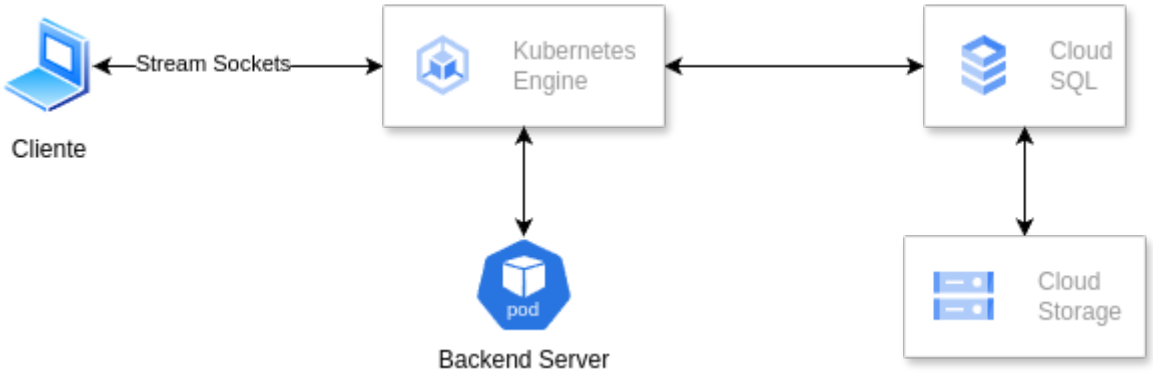
Documentación Arquitectura Proyecto 3 Sistemas Distribuidos

Integrantes

- **Arturo Alvarez - 2020407013**

Diagrama de la Arquitectura



Descripción de la Arquitectura

Cliente (Frontend)

- Los clientes (usuarios) se conectan al servidor de la aplicación mediante stream sockets utilizando el protocolo TCP.
- Cada cliente abre una conexión dedicada al servidor para enviar y recibir mensajes en tiempo real.

Servidor de Aplicación (Backend)

- Desplegado como un conjunto de pods dentro de un clúster de Kubernetes en Google Kubernetes Engine (GKE).
- Kubernetes se encarga de escalar horizontalmente (más pods) cuando aumenta la carga, y balancear el tráfico entre ellos.
- Se utiliza un Service tipo LoadBalancer para exponer la aplicación a los clientes externos.
- La aplicación se encarga de manejar las conexiones de los clientes, gestionar la lógica de la aplicación y actuar como intermediario entre los clientes y la base de datos.
- Maneja múltiples sockets simultáneamente mediante hilos o asincronía.
- Permite la comunicación entre clientes a través de un sistema de canales de chat:
    - Recibe mensajes de un cliente.
    - Valida y guarda los mensajes en la base de datos.
    - Retransmite los mensajes a los clientes destinatarios.

Servidor de Base de Datos (Google Cloud)

- Kubernetes y la aplicación se conectan a la base de datos usando una red privada para seguridad.
- Se crean replicas de la base de datos para alta disponibilidad y tolerancia a fallos.
- Almacena los datos persistentes, como:

    - Mensajes enviados entre los usuarios.
    - Información de los usuarios.

Cloud Storage (Google Cloud)

- Se utiliza para respaldar los mensajes de chat y los archivos multimedia enviados por los usuarios.

SLA para la Aplicación de Chat

Disponibilidad del Servicio

El servicio estará disponible al menos el 99.9% del tiempo en un mes calendario.

Tiempo de Respuesta del Servidor

El servidor responderá a las solicitudes de conexión en menos de 300 ms en al menos el 95% de los casos.

Recuperación ante Fallas

En caso de una interrupción del servicio, el tiempo máximo para la restauración completa será de 60 minutos.

Mantenimiento Programado

El mantenimiento programado no excederá las 2 horas por mes y se notificará al cliente con al menos 48 horas de anticipación.

SLO para la Aplicación de Chat

SLO de Disponibilidad

- Objetivo: Mantener el servicio operativo al menos el 99.9% del tiempo mensual.
- Métrica: Tasa de tiempo activo basada en los logs del sistema.

SLO de Tiempo de Respuesta

- Objetivo: Responder a solicitudes en menos de 300 ms para el 95% de las solicitudes.
- Métrica: Latencia medida por un sistema de monitoreo como Prometheus o Google Cloud Monitoring.

SLO de Manejo de Carga

- Objetivo: Manejar hasta 1,000 usuarios concurrentes sin degradación notable del rendimiento.
- Métrica: Tasa de errores y tiempo de respuesta promedio bajo condiciones de carga.

SLO de Persistencia de Mensajes

- Objetivo: Garantizar que los mensajes se escriban en la base de datos en menos de 50 ms en el 99% de los casos.
- Métrica: Tiempos de escritura medidos desde el backend.

SLO de Recuperación

- Objetivo: Restaurar el servicio en un máximo de 60 minutos tras una interrupción.
- Métrica: Tiempo transcurrido desde la detección del problema hasta la restauración.