Artículo
Ricardo Paiva · 18 ago, 2020 Lectura de 32 min

La tecnología de InterSystems en Amazon EC2: Arquitectura de referencia

¡Hola Comunidad!

Las empresas necesitan crecer y administrar sus infraestructuras informáticas globales de forma rápida y eficiente, al mismo tiempo que optimizan y administran la manera en que invierten y gastan sus fondos. Los servicios de computación y de almacenamiento de Amazon Web Services (AWS) y Elastic Compute Cloud (EC2) satisfacen las necesidades de las aplicaciones más exigentes basadas en InterSystems IRIS, al proporcionar una infraestructura informática global sumamente sólida.
La infraestructura de Amazon EC2 permite que las empresas se provean rápidamente de la capacidad de computación y/o aumenten de forma rápida y flexible la infraestructura con la que cuentan en sus instalaciones, dentro de la nube. AWS proporciona un amplio conjunto de servicios y de robustos mecanismos para seguridad, establecimiento de redes, computación y almacenamiento.

En el centro de AWS está Amazon EC2. Una infraestructura informática en la nube, que es compatible con una gran variedad de sistemas operativos y configuraciones (por ejemplo, CPU, RAM, red). AWS proporciona imágenes de máquinas virtuales (VM) configuradas previamente, conocidas como Amazon Machine Images, o AMI, con sistemas operativos invitados, incluyendo varias distribuciones y versiones de Linux® y de Windows. Pueden tener programas adicionales, usados como base para las instancias virtuales que se ejecutan en AWS. Puedes utilizar estas AMIs como punto de partida para crear instancias e instalar, o configurar software adicional, datos, etc. con la finalidad de generar otras AMI para aplicaciones o cargas de trabajo específicas.

Al igual que con cualquier plataforma o modelo de implementación, es necesario ser cuidadoso, a fin de garantizar que se tengan en cuenta todos los elementos que conforman el entorno de una aplicación, como el rendimiento, la disponibilidad, las operaciones y la administración de los procesos.

En este documento se tratarán aspectos específicos de cada una de las siguientes áreas:

- **Instalación y configuración de la red.** En esta sección se habla de la configuración de la red para las aplicaciones basadas en InterSystems IRIS dentro de AWS, incluidas las subredes que soportan los grupos de servidores lógicos para las diferentes capas y funciones que están dentro de la arquitectura de referencia.
- **Instalación y configuración del servidor.** Esta sección comprende los servicios y recursos que participan en el diseño de los distintos servidores para cada capa. También incluye la arquitectura para una mayor accesibilidad en todas las zonas disponibles.
- **Seguridad.** Esta sección trata sobre los mecanismos de seguridad en AWS, incluyendo cómo configurar la seguridad de la instancia y de la red para habilitar las autorizaciones de acceso a las soluciones globales, así como entre las capas e instancias.
- **Implementación y administración.** En esta sección se proporcionan detalles sobre empaquetado, implementación, supervisión y administración.

Escenarios de arquitectura e implementación

En este documento se proporcionan varias arquitecturas de referencia dentro de AWS como ejemplos para proporcionar aplicaciones robustas, de alto rendimiento y disponibilidad, basadas en las tecnologías de InterSystems, incluyendo InterSystems IRIS, HealthShare, TrakCare y tecnologías embebidas asociadas, como DeepSee, iKnow, CSP, Zen y Zen Mojo.

Para entender cómo InterSystems IRIS y sus componentes asociados pueden alojarse en AWS, primero revisaremos la arquitectura y los componentes de una implementación típica en InterSystems IRIS, y analizaremos algunos escenarios y topologías frecuentes.

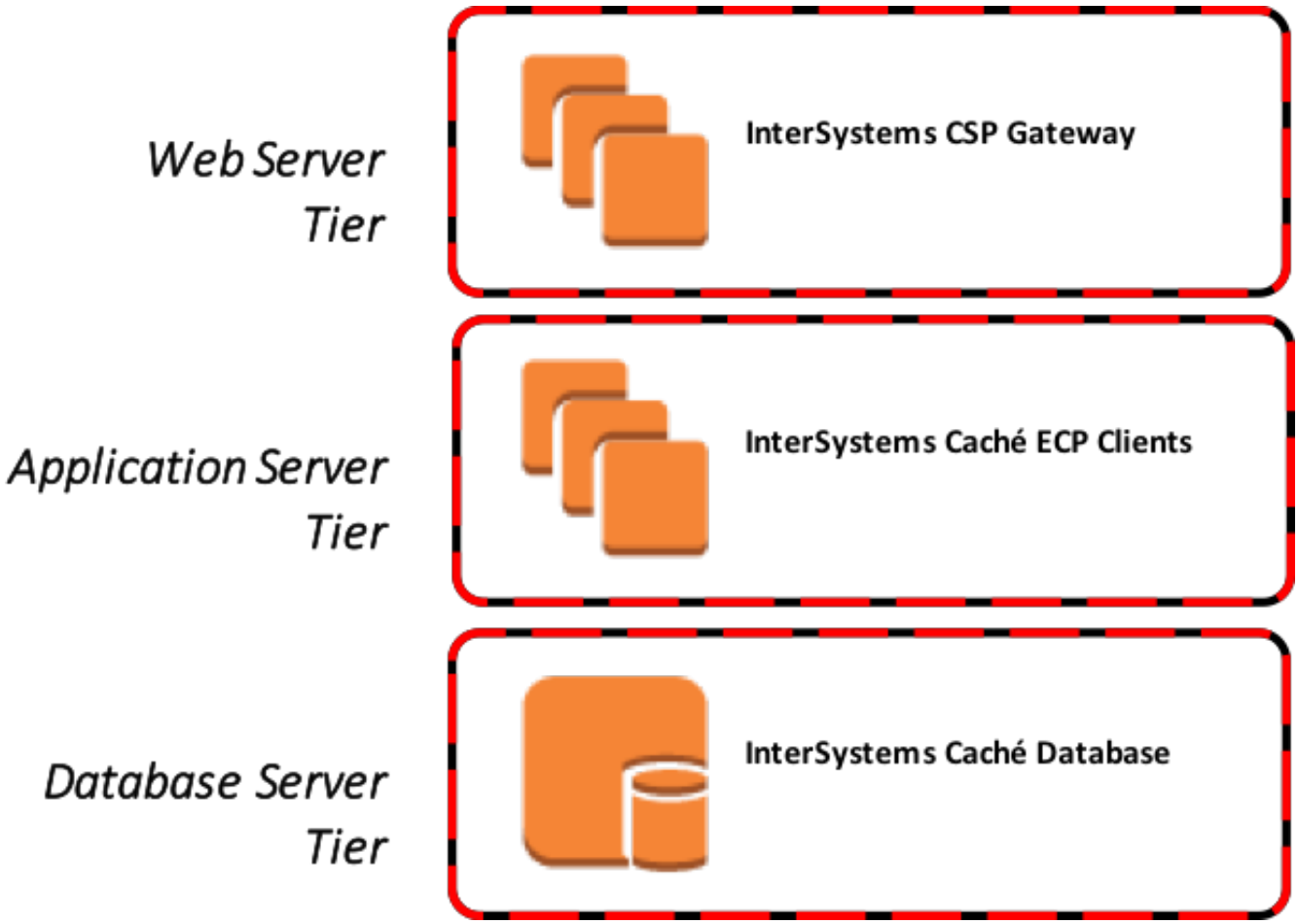
Análisis de la arquitectura de InterSystems IRIS

La plataforma de datos de InterSystems evoluciona continuamente para proporcionar un sistema avanzado de administración de bases de datos y un entorno de programación de aplicaciones que sea rápido, con la finalidad de lograr avances en el procesamiento y análisis de modelos de datos complejos, así como en el desarrollo de aplicaciones web y para dispositivos móviles.

Se trata de una nueva generación de tecnología de bases de datos, que proporciona múltiples formas de acceder a los datos. Los datos solamente se describen una vez en un único diccionario de datos integrado y están disponibles inmediatamente, utilizando el acceso para objetos, un SQL de alto rendimiento y un potente acceso multidimensional - todos ellos pueden acceder simultáneamente a los mismos datos.

Los servicios y las capas de componentes de la arquitectura de alto nivel de InterSystems IRIS que están disponibles se muestran en la Figura 1. Estas capas generales también aplican a los productos TrakCare y HealthShare de InterSystems.

Figura-1: Capas de los componentes de alto nivel



Escenarios más frecuentes para la implementación

Existen numerosas combinaciones posibles para la implementación, sin embargo, en este documento se incluirán dos escenarios: un modelo híbrido y un modelo completo de alojamiento en la nube.

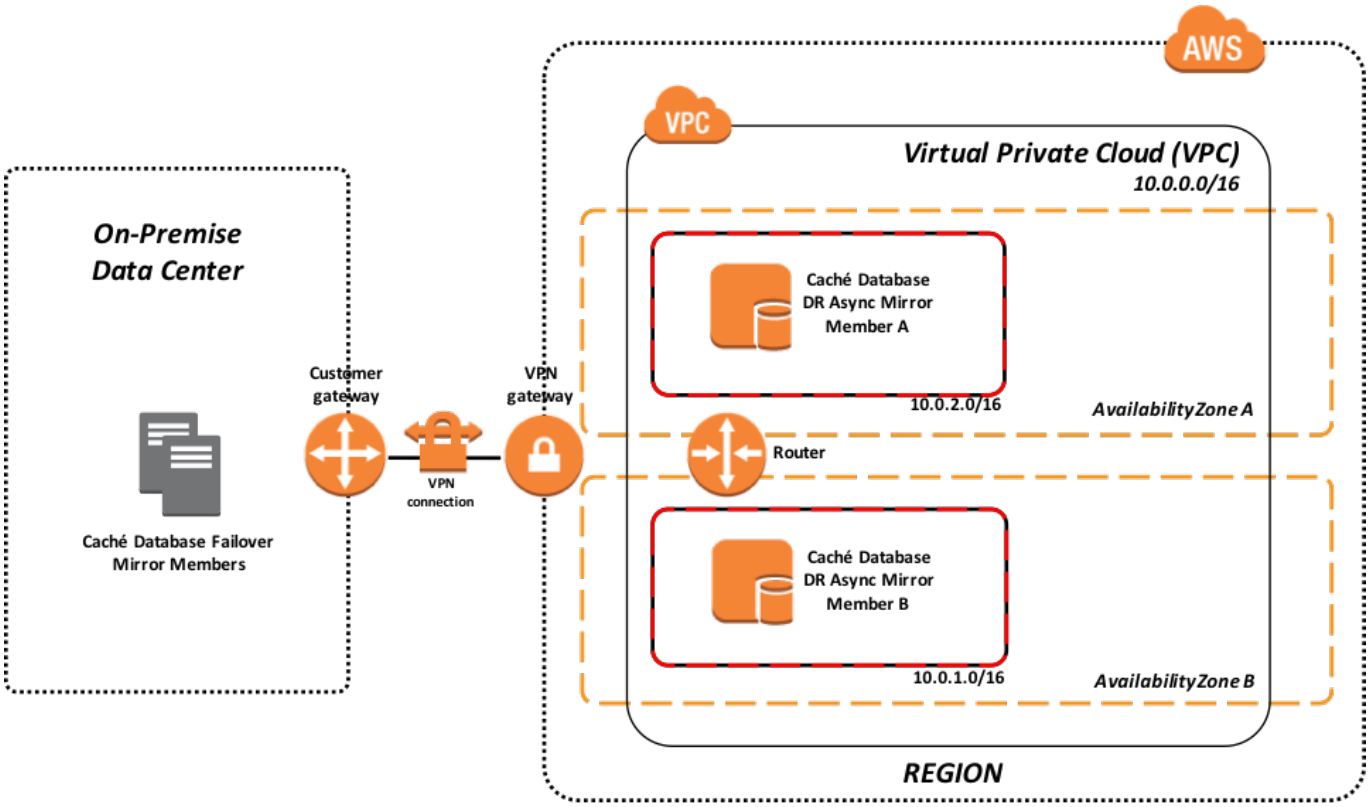
Modelo híbrido

En este escenario, una empresa quiere aprovechar tanto los recursos empresariales que se encuentran en su infraestructura como los recursos de AWS EC2, tanto para la recuperación en caso de desastres, como para las contingencias que surjan durante el mantenimiento interno, las iniciativas de reconfiguración o el aumento de sus capacidades a corto/largo plazo cuando sea necesario. Este modelo puede ofrecer un alto nivel de disponibilidad para la continuidad de negocio y la recuperación en caso de desastres, alojando en EC2 membro(s) de un mirror manteniendo los demás en el centro de datos de la empresa.

La conectividad para este modelo, en este escenario, depende de la presencia de un túnel VPN entre la implementación en las instalaciones y la/s zona/s donde AWS está disponible para presentar sus recursos como una extensión del centro de datos de la empresa. Existen otros métodos de conectividad como AWS Direct Connect, que no están incluídos en este documento. Puedes obtener más información sobre AWS Direct Connect aquí.

Los detalles para configurar este ejemplo de Amazon Virtual Private Cloud (VPC), para soportar la recuperación en caso de desastres en las instalaciones de su centro de datos, se pueden consultar aquí.

Figura-2: Modelo híbrido usando AWS VPC para la recuperación en caso de desastres en las instalaciones físicas

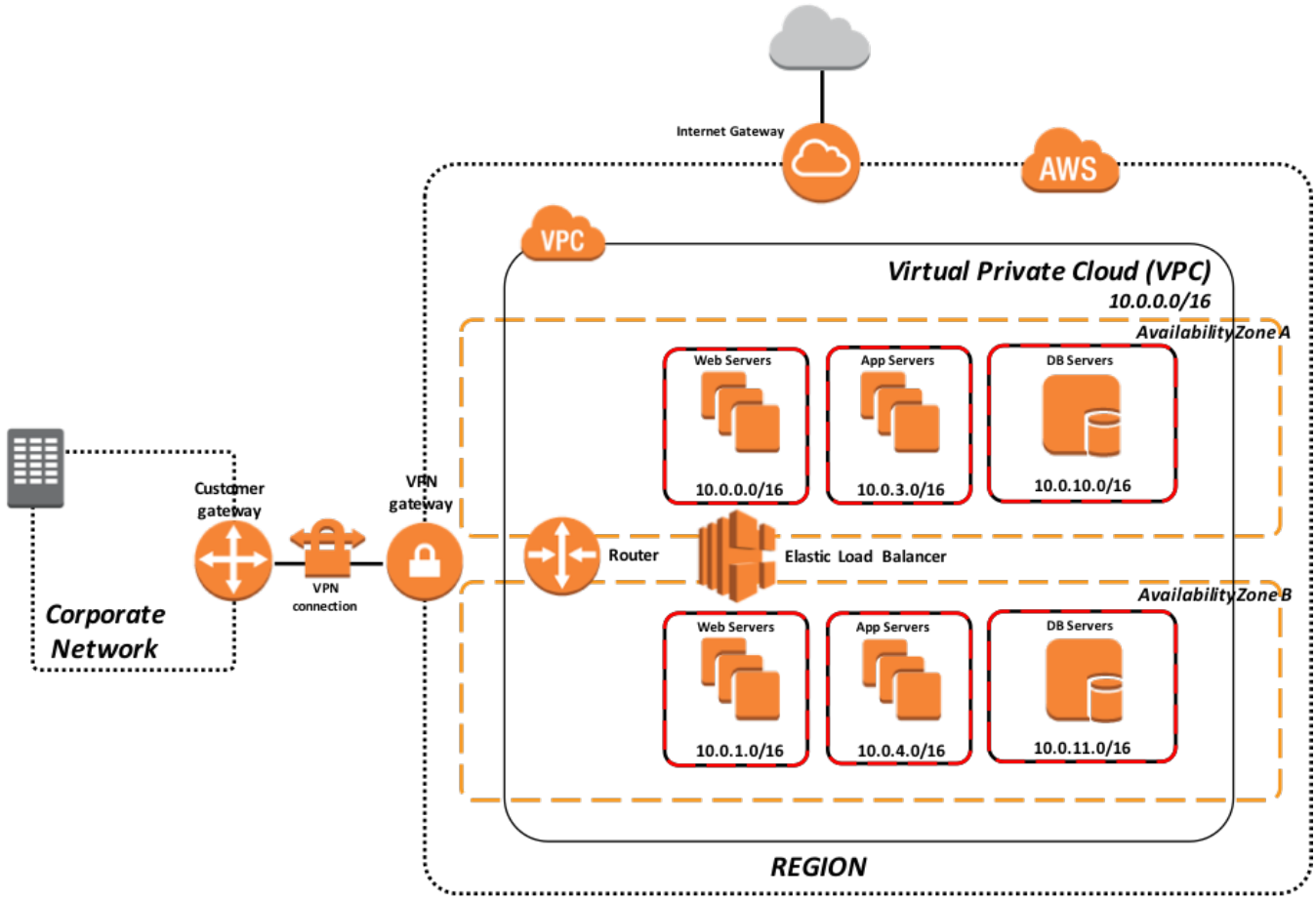


En el ejemplo anterior se muestran un par de réplicas en un procedimiento contra fallos, que operan en las instalaciones del centro de datos con una conexión VPN hacia su VPC de AWS. En la imagen del VPC se presentan varias subredes en zonas con doble disponibilidad para una cierta región AWS. Para proporcionar resiliencia, existen dos miembros "mirror" para la Recuperación en caso de desastres asíncrona (uno en cada zona de disponibilidad).

Modelo alojado en la nube

En este escenario, tu aplicación basada en InterSystems, incluyendo tanto las capas de datos como de presentación se mantiene completamente en la nube de AWS utilizando diversas zonas de disponibilidad dentro de una única región de AWS. También están disponibles el mismo túnel VPN, AWS Direct Connect, e incluso modelos de conectividad a Internet en estado puro.

Figura-3: Modelo alojado en la nube que soporta una carga de trabajo a plena producción



En el ejemplo anterior, en la Figura 3, se muestra un modelo que soporta una implementación de producción completa de la aplicación en tu VPC. Este modelo aprovecha las dos zonas de disponibilidad mediante la replicación síncrona entre las zonas de disponibilidad, junto con el balanceo de carga en los servidores web y los servidores de aplicaciones que están asociados como clientes ECP. Cada nivel está aislado en un grupo de seguridad separado para los controles de seguridad de la red. Las direcciones IP y los rangos de los puertos solo se abren cuando es necesario, en función de las necesidades de tu aplicación.

Almacenamiento y recursos computacionales

Almacenamiento

Existen varios tipos de opciones de almacenamiento disponibles. Para el propósito de esta arquitectura de referencia se discuten los volúmenes de almacenamiento del *Almacén elástico de bloques de Amazon* (Amazon EBS) y del *Almacén de instancias de Amazon EC2* (también llamadas unidades efímeras) para varios casos posibles de uso. Puedes encontrar información adicional sobre las diferentes opciones de almacenamiento aquí y aquí.

Almacenamiento elástico de bloques (EBS)

EBS ofrece almacenamiento persistente a nivel de bloques para utilizarlo con las instancias de Amazon EC2 (máquinas virtuales), las cuales pueden formatearse y organizarse como sistemas de archivos tradicionales, tanto en Linux como en Windows y, lo más importante, es que los volúmenes se almacenan fuera de la instancia, por lo que se mantienen independientemente de cuál sea la vida útil de una única instancia en Amazon EC2, y este es un aspecto importante para los sistemas de bases de datos.

Además, Amazon EBS oferece la capacidad de crear capturas, de un momento en el tiempo, de los volúmenes, que se mantienen en Amazon S3. Estas capturas se pueden utilizar como punto de partida para nuevos volúmenes en Amazon EBS y para proteger los datos con el fin de que permanezcan a largo plazo. La misma captura puede usarse para crear instancias de tantos volúmenes como se quieran. Estas capturas también se pueden copiar entre las distintas regiones de AWS, haciendo más sencillo aprovechar varias regiones de AWS para expansiones geográficas, migración entre centros de datos y recuperación en caso de desastres. El tamaño de los volúmenes de Amazon EBS varía desde 1 GB hasta 16 TB, y se asignan en incrementos de 1 GB.

Amazon EBS cuenta con tres tipos diferentes: Volúmenes magnéticos, de uso general (SSD) y de IOPS provisionadas (SSD). En las siguientes secciones ofrezco una breve descripción de cada uno de ellos.

Volúmenes magnéticos

Los volúmenes magnéticos ofrecen un almacenamiento rentable para aquellas aplicaciones que tengan requisitos de E/S moderados o por ráfagas. Los volúmenes magnéticos están diseñados para realizar aproximadamente 100 operaciones de entrada/salida por segundo (IOPS) en promedio, con una gran capacidad para alcanzar ráfagas de cientos de IOPS. Los volúmenes magnéticos también están muy bien adaptados para utilizarse como volúmenes de arranque, donde la capacidad de las ráfagas proporciona tiempos de inicio rápidos para las instancias.

De uso general (SSD)

Los volúmenes de uso general (SSD) ofrecen un almacenamiento rentable que resulta ideal para una gran variedad de cargas de trabajo. Estos volúmenes realizan latencias de milisegundos de un solo dígito, tienen la capacidad de producir ráfagas de hasta 3.000 IOPS durante largos periodos de tiempo y un rendimiento de referencia de 3 IOPS/GB, pero puede alcanzar un máximo de hasta10,000 IOPS (a 3.334 GB). Los volúmenes de uso general (SSD) pueden variar en tamaño desde 1 GB hasta 16 TB.

De IOPS provisionadas (SSD)

Los volúmenes IOPS provisionados (SSD) están diseñados para ofrecer un alto rendimiento predecible para cargas de trabajo de E/S intensivas, como las cargas de trabajo para bases de datos que son sensibles al rendimiento del almacenamiento y a la consistencia en el rendimiento de acceso aleatorio mediante el E/S. Cuando creas un volumen, especificas una tasa de IOPS y después Amazon EBS lo entrega con un 10 por ciento del rendimiento de IOPS provisionado el 99.9 por ciento del tiempo, durante un año determinado. Un volumen de IOPS provisionado (SSD) puede variar en tamaño desde 4 GB hasta 16 TB, y puedes aprovisionar hasta 20,000 IOPS por volumen. La proporción de IOPS provisionados con respecto al tamaño del volumen solicitado puede ser como máximo de 30. Por ejemplo, un volumen con 3,000 IOPS debe tener al menos 100 GB de tamaño. Los volúmenes de IOPS provisionados (SSD) tienen un rango límite de rendimiento de 256 KB por cada IOPS provisionado, hasta un máximo de 320 MB/segundo (a 1,280 IOPS).

Las arquitecturas comentadas en este documento utilizan volúmenes EBS, ya que son los más adecuados para las cargas de trabajo de producción que requieren poca latencia y que además sea predecible en las operaciones de entrada/salida por segundo (IOPs) y en el rendimiento. Cuando se selecciona un tipo de Máquina virtual (VM) en particular, se debe tener cuidado, porque no todos los tipos de instancia EC2 pueden tener acceso al almacenamiento EBS.
Nota: Debido a que los volúmenes de Amazon EBS son dispositivos conectados a la red, otra red de E/S desarrollada por una instancia de Amazon EC2, y también la carga total de la red compartida, pueden afectar el rendimiento de los volúmenes individuales de Amazon EBS. Para permitir que tus instancias de Amazon EC2 utilicen completamente los IOPs Provisionados en un volumen de Amazon EBS, puedes ejecutar los tipos de instancias seleccionadas en Amazon EC2 como instancias optimizadas de Amazon EBS.

Los detalles sobre los volúmenes de EBS se pueden ver aquí.

Almacenamiento de instancias EC2 (unidades efímeras)

El almacenamiento de instancias EC2 consiste en bloques de almacenamiento en un disco, el cual se configuró e instaló previamente en el mismo servidor físico que aloja la instancia operativa de Amazon EC2. La cantidad que se proporciona para el almacenamiento en el disco varía según el tipo de instancia en Amazon EC2. En las familias de instancias Amazon EC2 que proporcionan almacenamiento de instancias, las instancias más grandes tienden a proporcionar más volúmenes y una mayor capacidad de almacenamiento.

En las familias de Amazon EC2, existen instancias de almacenamiento optimizado (I2) y de almacenamiento intensivo (D2), que proporcionan almacenamiento de instancias que tienen un propósito especial, que están orientadas a casos de uso específicos. Por ejemplo, las instancias I2 brindan almacenamiento muy rápido en discos duros sólidos (SSD), capaces de mantener más de 365,000 IOPS de lectura aleatoria y 315,000 IOPS de escritura, con modelos a precios muy atractivos.

A diferencia de los volúmenes EBS, el almacenamiento no es permanente y solo se puede utilizar durante el tiempo de vida útil de la instancia, además no se puede separar o adjuntar a otra instancia. El almacenamiento de instancias está pensado para el almacenamiento temporal de la información que cambia constantemente. En el ámbito de las tecnologías y los productos de InterSystems, el uso de InterSystems IRIS o Health Connect como un Bus de servicios empresariales (ESB), los servidores de aplicaciones que utilizan Enterprise Cache Protocol (ECP), o el uso de servidores web con Web/CSP Gateway son excelentes ejemplos en los que podría utilizarse este tipo de almacenamiento. Además, las instancias que se optimizaron para este tipo de almacenamiento, junto con el uso de las herramientas para el aprovisionamiento y la automatización simplifican su eficacia y son compatibles con su elasticidad.

Los detalles sobre los volúmenes de volúmenes de almacenamiento de instancias se pueden ver aquí.

Cómputo

Instancias EC2

Existen numerosos tipos de instancias disponibles que están optimizadas para utilizarse en diferentes casos. Los tipos de instancia permiten combinaciones variadas de CPU, memoria, almacenamiento y networking, lo que permite formar una infinidad de combinaciones para dimensionar de forma correcta los recursos necesarios para tu aplicación.

En este documento, haré referencia a los tipos de instancias M4 de uso general en Amazon EC2 como maneras para ajustar el tamaño de un entorno, ya que estas instancias proporcionan capacidades y optimizaciones del volumen EBS. Las alternativas son posibles en función de los requerimientos de capacidad y el precio de los modelos de tu aplicación.

Las instancias M4 son la generación más reciente de las instancias de uso general . Esta familia proporciona un equilibrio entre cómputo, memoria y recursos de red, y es una buena opción para muchas aplicaciones. El rango de sus capacidades varía desde 2 hasta 64 procesadores virtuales y desde 8 hasta 256 GB de memoria, con el correspondiente ancho de banda dedicado a EBS.

Además de los tipos de instancias individuales, también existen clasificaciones por niveles como los Host dedicados, las instancias de Spot, las instancias reservadas y las instancias dedicadas, cada una de ellas con diferentes rangos de precio, rendimiento e independencia.

Se puede confirmar la disponibilidad y los detalles de las instancias disponibles actualmente aquí.

Disponibilidad y Operaciones

Balance de carga en servidores web/aplicaciones

Puede ser necesario equilibrar la carga en los servidores web externos e internos para tus aplicaciones basadas en InterSystems IRIS. Los balanceadores de carga externos se utilizan para acceder a Internet o a las redes de área extensa (como VPN o Direct Connect) y los balanceadores de carga internos se utilizan potencialmente para

el tráfico interno. El balance de carga elástico de AWS ofrece dos tipos de balanceadores: el de carga para aplicaciones y el de carga clásico.

Balanceador de carga clásico

El balanceador de carga clásico enruta el tráfico basado en la información a nivel de la aplicación o de la red, y es perfecto para balancear las cargas simples de tráfico a lo largo de varias instancias EC2 donde se necesite de una alta disponibilidad, escalado automático y seguridad robusta. Los detalles y las características específicas pueden encontrarse aquí.

Balanceador de carga para aplicaciones

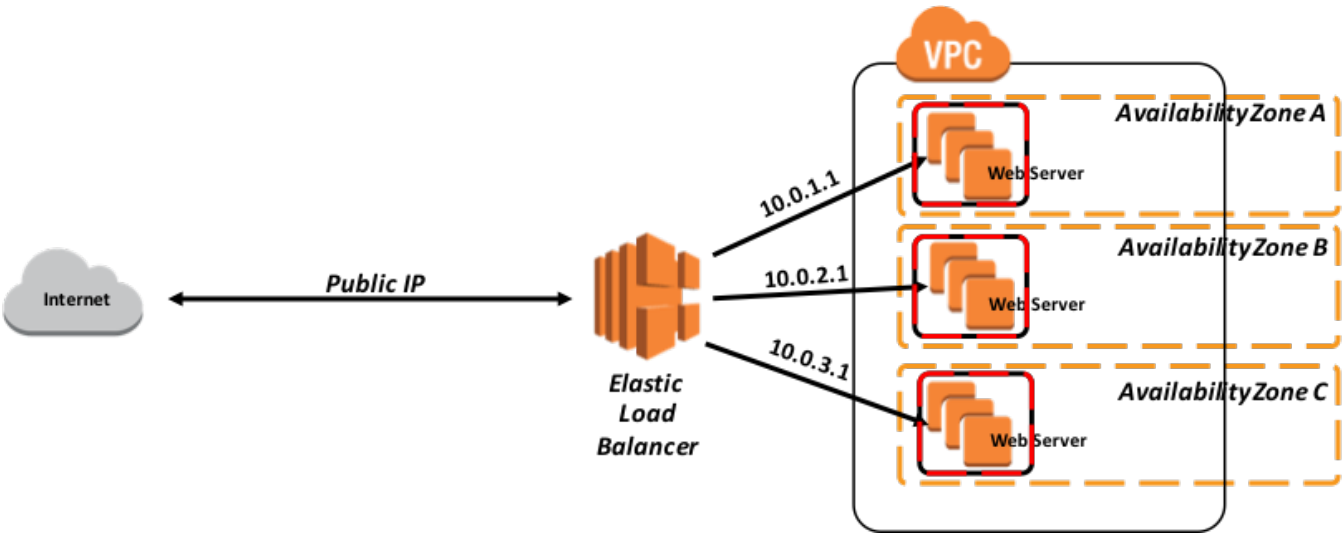
Un Balanceador de carga para aplicaciones es una opción adicional para el servicio que proporciona el Balanceador de carga elástico, que opera en la capa de la aplicación y permite definir reglas de enrutamiento basadas en el contenido de varios servicios o en los contenedores que se ejecutan en una o más instancias de Amazon EC2. Además, hay soporte para WebSockets y HTTP/2. Los detalles y las características específicas pueden encontrarse aquí.

Ejemplo

En el siguiente ejemplo, se define un conjunto de tres servidores web con una zona de disponibilidad separada para cada uno, a fin de que proporcionen los niveles más altos de disponibilidad. Los balanceadores de carga del servidor web deben estar configurados con *Sesiones sticky* para permitir la capacidad de fijar las sesiones del usuario en instancias específicas de EC2 utilizando cookies. El tráfico se enrutará hacia las mismas instancias conforme el usuario continúe accediendo a tu aplicación.

En el siguiente diagrama, la Figura 4 muestra un ejemplo simple del Balanceador de Carga Clásico dentro de AWS.

Figura 4: Ejemplo de un Balanceador de Carga Clásico



Mirroring de la bases de datos

Cuando se hace la implementación de InterSystems IRIS con base en las aplicaciones de AWS, proporcionar una alta disponibilidad para el servidor de la plataforma de datos requiere que se haga un mirror sincrónico de la base de datos, con la finalidad de ofrecer una alta disponibilidad en una región principal de AWS dada, y potencialmente, un mirror asincrónico de la base de datos para replicar datos hacia un modo de espera activa (hot standby) en una región secundaria de AWS para la recuperación en caso de desastre, pero esto dependerá de los requisitos del contrato de servicio durante el tiempo de funcionamiento.

Un mirror de la base de datos consiste en una agrupación lógica de dos sistemas de base de datos, conocidos como miembros de respuesta ante fallos. Son sistemas físicamente independientes que están conectados solamente por una red. Después de mediar entre los dos sistemas, el mirror designa automáticamente a uno de ellos como el sistema principal, mientras que el otro miembro se convierte automáticamente en el sistema de backup. Las estaciones de trabajo externas de los clientes u otros equipos se conectan a la réplica mediante la IP virtual (VIP) de la réplica, que se especifica durante la configuración. El mirror VIP se une automáticamente a una interfaz en el sistema principal del mirror.
Nota: En AWS, no es posible configurar el mirror VIP de la forma tradicional, por lo que se diseñó una solución alternativa. Sin embargo, el mirroring es compatible con todas las subredes.

La recomendación actual para implementar un mirror de la base de datos en AWS es configurar tres instancias (la primaria, la de bakcup y la de mediación) en la misma Red virtual privada en la nube (VPC) a través de tres zonas diferentes de disponibilidad. Esto asegura que, en cualquier momento, AWS garantizará la conectividad externa mediante al menos dos de estas máquinas virtuales con un SLA del 99.95%. Esto proporciona la independencia necesaria y la redundancia de los datos en la misma base de datos. Se puede encontrar información detallada sobre los contratos de servicio (SLA) para AWS EC2 aquí.

No hay un límite superior estricto para la latencia de la red entre los miembros de respuesta ante fallos. **El impacto que tenga aumentar la latencia difiere según la aplicación**. Si el tiempo de ida y vuelta entre los miembros de respuesta ante fallos es similar al tiempo de servicio de escritura en el disco, no se espera que haya ningún impacto. Sin embargo, el tiempo de ida y vuelta puede ser una preocupación cuando la aplicación debe esperar a que los datos se vuelvan persistentes (a veces conocido como registro de sincronización). Más detalles sobre el mirroring de la base de datos y la latencia de la red pueden encontrarse aquí.

Dirección IP Virtual IP y Sistema automático de respuesta ante fallos

La mayoría de los proveedores de IaaS en la nube carecen de la capacidad para proporcionar una dirección IP virtual (VIP), que es la que normalmente se utiliza en los diseños de respuesta ante fallos para las bases de datos. Para solucionar este problema, varios de los métodos de conectividad que se utilizan más frecuentemente, específicamente para los clientes ECP y Web Gateway, han sido mejorados dentro de InterSystems IRIS y HealthShare para que ya no dependan de las capacidades de la VIP, de manera que puedan darse cuenta del mirror.

Los métodos de conectividad como xDBC, sockets TCP/IP directos, u otros protocolos de conexión directa, aún requieren el uso de una VIP. Para solucionar estos problemas, la tecnología de mirroring de bases de datos de InterSystems permite ofrecer una respuesta automática ante fallos para esos métodos de conectividad dentro de AWS. Todo esto lo hace posible utilizando las API para interactuar con un Balanceador de carga elástico (ELB) de AWS, con el fin de lograr una funcionalidad que sea similar a la de una VIP, brindando así un diseño completo, robusto y de alta disponibilidad en AWS.

Además, AWS introdujo recientemente un nuevo tipo de ELB llamado Balanceador de carga de aplicaciones. Este tipo de balanceador se ejecuta en el *Layer* 7 y admite el enrutamiento basado en el contenido, además de las aplicaciones que se ejecutan en contenedores. El enrutamiento basado en el contenido es especialmente útil para los proyectos de tipo Big Data que utilizan el particionamiento de los datos o aquellos datos en los que se implementó la técnica de sharding.

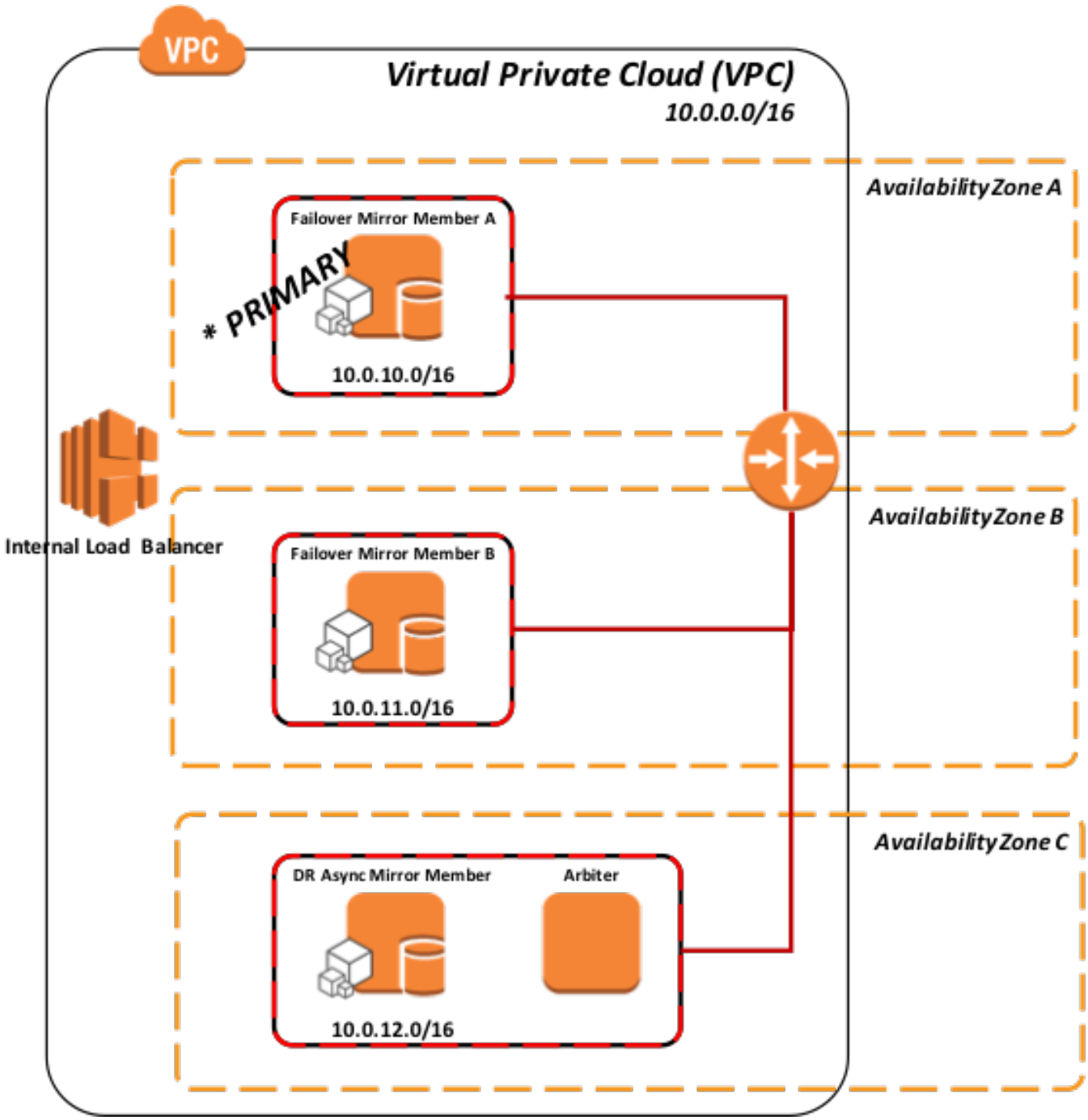
Al igual que con la IP virtual, este es un cambio abrupto en la configuración de la red y no implica ninguna lógica en las aplicaciones para informar a los clientes que ya existen y están conectados al miembro "mirror" principal que falló, o sea que se produjo un *failover*. Dependiendo de la naturaleza del fallo, esas conexiones pueden terminar como consecuencia del fallo en sí, debido al tiempo de espera o por error de la aplicación, como resultado de que la nueva instancia principal obliga a la vieja instancia principal a dejar de funcionar, o por el vencimiento del temporizador TCP *keep-alive* utilizado por el cliente.

Como consecuencia, es posible que los usuarios tengan que volver a conectarse e iniciar sesión. El funcionamiento que tenga tu aplicación es el que determinaría este comportamiento. Los detalles acerca de los diferentes tipos de ELB que están disponibles pueden encontrarse aquí.

Llamada de la instancia en AWS EC2 para el método del balanceador de carga elástico de AWS

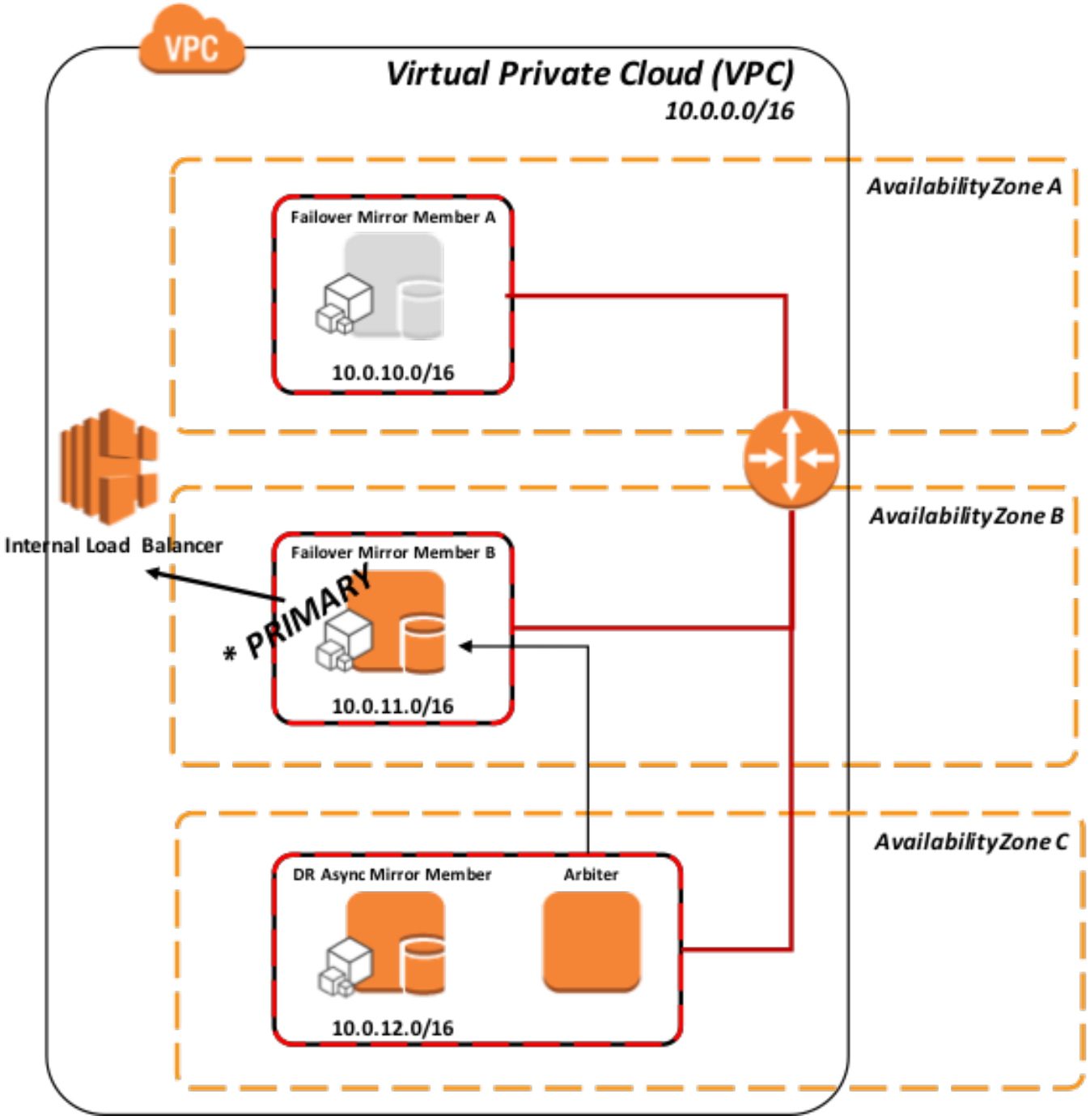
En este modelo, el ELB puede tener un grupo de servidores definido ya sea por miembros mirror que participan en un procedimiento contra fallos, como por miembro(s) mirror asíncronos que se utilizan potencialmente durante la Recuperación en caso de desastres (DR) con solo una de las entradas activas, que es el miembro principal actual, o solo un grupo de servidores con una única entrada para el miembro mirror activo.

Figura 5: Método de la API para interactuar con el Balanceador de Carga Elástico (interno)



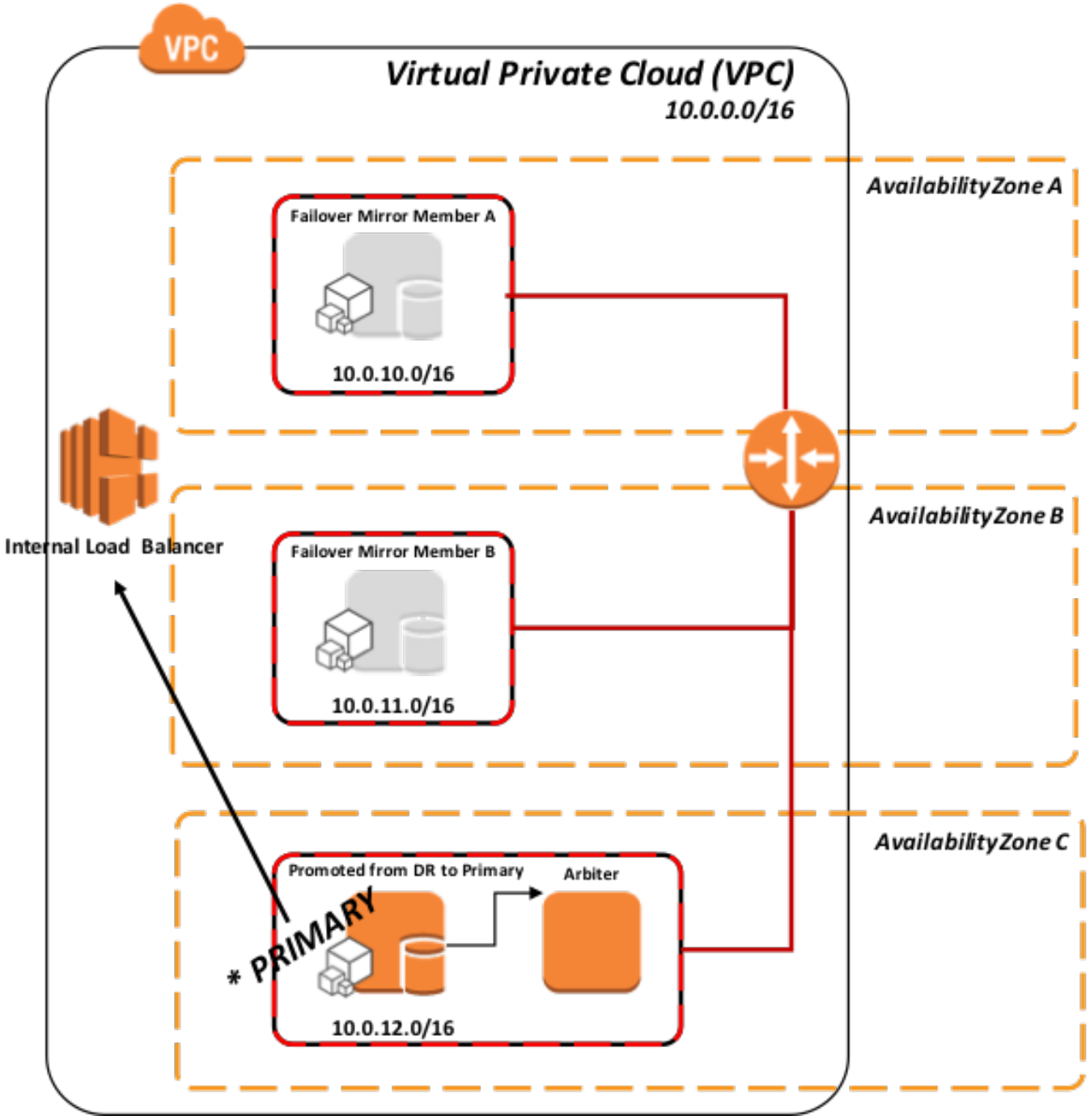
Cuando un miembro mirror se convierte en el miembro mirror principal, se emite una llamada de la API desde su instancia de EC2 hacia el AWS ELB para ajustar/dar instrucciones al ELB del nuevo miembro mirror principal.

Figura 6: Respuesta ante fallos de un Miembro B usando una API con el Balanceador de Carga



Este mismo modelo se aplica a la promoción de un miembro (asincrónico) de recuperación en caso de desastres (DR) en caso de que no estén disponibles tanto el miembro mirror principal como el de backup.

Figura-7: Promoción del miembro DR Asincrónico (DR) a miembro primário mediante API del Balanceador



Según el procedimiento estándar de DR recomendado, en la Figura 6 anterior, la promoción del miembro de DR implica una decisión humana debido a la posibilidad de pérdida de datos por la replicación asincrónica. Sin embargo, una vez que se toma esa acción, no se requiere ninguna acción administrativa en el ELB. Enruta automáticamente el tráfico una vez que se llama a la API durante la promoción.

Detalles de la API

Esta API para llamar al recurso del balanceador de carga de AWS se define en la rutina ^ ZMIRROR específicamente como parte de la llamada al procedimiento:

$$CheckBecomePrimaryOK^ZMIRROR()

Dentro de este procedimiento, inserte cualquier lógica de API o métodos que elija usar desde la API REST de AWS ELB, la interfaz de línea de comandos, etc. Una forma efectiva y segura de interactuar con ELB es usar roles de AWS Identity and Access Management (IAM) para que pueda no tiene que distribuir credenciales a largo plazo a una instancia EC2. El rol de IAM proporciona permisos temporales que InterSystems IRIS puede usar para

interactuar con AWS ELB. Los detalles para usar las funciones de IAM asignadas a sus instancias EC2 se pueden encontrar aquí.

Método de sondeo de AWS Elastic Load Balancer
Un método de sondeo que utiliza la página mirrorstatus.cxw de Web Gateway se puede utilizar como método de sondeo en el monitor de estado de ELB para cada miembro agregado al grupo de servidores de ELB. Solo el miembro primario responderá "Éxito", dirigiendo así el tráfico de red solo al miembro primario activo.

Este método no requiere que se agregue ninguna lógica a ^ ZMIRROR. Hay que tener en cuenta que la mayoría de los dispositivos de red de balanceo de carga tienen un límite en la frecuencia de ejecución de la verificación de estado. Normalmente, la frecuencia más alta no es inferior a 5 segundos, lo que suele ser aceptable para admitir la mayoría de los acuerdos de nivel de servicio de tiempo de actividad.

Una solicitud HTTP para el siguiente recurso probará el estado de miembro mirror de la configuración LOCAL de InterSystems IRIS.

/csp/bin/mirrorstatus.cxw

Para todos los demás casos, la ruta a estas solicitudes de estado Mirror debe resolverse en el servidor y NameSpace apropiados utilizando el mismo mecanismo jerárquico que se utiliza para solicitar páginas CSP reales.

Ejemplo: para probar el estado del Mirror de las aplicaciones de servicio de configuración en la ruta /csp/user/:

/csp/user/mirrorstatus.cxw
Nota: No se consumen licencias CSP al invocar la revisión del Mirror Status.

Dependiendo de si la instancia de destino es el miembro principal activo o no, el Gateway devolverá una de las siguientes respuestas de CSP:

**** Éxito (Miembro Primário)**

```
==============================
  HTTP/1.1 200 OK
  Content-Type: text/plain
  Connection: close
  Content-Length: 7
  SUCCESS
```

**** Fallo (No es el Miembro Primário)**

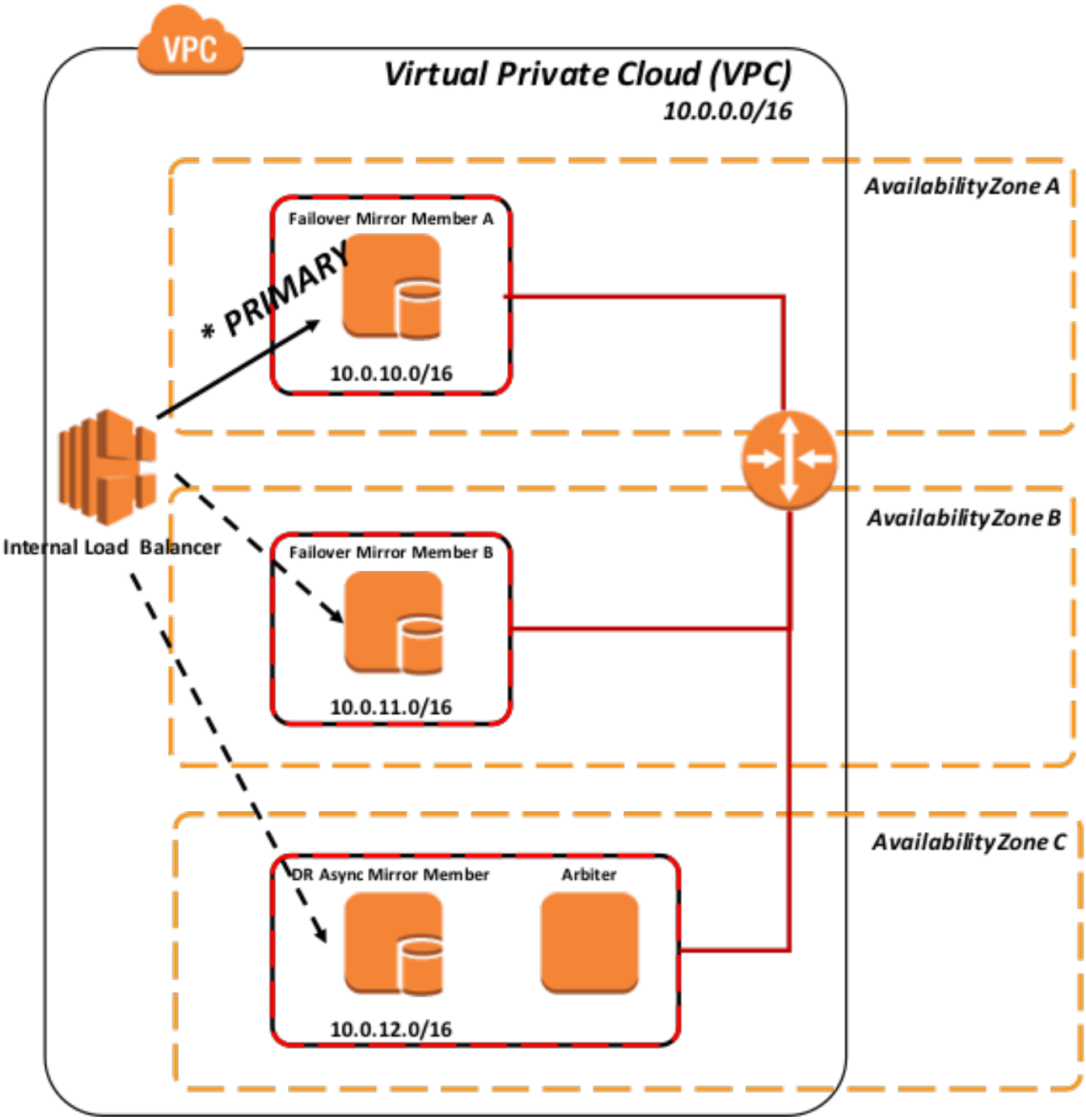
```
==============================
  HTTP/1.1 503 Service Unavailable
  Content-Type: text/plain
  Connection: close
  Content-Length: 6
  FAILED
```

**** Fallo (El servidor no soporta la petición *MirrorStatus.cxw*)**

```
==============================

HTTP/1.1 500 Internal Server Error

Content-Type: text/plain

Connection: close

Content-Length: 6

FAILED
```

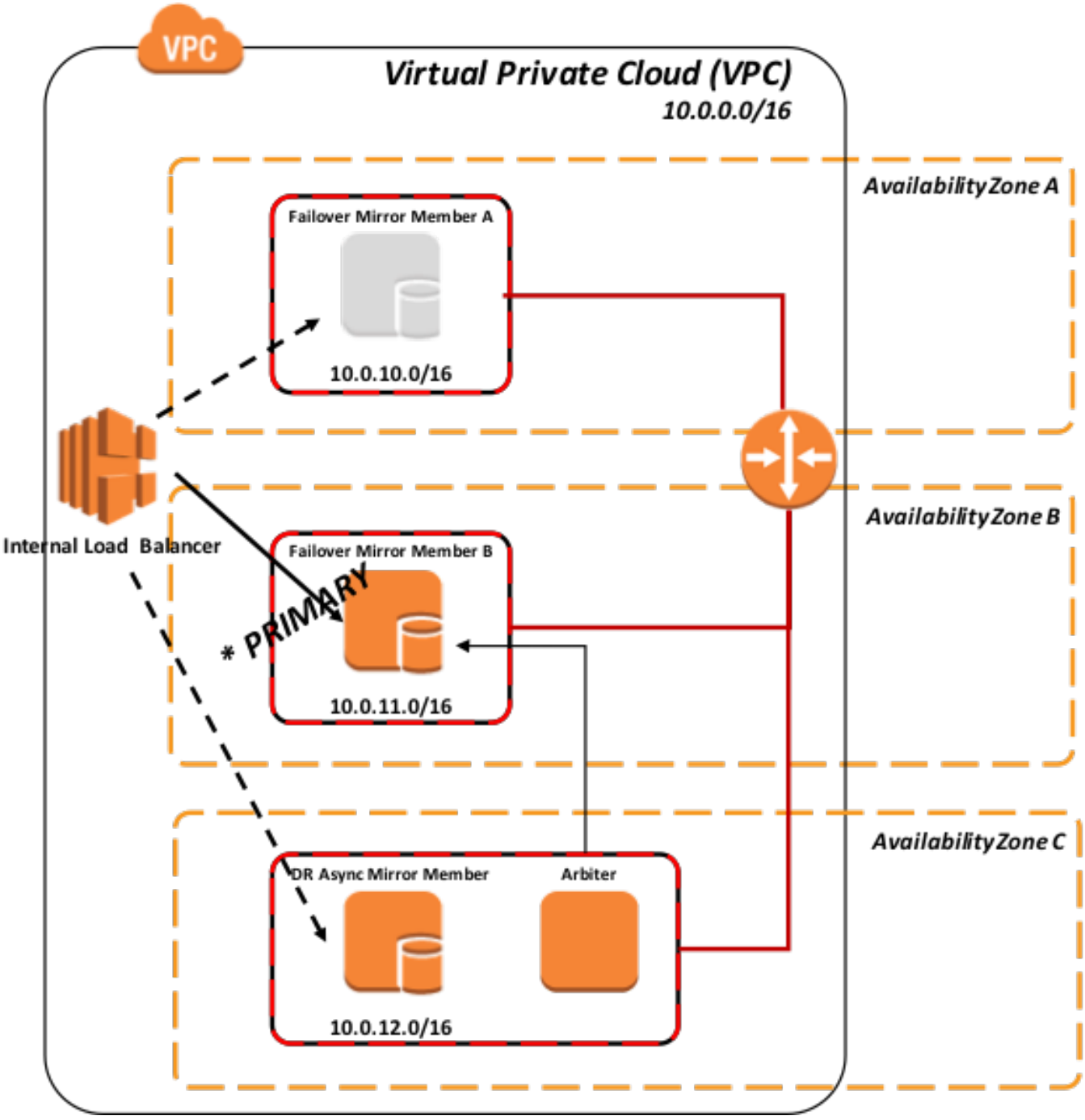
Las siguientes figuras ilustran los distintos escenarios de uso del método de sondeo.

Figura-8: Sondeo a todos los miembros del mirror



Como muestra la Figura 8 anterior, todos los miembros están operativos y solo el miembro principal está devolviendo "Éxito" al balanceador de carga, por lo que el tráfico de red se dirigirá solo a este miembro.

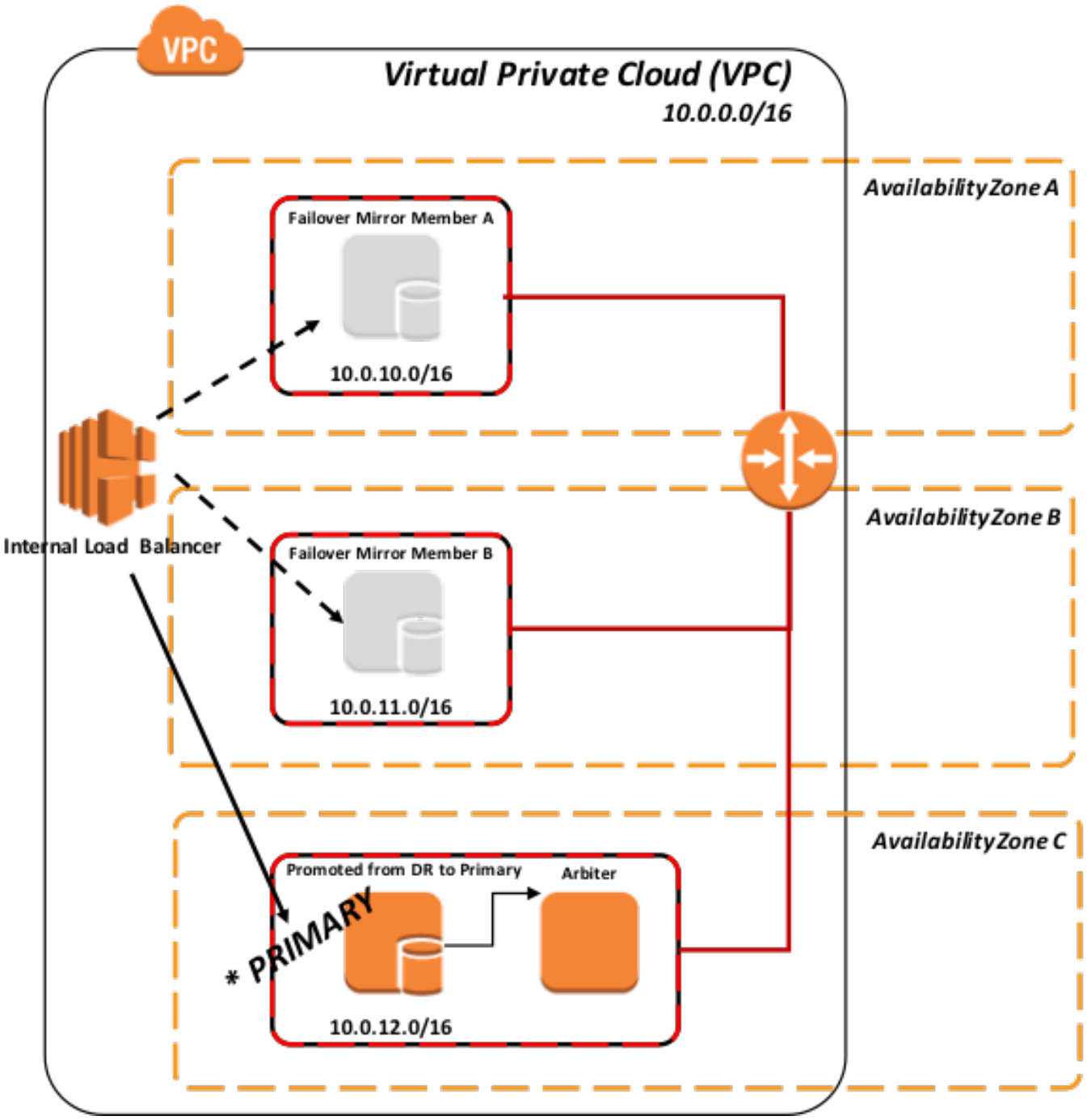
Figura-9: Failover al Miembro B por medio del sondeo



El siguiente diagrama demuestra la promoción de miembros asíncronos de recuperación ante desastres en el grupo con balanceo de carga, esto normalmente supone que el mismo dispositivo de red con balanceo de carga está dando servicio a todos los miembros (los escenarios divididos geográficamente se tratan más adelante en este artículo).

Según el procedimiento estándar de DR recomendado, la promoción del miembro de DR implica una decisión humana debido a la posibilidad de pérdida de datos por a la replicación asincrónica. Sin embargo, una vez que se toma esa acción, no se requiere ninguna acción administrativa en el ELB. Descubre automáticamente el nuevo primario.

Figura-10: Failover y Promoción del DR Asincrónico mediante sondeo



Backup y Restore

Hay varias opciones disponibles para las operaciones de backup. Las siguientes tres opciones son viables para su implementación de AWS con productos InterSystems. Las dos primeras opciones incorporan un procedimiento de tipo *snapshot* que implica suspender las escrituras de la base de datos en el disco antes de la creación del *snapshot* y luego reanudar las actualizaciones una vez que se haya realizado correctamente. Se toman los siguientes pasos de alto nivel para crear una copia de seguridad limpia utilizando cualquiera de los métodos de *snapshot*:

- Pausar las escrituras en la base de datos mediante la llamada a la API Freeze de la base de datos.
- Cree *snapshots* del sistema operativo + discos de datos.
- Reanudar las escrituras a través de la llamada API de la base de datos.
- Copia del backup a la ubicación de la copia de seguridad

Se pueden agregar pasos adicionales, como verificaciones de integridad, en un intervalo periódico para garantizar una copia de seguridad limpia y consistente.

Los puntos de decisión sobre qué opción utilizar dependen de los requisitos operativos y las políticas de su organización. InterSystems está disponible para discutir las diversas opciones con más detalle.

EBS Snapshots

Los *snapshots* de EBS son formas muy rápidas y eficientes de crear una imagen en un momento determinado en un almacenamiento de Amazon S3 de alta disponibilidad y menor costo. Los *snapshots* de EBS junto con las capacidades de API de congelación y descongelación externas de InterSystems permiten una verdadera resistencia operativa 24x7 y la garantía de copias de seguridad regulares limpias. Existen numerosas opciones para automatizar el proceso utilizando los servicios proporcionados por AWS, como Amazon CloudWatch Events o soluciones de terceros disponibles en el mercado, como Cloud Ranger o N2W Software Cloud Protection Manager, por nombrar algunos.

Además, puede crear mediante programación su propia solución de copia de seguridad personalizada mediante el uso de llamadas API directas de AWS. Los detalles sobre cómo aprovechar las API están disponibles aquí y aquí.
Nota: InterSystems no respalda ni valida explícitamente ninguno de estos productos de terceros. Las pruebas y la validación dependen del cliente.

Logical Volume Manager Snapshots

Alternativamente, muchas de las herramientas de copia de seguridad de terceros disponibles en el mercado pueden utilizarse mediante la implementación de agentes de copia de seguridad individuales dentro de la propia máquina virtual y aprovechando las copias de seguridad a nivel de archivo junto con las instantáneas de Linux Logical Volume Manager (LVM) o el Servicio de Snapshots de Volumen de Windows (VSS).

Uno de los principales beneficios de este modelo es la capacidad de tener restauraciones a nivel de archivo de instancias basadas en Linux y Windows. Un par de puntos a tener en cuenta con esta solución; Dado que AWS y la mayoría de los demás proveedores de nube de IaaS no proporcionan medios de cinta, todos los repositorios de respaldo están basados en disco para el archivado a corto plazo y tienen la capacidad de aprovechar el almacenamiento de bajo costo de Amazon S3 y, finalmente, Amazon Glacier para la retención a largo plazo (LTR). Si utiliza este método, se recomienda encarecidamente utilizar un producto de respaldo que admita tecnologías de deduplicación para hacer el uso más eficiente de los repositorios de respaldo basados en disco.

Algunos ejemplos de estos productos de respaldo con soporte en la nube incluyen, entre otros: Commvault, EMC Networker, HPE Data Protector y Veritas Netbackup.
Nota: InterSystems no respalda ni valida explícitamente ninguno de estos productos de terceros. Las pruebas y la validación dependen del cliente

Copia de Seguridad en Línea

Para implementaciones pequeñas, la función integrada de backup en línea también es una opción viable. La utilidad de backup en línea de la base de datos de InterSystems respalda los datos en los archivos de la base de datos capturando todos los bloques en las bases de datos y luego escribe la salida en un archivo secuencial. Este mecanismo de backup patentado está diseñado para no causar tiempo de inactividad a los usuarios del sistema de producción.

En AWS, una vez finalizada la copia de seguridad en línea, el archivo de salida de la copia de seguridad y todos los demás archivos en uso por el sistema deben copiarse en un EC2 que actúa como un recurso compartido de archivos (CIFS/NFS). Este proceso debe programarse y ejecutarse dentro de la máquina virtual.

La copia de seguridad en línea es el enfoque de nivel de entrada para los sitios más pequeños que desean implementar una solución de bajo costo para la copia de seguridad. Sin embargo, a medida que las bases de datos aumentan de tamaño, las copias de seguridad externas con tecnología de instantáneas se recomiendan como una mejor práctica con ventajas que incluyen la copia de seguridad de archivos externos, tiempos de

restauración más rápidos y una vista de los datos y herramientas de administración de toda la empresa.

Recuperación ante Desastres (DR)

Al implementar una aplicación basada en InterSystems IRIS en AWS, se recomienda que los recursos de recuperación ante desastres, incluida la red, los servidores y el almacenamiento, estén en una región de AWS diferente o en zonas de disponibilidad separadas como mínimo. La capacidad requerida en la región DR AWS designada depende de las necesidades de su organización. En la mayoría de los casos, se requiere el 100% de la capacidad de producción cuando se opera en modo DR; sin embargo, se puede aprovisionar menor capacidad hasta que se necesite más como modelo elástico. La menor capacidad puede presentarse en forma de menos servidores web y de aplicaciones y, potencialmente, incluso se puede utilizar un tipo de instancia EC2 más pequeño para el servidor de base de datos y, tras la promoción, los volúmenes de EBS se adjuntan a un tipo de instancia EC2 grande.

La duplicación de bases de datos asincrónica se utiliza para replicar continuamente en las instancias EC2 de la región DR AWS. La duplicación utiliza *journals* de transacciones de la base de datos para replicar las actualizaciones a través de una red TCP / IP de una manera que tiene un impacto mínimo en el rendimiento del sistema principal. Se recomienda encarecidamente configurar la compresión y el cifrado de archivos de *journals* con estos miembros asíncronos de DR.

Todos los clientes externos en la Internet pública que deseen acceder a la aplicación serán enrutados a través de Amazon Route53 como un servicio DNS adicional. Amazon Route53 se utiliza como conmutador para dirigir el tráfico al centro de datos activo actual. Amazon Route53 realiza tres funciones principales:

- **Registro de dominio:** Amazon Route53 le permite registrar nombres de dominio como example.com.
- **Servicio de sistema de nombres de dominio (DNS)**: Amazon Route53 traduce nombres de dominios sencillos como www.example.com en direcciones IP como 192.0.2.1. Amazon Route53 responde a las consultas de DNS mediante una red global de servidores DNS autorizados, lo que reduce la latencia.
- **Verificación de estado**: Amazon Route53 envía solicitudes automatizadas a través de Internet a su aplicación para verificar que sea accesible, disponible y funcional.

Los detalles de estas funciones se pueden encontrar aquí.

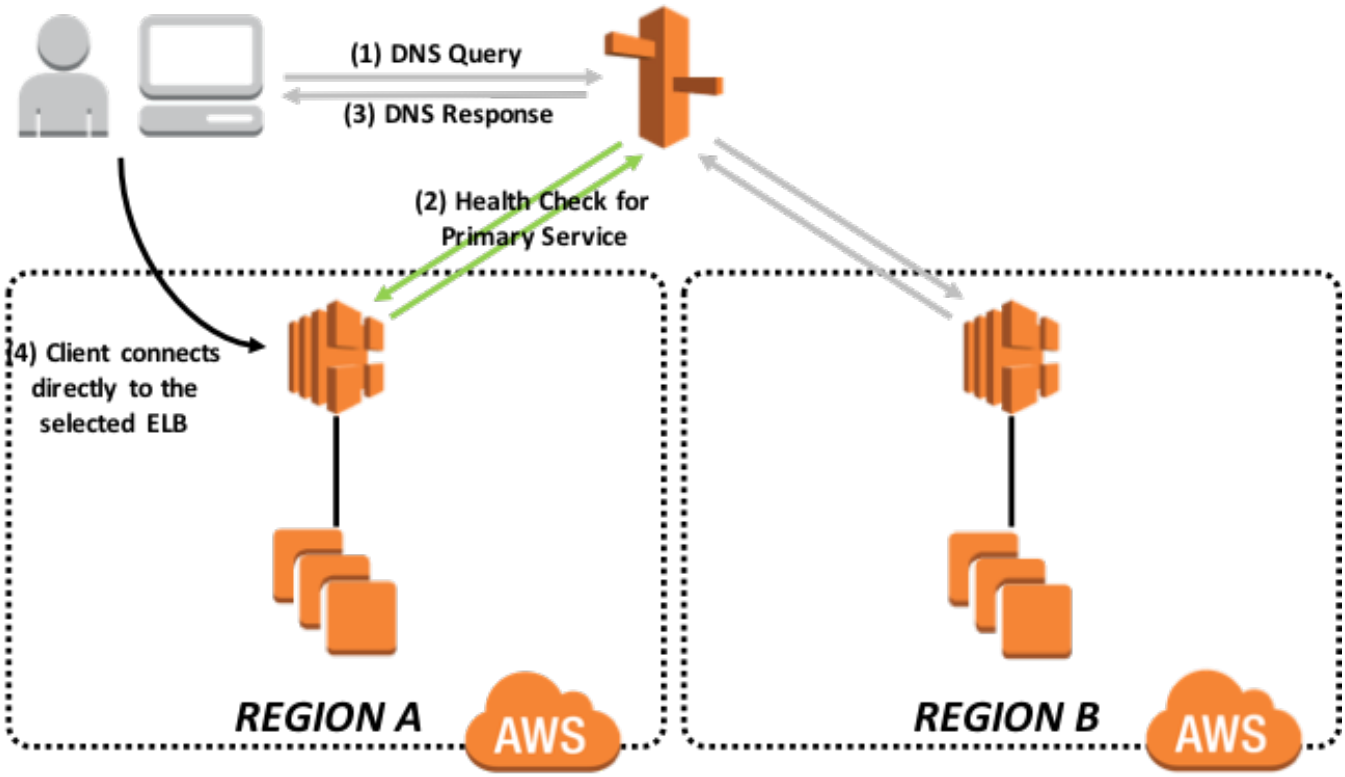
A los efectos de este documento, se analizarán la conmutación por error de DNS y la verificación del estado de Route53. Los detalles de la supervisión de Health Check y la conmutación por error de DNS se pueden encontrar aquí y aquí.

Route53 funciona realizando solicitudes regulares a extremo y luego verificando la respuesta. Si un extremo no proporciona una respuesta válida, ya no se incluye en las respuestas de DNS, que en su lugar devolverá un extremo alternativo disponible. De esta manera, el tráfico de usuarios se dirige lejos de los extremos que fallan y hacia los extremos que están disponibles.

Usando los métodos anteriores, el tráfico solo se permitirá a una región específica y un miembro específico. Esto está controlado por la definición de extremo, que es una página mirrorstatus.cxw discutida anteriormente en este artículo presentado desde InterSystems Web Gateway. Solo el miembro principal informará "Éxito" como HTTP 200 de la verificación de estado.

El siguiente diagrama muestra a alto nivel la política de enrutamiento de conmutación por error. Los detalles de este método y otros se pueden encontrar aquí.

Figura-11: Política de Rutina de Conmutación por Error de Amazon Route53



En un momento dado, solo una de las regiones informará el monitoreo del extremo en tiempo real. Esto asegura que el tráfico solo fluya a una región en un momento dado. No se necesitan pasos adicionales para la conmutación por error entre las regiones, ya que la supervisión del punto de conexión detectará que la aplicación en la región de AWS primaria designada está inactiva y la aplicación ahora está activa en la región de AWS secundaria. Esto se debe a que el miembro asíncrono de DR se ha promovido manualmente a primario, lo que permite que Web Gateway informe de HTTP 200 a la supervisión del extremo de Elastic Load Balancer.

Existen muchas alternativas a la solución descrita anteriormente y se pueden personalizar según los requisitos operativos de su organización y los acuerdos de nivel de servicio.

Monitorización

Amazon CloudWatch está disponible para brindar servicios de monitoreo para todos sus recursos en la nube de AWS y sus aplicaciones. Amazon CloudWatch se puede utilizar para recopilar y rastrear métricas, recopilar y monitorear archivos de registro, configurar alarmas y reaccionar automáticamente a los cambios en los recursos de AWS. Amazon CloudWatch puede monitorear recursos de AWS como Amazon EC2

Instancias, así como métricas personalizadas generadas por sus aplicaciones y servicios, y cualquier archivo de registro que generen sus aplicaciones. Puede utilizar Amazon CloudWatch para obtener visibilidad de todo el sistema sobre la utilización de recursos, el rendimiento de las aplicaciones y el estado operativo. Detalles pueden ser encontrados aqui.

Aprovisionamiento automatizado

Actualmente, existen numerosas herramientas disponibles en el mercado, incluidas Terraform, Cloud Forms, Open Stack y CloudFormation de Amazon. El uso de estos y el acoplamiento con otras herramientas como Chef, Puppet, Ansible y otros pueden proporcionar DevOps completo de soporte de infraestructura como código o simplemente iniciar su aplicación de una manera completamente automatizada. Los detalles de Amazon CloudFormation se pueden encontrar aquí.

Conectividad de Red

Según los requisitos de conectividad de la aplicación, hay varios modelos de conectividad disponibles mediante Internet, VPN o un enlace dedicado mediante Amazon Direct Connect. El método a elegir dependerá de la aplicación y las necesidades del usuario. El uso de ancho de banda para cada uno de los tres métodos varía, y es mejor consultar con su representante de AWS o con la Consola de administración de Amazon para confirmar las opciones de conectividad disponibles para una región determinada.

Seguridad

Se debe tener cuidado al decidir implementar una aplicación en cualquier proveedor público de nube IaaS. Se deben seguir las políticas de seguridad estándar de tu organización, o las nuevas desarrolladas específicamente para la nube. También deberá comprender la soberanía de los datos, que es relevante cuando los datos de una organización se almacenan fuera de su país y están sujetos a las leyes del país en el que residen los datos. Las implementaciones en la nube tienen el riesgo adicional de que los datos estén ahora fuera de los centros de datos de los clientes y del control de seguridad física. Se recomienda encarecidamente uso de cifras. Tanto para la base de datos y *journals* para datos en reposo como para los datos en tránsito (comunicaciones de red) mediante cifrado AES y SSL/TLS respectivamente.

Al igual que con toda la administración de claves de cifrado, los procedimientos adecuados deben documentarse y seguirse de acuerdo con las políticas de tu organización para garantizar la seguridad de los datos y evitar el acceso no deseado a los datos o las violaciones de seguridad.

Amazon proporciona una amplia documentación y ejemplos para proporcionar un entorno operativo altamente seguro para sus aplicaciones basadas en tecnologia InterSystems. Asegúrese de revisar Identity Access Management (IAM) para conocer los diversos temas de discusión que se encuentran aquí.

Ejemplos de Diagramas de Arquitectura

El diagrama siguiente ilustra una instalación típica de produtos InterSystems que proporcionan alta disponibilidad en forma de duplicación de base de datos (tanto conmutación por error síncrona - *failover* - como DR asincrónica), servidores de aplicaciones que utilizan ECP y varios servidores web con equilibrio de carga.

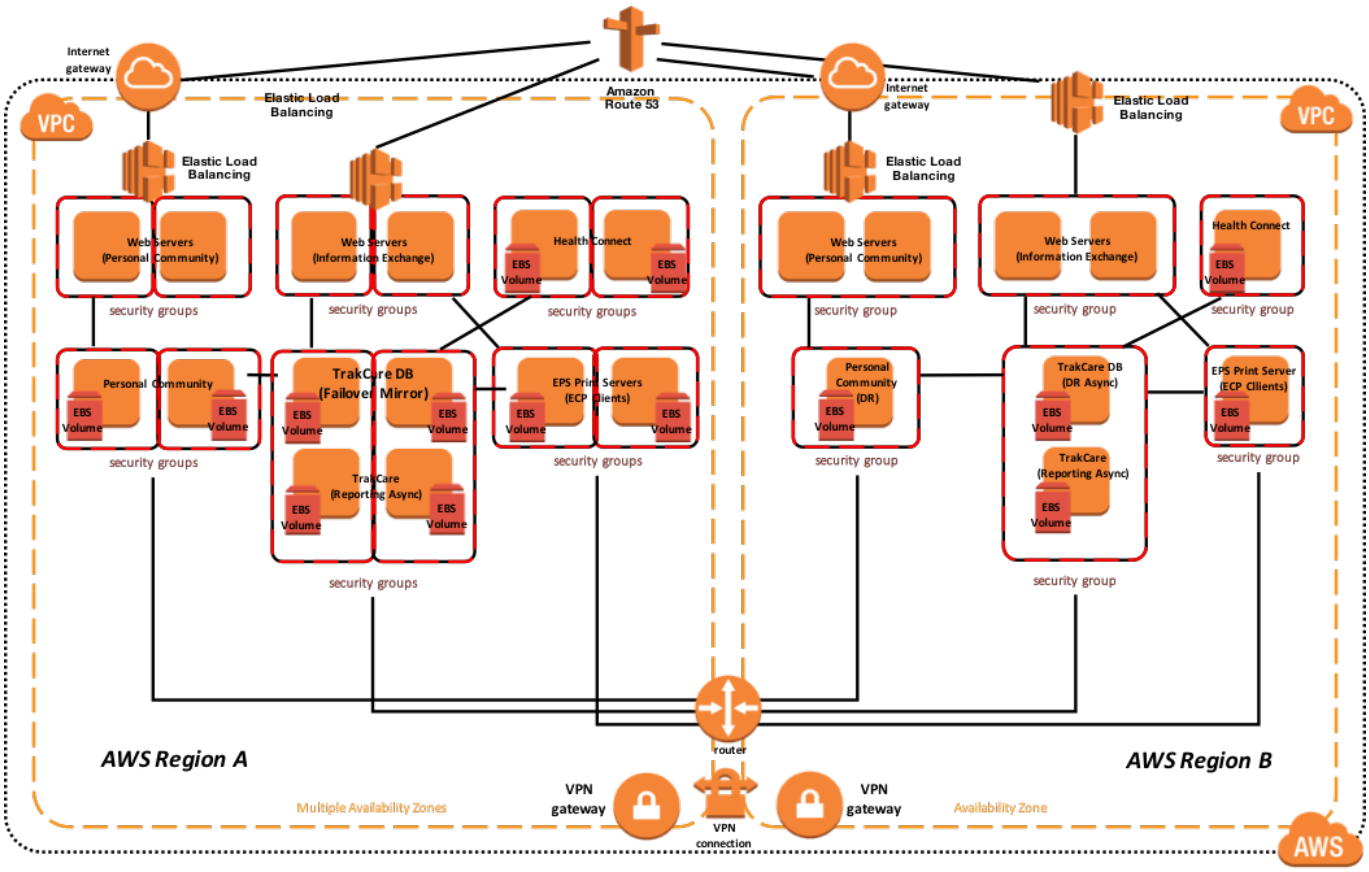
Ejemplo de TrakCare

El siguiente diagrama ilustra una implementación típica de TrakCare con varios servidores web con balanceo de carga, dos servidores de impresión como clientes ECP y una configuración de mirror de base de datos. La dirección IP virtual solo se usa para la conectividad no asociada con ECP o Web Gateway. Los clientes de ECP y Web Gateway son compatibles con el mirror y no requieren una VIP.

Si está utilizando Direct Connect, hay varias opciones que incluyen múltiples circuitos y acceso a múltiples regiones que se pueden habilitar para escenarios de recuperación de desastres. Es importante trabajar con los proveedores de telecomunicaciones para comprender los escenarios de alta disponibilidad y recuperación ante desastres que admiten.

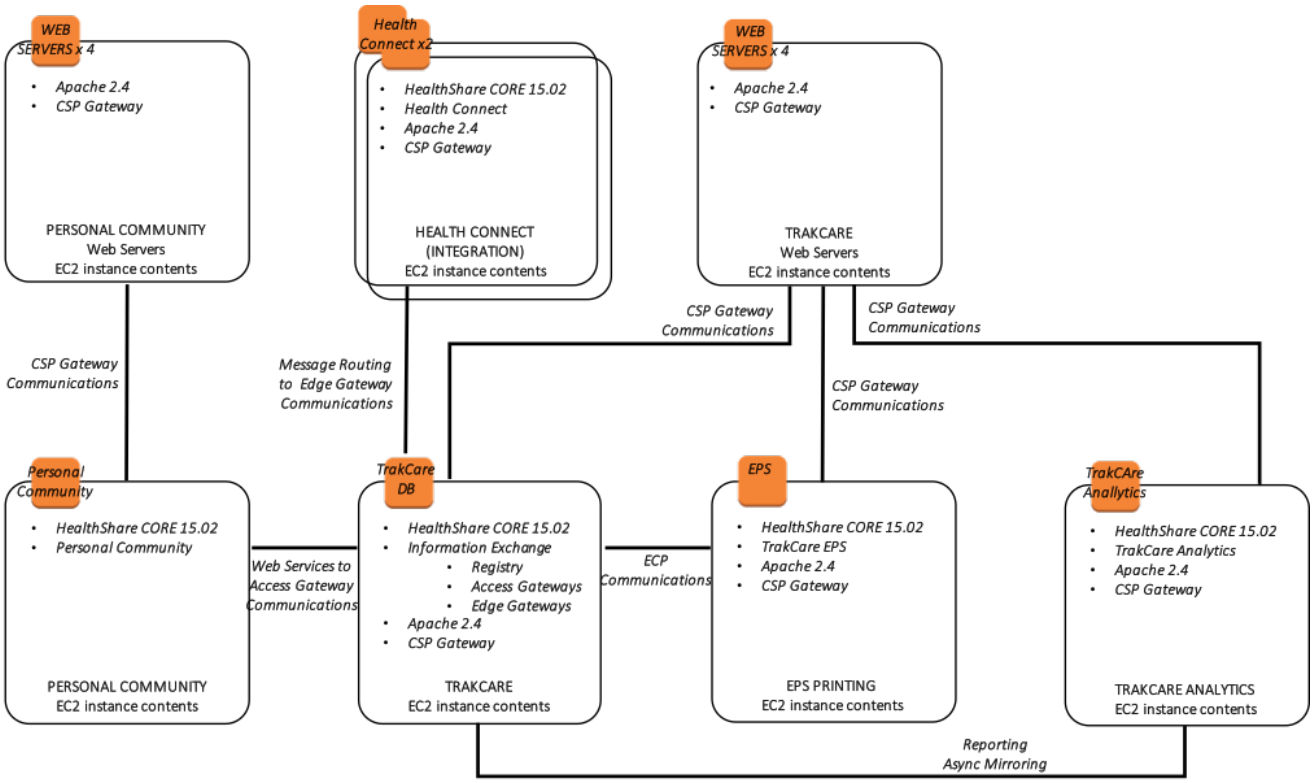
El siguiente diagrama de arquitectura de referencia de muestra incluye alta disponibilidad en la región activa o primaria y recuperación ante desastres en otra región de AWS si la región primaria de AWS no está disponible. También dentro de este ejemplo, los miembros del mirror de la base de datos contienen los namespaces de TrakCare DB, TrakCare Analytics e Integration, todo dentro de ese único mirror.

Figura-12: TrakCare AWS Diagrama de Arquitectura de Referencia – Arquitectura Física



Además, se proporciona el siguiente diagrama que muestra una vista más lógica de la arquitectura con los productos de software de alto nivel asociados instalados y su propósito funcional.

Figura-13: TrakCare Diagrama de Arquitectura de Referencia – Arquitectura Lógica

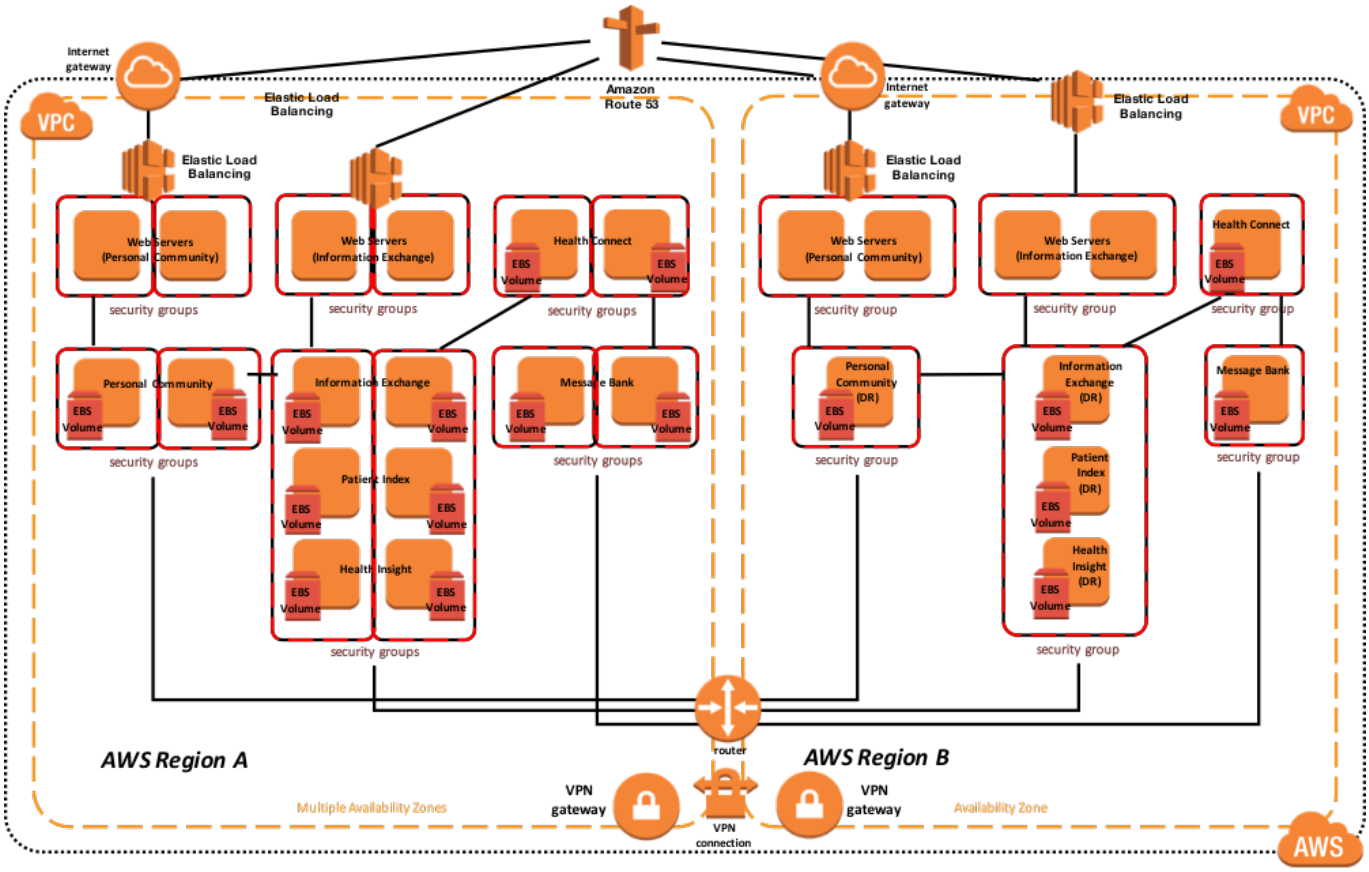


[Ejemplo de HealthShare]

El siguiente diagrama ilustra una implementación típica de HealthShare con varios servidores web con balanceo de carga, con varios productos de HealthShare incluidos: Information Exchange, Patient Index, Personal Community, Health Insight y Health Connect. Cada uno de esos productos respectivos incluye un par de réplicas de bases de datos para alta disponibilidad dentro de múltiples zonas de disponibilidad. La dirección IP virtual solo se usa para la conectividad no asociada con ECP o Web Gateway. Las puertas de enlace Web utilizadas para las comunicaciones de servicios web entre los productos HealthShare son compatibles con el mirror y no requieren un VIP.

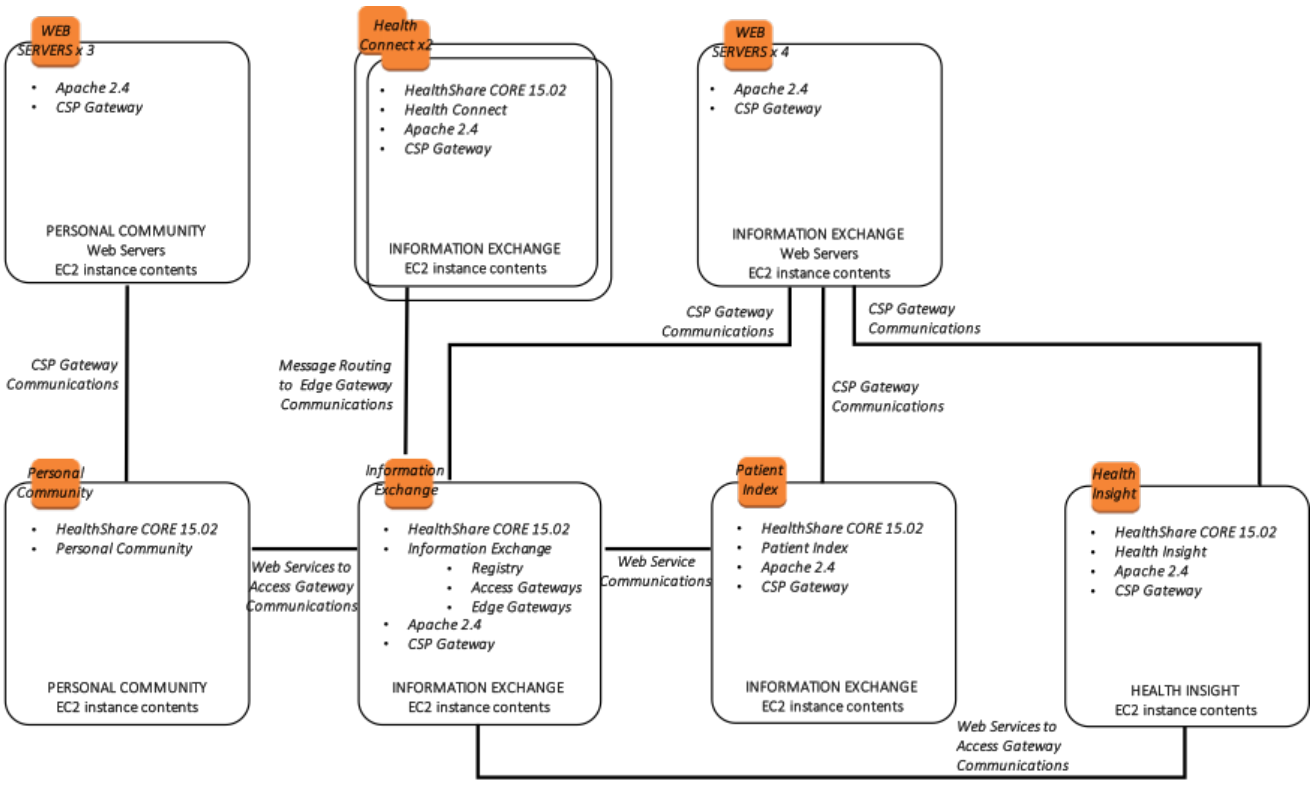
El siguiente diagrama de arquitectura de referencia de muestra incluye alta disponibilidad en la región activa o primaria, y recuperación ante desastres en otra región de AWS si la región principal no está disponible.

Figura-14: HealthShare AWS Diagrama de Arquitectura de Referencia – Arquitectura Física



Además, se proporciona el siguiente diagrama que muestra una vista más lógica de la arquitectura con los productos de software de alto nivel asociados instalados, los requisitos y métodos de conectividad y el propósito funcional respectivo.

Figura-15: HealthShare AWS Diagrama de Arquitectura de Referencia – Arquitectura Lógica



#Administración del sistema #AWS #iFind #Nube #Caché

URL de fuente:https://es.community.intersystems.com/post/la-tecnolog%C3%ADa-de-intersystems-en-amazon-ec2-arquitectura-de-referencia