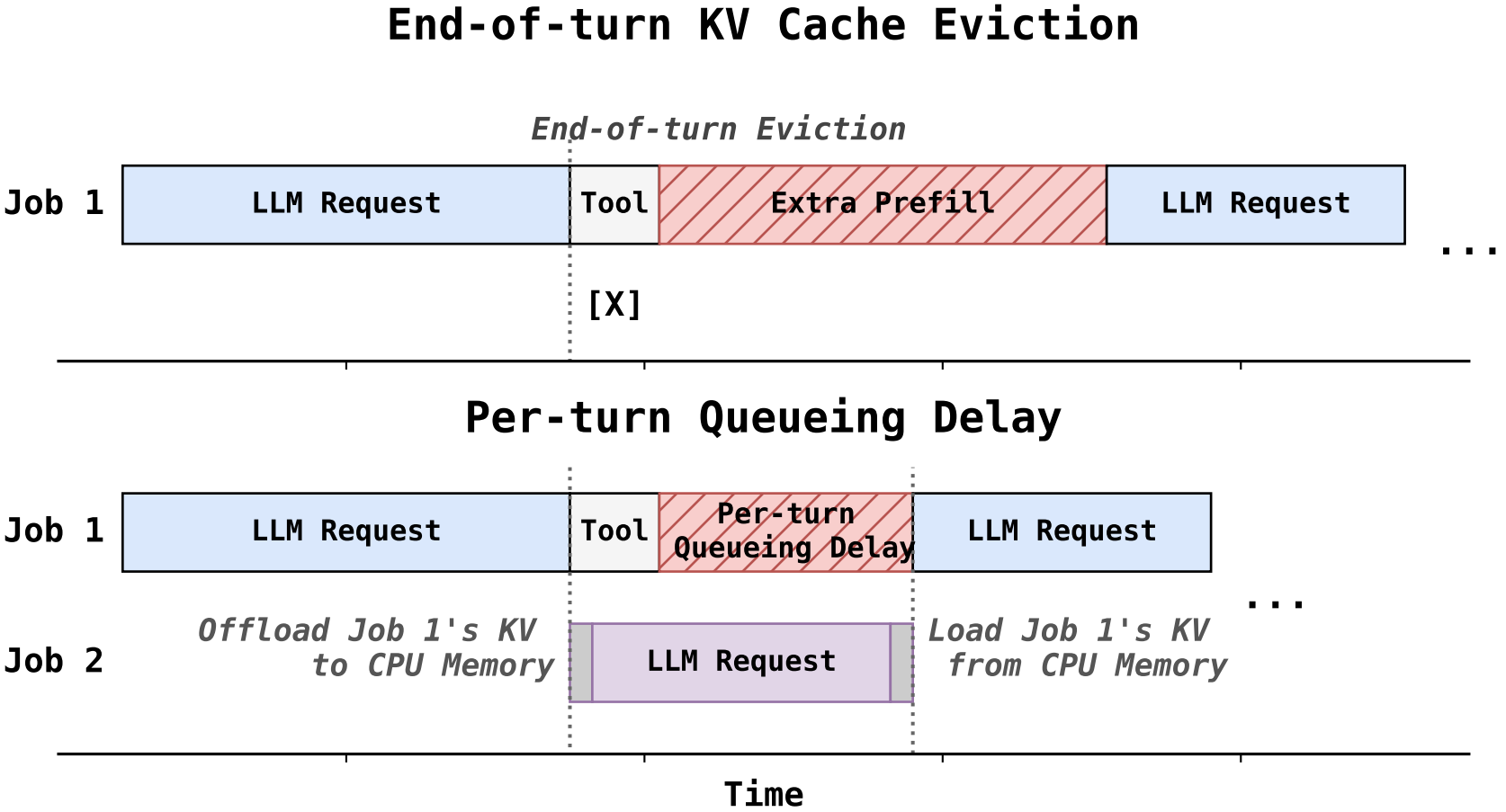
End-of-turn KV Cache Eviction
End-of-turn Eviction
Job 1
LLM Request
Tool
Extra Prefill
LLM Request
...
[X]
Per-turn Queueing Delay
Job 1
LLM Request
Tool
Per-turn Queueing Delay
LLM Request
...
Offload Job 1's KV to CPU Memory
Job 2
LLM Request
Load Job 1's KV from CPU Memory
Time