Cuando un servidor de Internet se vuelve lento debido a la congestión de información, la solución más obvia es ampliar la memoria, ampliar el disco duro o actualizar el procesador. Pero el tráfico de Internet está en constante crecimiento y lo anteriormente expuesto es sólo una solución temporal. La opción más razonable es, a largo plazo, configurar más servidores y repartir las peticiones de los clientes entre ellos. Esto incrementa la velocidad de acceso del usuario al servidor, mejora la fiabilidad del sistema y la tolerancia a fallos, permitiendo reparar o mantener cualquier servidor en línea sin que afecte al resto del servicio.

Así es como muchos de los sitios web más conocidos pueden abarcar toda la demanda que reciben, gracias al uso de un número de servidores "espejo" (mirrors) llamados granjas de servidores. Sin embargo, una vez utilizados varios servidores para responder todas las peticiones que van a una dirección, ¿Cómo las dividimos entre los distintos servidores? ¿Cómo podemos saber qué rendimiento ofrecemos y qué tiempo de CPU está generando cada petición? Simplemente el conectar más servidores a una red, no asegura la mejora del servicio.

PERO ¿QUÉ ES REALMENTE?

De forma sencilla, el balanceo de carga es la manera en que las peticiones de Internet son distribuídas sobre una fila de servidores. Existen varios métodos para realizar el balanceo de carga. Desde el simple "Round Robin" (repartiendo todas las peticiones que llegan de Internet entre el número de servidores disponibles para dicho servicio) hasta los equipos que reciben las peticiones, recogen información, en tiempo real, de la capacidad operativa de los equipos y la utilizan para enrutar dichas peticiones individualmente al servidor que se encuentre en mejor disposición de prestar el servicio adecuado.

Los balanceadores de carga pueden ser soluciones hardware, tales como routers y switches que incluyen software de balanceo de carga preparado para ello, y soluciones software que se instalan en el back end de los servidores.

MÉTODO ROUND ROBIN

El método más simple de todos, es la solución Round Robin. Las peticiones clientes son distribuídas equitativamente entre todos los servidores existentes. Este metodo ciclico no tiene en cuenta las condiciones y carga de cada servidor. Esto puede llevarnos a tener servidores que reciben peticiones de carga mucho mayor, mientras tenemos servidores que apenas se encuentran utilizando recursos.

Otra limitación es que los problemas de los servidores no son recogidos inmediatamente. Esto puede llevarnos a estar enviando peticiones a un servidor que se encuentra fuera de servicio o que responde lentamente. Finalmente, el método Round Robin no aprovecha las diferentes prestaciones de los servidores (un PC Pentium normal puede estar obteniendo tantas peticiones como un multiprocesador Sun situado junto a él).

PRIMERA GENERACION DE BALANCEO DE CARGA

Las soluciones "reales" de balanceo de carga necesitan descubrir el rendimiento del servidor. La primera generación puede detectar el rendimiento del servidor via "passive polling", lo que significa que el balanceador de carga mide el tiempo de respuesta de los servidores y por ello tiene una idea de cómo están funcionando. De nuevo, tampoco se tiene en cuenta la variedad de servidores empleados. Además, sólo descubre que los servidores tienen un problema después de que se producen retrasos o, en el peor de los casos, cuando los servidores están completamente caídos.

SEGUNDA GENERACION DE BALANCEO

El balanceo de carga más seguro sólo se puede conseguir considerando el uso real de los servidores, permitiendo que los recursos existentes se empleen al máximo, al conocer cómo están siendo utilizados estos recursos incluso antes de que las peticiones de los clientes lleguen a ellos. El tráfico se enruta proactivamente, cambiando el antiguo concepto existente de balanceo de carga, hacia una solución de optimización del servidor, consiguiendo el mejor resultado posible con la tecnología disponible.

Para lograrlo, el balanceador de carga continuamente realiza peticiones de datos de cada servidor en la granja de servidores para monitorizar sus condiciones y direccionar las peticiones de los clientes hacia el servidor que se encuentre más disponible y en mejor estado para responder a dichas peticiones. Los parámetros solicitados, dependen del producto utilizado. Normalmente se emplea la utilización de la CPU del servidor, el uso de memoria y el número de conexiones abiertas.

La segunda generación de balanceadores posee funciones de mensajería, informando si los servidores están fuera de servicio, y si es así, cuándo serán devueltos a producción. La mayoría de los servidores "revividos" pasan un período de prueba durante el cuál no se llenan completamente de peticiones.

Se puede incluso desconectar los servidores para repararse o para realizar el mantenimiento, a través del método de "apagado progresivo". El servidor, a partir de ese momento, no acepta nuevas peticiones pero permanece activo hasta que las transacciones de comercio electrónico y las descargas que se estén produciendo finalicen.

La segunda (algunas veces considerada la primera) regla más importante de una solución de balanceo de carga, es incrementar la fiabilidad del sitio web y del contenido y los servicios que está ofreciendo. Normalmente la segunda generación de balanceadores de carga hardware se vende en parejas, es decir, dos equipos iguales. Uno de ellos es la unidad activa y el segundo la unidad de repuesto o de back up. Una unidad de back up en modo stand by (en espera) con una misma dirección IP y MAC significa que incluso cuando el balanceador se ve afectado por un incidente como puede ser un fallo de cableado, fuego o error humano, hay una unidad de repuesto pre-configurada que pasa a ser operacional de forma inmediata.