Spanner, TrueTime & The CAP Theorem

Eric Brewer

Technical Report

https://research.google.com/pubs/pub45855.html, 2017

Resumen

Spanner es una base de datos global de alta disponibilidad de Google.
Gestiona datos replicados a gran escala, tanto en términos de tamaño de datos como de volumen de transacciones.
Asigna marcas de tiempo en tiempo-real globalmente consistente a cada dato escrito en él, y los clientes pueden hacer lecturas globalmente consistentes en toda la base de datos sin bloqueo.

Teorema CAP

Presentado como una conjetura por Eric Brewer en el año 2000 durante el PODC.
En 2002, Seth Gilbert y Nancy Lynch, publicaron una demostración formal de la conjetura, convirtiéndola en un teorema[1].

El teorema CAP dice que solo puede tener dos de las tres propiedades deseables de:

- C: Consistencia(serializabilidad).
- A: 100% de disponibilidad, tanto para lecturas como para actualizaciones.
- P: tolerancia a particiones de red.

1.- *ACM SIGACT News*, Volume 33 Issue 2 (2002), pg. 51-59.

Teorema CAP

De acuerdo al teorema se pueden tener tres tipos de sistemas: CA, CP y AP. Teniendo en cuenta que no es forzoso tener 2 de 3 propiedades, se pueden tener sistemas de 1 o cero propiedades.

En ausencia de falla de la red, es decir, cuando el sistema distribuido se está ejecutando normalmente, se puede satisfacer tanto la disponibilidad como la consistencia.

Explicación

En presencia de una partición, tenemos dos opciones: consistencia o disponibilidad.

- Al elegir la consistencia, el sistema devolverá un error o un tiempo de espera si no se puede garantizar que la información particular esté actualizada debido a la partición de la red.

- Al elegir la disponibilidad, el sistema siempre procesará la consulta e intentará devolver la versión disponible más reciente de la información, incluso si no puede garantizar que esté actualizada debido a la partición de la red.

Spanner y las particiones

Spanner afirma ser consistente y altamente disponible, lo que implica que no hay particiones.
Sin embargo las particiones ocurren y técnicamente es un sistema CP.

Spanner asume su alta disponibilidad de las siguientes maneras:

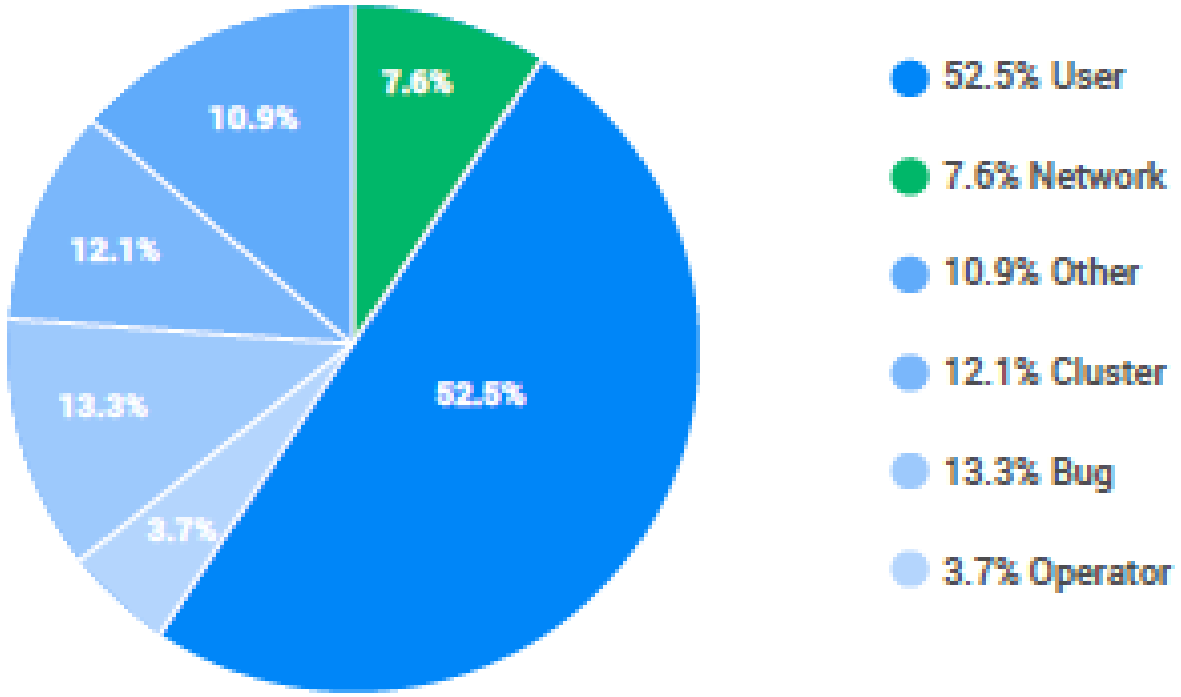
- Teniendo una disponibilidad muy alta , con 1 falla en 10^5 (downtime por año de 5.26 min)

De tal manera que los usuarios ignoren las fallas.

- Haciendo un refinamiento del origen de las fallas, ya que no todas se deben a Spanner.
- Para Spanner no todas las fallas se deben a particiones, es de este modo que Spanner se asume CA
- Red Global Privada de Google

Incidentes en Spanner

Un incidente es un inesperado evento, pero no todos los incidentes son interrupciones.



La Red Global Privada de Google

La red de área amplia de Google, y su experiencia en el área operativa limitan en gran medida las particiones, y así permitir alta disponibilidad.

- Spanner no se ejecuta en Internet público.
- Spanner fluye solo a través de enrutadores y enlaces controlados por Google

Aunque la red puede reducir en gran medida las particiones, Google define regiones para proporcionar un equilibrio entre latencia y tolerancia a fallas.

Funcionamiento de Spanner

Spanner utiliza el protocolo de dos fases (2PC) y bloqueo estricto de dos fases para garantizar el aislamiento y fuerte consistencia.

El protocolo 2PC requiere que todos sus miembros se encuentren disponibles.
Spanner mitiga esto al hacer que cada miembro sea un grupo de Paxos, asegurando así que cada miembro de 2PC sea altamente disponible incluso si algunos de sus participantes de Paxos están inactivos.
Los datos se dividen en grupos que forman la unidad básica de colocación y replicación.

Funcionamiento de Spanner

Spanner elige C sobre A por las siguientes opciones:

• Uso de grupos de Paxos para lograr consenso sobre una actualización; si el líder no puede mantener un quórum debido a una partición, las actualizaciones están estancadas y el sistema no está disponible(Teorema CAP). Finalmente, puede surgir un nuevo líder, pero eso también requiere una mayoría.

• El uso de la confirmación en dos fases para transacciones entre grupos también significa que una partición deLos miembros pueden evitar los commits

Funcionamiento de Spanner

Para el caso de lecturas de Snapshots Spanner mantiene varias versiones a lo largo del tiempo, cada una con una marca de tiempo y, por lo tanto, puede responder las lecturas de Snapshots con la versión correcta.

Una lectura de Snapshot funcionará si:

1. Hay al menos una réplica para cada grupo en el lado inicial de la partición.
2. La marca de tiempo está en el pasado para esas réplicas.

TrueTime

TrueTime es un reloj global sincronizado con un error acotado distinto de cero: devuelve un intervalo de tiempo que garantiza que contiene la hora real del reloj durante algún tiempo durante la ejecución de la llamada.

Si dos intervalos no se superponen:
Sabemos que las llamadas se ordenaron definitivamente en tiempo real.

Si dos intervalos se superponen: No sabemos el valor actual de orden.

True Time

Spanner obtiene su consistencia externa por medio de TrueTime.

La consistencia externa invariante de Spanner es que para dos transacciones, T1 y T2 (incluso si están en lados opuestos del mundo):

Si T2 comienza a hacer commit después de que T1 termina de hacer commit, entonces la marca de tiempo para T2 es mayor que marca de tiempo para T1.

TrueTime garantiza que el reloj de Spanner se mantenga invariante.

Conclusión

Spanner afirma razonablemente ser un sistema de "CA” eficaz, usa el protocolo 2PC para lograr la serialización, pero usa TrueTime para la consistencia externa, lecturas consistentes sin bloqueo y Snapshots consistentes.