Explanation
=========
Diagram 1
It shows clients connecting to hosts. They do not connect to hosts directly; for sake of simplicity it is not shown. Most likely it connected through loadbalancer. It makes sense to connect to loadbalancer to get video details. It is a bad design though. The loadbalancer is meant to distribute load and it should do just that.

Diagram 2
There are clusters of microservices connected to each host. They pick up the video link and store it into a database. The database makes sure of the count. Microservices create a new record, if it does not exist. If a record exists then it increments count. This makes sure of multiple access not overwriting

Datastructure {
Video link: #string
}

Algorithm
Check record exists - yes then update the count; if no then create record and set count 0. These updates are done to the database. The database enters concurrency.

Diagram 3
Microservice picks from the database and reports count on video. It will report max access at any given time. Over a period the most watched video may change

Datastructure
Video structure {
Video link: #string
Video count: #int
}