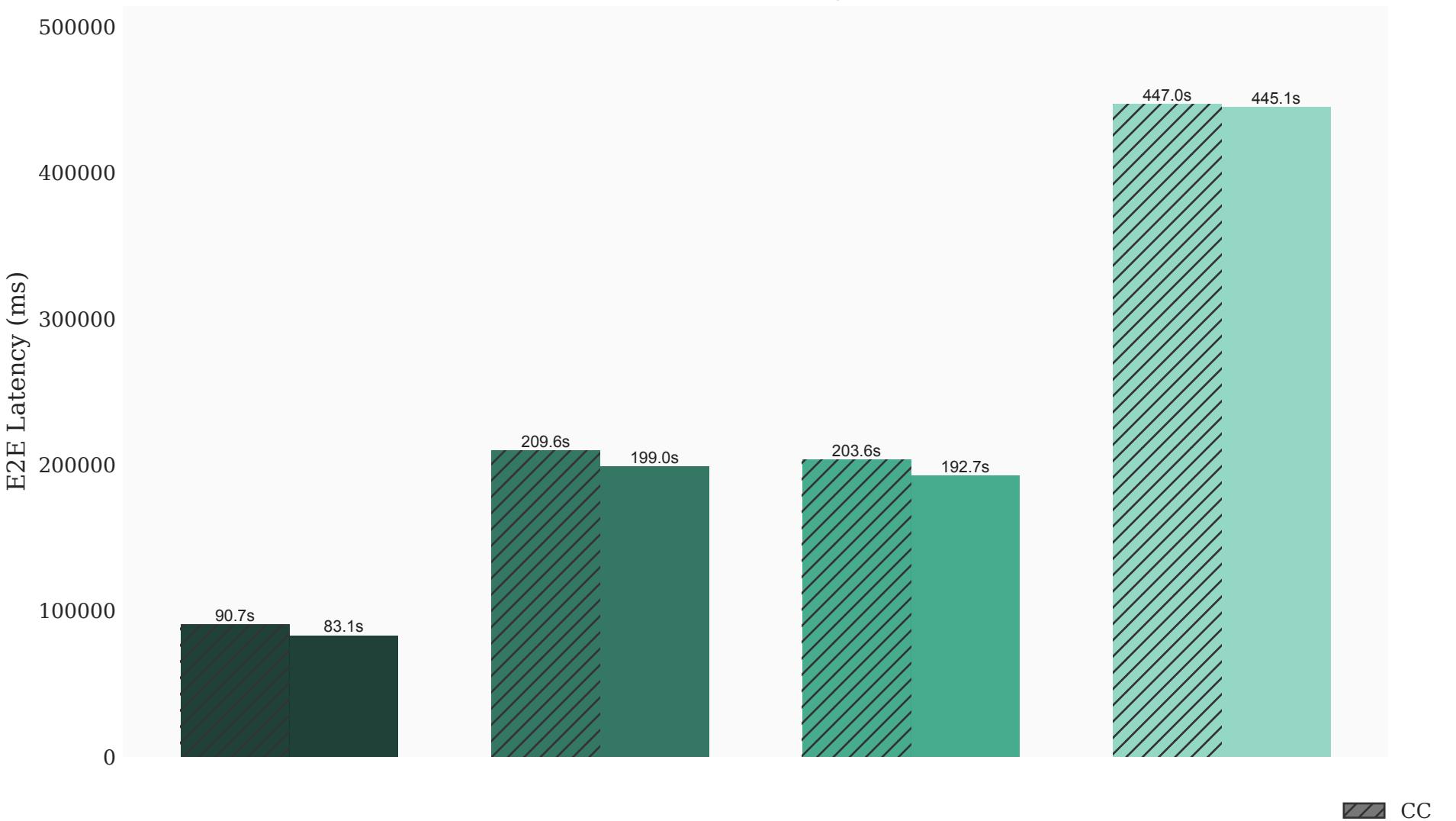
ShareGPT (Rate 1)



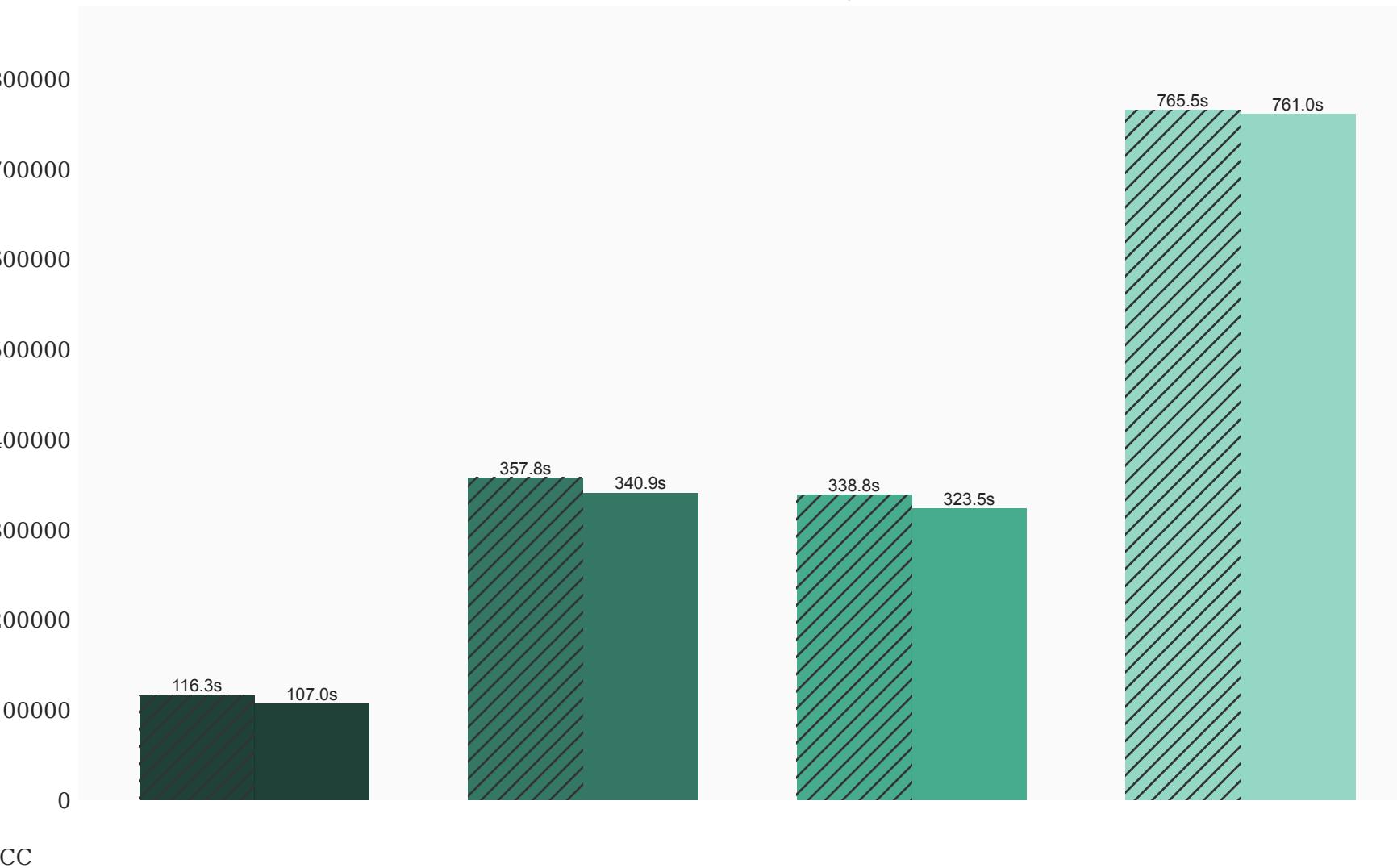
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Edit 10K Characters (Rate 100)

End-to-End Latency (Mean)



End-to-End Latency (P99)



LLama 3.1 8B

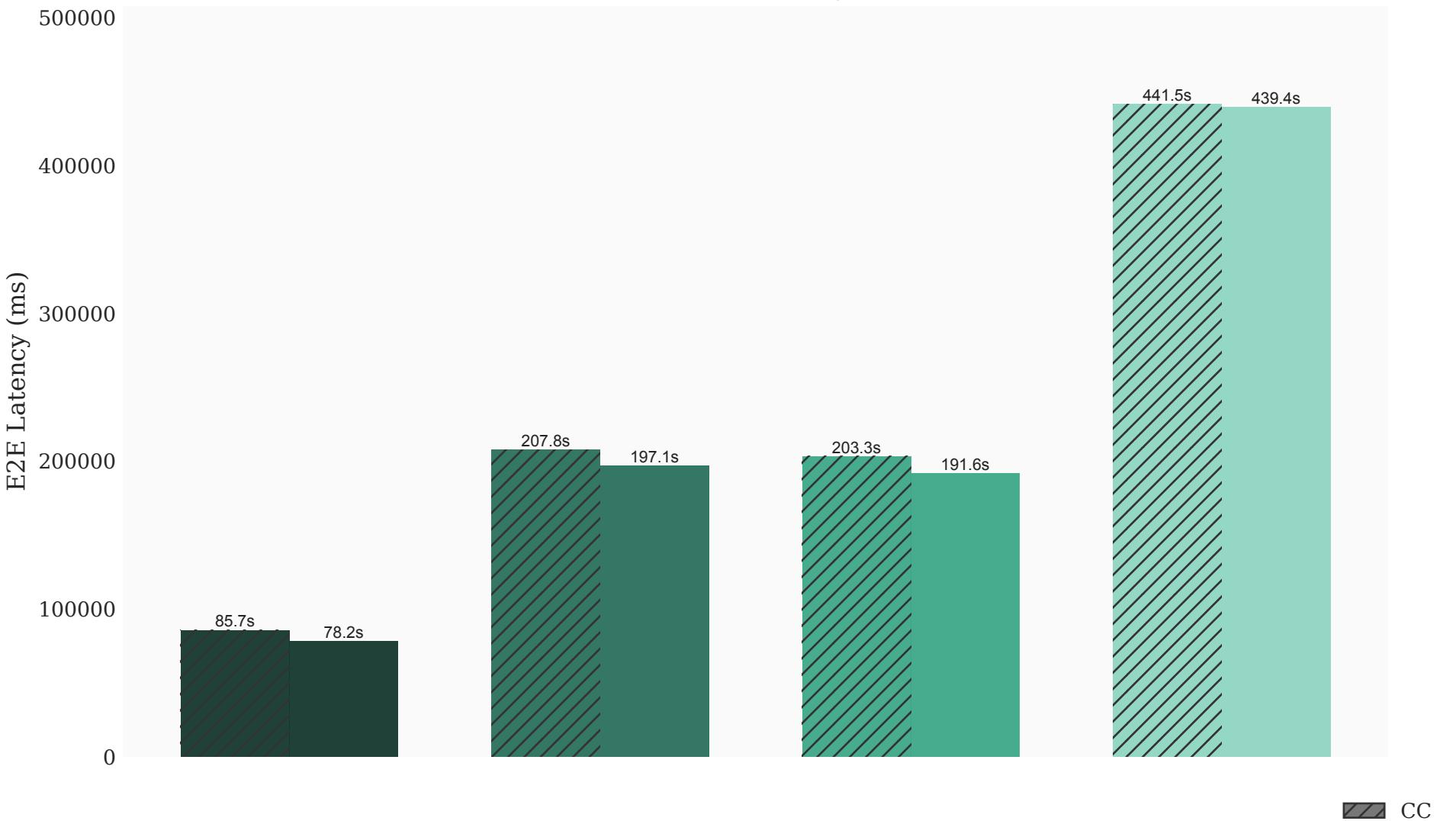
Mistral 3.1 24B

GPT OSS 120B

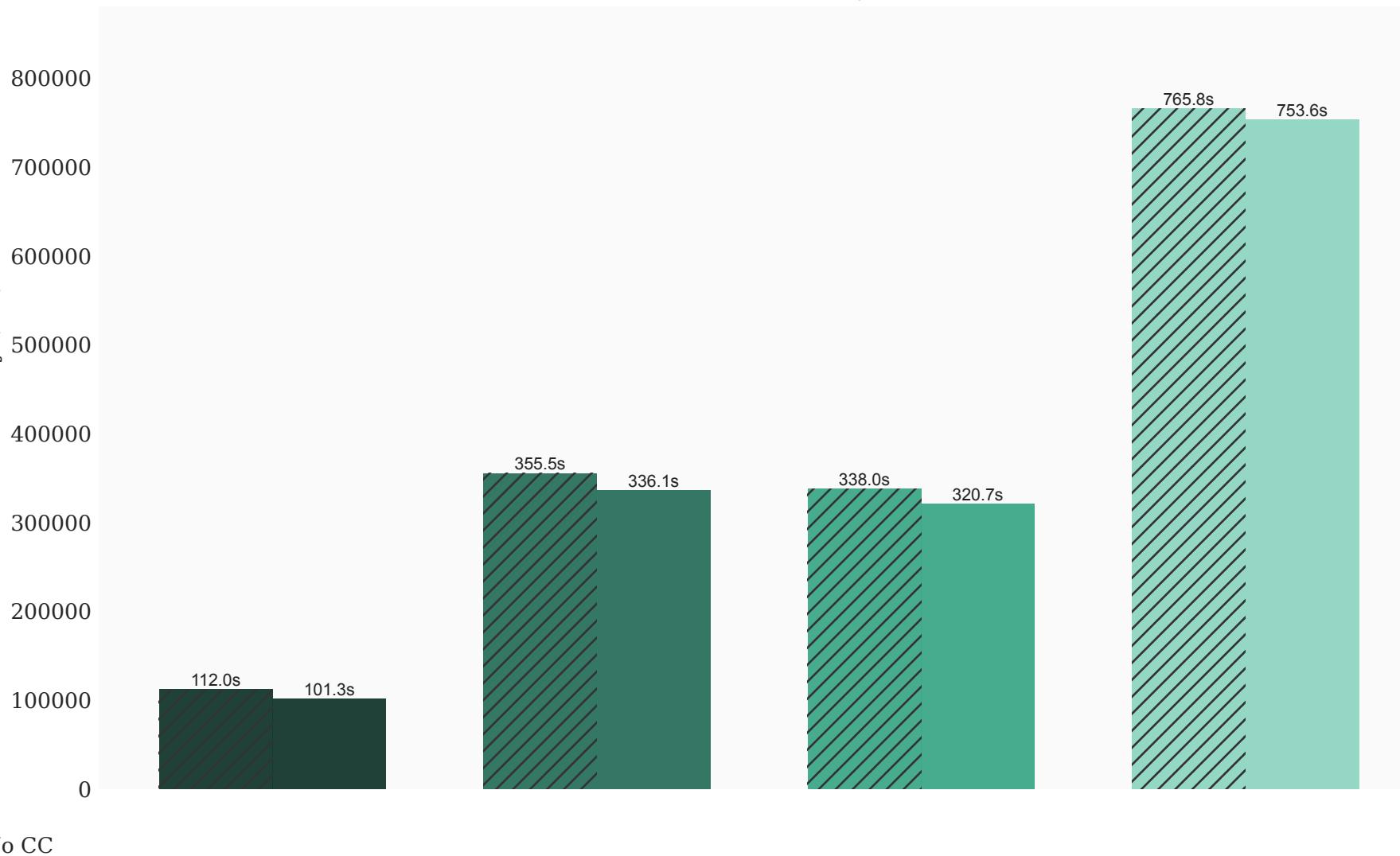
LLama 3.3 70B Int4

Edit 10K Characters (Rate 50)

End-to-End Latency (Mean)



End-to-End Latency (P99)



LLama 3.1 8B

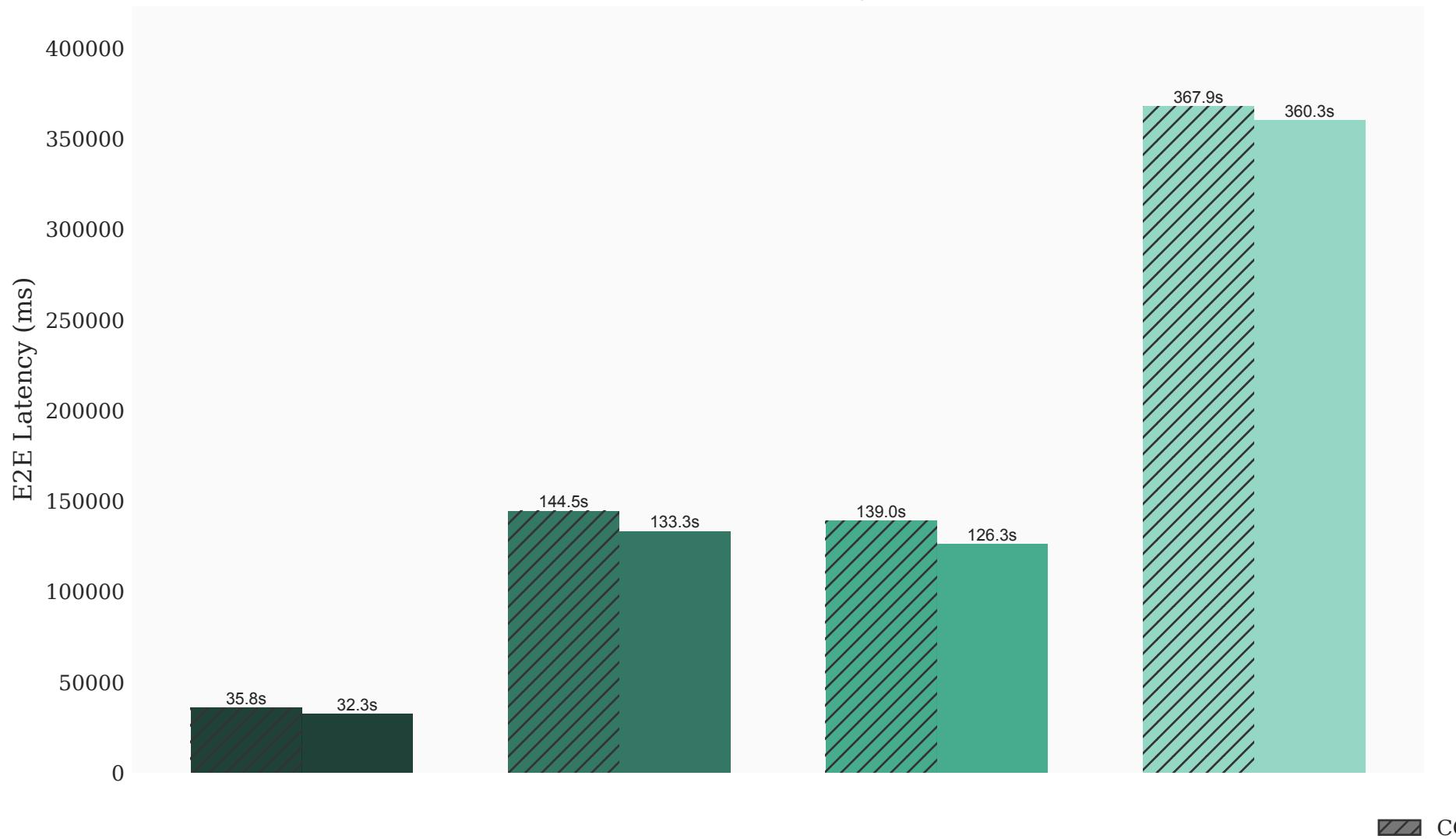
Mistral 3.1 24B

GPT OSS 120B

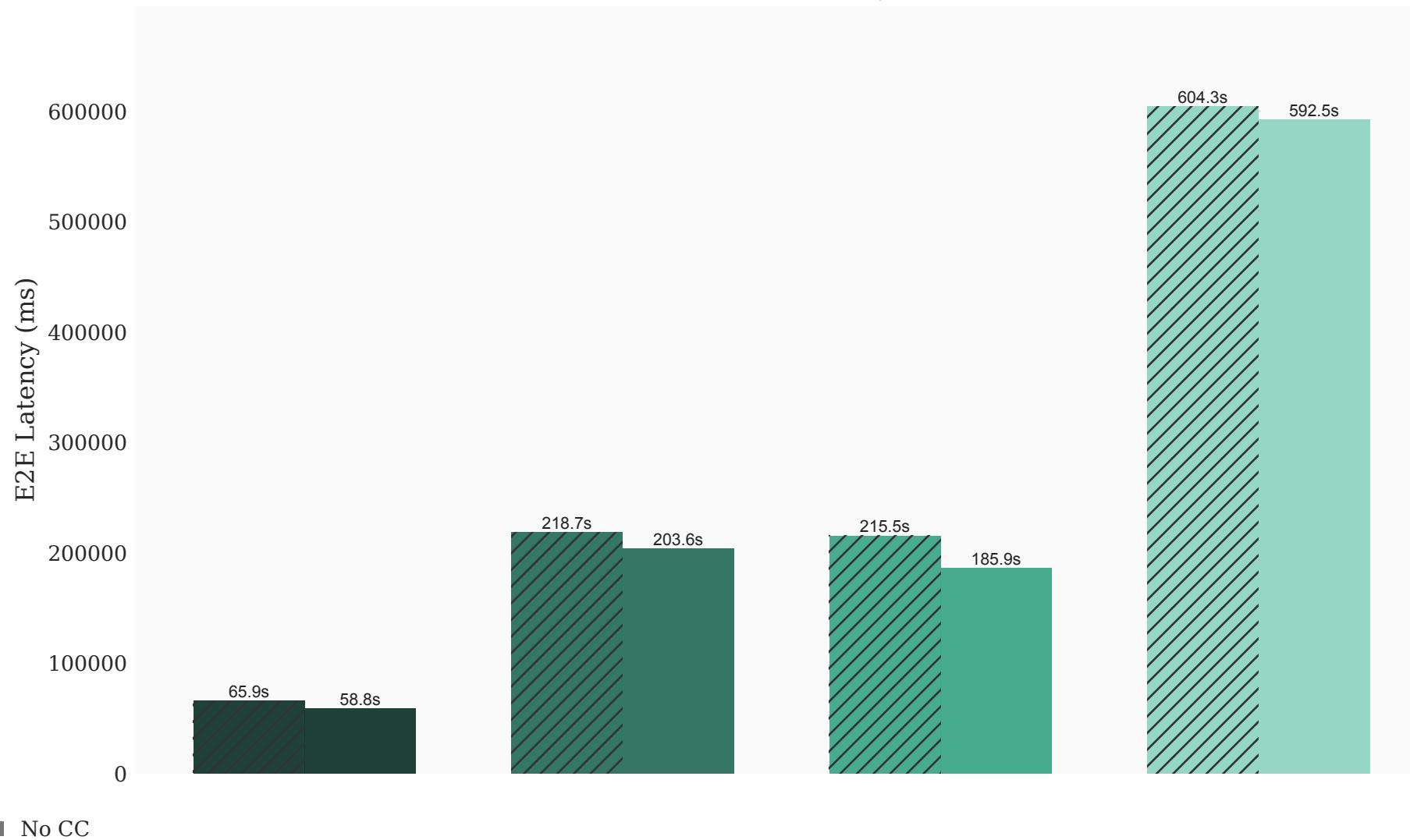
LLama 3.3 70B Int4

Edit 10K Characters (Rate 1)

End-to-End Latency (Mean)



End-to-End Latency (P99)



LLama 3.1 8B

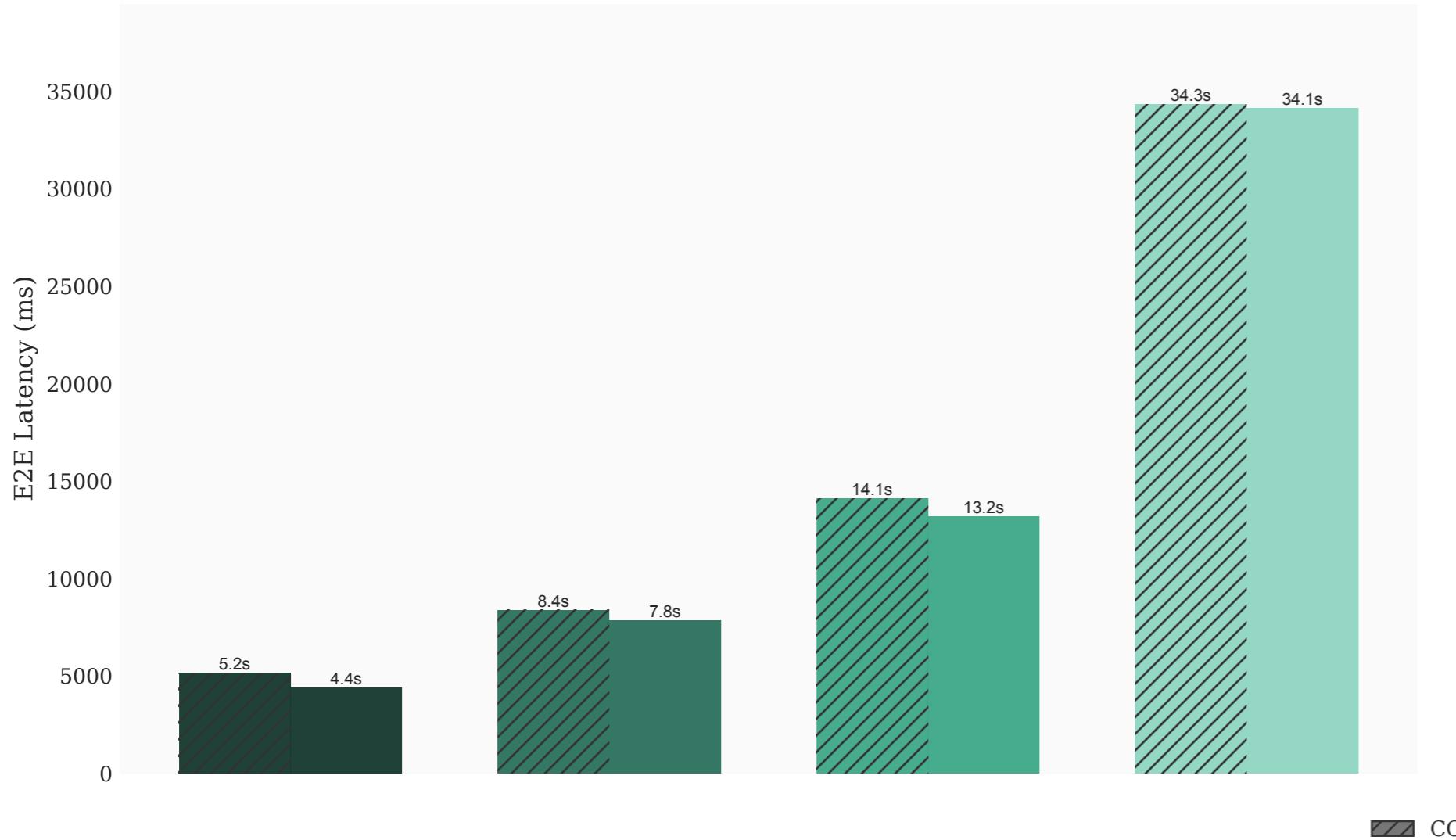
Mistral 3.1 24B

GPT OSS 120B

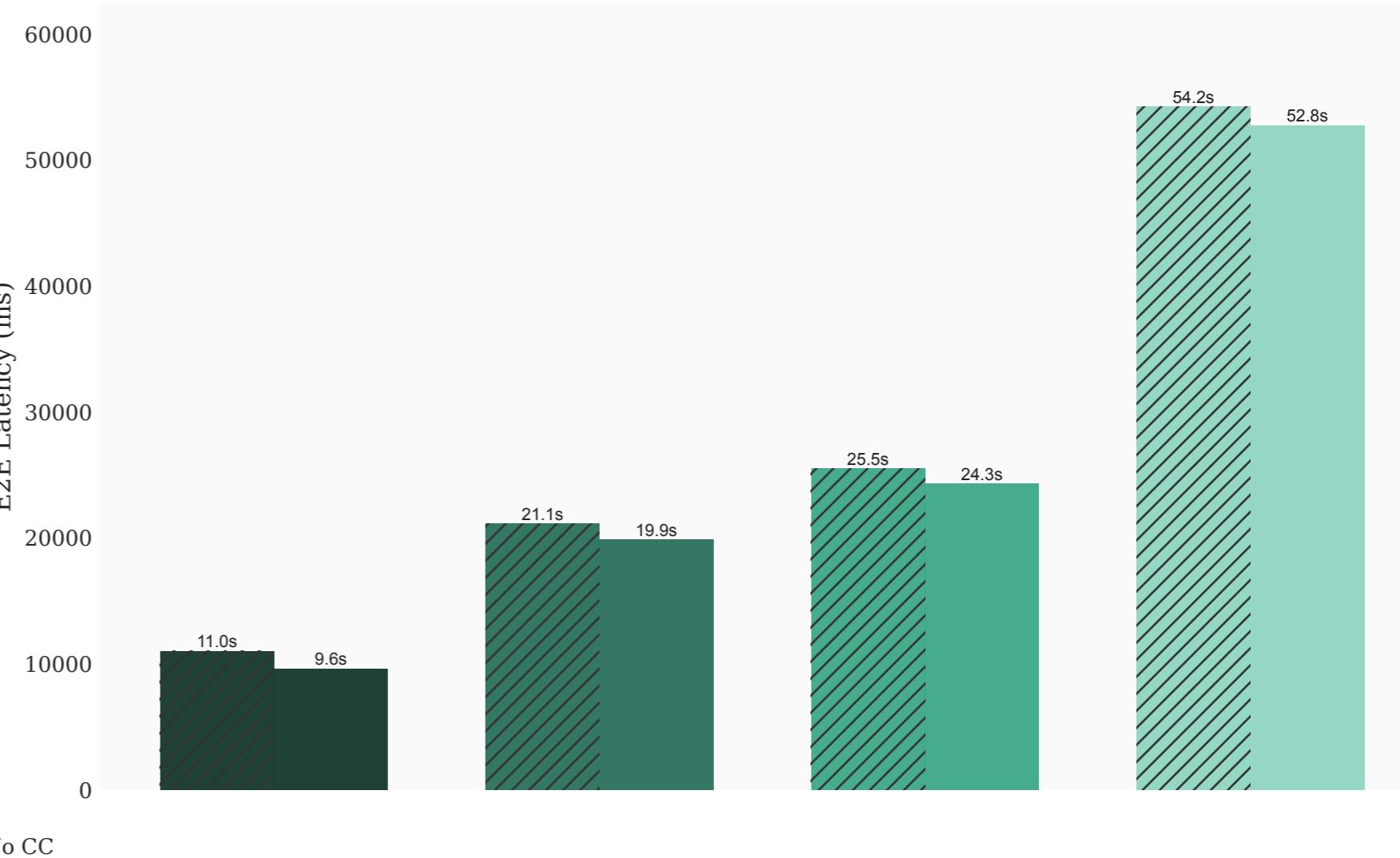
LLama 3.3 70B Int4

Numina Math (Rate 100)

End-to-End Latency (Mean)



End-to-End Latency (P99)



LLama 3.1 8B

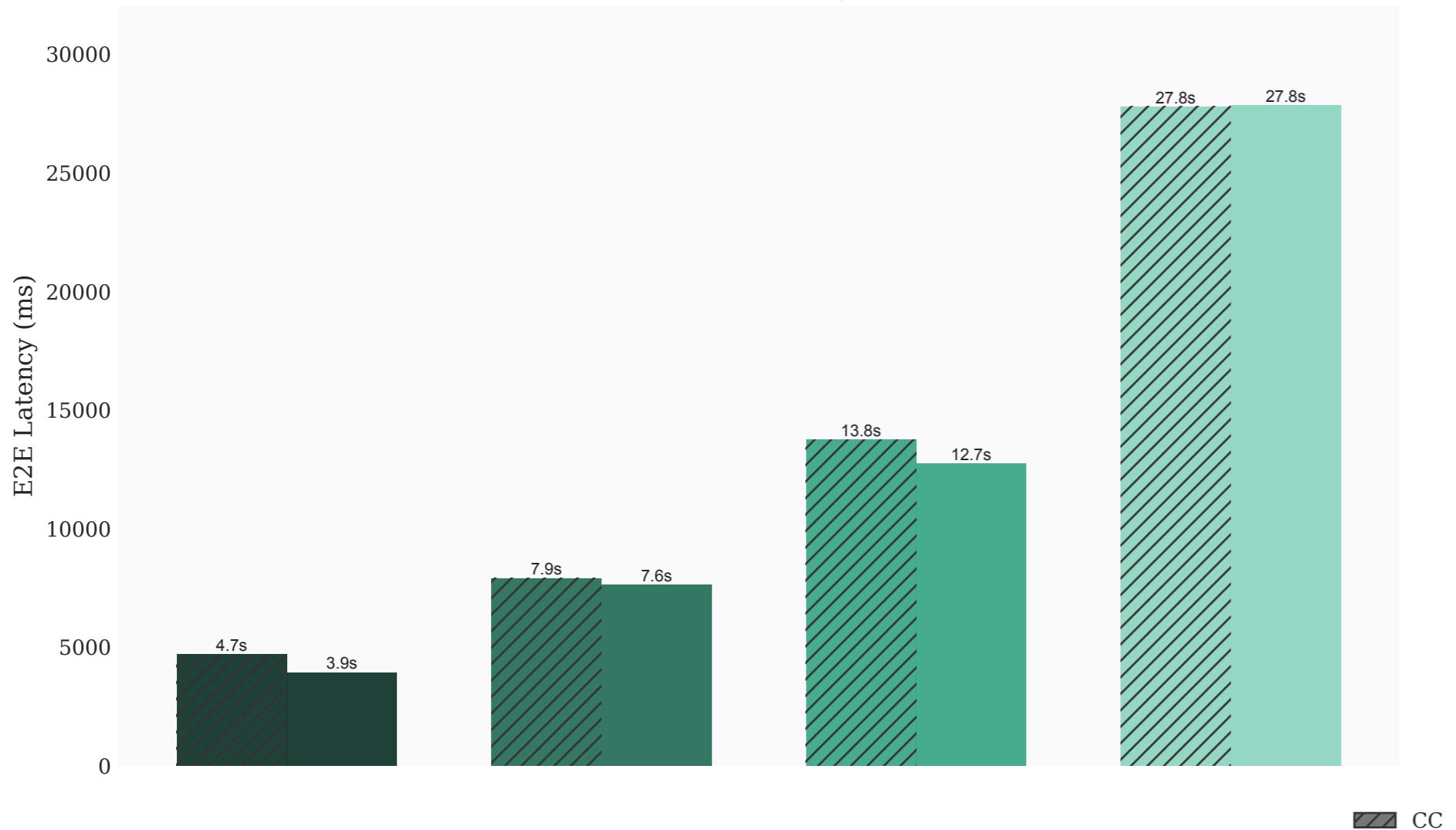
Mistral 3.1 24B

GPT OSS 120B

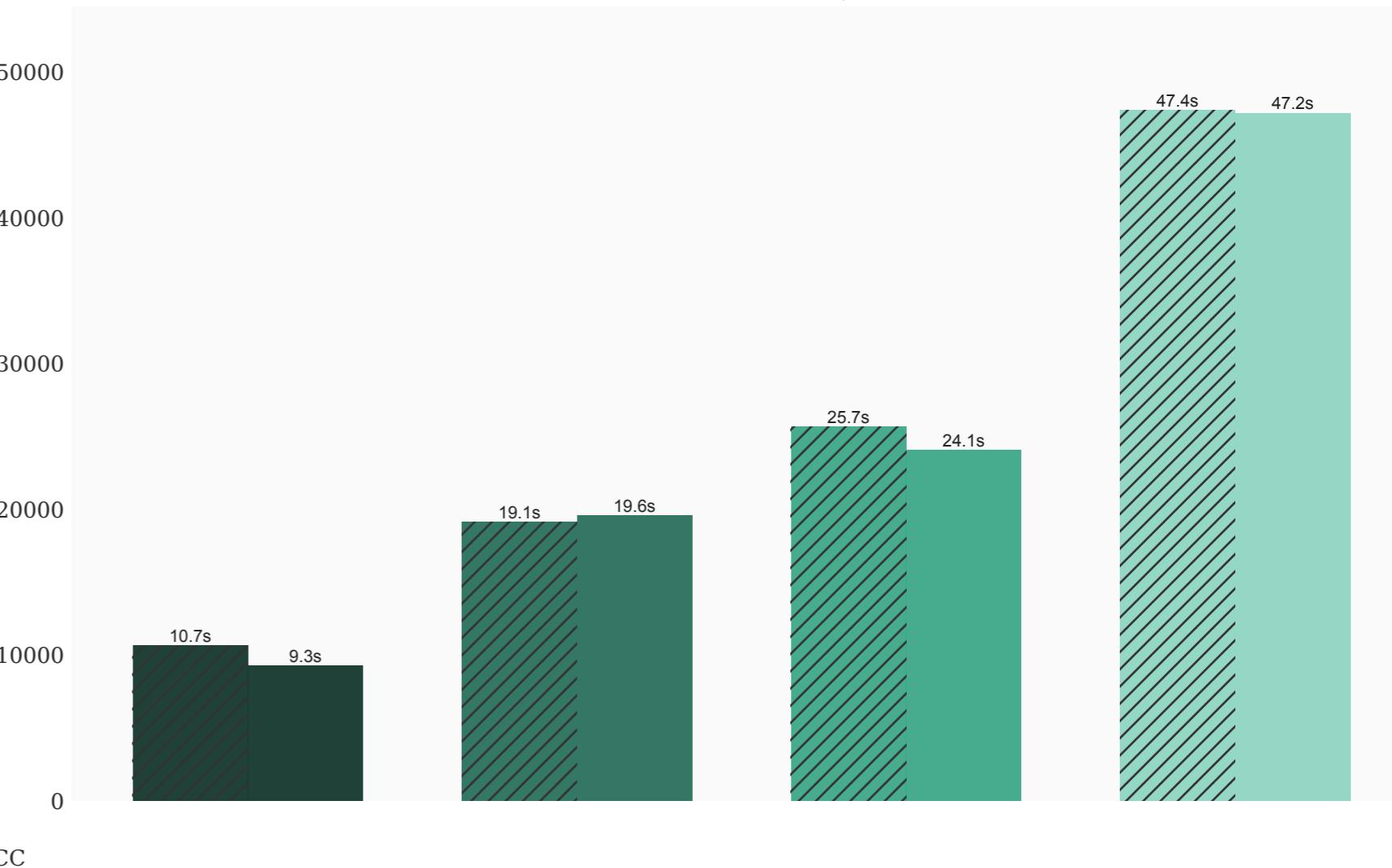
LLama 3.3 70B Int4

Numina Math (Rate 50)

End-to-End Latency (Mean)



End-to-End Latency (P99)



LLama 3.1 8B

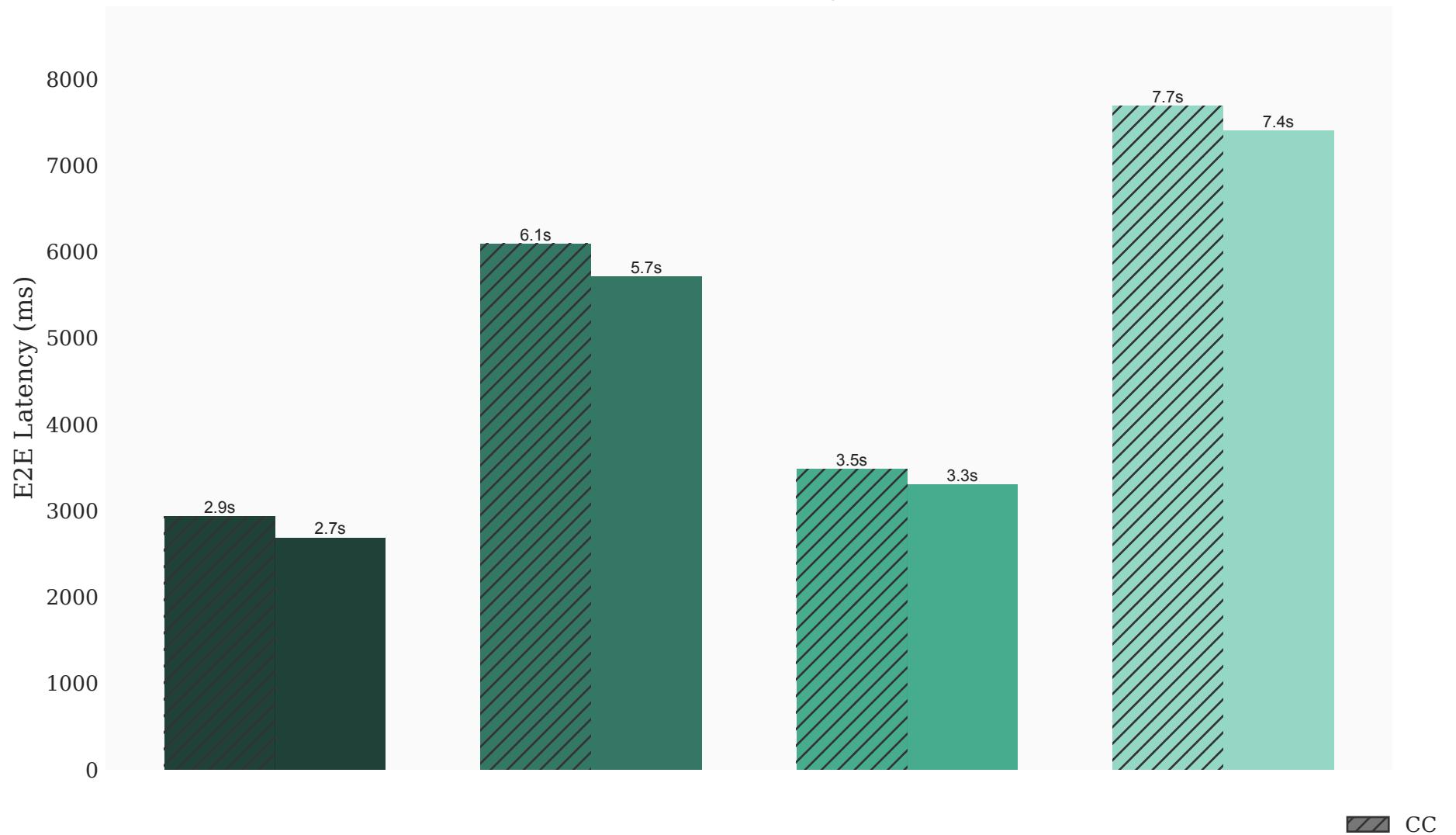
Mistral 3.1 24B

GPT OSS 120B

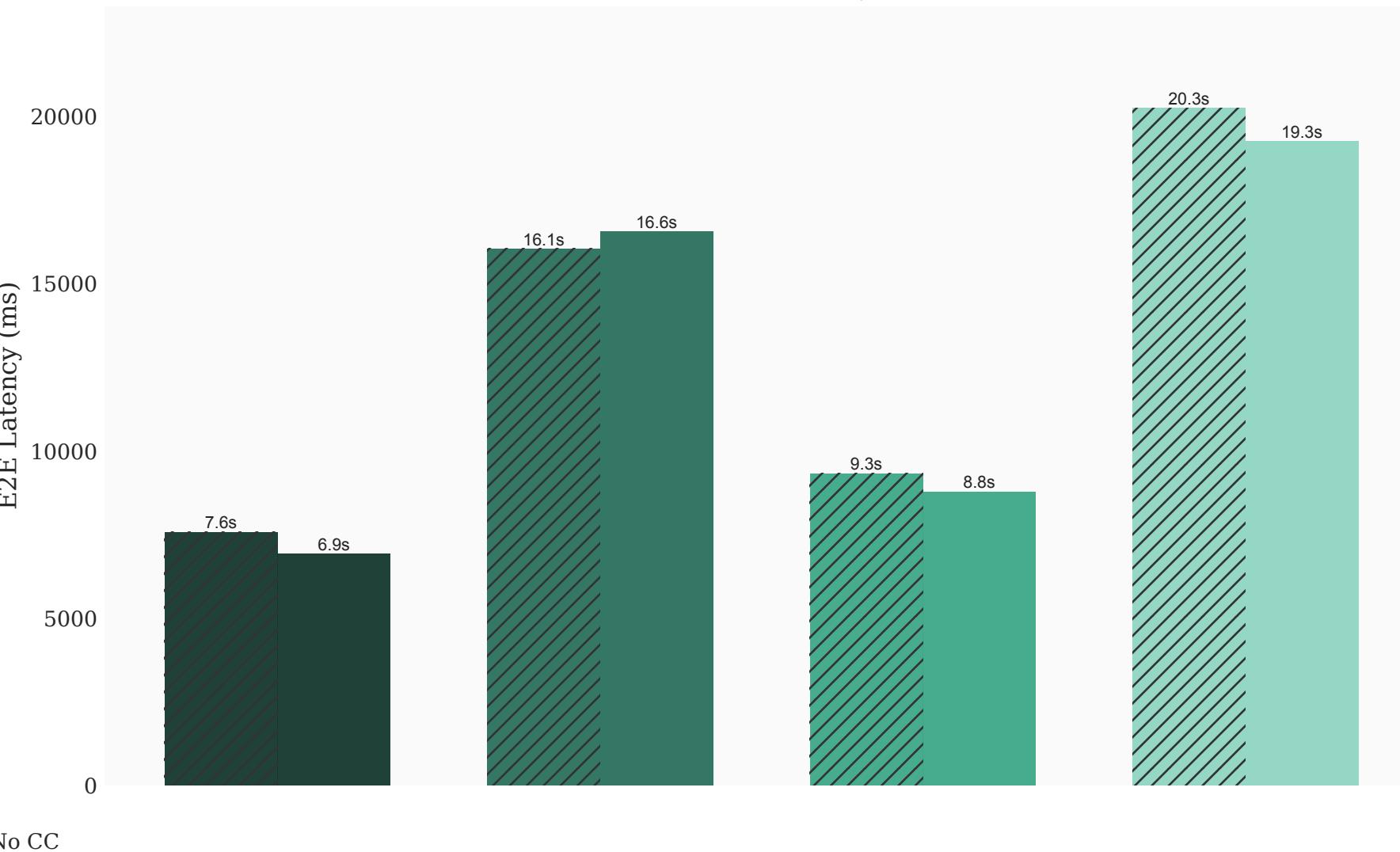
LLama 3.3 70B Int4

Numina Math (Rate 1)

End-to-End Latency (Mean)



End-to-End Latency (P99)



LLama 3.1 8B

Mistral 3.1 24B

GPT OSS 120B

LLama 3.3 70B Int4