Cómo reducir la latencia con Edge Computing y optimización de red

Cómo reducir la latencia con Edge Computing y optimización de red.Las empresas actuales a menudo viven o mueren por el rendimiento de su red. Enfrentando la presión de los clientes y las demandas de tiempo de actividad de SLA de sus clientes, las organizaciones buscan constantemente formas de mejorar la eficiencia de la red y ofrecer servicios mejores, más rápidos y más confiables.

Es por eso que la arquitectura informática de vanguardia se ha convertido en un tema nuevo y emocionante en el mundo de la infraestructura de red en los últimos años. Si bien el concepto no es necesariamente nuevo, los desarrollos en dispositivos de Internet de las cosas (IoT) y la tecnología del centro de datos lo han convertido en una solución viable por primera vez.

Edge computing reubica las funciones clave de procesamiento de datos desde el centro de una red hasta el borde, más cerca de donde se reúne y se entrega a los usuarios finales. Si bien hay muchas razones por las que esta arquitectura tiene sentido para ciertas industrias , la ventaja más obvia de la informática de borde es su capacidad para combatir la latencia. La latencia efectiva de solución de problemas a menudo puede significar la diferencia entre perder clientes y proporcionar servicios de alta velocidad y receptivos que satisfagan sus necesidades.

¿Qué es la latencia?

Ninguna discusión sobre la latencia estaría completa sin una breve descripción de la diferencia entre latencia y ancho de

banda . Aunque los dos términos a menudo se usan indistintamente, se refieren a cosas muy diferentes. El ancho de banda mide la cantidad de datos que pueden viajar a través de una conexión a la vez. Cuanto mayor sea el ancho de banda , más datos se pueden entregar.

En términos generales, un mayor ancho de banda contribuye a una mejor velocidad de la red porque más datos pueden viajar a través de las conexiones, pero el rendimiento de la red todavía está limitado por el rendimiento, que mide la cantidad de datos que pueden procesarse a la vez por diferentes puntos de una red. Ancho de banda creciente a un servidor de bajo rendimiento, entonces, no hará nada para mejorar el rendimiento porque los datos simplemente se bloquearán cuando el servidor intente procesarlo.

La latencia, por otro lado, es una medida de cuánto tiempo tarda un paquete de datos en viajar desde su punto de origen hasta su destino. Si bien el tipo de conexión es una consideración clave (los cables de fibra óptica transmiten datos mucho más rápido que el cobre convencional, por ejemplo), la distancia sigue siendo uno de los factores clave para determinar la latencia. Esto se debe a que los datos todavía están limitados por las leyes de la física y no pueden exceder la velocidad de la luz (aunque algunas conexiones se han acercado a ella). No importa cuán rápido pueda ser una conexión, los datos aún deben recorrer físicamente esa distancia, lo que lleva tiempo.

El tipo de conexión entre estos puntos es importante ya que los datos se transmiten más rápido a través de cables de fibra óptica que el cableado de cobre, pero la distancia y la complejidad de la red juegan un papel mucho más importante. Las redes no siempre enrutan los datos a lo largo de la misma ruta porque los enrutadores y conmutadores priorizan y evalúan continuamente dónde enviar los paquetes de datos que reciben. Es posible que la ruta más corta entre dos puntos no siempre esté disponible, lo que obliga a los paquetes de datos a

viajar una distancia más larga a través de conexiones adicionales, todo lo cual aumenta la latencia en una red.

Prueba de latencia de red

¿Cuánto tiempo? Hay algunas maneras fáciles de realizar una prueba de latencia de red para determinar qué tan grande es el impacto de la latencia en el rendimiento. Los sistemas operativos como Microsoft Windows, Apple OS y Linux pueden llevar a cabo un comando » traceroute «. Este comando supervisa el tiempo que tardan los enrutadores de destino en responder a una solicitud de acceso, medido en milisegundos. Sumar la cantidad total de tiempo transcurrido entre la solicitud inicial y la respuesta del enrutador de destino proporcionará una buena estimación de la latencia del sistema.

La ejecución de un comando traceroute no solo muestra cuánto tiempo tardan los datos en viajar de una dirección IP a otra, sino que también revela cuán compleja puede ser la creación de redes . Dos solicitudes idénticas podrían tener diferencias significativas en la latencia debido a la ruta que tomaron los datos para llegar a su destino. Este es un subproducto de la forma en que los enrutadores priorizan y dirigen diferentes tipos de datos. La ruta más corta puede no estar siempre disponible, lo que puede causar latencia inesperada en una red.

Latencia en juegos

Aunque muchas personas solo pueden escuchar acerca de la latencia cuando la culpan por sus desgracias en los juegos en línea, los videojuegos son en realidad un buen ejemplo para explicar el concepto.

En el contexto de un videojuego, la alta latencia significa que la entrada del controlador de un jugador tarda más en llegar a un servidor multijugador. Las conexiones de alta latencia producen un retraso significativo o un retraso entre

las entradas del controlador de un jugador y las respuestas en pantalla. Para un jugador con una conexión de baja latencia, estos oponentes parecen reaccionar lentamente a los eventos, incluso parados . Desde la perspectiva del jugador de alta latencia, otros jugadores parecen teletransportarse por toda la pantalla porque su conexión no puede entregar y recibir datos lo suficientemente rápido como para presentar información del juego proveniente del servidor.

Los jugadores a menudo se refieren a su «ping» cuando hablan de latencia. Una prueba de ping es similar a un comando «traceroute». La principal diferencia es que la prueba de ping también mide cuánto tiempo debe responder el sistema de destino (como un «ping» de sonda que se devuelve a la fuente después de rebotar en un objeto). Un ping bajo significa que hay muy poca latencia en la conexión. No es sorprendente, entonces, que el consejo sobre cómo los jugadores pueden reducir su ping implica cosas como eliminar impedimentos que podrían ralentizar los paquetes de datos, como firewalls (no recomendado), o acercar físicamente su computadora al enrutador de su hogar (probablemente insignificante, pero cada poquito podría ayudar en una partida clasificada de Overwatch).

Latencia en servicios de transmisión

La misma latencia que atormenta a los jugadores es responsable del contenido de transmisión fragmentada . Estas demoras de almacenamiento en búfer ya ocurren en el 29 por ciento de las experiencias de transmisión. Dado que se espera que el contenido de video represente el 67% del tráfico global de Internet (un estimado de 187 exabytes) para 2021, la latencia es un problema que bien podría volverse aún más común en el futuro cercano. Los estudios han demostrado que los usuarios de Internet abandonan los videos que almacenan en búfer o tardan en cargarse después de solo dos segundos de retraso. Las empresas que brindan servicios de transmisión deben

encontrar soluciones a este problema si esperan emprender la transformación digital empresarial que los mantendrá competitivos en el futuro.

Cómo mejorar la latencia

La latencia es ciertamente fácil de notar dado que demasiado puede causar tiempos de carga lentos, video o audio nervioso o solicitudes de tiempo de espera agotado. Sin embargo, solucionar el problema puede ser un poco más complicado ya que las causas a menudo se encuentran aguas abajo de la infraestructura de una empresa.

En la mayoría de los casos, la latencia es un subproducto de la distancia. Aunque las conexiones rápidas pueden hacer que las redes parezcan funcionar instantáneamente, los datos todavía están limitados por las leyes de la física. No puede moverse más rápido que la velocidad de la luz, aunque las innovaciones en la tecnología de fibra óptica le permiten llegar a aproximadamente dos tercios del camino . En las mejores condiciones, se necesitan datos de aproximadamente 21 milisegundos para viajar de Nueva York a San Francisco. Sin embargo, este número es engañoso. Varios cuellos de botella debido a las limitaciones de ancho de banda y al redireccionamiento cerca de los puntos finales de datos (el problema de la «última milla») pueden agregar entre 10 y 65 milisegundos de latencia.

Reducir la distancia física entre la fuente de datos y su destino final es la mejor estrategia para reducir la latencia. Para los mercados e industrias que dependen del acceso más rápido posible a la información, como dispositivos IoT o servicios financieros, esa diferencia puede ahorrarles a las compañías millones de dólares. La velocidad, entonces, puede proporcionar una ventaja competitiva significativa para las organizaciones dispuestas a comprometerse con ella.

Cómo reducir la latencia con Edge Computing

La arquitectura de computación perimetral ofrece una solución innovadora para el problema de la latencia y cómo reducirla. Al ubicar las tareas de procesamiento clave más cerca de los usuarios finales, la informática de borde puede ofrecer servicios más rápidos y más receptivos. Los dispositivos IoT proporcionan una forma de llevar estas tareas al borde de una red.

Los avances en la tecnología de procesador y almacenamiento han hecho que sea más fácil que nunca aumentar la potencia de los dispositivos habilitados para Internet, lo que les permite procesar gran parte de los datos que recopilan localmente en lugar de transmitirlos de vuelta a los servidores centralizados de computación en la nube para su análisis. Al resolver más procesos más cerca de la fuente y transmitir muchos menos datos al centro de la red, los dispositivos IoT pueden mejorar en gran medida la velocidad de rendimiento. Esto será de vital importancia para la tecnología, como los vehículos autónomos , donde unos pocos milisegundos de retraso podrían ser la diferencia entre un viaje seguro a una reunión familiar y un accidente fatal.

Por supuesto, no todas las transformaciones digitales empresariales se realizarán a través de dispositivos IoT. Los servicios de transmisión de video, por ejemplo, necesitan un tipo diferente de solución. Los centros de datos perimetrales, instalaciones más pequeñas y especialmente diseñadas ubicadas en mercados emergentes clave, facilitan la transmisión de video y audio al almacenar en caché el contenido de alta demanda mucho más cerca de los usuarios finales. Esto no solo garantiza que los servicios populares se entreguen más rápido, sino que también libera ancho de banda para entregar contenido desde ubicaciones más distantes.

Por ejemplo, si los diez programas principales de Netflix se transmiten desde una instalación de hiperescala en la ciudad de Nueva York, pero pueden almacenar en caché ese mismo contenido en una instalación periférica fuera de Pittsburgh, los usuarios finales en ambos mercados podrán transmitir contenido de manera más eficiente porque las fuentes de transmisión se distribuyen más cerca de los consumidores.

Las experiencias de juego en línea (como Roblox) también pueden ayudar a reducir la latencia para sus usuarios al colocar los servidores en centros de datos periféricos más cerca de donde se encuentran los jugadores. Si los jugadores de una región en particular inician sesión en servidores a los que se puede acceder con una latencia mínima, tendrán una experiencia mucho más agradable que si estuvieran luchando constantemente para lidiar con las altas tasas de ping que resultan del uso de servidores en el otro lado de la red del país.

Consejos y herramientas adicionales para solucionar problemas de latencia de red

Si bien la simple reducción de la distancia que deben recorrer los datos es a menudo la mejor manera de mejorar el rendimiento de la red, existen algunas estrategias adicionales que pueden reducir sustancialmente la latencia de la red.

Cambio de etiqueta multiprotocolo (MPLS)

La optimización efectiva del enrutador también puede ayudar a reducir la latencia. La conmutación de etiquetas multiprotocolo (MPLS) mejora la velocidad de la red al etiquetar paquetes de datos y enrutarlos rápidamente a su próximo destino. Esto permite que el siguiente enrutador simplemente lea la información de la etiqueta en lugar de tener que buscar en las tablas de enrutamiento más detalladas del paquete para determinar dónde debe ir a continuación. Si bien no es aplicable para todas las redes, MPLS puede reducir considerablemente la latencia al agilizar las tareas del enrutador.

Cableado de conexión cruzada

En un centro de datos de colocación neutral para el operador, los clientes de colocación a menudo necesitan conectar sus redes híbridas y de múltiples nubes a una variedad de proveedores de servicios en la nube . En circunstancias normales, se conectan a estos servicios a través de un ISP, lo que los obliga a usar Internet público para establecer una conexión. Sin embargo, las instalaciones de colocación ofrecen cableado de conexión cruzada , que es simplemente un tendido de cable dedicado desde el servidor de un cliente al servidor de un proveedor de la nube. Con la distancia entre los servidores a menudo medida en simples centímetros o metros, la latencia se reduce considerablemente, lo que permite tiempos de respuesta mucho más rápidos y un mejor rendimiento general de la red.

Cableado de interconexión directa

Cuando el cableado de conexión cruzada en un entorno de colocación no es posible, existen otras formas de simplificar las conexiones para reducir la latencia. Las conexiones directas a proveedores en la nube, como Microsoft Azure ExpressRoute , no siempre resuelven los desafíos que plantea la distancia, pero el cableado de interconexión punto a punto significa que los datos siempre viajarán directamente desde el servidor del cliente al servidor en la nube. A diferencia de una conexión a Internet convencional, no hay que considerar el enrutamiento de ruta, lo que significa que los datos no se redirigirán cada vez que se envíe un paquete a través de la red.

Construyendo un futuro más rápido

Los centros de datos de colocación ofrecen una serie de herramientas valiosas para solucionar problemas de latencia de red. Aunque la tecnología puede no existir (todavía) para

enviar y recibir datos a través de una red de manera instantánea, las estrategias como la computación de borde y el cableado de conexión cruzada brindan a los clientes de colocación opciones eficaces para combatir la latencia y brindar servicios más rápidos y confiables.

La combinación de centros de datos periféricos y dispositivos IoT tiene el potencial de transformar la forma en que las empresas construyen su arquitectura de red . Edge computing abre una nueva gama de opciones sobre cómo reducir la latencia y brindar servicios de manera más eficiente a los usuarios finales. En un mercado cada vez más impulsado por cortos períodos de atención, es muy probable que la velocidad continúe siendo un diferenciador clave, lo que hace que las estrategias informáticas de vanguardia sean cada vez más vitales para las empresas en muchas industrias.

Si tu negocio opera en Colombia en la zona Andina, consulta con nosotros cómo disponer de un data center en Bogotá, Medellin o Cali puede reducir la latencia de tus datos.
Leer también:¿Qué es una estrategia de centro de datos múltiples y por qué necesita una?; 5 señales de que es hora de una migración del Centro de datos y cómo planificar una; Tecnología 5G la batalla por la latencia