YPF Ruta

Índice

Introducción

En este trabajo se desarrolla **YPF Ruta**, un sistema que permite a las empresas centralizar el pago y el control de gasto de combustilble para su flota de vehículos.
Las empresas tienen una cuenta principal y tarjetas asociadas para cada uno de los conductores de sus vehículos. Cuando un vehículo necesita cargar en cualquiera de las 1600 estaciones distribuídas alrededor del país, puede utilizar dicha tarjeta para autorizar la carga; siendo luego facturado mensualmente el monto total de todas las tarjetas a la compañía.

Aplicaciones

Server

El servidor consiste de un sistema distribuido en el que existen tres tipos diferentes de clústers de nodos:

- *Surtidores* en una estación.
- *Nodos suscriptos a una tarjeta.*
- *Nodos líderes de tarjetas* que forman una *cuenta.*

Entidades que participan:

- **Surtidores.** Los surtidores corresponden a las máquinas interconectadas de manera *local* en una estación.
- **Estaciones/Nodos.** Los nodos representan estaciones de YPF. Dentro de una estación, uno de los surtidores tiene la responsabilidad de llevar a cabo la función del nodo en el sitema global.

Hay tres tipos de nodos:

- **Suscriptor (tarjeta).** Los nodos suscriptores mantienen informados a sus pares (otros nodos suscriptos a la misma tarjeta) sobre las actualizaciones al registro de la tarjetas a la que suscriben. Un nodo puede estar suscripto a varias tarjetas.
- **Líder (tarjeta).** Los nodos líder *lideran* un clúster de nodos suscriptores a una tarjeta; ésto es: tienen la responsabilidad de intercomunicar a los nodos del clúster y a su vez de informar sobre actualizaciones de la tarjeta al *nodo cuenta* cuando este así lo solicite. Un nodo líder es también un nodo suscriptor.
- **Cuenta.** Los nodos cuenta se comunican con un nodo líder de cada una de las tarjetas que le pertenecen a la cuenta. Un nodo cuenta **no** puede ser el líder de un clúster de nodos suscriptos a una tarjeta.

Cliente

El único cliente (fuera del servidor de YPF) es el **administrador**. El administrador puede

- Limitar los montos disponibles en su cuenta.
- Limitar los montos disponibles en las tarjetas de la cuenta.
- Consultar los saldos de las cuentas.
- Consultar los saldos de las tarjetas de la cuenta.
- Realizar la facturación de la cuenta.

Arquitectura del servidor

Como ya se mencionó, el servidor está implementado de manera distribuida. El foco principal del diseño de la arquitectura está en reducir la cantidad de mensajes entre nodos que tienen viajar en la red, partiendo de la arquitectura trivial: un grafo completo, con réplicas de la información del sistema en todos los nodos.

Se pueden hacer varias optimizaciones a partir de algunas observaciones del *modelo de negocio* del sistema. Existe localidad con respecto al posicionamiento geográfico de las estaciones; un conductor que aparece en una estación probablemente vuelva a aparecer en estaciones cercanas, y probablemente no aparazca en una estación en la otra punta del país (o al menos no con frecuencia significativa). Una forma de optimizar la comunicación entre nodos sería entonces tenerlos separados por cuentas: cada nodo tendría una réplica de la información de todas las tarjetas de la cuenta a la que pertenece y sólo debería comunicar a los otros nodos del clúster de la cuenta respecto de las actualizaciones de la misma.
El problema con esto último es que una empresa grande, con muchas tarjetas y muchos conductores a lo largo del país; tendría réplicas innecesarias: un conductor que vive en Salta probablemente no use una estación en Santa Cruz, sin embargo, si uno de sus compañeros de trabajo así lo hace, entonces el registro de su tarjeta estaría replicado en la estación de Santa Cruz.
La solución que se encontró es la de dividir los clústers por tarjeta y no por cuenta. Ahora bien, como también necesitamos centralizar la información de todas las tarjetas pertenecientes a una cuenta, surge la necesidad de los nodos *cuenta*. Para minimizar la comunicación de los nodos cuenta con los nodos de las tarjetas que le pertenecen, el rol de comunicador se centraliza en los nodos *líder tarjeta*.

Tipos de clúster

A continuación se explican más en profundidad cada uno de los tipos de clúster que se mencionaron.

Clúster de surtidores. Los surtidores en una estación están conectados de manera local y se encargan de mantener actualizado al surtidor líder del clúster para que este ejerza la función de nodo estación en el sistema global.

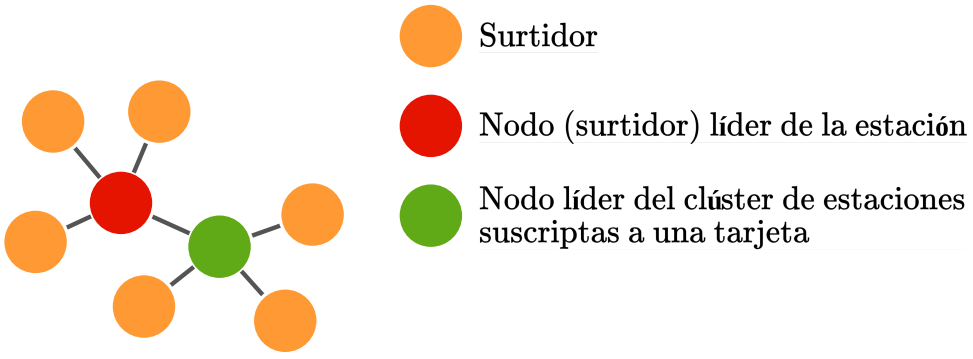

Figura 1: Dos estaciones, con cuatro surtidores cada una.

Clúster de nodos suscriptos a una tarjeta. Los nodos suscriptos a una tarjeta informan a sus pares de las actualizaciones en los registros de las tarjetas a las que suscriben. Hay un líder del clúster y los *súbditos* se encargan de elegirlo al principio de la ejecución y en caso de que el mismo deje de estar activo.

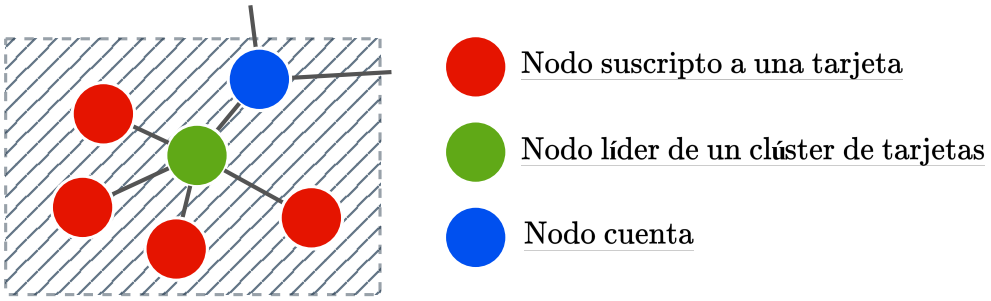

Figura 2: Clúster de nodos suscriptos a una tarjeta.

Clúster de cuenta. El clúster de nodos líderes de tarjetas tienen su propio líder: el *nodo cuenta*. Dentro de éste clúster se mantiene actualizado al nodo cuenta ante cualquier cambio en alguno de los registros de las tarjetas que conforman la cuenta. Los *súbditos* eligen un líder al principio de la ejecución y en caso de que el mismo deje de estar activo. Las actualizaciones son comunicadas sólo cuando el nodo cuenta así lo solicita.

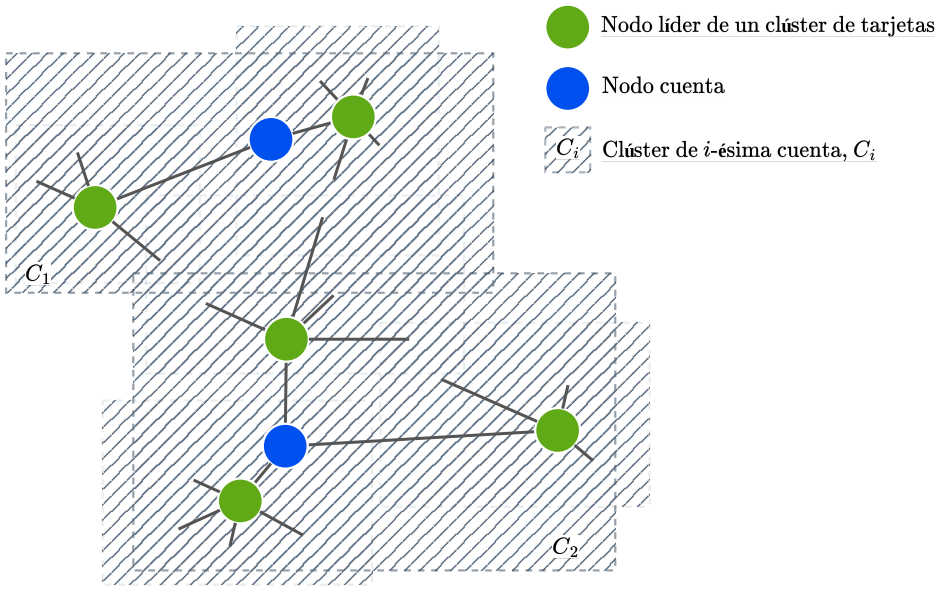

Figura 3: Clúster de cuenta.

Vista de águila

Cabe recalcar que los nodos cuenta (azules) no pueden ser nodos líderes de tarjetas (verdes). Por otro lado, los nodos líder tarjeta (verdes) siempre son suscriptores a la tarjeta que lideran (rojos); más aún, todos los nodos del sistema cumplen mínimamente con el rol de suscriptor.
En resumen:

- los nodos cuenta y los nodos líder tarjeta ejecutan también la responsabilidad de nodos suscriptores,
- los nodos líder tarjeta son, en particular, suscriptores a la tarjeta que lideran (también pueden estar suscriptos a otras tarjetas)
- y los nodos cuenta no pueden ser nodos líder. Si un nodo líder tarjeta asume la responsabilidad de ser un nodo cuenta, entonces tiene que delegar la responsabilidad de líder tarjeta a otro nodo del clúster de suscriptores a la tarjeta; de donde surge una última regla:
- un clúster de nodos suscriptos a una tarjeta tiene que cumplir con una cantidad mínima. En caso de no hacerlo, se invita a un nodo del sistema a suscribirse a la tarjeta.

Agrupando los niveles de clúster (y obviando los surtidores), la vista general de una posible configuración del sistema se ve de la siguiente forma:

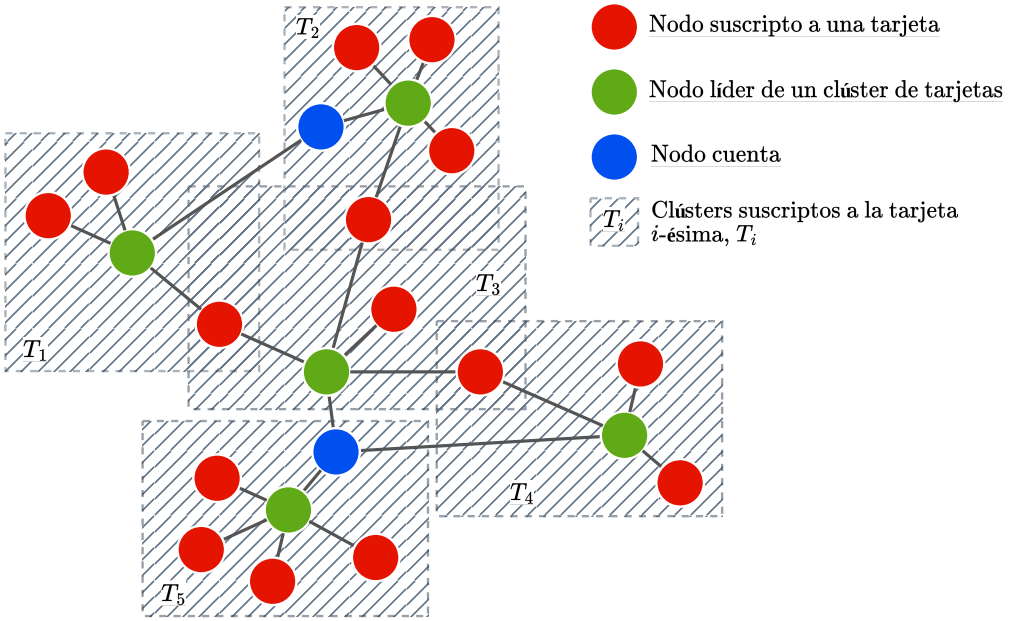

Figura 4: *Overview* del sistema distribuido global.

Paseo por varios casos de uso

1. *Un conductor usa su tarjeta por primera vez en el surtidor de una estación.*

1. El conductor le da su tarjeta al cajero, que usa la terminal de cobro de la columna del surtidor que usó para cargar nafta. El surtidor necesita saber si el cobro puede o no ser efectuado. Para ello revisa la información de la tarjeta, como no la tiene en guardada, la solicita. El mensaje utilizado para la solicitud es delegado al nodo central de la estación. En este punto ya nos encontramos en el sistema distribuido de estaciones.

2. Una vez que el nodo estación recibe el mensaje con la solicitud de información de la tarjeta, envía el mensaje a sus estaciones vecinas, y así lo hacen estas últimas, propagando el mensaje como un *virus*. El mensaje que se propaga contiene, además de la solicitud en sí misma, las direcciones a las que ya se propagó; para evitar demasiados mensajes redundantes.
   Como esta es la primera vez que la tarjeta es utilizada, ningún nodo va a contestar con su información y por lo tanto el nodo de la estación original genera el registro de la tarjeta.
   En caso de que el mensaje llegue a un nodo cuenta al que le pertenece la tarjeta, el mismo puede rápidamente contestar si la tarjeta ya existe o no.

3. Una vez generado el registro, se deben tener un mínimo de nodos suscriptos a la misma, un nodo cuenta líder y un nodo cuenta generado para la tarjeta. Como ningún nodo cuenta contestó, y no ningún otro nodo tenía la tarjeta, se generan ambos. Además se invitan a la lista de suscripción al top N nodos más cercanos para replicar en ellos la información del registro de la misma, y también porque el sistema no acepta un nodo que sea cuenta y líder tarjeta en simultáneo.

4. Con todas las condiciones del sistema distribuido en orden, la estación procede a realizar el cobro para luego actualizar a los suscriptores de la tarjeta (que acaban de generarse).

2. *Un conductor usa su tarjeta en el surtidor de una estación a la que frecuenta.* Si un conductor utiliza su tarjeta en una estación a la que va con frecuencia, entonces ésta estación ya tiene cargado el registro de la tarjeta. Aún así, se necesita saber si a la cuenta le queda monto para realizar el cobro, para esto se procede de la siguiente manera:

1. El surtidor envía la consulta de saldo de cuenta al nodo líder de la estación.
2. El nodo líder de la estación envía la consulta de saldo de cuenta al nodo líder tarjeta.
3. El nodo líder tarjeta envía la consulta al nodo cuenta.
4. El nodo cuenta consulta las actualizaciones de los nodos líder del resto de tarjetas, computa la respuesta y se la envía al nodo líder tarjeta que le hizo la consulta.

3. *Un conductor usa su tarjeta en una nueva estación nueva, habiéndola usado en otras.* Si un conductor usa su tarjeta en una nueva estación, es decir, en una estación en la que todavía no la había usado, entonces la estación no va a contar con el registro de la tarjeta y por tanto propagará la consulta como en el caso 1. Ésta vez si va a recibir una respuesta de una de los nodos que estén suscriptos a la tarjeta, por lo que

1. gestiona el cobro como en 2,

2. envía el mensaje de *suscripción*,
3. invita a sus nodos cercanos,
4. y actualiza a la lista de nodos suscriptos por el cobro realizado.

Time-to-leave (TTL)

Supongamos que un conductor utiliza siempre su tarjeta en las estaciones cercanas a su casa en Córdoba. Si el conductor se va, de manera espontánea, de viaje a Formosa (por trabajo, si no no usaría la tarjeta de la empresa…), entonces probablemente utilice varias estaciones entre Córdoba y Formosa. Cuando vuelva de de su jornada laboral (o de sus vacaciones si no hizo un buen uso de la tarjeta), no volvería a usar su tarjeta en las estaciones en las que la usó para viajar a Formosa.
Sería un desperdicio de recursos—mínimos en memoria, pero sí significativos para la comunicación en la red—tener un nodo suscrito a la lista de una tarjeta si éste no fuera a volver a ser utilizado.
Por esto se introduce el campo **TTL** en los registros de las tarjetas. Si un nodo es actualizado de manera *externa*, es decir, se actualiza la información de un registro de una de sus tarjetas sin que la tarjeta haya efectuado la carga en esa estación; un número mayor a TTL veces, entonces se elimina de la lista de suscripción de la tarjeta. De esta forma, evitamos que con el paso del tiempo el sistema gaste recursos actualizando a estaciones a las que no les debería importar el registro de una tarjeta.

Node failure recovery

Hasta ahora sólo consideramos los casos felices del funcionamiento del sistema, pero en la realidad los nodos pueden fallar. A continuación detallamos lo que pasaría en caso de que cada uno de los distintos tipos de nodos falle, a partir de la siguiente configuración arbitraria del sistema:

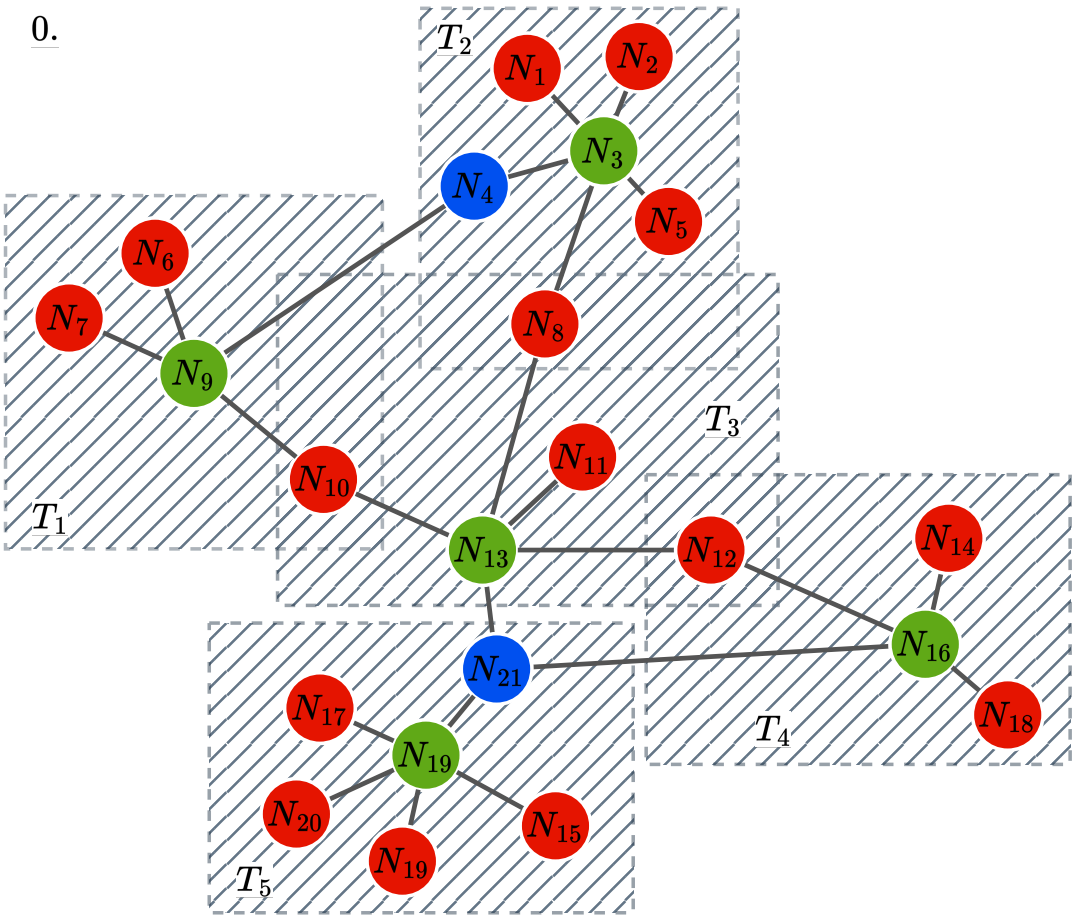

Figura 5: Estado inicial del sistema. Dos cuentas, una con las tarjetas T_1 y T_2 y la otra con T_3, T_4 y T_5.

Se cae N_1: nodo suscriptor. Que se caiga un nodo suscriptor no representa un problema demasiado grande. En este caso la estación va a tener que guardarse las actualizaciones a la tarjeta, sin poder realizar las consultas de suficiencia de saldo en las mismas o en sus cuentas. No hay nada más que hacer puesto que la única responsabilidad del nodo suscriptor es comunicar al nodo líder y no hay nunca posibilidad de que esto así ocurra.
Cuando el nodo vuelve a la vida, tiene que preguntar quién es el leader, enviarle sus actualizaciones de la tarjeta a la que el clúster suscribe para que este actualice al resto de nodos en el clúster y al nodo recuperado, a este último con la agregación de las actualizaciones que acaba de enviar y las que se efectuaron durante su baja.

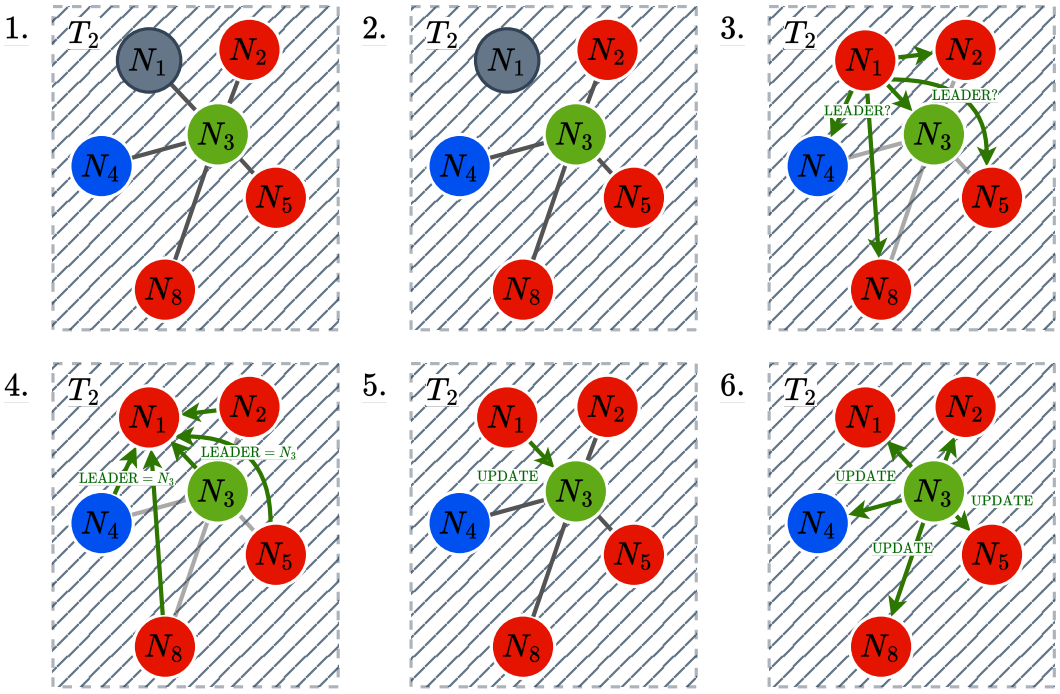

Figura 6: *Recovery* de la falla en N_1.

Se cae N_{13}: nodo líder tarjeta. Que se caiga un nodo líder tarjeta representa un mayor problema ya que su responsabilidad es la de centralizar la información generada por un clúster sobre una tarjeta y estar disponible para cuando el nodo cuenta al que pertenece la tarjeta consulte la información de la misma. En este caso se usa el algoritmo de elección de líder *Bully* y se comunica al nodo cuenta sobre el líder elegido.
En caso de que sea el nodo cuenta quien se entera de la baja del nodo líder, simplemente envía un mensaje de elección de líder, sin participar de la elección, y recibir el líder elegido al final de la misma.

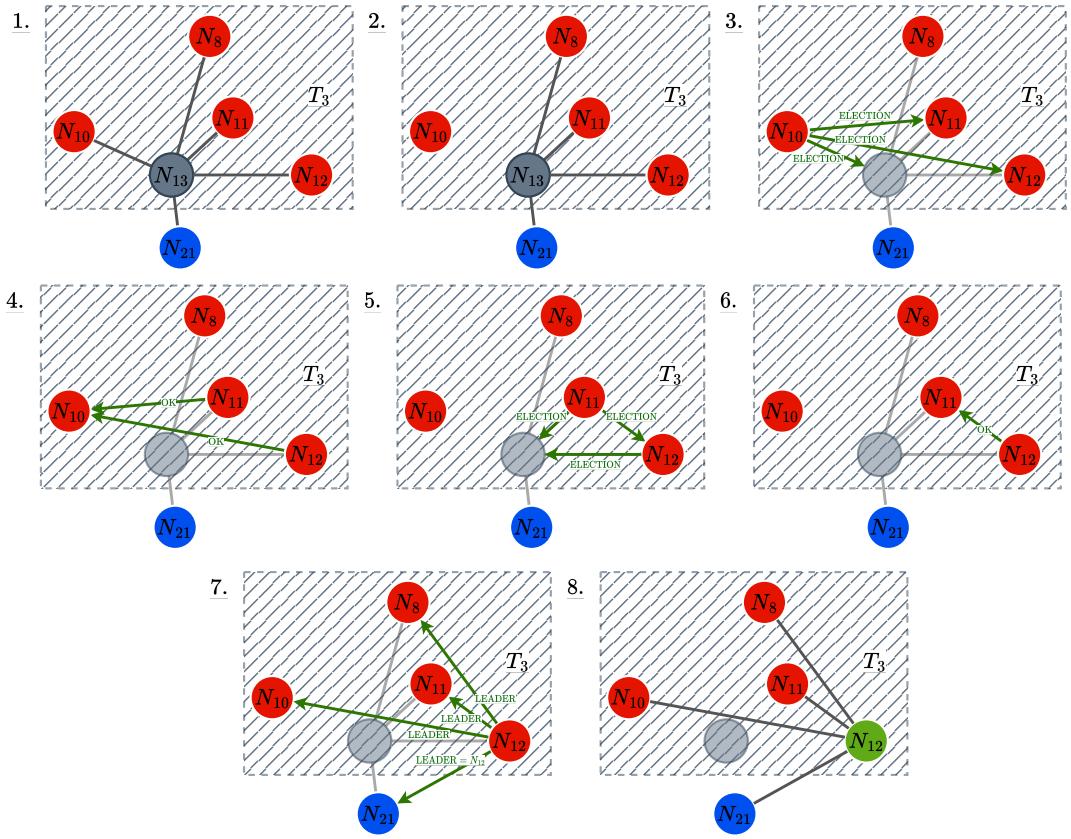

Figura 7: *Recovery* de la falla en N_{13}.

Se cae N_{22}: nodo cuenta. Este es el caso más complicado, ya que el nodo cuenta es el tipo de nodo con mayor responsabilidad del sistema. La dinámica de recovery de este caso es muy similar a la de cuando se cae un nodo líder tarjeta, sumando una re-elección del nodo líder tarjeta ya que las responsabilidades líder tarjeta y cuenta no son compatibles.

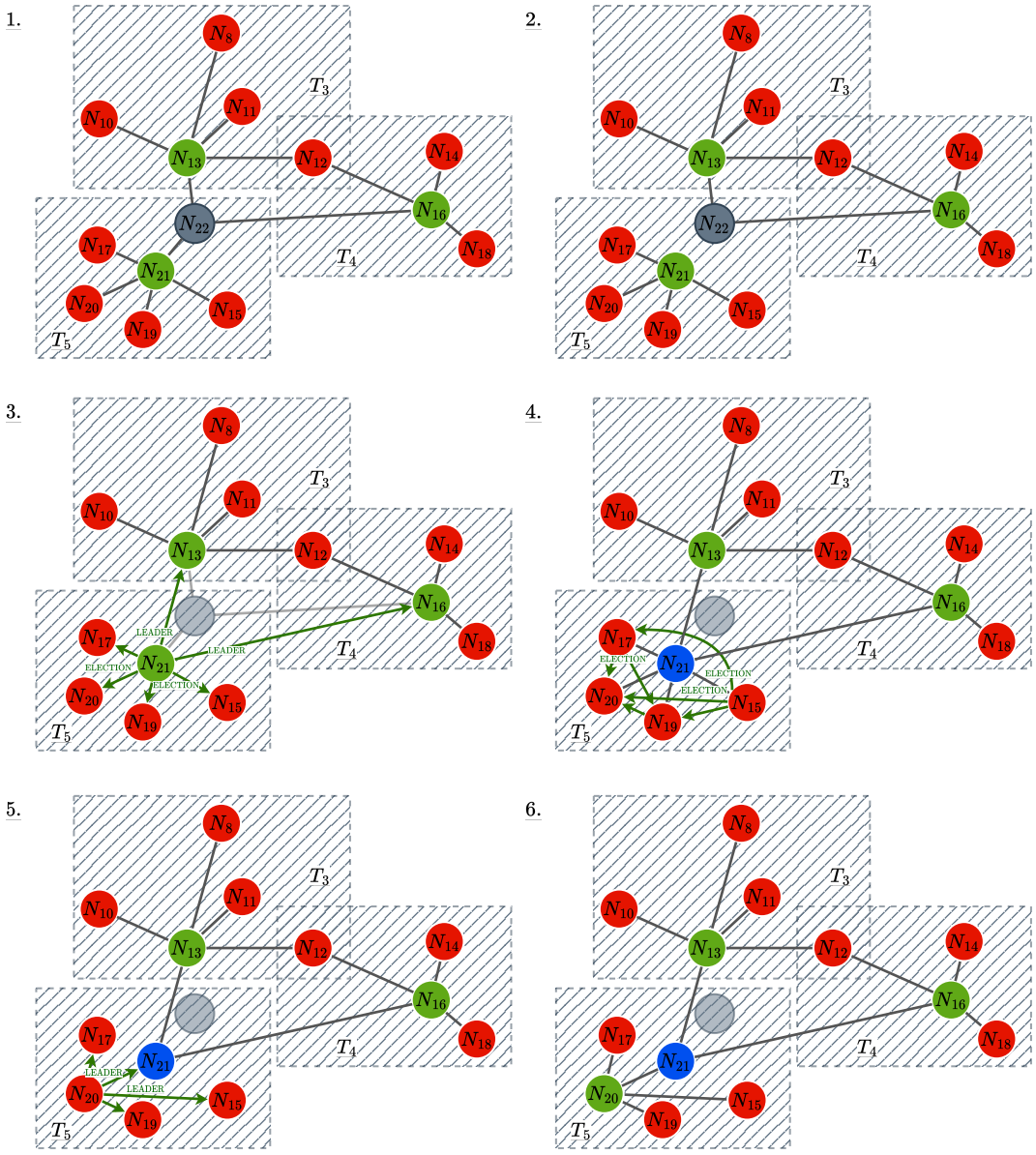

Figura 8: *Recovery* de la falla en N_{22}. En este diagrama se obvia el algoritmo *bully* para elegir nodo líder del clúster suscripto a la tarjeta T_5 puesto que ya se mostró en mayor detalle en el caso anterior. (4.) Existen optimizaciones como hacer que N_{21} mande un sólo mensaje de ELECTION a los nodos del clúster, pero en sí la idea es logar que los nodos elijan a un nuevo líder, ya que los nodos cuenta no pueden ser nodos líder tarjeta. (5.) Notar además que es N_{21} quien se encarga de poner al clúster de suscriptores T_5 en modo elección, para no quedar elegido, siendo que es el de mayor ID, puede simplemente no contestar, o contestar con mensaje del tipo CANNOT.

Flujo de las consultas de los clientes

- Cuando un **administrador** hace una consulta de ya sea su cuenta principal o una de sus tarjetas, propaga el mensaje desde el nodo más cercano hasta su ubicación. Cada uno de los nodos del server que recibe la consulta, checkea si tiene o no el registro de la tarjeta, si no la tiene propaga el mensaje. Eventualmente uno de los nodos que recibe la consulta contiene la información y le contesta al administrador, (TODO: ya sea directamente o por medio de un nodo que mantenga una pared entre cliente y los nodos de las estaciones).
- Para actualizar el límite de cuenta o de tarjeta, se procede como en el caso anterior sólo que ahora contestan nodos cuenta o nodos suscritos a la tarjeta, respectivamente.

Modelo de actores

Como probablemente ya se haya inferido, el modelo que plantea el sistema es un modelo de actores, ejecutado sobre un sistema distribuido.

Actores

Los actores del modelo son: - **nodo estación suscriptor**: nodo estación que está suscripto a un registro de una tarjeta. - **nodo estación lider**: nodo estación que está suscripto a un registro de una tarjeta y que tiene como responsabilidad mantener actualizado al nodo cuenta. - **nodo cuenta**: nodo que tiene direcciones de nodos estácion líder de manera tal que la unión de las suscripciones de esos nodos equivalga al conjunto de tarjetas de la cuenta. - **nodo surtidor**: nodo que se ejecuta en la red local de una estación y que sólo realiza consultas al nodo de la estación.

Los actores estación suscriptor, estación lider y estación cuenta pueden correrse sobre uno o distintos nodos, idealmente, y como consecuencia de los algoritmos de actualización del sistema, *un nodo cuenta siempre va a ejecutarse en el mismo nodo que un nodo estación líder*.

Mensajes

- **Cobrar**: el mensaje que envía el nodo surtidor al nodo central de su estación.
- **Consulta Registro**: el mensaje que se propaga como consecuencia de que una estación no conozca una tarjeta que le llega de uno de sus surtidores.
- **Registro**: La contestación al mensaje anterior.
- **Suscripción**: El mensaje que un nodo envía a los suscriptores de una tarjeta para que lo agreguen a la lista de suscriptores en sus registros de esa tarjeta.
- **Actualización**: el mensaje que se propaga a los nodos suscriptos a una tarjeta cada vez que esta se actualiza en cualquiera de los nodos que están suscriptos a ellas.
- TODO: **Líder estación**: Mensaje que se propaga para elegir un líder en un clúster de nodos suscriptos a una tarjeta.
- TODO: **Líder cuenta**: Mensaje que se propaga para elegir un líder en un conjunto de clústers de nodos suscriptos a las tarjetas que componen una cuenta principal.