Lab #2: Deploying Request Splitters & load Balancers in Kubernetes

SOFE 4790U : Distributed Systems

Due: Oct 2, 2022

Joshua White (100747854)

Discussion

During this lab multiple yaml files were developed to provide the system with a factor of load distribution. These files were calibrated so that 10% of the users requests would be sent and filled by an external server. This is done by creating a request splitting file which is operated by an ambassador file which will deploy and utilize it once they are associated with one another.

Why Auto Scaling is Used

Auto scaling is used to ensure that users always have a sufficient amount of resources. In a distributed system without autoscaling resources must be allocated to user requests manually meaning a user will likely be given excess resources for their request which will not change until they reach their resource limitation and request more or if they proactively request and are granted more. In an auto scaling system resources are automatically deployed to the user upon a request and react to the demand the user puts on the system.

Differences Between Load Balancing & Auto Scaling

Load balancing is the function of a portion of the software side of a distributed system. This portion of the software responds to users requests to access the system and ensures that the combination of requests from the users are sply up across multiple machines to see that one machine is not given to much more then another so that all users can receive the best support and processing available. Auto scaling references the portion of software which responds to individual user requests and how much computation power that users needs at a time, for instance a user making 90 requests per second while increasing at an additional 2 requests each second will be given enough resources to insure that hardware can account for the growing number of requests while also being ready for a sudin jump in demand which may exceed the 2 request per second increase rate.