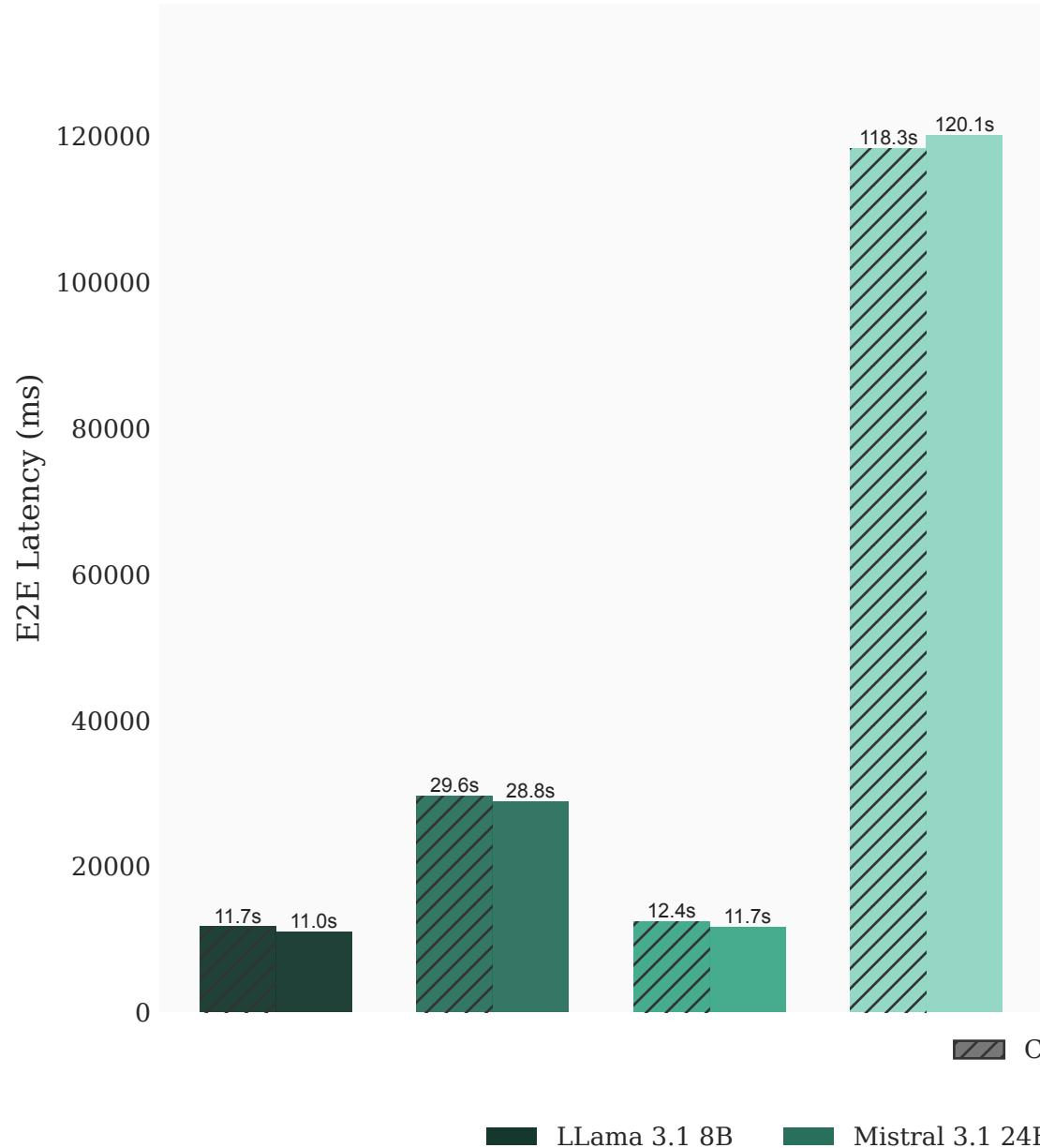
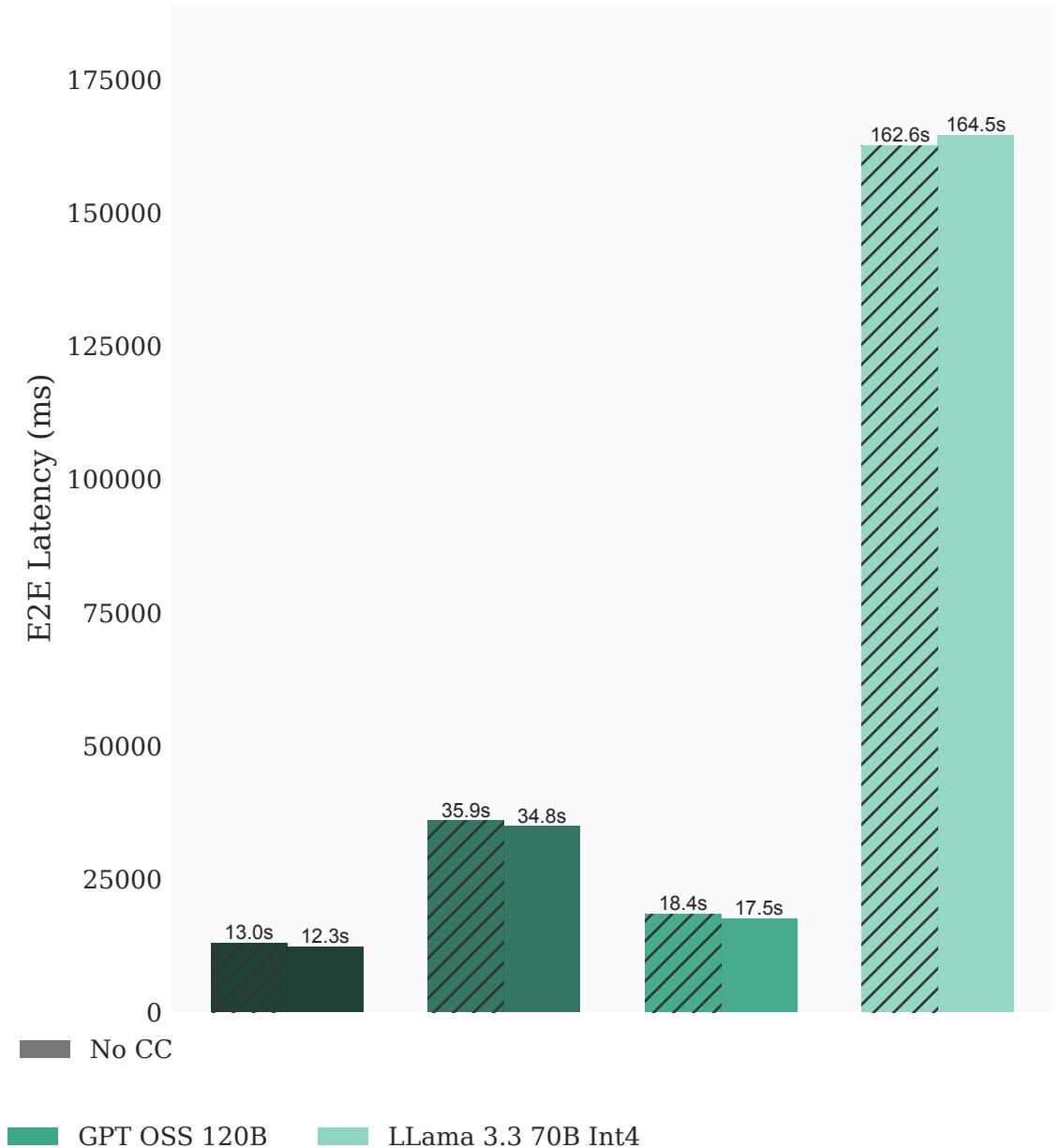


Random (1500 \Rightarrow 250) (100 Request Rate)

End-to-End Latency (Mean)

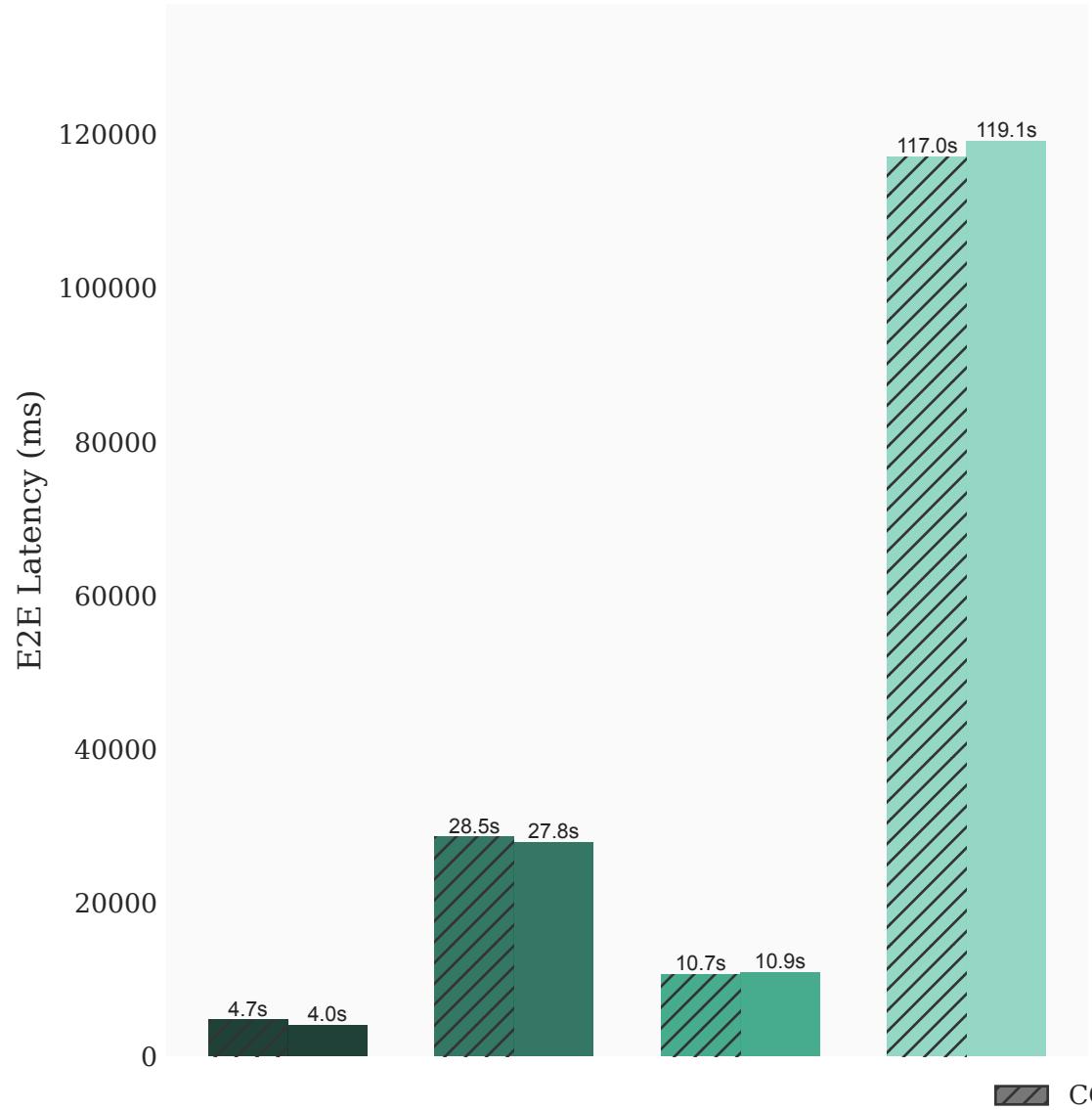


End-to-End Latency (P99)

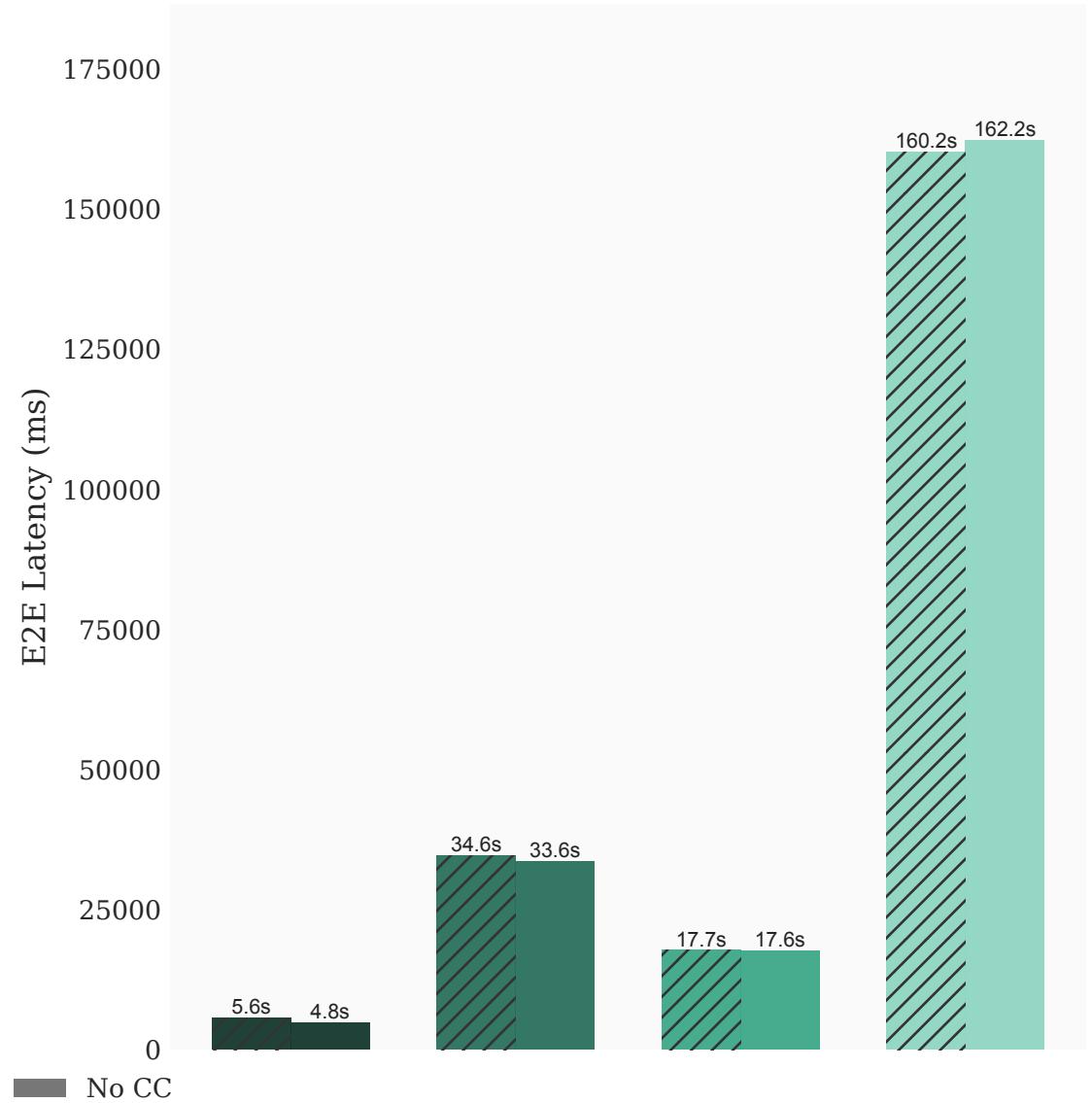


Random (1500 \Rightarrow 250) (50 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

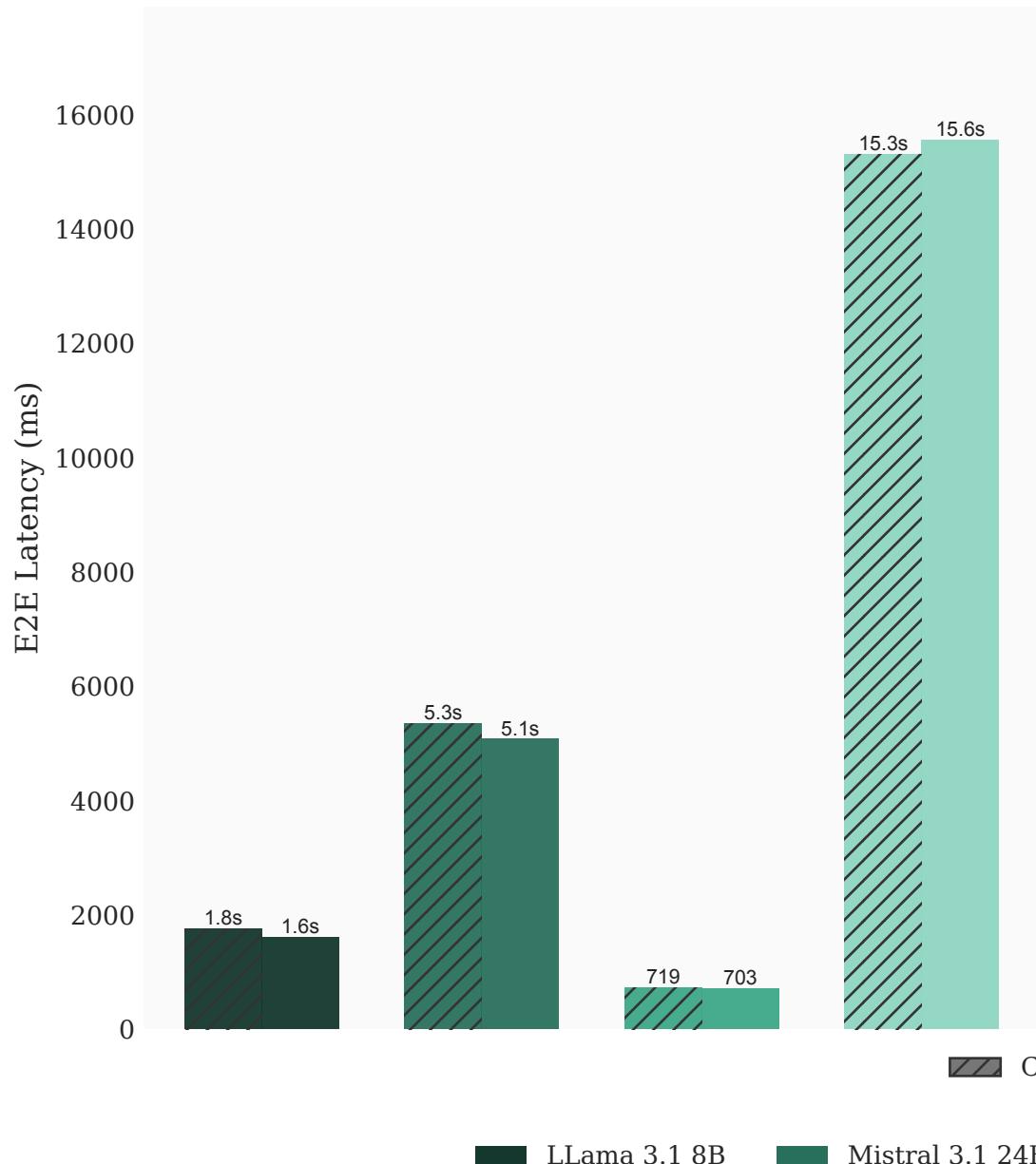
■ Mistral 3.1 24B

■ GPT OSS 120B

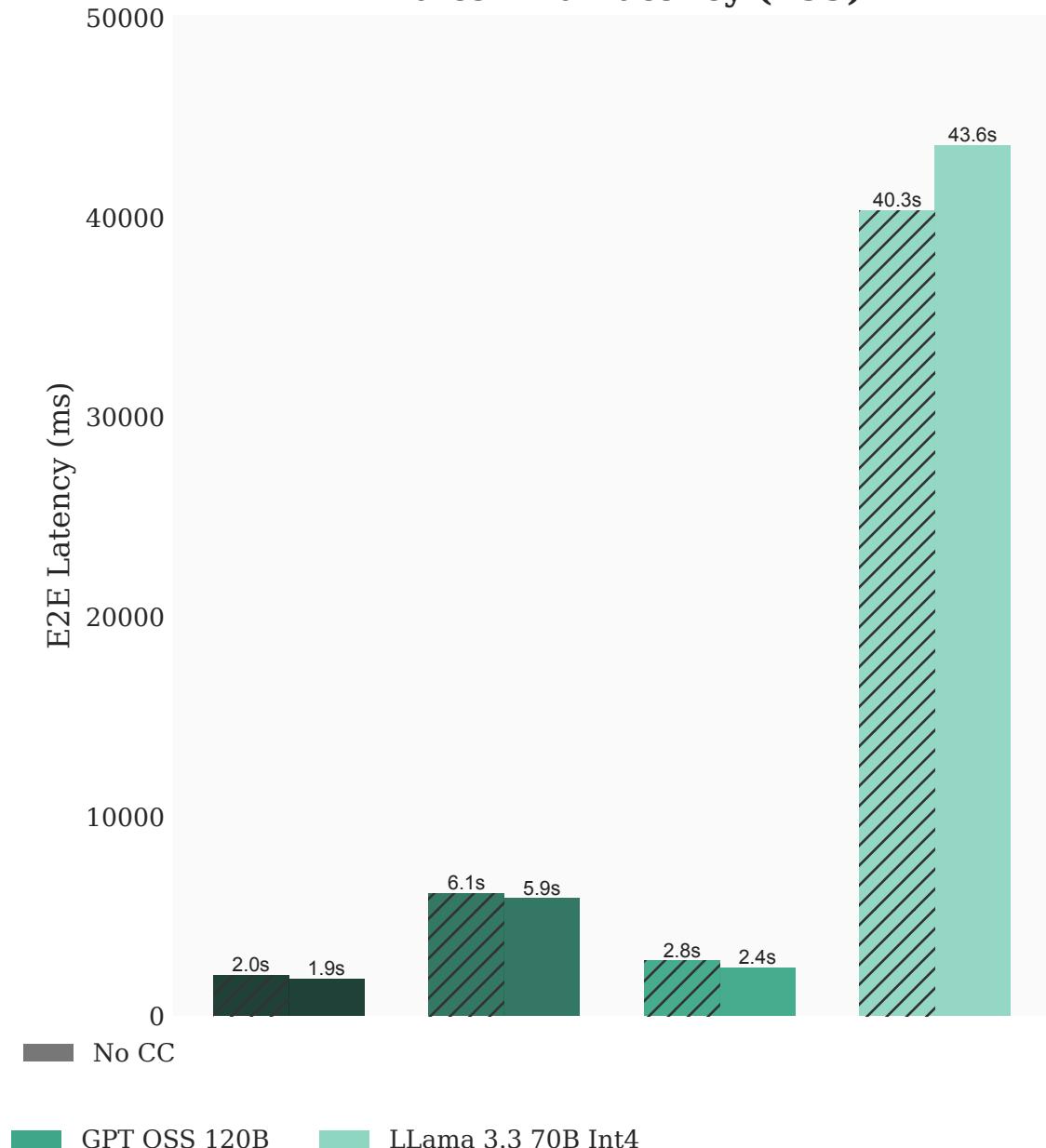
■ LLama 3.3 70B Int4

Random (1500 \Rightarrow 250) (Single Request)

End-to-End Latency (Mean)

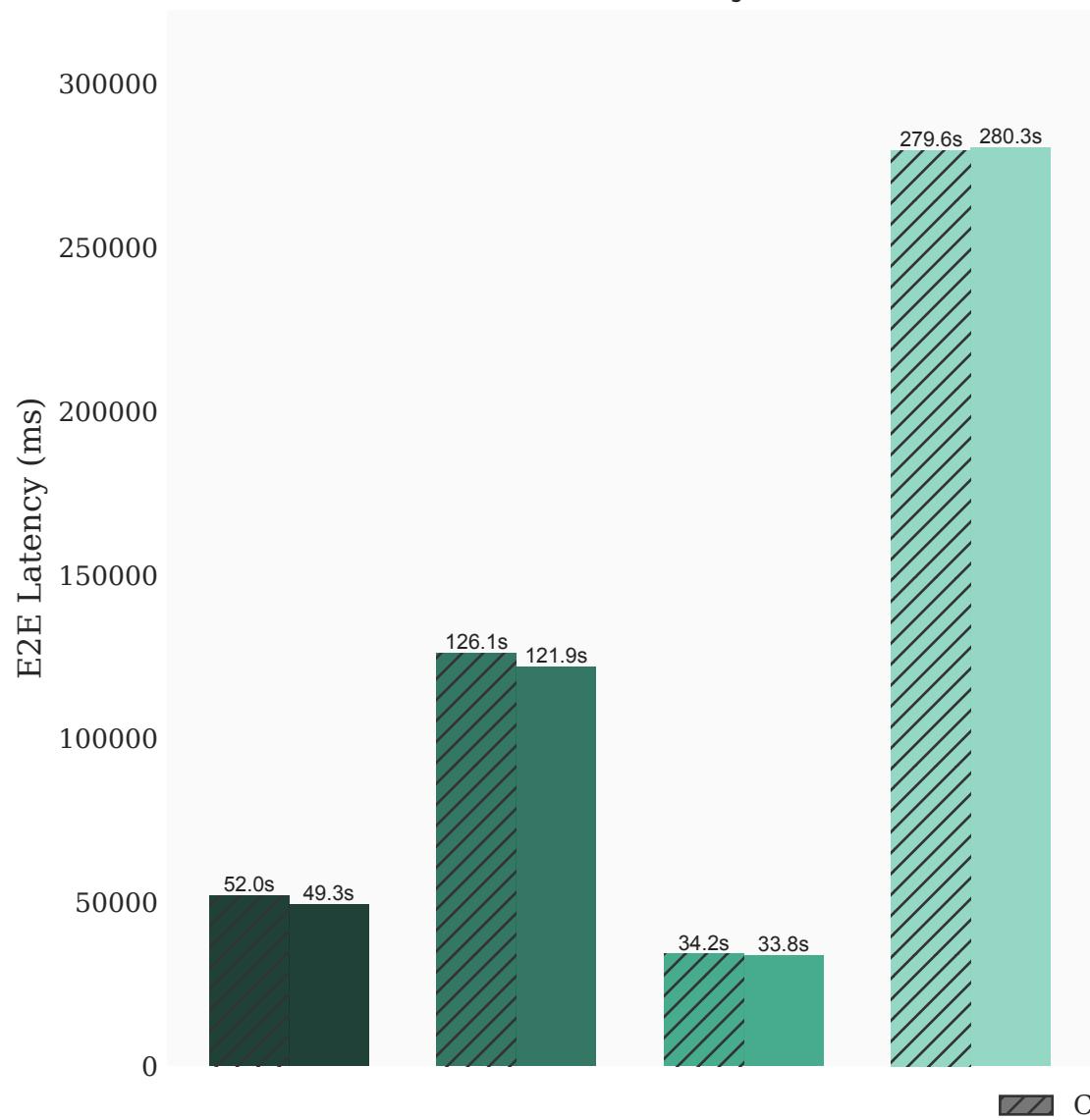


End-to-End Latency (P99)

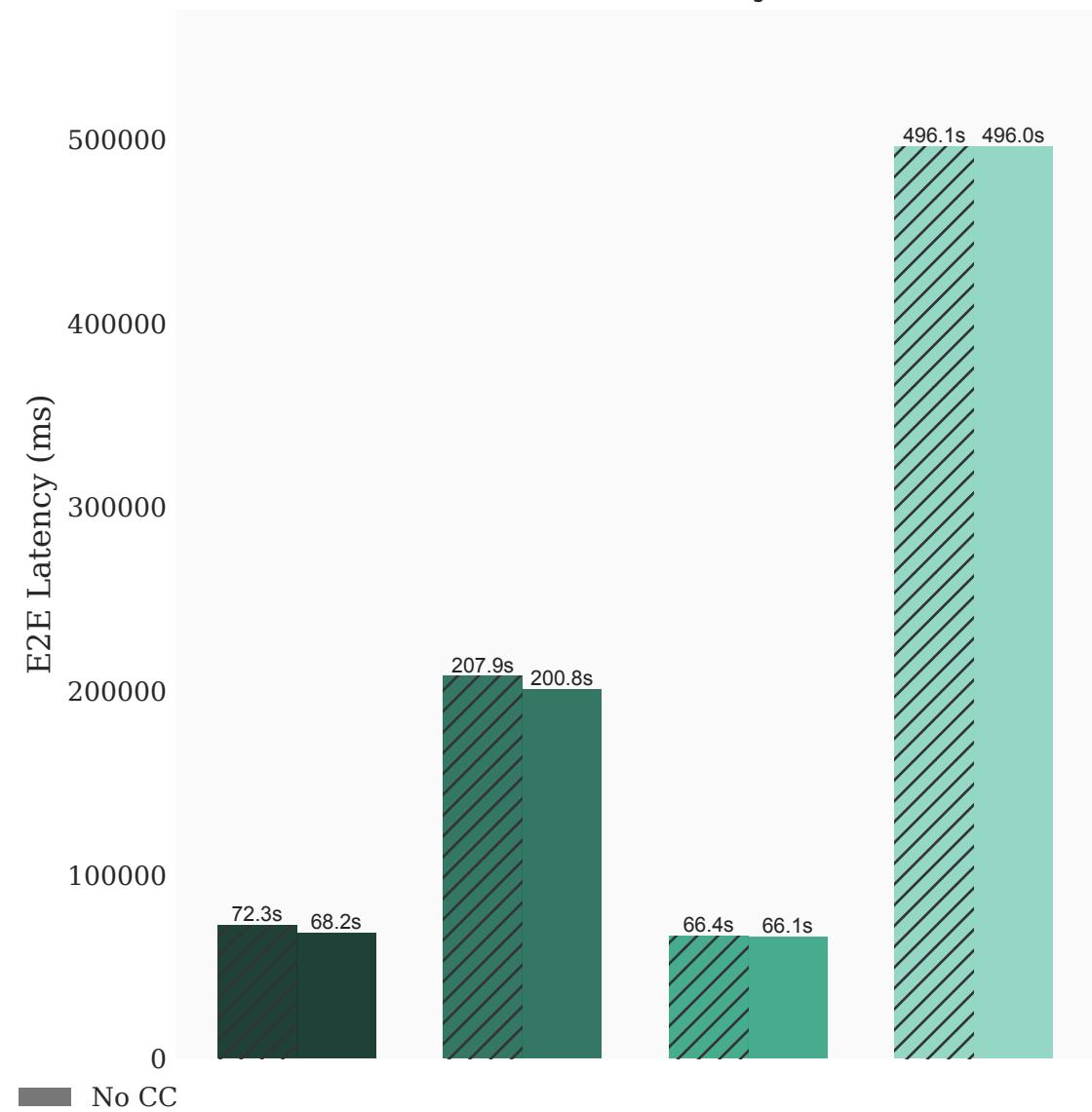


Random (4000 \Rightarrow 1000) (100 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

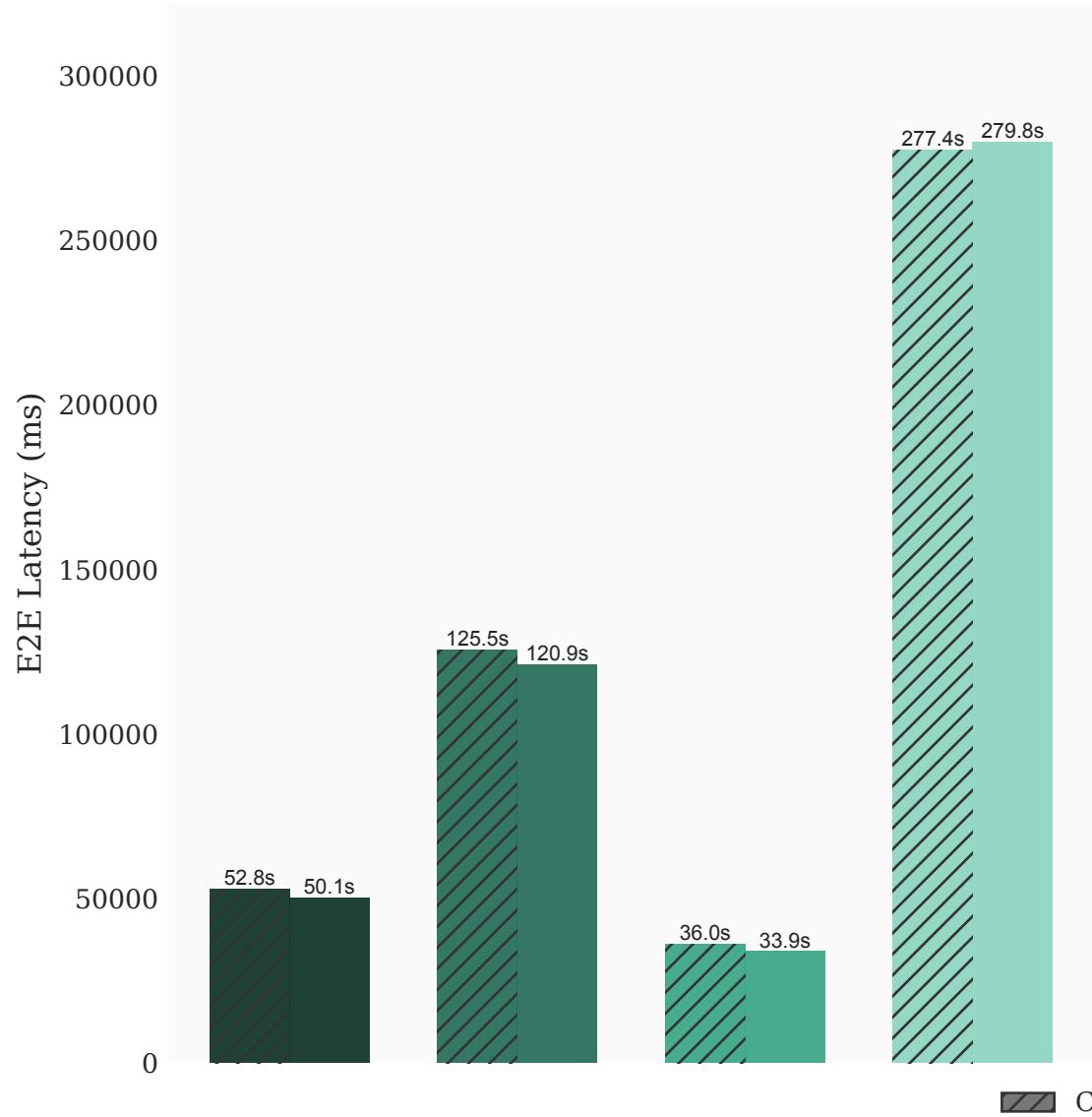
■ Mistral 3.1 24B

■ GPT OSS 120B

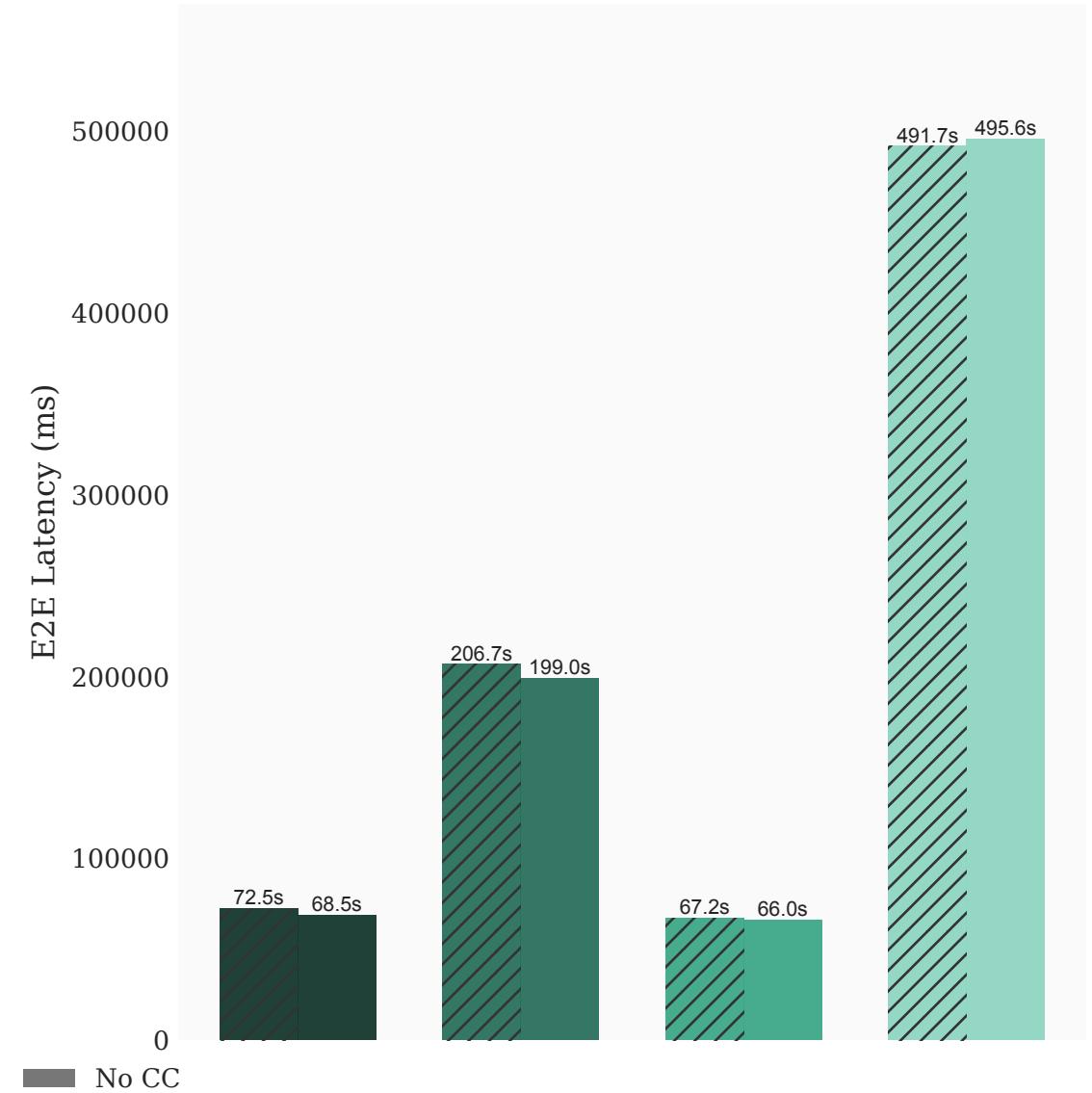
■ LLama 3.3 70B Int4

Random (4000 \Rightarrow 1000) (50 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



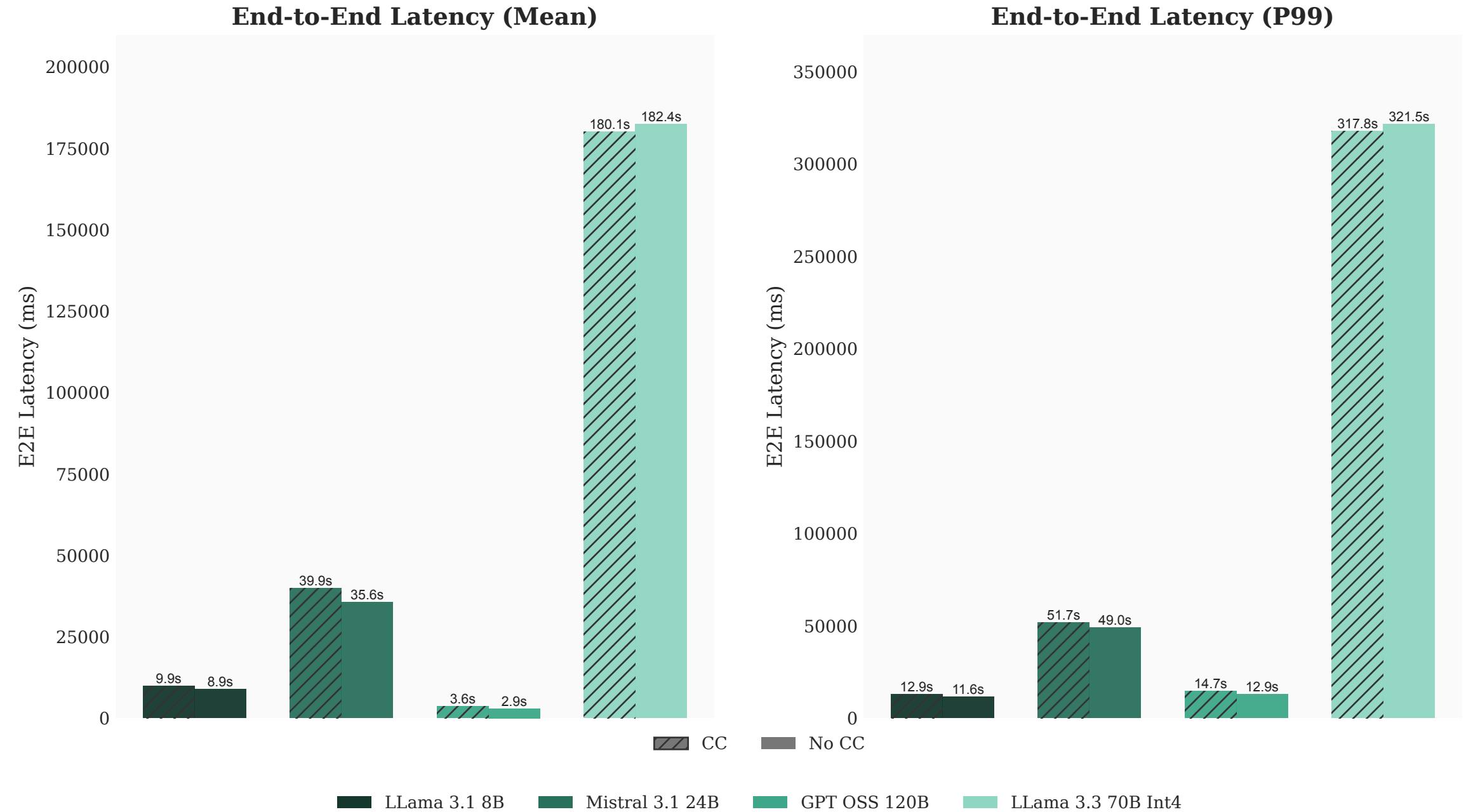
■ LLama 3.1 8B

■ Mistral 3.1 24B

■ GPT OSS 120B

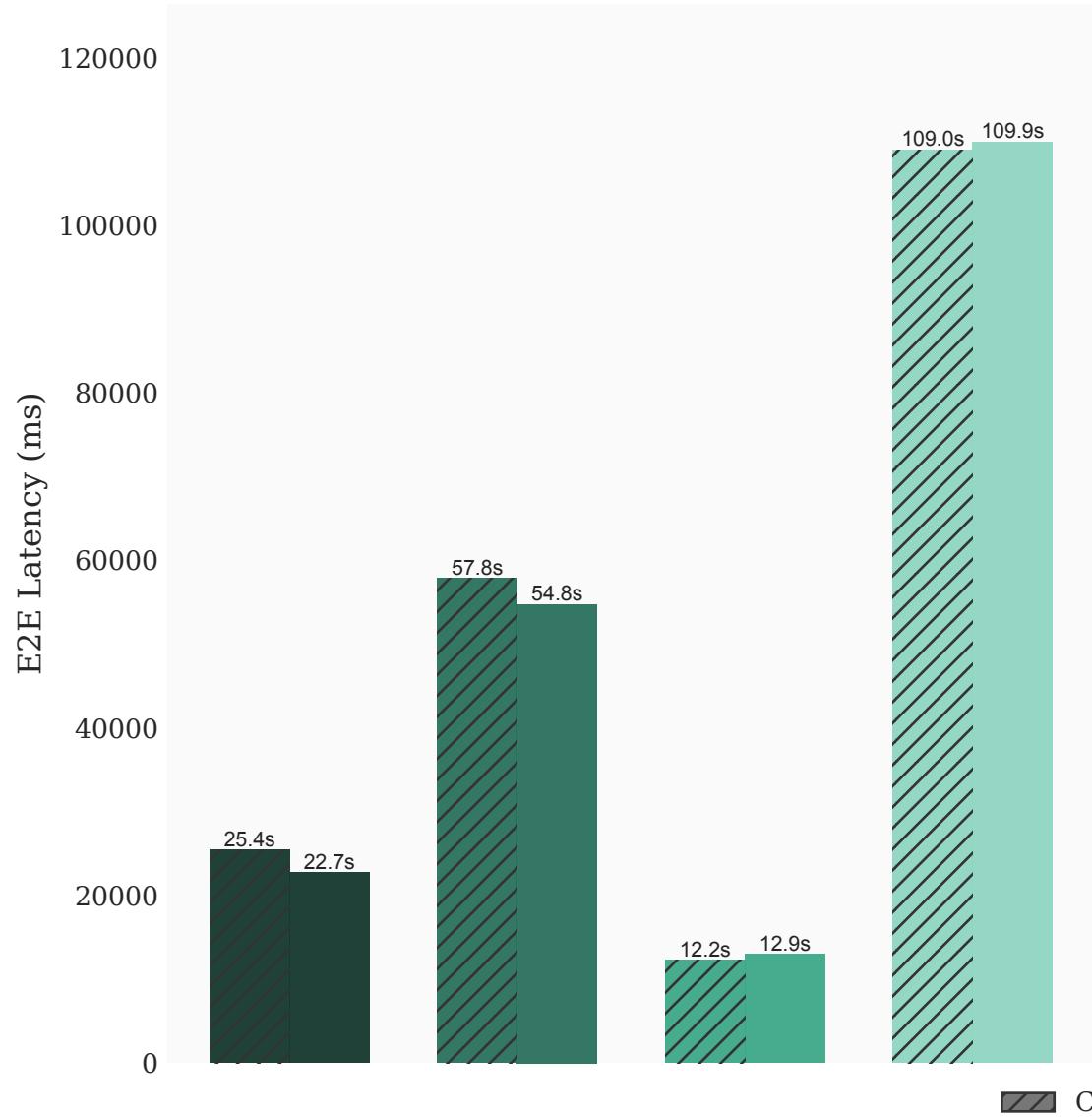
■ LLama 3.3 70B Int4

Random (4000 \Rightarrow 1000) (Single Request)

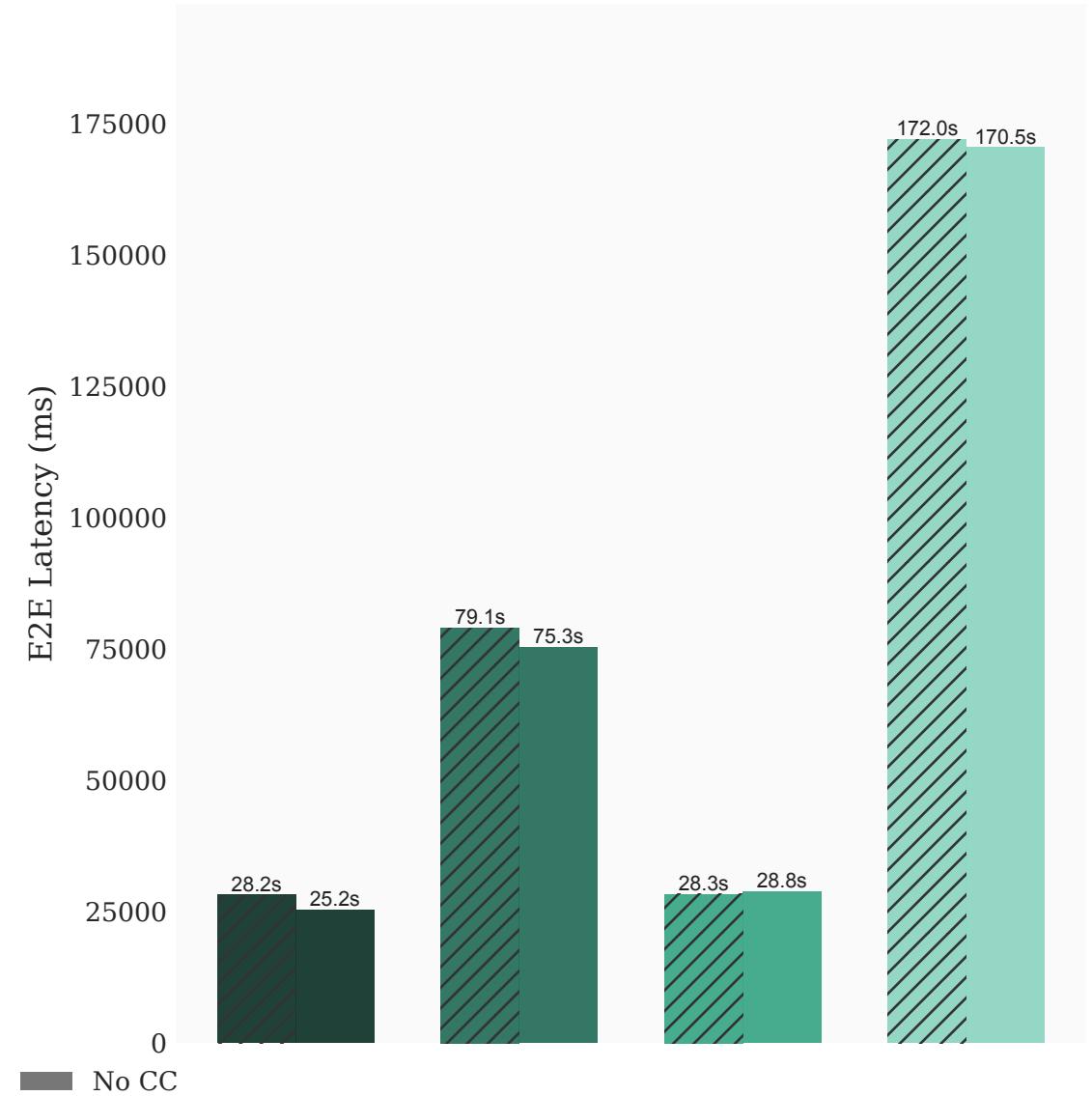


Random (1000 \Rightarrow 1000) (100 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

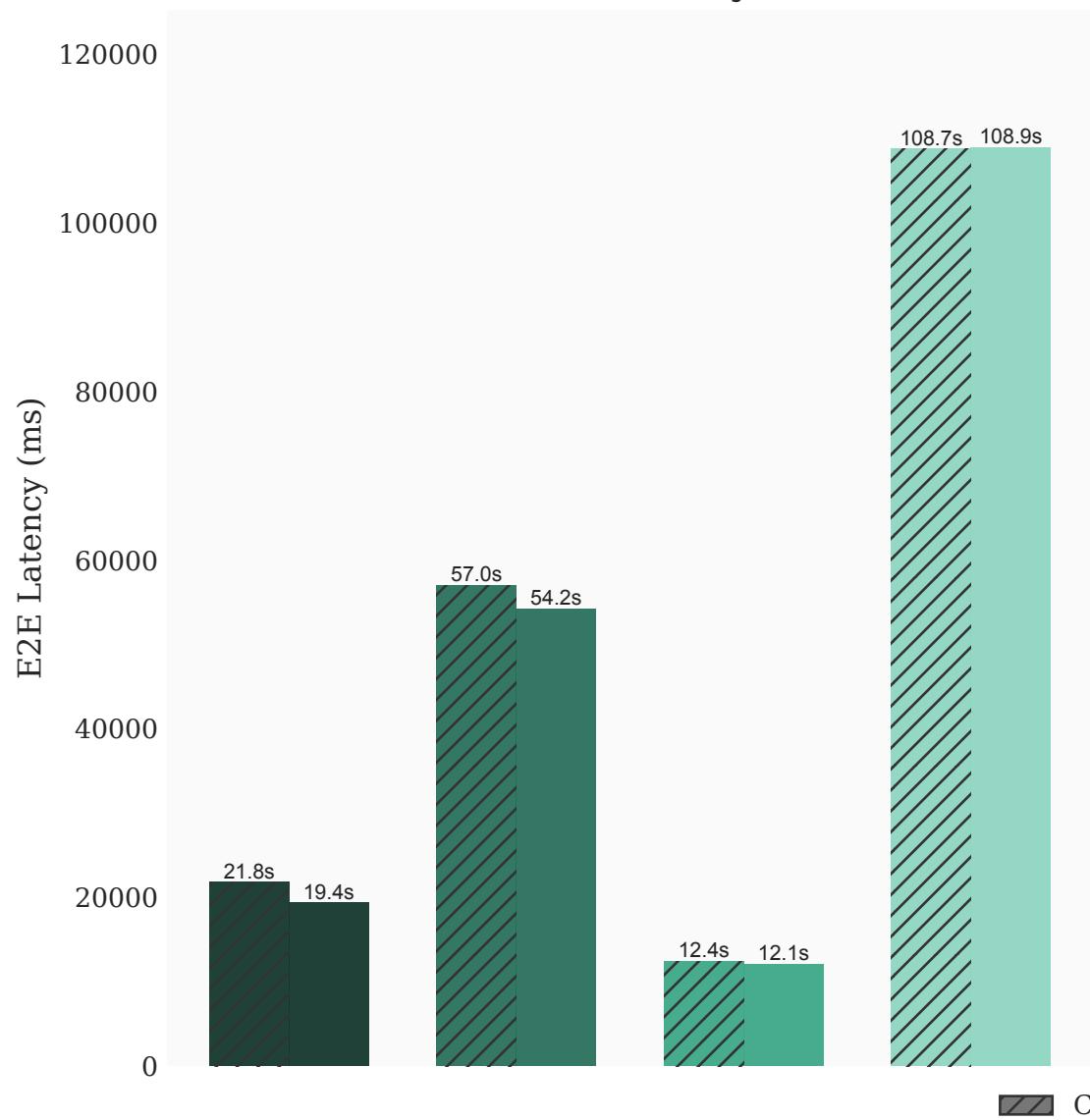
■ Mistral 3.1 24B

■ GPT OSS 120B

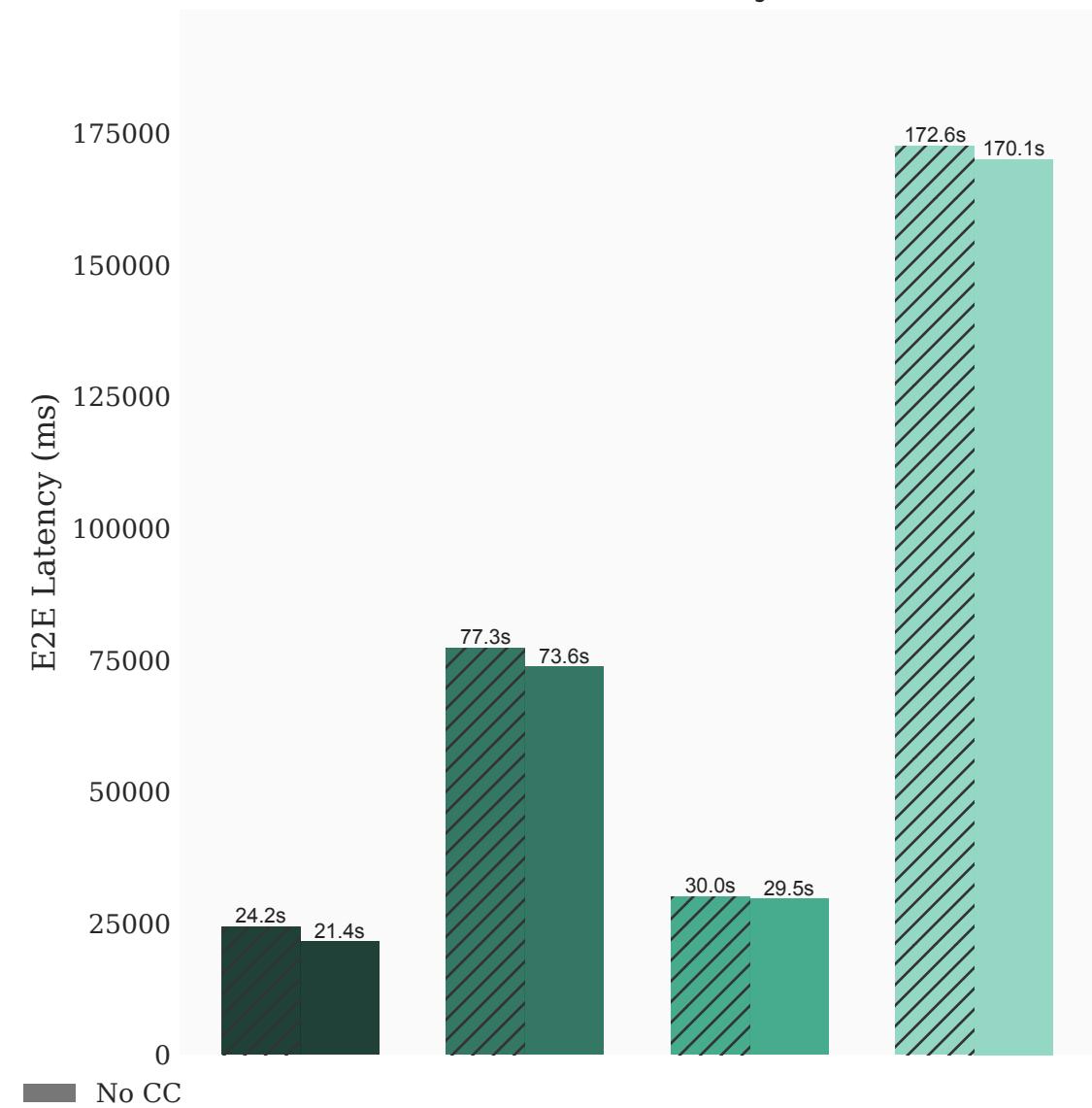
■ LLama 3.3 70B Int4

Random (1000 ⇒ 1000) (50 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

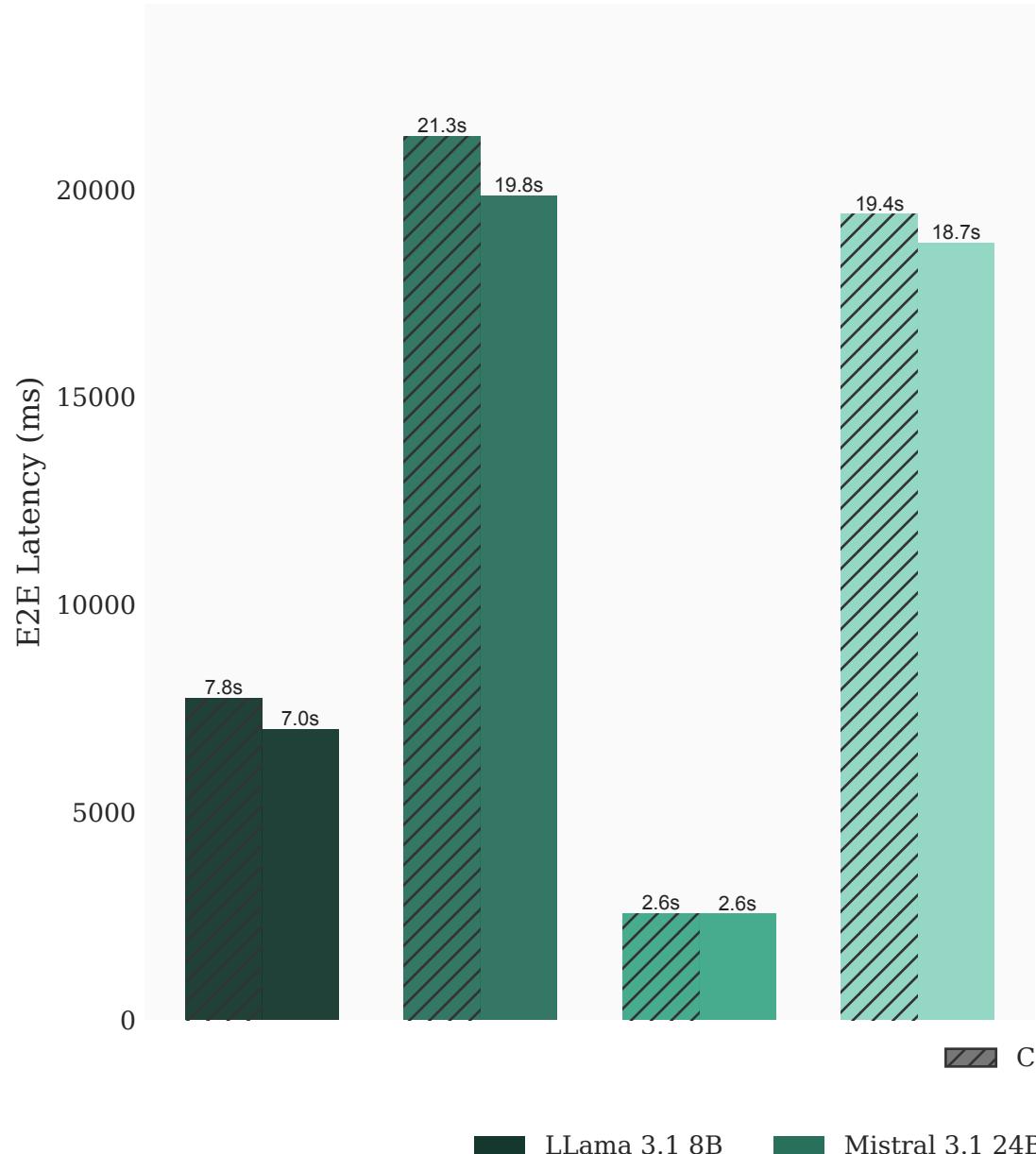
■ Mistral 3.1 24B

■ GPT OSS 120B

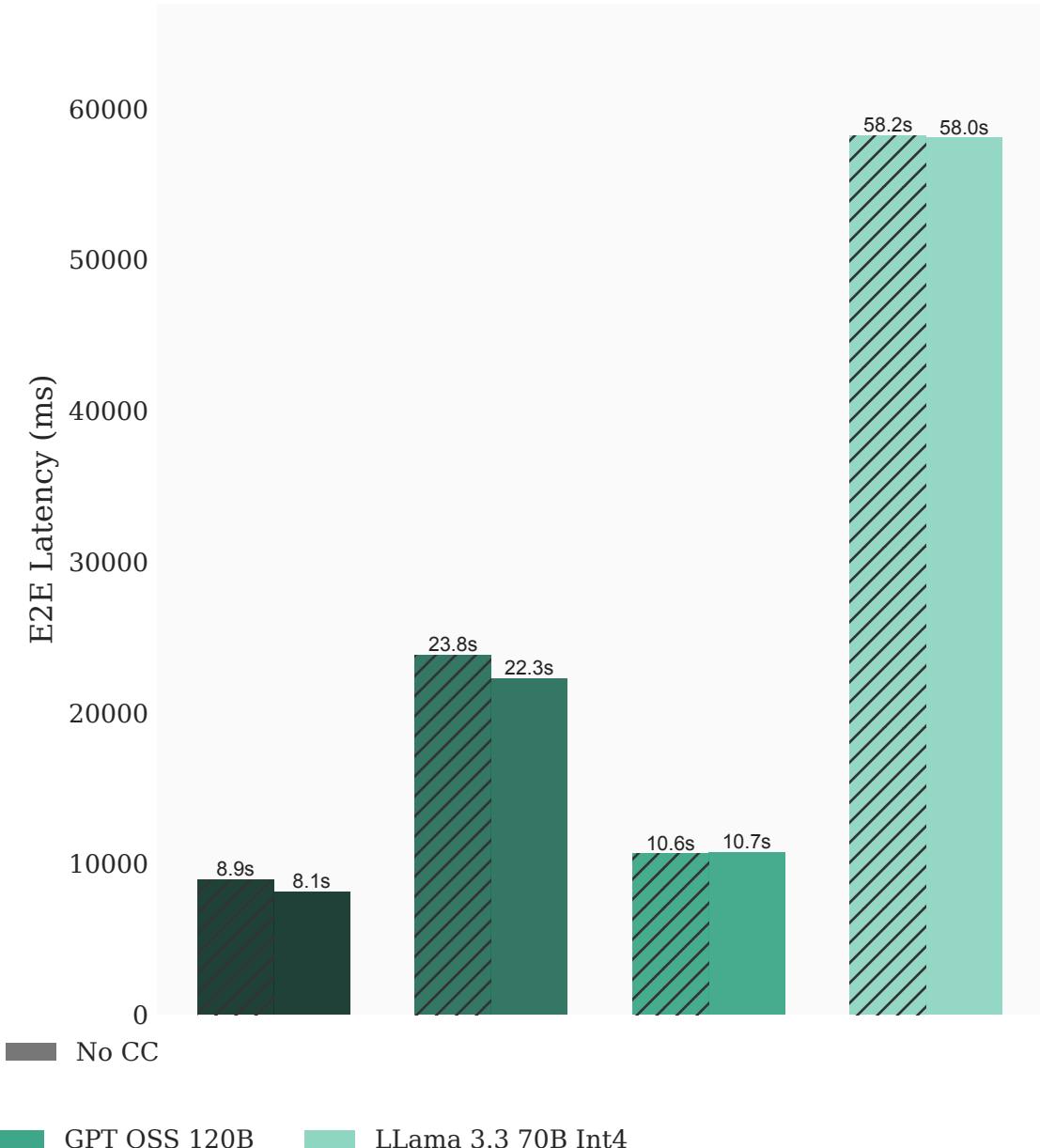
■ LLama 3.3 70B Int4

Random (1000 \Rightarrow 1000) (Single Request)

End-to-End Latency (Mean)

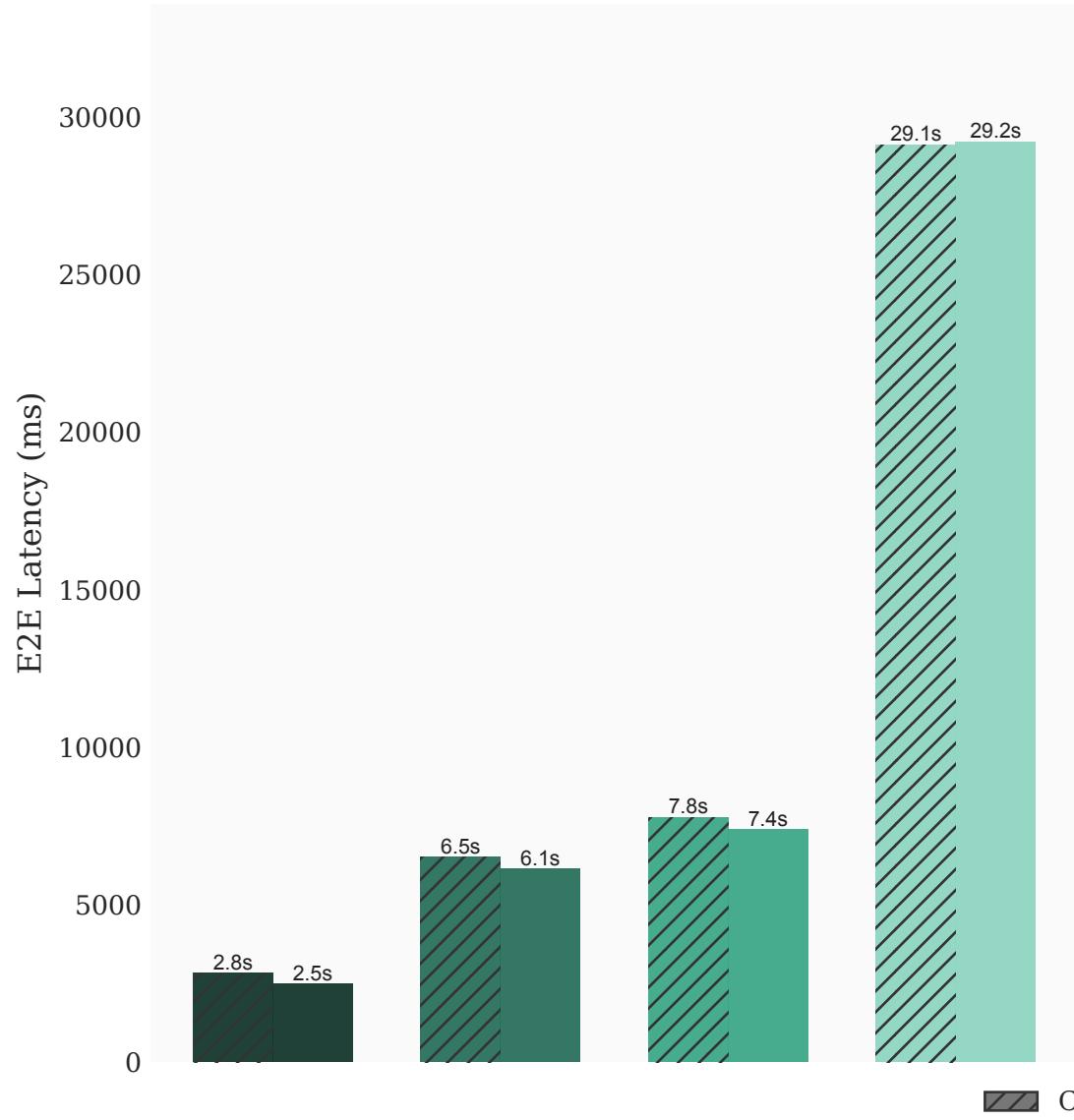


End-to-End Latency (P99)

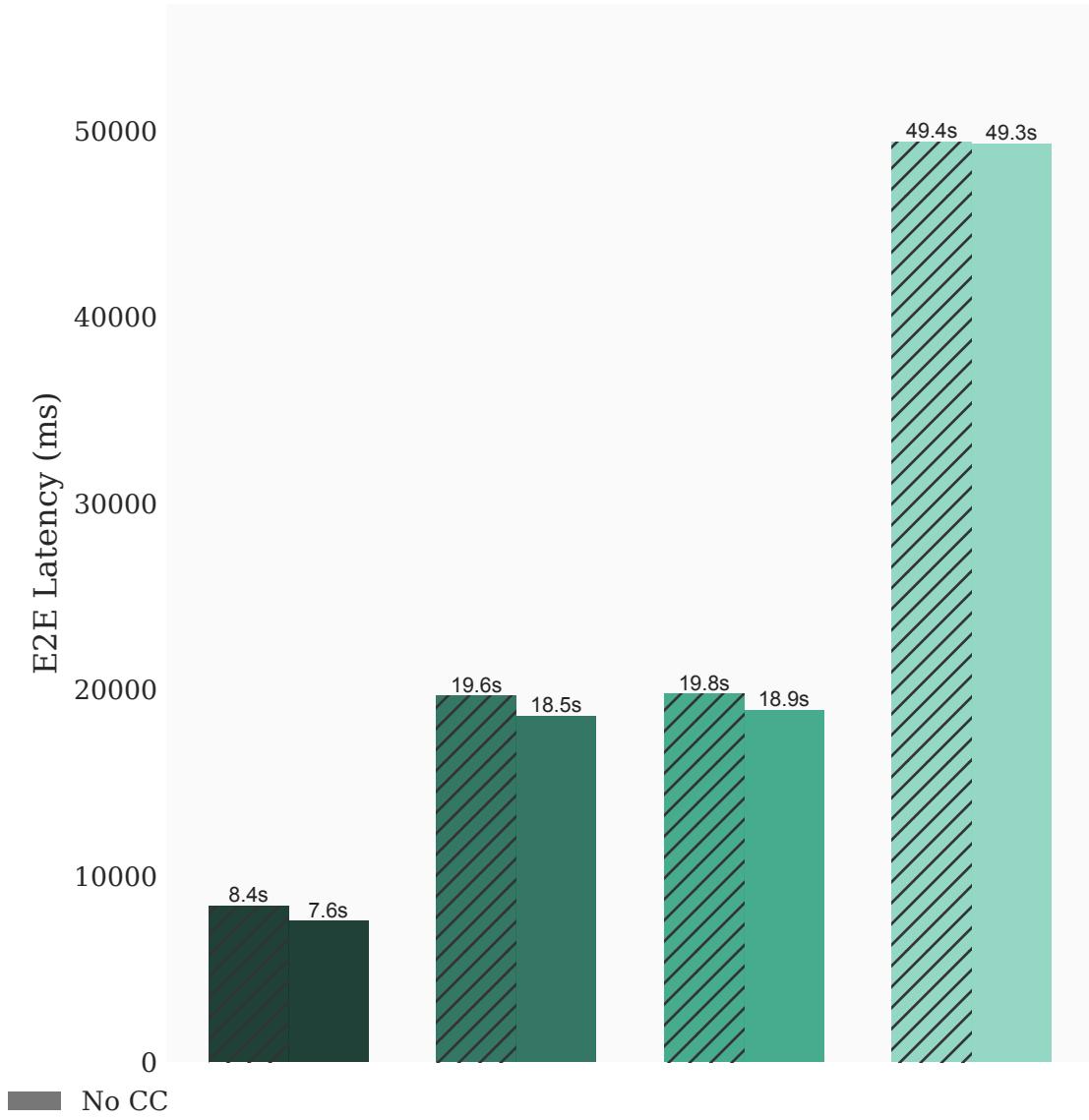


ShareGPT (100 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



LLama 3.1 8B

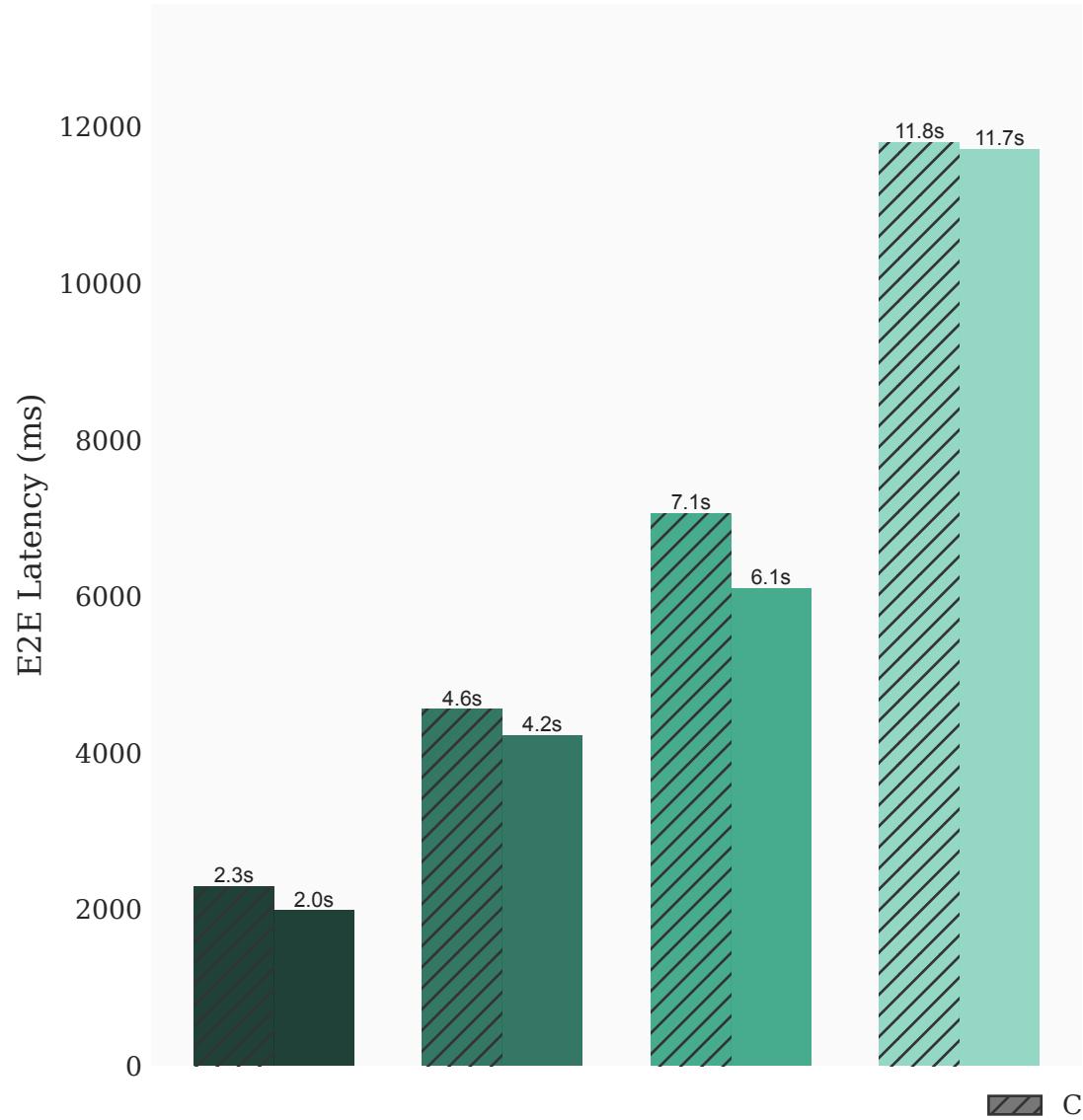
Mistral 3.1 24B

GPT OSS 120B

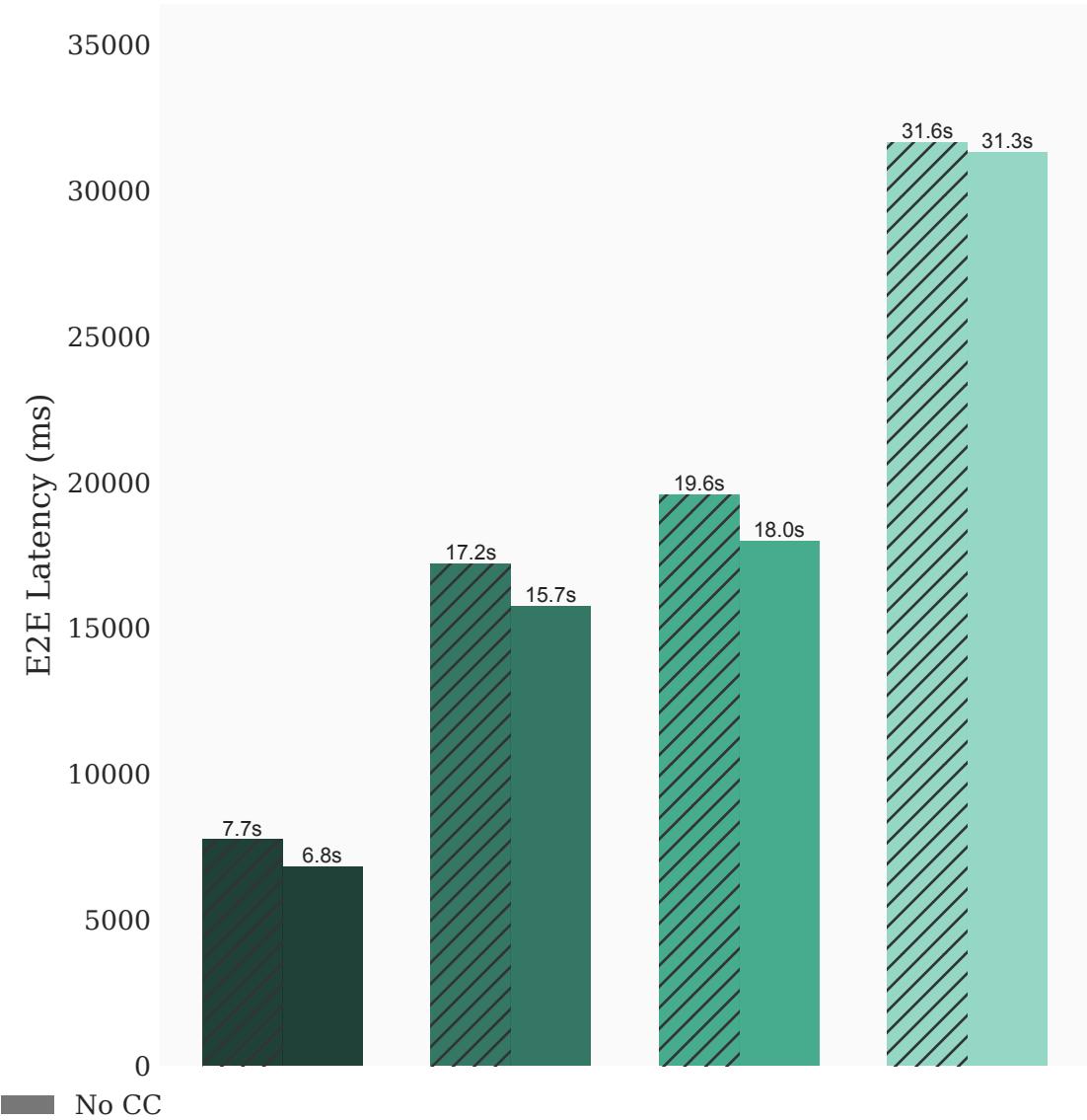
LLama 3.3 70B Int4

ShareGPT (50 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

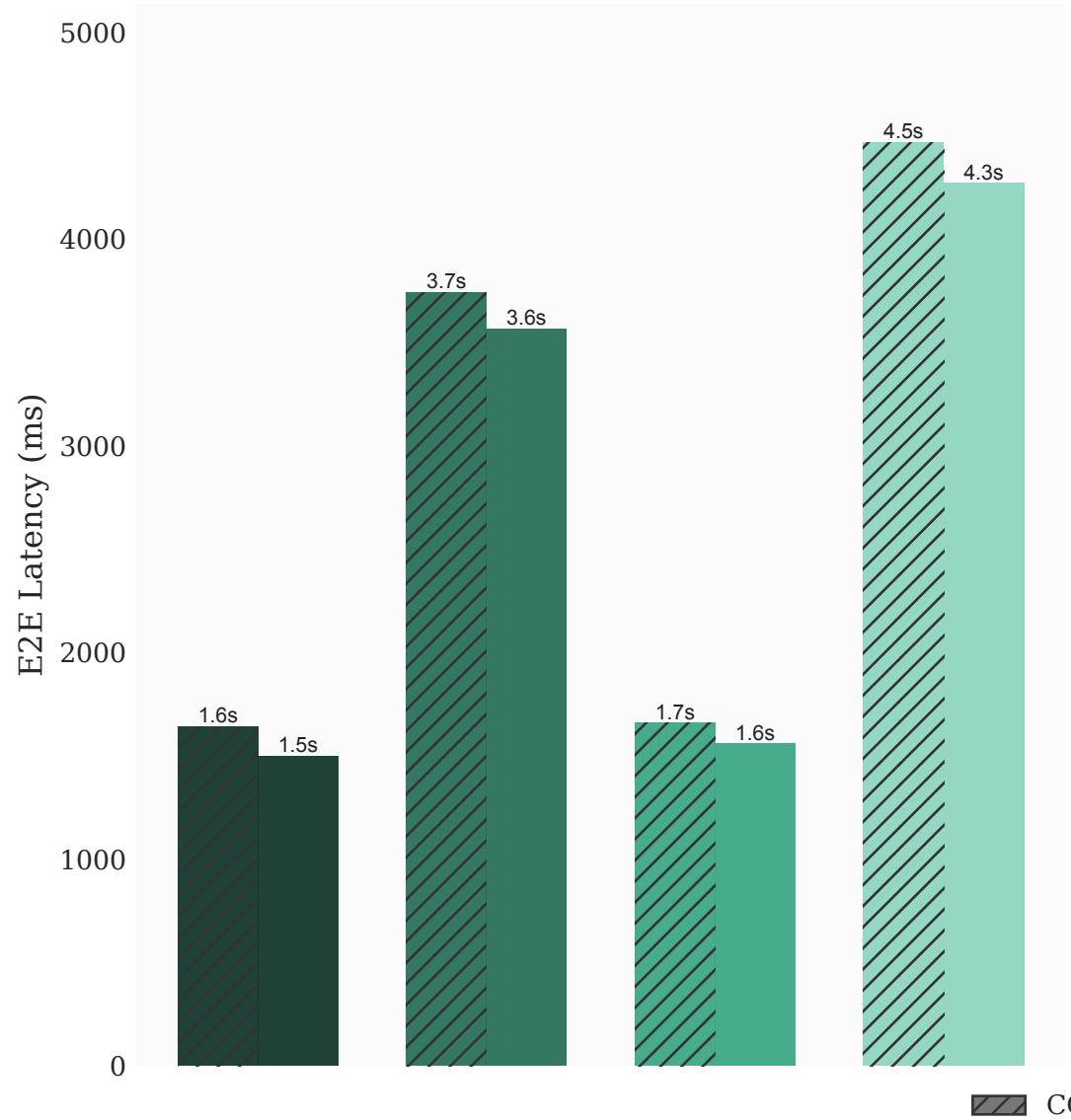
■ Mistral 3.1 24B

■ GPT OSS 120B

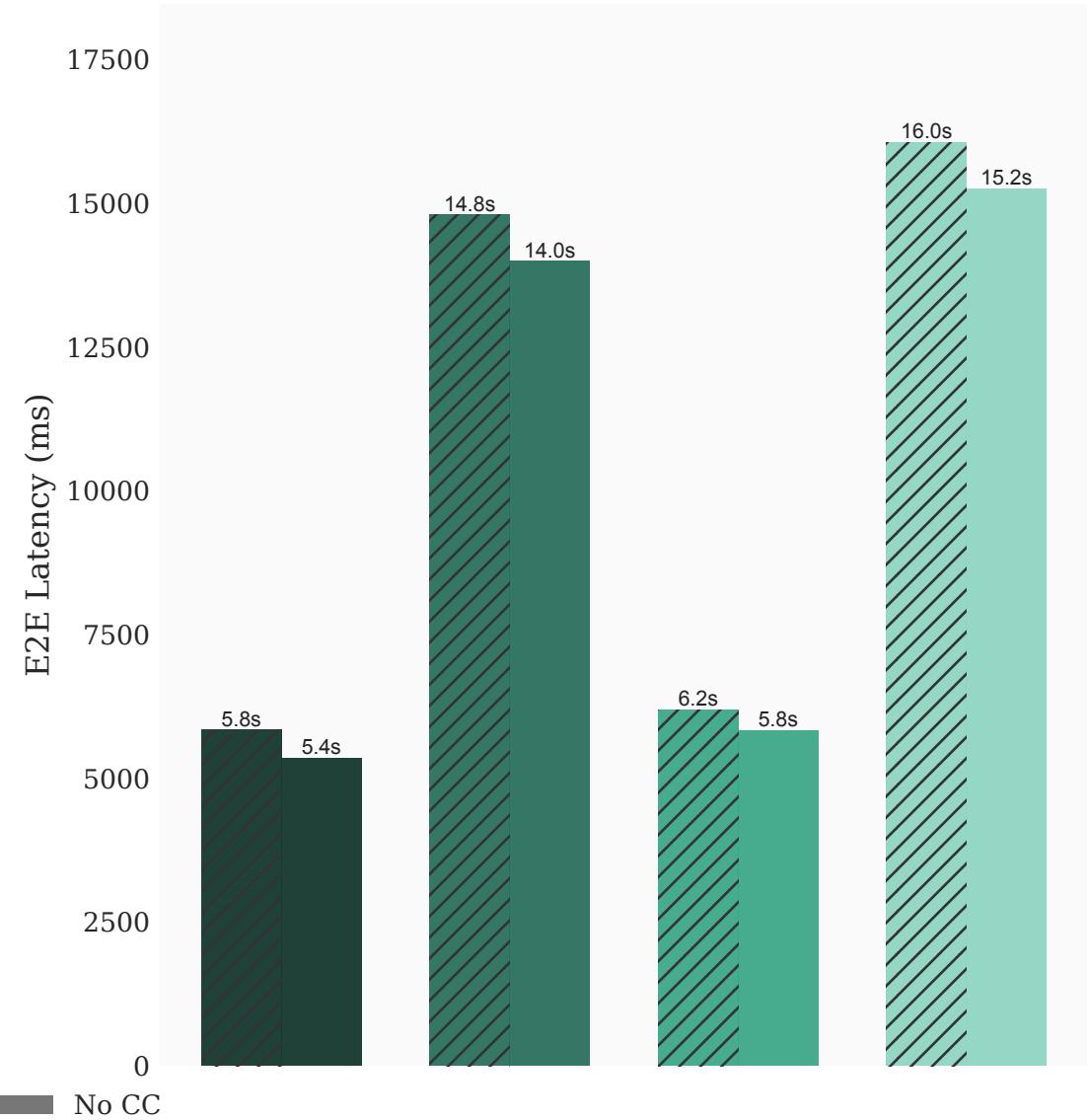
■ LLama 3.3 70B Int4

ShareGPT (Single Request)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

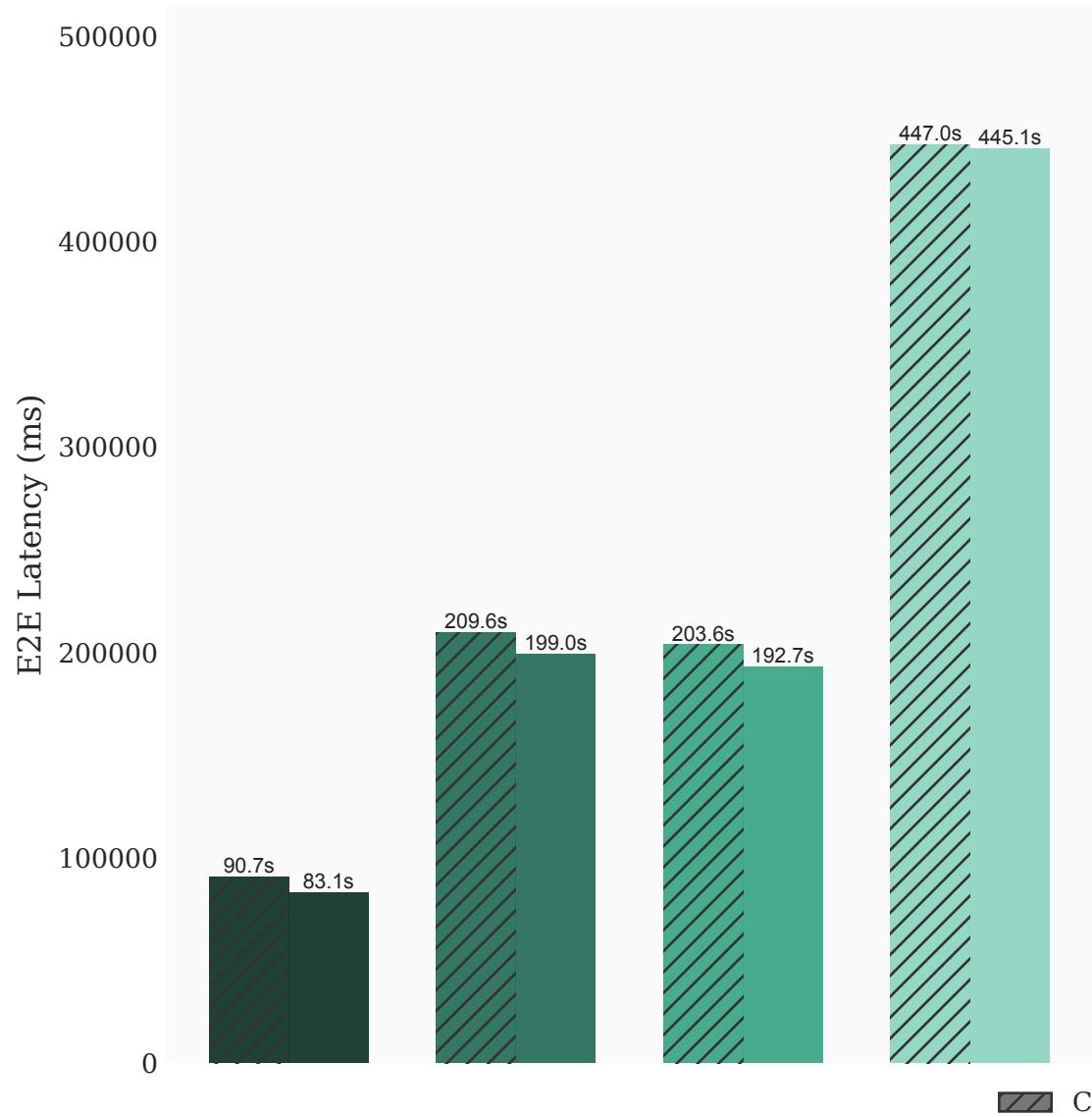
■ Mistral 3.1 24B

■ GPT OSS 120B

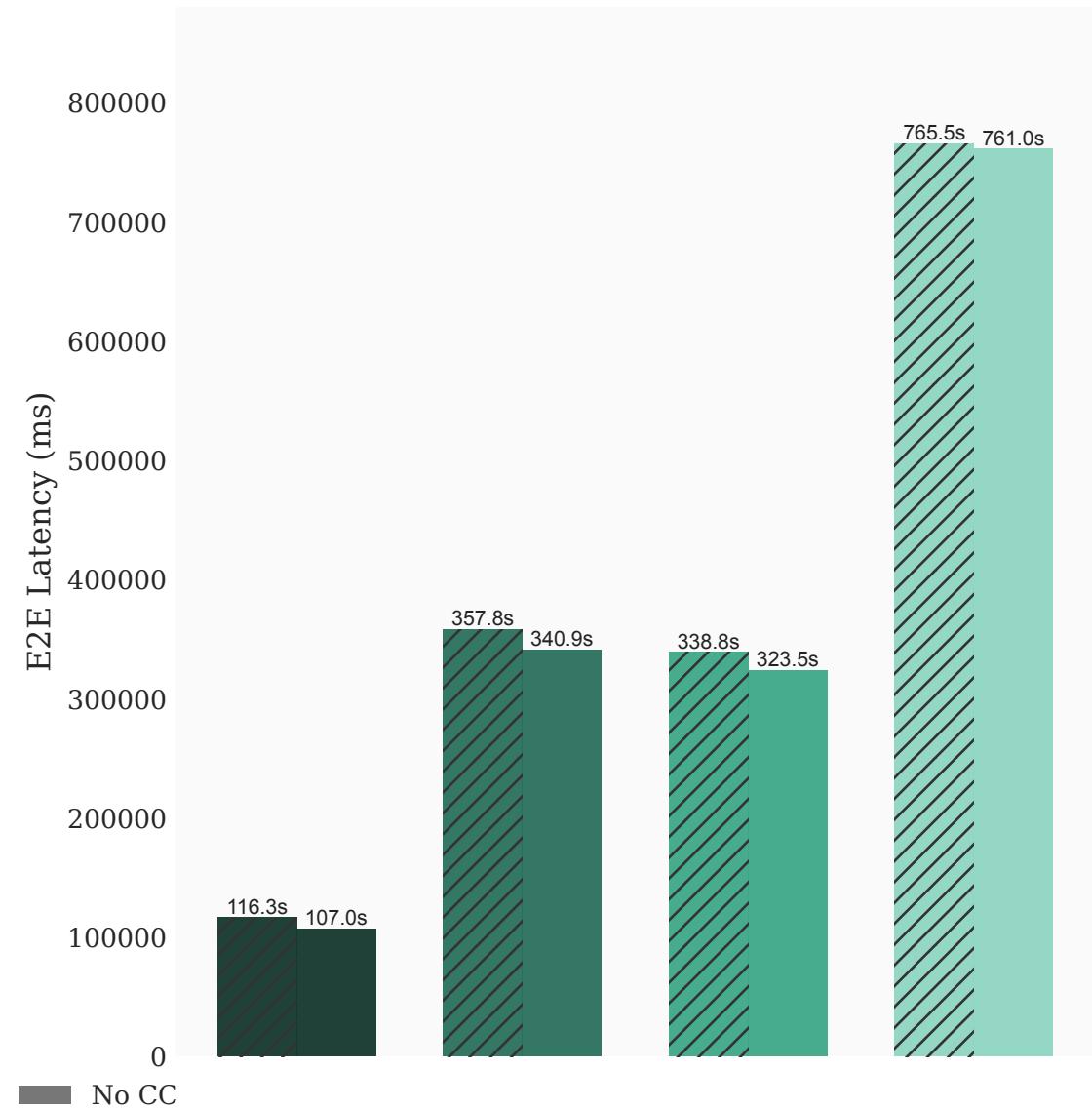
■ LLama 3.3 70B Int4

Edit 10K Characters (100 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

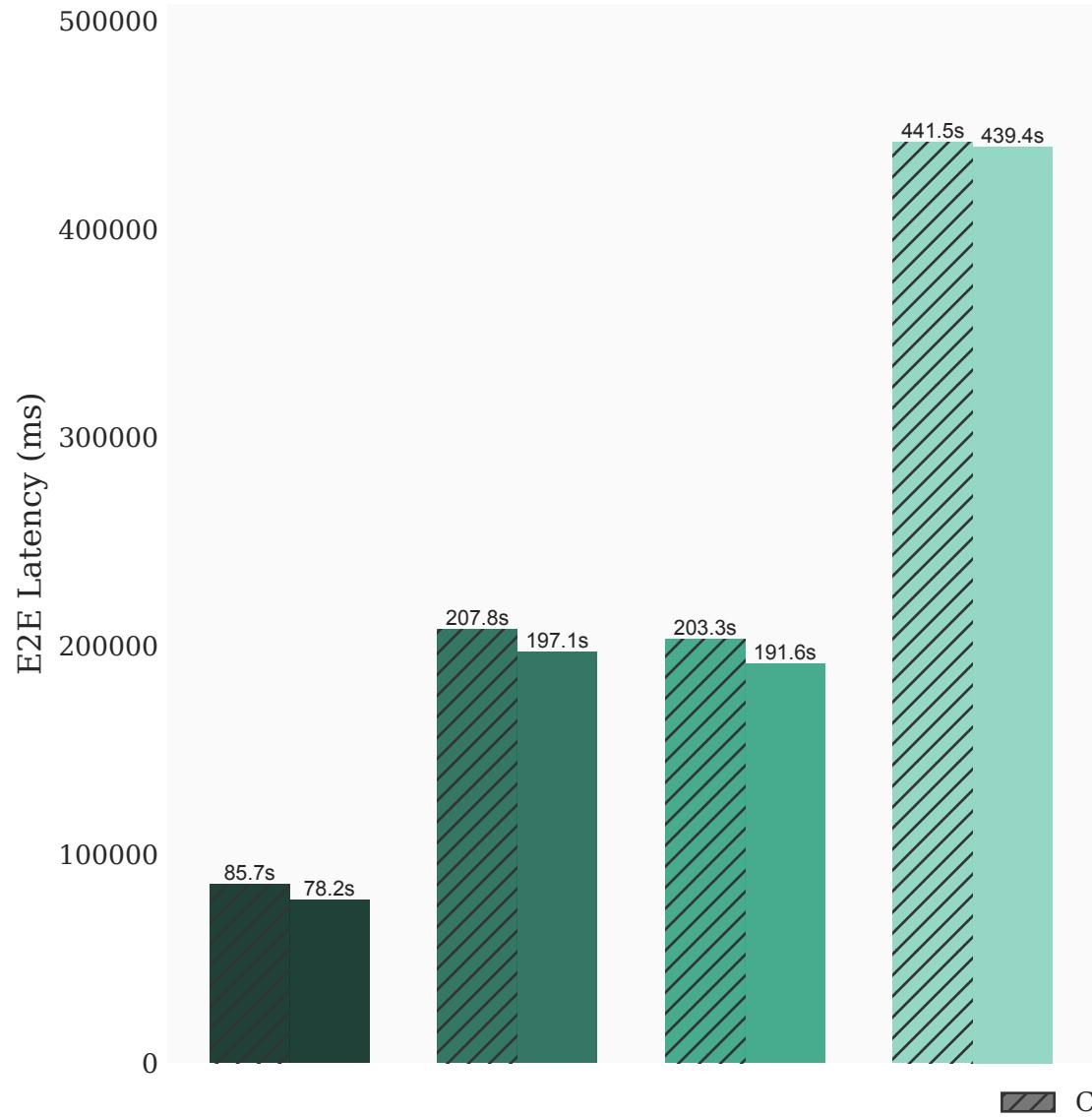
■ Mistral 3.1 24B

■ GPT OSS 120B

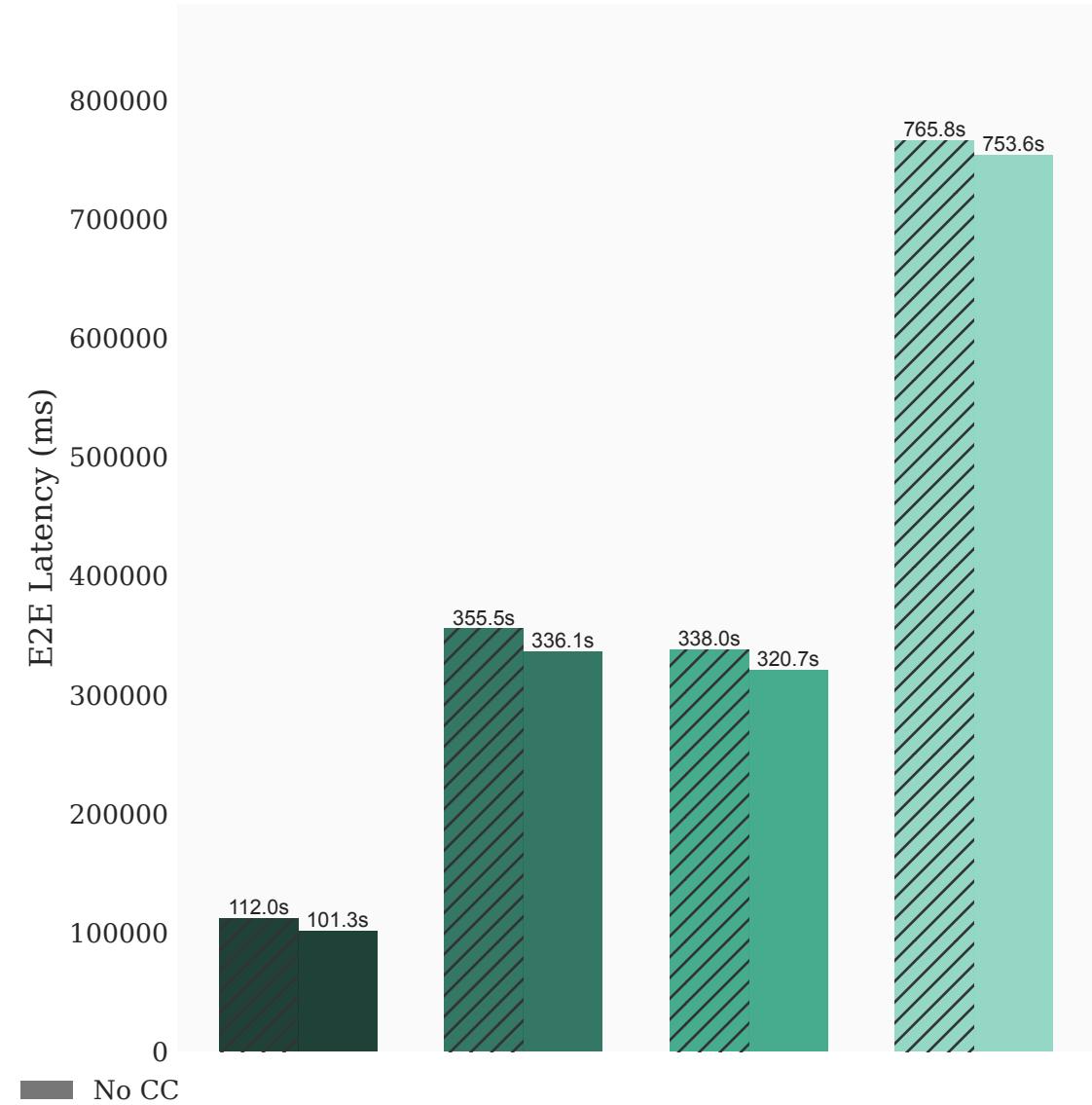
■ LLama 3.3 70B Int4

Edit 10K Characters (50 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

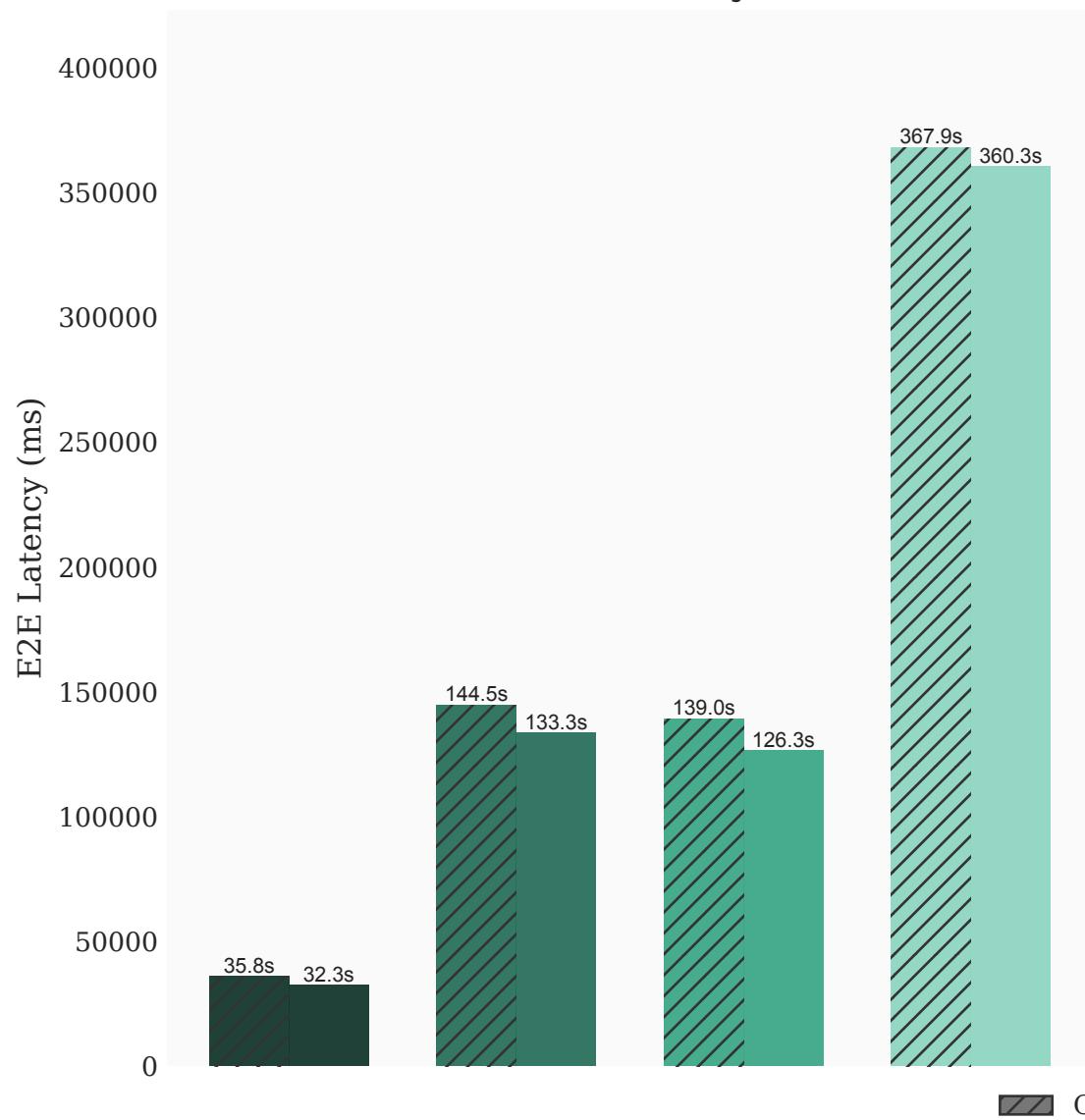
■ Mistral 3.1 24B

■ GPT OSS 120B

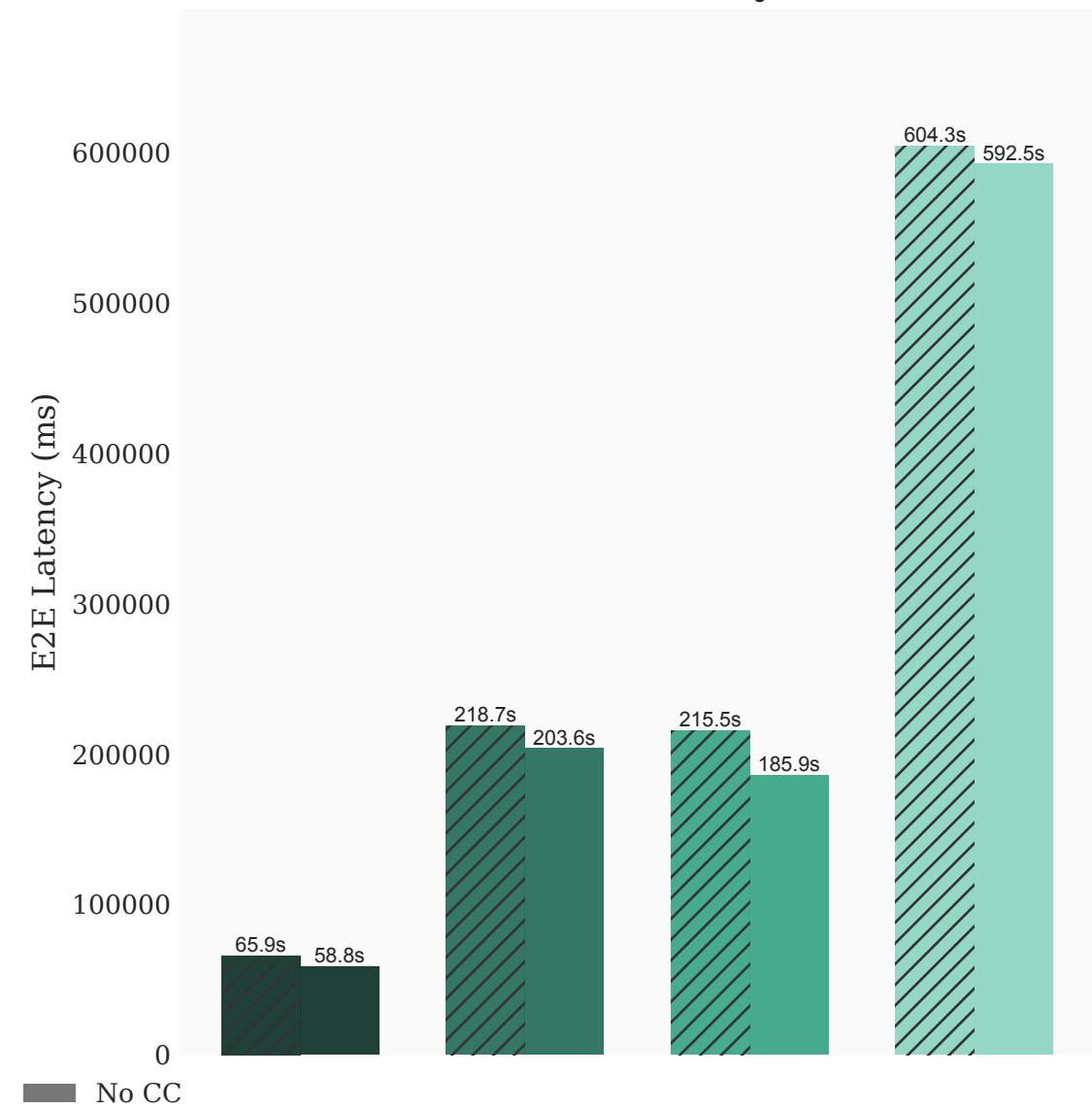
■ LLama 3.3 70B Int4

Edit 10K Characters (Single Request)

End-to-End Latency (Mean)



End-to-End Latency (P99)



LLama 3.1 8B

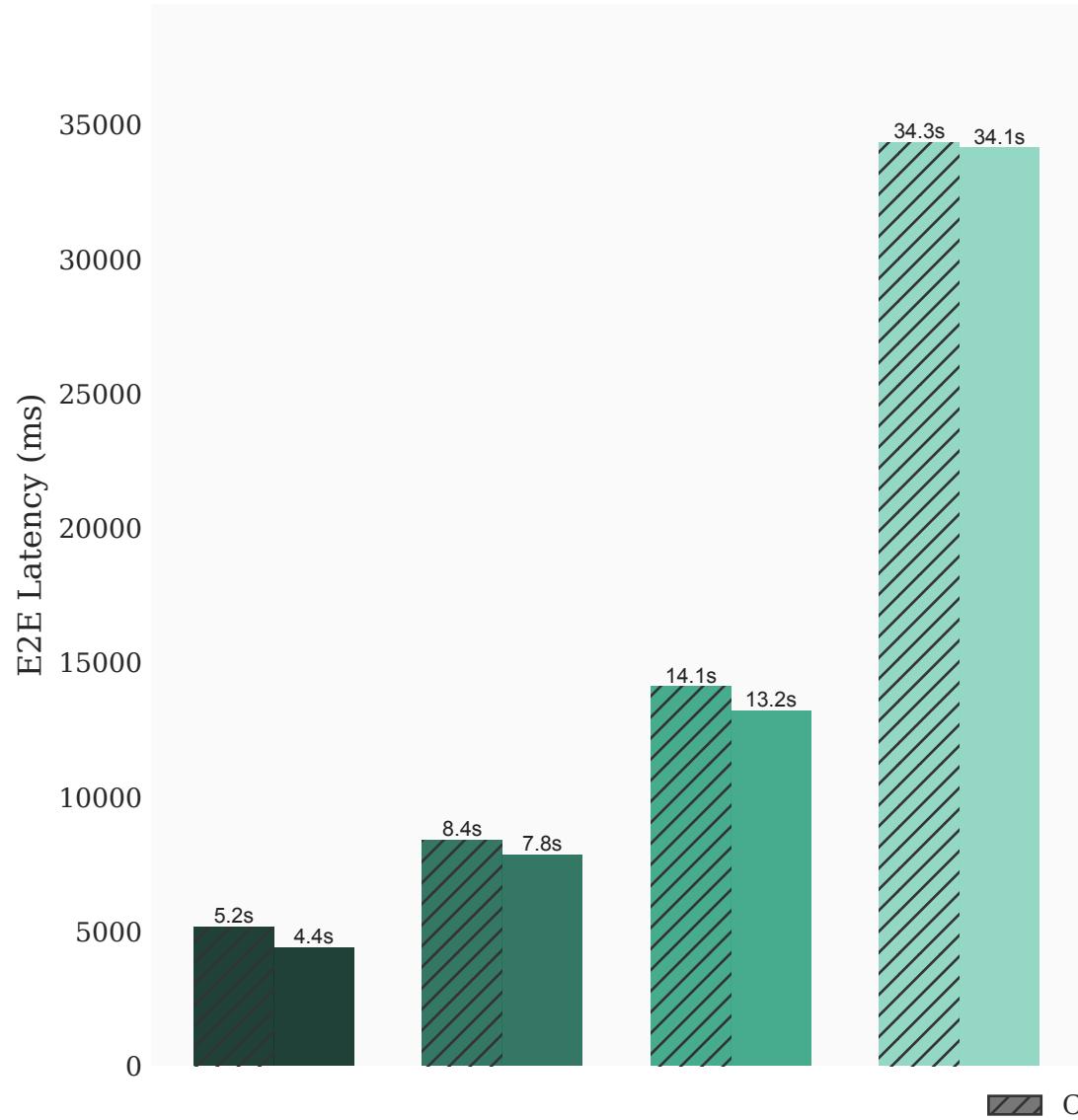
Mistral 3.1 24B

GPT OSS 120B

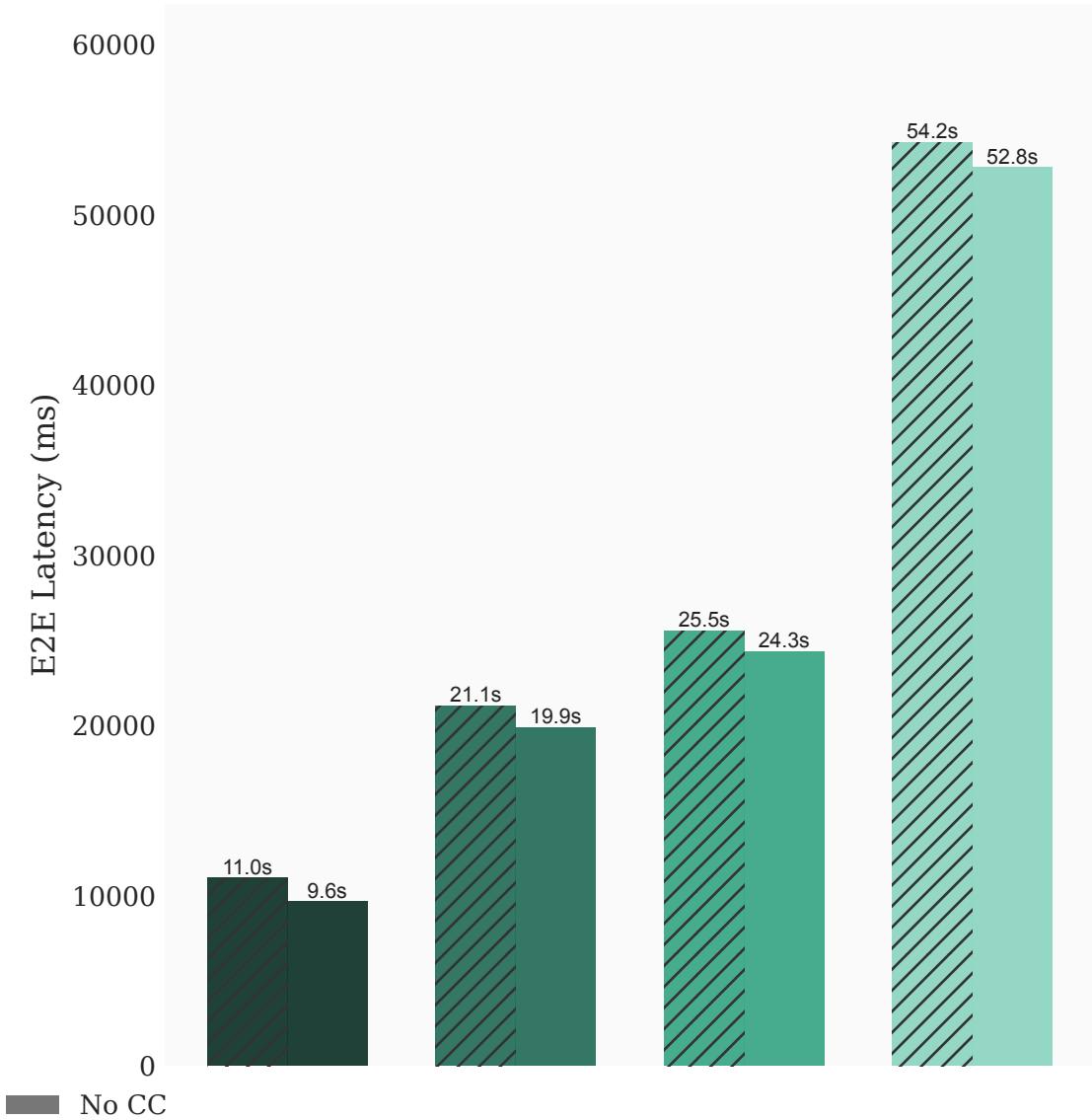
LLama 3.3 70B Int4

Numina Math (100 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

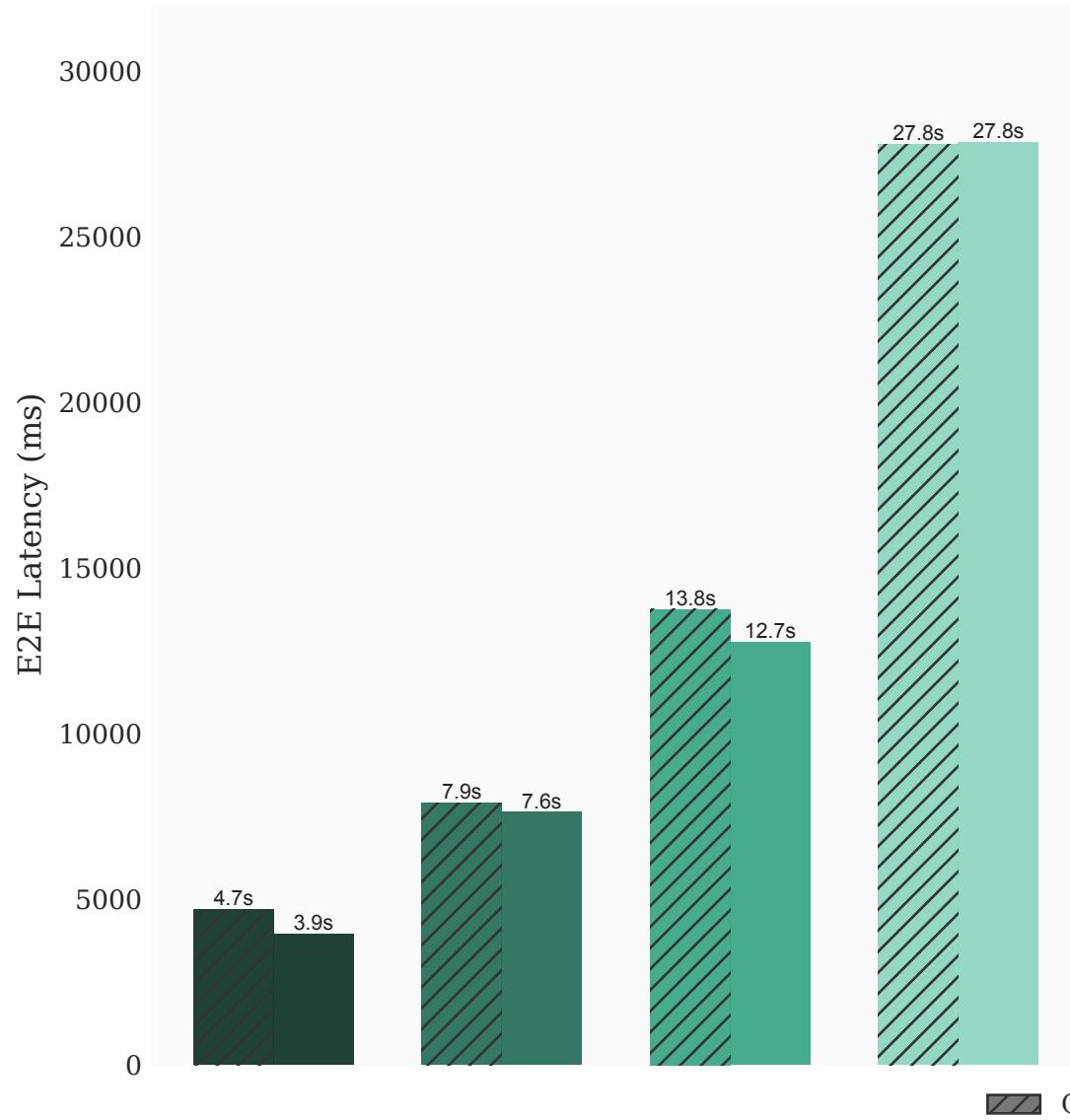
■ Mistral 3.1 24B

■ GPT OSS 120B

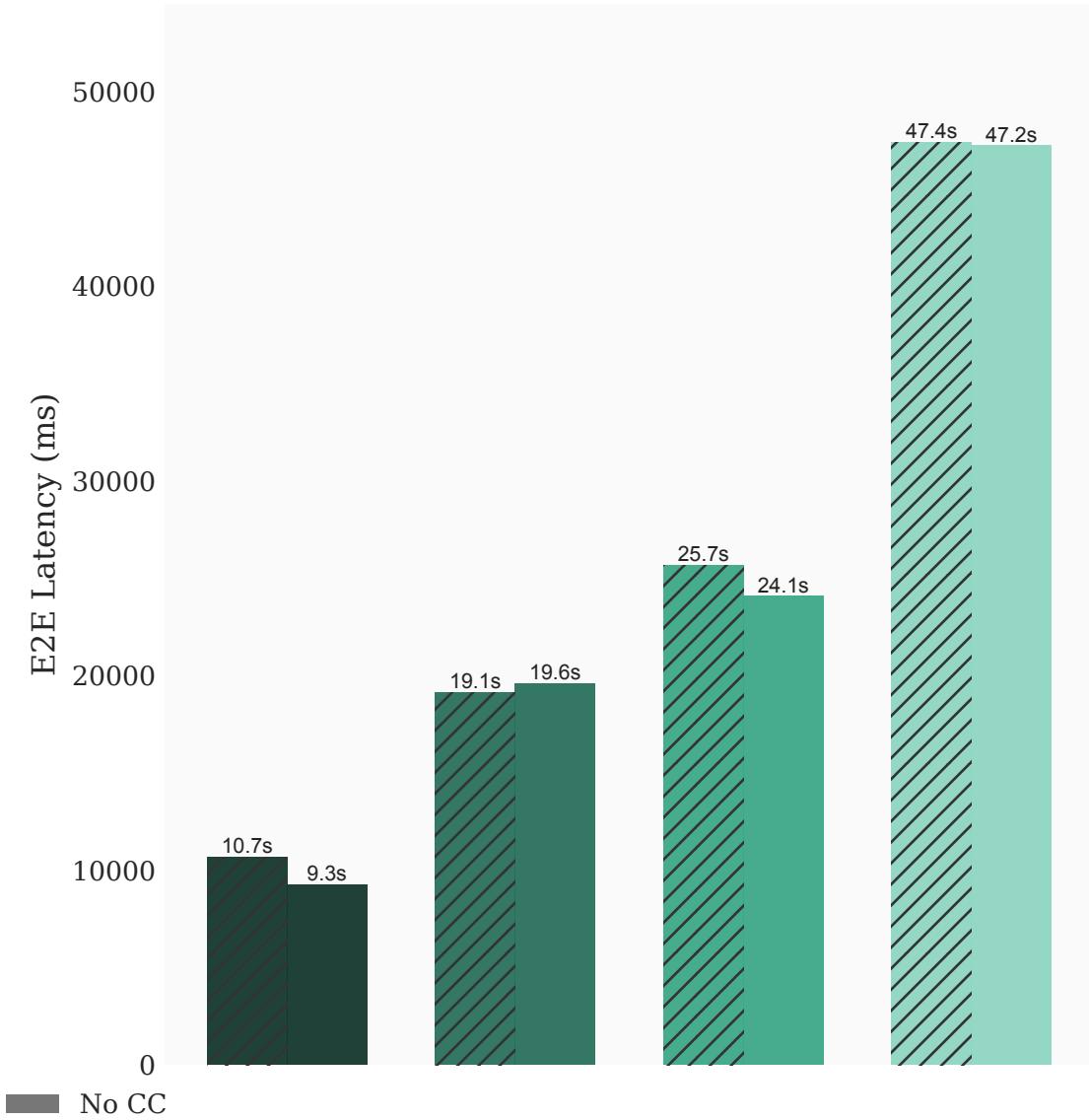
■ LLama 3.3 70B Int4

Numina Math (50 Request Rate)

End-to-End Latency (Mean)



End-to-End Latency (P99)



■ LLama 3.1 8B

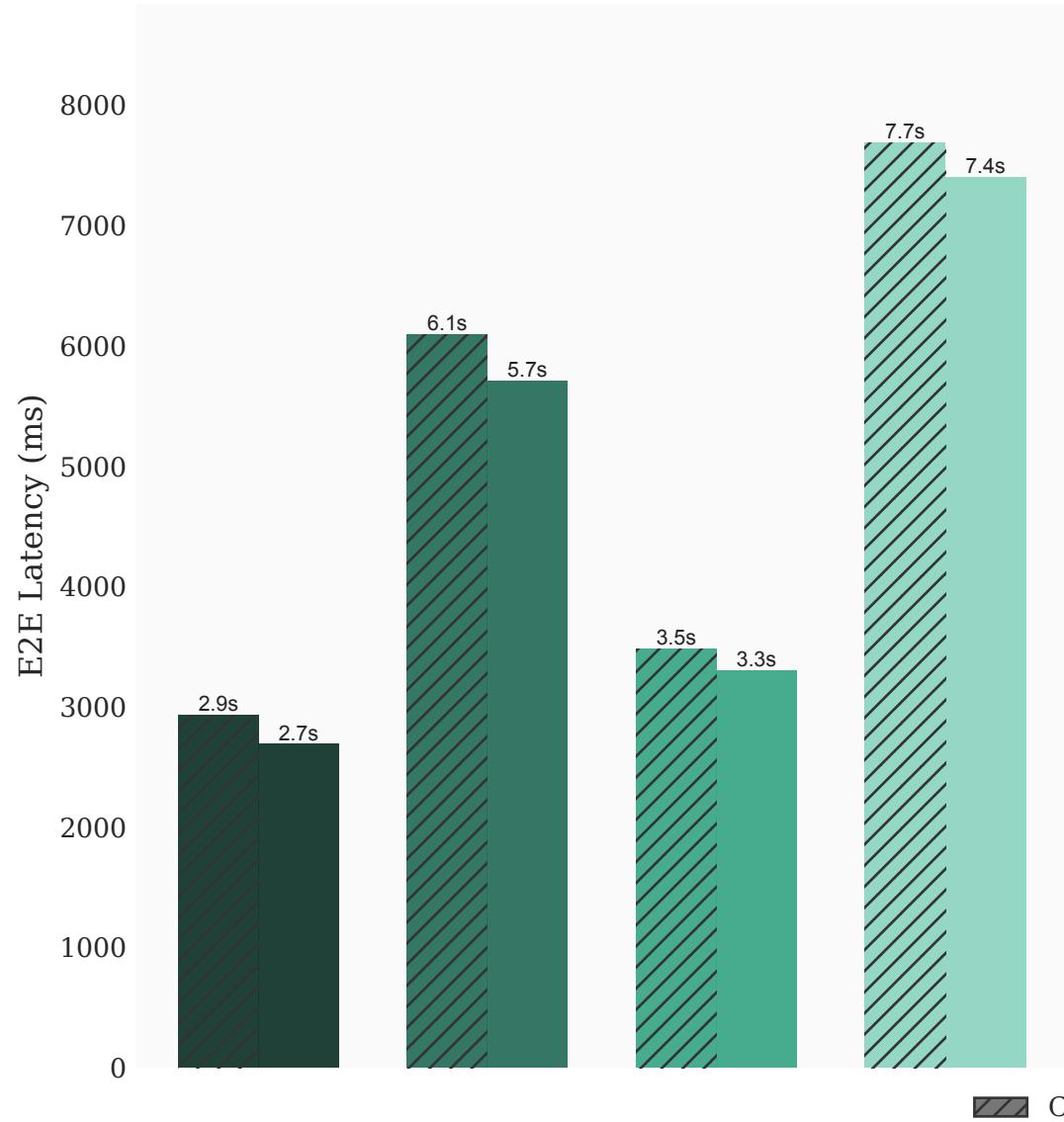
■ Mistral 3.1 24B

■ GPT OSS 120B

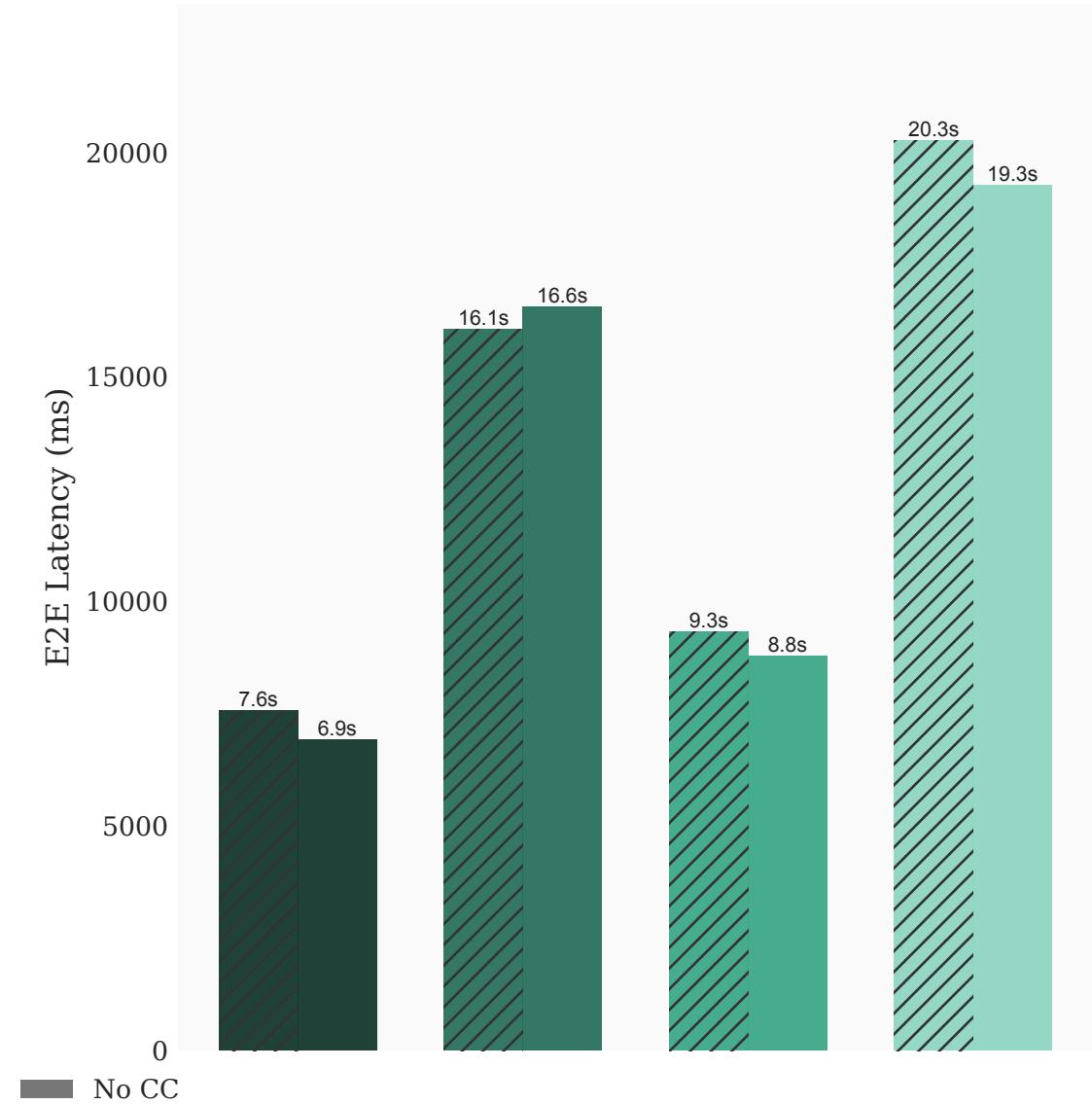
■ LLama 3.3 70B Int4

Numina Math (Single Request)

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