



Important information in Replica\_Scheduler:

Water\_mark stock: The memory left for decoded tokens.

Allocation\_map: store request information of running batches.

Max\_micro\_batch\_size: number of requests can be supported in each stage

Config.batch\_size\_cap: number of requests can be supported in whole replica

Max\_micro\_batch\_size=Config.batch\_size\_cap / num\_stage

Preemted\_requests: Requests are not finished due to the memory blow up.

