```
l=[1,2,4,8]
index=0
pre=redis.get('hits')
while(True):
    time.sleep(10)
    #monitoring interval, we are monitoring hits happened in 10 seconds
    now=redis.get('hits')
    diff=now-pre
    #get the number of hits happen in 10 seconds
    if diff>2*scale and index<3:
    # we assume more than 2 requests for each service will trigger the scale up since the average computation time is around 4 seconds
        Scale up:
            Increase the index by one, scale with l[index]
    #the scaling policy is double the services to scale up fast enough
    elif diff<2 and index >= 1:
        Scale down:
            Decrease the index by one, scale with l[index]
    #the scaling policy is half the services to scale down fast enough
    # we only scale down when only a few requests happens in 10 seconds. In case of a short dip in user requests.
    pre=now
```