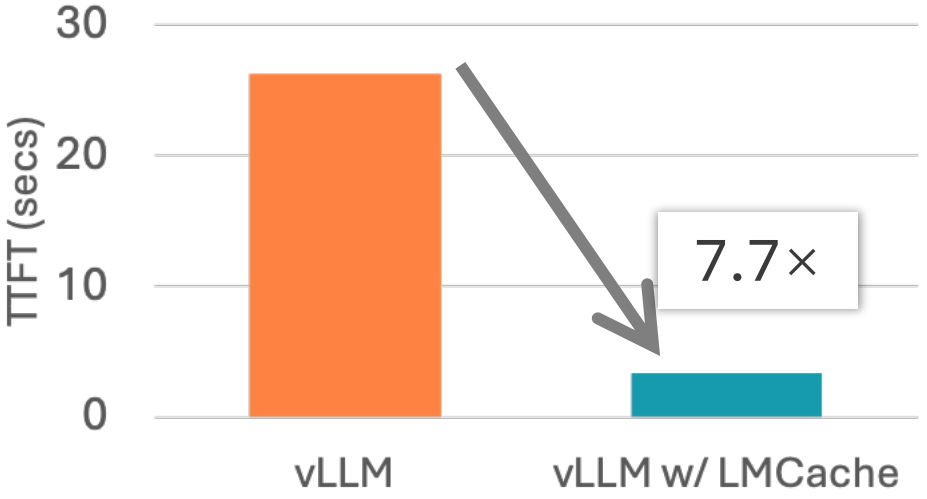
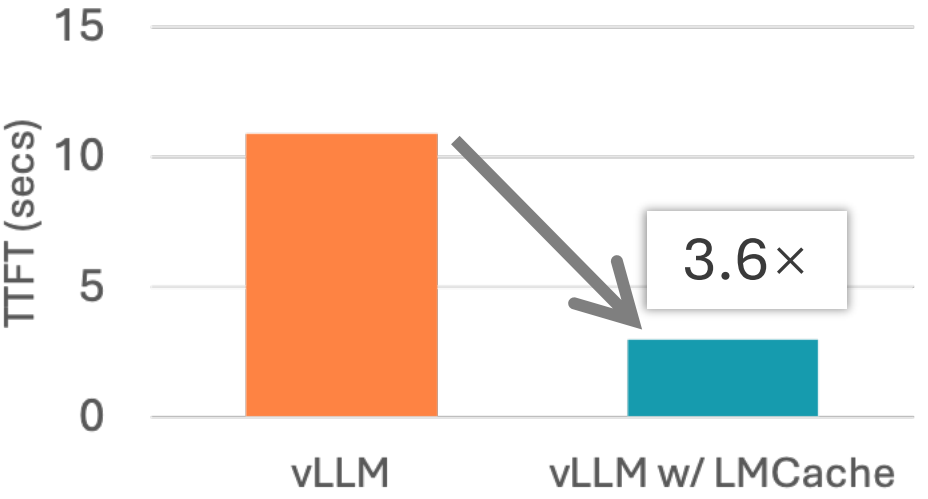


Use case 1: **long context**
*Context length**: 25K** tokens*
(Llama 70B on A40)



Use case 2: **RAG**
*Retrieved chunks**: 4 x 2K** tokens*
(Llama 70B on A40)

LMCache drastically reduces the prefill delay (TTFT) by reusing KV caches