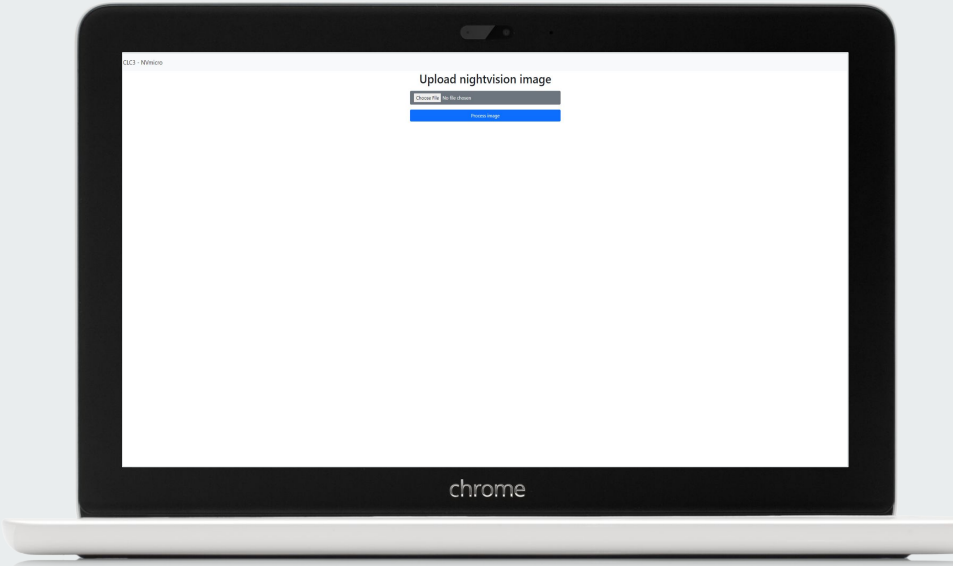
Project nvmicro

Jonas Haslauer & Kilian Straka

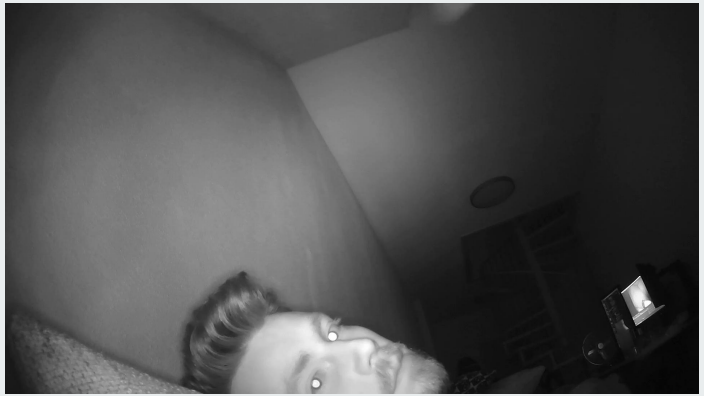


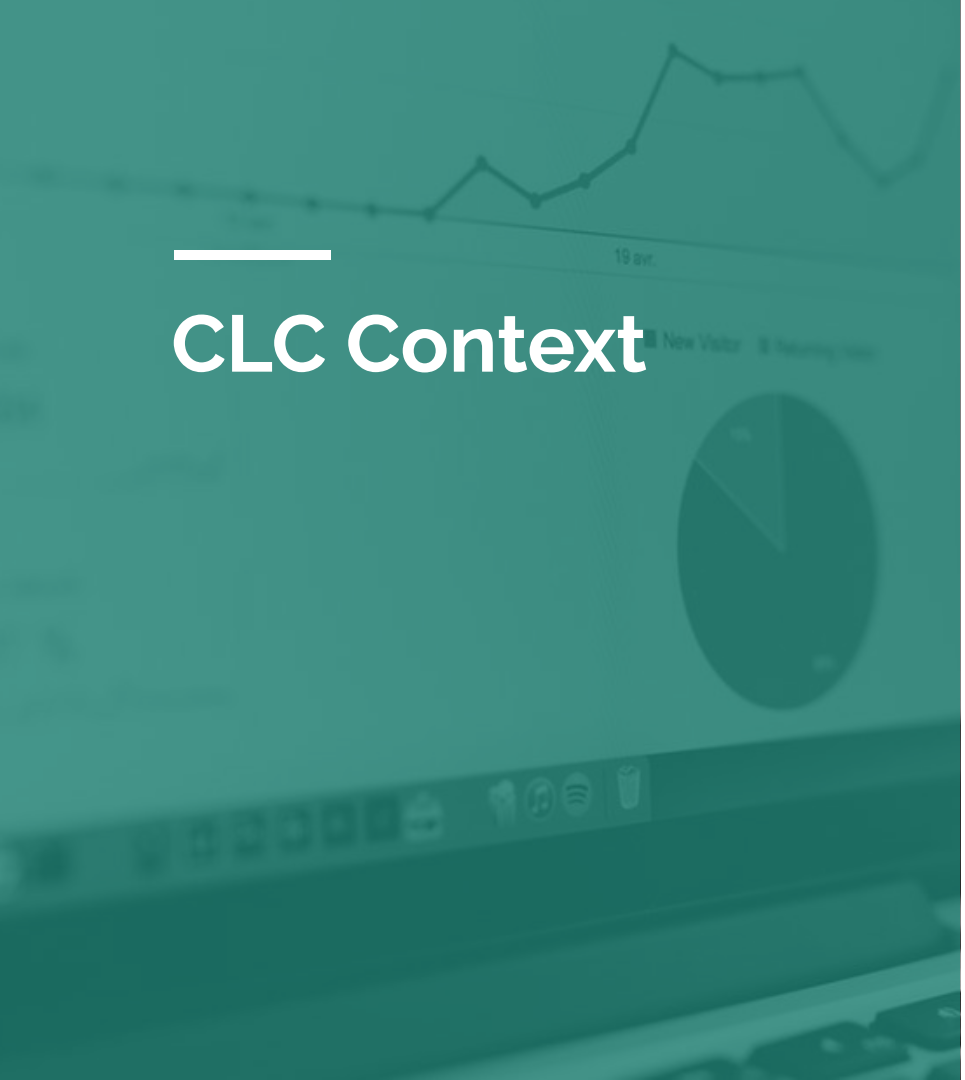
Outline

Problem statement

Provide a service that analyzes the eye-state of humans in night vision images.

An open eye while sleeping can indicate epileptic episodes.





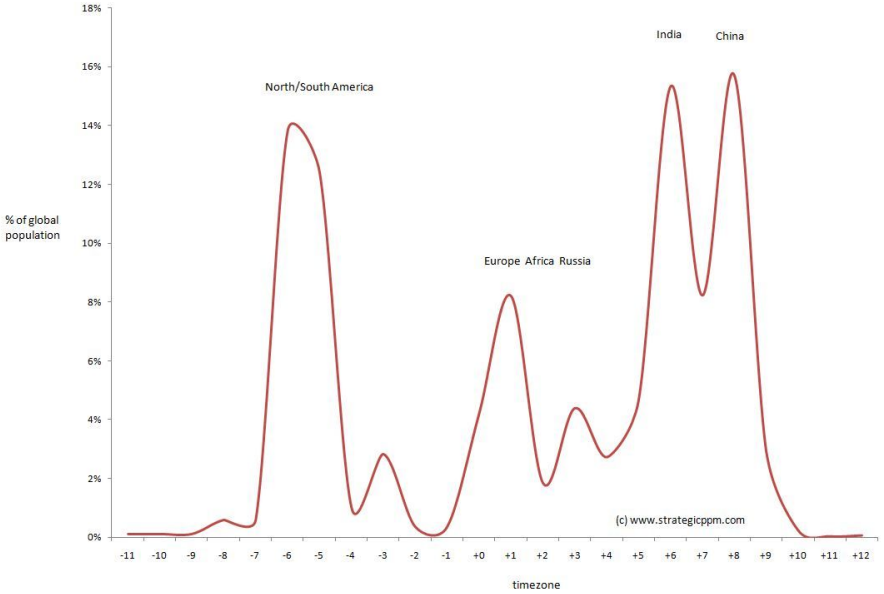
CLC Context

- The service is provided in a GKE-Cluster
- Horizontal-Pod-Autoscaler (HPA) scales up the number of replicas in situations of high demand (and vice versa for low demand)
- HPA bases its decision on custom metrics defined by us
- VPA manages requested resources

But... Why?

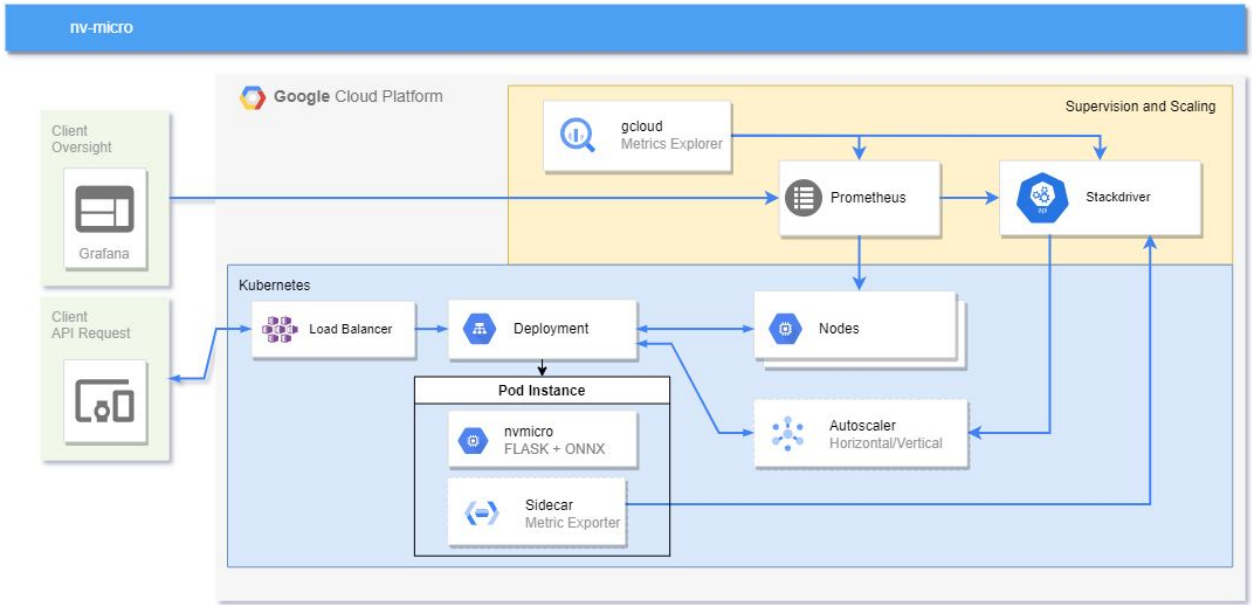
To provide a service like this on a global scale fluctuations in demand have to be accounted for.

Distribution of global population in the different time zones



https://strategicppm.files.wordpress.com/2011/03/pop1.jpg

Architecture



Horizontal Pod Autoscaler

- ➔ Sits in Kubernetes control plane
- ➔ Periodically checks a metric against a desired value
- ➔ Dynamically changes the number of replicas for a deployment
- ➔ Input is a metric (out of the box only CPU util. and RAM usage)
- ➔ With custom metric stackdriver adapter *everything is possible*
- ➔ Unschedulable pods are getting scheduled on next available node

Vertical Pod Autoscaler

- Restarts pods with updated CPU and RAM requests based on update strategy
  - Auto: pods get restarted with updated requests after they exceeded the boundaries for a while
  - Initial: only newly scheduled pods will be updated with resource requests
- This is used to efficiently provision the resources of a node
- There are two conditions to consider
  - A pod should have enough resources to work properly
  - A pod should not request more resources than it needs

Live demo

Lessons Learned

- Gcloud is complex
- Making the various services work together can be challenging
- Trial and error in the cloud is very time intensive
- Everything sounds easy in the tutorials but reality might look different
- Blindly following tutorials won't work out for your specific use case in most cases
- Dependency optimization is important

Questions?