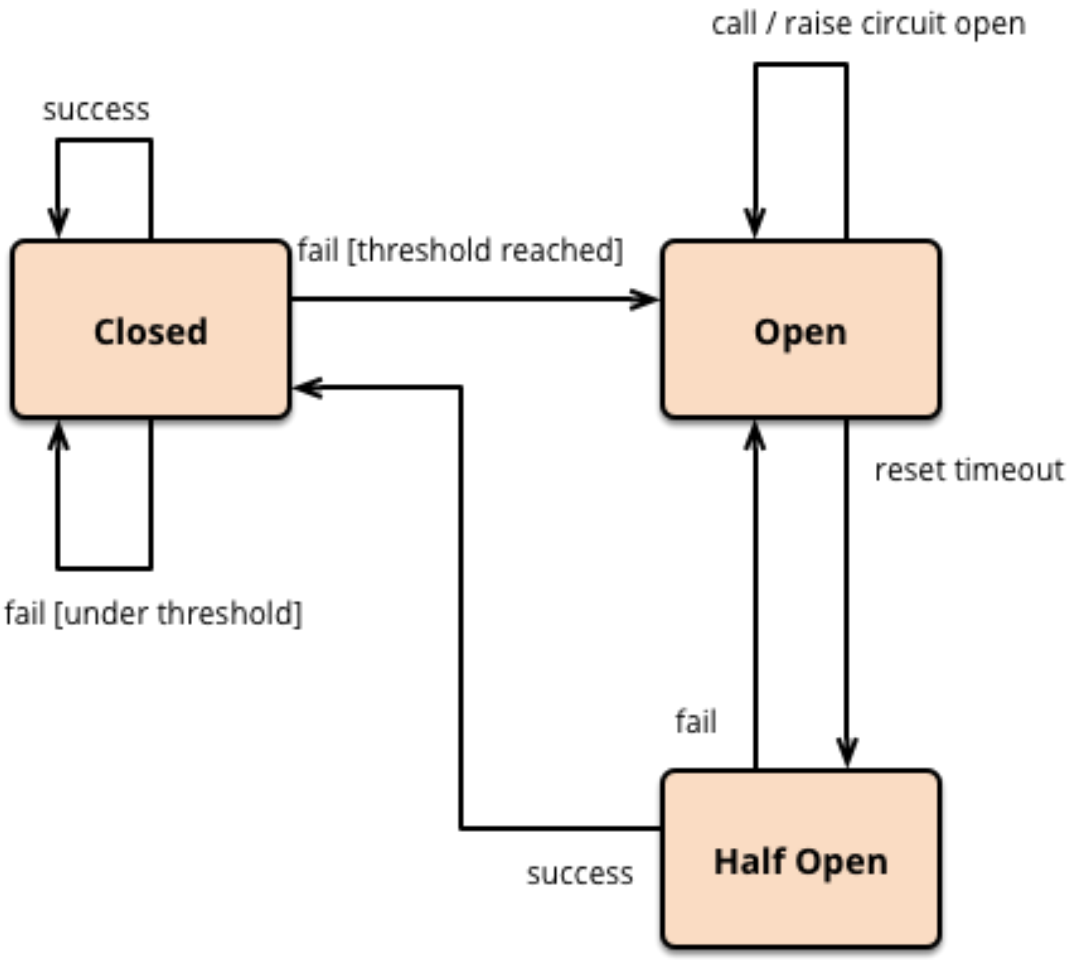
ARQUITECTURA YOLO

Teniendo en cuenta que YOLO solo puede manejar 10 solicitudes simultáneamente, y que el número de solicitudes por segundo es de al menos 5, considero que el tiempo máximo que debería tomar una transacción es de 2 segundos para evitar una cola en las solicitudes.
Como el tiempo por cada solitud no supera el tiempo de espera de 30 segundos del API WTF, se debe controlar el tiempo de ejecución y asegurarse de que no se supere máximo de 2 segundos, y se debe monitorear las solicitudes a WTF para que cuando falle se retengan, y se envíen en cuando WTF este disponible otra vez.
Uno de los patrones mas conocidos para poder afrontar esto seria el **Circuit Breaker.**



Cerrado: cuando todo es normal, el disyuntor permanece cerrado y todas las llamadas a los servicios ocurren normalmente. Si el número de fallas excede un límite predeterminado, el estado cambia a Abierto.

Abierto: en este estado, el disyuntor no ejecutará la llamada de servicio y devolverá un error manejado (hay casos que en su lugar pueden devolver información de caché).

Medio abierto: después de un período, el estado cambia a Medio abierto para probar si el problema original aún ocurre. Si ocurre una sola falla, el estado cambiará a Abrir nuevamente. Si tiene éxito, vuelve a la normalidad (Cerrado).