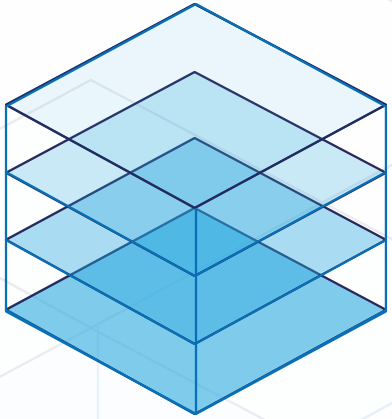


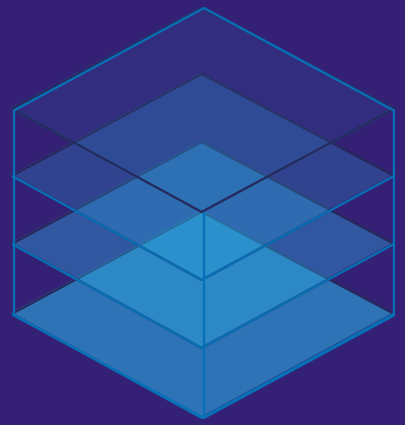
MISO

Maestría en Ingeniería de Software

# Entrega 2 Sistema Conversión Cloud

## Despliegue Básico en la Nube Pública

Camilo Ramírez Restrepo
Laura Daniela Molina Villar
Leidy Viviana Osorio Jiménez
Tim Ulf Pambor
Shadit Perez

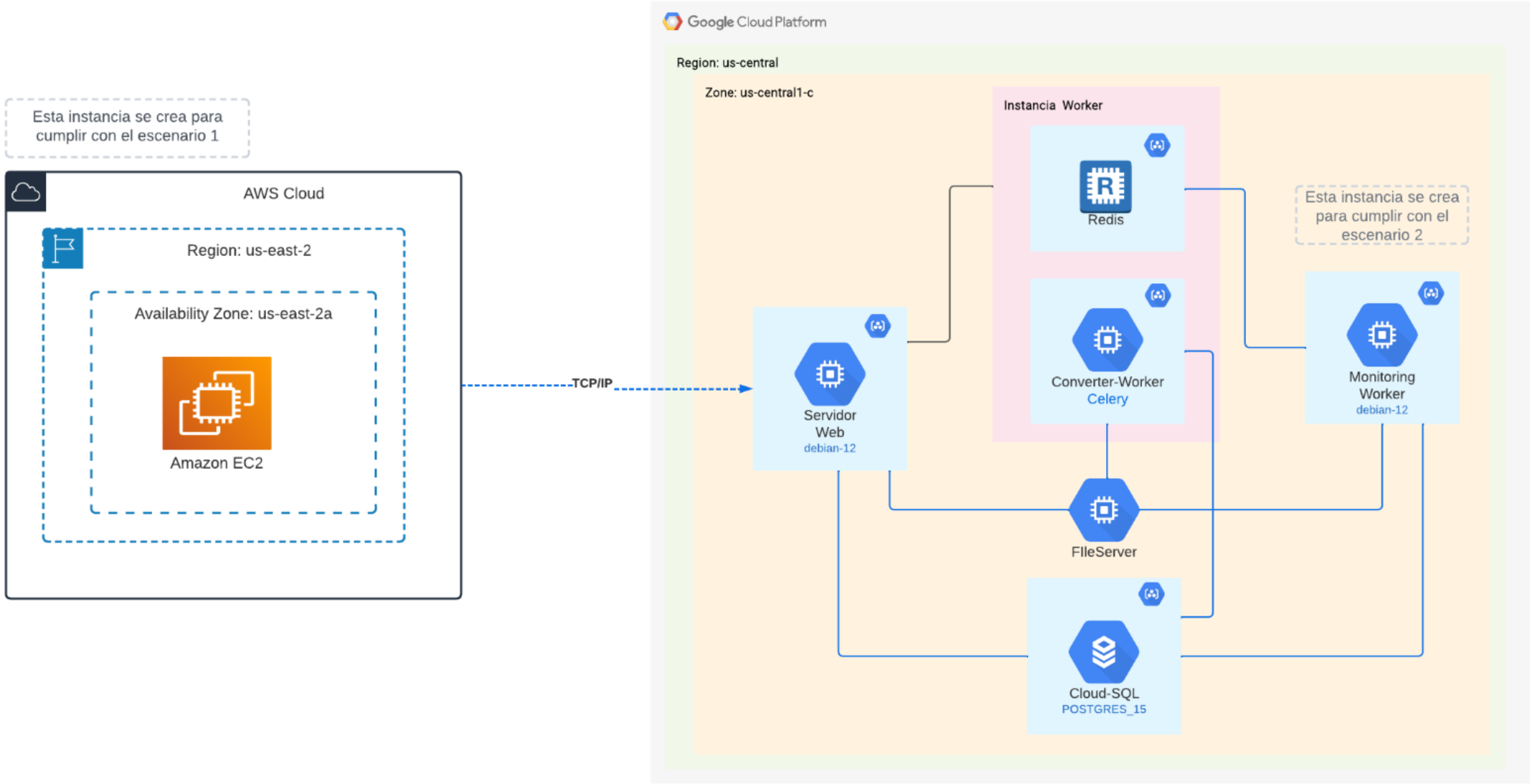


MISO

Maestría en Ingeniería de Software

# Arquitectura

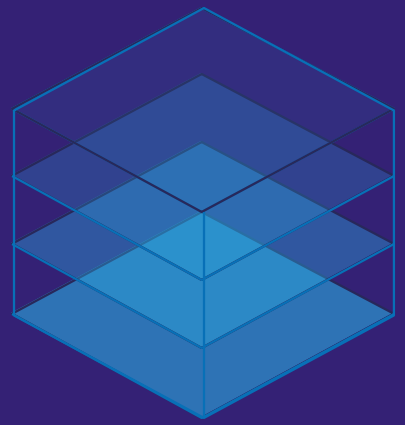
# Arquitectura



# Configuración del sistema

- **Servidores EC2:**
  - Instancia tipo t2.medium
  - Capacidad de CPU: Cada instancia t2.medium proporciona 4 vCPUs, lo que permite ejecutar múltiples instancias en paralelo para simular usuarios concurrentes.
  - Sistema Operativo: Se utilizará el sistema operativo Debian 12 en las instancias EC2 para ejecutar las pruebas.
  - Configuración de Red: Las instancias EC2 proporcionan una conexión de red hasta 10Gbps para hacer frente al ancho de banda necesario para enviar los videos.
  - Se instala sobre la instancia JMeter

- **Entorno GCP**
  - Google Cloud Monitoring.
  - Compute Engine utilizando una instancia de e2-highcpu-2 con Debian 12. Este servidor tiene etiquetas para permitir el tráfico HTTP y recibe información relacionada con la base de datos y el servidor de archivos.
  - Instancia worker con dos contenedores, un contenedor con Redis, otro contenedor con Celery
  - Compute Engine llamado "Fileserver". Este servidor utiliza una instancia de máquina virtual e2-micro con el sistema operativo Debian 12.
  - Compute Engine llamado "monitoring-worker, con una instancia e2-highcpu-2. Este servidor también recibe información relacionada con la base de datos y el servidor de archivos.
  - Cloud SQL Postgress, esta base de datos se crea con 1 vCPU y 4 GB de memoria
- **Aplicación Flask**



MISO

Maestría en Ingeniería de Software

# Experimento 1 – Agregar tarea a la cola

# Experimento 1

El objetivo de este plan es evaluar la capacidad de la aplicación Cloud conversión tool y su infraestructura de soporte en un entorno tradicional, para determinar sus máximos aceptables. El objetivo es comprender cómo la aplicación responde a diferentes niveles de carga de usuarios y cuál es su capacidad máxima.

- Configuración Aws y Jmeter: Documento Configuración
- Resultados: Datos Finales

# Experimento 1

## Métricas

**Throughput:** cantidad de peticiones procesadas por minuto.

**Tiempo de respuesta (P95):** percentil 95% de tiempo máximo que tarda la aplicación en procesar una petición

**Tiempo de respuesta (P99):** percentil 99% de tiempo máximo que tarda la aplicación en procesar una petición

**Utilización de recursos:** monitoreo de la CPU, memoria y uso de red durante las pruebas.

## Criterios de aceptación

**Throughput:** La aplicación debe ser capaz de procesar 100 peticiones por minuto.

**Tiempo de respuesta:**

- El tiempo de respuesta de la aplicación en todos los escenarios de prueba no debe superar los 0.5 segundos en el 95% de las transacciones, por petición.
- El tiempo de respuesta en ningún escenario de prueba debe superar los 4 segundos en el 99% de las transacciones.

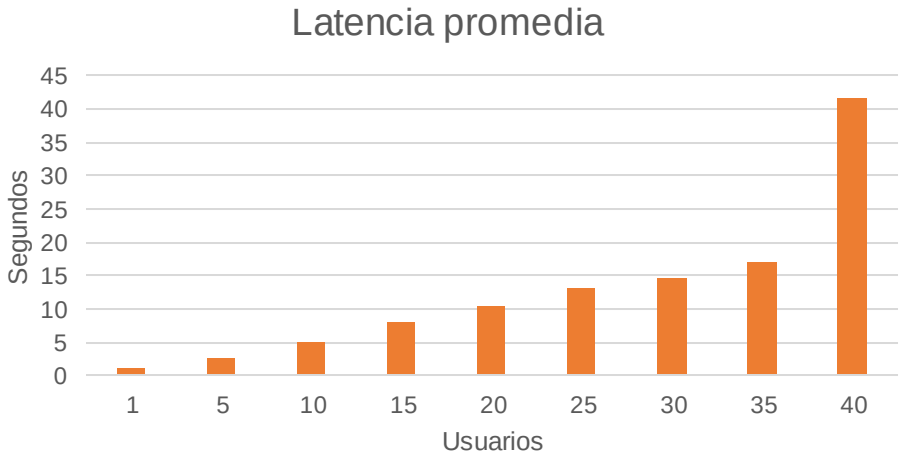
**Utilización de recursos:** Durante las pruebas con 100 peticiones concurrentes, la CPU del servidor alcanza un pico de 80% y la memoria se mantendrá < 80% de uso.

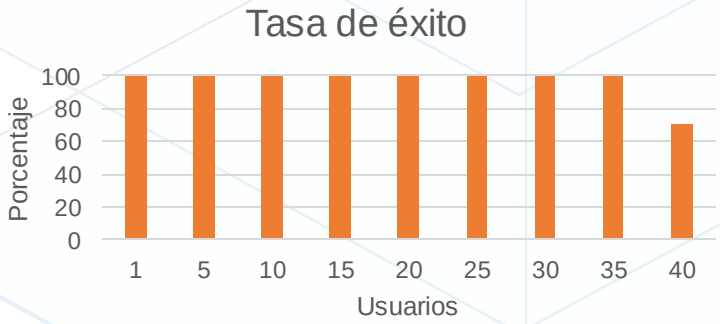
# Despliegue Arquitectura GCP

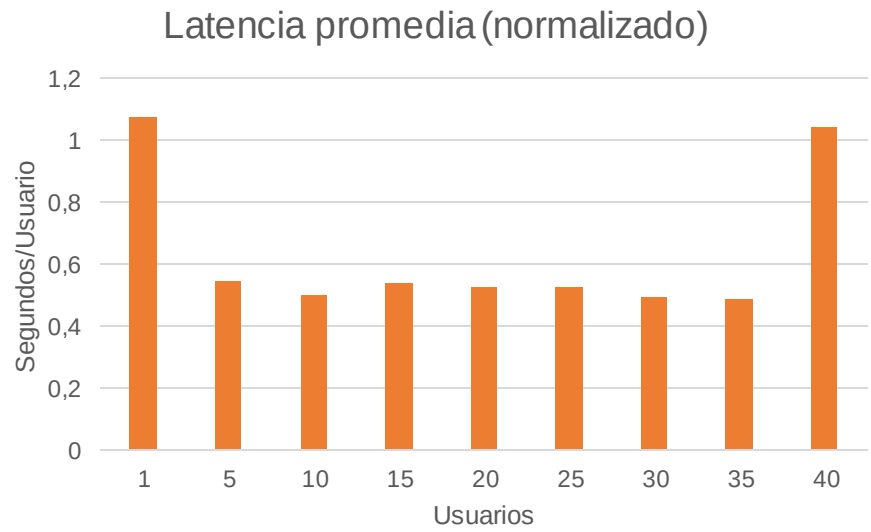
Configuración GCP: Documento Configuración

La latencia media aumenta linealmente con el número de usuarios. Esto se puede atribuirse principalmente al hecho de que el ancho de banda de red entre la máquina en AWS y GCP es fijo y no aumenta con los usuarios, lo que resulta en menos ancho de banda por usuario con un número creciente de usuarios. Este efecto puede reducirse normalizando por usuario.

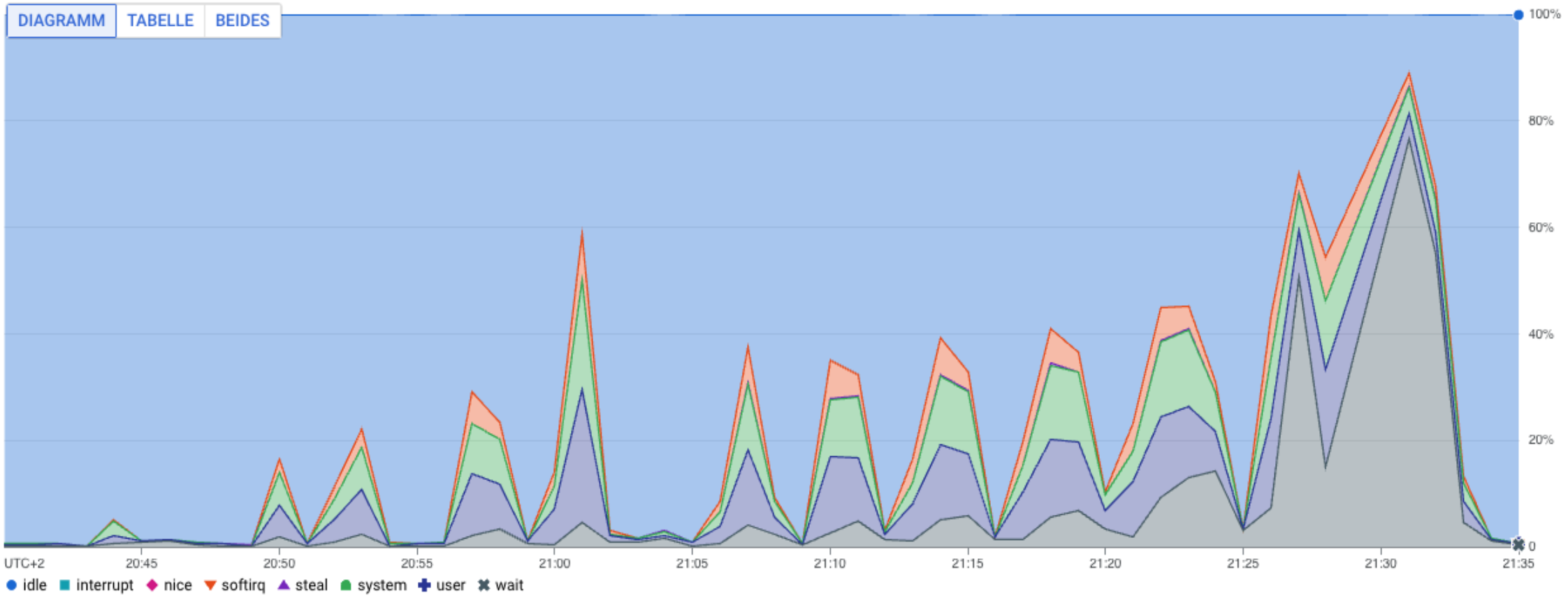
En la latencia normalizada se evidencia que con un solo usuario no se puede aprovechar al máximo las capacidades del sistema. Con más usuarios la latencia es aproximadamente es constante hasta el sistema colapsa con 40 usuarios, lo que también se evidencia en la tasa de éxito.



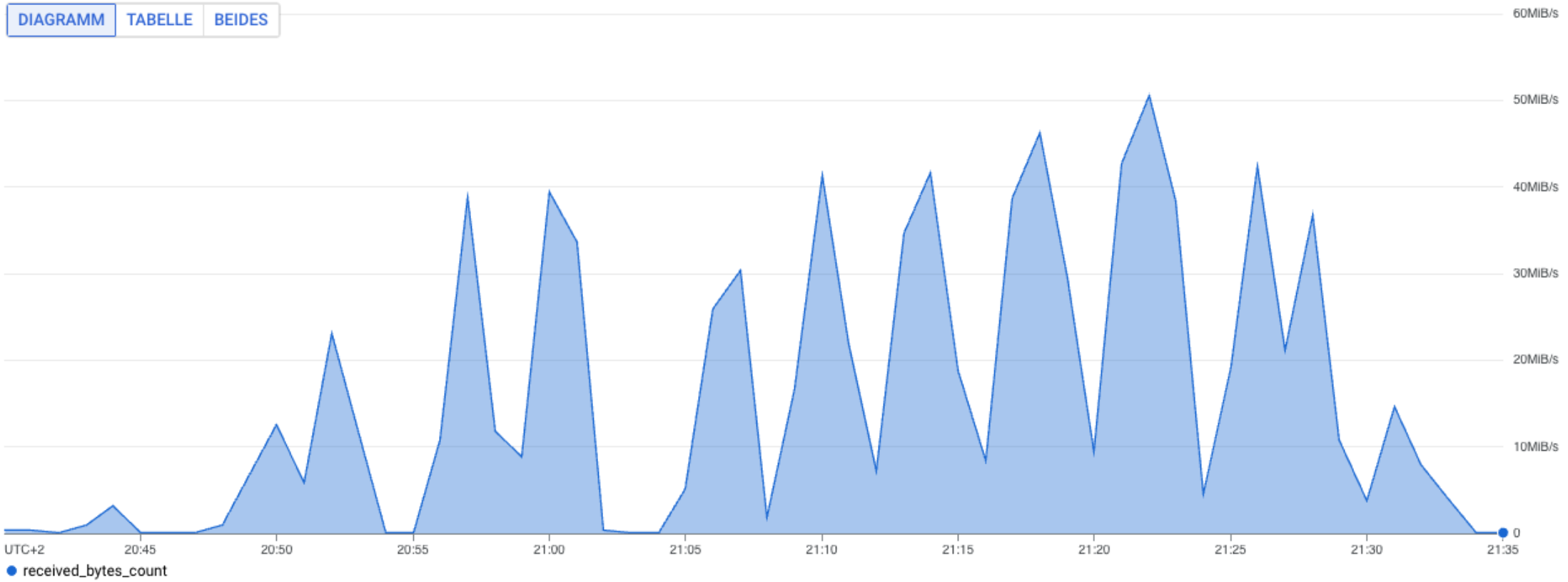



# Recursos utilizados Servidor Web - CPU

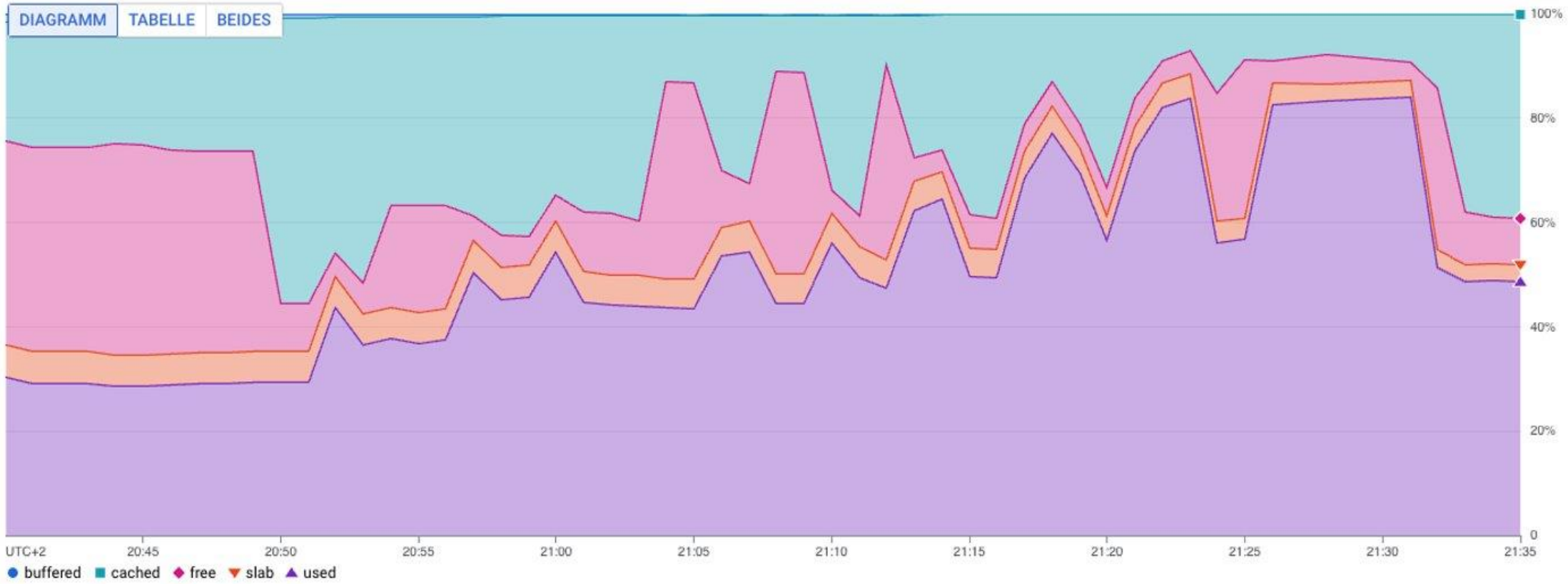



1 5 10 15 20 25 30 35 40 usuarios

# Recursos utilizados Servidor Web - Red Recibido



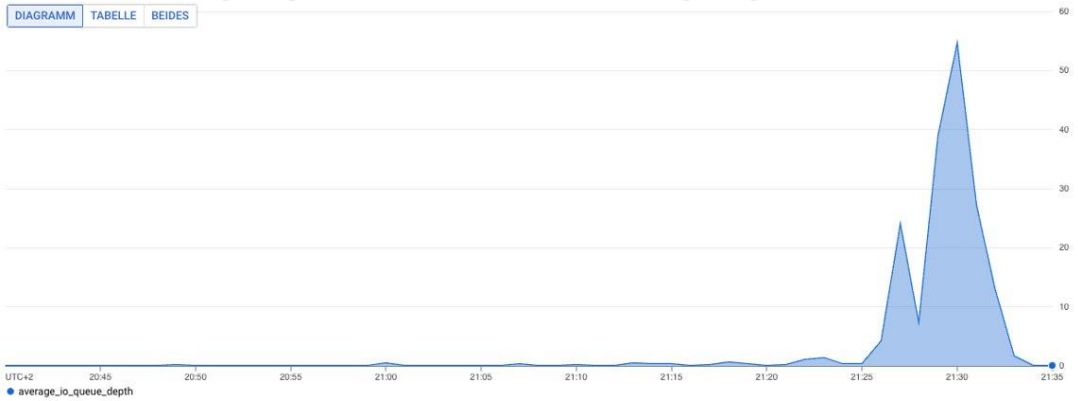
# Recursos utilizados Servidor Web - RAM



# Recursos utilizados Servidor Web

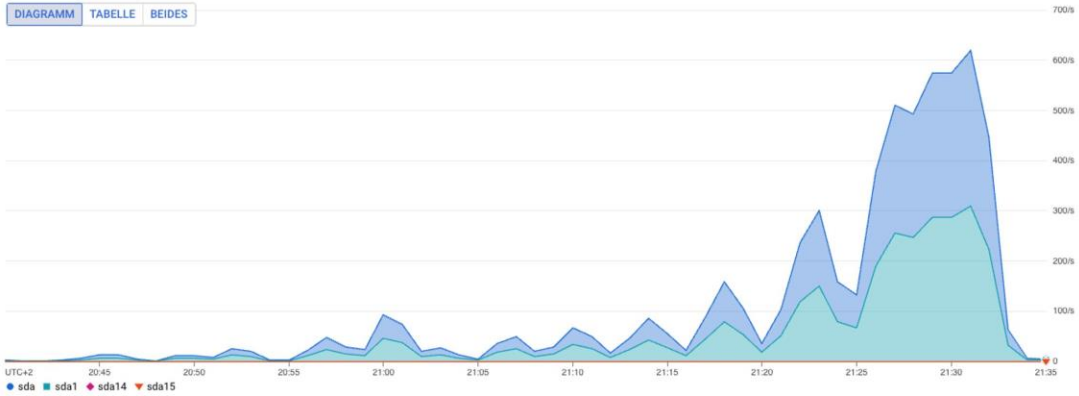
## IO Queue

Número de peticiones de disco esperando a ser procesados



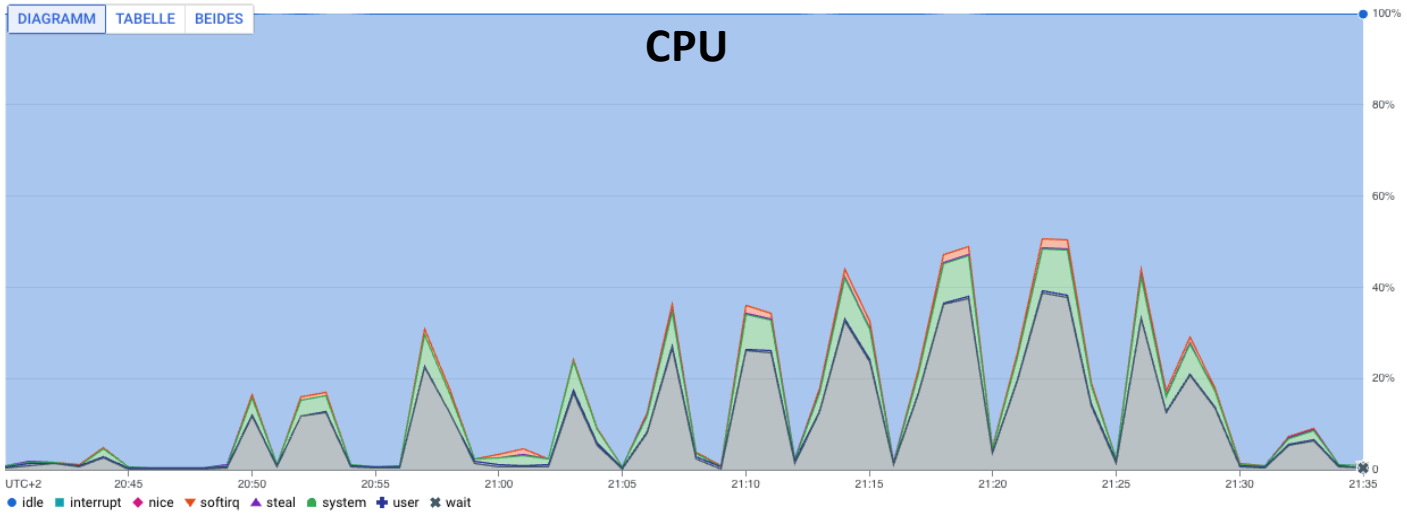

## Web IOPS

Número de peticiones de disco que se está procesando

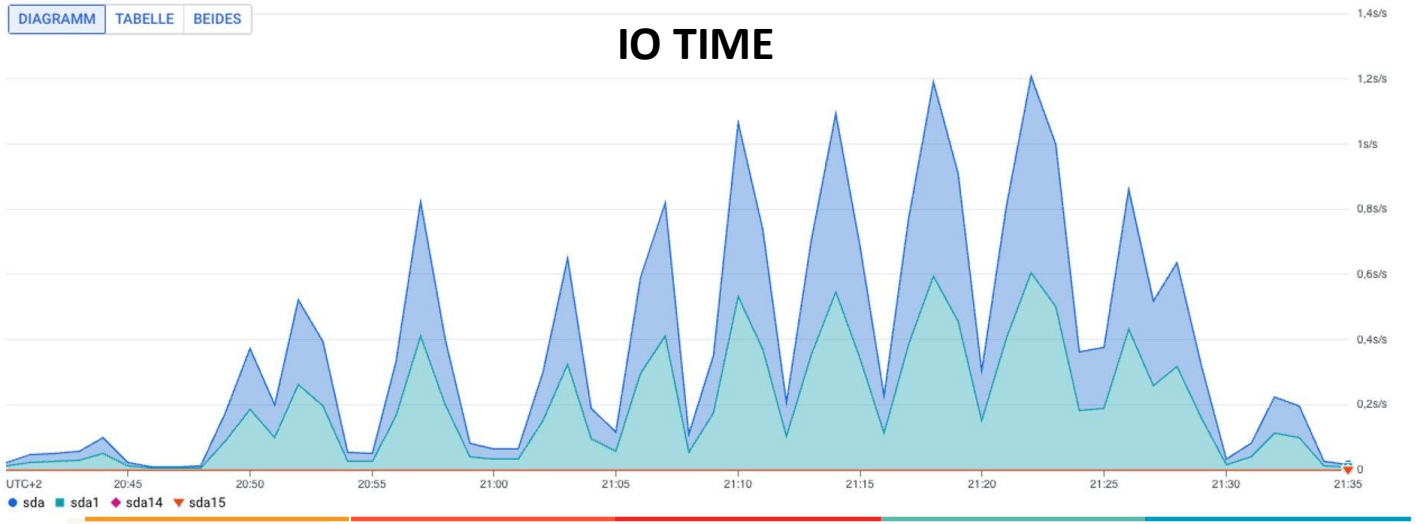



Se evidencia el colapso del sistema, porque con 40 peticiones no había suficiente memoria y el sistema empezó a hacer swap (usar el espacio de intercambio). Este proceso es lento y el disco no podía procesar las peticiones, lo que generó un pico en el número de peticiones de disco.

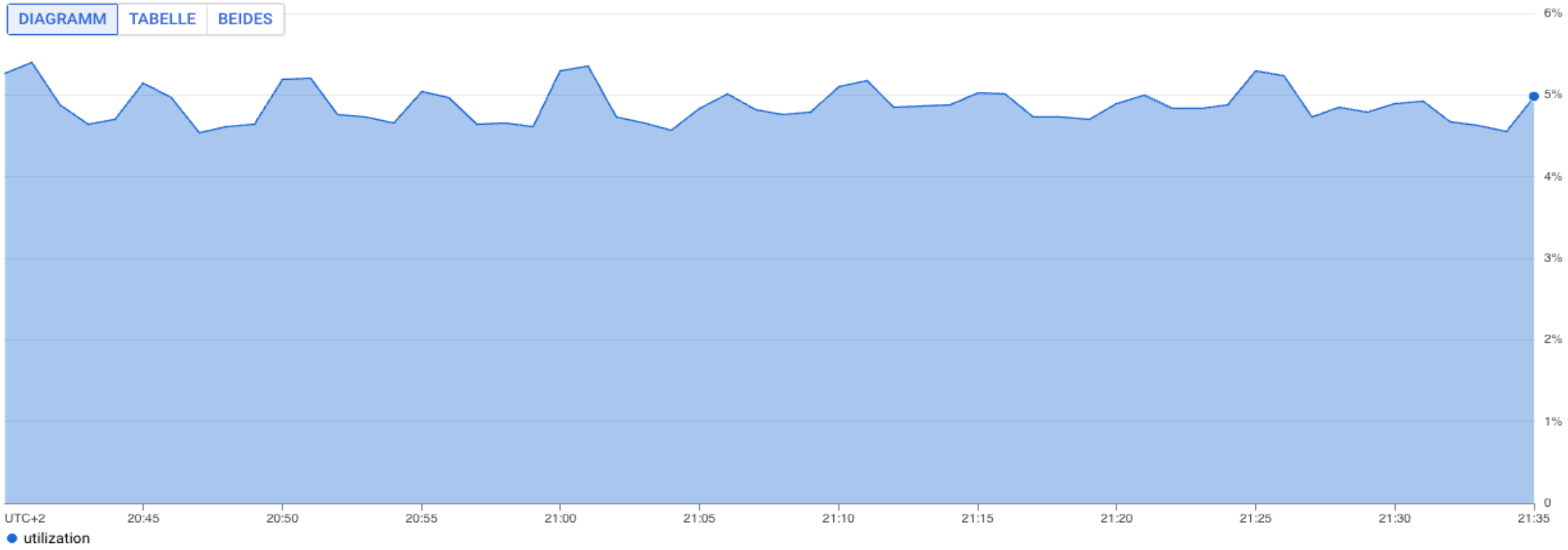
# Recursos utilizados Fileserver




En la gráfica se puede ver los picos porque recibe más videos en paralelo, pero no se muestra saturación del Fileserver con respecto a las operaciones de disco o CPU.

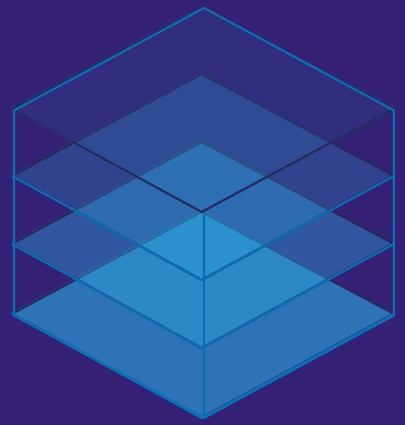


# Recursos utilizados Base de datos - CPU



# Reflexiones y conclusiones

- El uso de la CPU del servidor web supero el criterio de aceptación definido en 80%.
- El uso de la Memoria RAM del servidor web supera el criterio de aceptación definido en 80%.
- La aplicación no fue capaz de procesar 100 peticiones por minuto, porque con 40 peticiones no había suficiente memoria.
- El servidor Fileserver no muestra saturación con respecto a las operaciones de disco.



MISO

Maestría en Ingeniería de Software

# Experimento 2 – Conversión de videos

# Experimento 2

Análisis de desempeño del componente encargado de convertir videos, actualizar el estado en la base de datos y guardar el video convertido, de manera que el video quede listo para ser solicitado por el usuario.

• Configuración:

# Experimento 2

## Métricas

**Throughput**: Cantidad de peticiones por minuto en la conversión de videos.

**Tiempo de respuesta promedio:** Tiempo promedio que tarda la aplicación en procesar las tareas de conversión**.**

**Tiempo de respuesta (P95)**: Percentil 95% de tiempo máximo que tarda la aplicación en procesar una tarea.

## Criterios de aceptación

**Throughput:** Capacidad de procesar 100 peticiones por minuto.

**Tiempo de respuesta**

- No debe superar los 0.5 segundos en el 95% de las transacciones, por petición.
- No debe superar los 4 segundos en el 99% de las transacciones.

**Utilización de recursos:** Durante las pruebas con 100 peticiones concurrentes, la CPU del servidor alcanza un pico de 80% y la memoria se mantendrá < 80% de uso.

# Resultados

El experimento constó de 3 casos, que fueron ejecutados 5 veces cada uno:

- **Caso 1:** 5 peticiones concurrentes, 10 peticiones en total
- **Caso 2:** 10 peticiones concurrentes, 20 peticiones en total
- **Caso 3:** 20 peticiones concurrentes, 40 en total

Las métricas arrojadas en cada iteración se pueden observar en el siguiente link: Resultados escenario 2.xlsx

Igualmente, por cada iteración y caso se generaron 3 gráficas:

- Petición vs Tiempo de respuesta
- Uso CPU
- Uso RAM

Los resultados con sus respectivas gráficas se pueden ver en el siguiente enlace: Resultados escenario 2 con gráficas.docx

# Análisis de resultados

- Para todos los casos evaluados, **el uso de la memoria RAM cumple con el criterio de aceptación planteado** en la formulación del escenario (no exceder el 80% de su capacidad) ya que el pico máximo presentado en los 3 casos no fue superior al 6%, indicando que este no es un cuello de botella del sistema.

- **El uso de CPU se mantuvo alto en todos los casos**, oscilando entre 88% y el 89% de su capacidad, por lo tanto, no se cumplió con el criterio de aceptación planteado en la formulación del escenario (no exceder el 80% de capacidad), e indica que este es un cuello de botella en el proceso.

- **En ninguno de los 3 casos se cumplió con el criterio de aceptación planteado para el tiempo de respuesta de cada petición** (menor o igual a 40 segundos), por el contrario, aumentó cada vez que se realizaban más peticiones. Este comportamiento se debe a que la cola de tareas se llenaba cada vez más de peticiones, resultando en un tiempo de espera mayor.

- **El número de peticiones atendidas por minuto se mantuvo constante entre 0.79 y 0.81** en cada uno de los tres casos, dando con el incumplimiento del criterio de aceptación planteado para esta métrica (mínimo 100 peticiones por minuto) y reflejando la limitación de procesamiento para procesar peticiones adicionales.

# Reflexiones y conclusiones

- Todo el componente de conversión de videos del sistema, tal como se encuentra en su estado actual, representa un cuello de botella para el sistema, ya que la capacidad de procesamiento de la máquina virtual que lo aloja presentó máximos durante todo el experimento de manera constante, sin importar el caso probado, por lo cual debe ser un punto de mejora en el futuro.

- Los tiempos de respuesta a las peticiones fueron altos desde el principio del experimento ya que el conversor de tareas va recibiendo dichas tareas a medida de sus capacidades, dando como resultado un atascamiento de tareas cada vez que se agregan más y más a la cola.