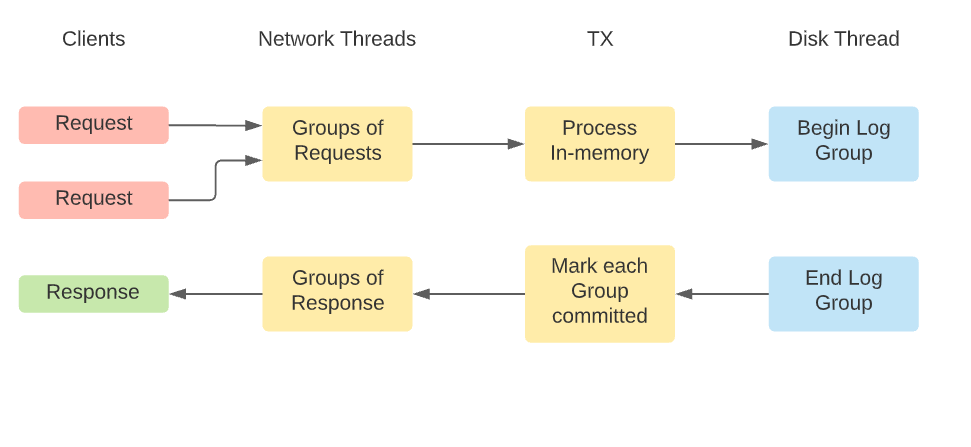
Approach 1:

Each thread is working in parallel and does not interfere with other threads making their job. The more parallel and heavier the workload is, the fewer syscalls per request and the more requests per second the system can handle.

At the same time, the latency is good because threads don't wait for other threads, they just handle as much work as they have right now, and while they're handling it, a new piece of work is being prepared in parallel.



Approach 2: We can use a load balancer which will  evenly distributes incoming traffic among web servers that are defined in a load-balanced set.

