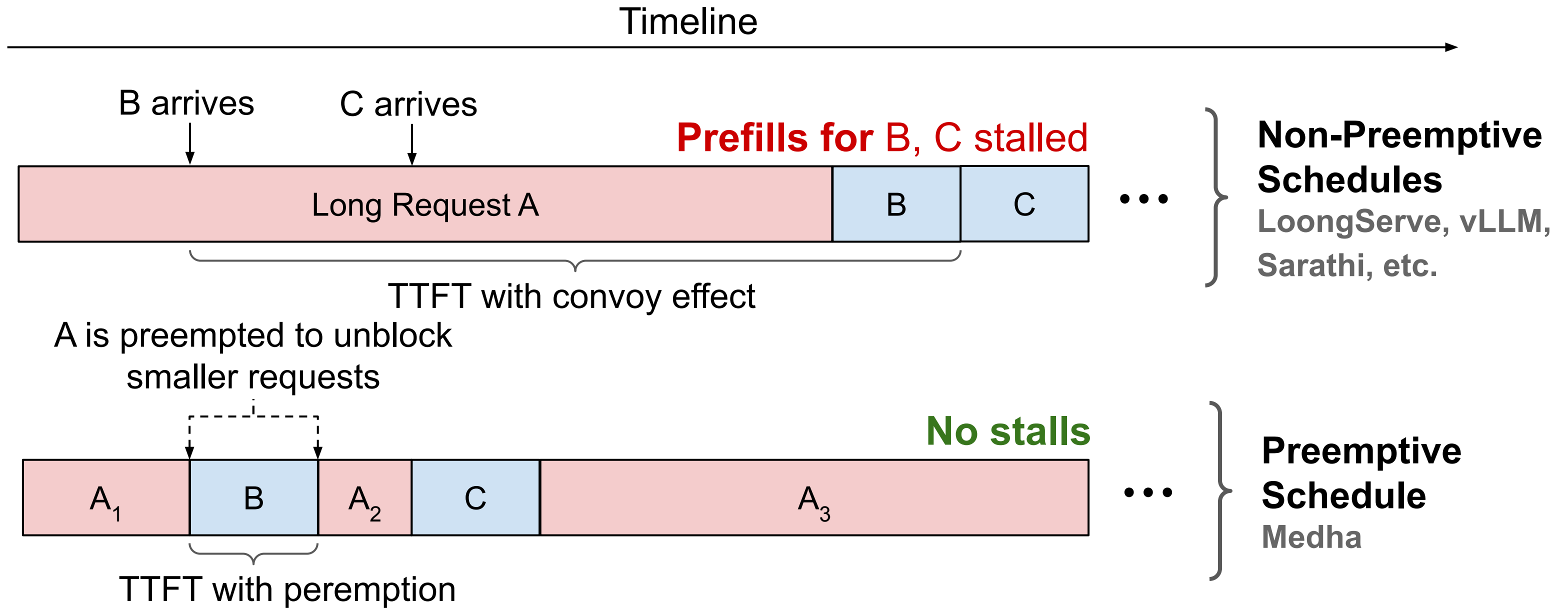
Timeline
B arrives
C arrives
Prefills for B, C stalled
Long Request A
B
C
• • •
Non-Preemptive Schedules
LoongServe, vLLM, Sarathi, etc.
TTFT with convoy effect
A is preempted to unblock smaller requests
No stalls
A_1
B
A_2
C
A_3
• • •
Preemptive Schedule
Medha
TTFT with peremption