App Engine is Google's platform as a service, you don't have to manage servers.(PAAS)

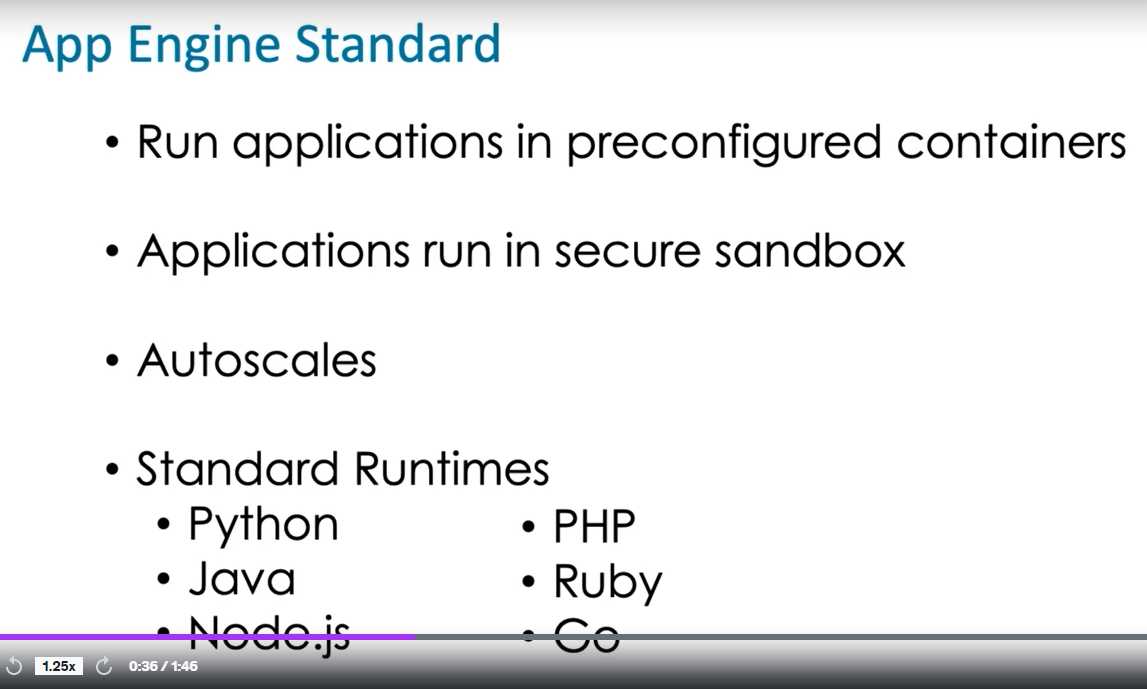
This is considered a serverless environment.

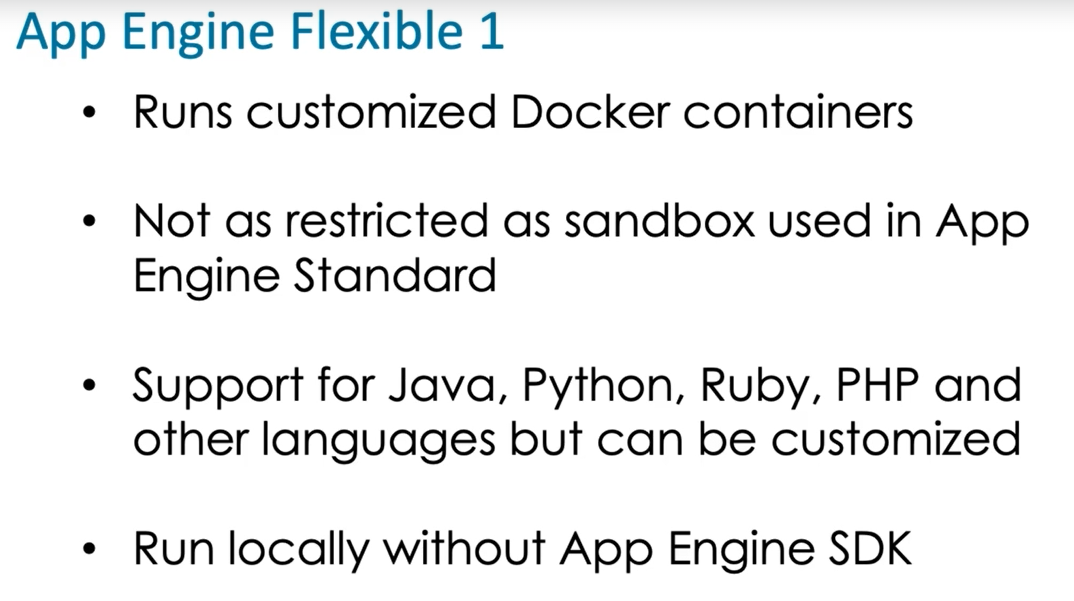
App Engine is well suited for running microservices.

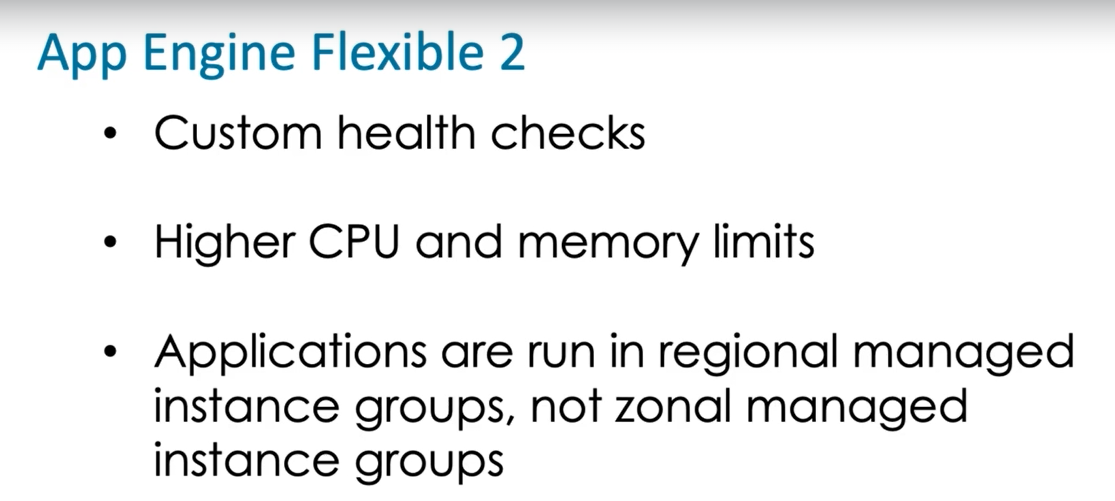
Now, originally, App Engine was limited to a small set of language-specific runtimes, but now there's a new version known as App Engine Flexible that can actually run custom containers .

App Engine is designed for applications written in supported languages, including Python 3, that need to run at low cost, and need to scale in response to rapid increases in load. App Engine is a managed service and as such minimizes operational overhead . Compute Engine and Kubernetes Engine both require more management overhead. Cloud Functions are used to respond to events in GCP, not to execute a continually running application.

a new version of App Engine called App Engine flexible, and the original is called App Engine standard.







Deploying an application to app engine :

gcloud components install app-engine-python - install specific components in gcloud cli.

gcloud app deploy app.yaml - **Deploy your application to App Engine** using the gcloud app deploy command. The deployment command automatically builds a container image by using the Cloud Build service and then deploys that image to the App Engine flexible environment.

gcloud app browse - Open the current app in a web browser.

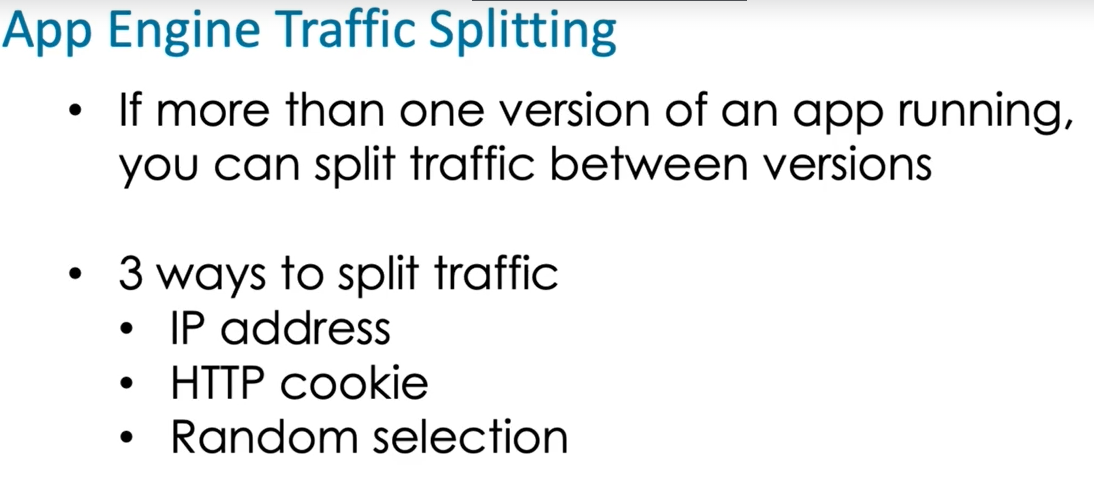
advantages of using App Engine is that it manages scaling for us,

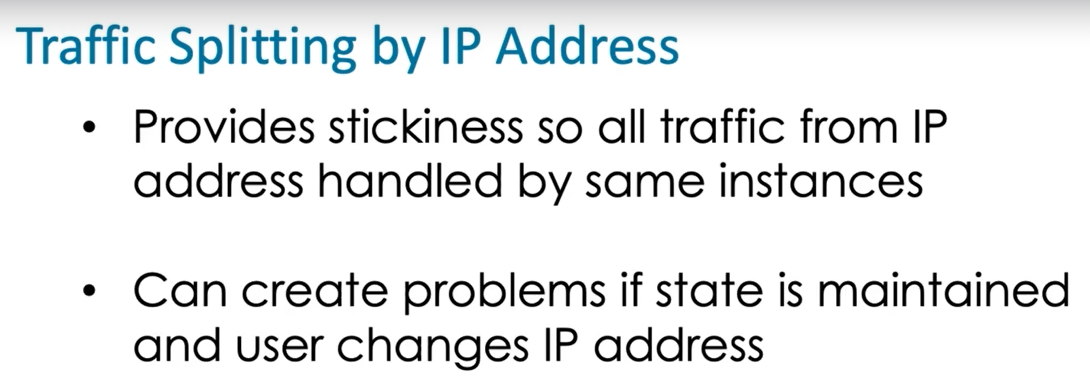
So, applications execute on App Engine managed instances. As you'll probably recall,

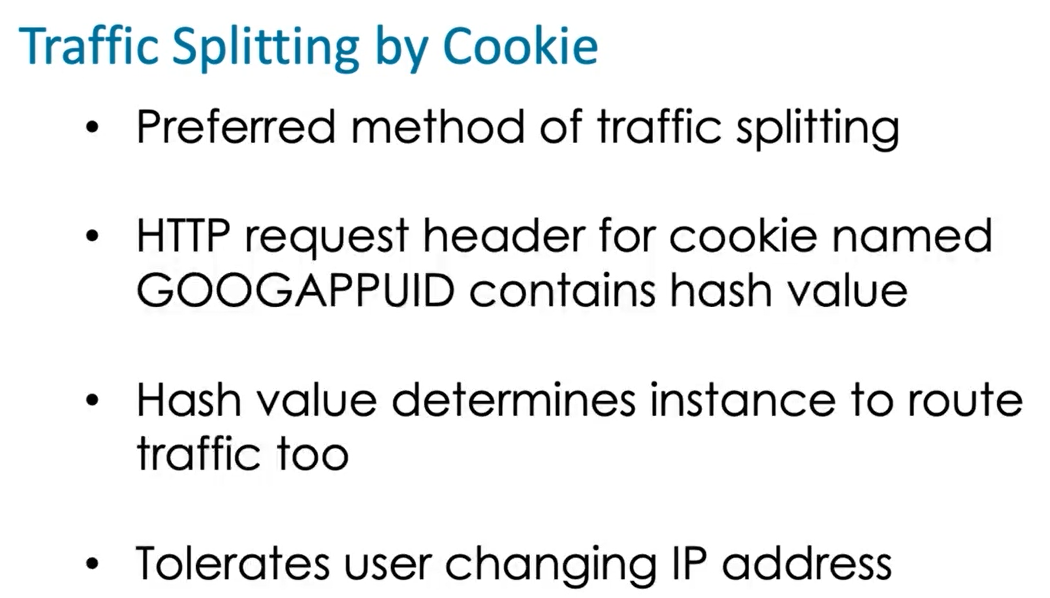
managed instance groups support auto-scaling. So it's no surprise Google takes advantage of instance group auto-scaling to support App Engine scaling, and scale is based on load when running dynamic instances. Now, you can configure what are known as resident instances, which run all the time, but when auto-scaling is enabled, we use dynamic instances. So you can basically go down to zero instances, in fact.

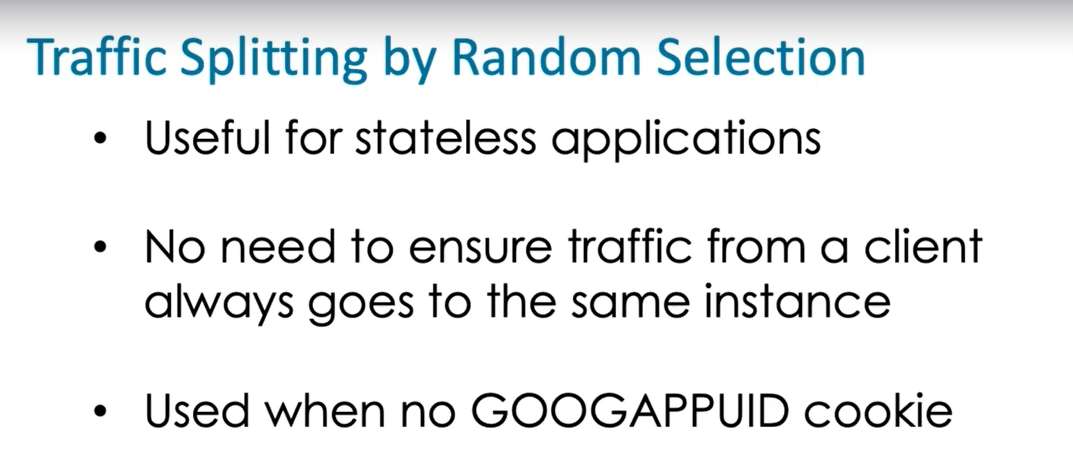
Now, when you do specify auto-scaling parameters, you provide them in the app.yaml file, and so some of the configuration options you might wanna take a look at is target CPU utilization and target throughput utilization. You can also look at max concurrent requests and then also look at latency, like maximum pending latency and minimum pending latency. So again, by using App Engine YAML configuration parameters, you can kind of tweak how auto-scaling works within App Engine.

Splitting traffic in app engine :









Resident instances run at all times regardless of the load on the application. Unmanaged instance groups are not used by App Engine and changing the health check to always return a healthy status is not something that can be done with certainty since the instance running the health check may not be healthy, which means it may not be able to return a value.

Deleting an App Engine instance is something an administrator could do but is not needed by someone who only needs to review code.

App.yaml is the configuration file in App Engine that specifies the runtime environment.

App Engine Standard which is a platform as a service offering that runs applications in sandboxes configured by Google. App Engine Flexible is a PaaS offering but users configure their own containers rather than run in a Google-configured sandbox.

App Engine Flexible does not restrict applications to Python, Go, Php, and Javascript.