Creating operational setup for high-available and scalable REST API

I would choose to use kubernetes for deployment of storage and service layer. I would in production use GKE but has in this test used my own Kubernetes cluster.

The diagram for the solution is shown in following diagram:

The cassandra deployment is created using StatefulSets and each has a local Persistent Volume (static created). It is possible to create a more dynamic way of creating a local PV when a node is

created but this is beyond current scope. Cassandra is a headless service and thus exposed using an kubernetes dns (cassandra). This limits the possibility for load balancing, and should be fixed in a production setup.

The gameapi is created using ReplicaSets and these is fronted by a Load Balancer service, which only provides a external ip to a real Load Balancer.

The following should be done before moving to production:

- Move to a Cloud Provider (like GKE) that implements provision of new nodes
- Implement autoscaling of GAMEAPI so we runs enough but not too many replicas.
- Implement autoscaling but only up of cassandra pods. Scaling down of persistent storage should be done manually if at all.
- Performance test the setup
- Add monitoring
- Multi datacenter and region for high availability
- Geographic distribution to minimize latency for close clients.