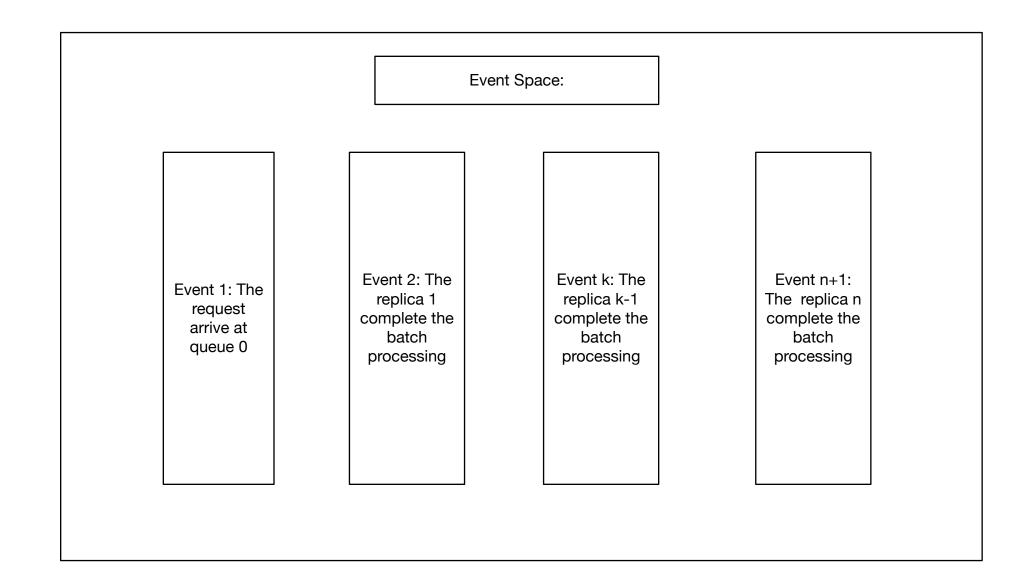
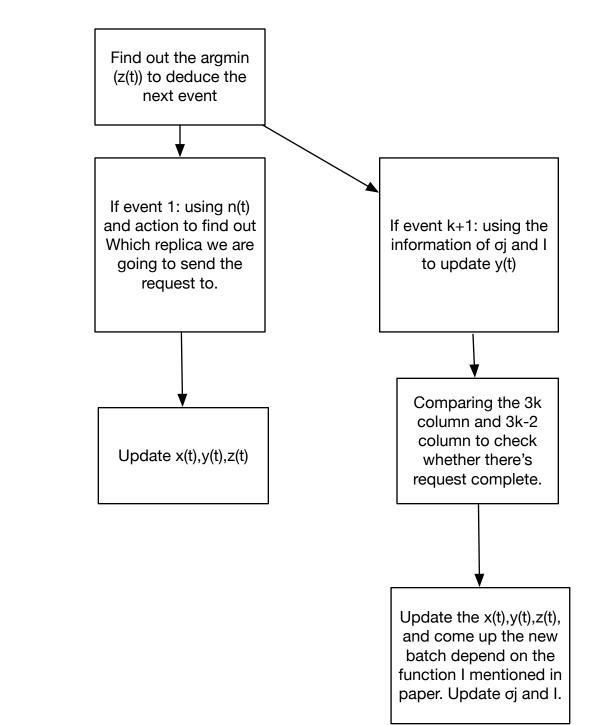


Action Space: Allocate the request to different sever, here I use the {0,1} ^n*N matrix, where N is the total number of requests arrive.





If we want to change the batching policy, we can use different function here.

Hidden State: σj:the vector to represent number of requests of processing batch in each replica

I ∈ {0, 1}^n:the vector to represent whether each replica is serving queue 1 or queue 2, where 0 indicates queue 1 and 1 indicates queue 2.

N(t): represent how many request arrive before time t.

Parameter: ·

- number of tokens in block
- NumBlocks:The total number of blocks in replica
- Watermarkblocks: The number of blocks reserved in replica
- Maxtokensinbatch: The maximum number of blocks in Batch
- Configbatchsize: The maximum number of requests in replica
- maxmicrobatchsize: The maximum number of requests in a batch