	Prefill Latency	**Decode Latency**
1M Tokens 5 Books 700K Words	**14** Secs	**64** Tokens/s
5M Tokens 4 Audiobooks 55 Hours	**3.5** Mins	**56** Tokens/s
10M Tokens 4 Movies 10 Hours	**10.6** Mins	**40** Tokens/s