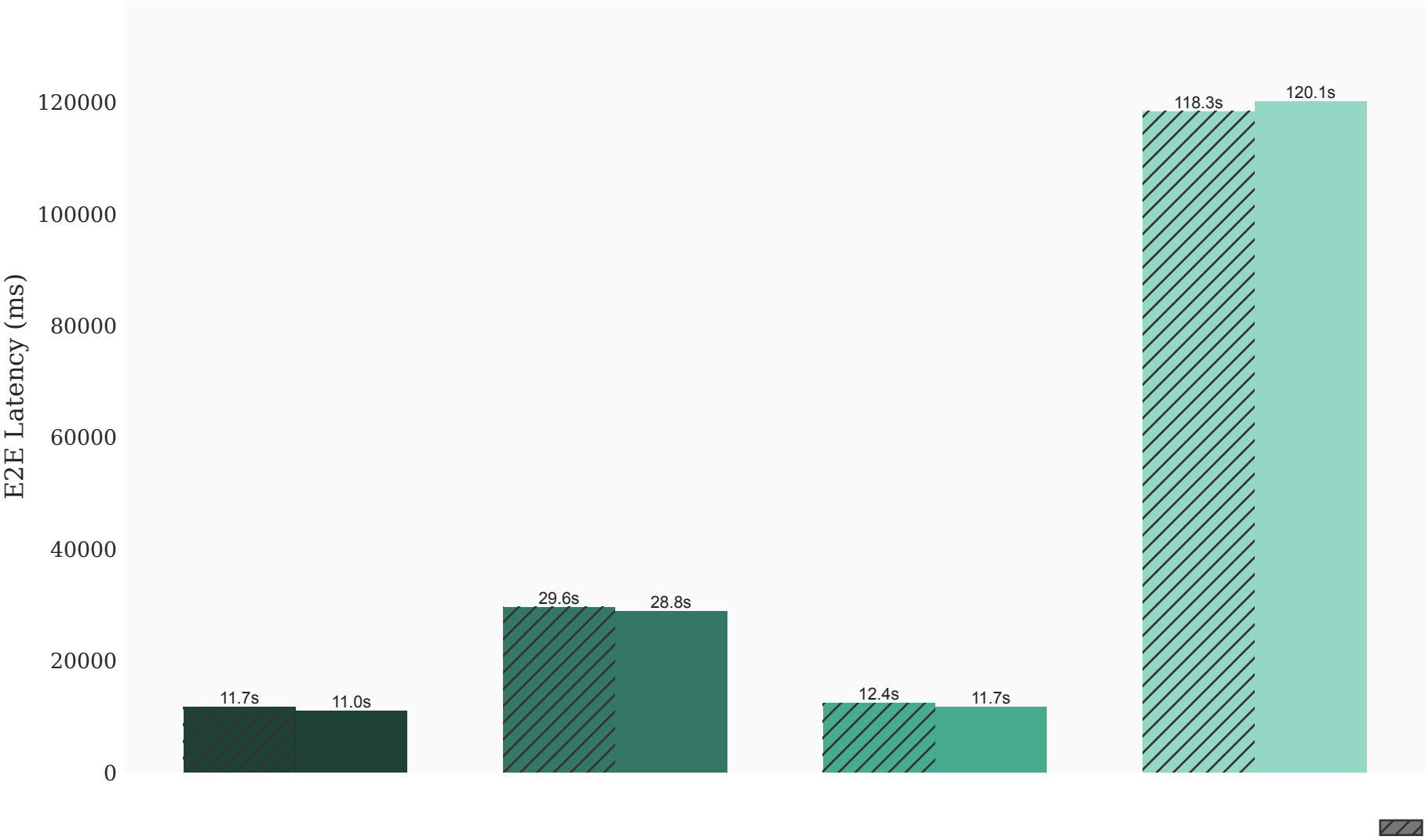
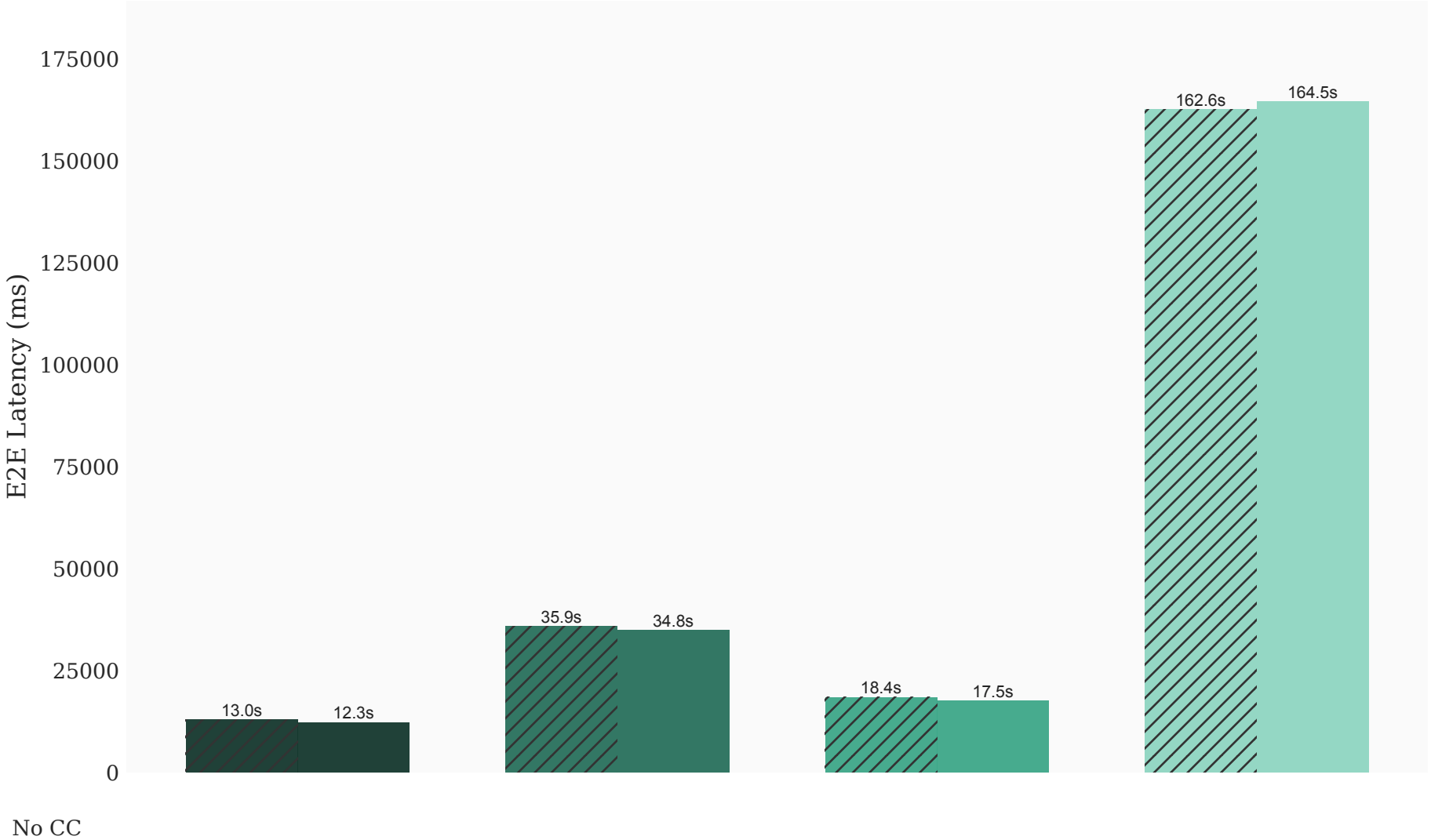


Random (1500 ⇒ 250) (100 Concurrent Requests)

End-to-End Latency (Mean)



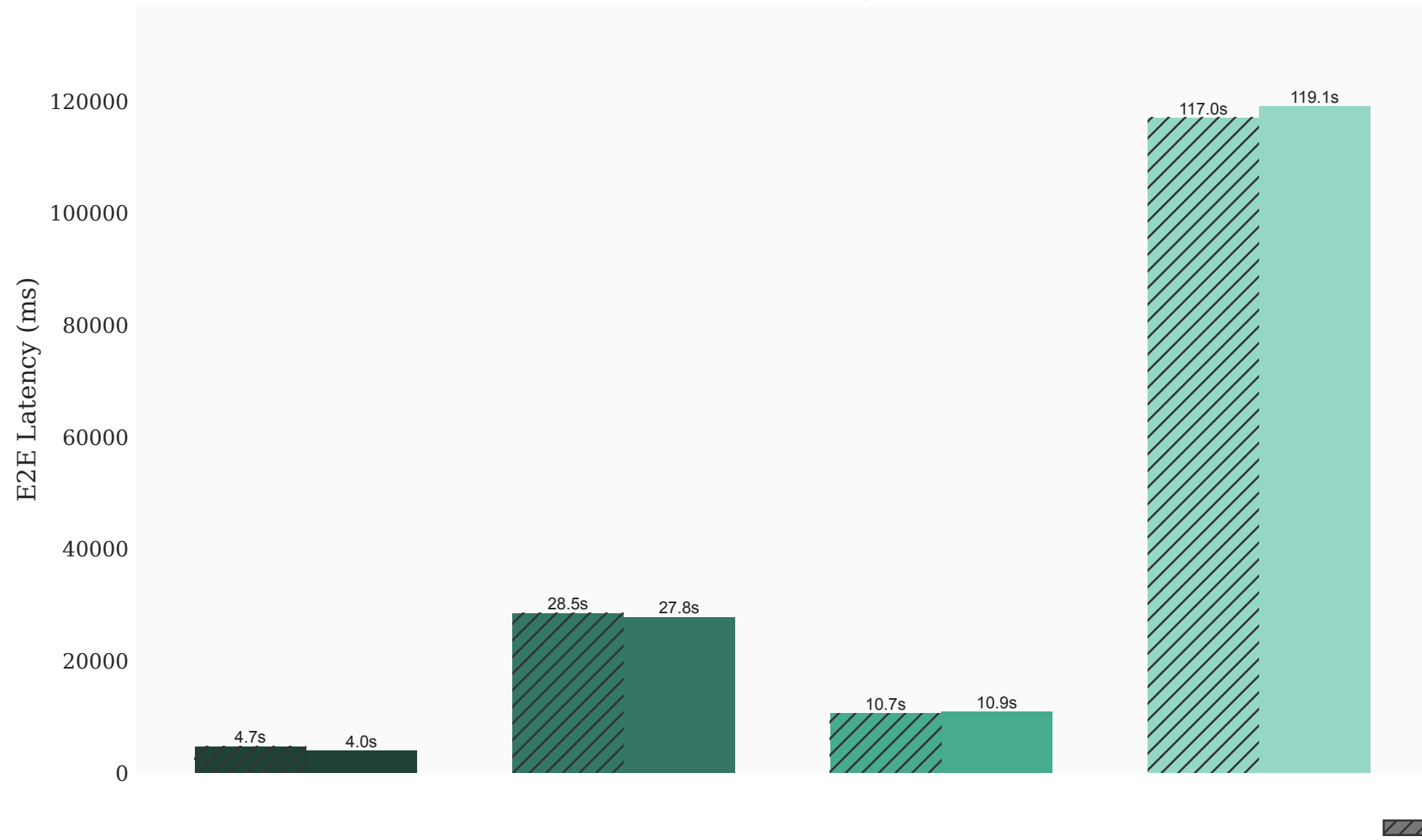
End-to-End Latency (P99)



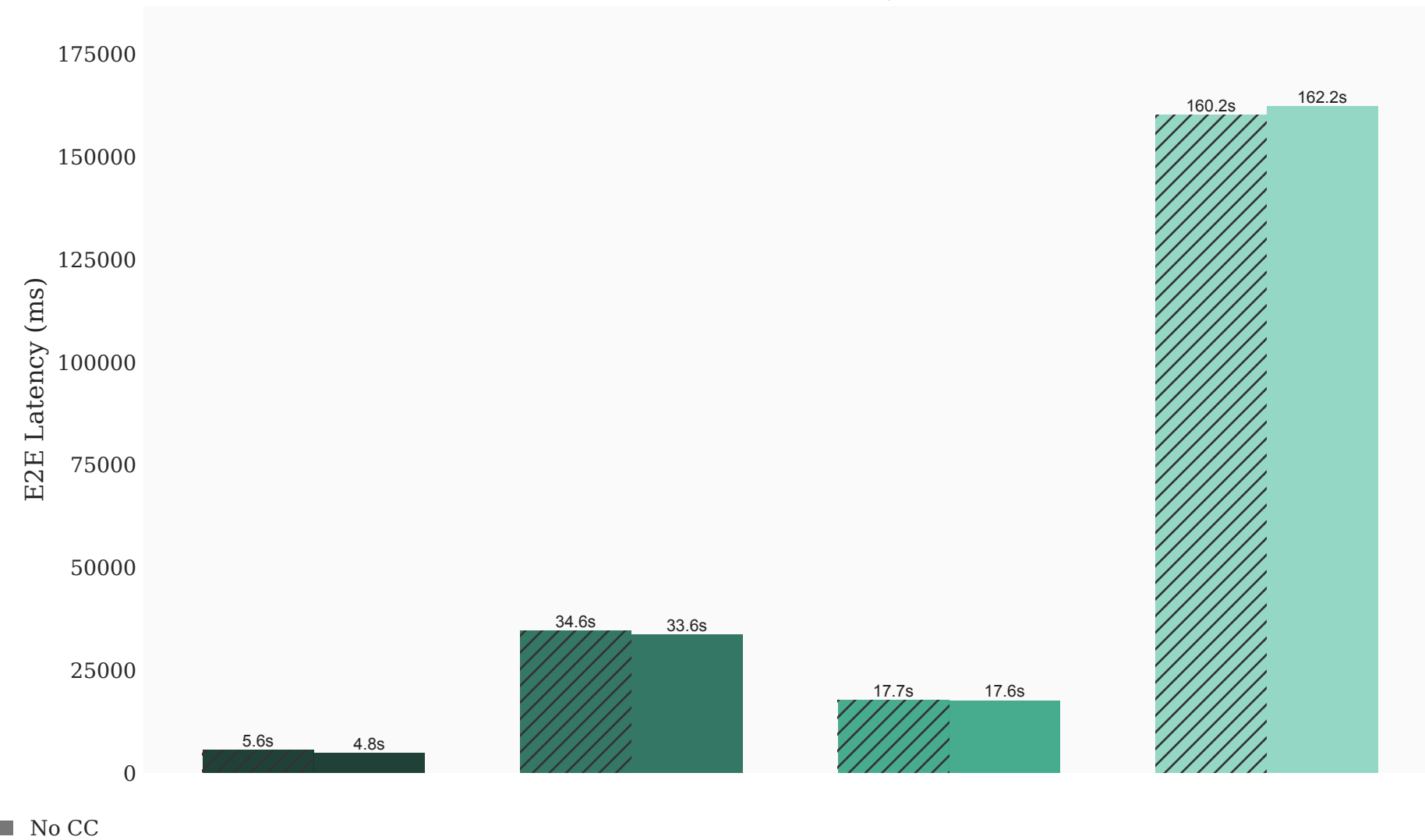
LLama 3.1 8B Mistral 3.1 24B GPT OSS 120B LLama 3.3 70B Int4

Random (1500 \Rightarrow 250) (50 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

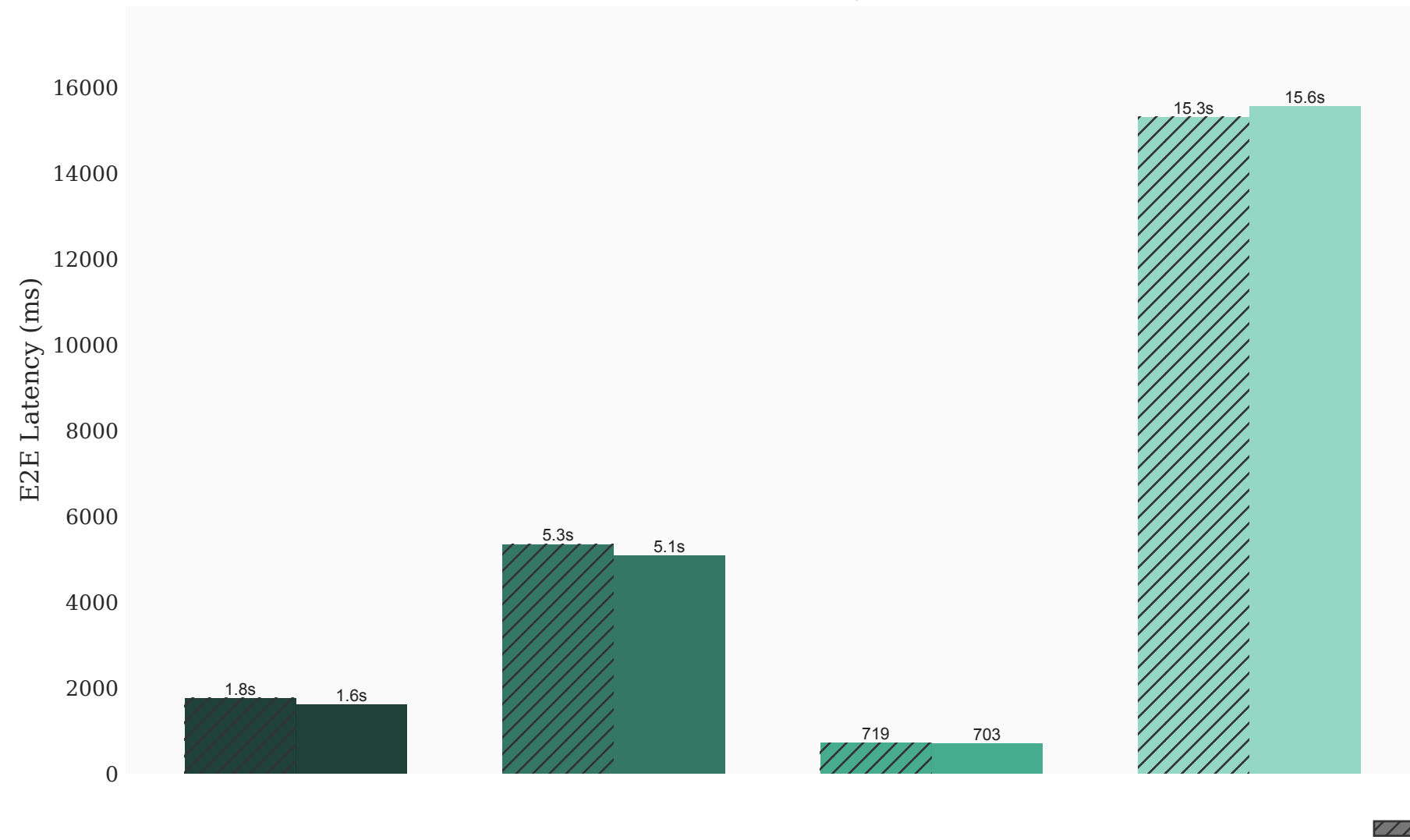


■ CC ■ No CC

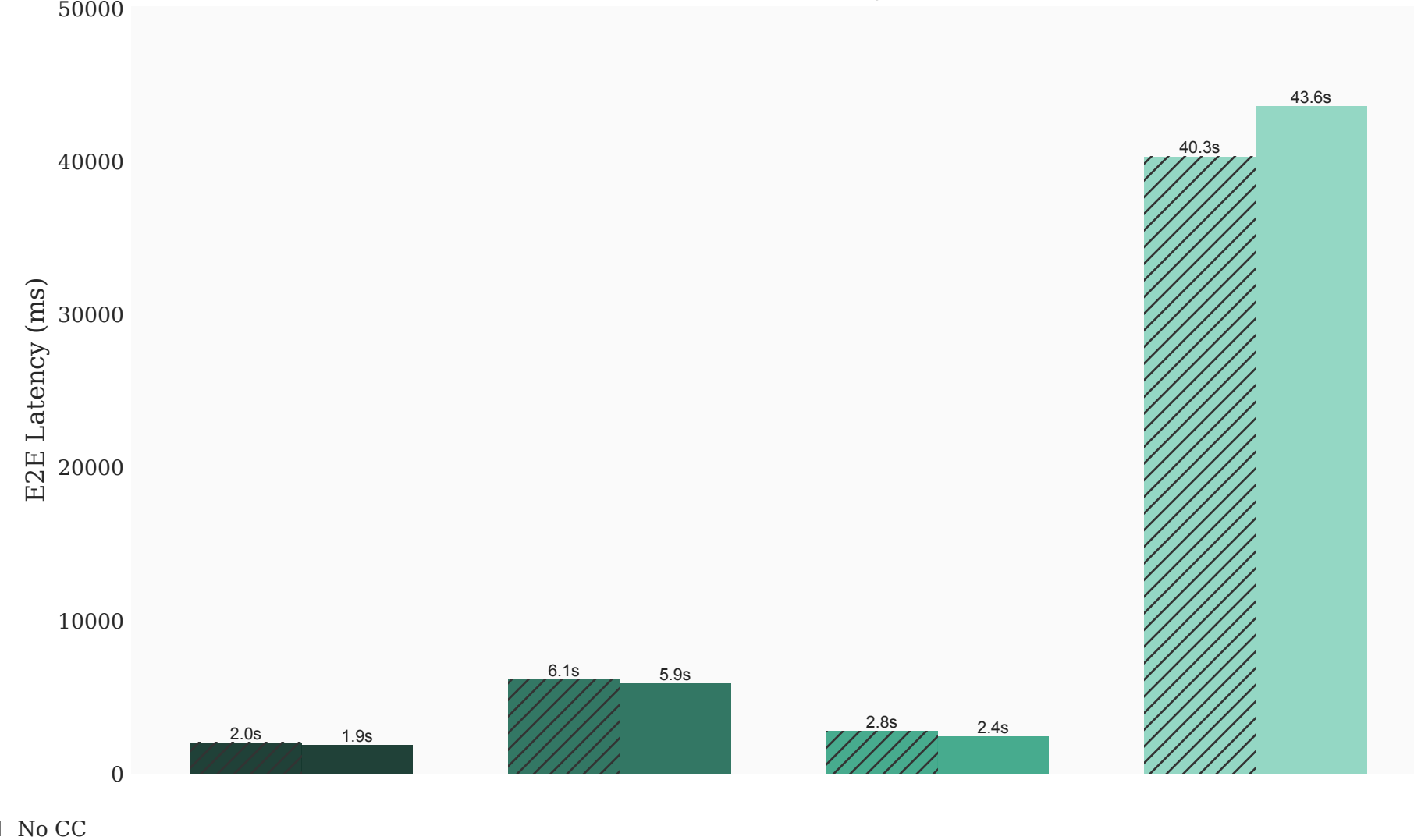
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (1500 \Rightarrow 250) (1 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

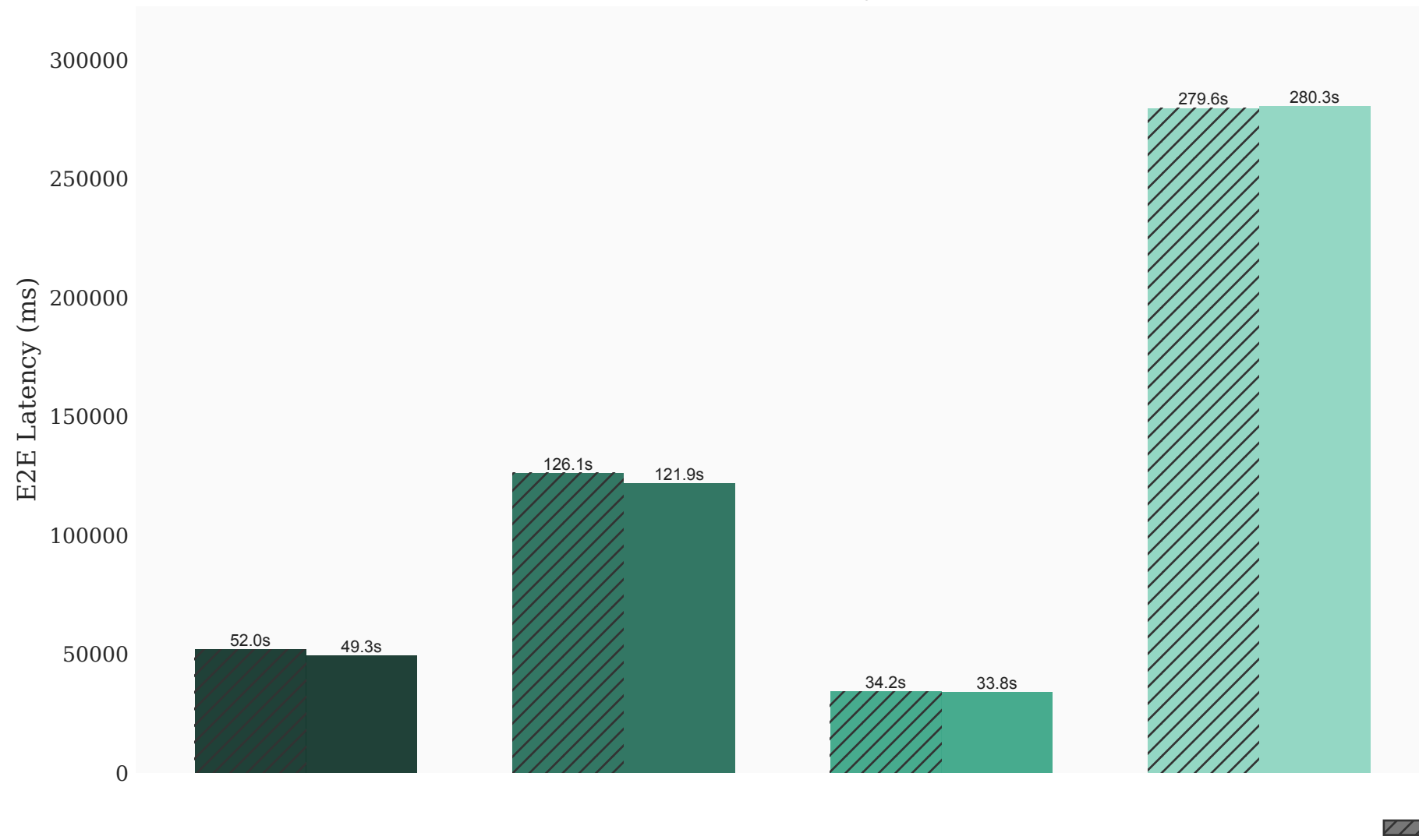


■ CC ■ No CC

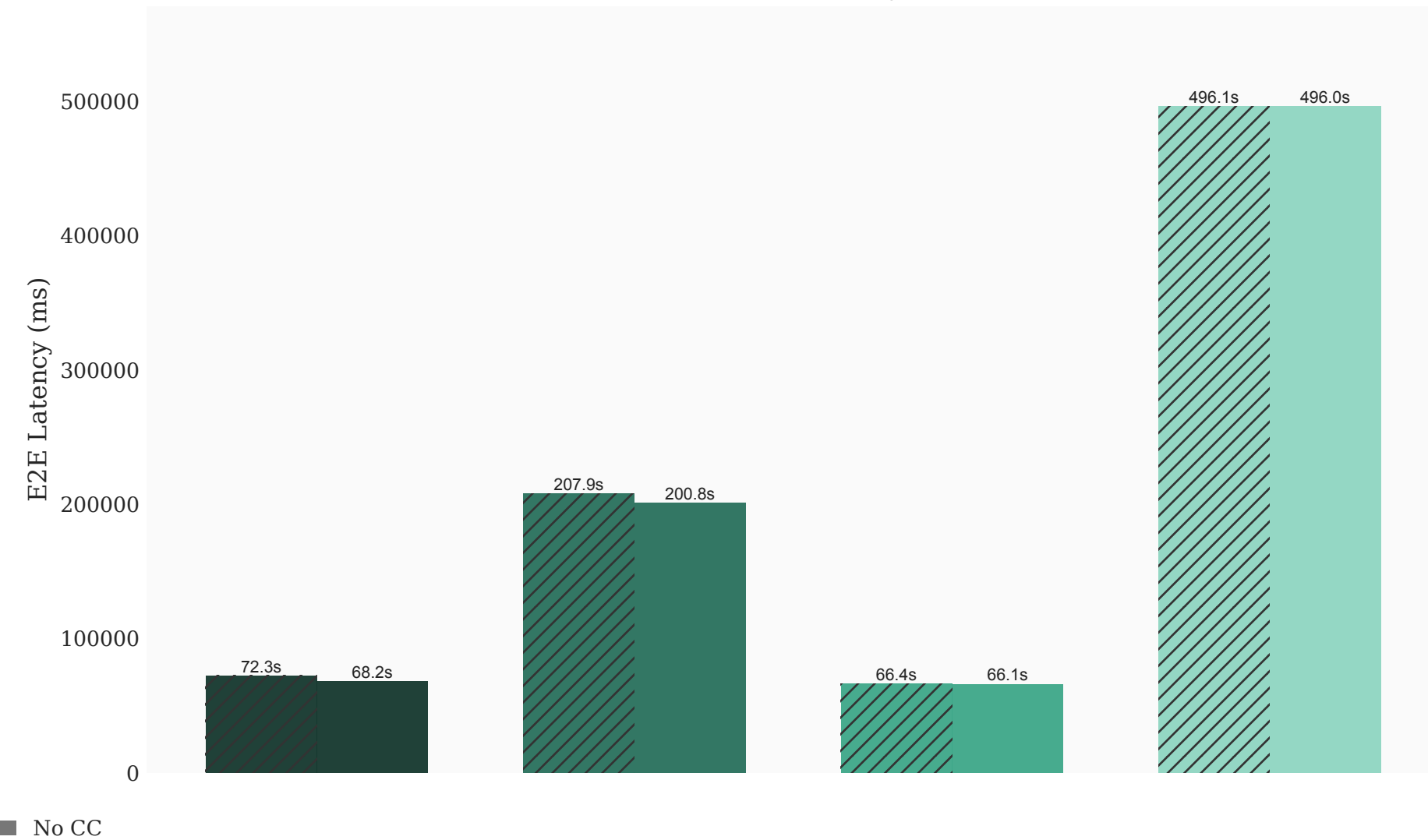
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Random (4000 \Rightarrow 1000) (100 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

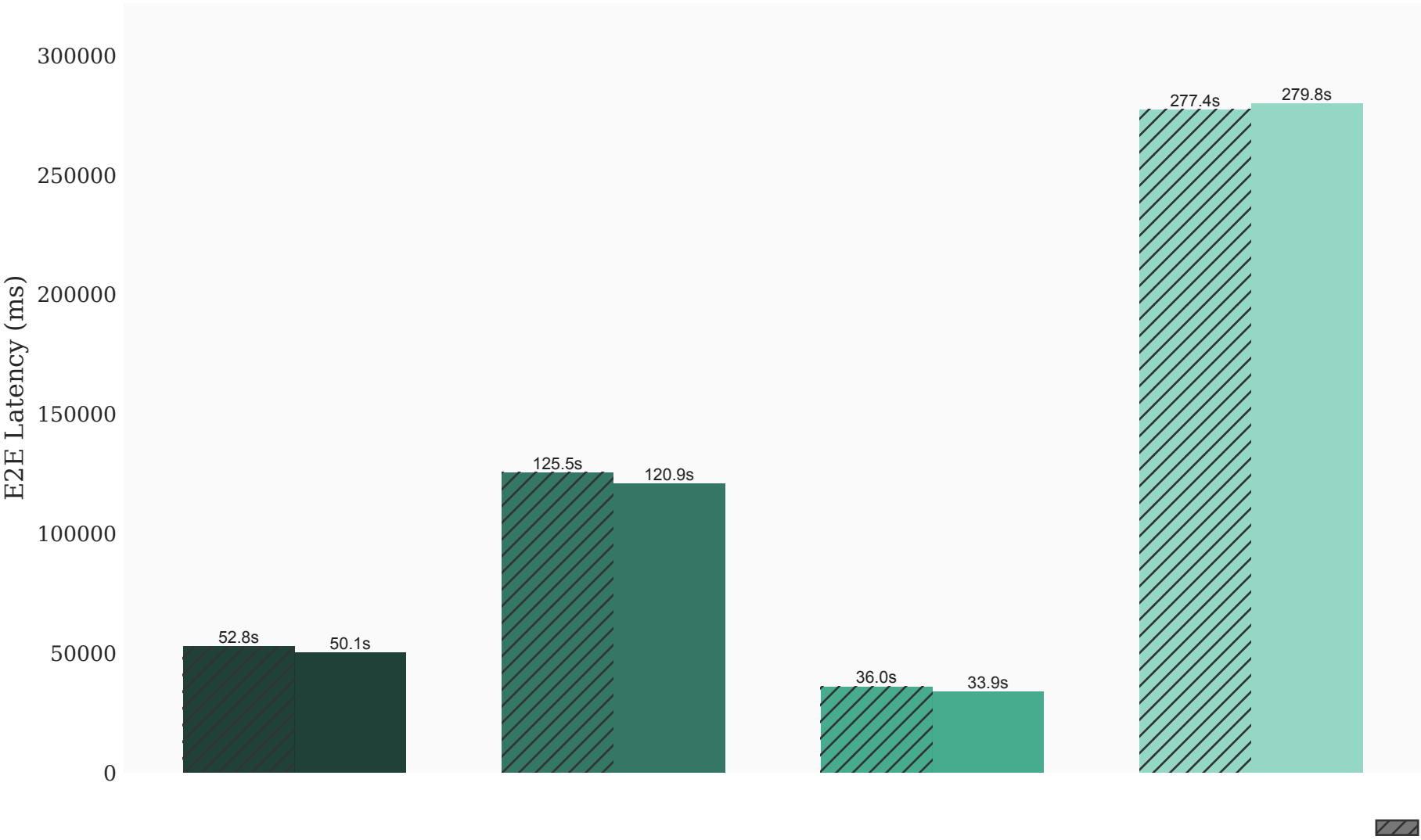


CC No CC

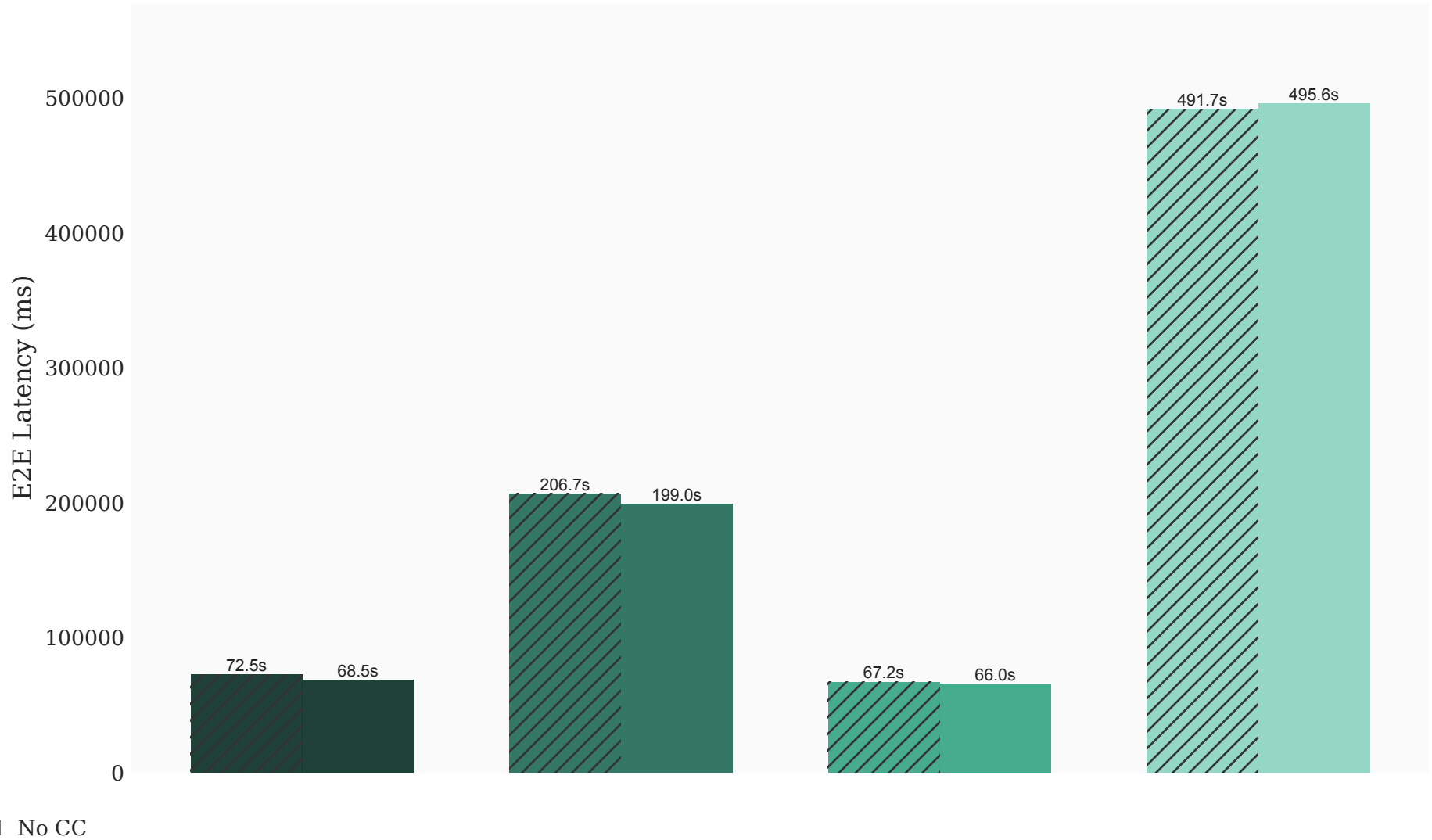
Llama 3.1 8B Mistral 3.1 24B GPT OSS 120B Llama 3.3 70B Int4

Random (4000 \Rightarrow 1000) (50 Concurrent Requests)

End-to-End Latency (Mean)



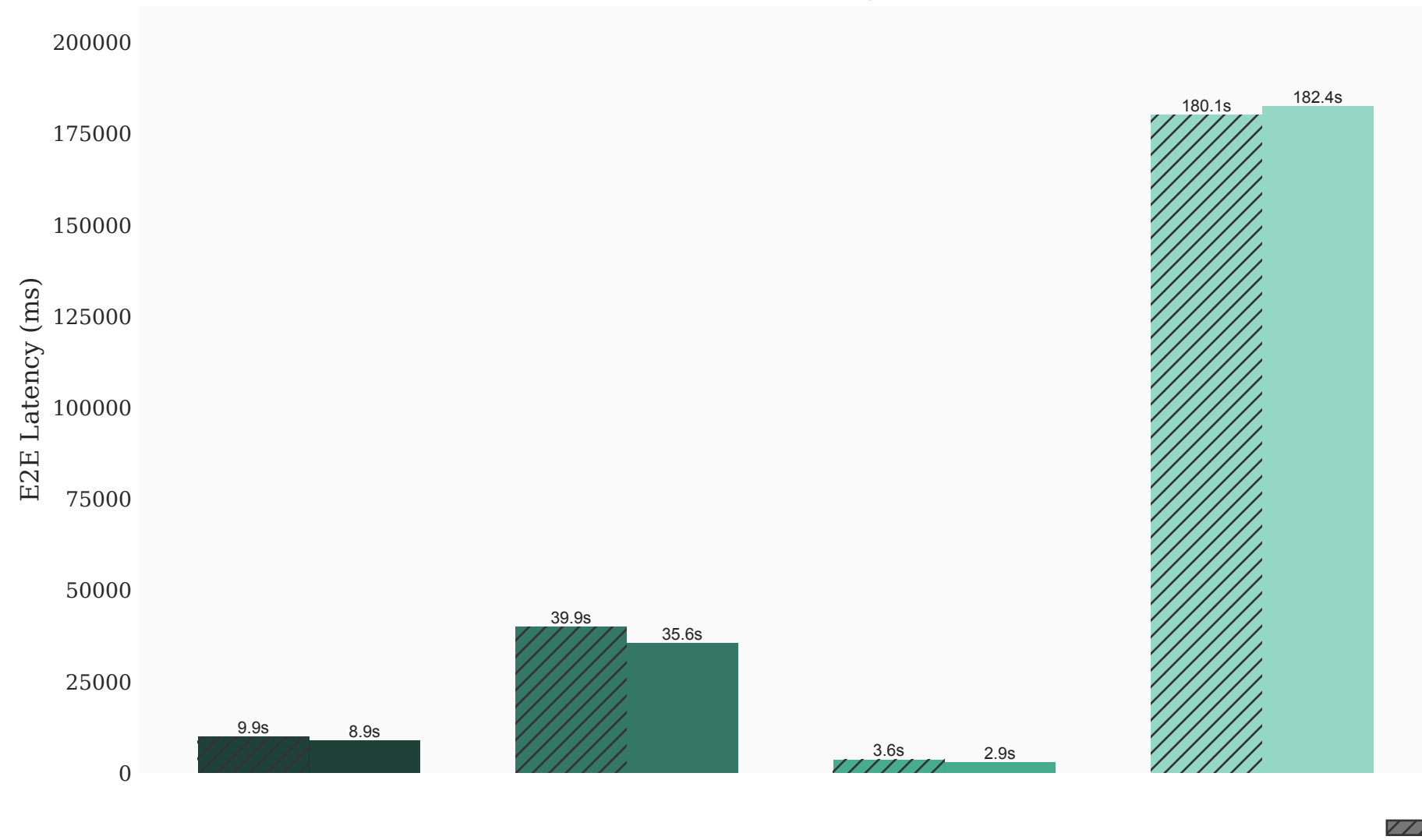
End-to-End Latency (P99)



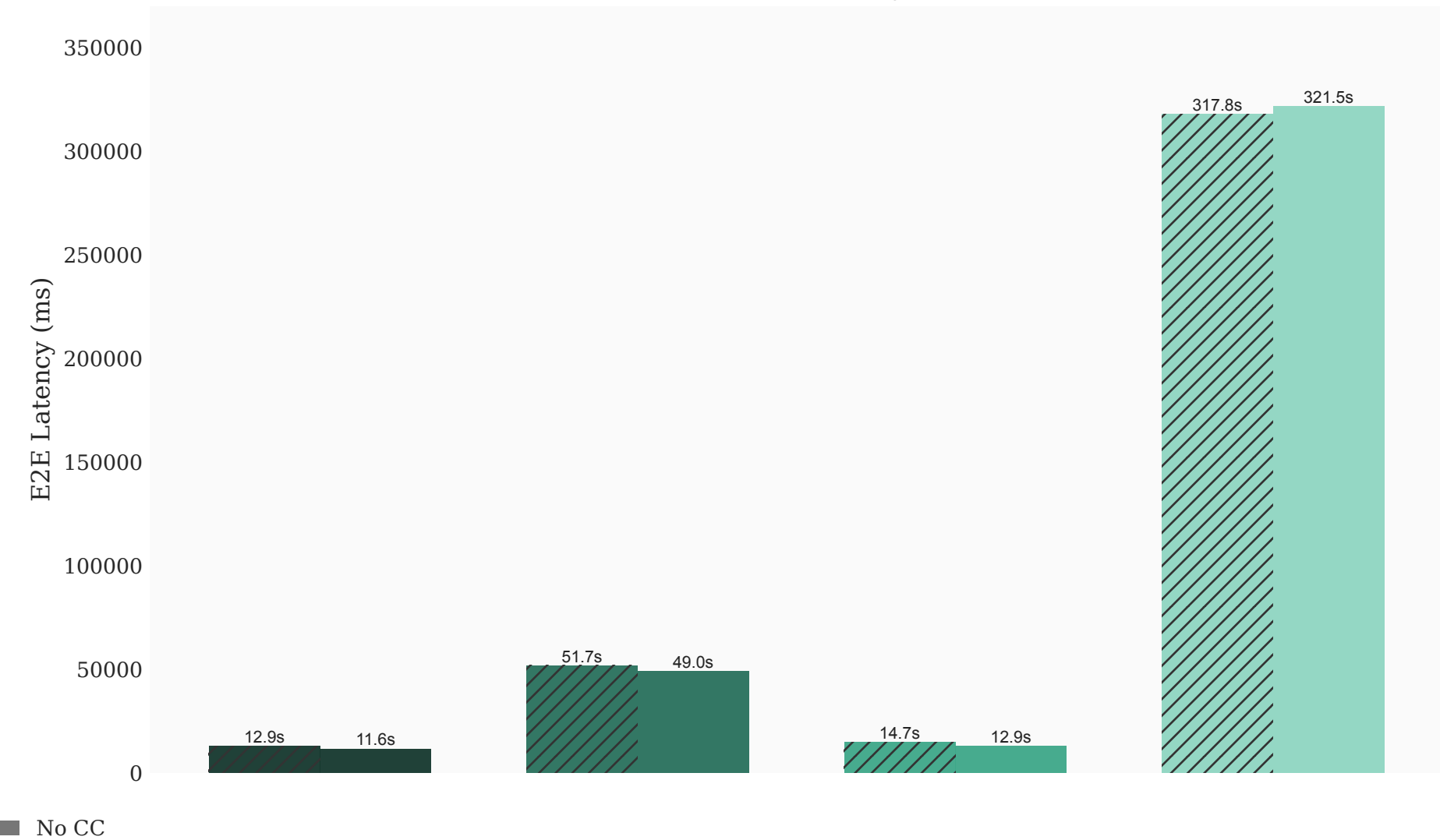
LLama 3.1 8B Mistral 3.1 24B GPT OSS 120B LLama 3.3 70B Int4

Random (4000 \Rightarrow 1000) (1 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

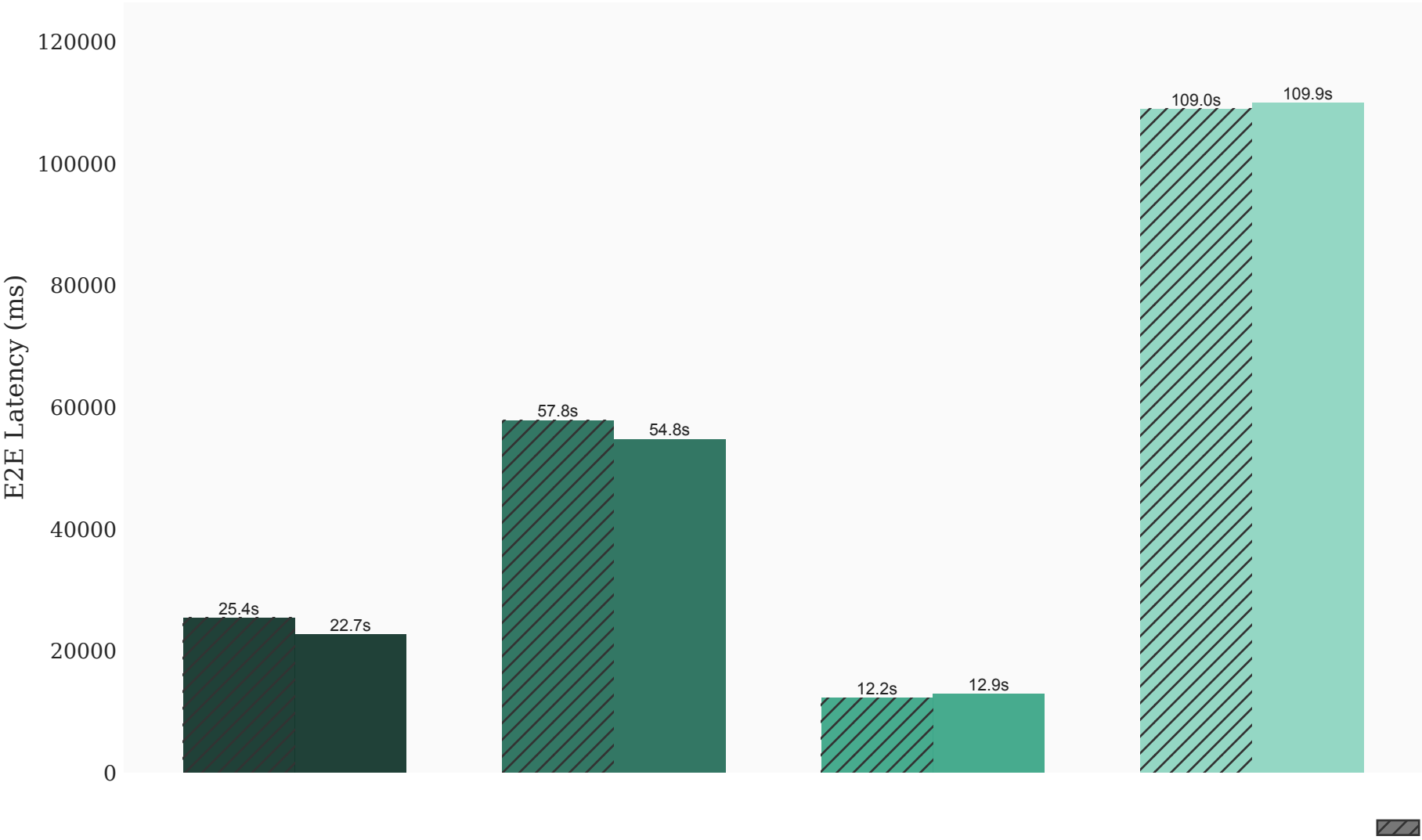


■ CC ■ No CC

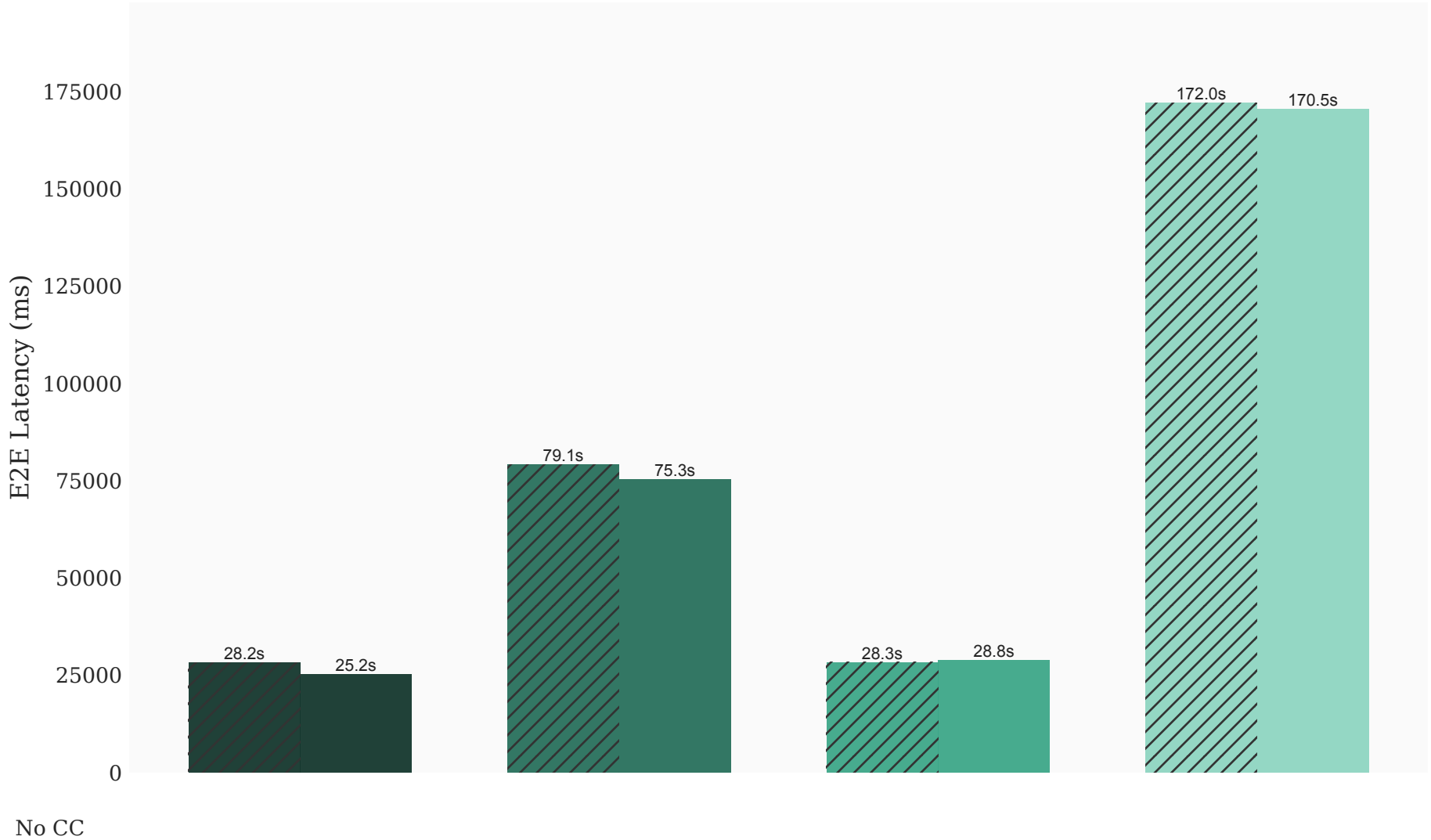
■ Llama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ Llama 3.3 70B Int4

Random (1000 \Rightarrow 1000) (100 Concurrent Requests)

End-to-End Latency (Mean)



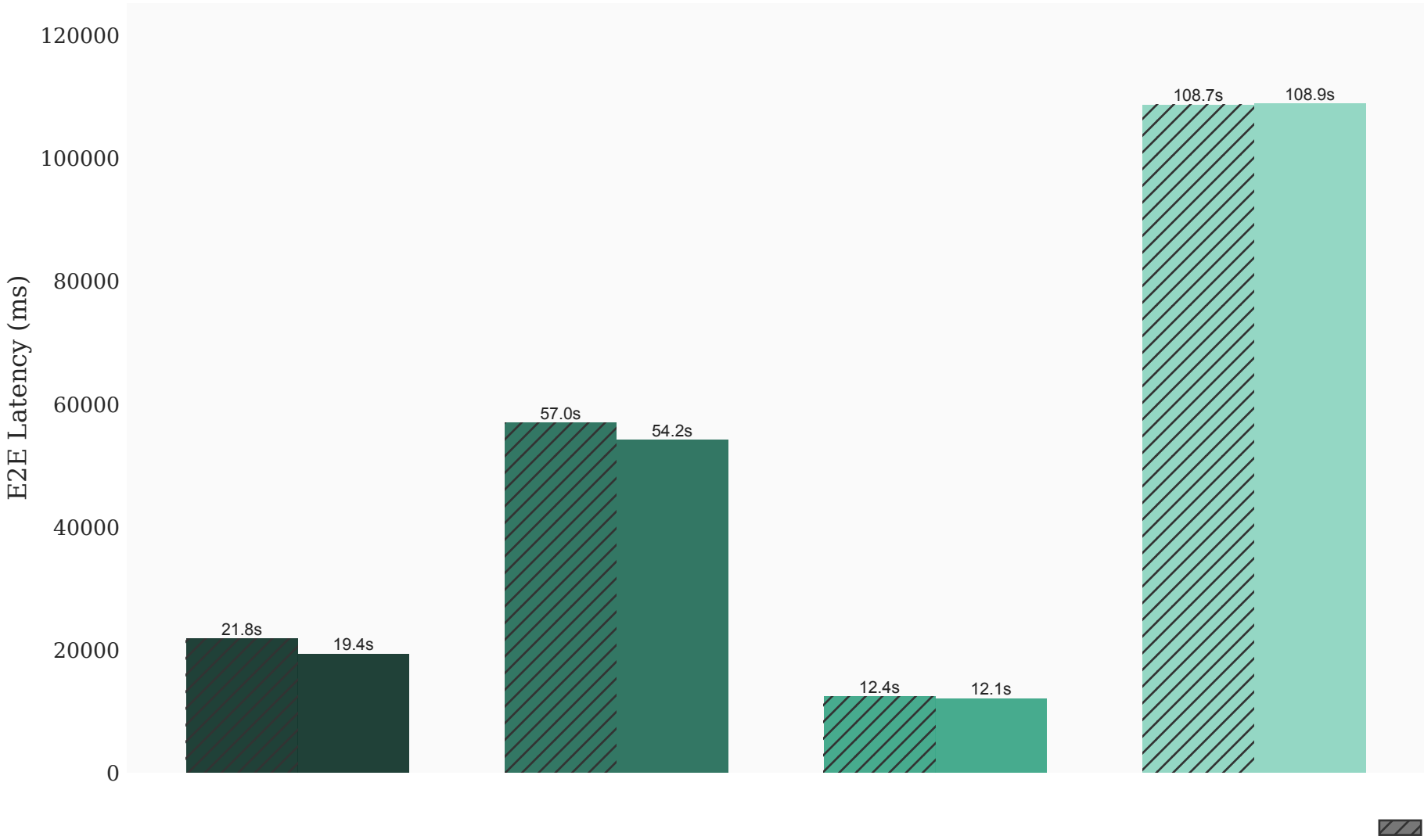
End-to-End Latency (P99)



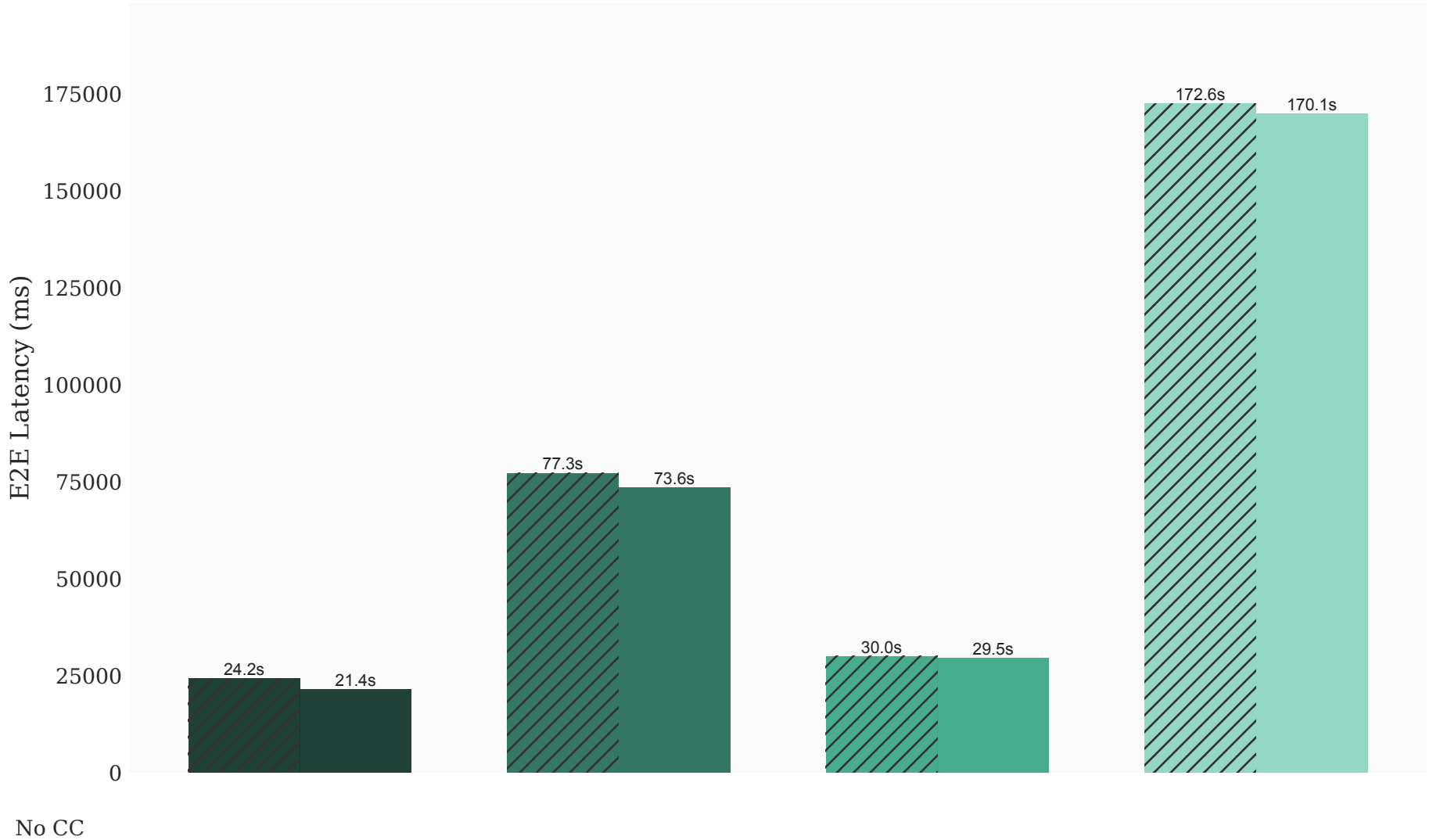
■ Llama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ Llama 3.3 70B Int4

Random (1000 ⇒ 1000) (50 Concurrent Requests)

End-to-End Latency (Mean)



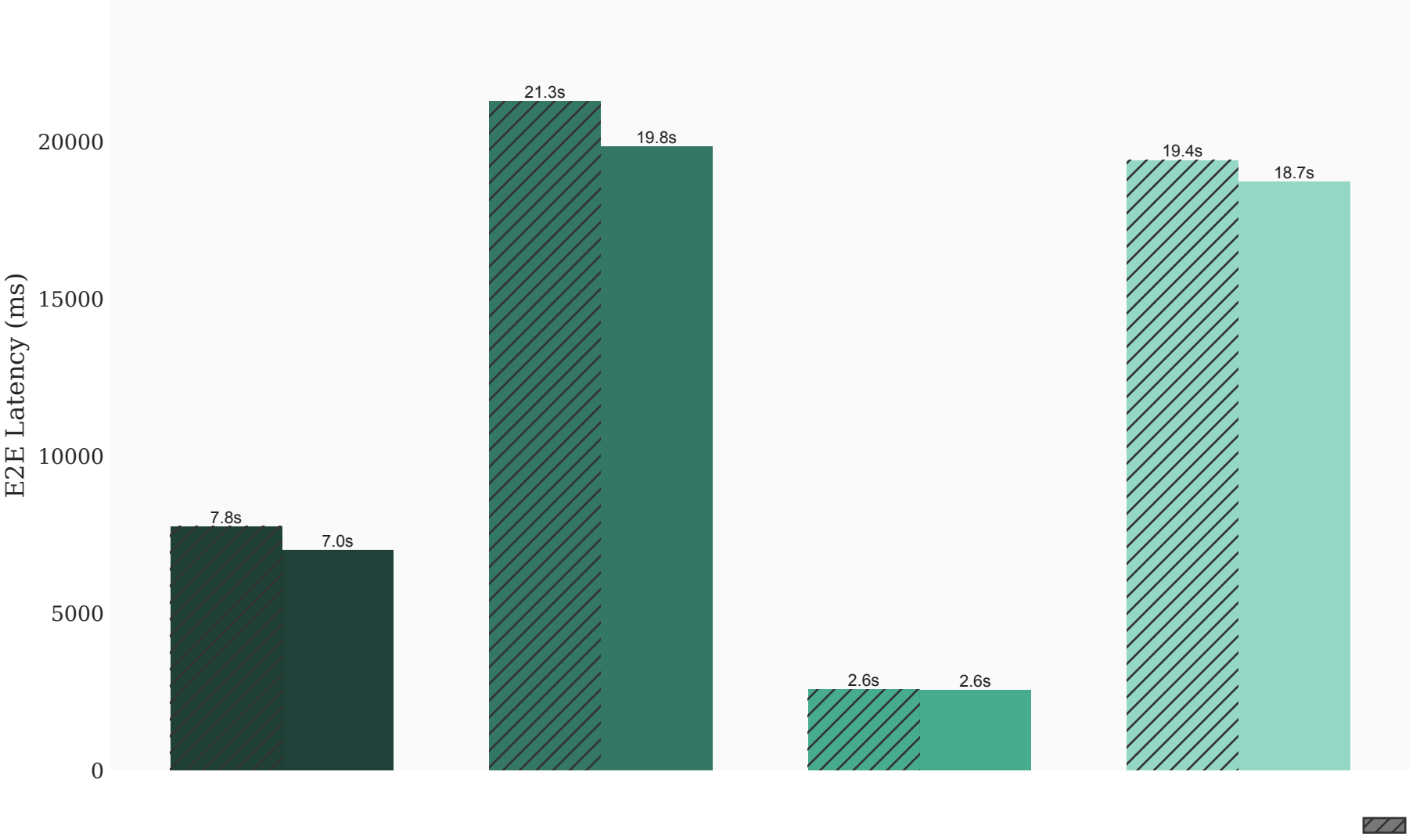
End-to-End Latency (P99)



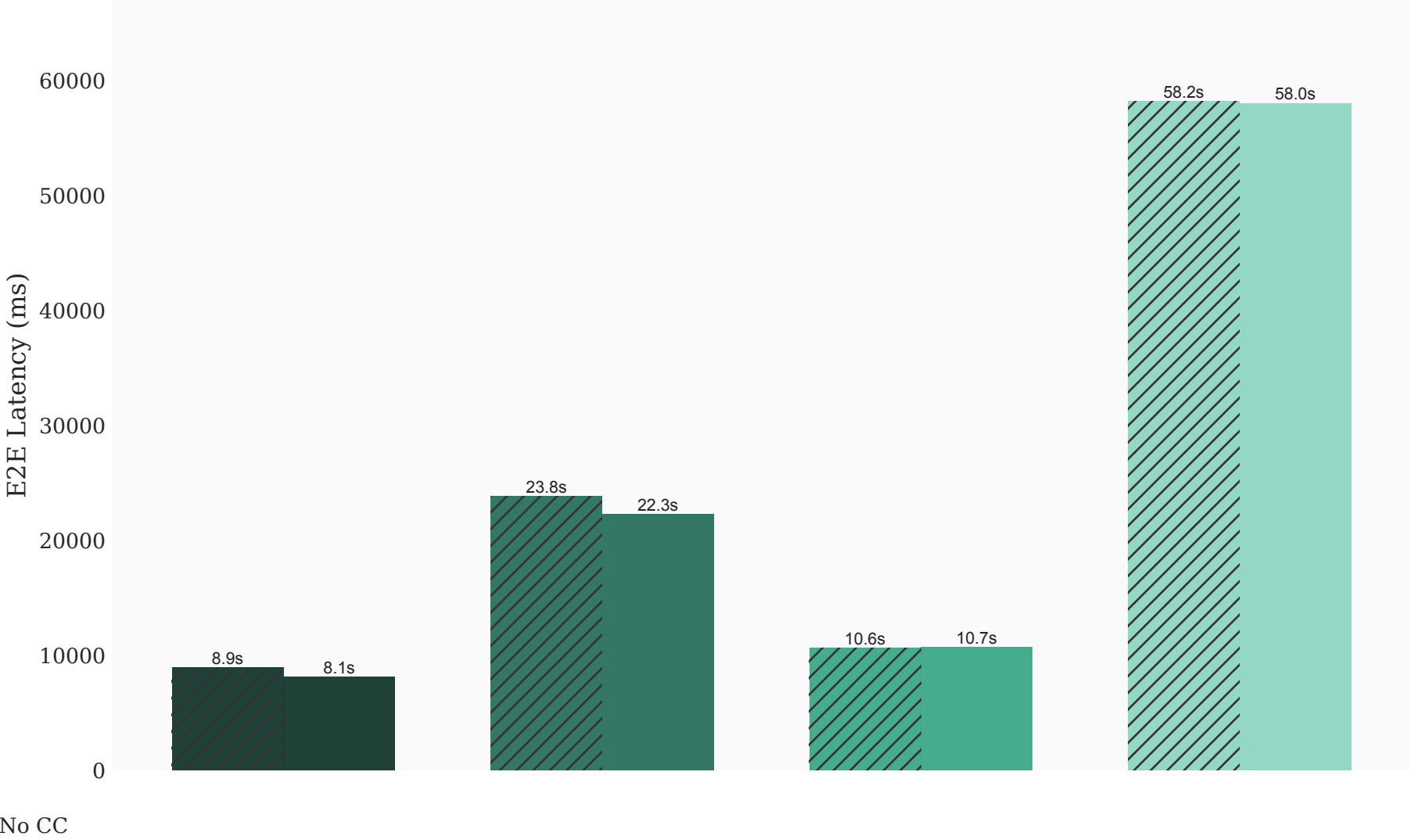
LLama 3.1 8B Mistral 3.1 24B GPT OSS 120B LLama 3.3 70B Int4

Random (1000 ⇒ 1000) (1 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

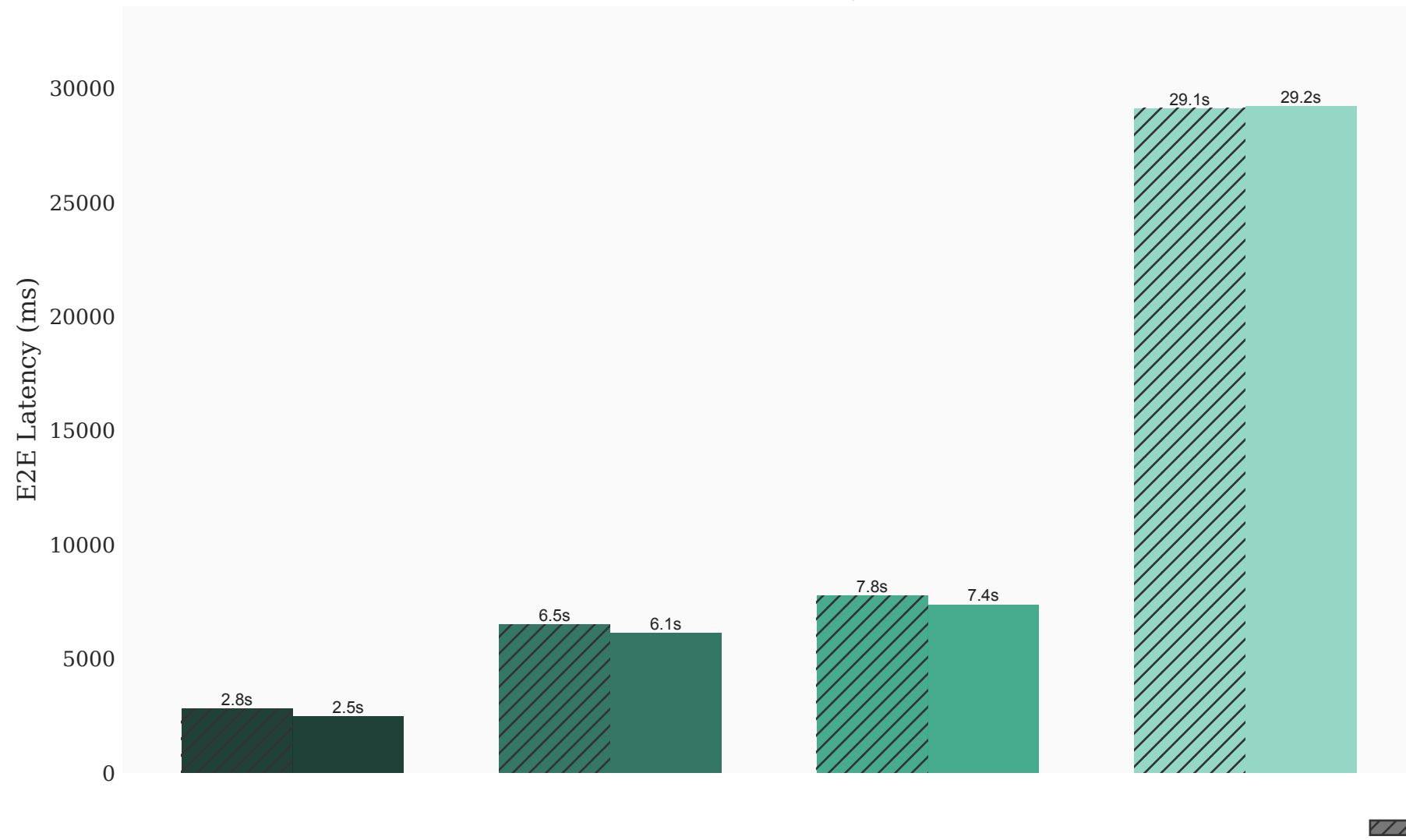


■ CC ■ No CC

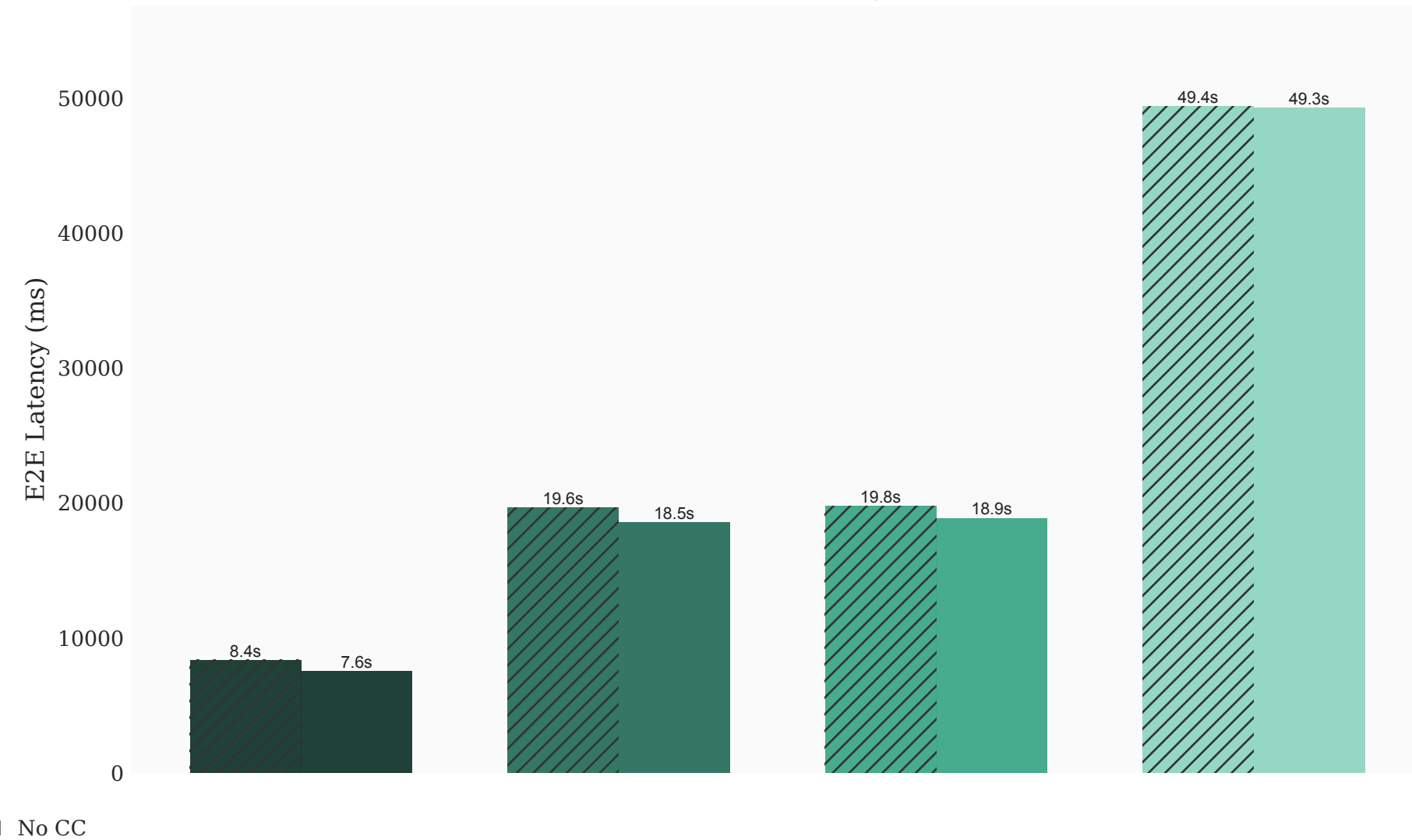
■ Llama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ Llama 3.3 70B Int4

ShareGPT (100 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

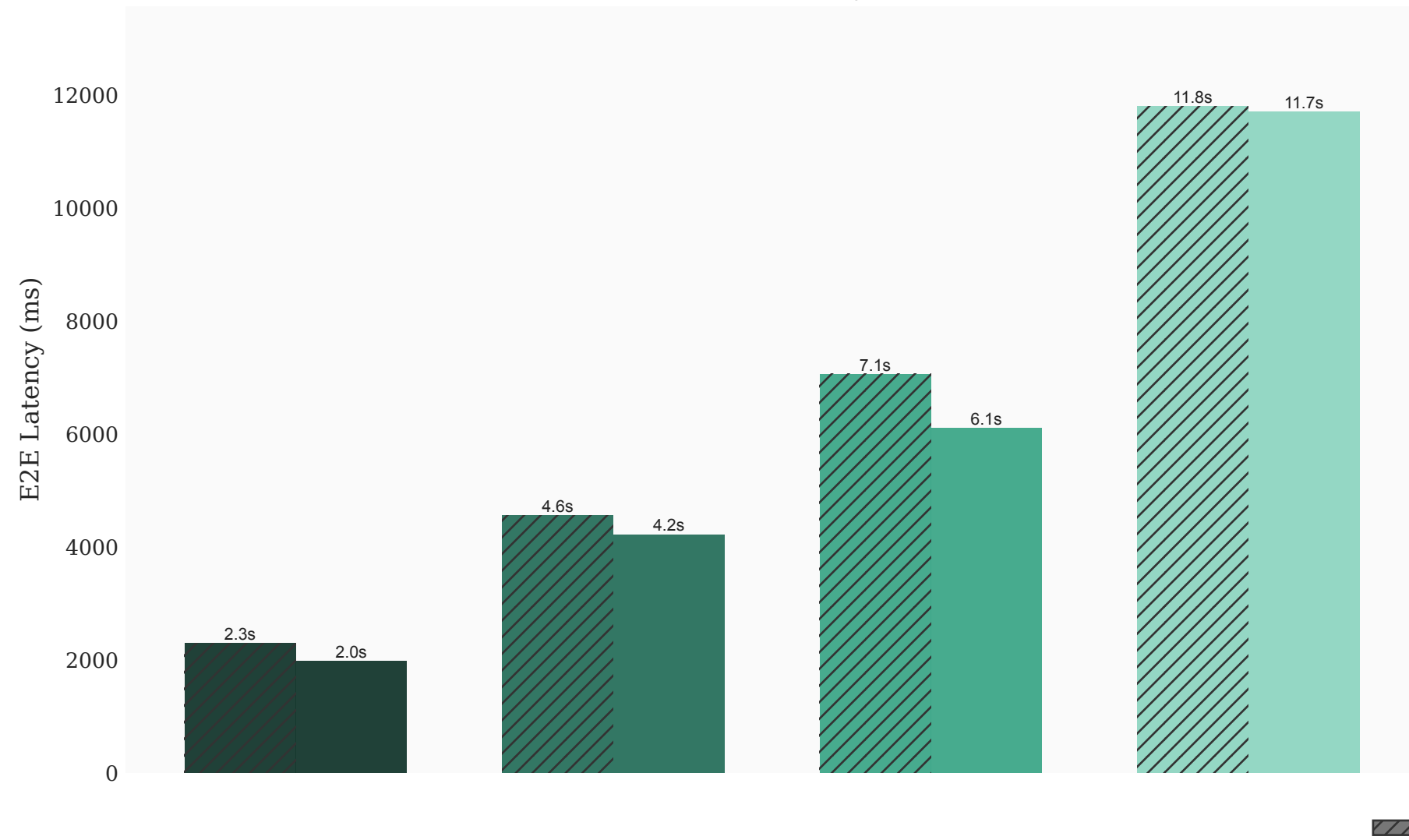


■ CC ■ No CC

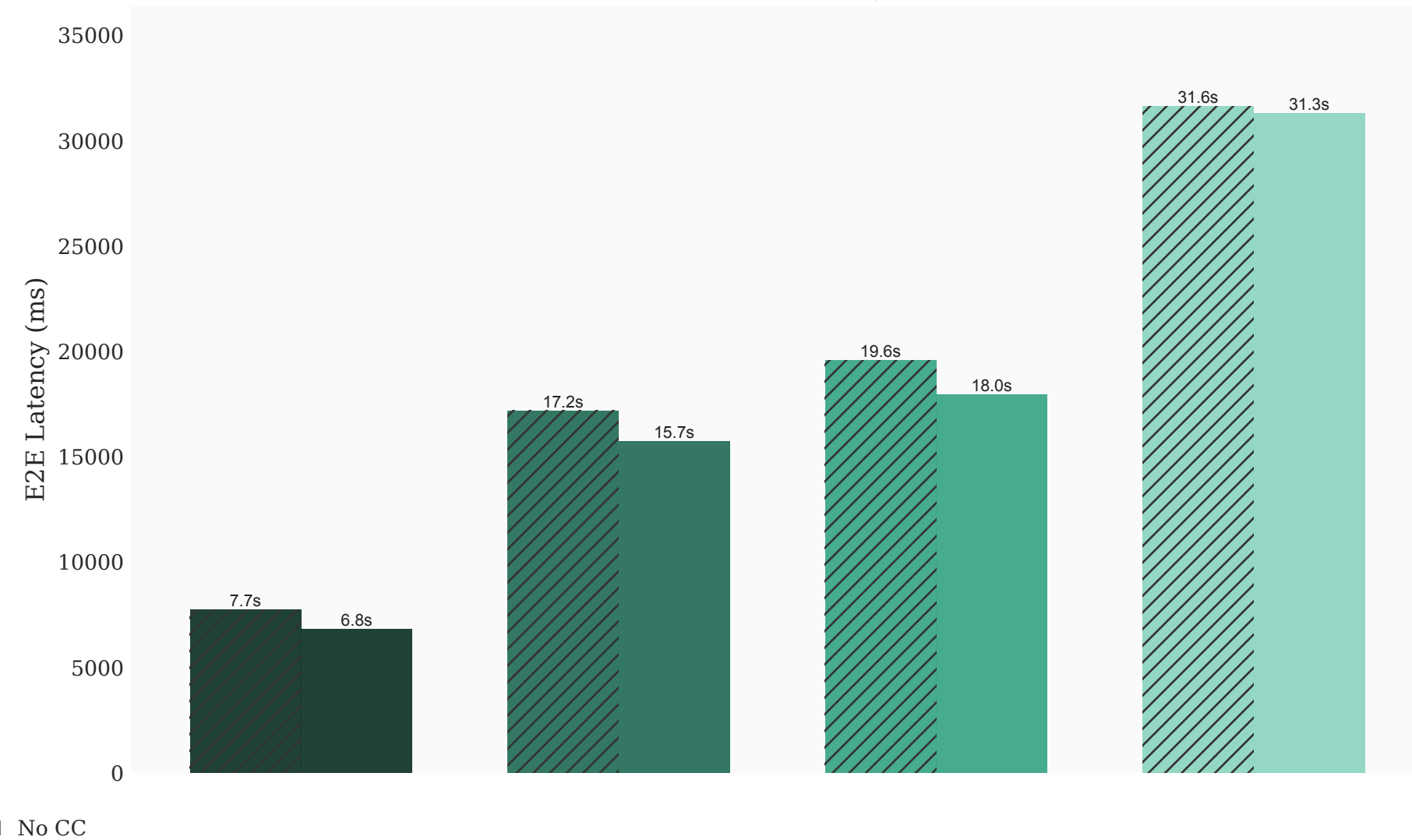
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

ShareGPT (50 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

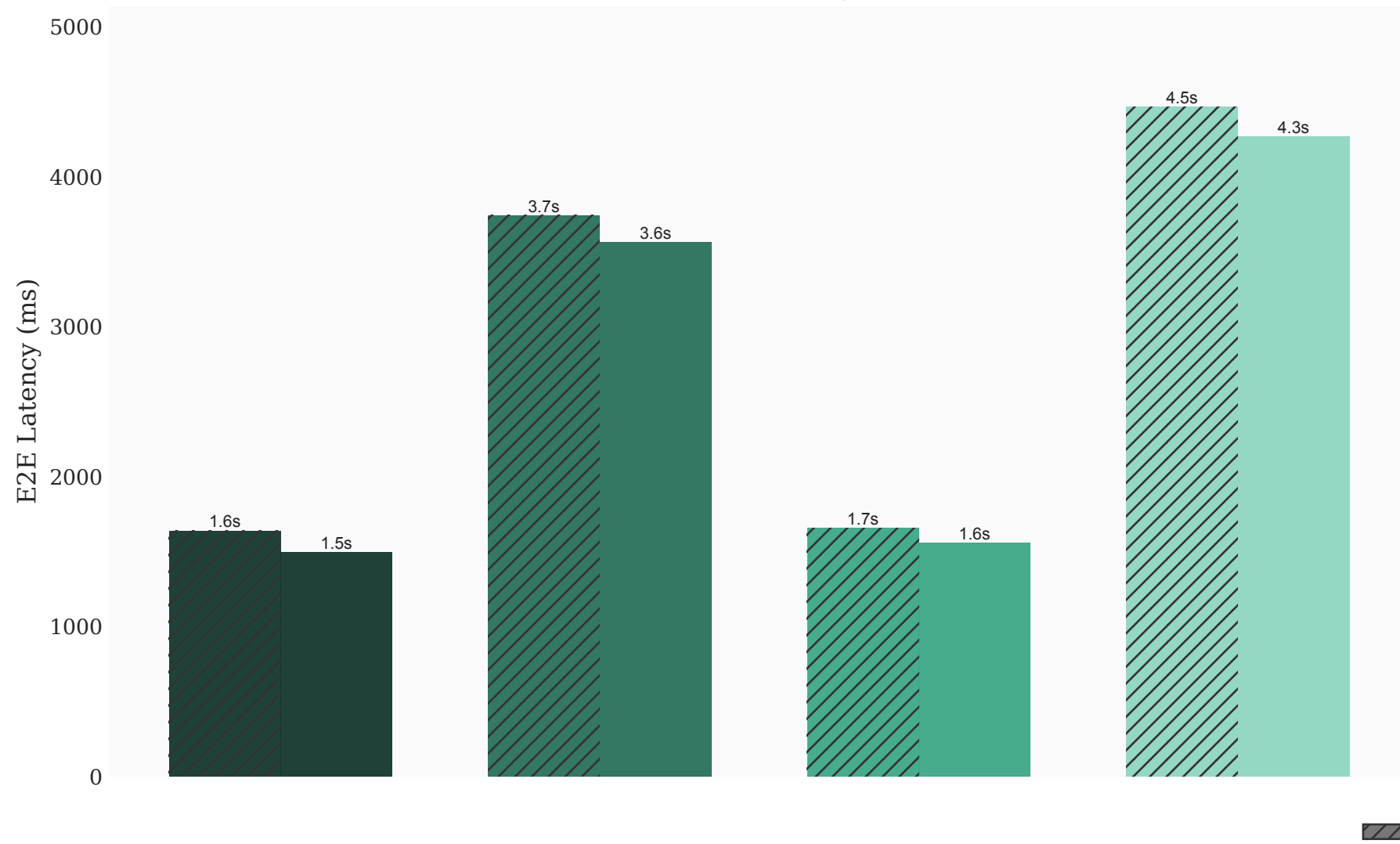


■ CC ■ No CC

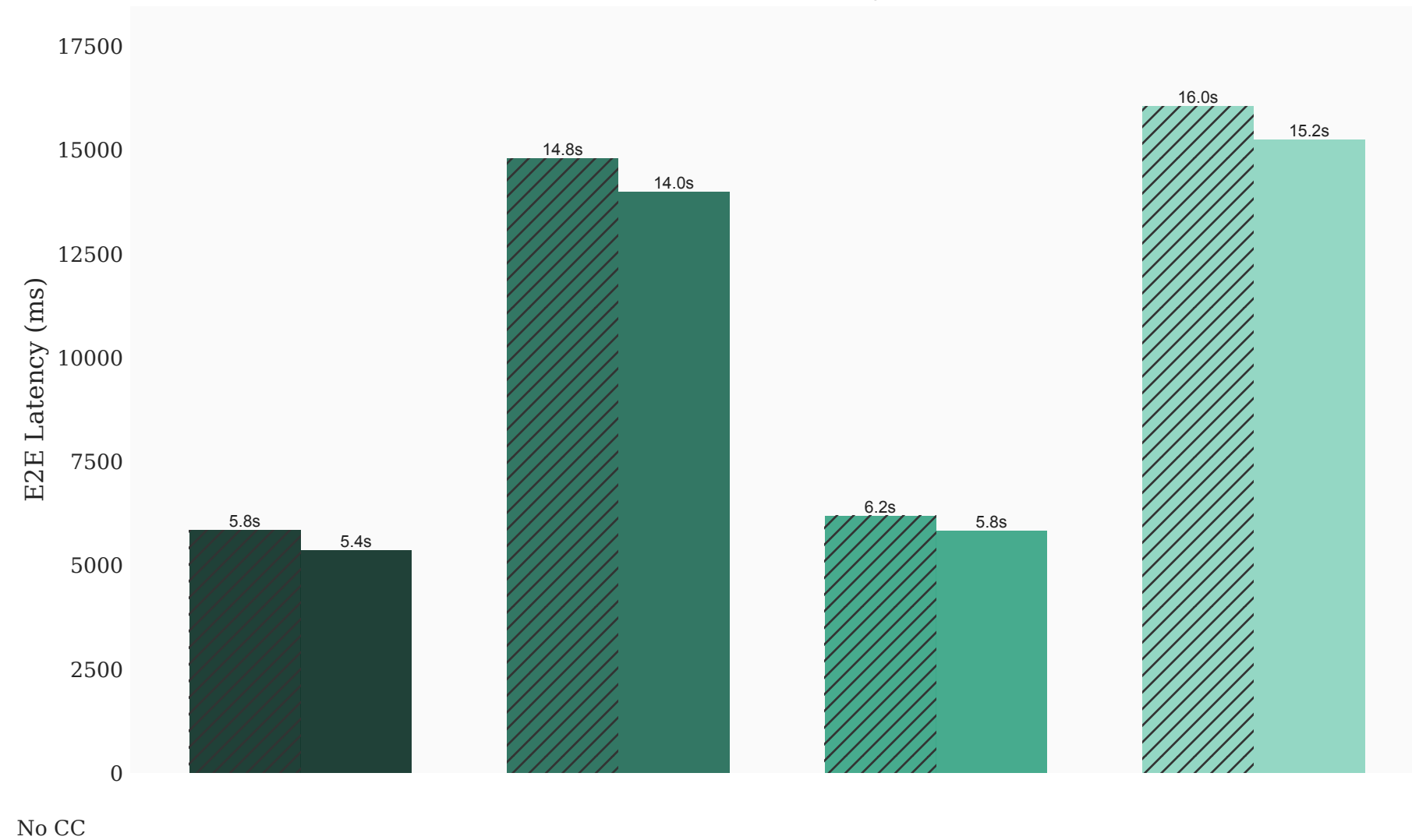
■ Llama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ Llama 3.3 70B Int4

ShareGPT (1 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

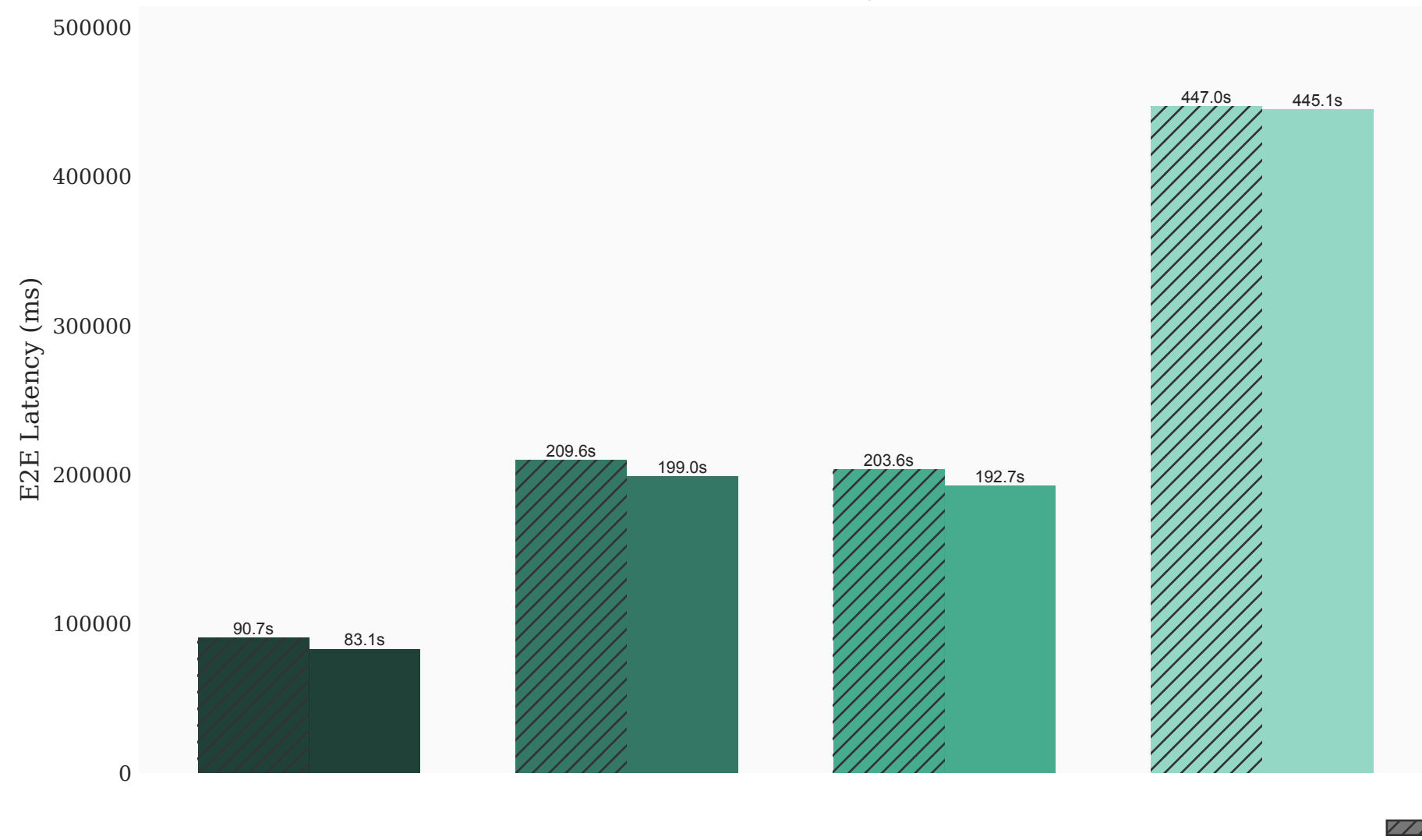


■ CC ■ No CC

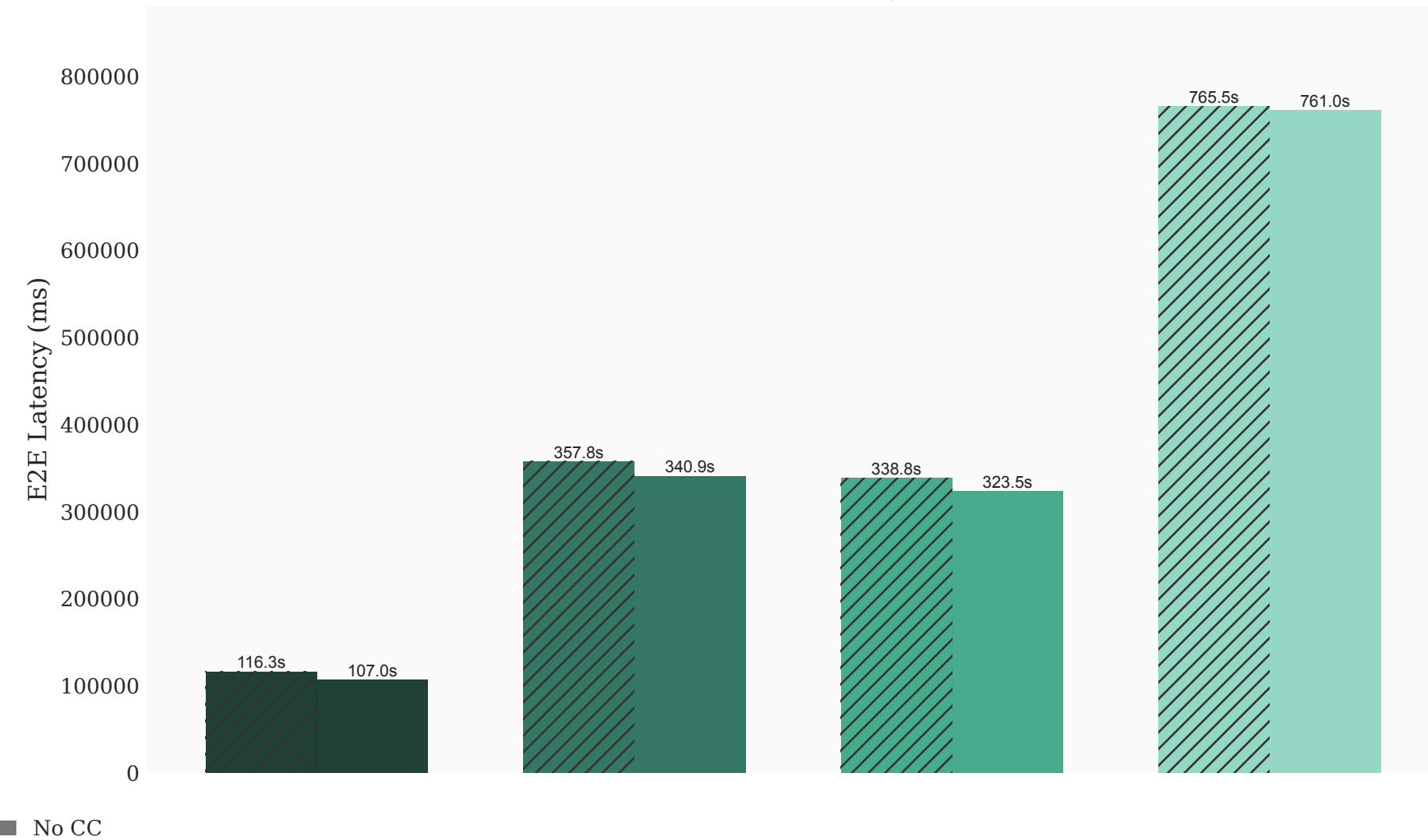
■ Llama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ Llama 3.3 70B Int4

Edit 10K Characters (100 Concurrent Requests)

End-to-End Latency (Mean)



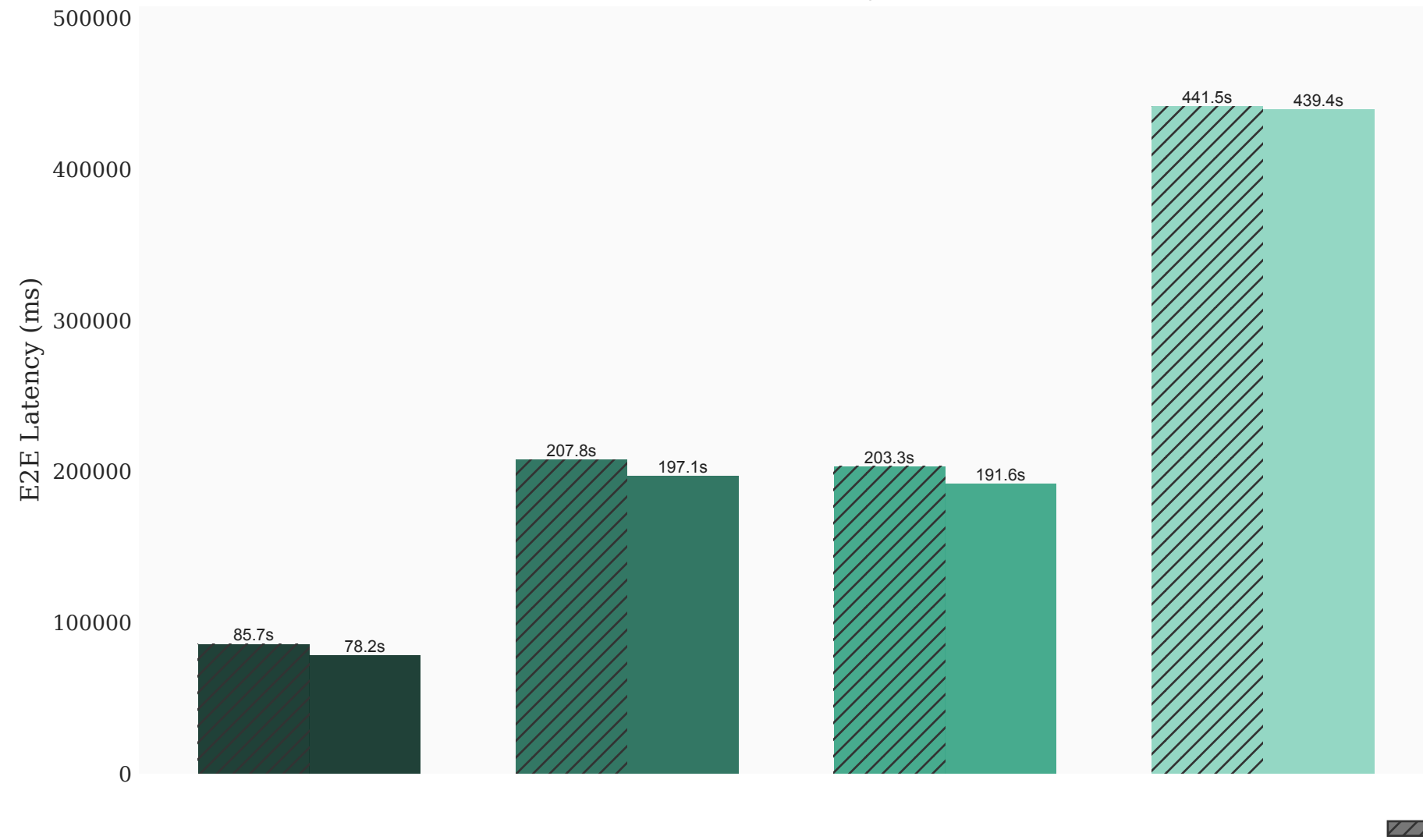
End-to-End Latency (P99)



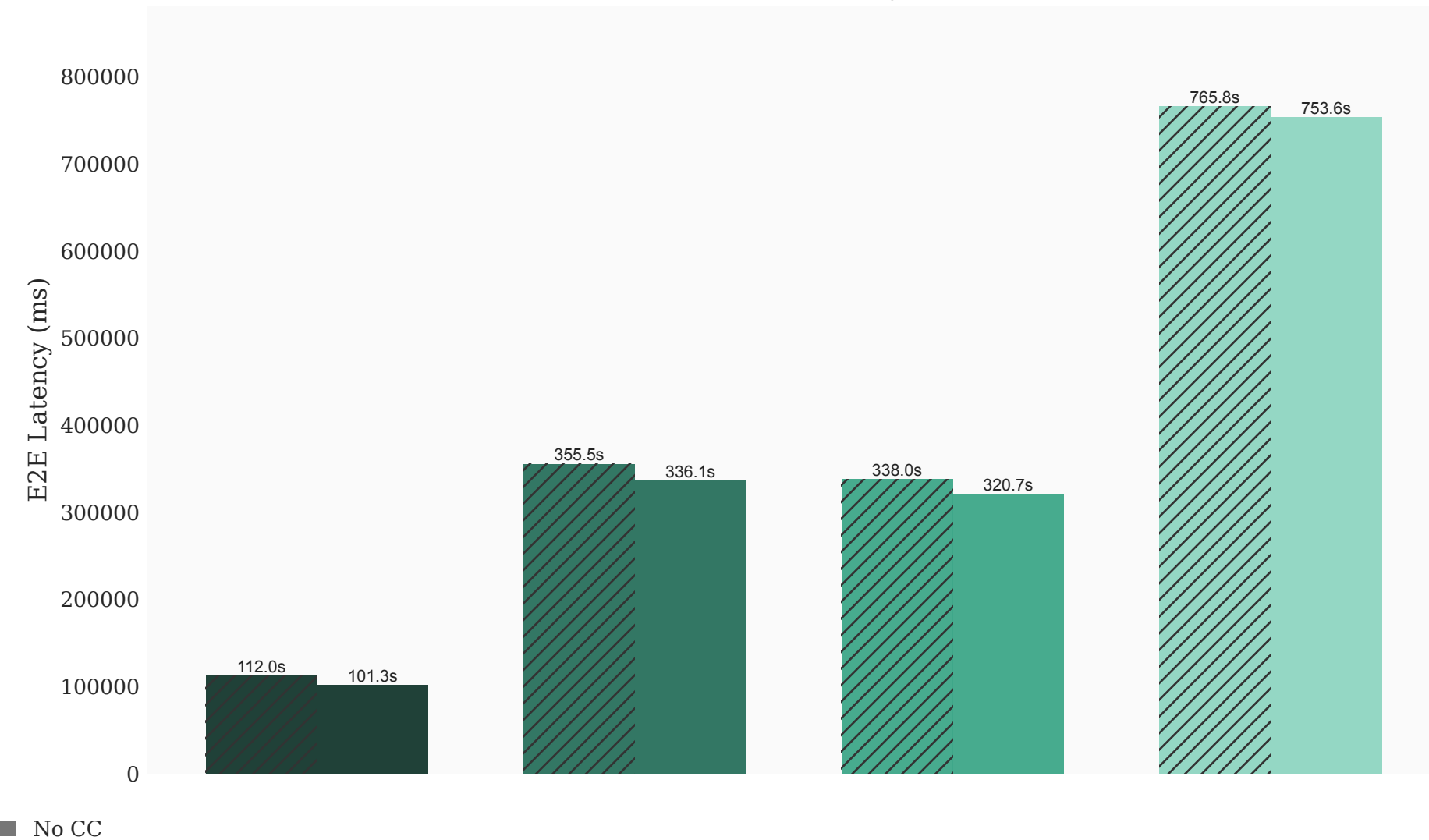
■ Llama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ Llama 3.3 70B Int4

Edit 10K Characters (50 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

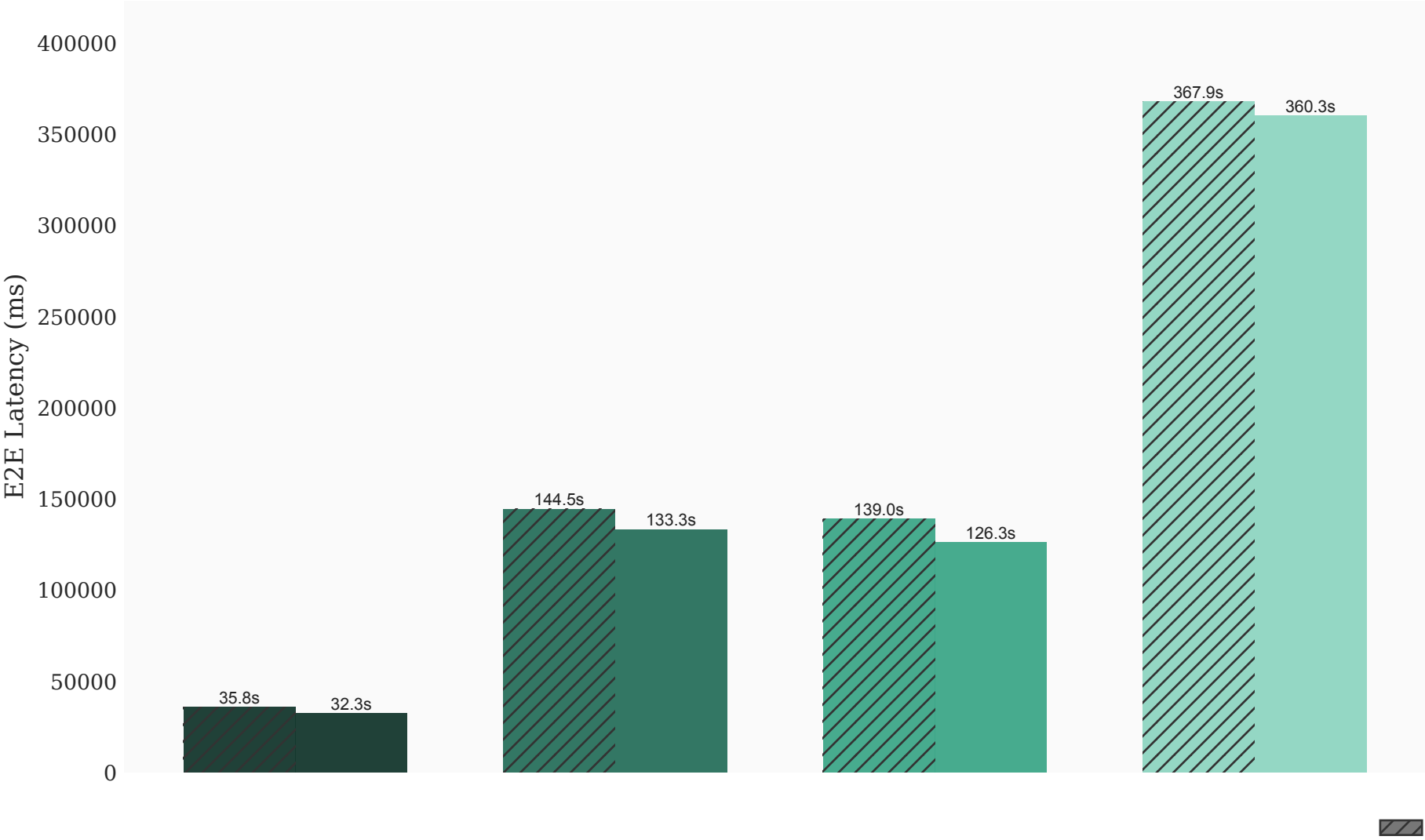


■ Llama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ Llama 3.3 70B Int4

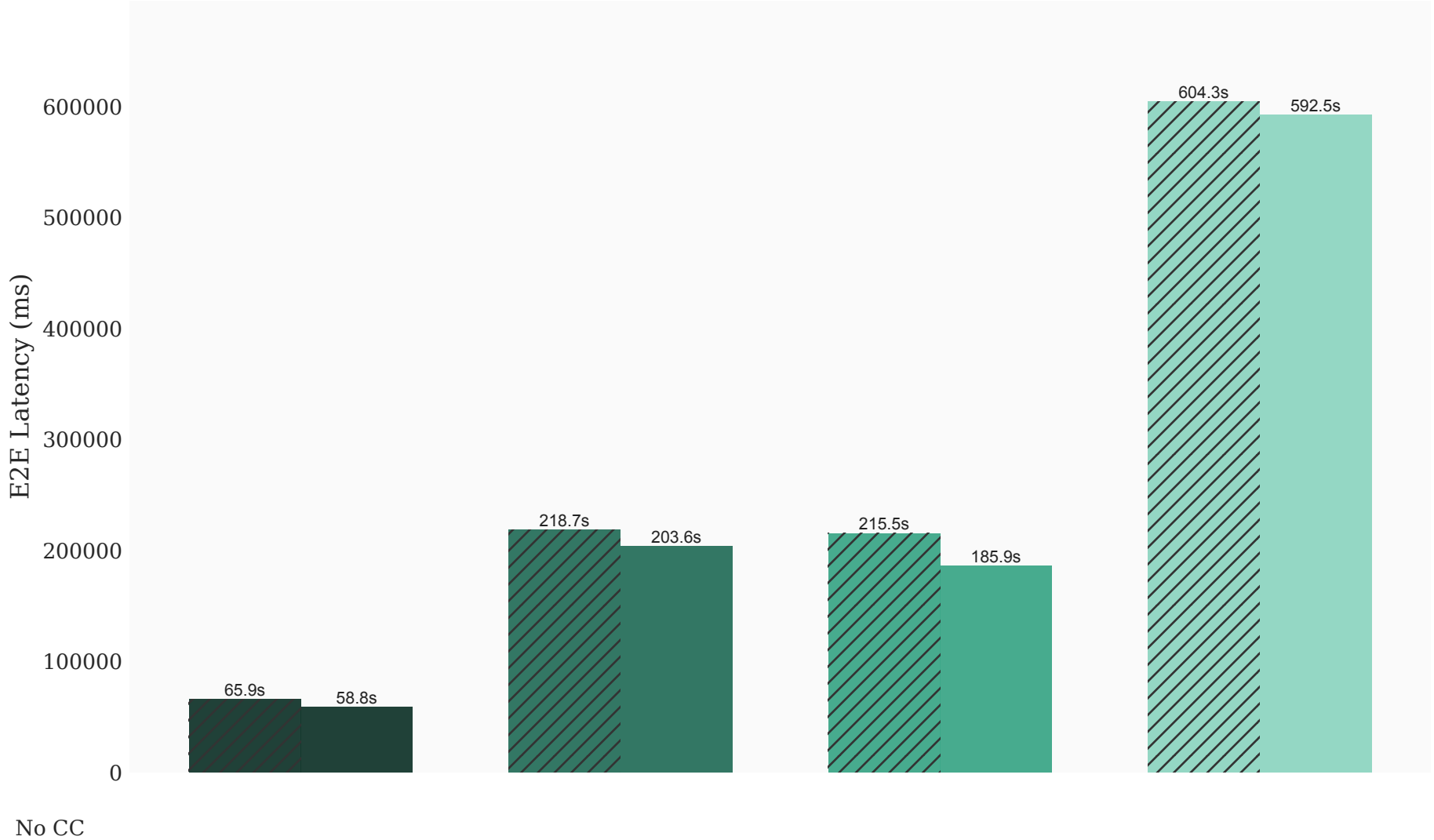
▨ CC ■ No CC

Edit 10K Characters (1 Concurrent Requests)

End-to-End Latency (Mean)



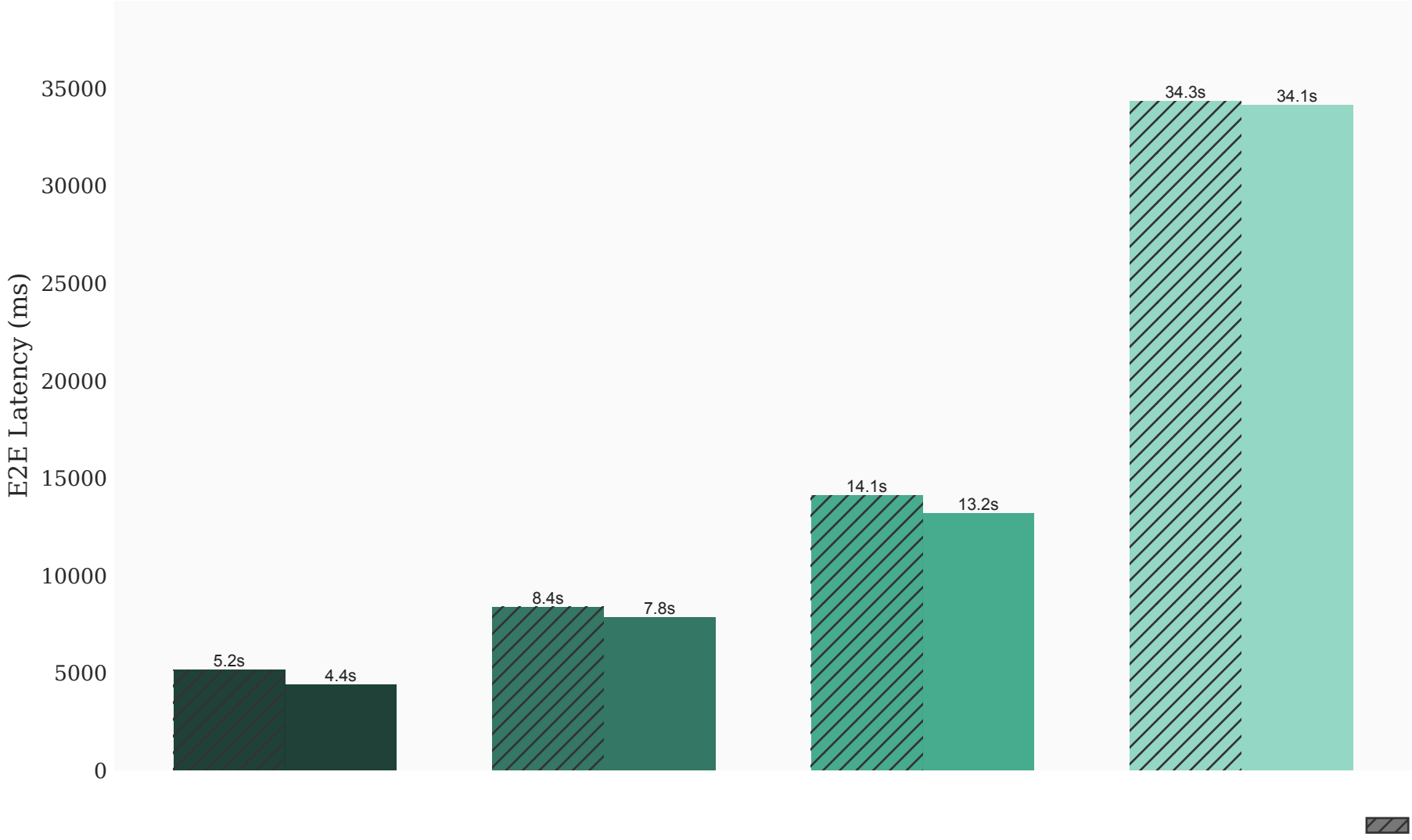
End-to-End Latency (P99)



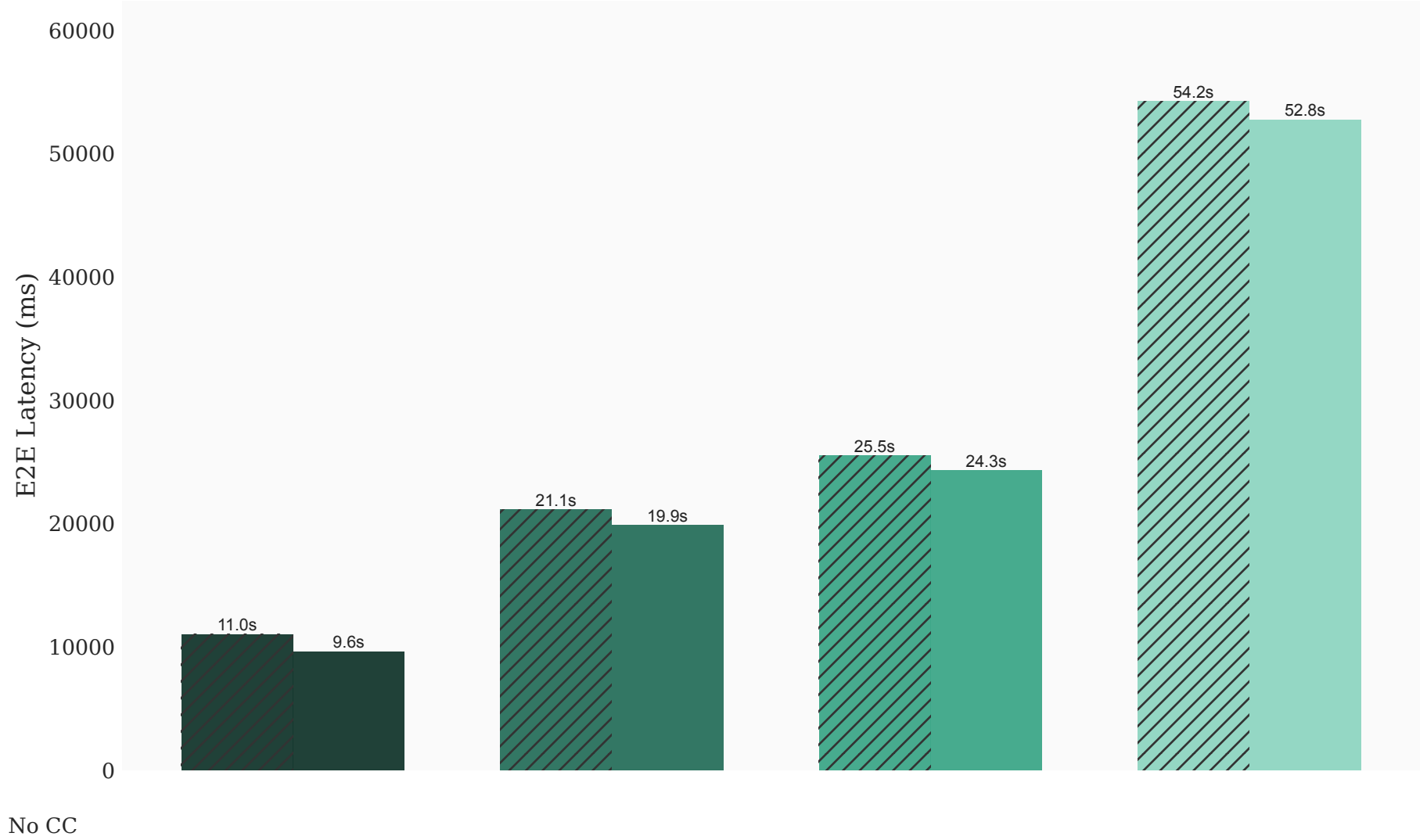
Llama 3.1 8B Mistral 3.1 24B GPT OSS 120B Llama 3.3 70B Int4

Numina Math (100 Concurrent Requests)

End-to-End Latency (Mean)



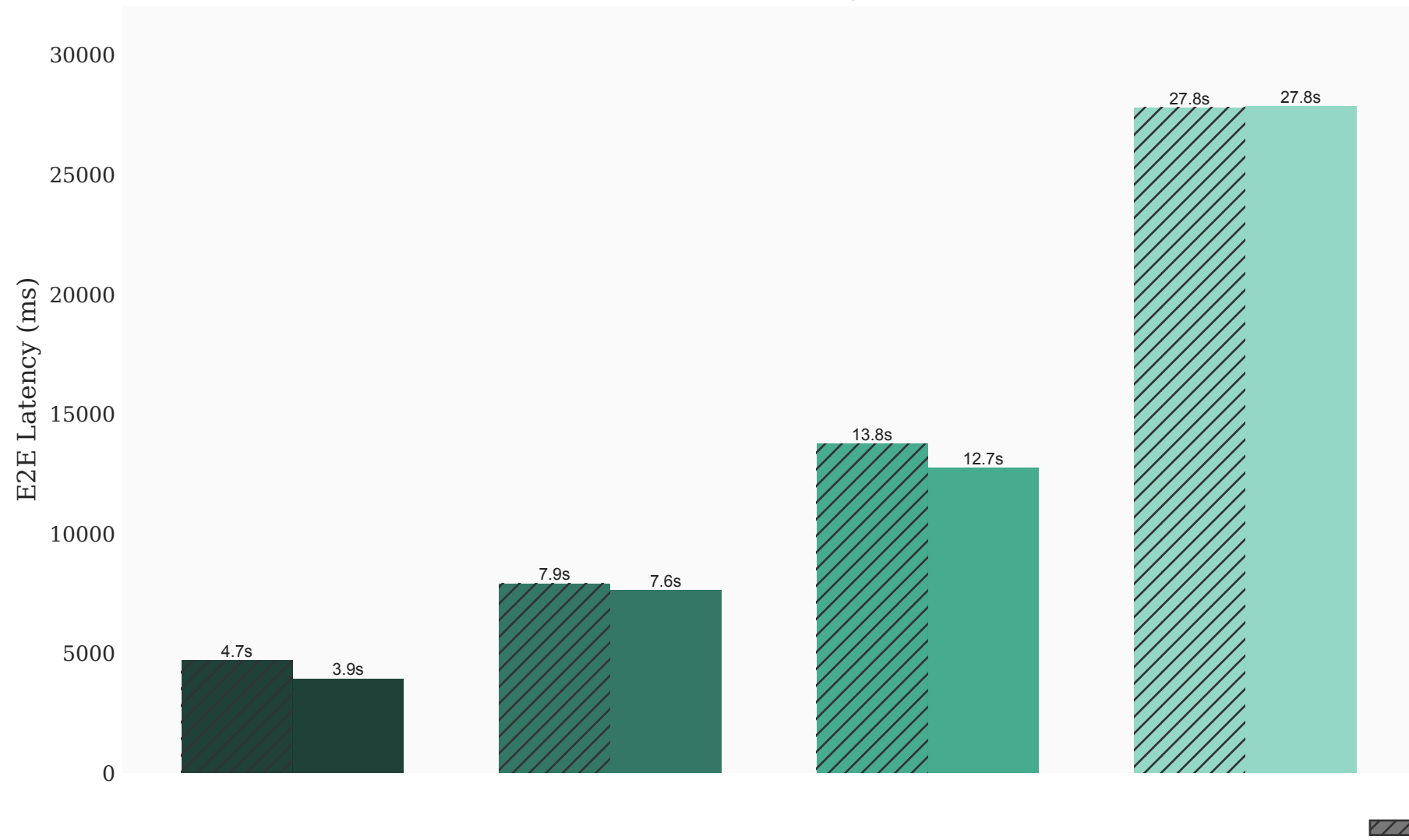
End-to-End Latency (P99)



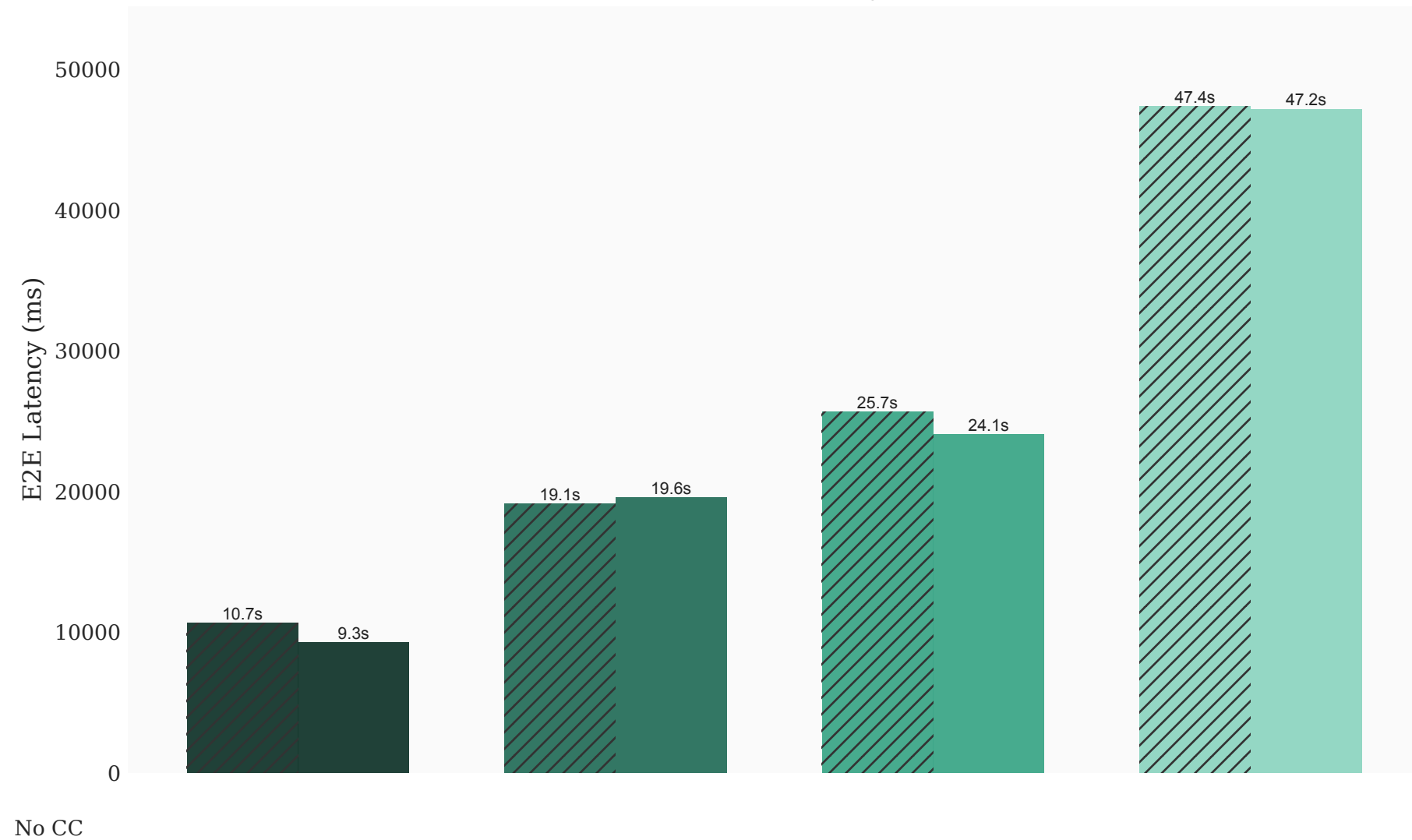
LLama 3.1 8B Mistral 3.1 24B GPT OSS 120B LLama 3.3 70B Int4

Numina Math (50 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

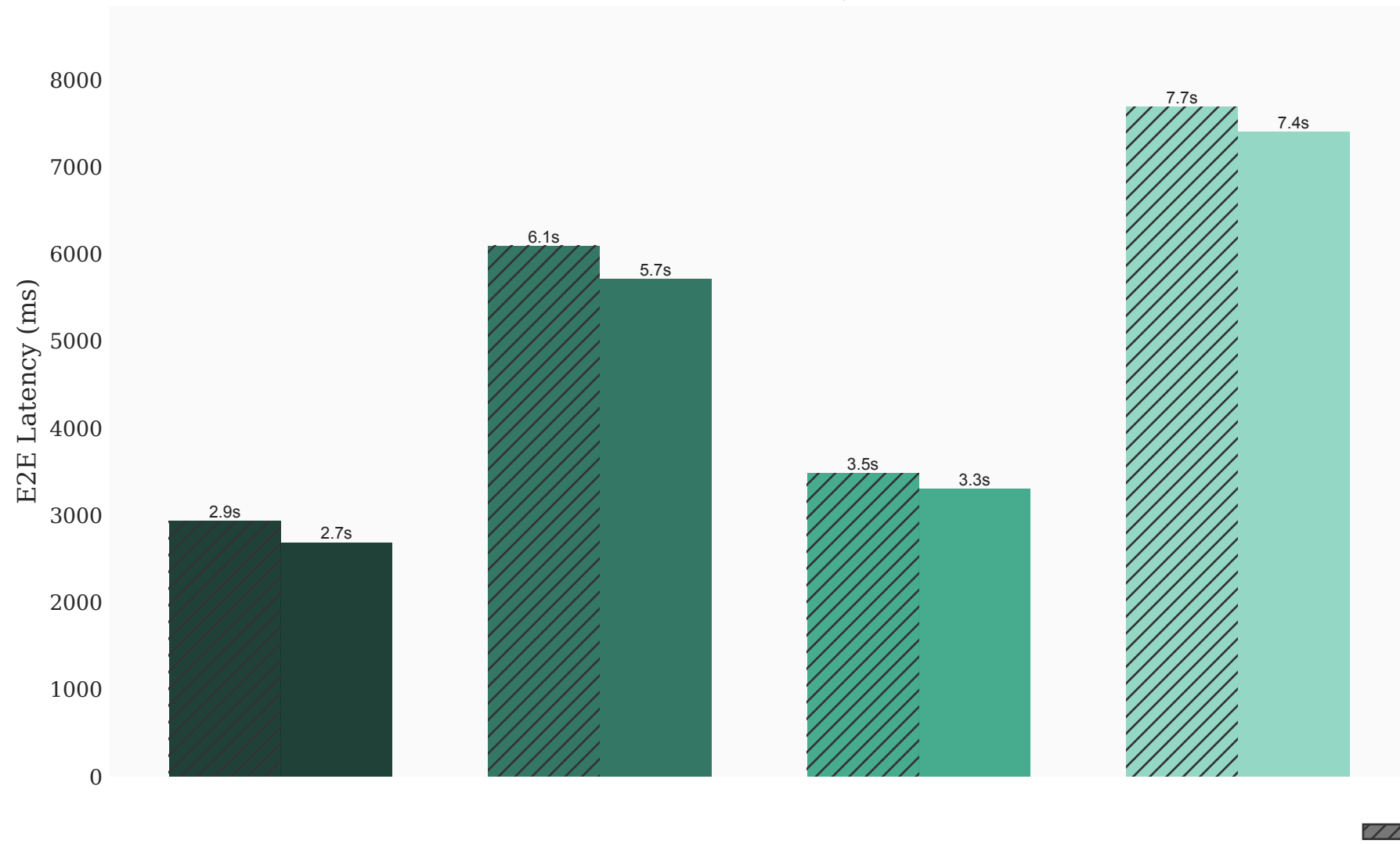


■ CC ■ No CC

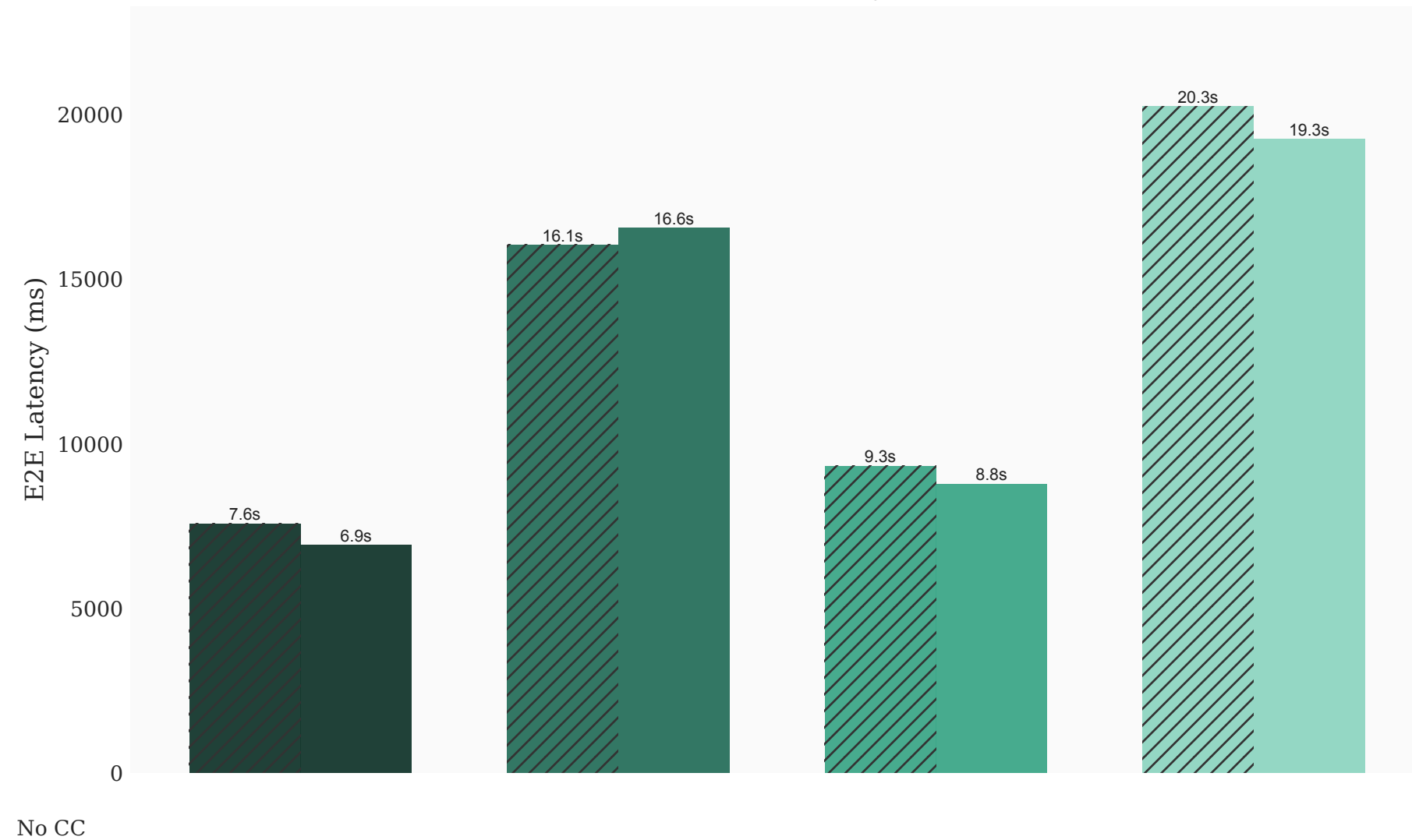
■ LLama 3.1 8B ■ Mistral 3.1 24B ■ GPT OSS 120B ■ LLama 3.3 70B Int4

Numina Math (1 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