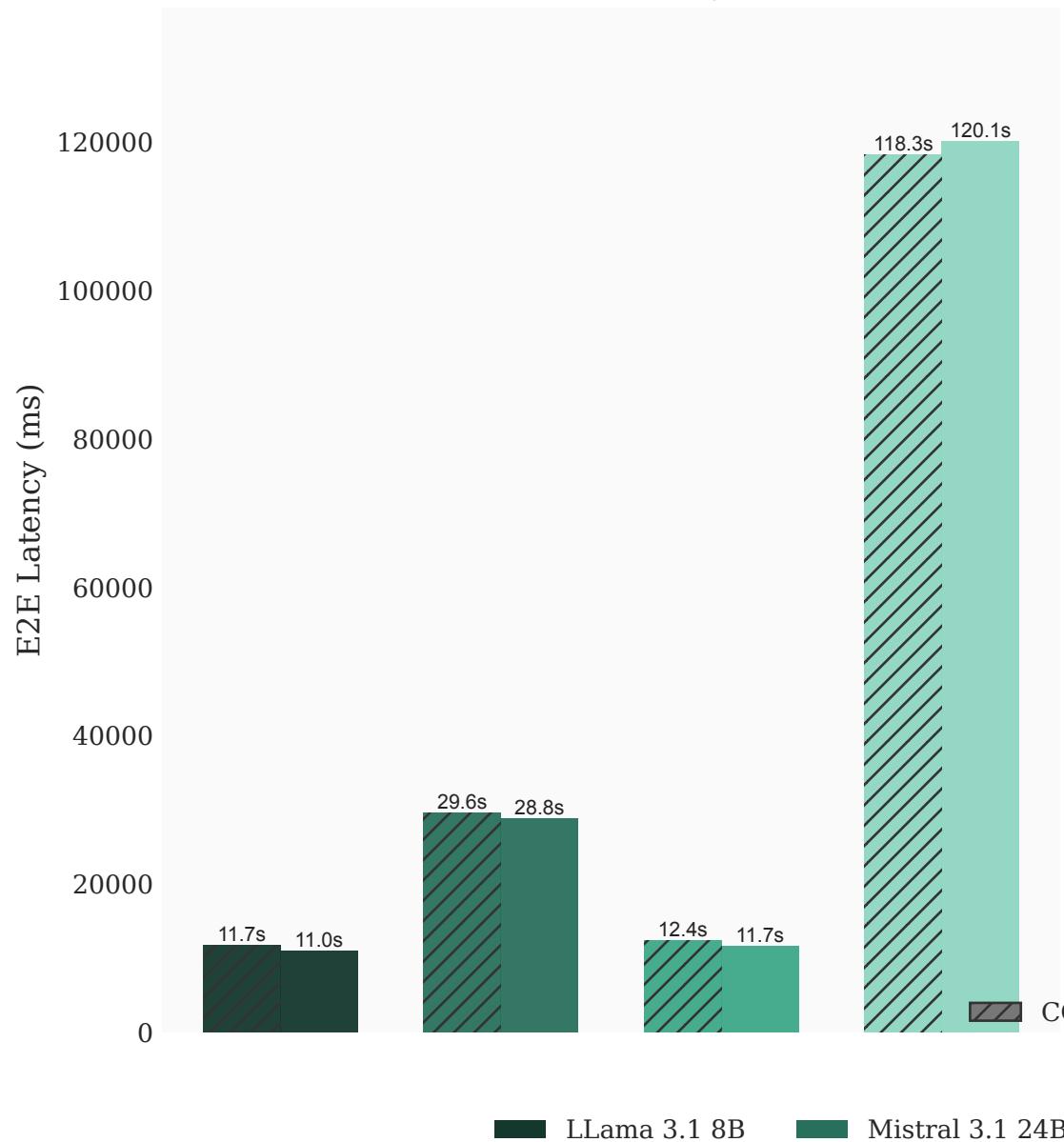
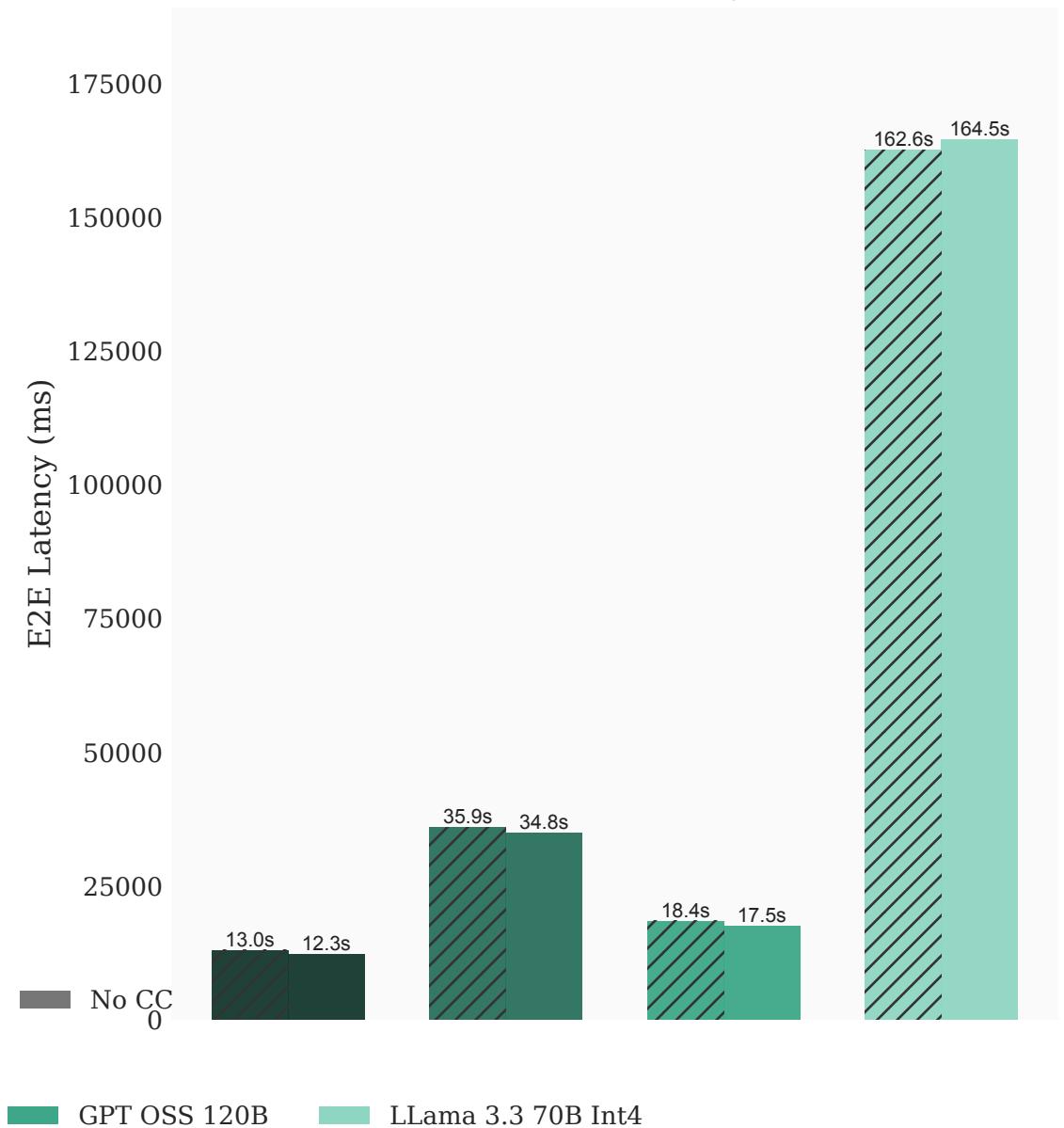


Random (1500 \Rightarrow 250) (100 Concurrent Requests)

End-to-End Latency (Mean)

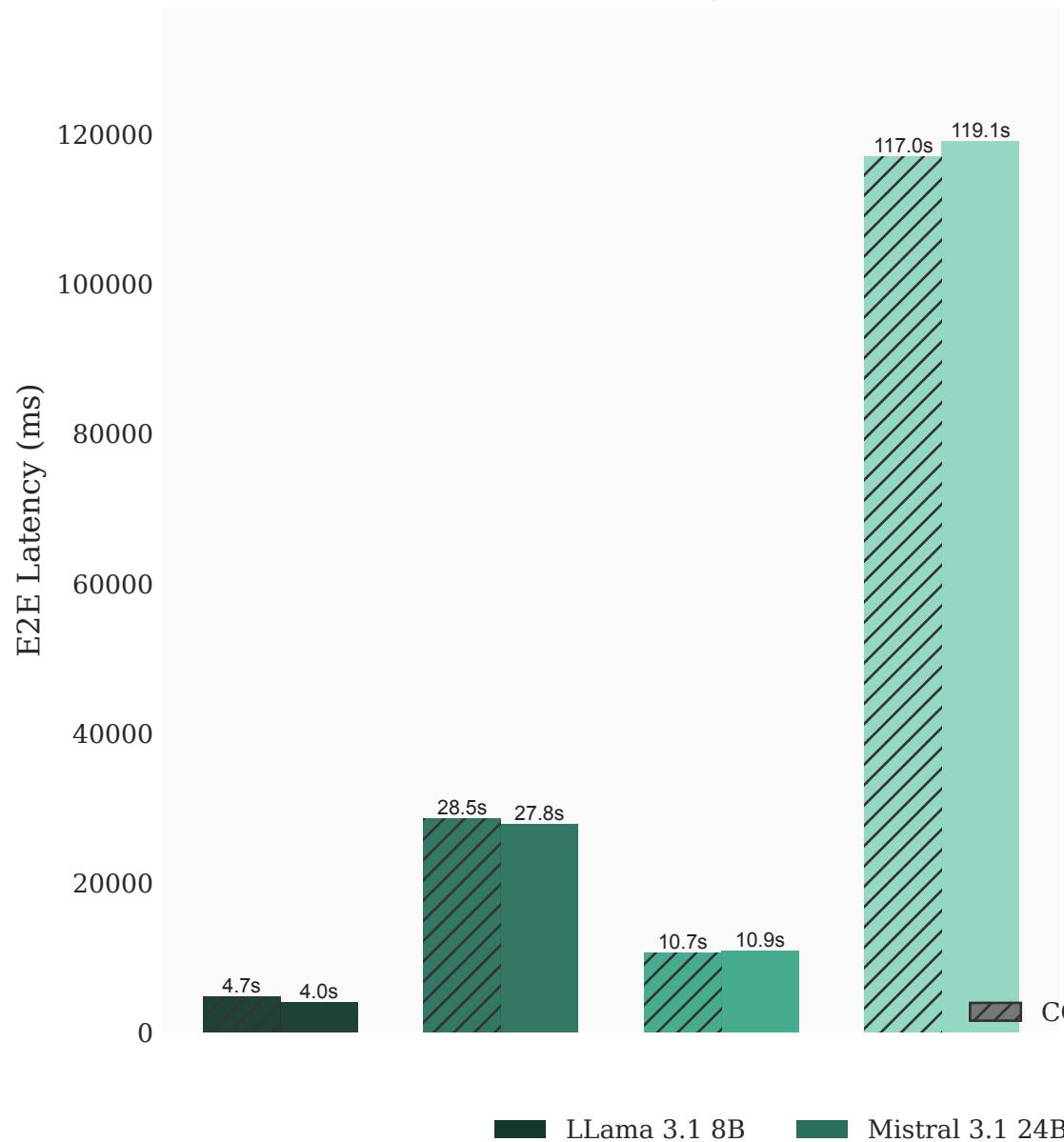


End-to-End Latency (P99)

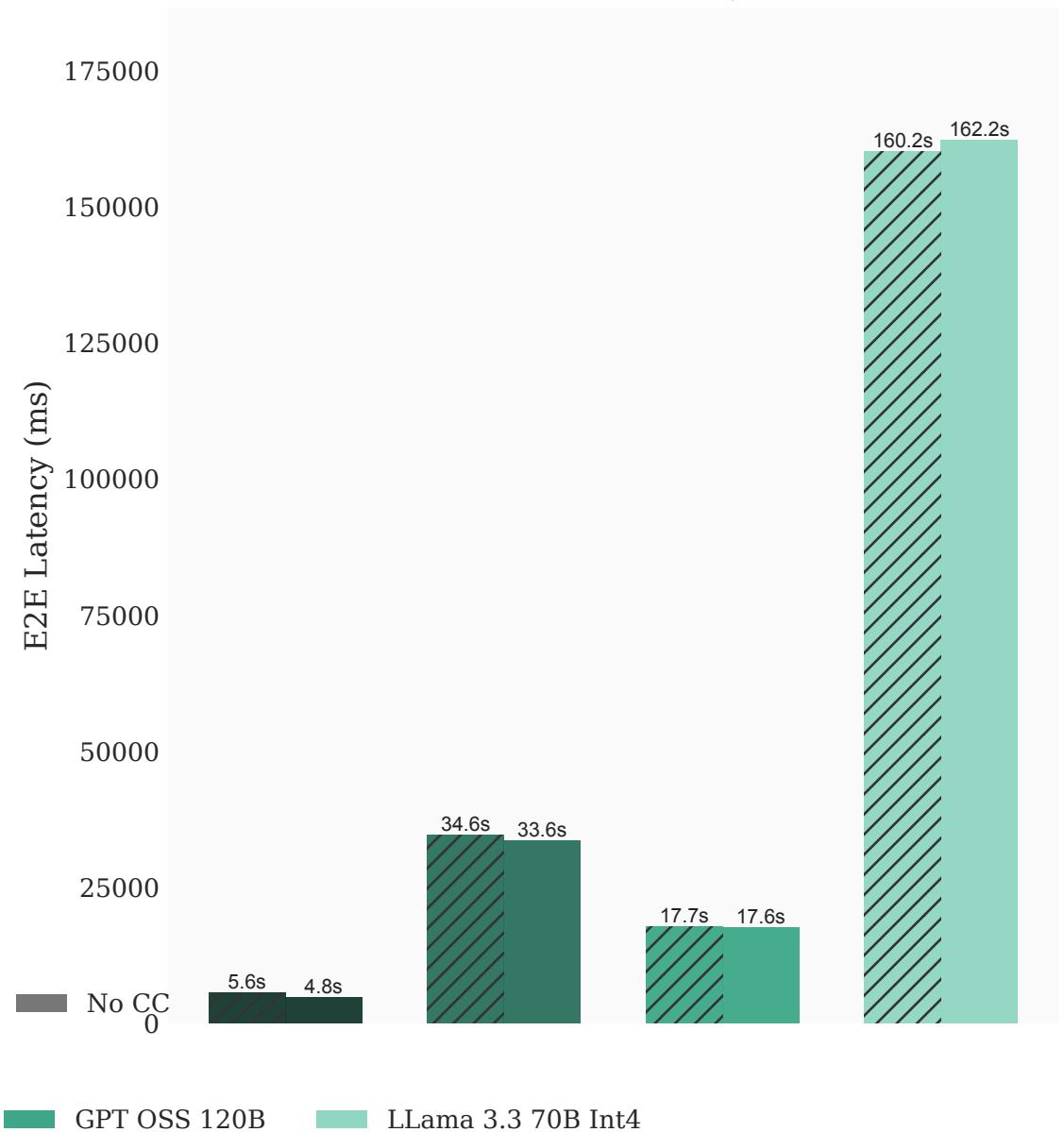


Random (1500 \Rightarrow 250) (50 Concurrent Requests)

End-to-End Latency (Mean)

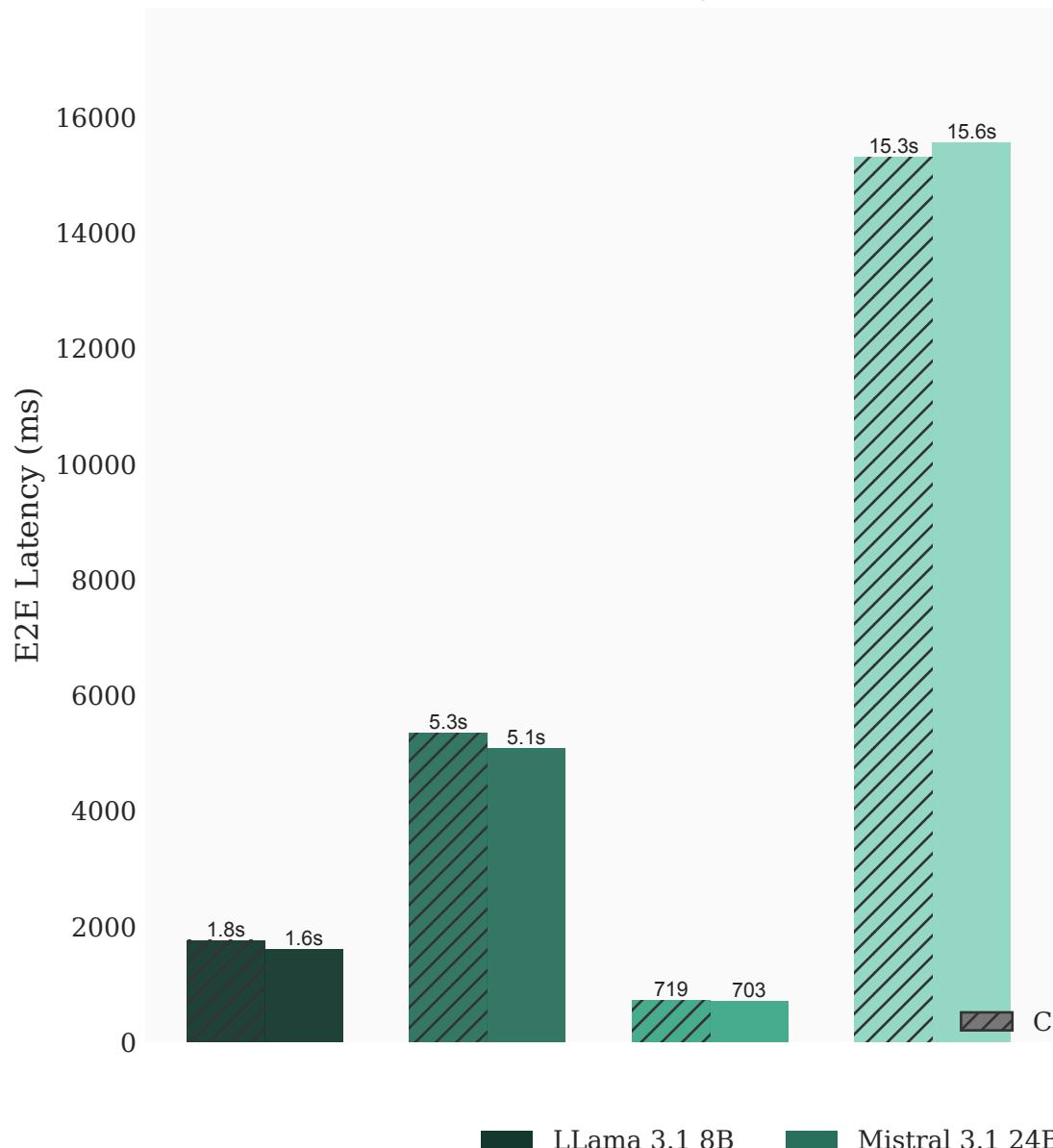


End-to-End Latency (P99)

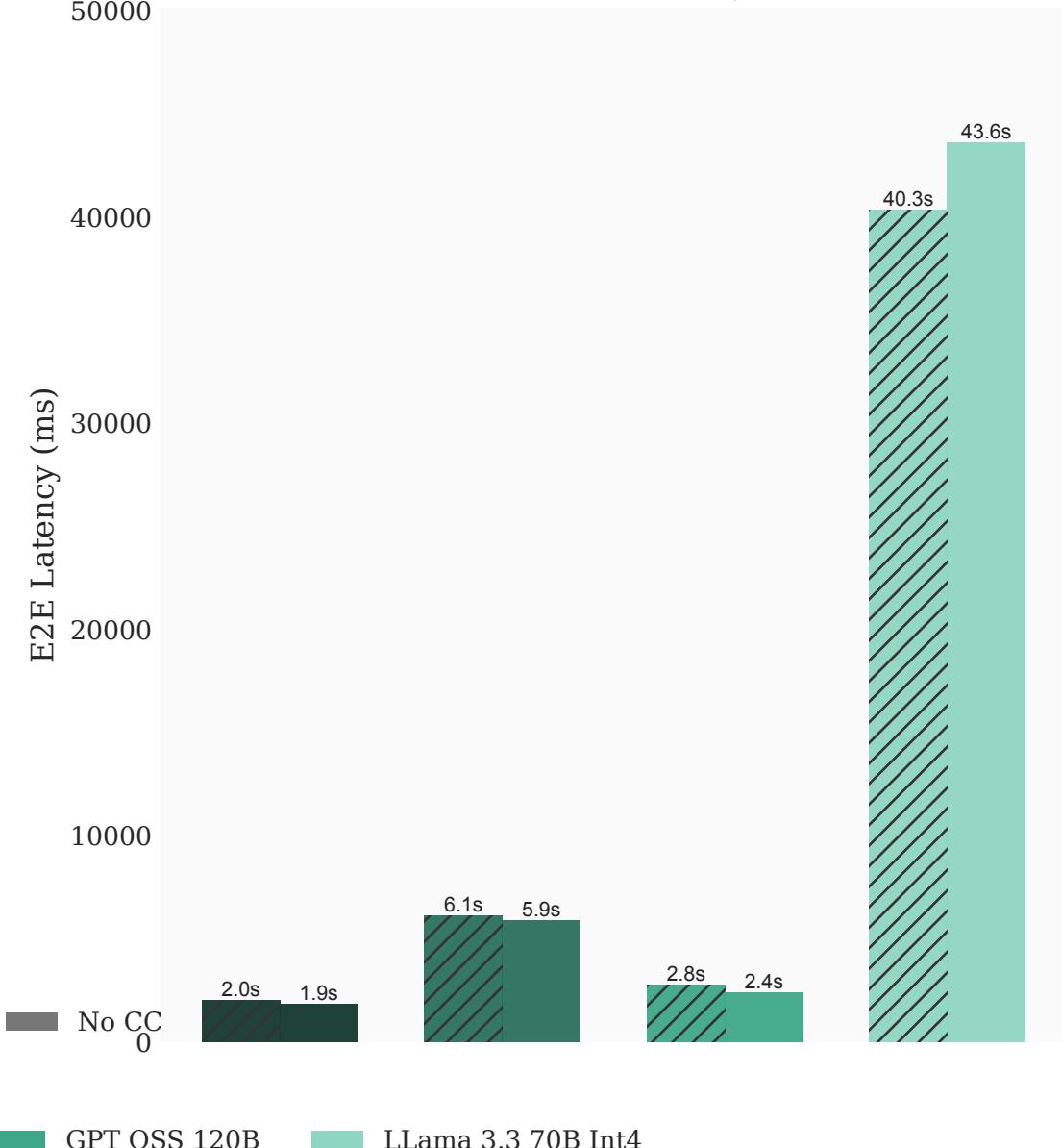


Random (1500 \Rightarrow 250) (1 Concurrent Requests)

End-to-End Latency (Mean)

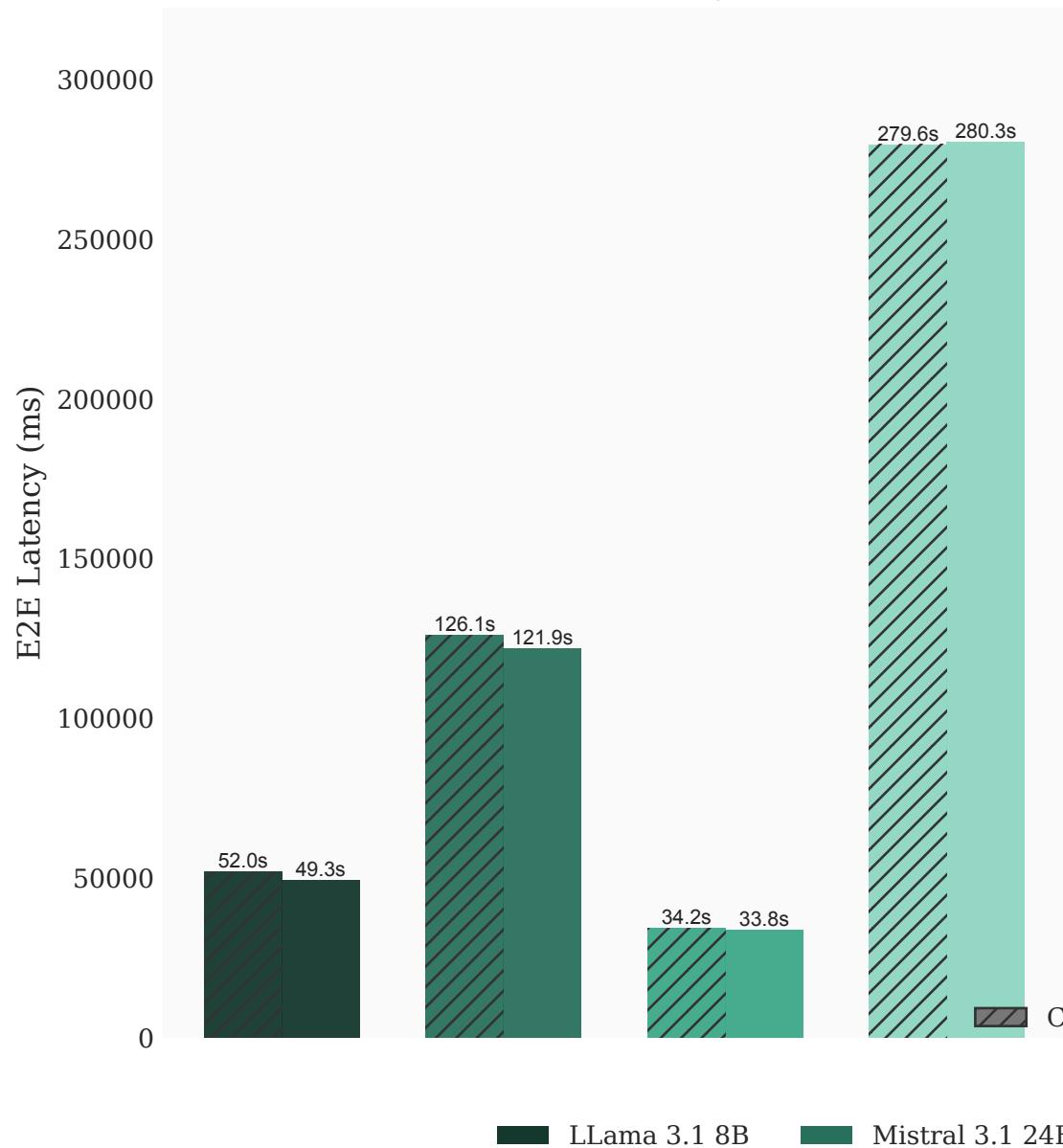


End-to-End Latency (P99)

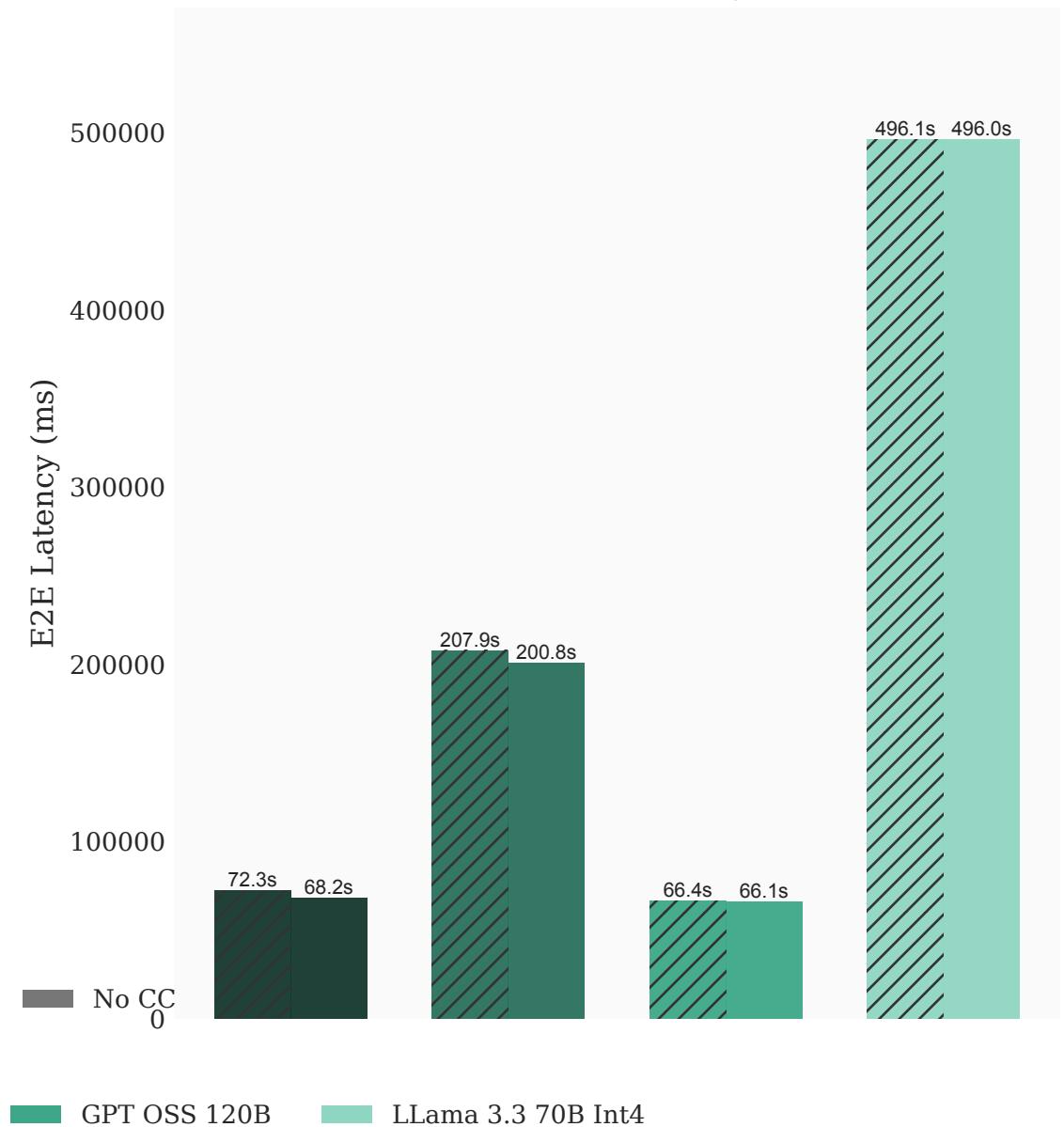


Random (4000 \Rightarrow 1000) (100 Concurrent Requests)

End-to-End Latency (Mean)

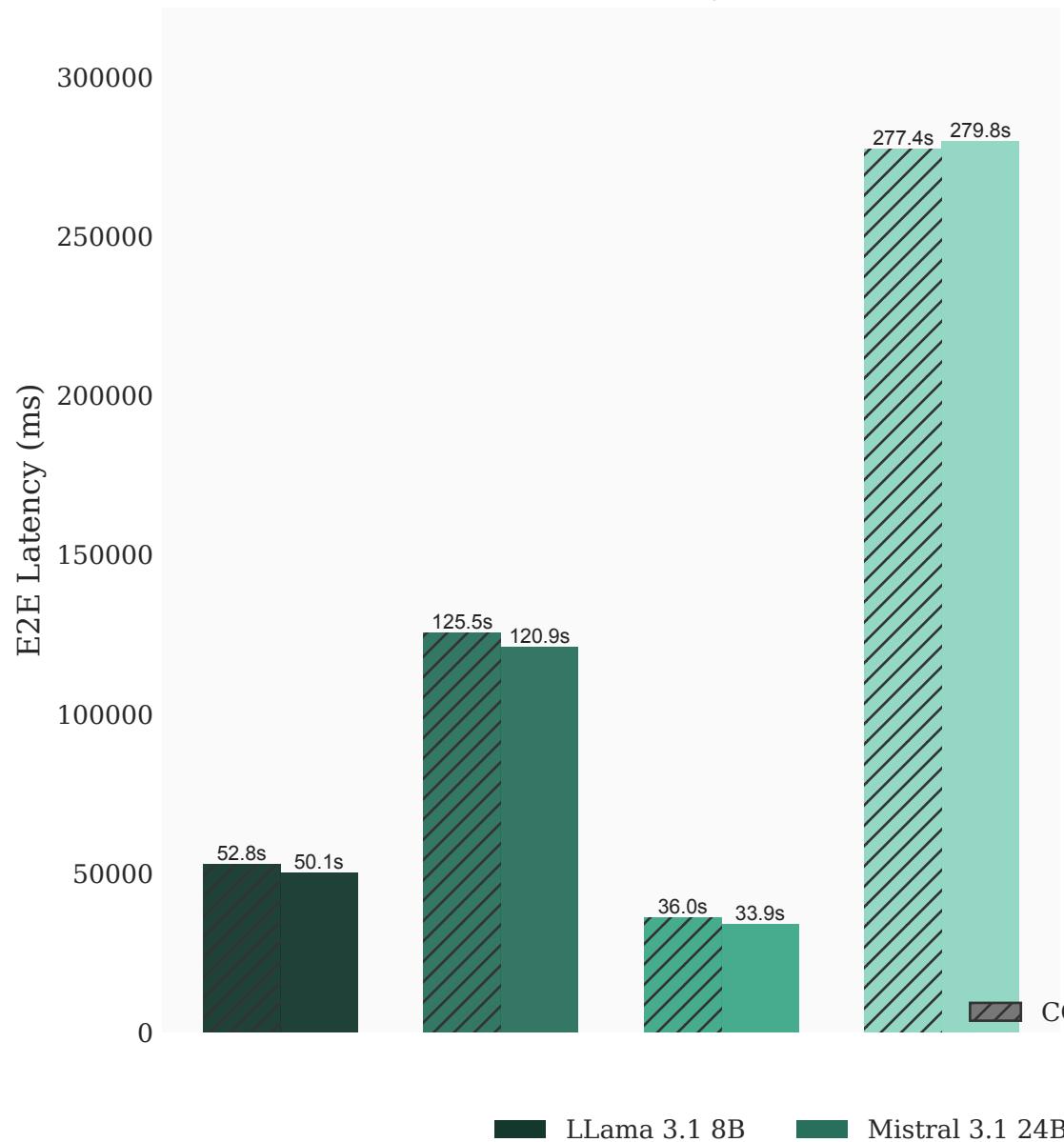


End-to-End Latency (P99)

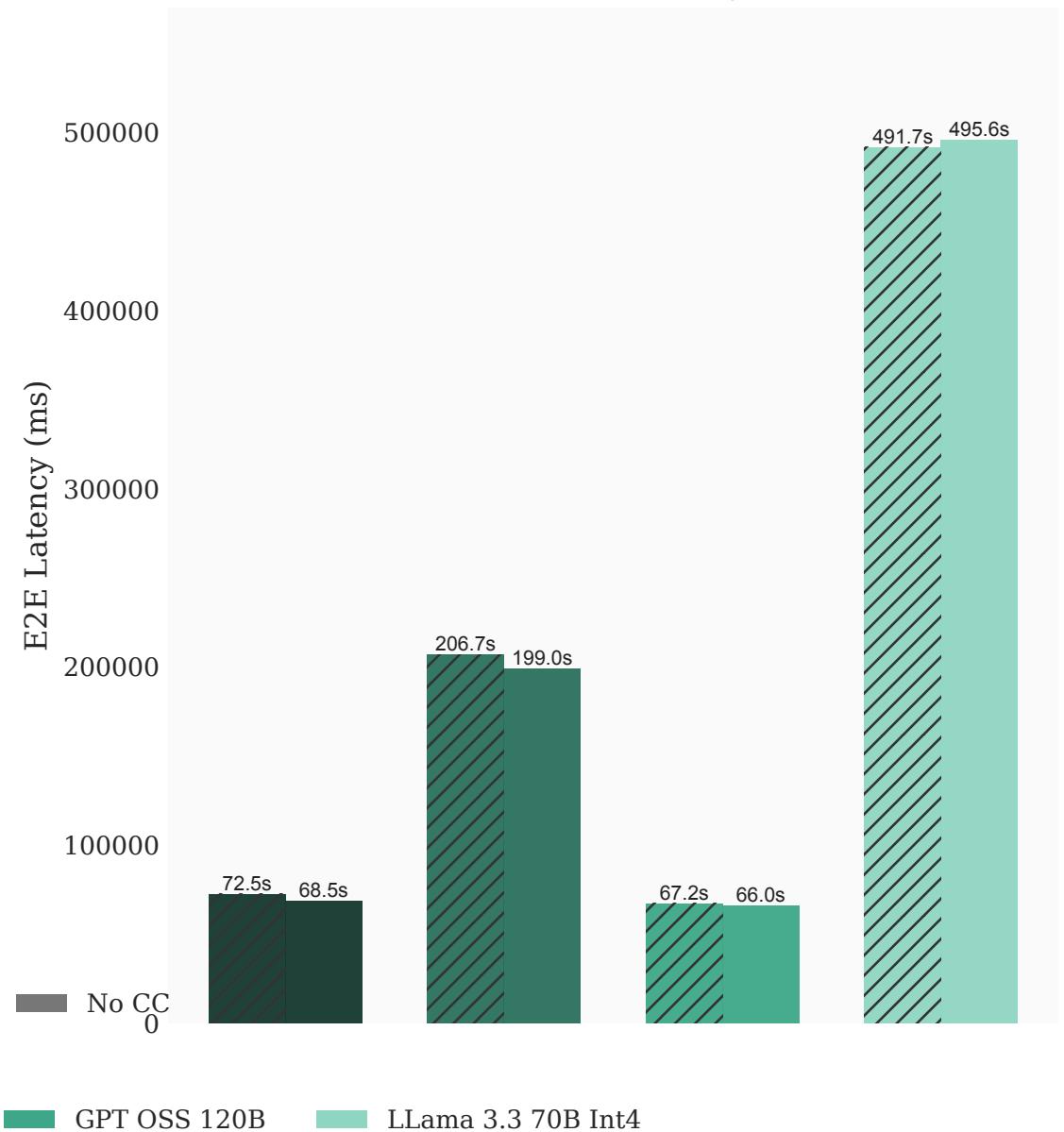


Random (4000 \Rightarrow 1000) (50 Concurrent Requests)

End-to-End Latency (Mean)

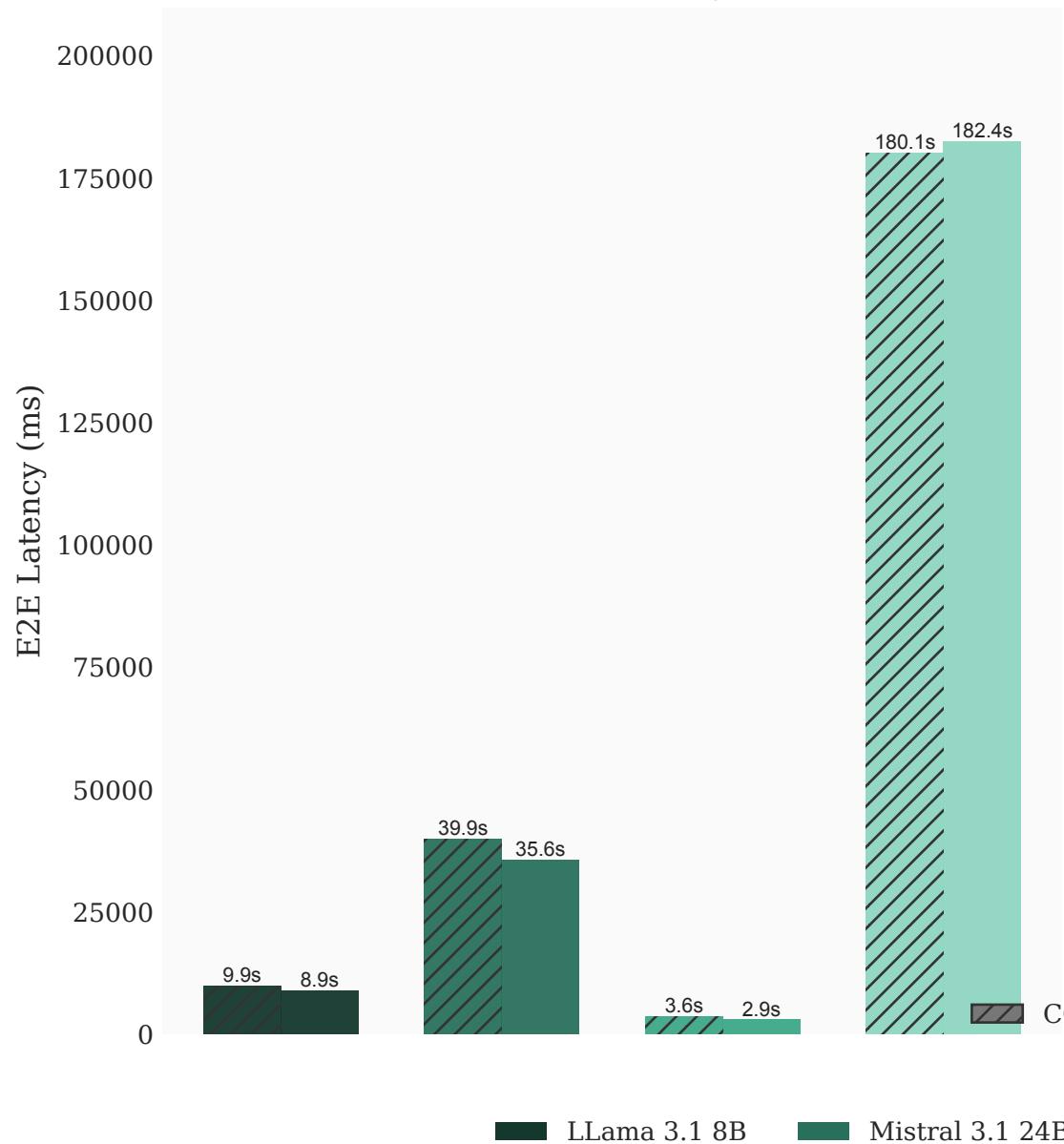


End-to-End Latency (P99)

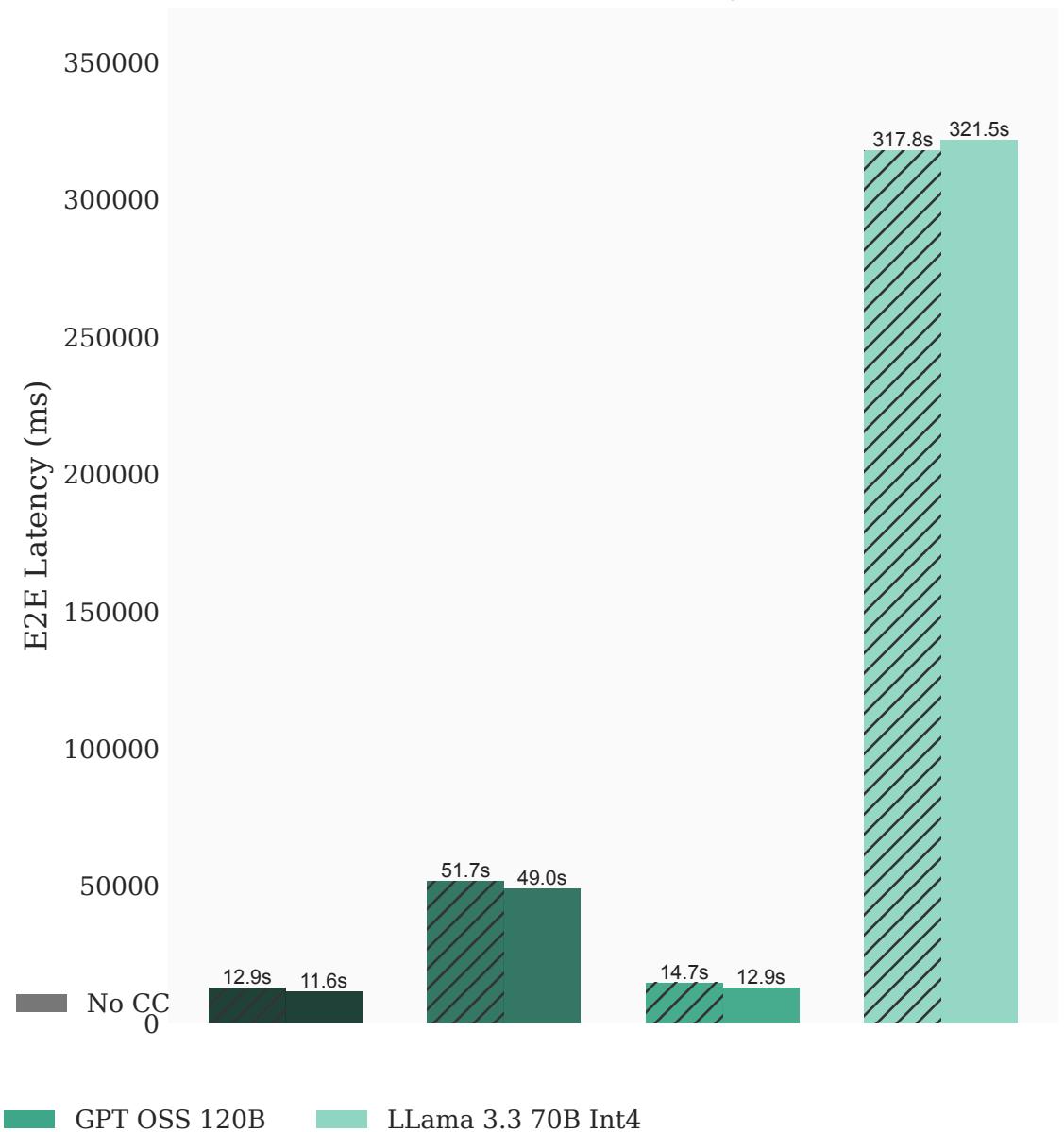


Random (4000 \Rightarrow 1000) (1 Concurrent Requests)

End-to-End Latency (Mean)

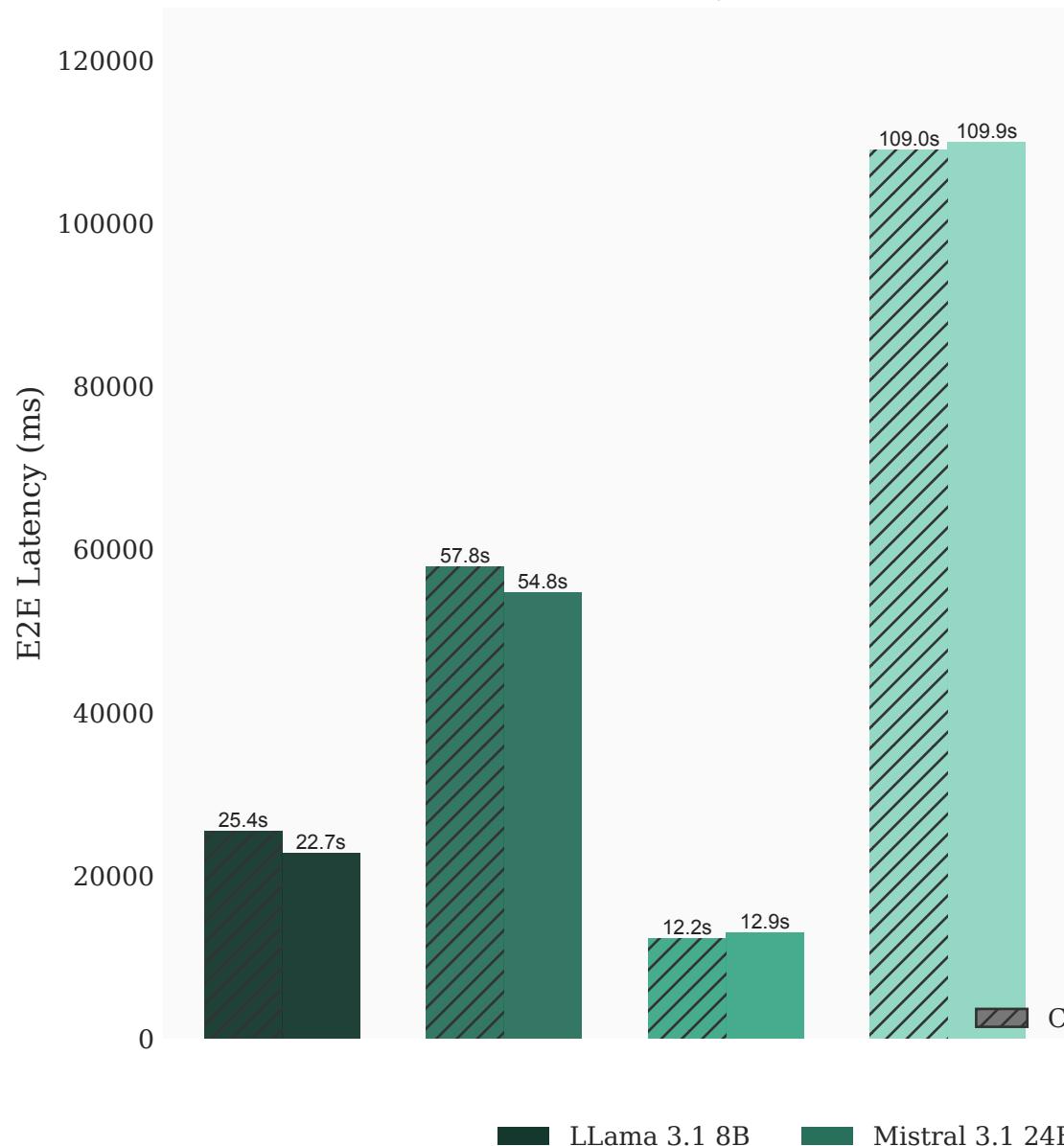


End-to-End Latency (P99)

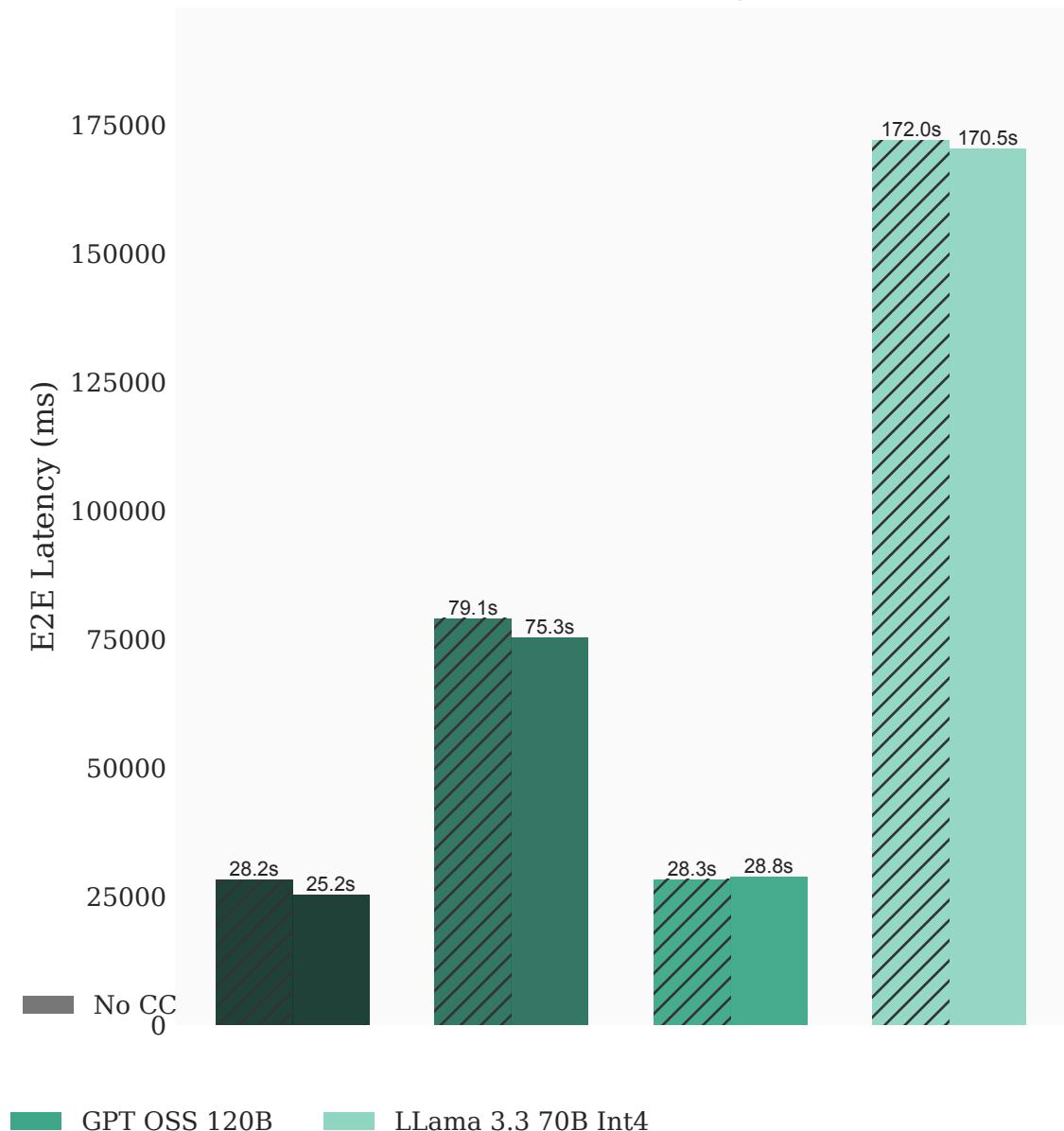


Random (1000 ⇒ 1000) (100 Concurrent Requests)

End-to-End Latency (Mean)

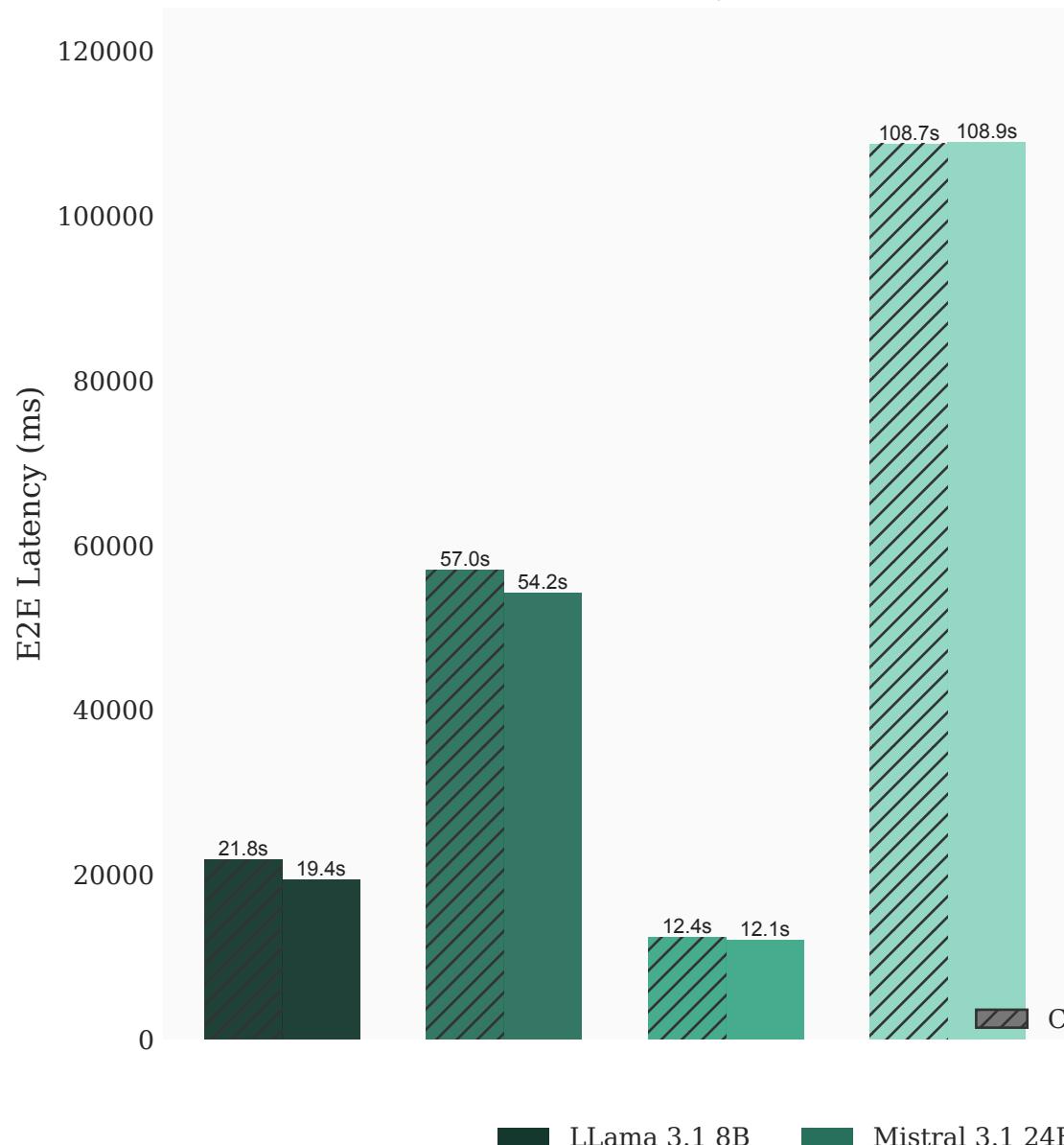


End-to-End Latency (P99)

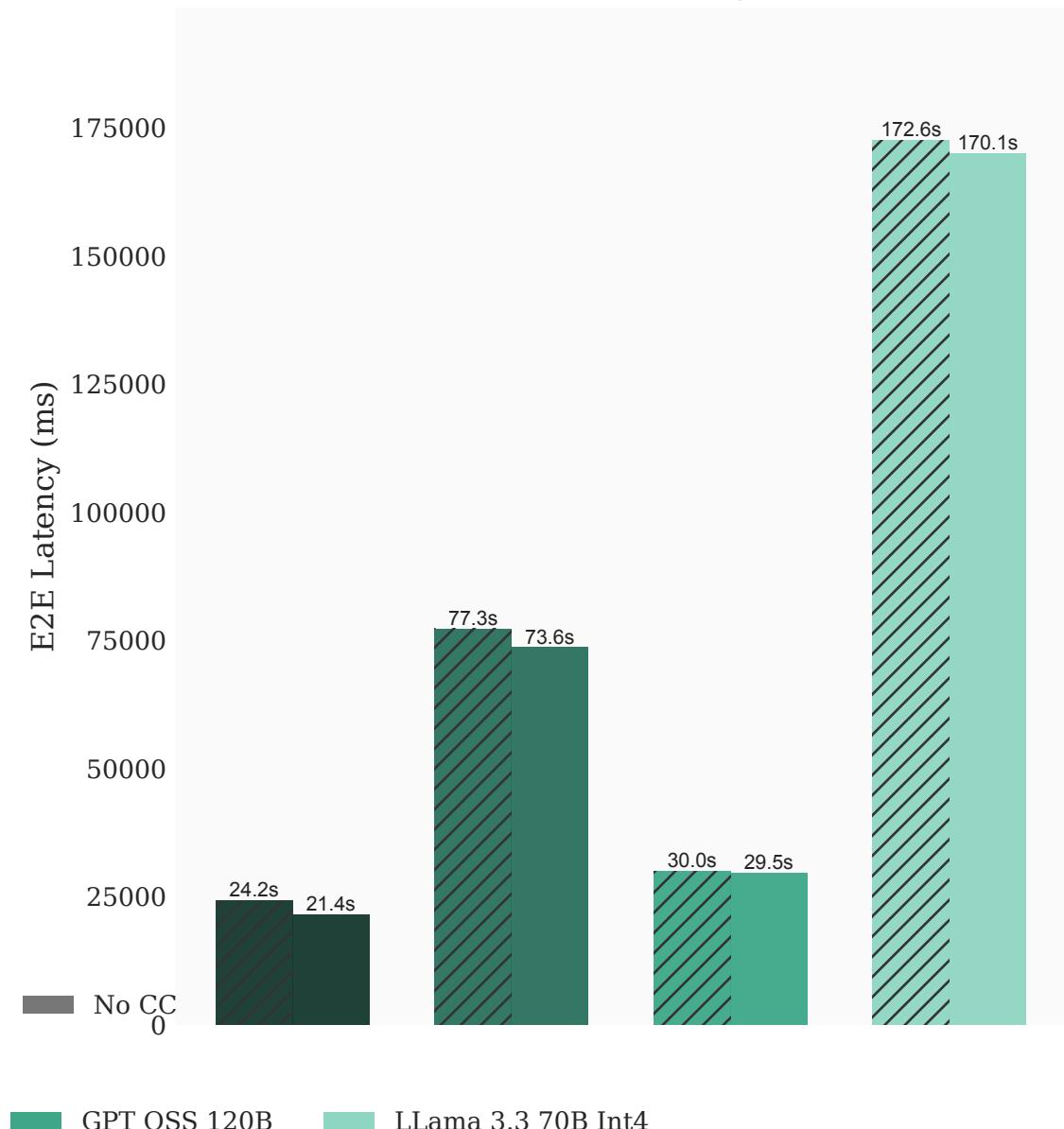


Random (1000 \Rightarrow 1000) (50 Concurrent Requests)

End-to-End Latency (Mean)

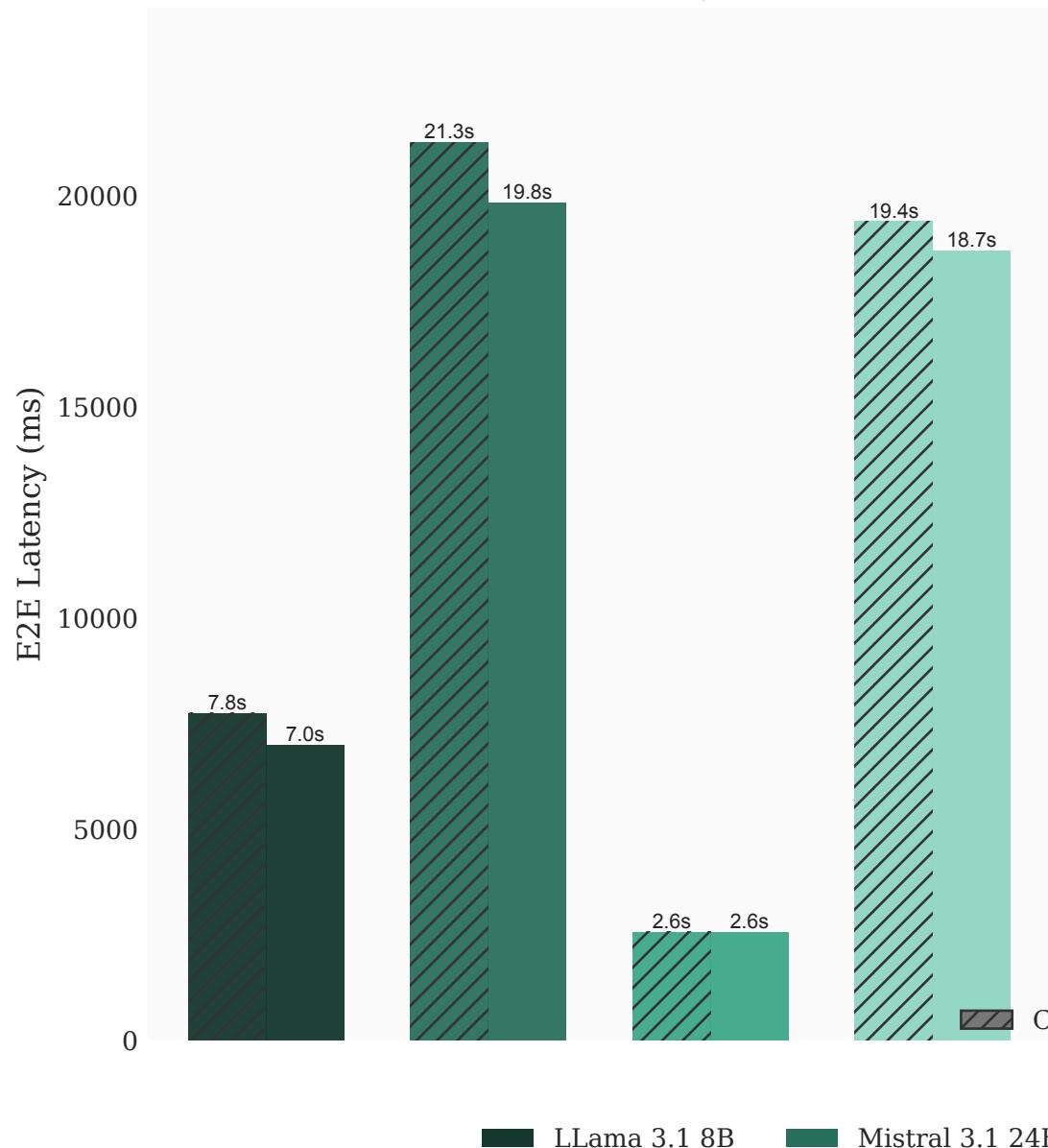


End-to-End Latency (P99)

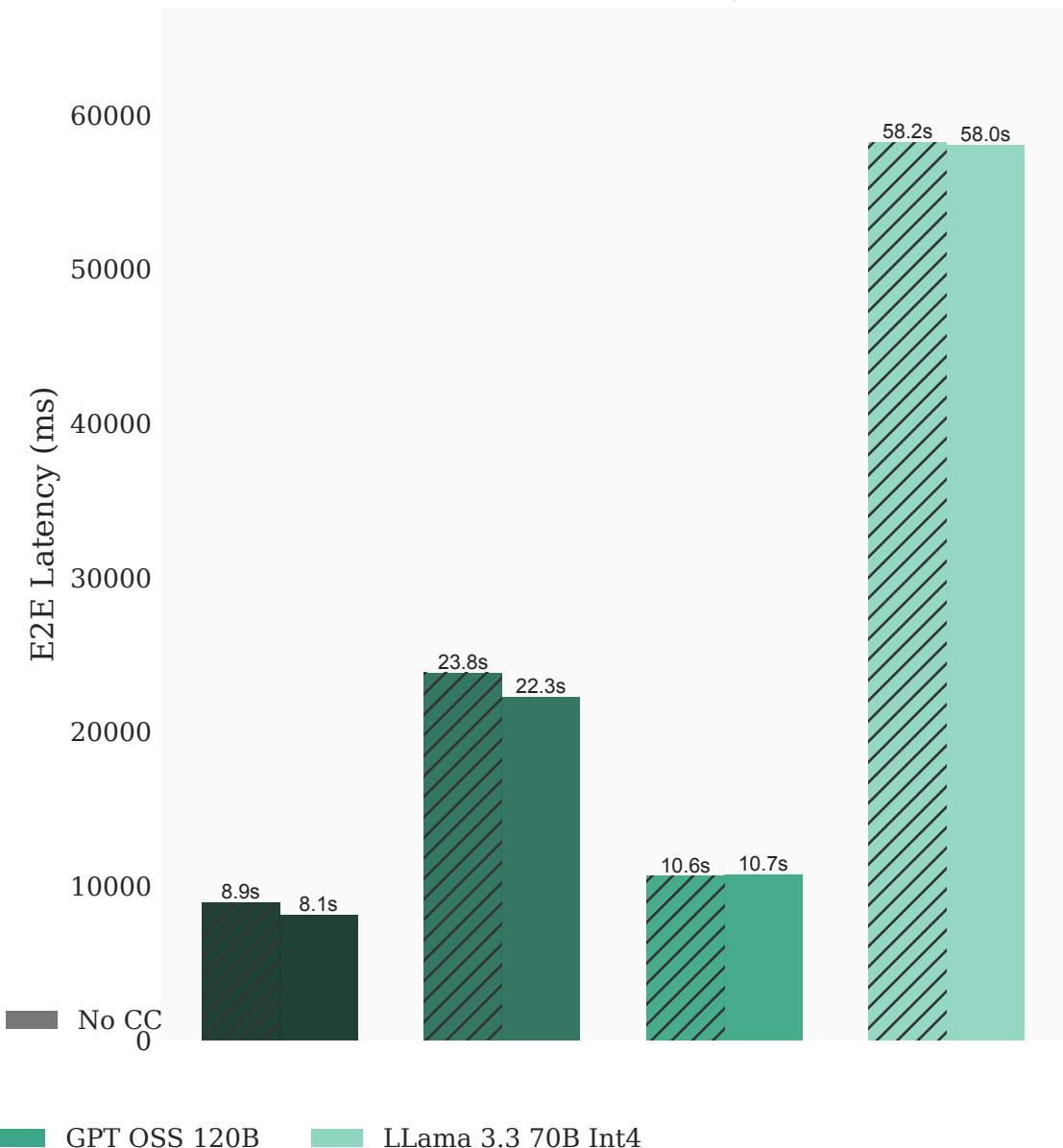


Random (1000 \Rightarrow 1000) (1 Concurrent Requests)

End-to-End Latency (Mean)

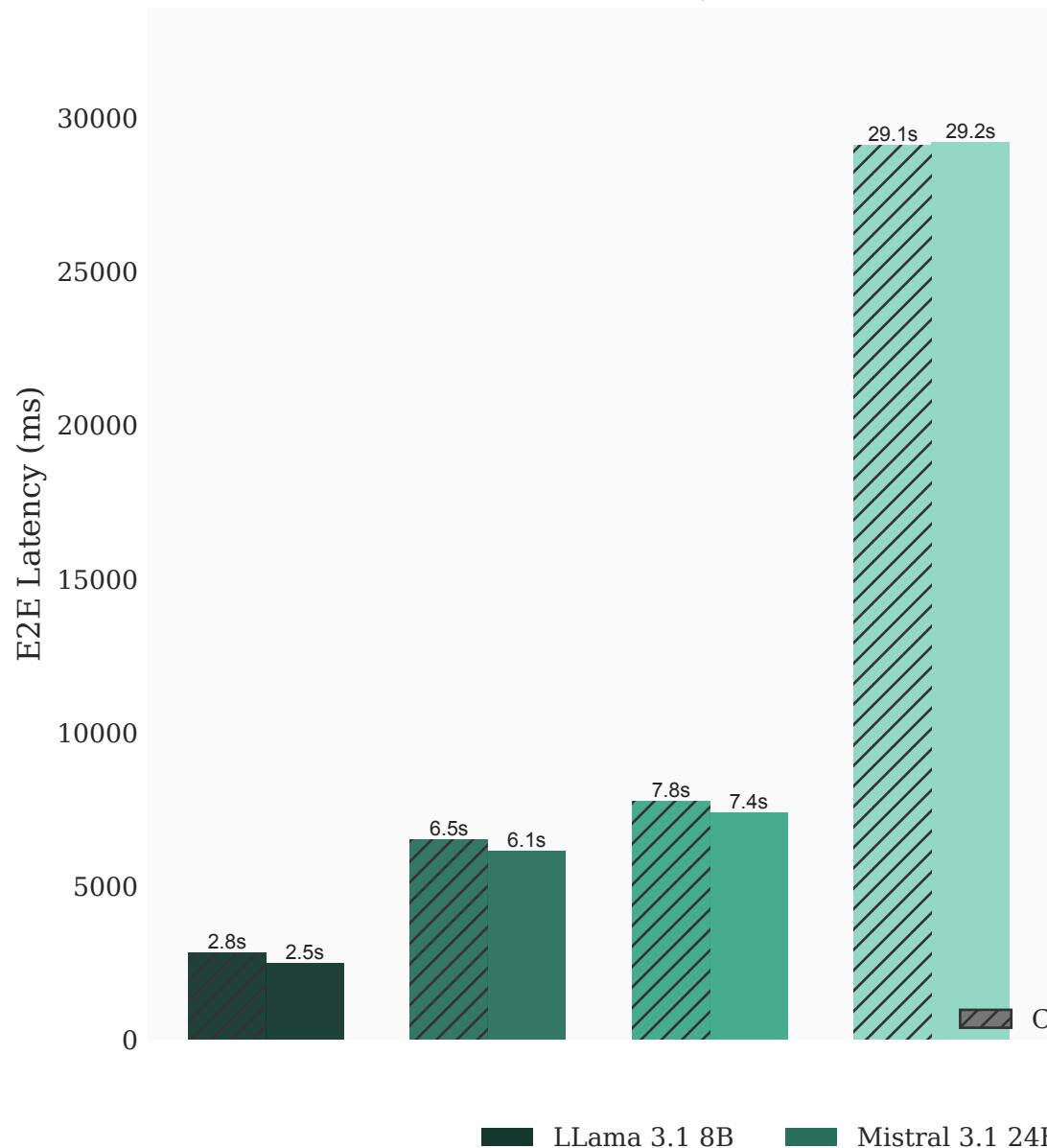


End-to-End Latency (P99)

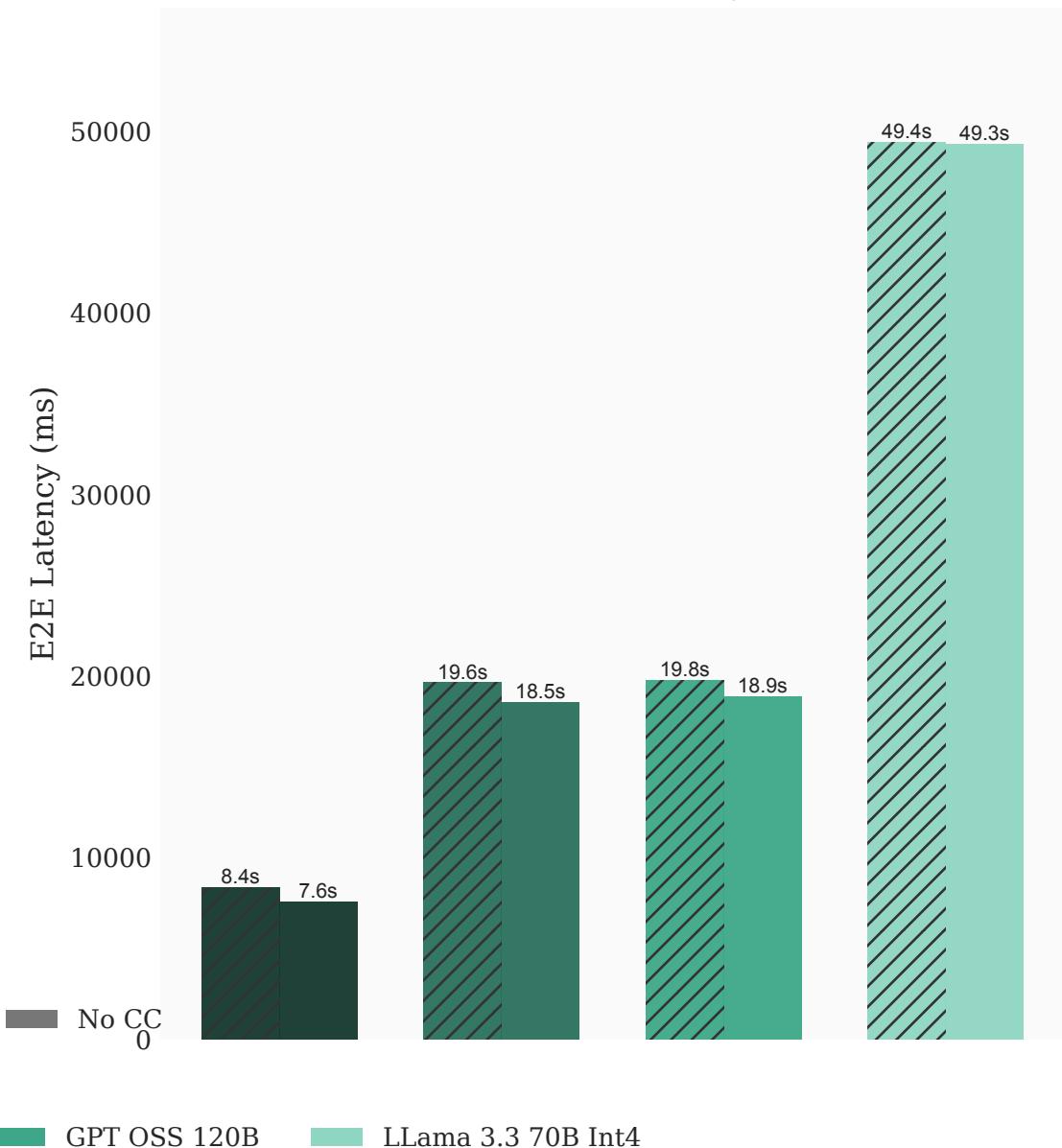


ShareGPT (100 Concurrent Requests)

End-to-End Latency (Mean)

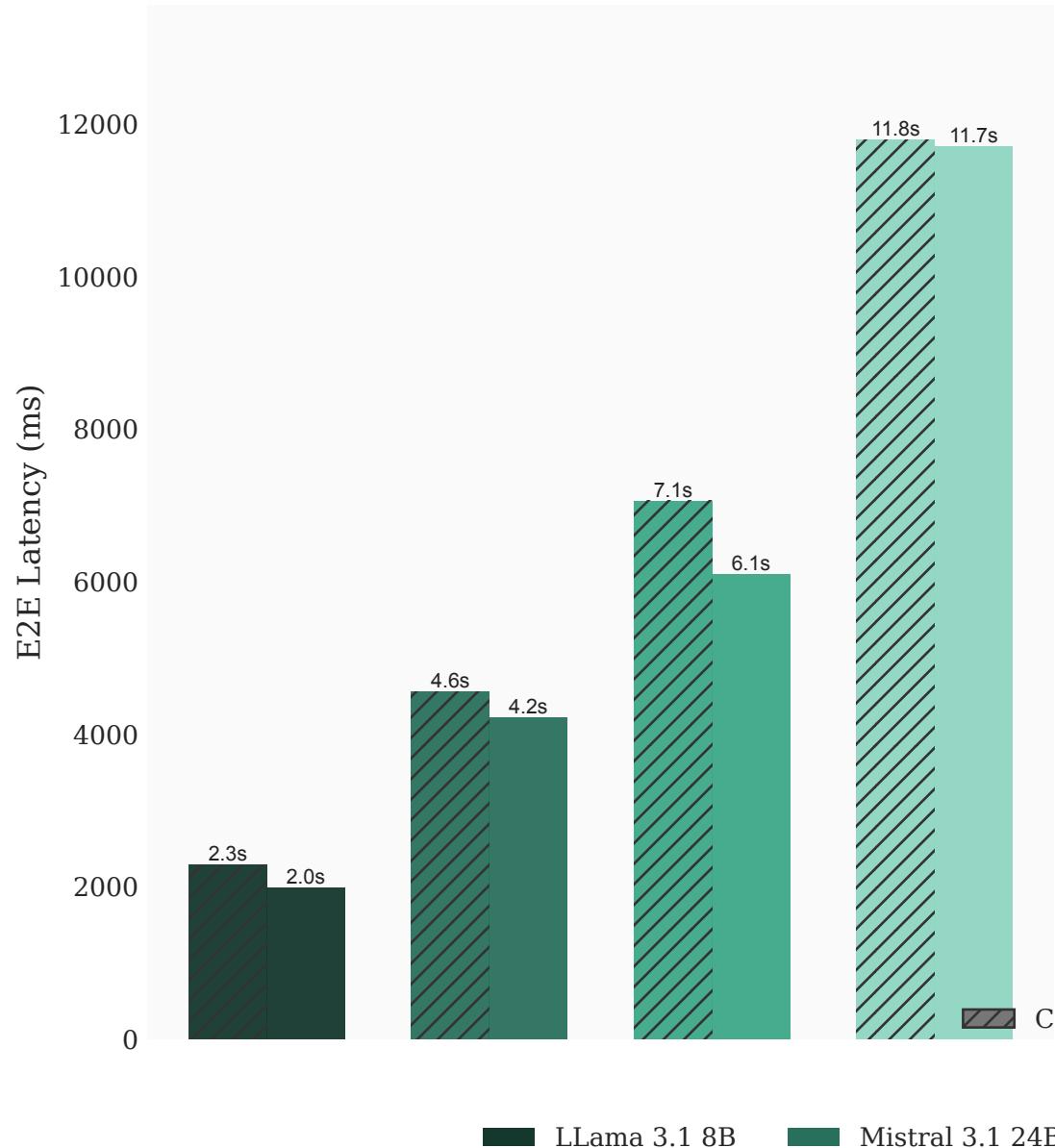


End-to-End Latency (P99)

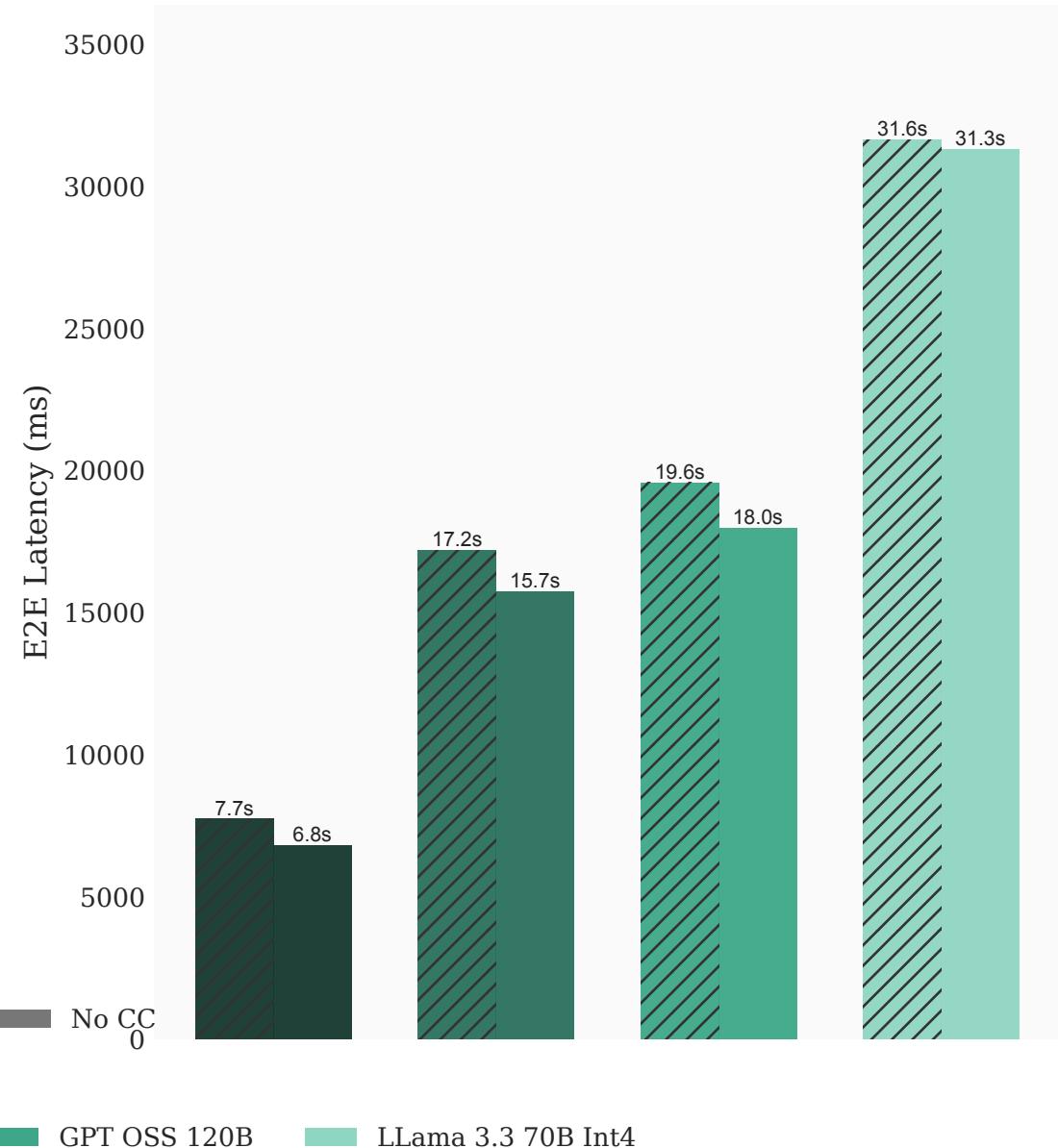


ShareGPT (50 Concurrent Requests)

End-to-End Latency (Mean)

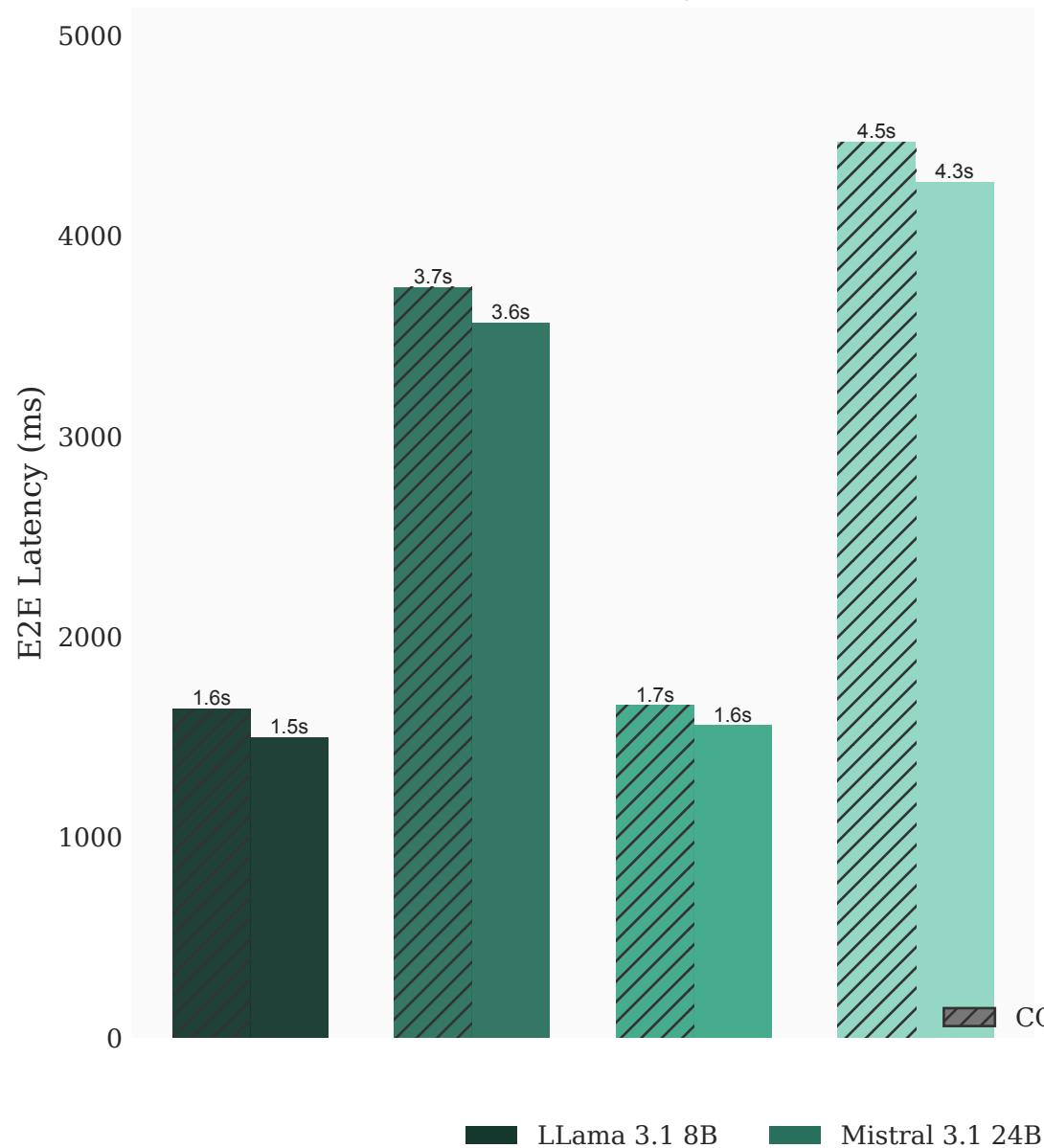


End-to-End Latency (P99)

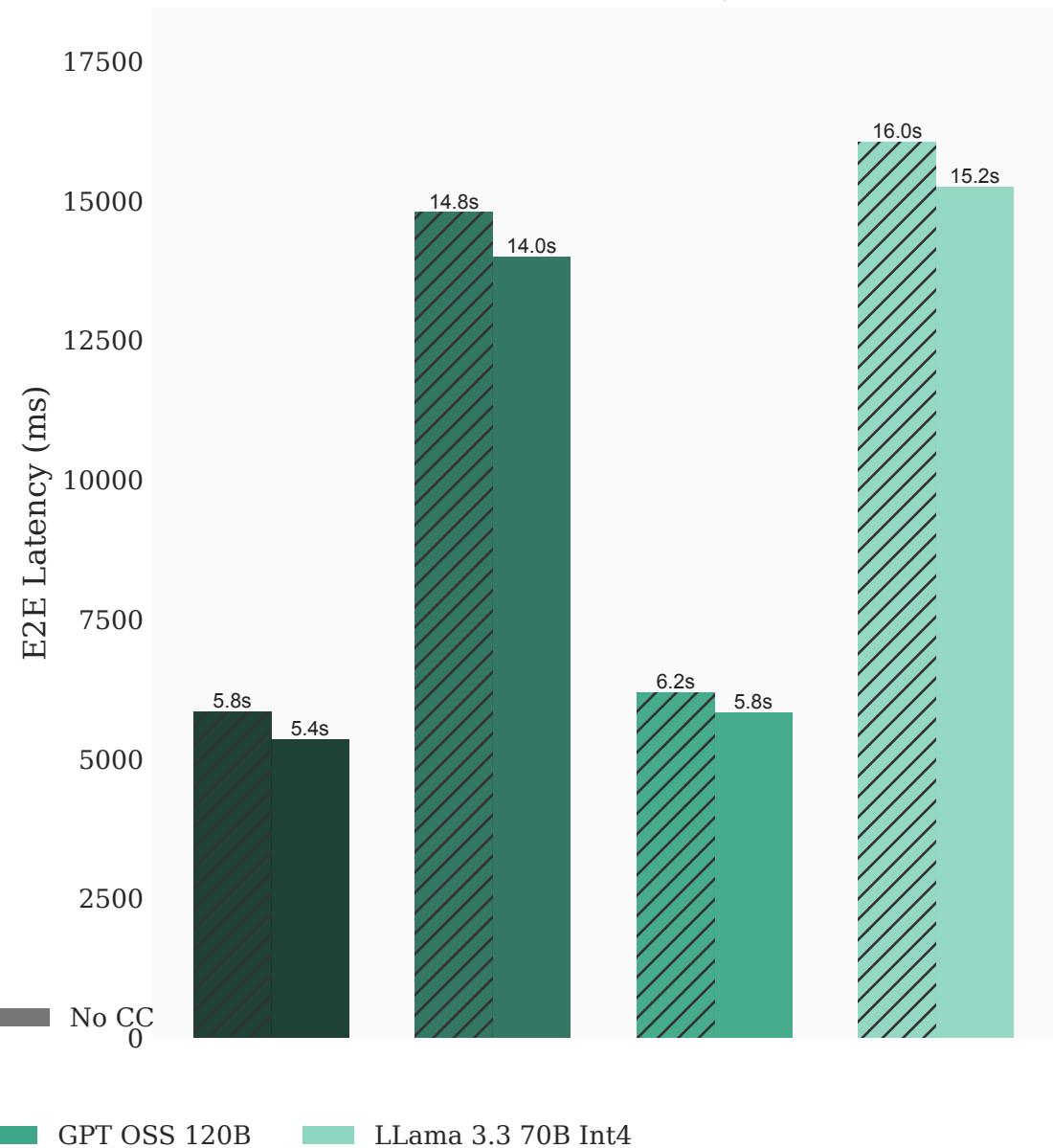


ShareGPT (1 Concurrent Requests)

End-to-End Latency (Mean)

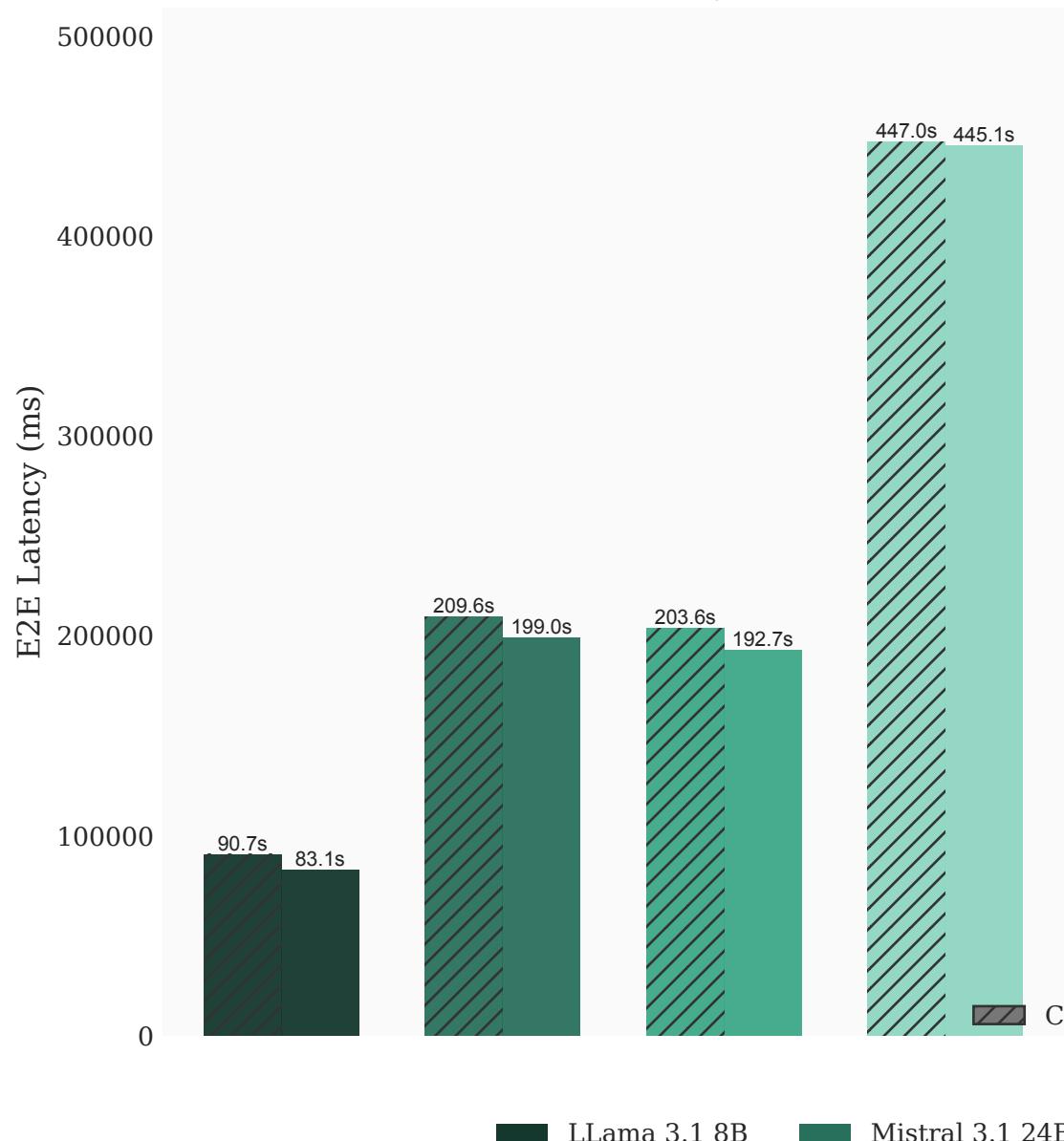


End-to-End Latency (P99)

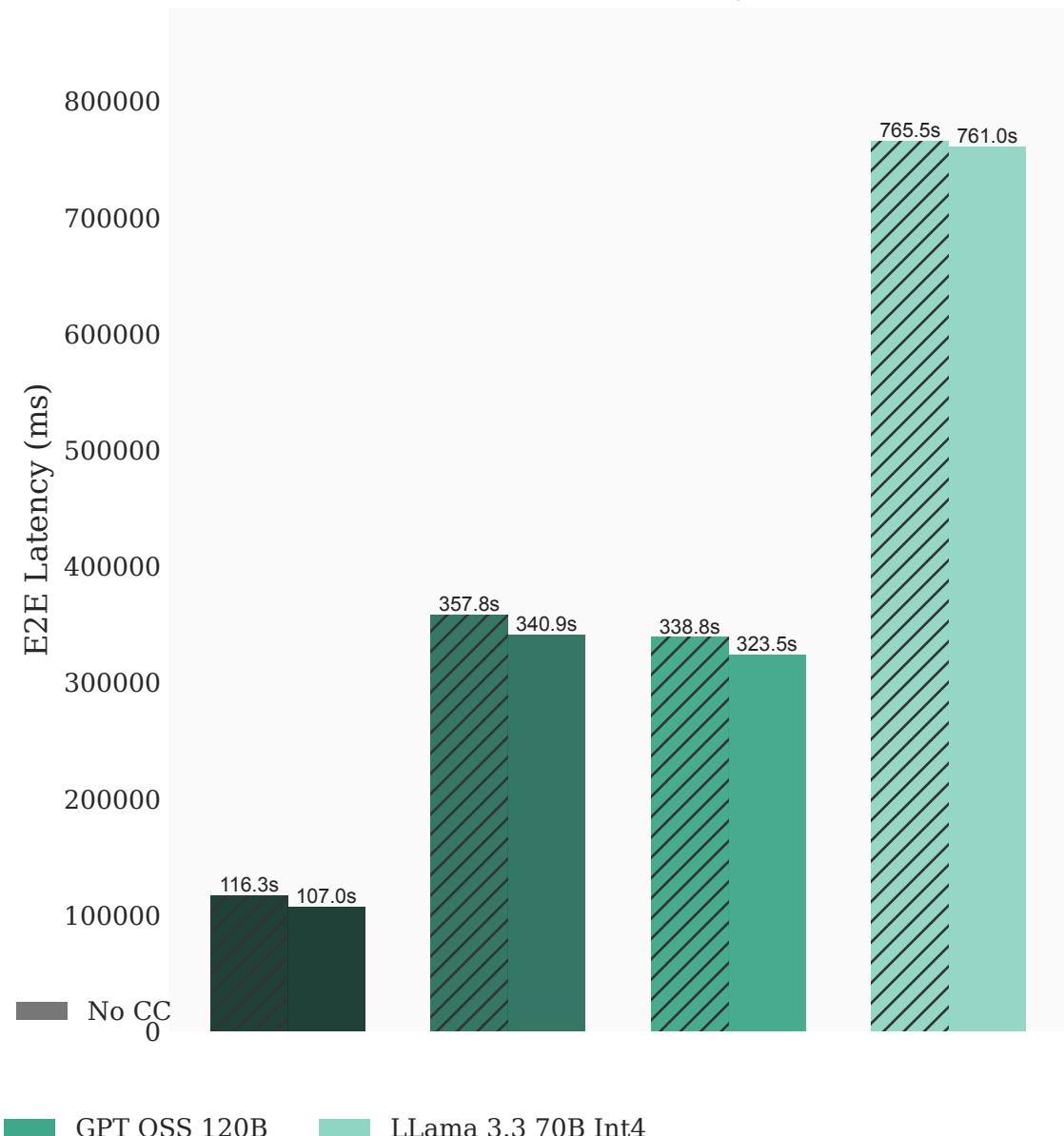


Edit 10K Characters (100 Concurrent Requests)

End-to-End Latency (Mean)

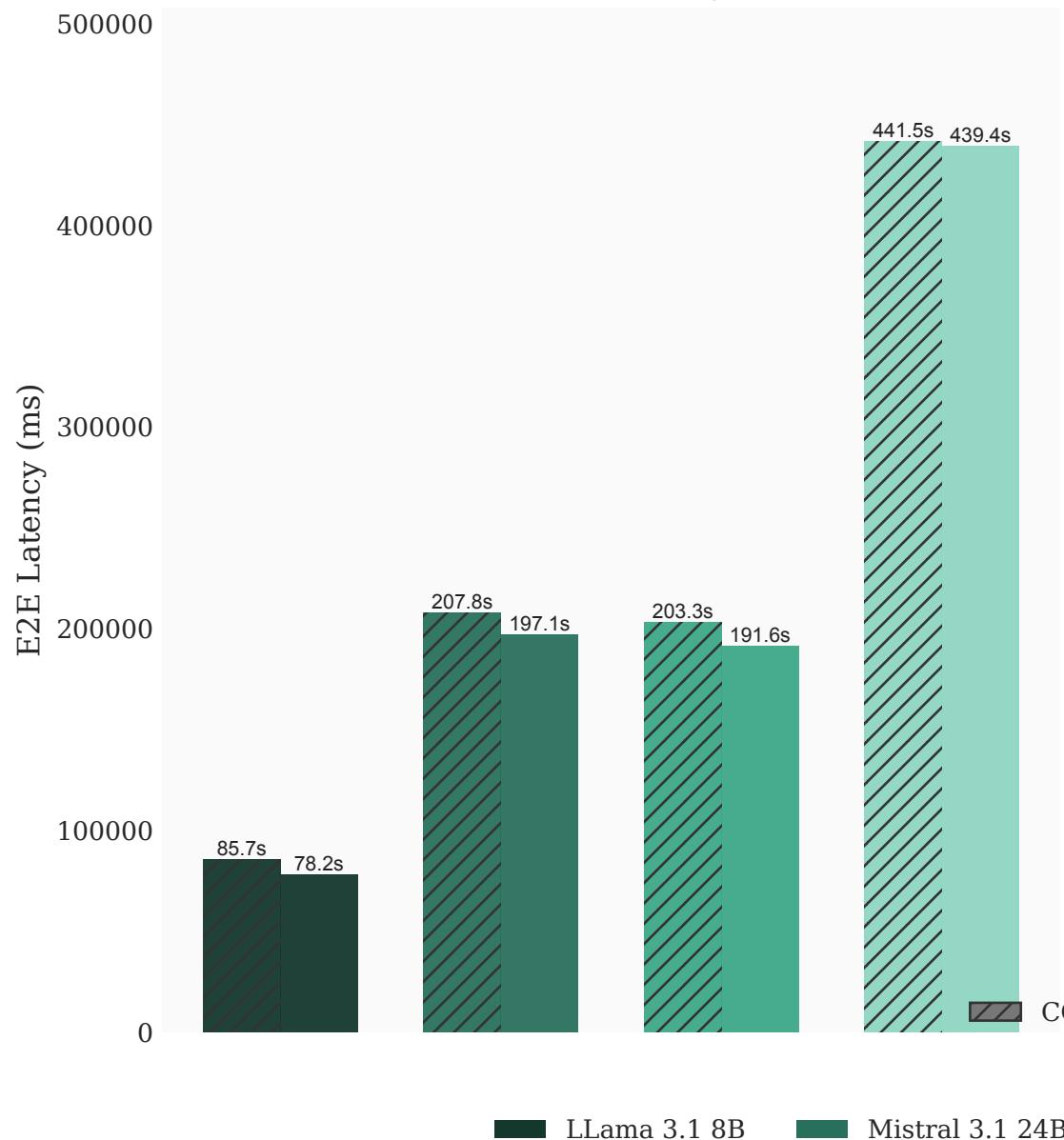


End-to-End Latency (P99)

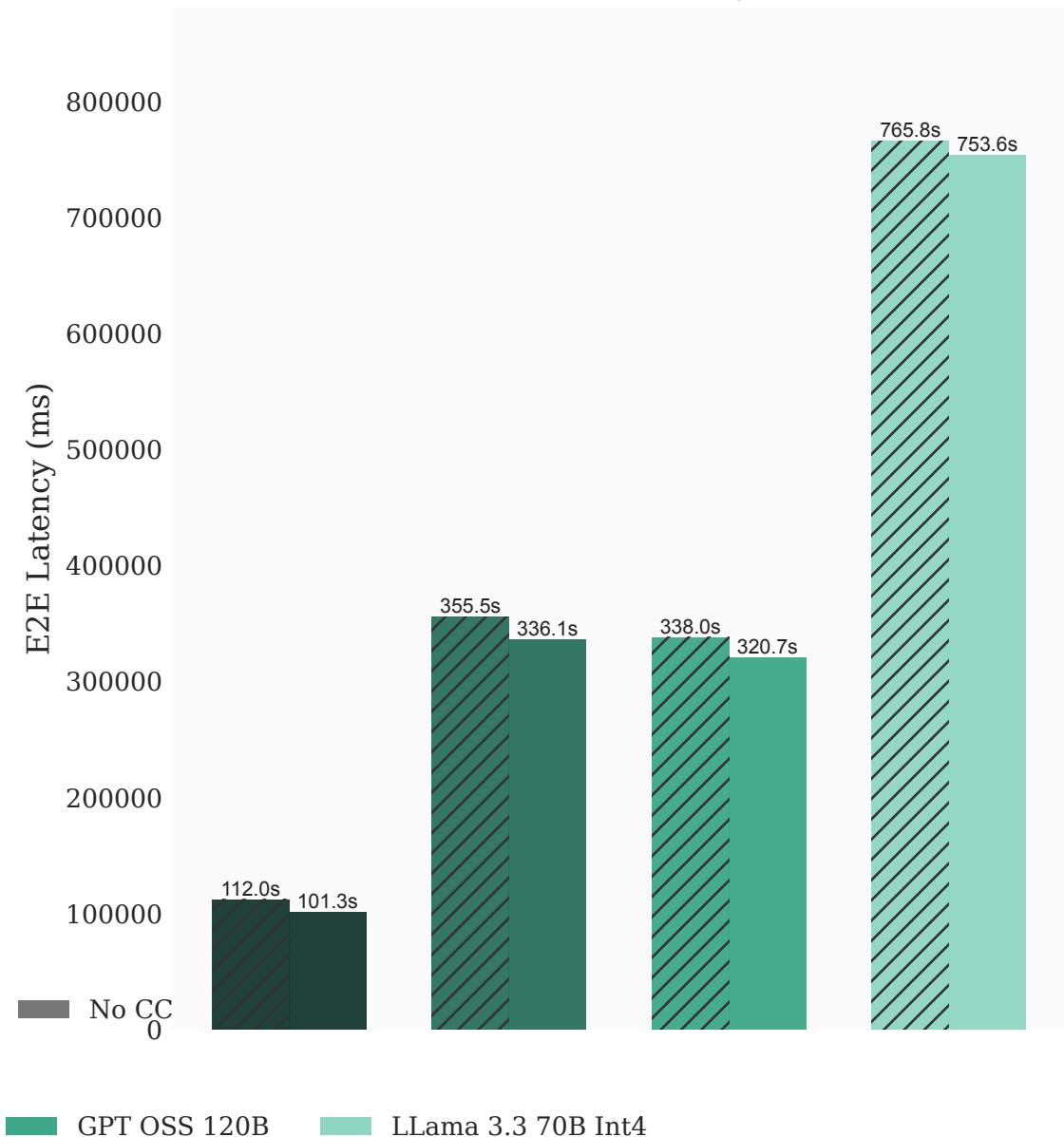


Edit 10K Characters (50 Concurrent Requests)

End-to-End Latency (Mean)

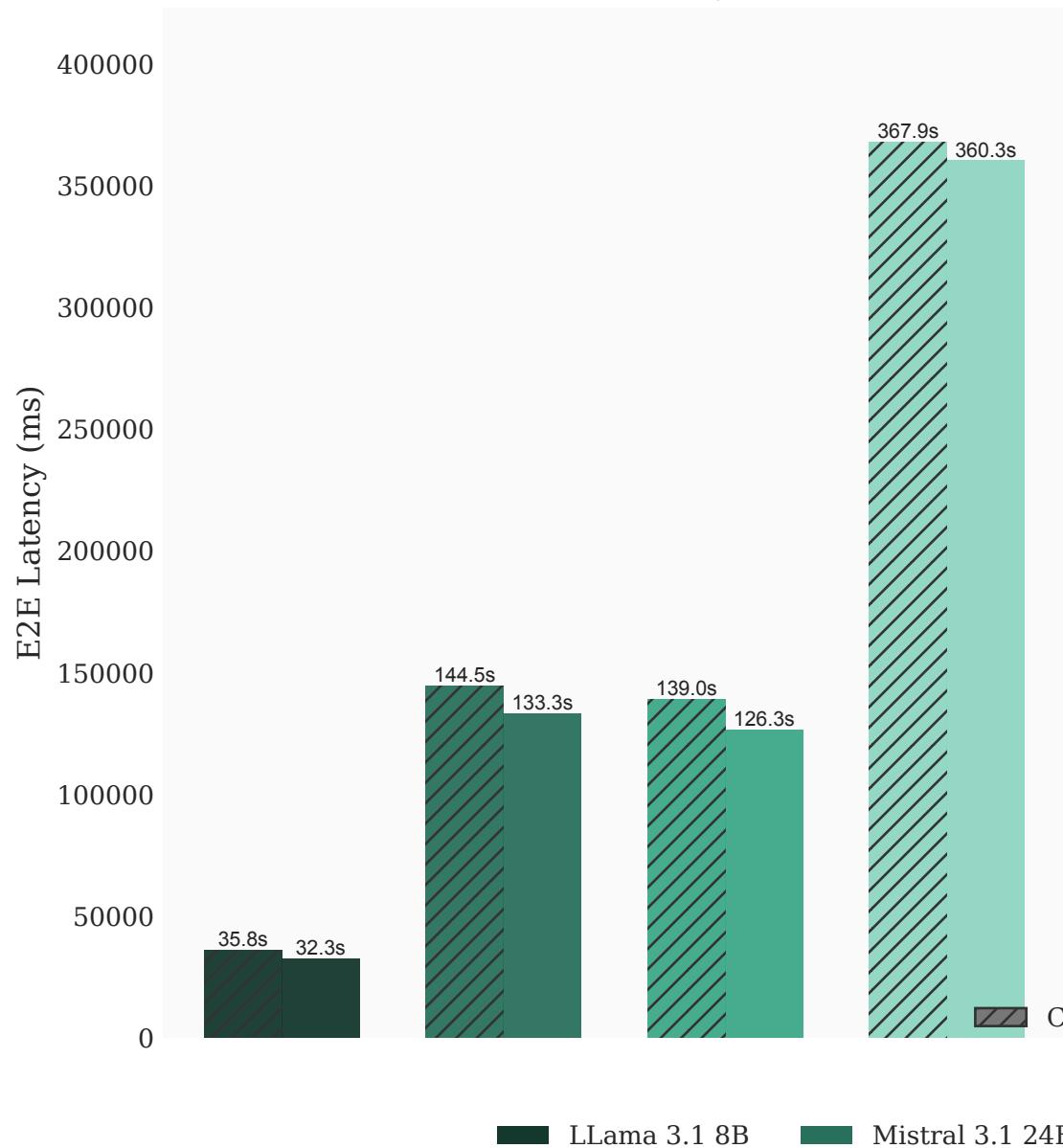


End-to-End Latency (P99)

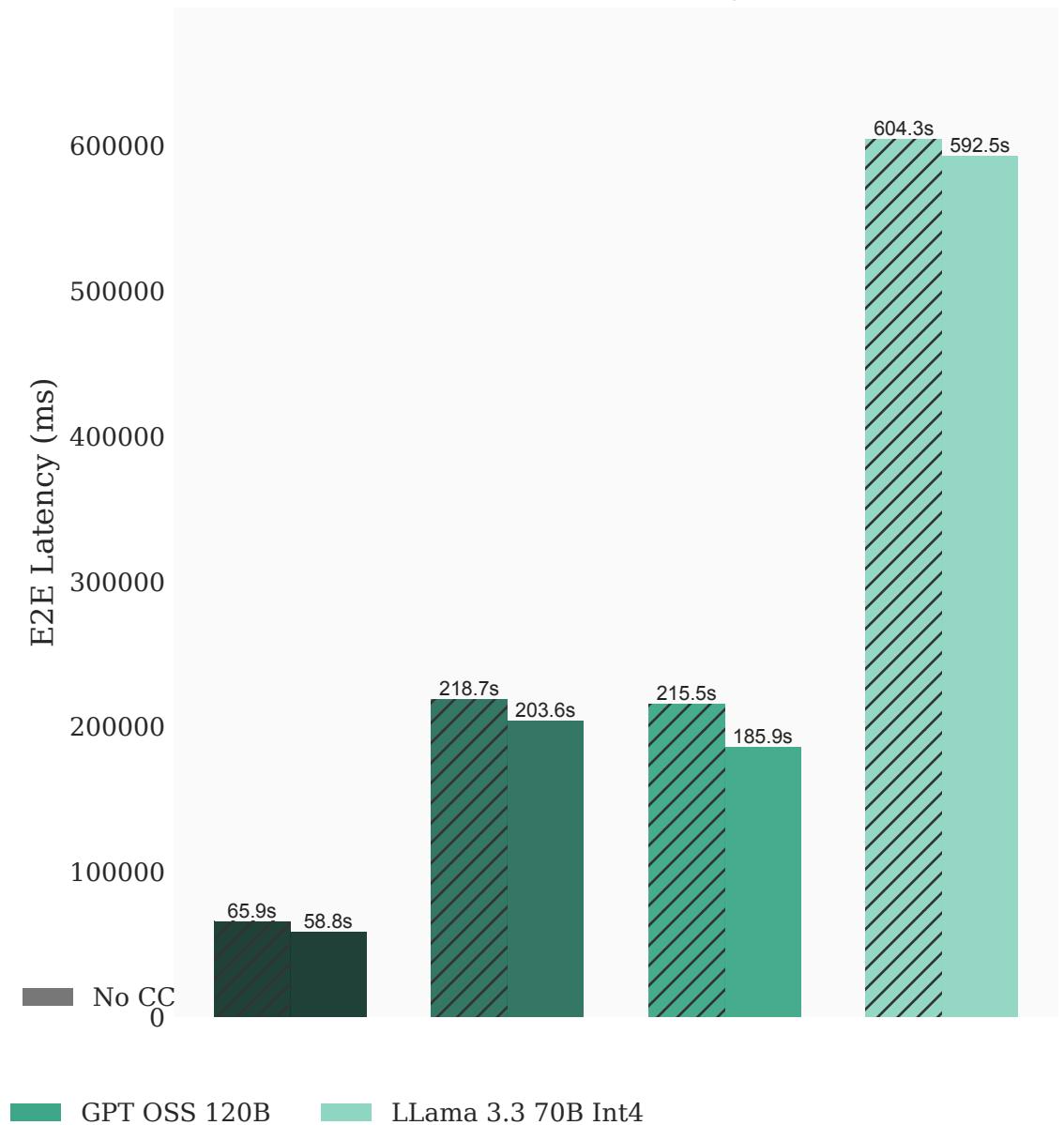


Edit 10K Characters (1 Concurrent Requests)

End-to-End Latency (Mean)

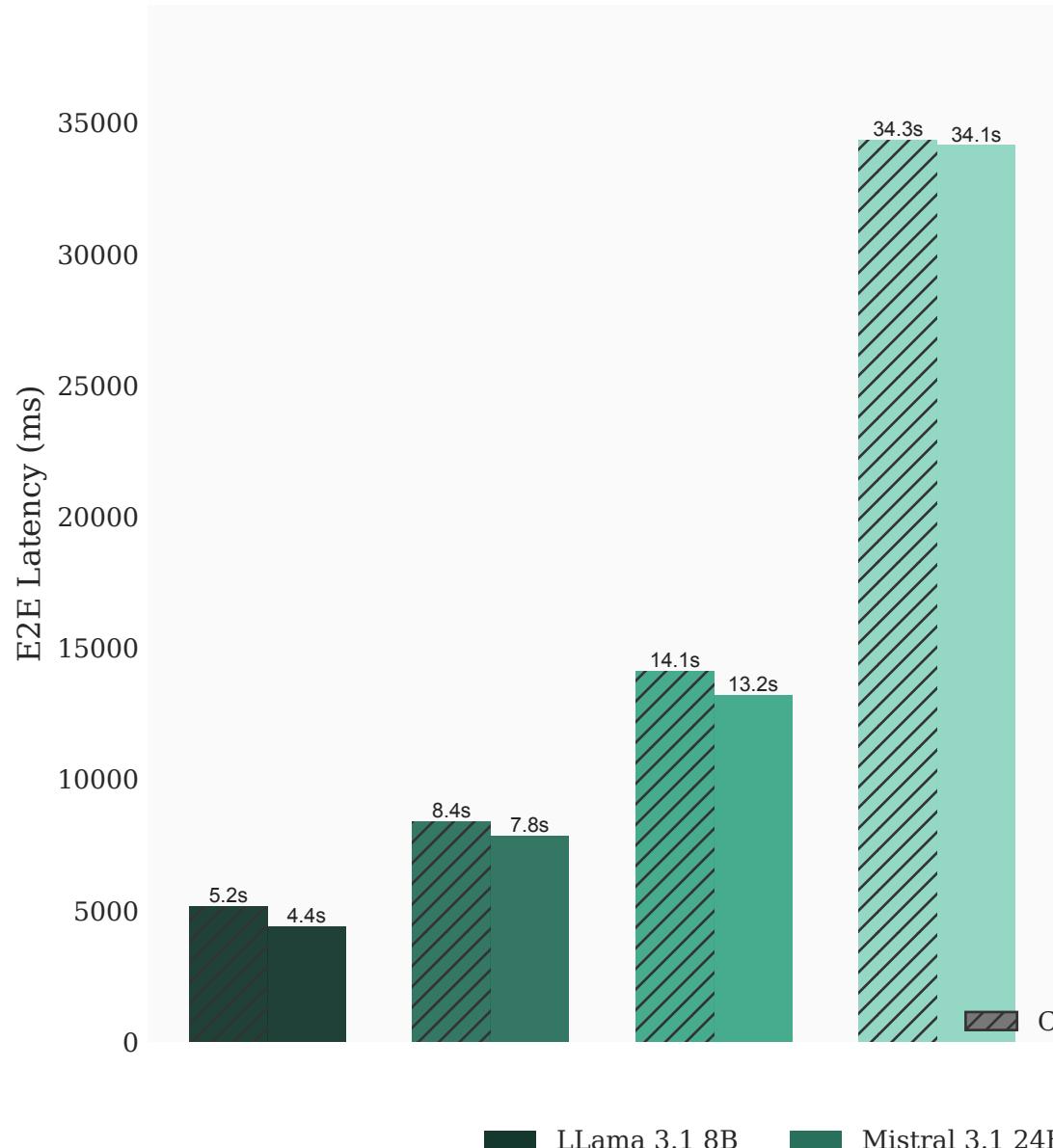


End-to-End Latency (P99)

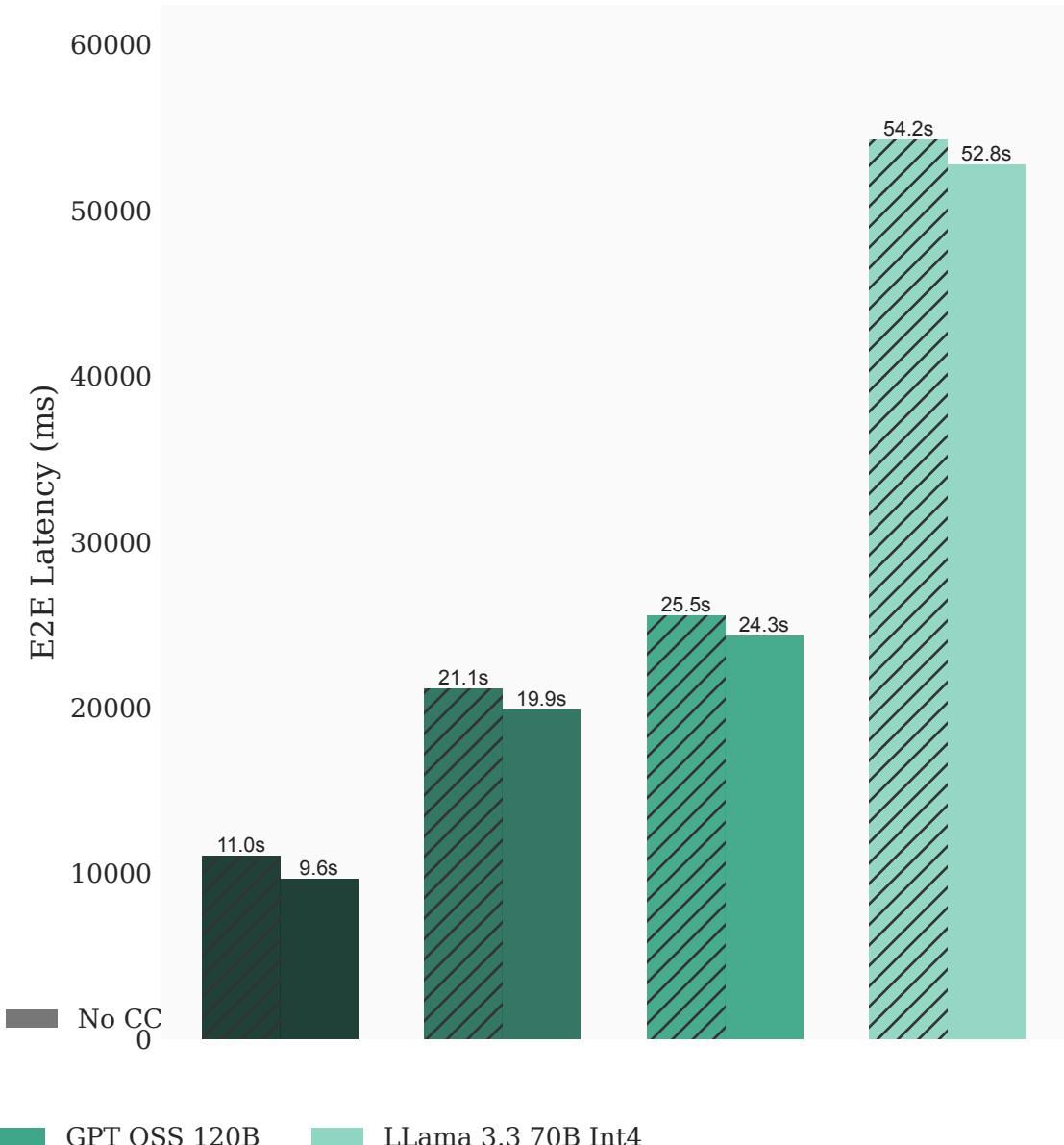


Numina Math (100 Concurrent Requests)

End-to-End Latency (Mean)

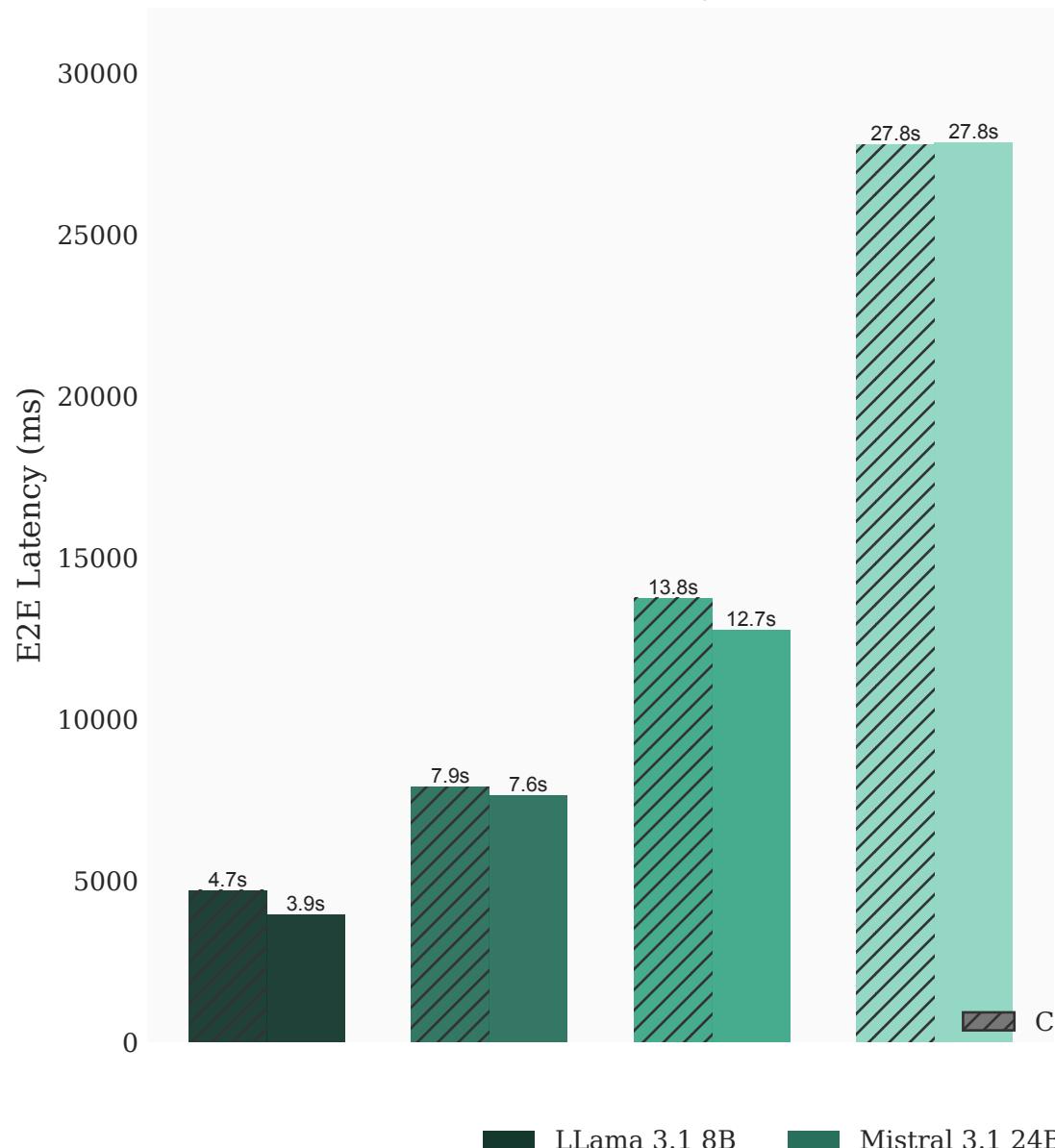


End-to-End Latency (P99)

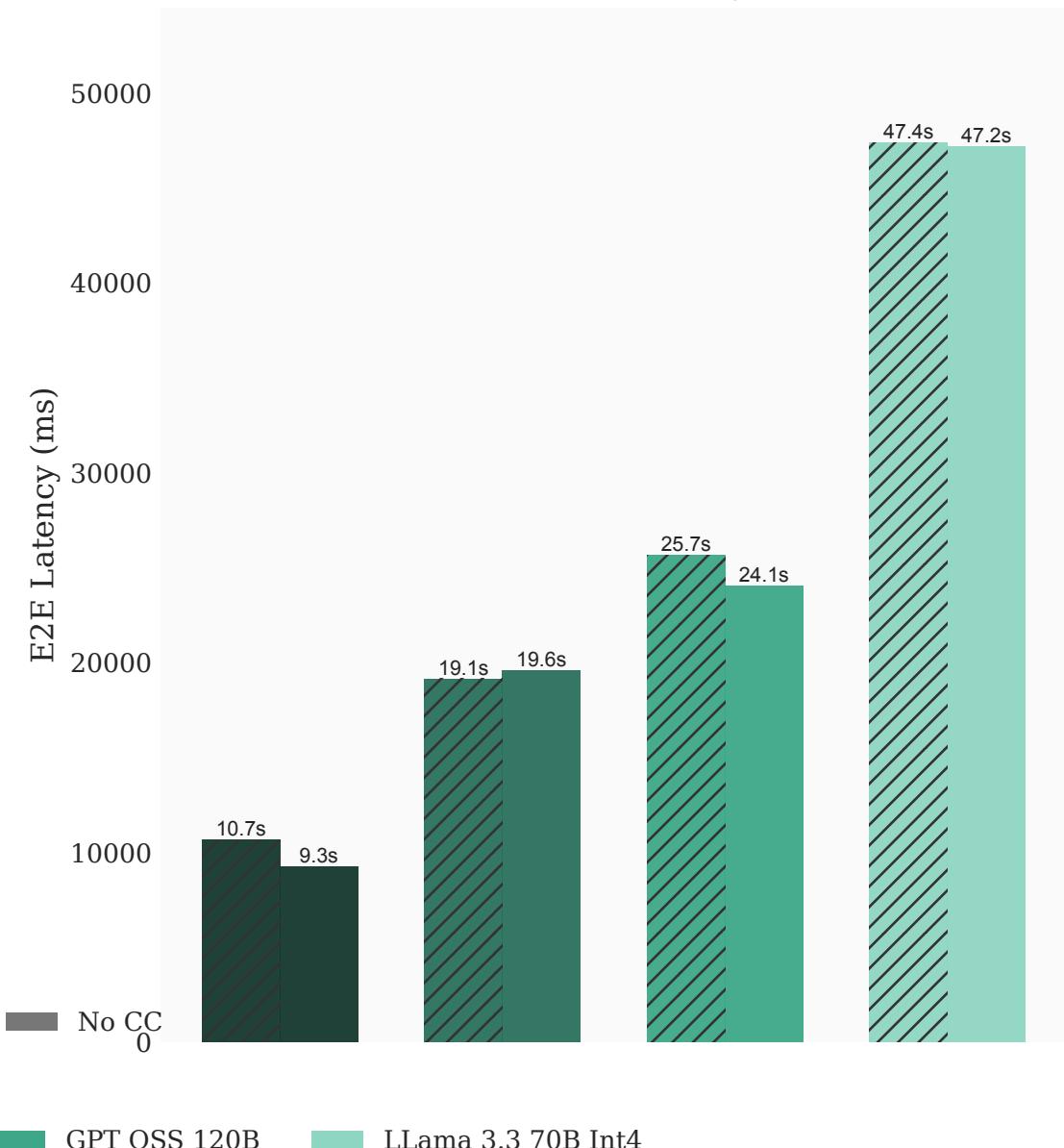


Numina Math (50 Concurrent Requests)

End-to-End Latency (Mean)

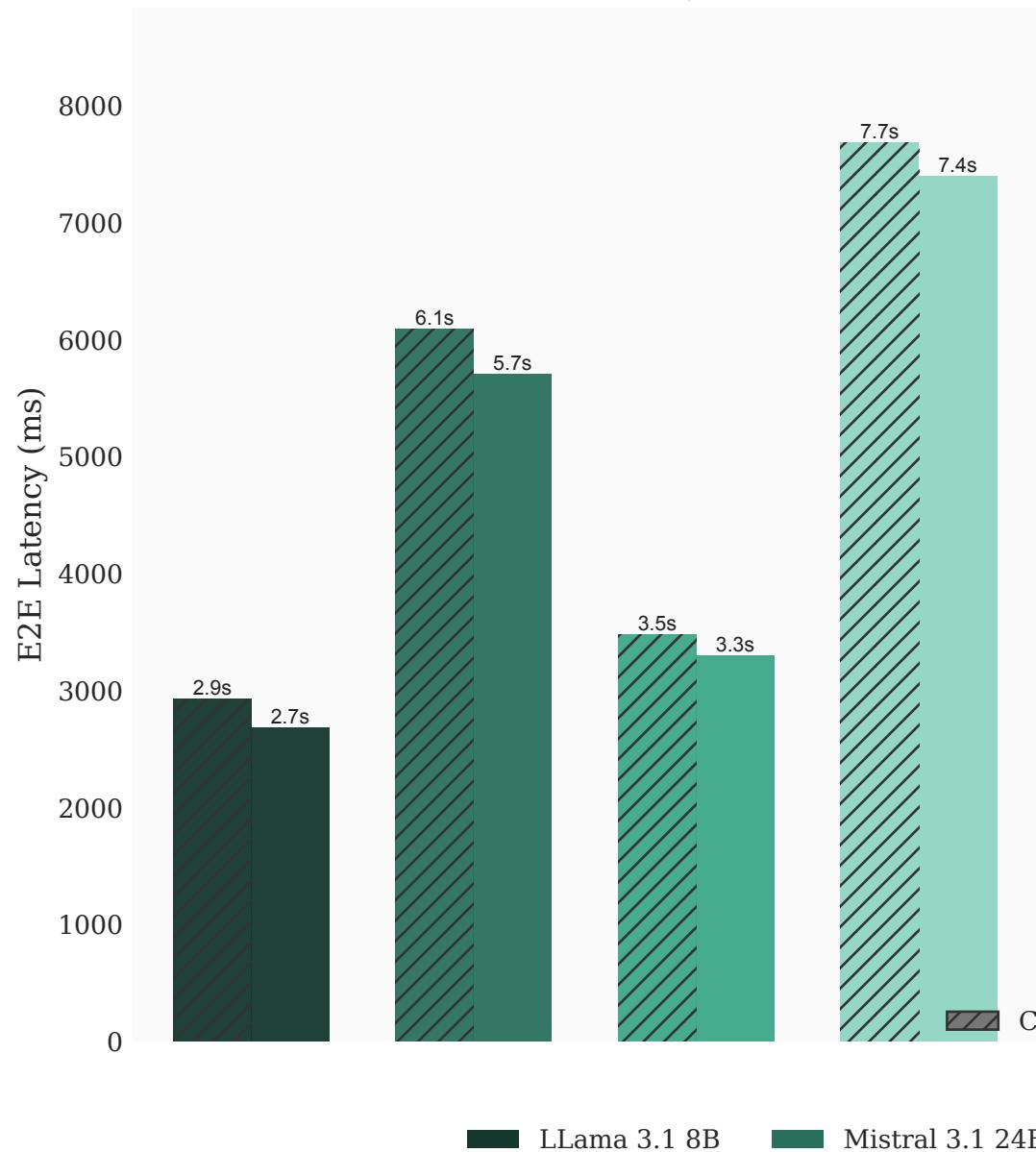


End-to-End Latency (P99)



Numina Math (1 Concurrent Requests)

End-to-End Latency (Mean)



End-to-End Latency (P99)

