Pilar de eficiencia de rendimiento

Marco de buena arquitectura de AWS

Julio de 2020

This paper has been archived.

The latest version is now available at:

https://docs.aws.amazon.com/es_es/wellarchitected/latest/performance-efficiency-pillar/welcome.html

Avisos

Los clientes son responsables de hacer su propia evaluación independiente de la información en este documento. Este documento: (a) es solo para fines informativos, (b) representa las ofertas y prácticas actuales de productos de AWS, que están sujetas a cambios sin previo aviso y (c) no crea compromisos o garantías de parte AWS y sus afiliados, proveedores o licenciatarios. Los productos o servicios de AWS se ofrecen “como son”, sin garantías, representaciones o condiciones de ningún tipo, ya sean expresas o implícitas. Las responsabilidades y obligaciones de AWS frente a sus clientes se rigen por los acuerdos celebrados con AWS y este documento no forma parte de ningún acuerdo entre AWS y sus clientes, ni lo modifica.

Contenido

Resumen

Este documento técnico se centra en el pilar de eficiencia de rendimiento de Amazon Web Services (AWS) Marco de buena arquitectura. Proporciona orientación para ayudar a los clientes a aplicar las prácticas recomendadas en el diseño, la entrega y el mantenimiento de entornos de AWS.

El pilar de eficiencia de rendimiento aborda las prácticas recomendadas para administrar entornos de producción. Este documento no cubre el diseño ni la administración de procesos y entornos no productivos, como la integración o entrega continua.

Introducción

El Marco de Buena Arquitectura de AWS lo ayuda a comprender las ventajas y desventajas de las decisiones que toma cuando crea cargas de trabajo en AWS. La utilización del marco lo ayuda a aprender las prácticas recomendadas de arquitectura para diseñar y operar cargas de trabajo confiables, seguras, eficientes y rentables en la nube. El marco ofrece una forma para que pueda medir de manera consistente sus arquitecturas en función de las prácticas recomendadas e identificar las áreas de mejora. Creemos que tener cargas de trabajo de buena arquitectura aumenta considerablemente la probabilidad del éxito empresarial.

El marco se basa en cinco pilares:

- Excelencia operativa
- Seguridad
- Fiabilidad
- Eficiencia de rendimiento
- Optimización de costos

Este documento se enfoca en la aplicación de los principios del pilar de eficiencia de rendimiento a las cargas de trabajo. En entornos locales tradicionales, es un desafío lograr un rendimiento alto y duradero. El uso de los principios en este documento lo ayudará a crear arquitecturas en AWS que otorgan un rendimiento sostenido de manera eficiente a lo largo del tiempo.

Este documento está destinado a aquellos que tienen roles de tecnología, como directores de tecnología (CTO), arquitectos, desarrolladores y miembros de equipos operativos. Después de leer este documento, comprenderá las prácticas recomendadas y estrategias de AWS que se deben utilizar cuando se diseña una arquitectura de nube con rendimiento. Este documento no ofrece detalles de implementación o patrones de arquitectura. Sin embargo, incluye referencias a recursos apropiados.

Eficiencia de rendimiento

El pilar de eficiencia de rendimiento se centra en el uso eficaz de recursos informáticos para cumplir con los requisitos y la forma de mantener la eficiencia a medida que las demandas cambian y las tecnologías evolucionan.

Principios de diseño

Los siguientes principios de diseño pueden ayudarlo a lograr y mantener cargas de trabajo eficientes en la nube.

- **Democratice las tecnologías avanzadas**: facilita la implementación de tecnología avanzada para su equipo mediante la delegación de tareas complejas a su proveedor de nube. En lugar de pedirle a su equipo de TI que aprenda sobre el alojamiento y la ejecución de una nueva tecnología, considere consumir la tecnología como un servicio. Por ejemplo, las bases de datos NoSQL, la transcodificación de medios y el aprendizaje automático son tecnologías que requieren conocimientos especializados. En la nube, estas tecnologías se convierten en servicios que su equipo puede consumir, lo que les permite centrarse en el desarrollo del producto en lugar del aprovisionamiento y administración de recursos.
- **Globalícese en minutos**: la implementación de su carga de trabajo en varias regiones de AWS en todo el mundo le permite proporcionar baja latencia y una mejor experiencia para sus clientes a un costo mínimo.
- **Utilice arquitecturas sin servidor**: las arquitecturas sin servidor eliminan la necesidad de ejecutar y mantener servidores físicos para actividades informáticas tradicionales. Por ejemplo, los servicios de almacenamiento sin servidor pueden actuar como sitios web estáticos (eliminan la necesidad de servidores web) y los servicios para eventos pueden alojar un código. Esto elimina la carga operativa de administrar servidores físicos y puede reducir los costos transaccionales porque los servicios administrados operan a escala de la nube.
- **Experimente con más frecuencia**: con los recursos automatizables y virtuales, puede llevar a cabo con rapidez pruebas comparativas con diferentes tipos de instancias, almacenamiento o configuraciones.
- **Considere la afinidad mecánica**: utilice el enfoque tecnológico que mejor represente sus objetivos. Por ejemplo, tenga en cuenta los patrones de acceso de datos cuando selecciona las bases de datos o los enfoques de almacenamiento.

Definición

Céntrese en las siguientes áreas para lograr eficiencia de rendimiento en la nube:

- Selección
- Revisión
- Monitoreo
- Compensaciones

Adopte un enfoque basado en datos para crear una arquitectura de alto rendimiento. Recopile datos sobre todos los aspectos de la arquitectura, desde el diseño de alto nivel hasta la selección y configuración de tipos de recursos.

La revisión de sus opciones de forma regular, garantiza que se aproveche de la continua evolución de la nube de AWS. El monitoreo garantiza que esté al tanto de cualquier desviación del rendimiento esperado. Realice compensaciones en su arquitectura para mejorar el rendimiento, como el uso de compresión o almacenamiento en caché o la flexibilización de los requisitos de consistencia.

Selección

La solución óptima para una carga de trabajo particular varía y las soluciones suelen combinar múltiples enfoques. Las cargas de trabajo de buena arquitectura utilizan múltiples soluciones y permiten diferentes características para mejorar el rendimiento.

Los recursos de AWS están disponibles en muchos tipos y configuraciones, lo que facilita encontrar un enfoque que se ajuste a sus necesidades. También puede encontrar opciones que no son fáciles de lograr con la infraestructura en las instalaciones. Por ejemplo, un servicio administrado como Amazon DynamoDB ofrece una base de datos NoSQL totalmente administrada con latencia en milisegundos de un solo dígito en cualquier escala.

Selección de arquitectura de rendimiento

A menudo, se requieren múltiples enfoques para obtener un rendimiento óptimo en una carga de trabajo. Los sistemas de buena arquitectura utilizan múltiples soluciones y permiten que diferentes características mejoren el rendimiento.

Utilice un enfoque basado en datos para seleccionar los patrones y la implementación para su arquitectura y logre una solución rentable. Los arquitectos de soluciones de AWS, los socios de las arquitecturas de referencia de AWS y la red de socios de AWS (APN) pueden ayudarlo a seleccionar una arquitectura en función del conocimiento del sector, pero los datos obtenidos a través de pruebas de puntos de referencia o de carga serán necesarios para optimizar su arquitectura.

Es probable que su arquitectura combine varios enfoques de arquitectura diferentes (por ejemplo, impulsados por eventos, ETL o canalización). La implementación de la arquitectura utilizará los servicios de AWS que son específicos para la optimización del rendimiento de la arquitectura. En las siguientes secciones discutiremos los cuatro tipos de recurso principales a considerar (informática, almacenamiento, base de datos y red).

Comprenda los recursos y servicios disponibles: conozca y comprenda la amplia gama de servicios y recursos disponibles en la nube. Identifique los servicios relevantes y opciones de

configuración para la carga de trabajo y comprenda de qué manera puede lograr un rendimiento óptimo.

Si evalúa una carga de trabajo existente, debe generar un inventario de los diversos recursos de servicios que utiliza. El inventario lo ayuda a evaluar qué componentes se pueden reemplazar con los servicios administrados y las tecnologías más nuevas.

Defina un proceso para opciones de arquitectura: utilice la experiencia y el conocimiento interno de la nube o los recursos externos, como los casos de uso publicados, la documentación relevante o los documentos técnicos para definir un proceso a fin de elegir recursos y servicios. Debe definir un proceso que promueva la experimentación y los puntos de referencia con los servicios que se pueden utilizar en la carga de trabajo.

Cuando escribe historias de usuarios críticas para su arquitectura, debe incluir requisitos de rendimiento, como especificar con qué rapidez se debe ejecutar cada historia crítica. Para estas historias críticas, debe implementar experiencias de usuario con secuencias de comandos adicionales para asegurarse de tener visibilidad sobre cómo estas historias funcionan según sus requisitos.

Gestione los requisitos de costos en las decisiones: las cargas de trabajo suelen tener requisitos de costos para su funcionamiento. Utilice los controles de costos internos para seleccionar los tipos y tamaños de recursos según la necesidad de recursos prevista.

Determine qué componentes de la carga de trabajo pueden reemplazarse con servicios totalmente administrados, como bases de datos administradas, caché en memoria y otros servicios. La reducción de la carga de trabajo operativa le permite concentrar los recursos en los resultados comerciales.

Para las prácticas recomendadas de solicitud de costos, consulte la sección de *Recursos rentables* del documento técnico del Pilar de eficiencia de rendimiento.

Utilice políticas o arquitecturas de referencia: maximice el rendimiento y la eficiencia mediante la evaluación de políticas internas y arquitecturas de referencia existentes y utilice su análisis para seleccionar los servicios y configuraciones para la carga de trabajo.

Utilice la guía del proveedor de la nube o un socio adecuado: utilice los recursos de la empresa de la nube, como arquitectos de soluciones, servicios profesionales o un socio adecuado para guiar sus decisiones. Estos recursos pueden ayudar a revisar y mejorar su arquitectura para un rendimiento óptimo.

Póngase en contacto con AWS para obtener ayuda cuando necesite orientación adicional o información sobre el producto. Los arquitectos de soluciones de AWS y los servicios profesionales de AWS ofrecen orientación para la implementación de soluciones. Los socios de APN ofrecen experiencia en AWS para ayudarlo a desbloquear agilidad e innovación para su empresa.

Compare las cargas de trabajo existentes: compare el rendimiento de una carga de trabajo existente para comprender de qué manera rinde en la nube. Utilice los datos recopilados de los puntos de referencia para impulsar decisiones sobre arquitectura.

Utilice los puntos de referencia con pruebas sintéticas para generar datos sobre la forma en que rinden los componentes de la carga de trabajo. Por lo general, los puntos de referencia son más rápidos de configurar que la prueba de carga y se utilizan para evaluar la tecnología de un componente particular. Los puntos de referencia se suelen utilizar al principio de un nuevo proyecto, cuando falta una solución completa para cargar la prueba.

Puede crear sus propias pruebas de referencia personalizadas o utilizar una prueba estándar del sector, como TPC-DS para comparar las cargas de trabajo de almacenamiento de datos. Los puntos de referencia del sector son útiles cuando se comparan los entornos. Los puntos de referencia personalizados son útiles para apuntar a tipos específicos de operaciones que espera realizar en su arquitectura.

Cuando se compara, es importante precalentar el entorno de prueba para asegurar resultados válidos. Ejecute los mismos puntos de referencia varias veces para garantizar que capturó cualquier variación con el paso del tiempo.

Porque generalmente los puntos de referencia son más rápidos de ejecutar que las pruebas de carga, pueden utilizarse antes en la canalización de implementación y ofrecer retroalimentación más rápido en las variaciones de rendimiento. Cuando se evalúa un cambio significativo en un componente o servicio, una referencia puede ser una forma rápida de ver si se justifica el esfuerzo de hacer el cambio. La utilización de los puntos de referencia junto con las pruebas de carga es importante porque las pruebas de carga informan sobre cómo sus cargas de trabajo rendirán en producción.

Realice pruebas de carga a su carga de trabajo: implemente su última arquitectura de carga de trabajo en la nube con diferentes tipos y tamaños de recursos. Monitoree la implementación para capturar las métricas de rendimiento que identifican los cuellos de botella o los excesos de capacidad. Utilice esta información de rendimiento para diseñar o mejorar su selección de recursos y arquitectura.

La prueba de carga utiliza la carga de trabajo *real*, de esta manera puede ver de qué manera rinde su solución en un entorno de producción. Las pruebas de carga se deben ejecutar con versiones sintéticas o depuradas de los datos de producción (elimina la información confidencial o de identificación). Utilice los traspasos de usuario reproducidos o preprogramados a través de la carga de trabajo a escala que ejercita su arquitectura completa. De manera automática, llevan a cabo pruebas de carga de la canalización de entrega y comparan los resultados con los KPI predefinidos y los umbrales. Esto garantiza que siga obteniendo el rendimiento requerido.

Amazon CloudWatch puede recopilar métricas mediante los recursos en su arquitectura. También puede recopilar y publicar métricas personalizadas para los negocios de superficie

o métricas derivadas. Utilice CloudWatch para establecer alarmas que indiquen cuándo se alcanzan los límites y señalen que una prueba se encuentra fuera del rendimiento esperado.

Con los servicios de AWS, puede ejecutar entornos a escala de producción para probar su arquitectura de forma agresiva. Ya que solo paga por el entorno de prueba cuando lo necesita, puede llevar a cabo pruebas a escala completa a solo una parte del costo de usar un entorno en las instalaciones. Aproveche la nube de AWS para probar su carga de trabajo y ver dónde falla en el escalado o escala de una manera que no es lineal. Puede utilizar las instancias de spot de Amazon EC2 para generar cargas a bajo costo y descubrir cuellos de botellas antes de que se experimenten en la producción.

Cuando las pruebas de carga toman un tiempo considerable en ejecutarse, paralelícelas con copias múltiples de su entorno de prueba. Los costos serán similares, pero el tiempo de prueba se reducirá. (Cuesta lo mismo ejecutar una instancia EC2 por 100 horas que 100 instancias por una hora). También puede reducir los costos de prueba de carga con las instancias de spot y seleccionar regiones que tienen costos más bajos que las regiones que utiliza para la producción.

La ubicación de los clientes de pruebas de carga refleja la distribución geográfica de los usuarios finales.

Recursos

Consulte los siguientes recursos para obtener más información sobre las prácticas recomendadas de AWS para las pruebas de carga.

Videos

- Introducing The Amazon Builders' Library (DOP328)

Documentación

- Centro de arquitectura de AWS
- Optimización de rendimiento de Amazon S3
- Rendimiento del volumen de Amazon EBS
- AWS CodeDeploy
- AWS CloudFormation
- Prueba de carga de CloudFront
- Paneles de AWS CloudWatch

Selección de arquitectura informática

La opción de informática óptima para una carga de trabajo particular puede variar en función del diseño de la aplicación, los patrones de uso y los ajustes de configuración. Las arquitecturas pueden utilizar diferentes opciones de informática para varios componentes y permiten que distintas características mejoren el rendimiento. Si se elige la opción informática incorrecta para una arquitectura, esto puede dar lugar a una reducción de la eficiencia de rendimiento.

Evalúe las opciones informáticas disponibles: comprenda las características de rendimiento de las opciones relacionadas con la informática disponibles para usted. Conozca cómo funcionan las instancias, los contenedores y las funciones y cuáles son sus ventajas o desventajas para la carga de trabajo.

En AWS, la informática está disponible de tres formas: instancias, contenedores y funciones:

Instancias

Las instancias son servidores virtualizados, que le permiten cambiar sus capacidades con un botón o una llamada a la API. Como las decisiones de recursos en la nube no son fijas, puede experimentar con diferentes tipos de servidores. En AWS, estas instancias de servidores virtuales vienen en diferentes familias y tamaños y ofrecen una amplia variedad de capacidades, incluidas unidades de estado sólido (SSD) y unidades de procesamiento de gráficos (GPU).

Las instancias de servidores virtuales de Amazon Elastic Compute Cloud (Amazon EC2) vienen en diferentes familias y tamaños. Ofrecen una amplia variedad de capacidades, incluidas unidades de estado sólido (SSD) y unidades de procesamiento de gráficos (GPU). Cuando lanza una instancia EC2, el tipo de instancia que especifica determina el hardware de la computadora anfitriona utilizada para la instancia. Cada tipo de instancia ofrece diferentes capacidades de informática, memoria y almacenamiento. Los tipos de instancias se agrupan en familias de instancias en función de estas capacidades.

Utilice los datos para seleccionar el tipo de instancia EC2 óptimo para su carga de trabajo, asegúrese de tener las opciones de almacenamiento y red correctas y tenga en cuenta las configuraciones del sistema operativo que pueden mejorar el rendimiento de la carga de trabajo.

Contenedores

Los contenedores son un método de virtualización del sistema operativo que le permite ejecutar una aplicación y sus dependencias en procesos de recursos aislados.

Cuando ejecuta contenedores en AWS, tiene que tomar dos decisiones. Primero, elija si quiere administrar los servidores o no. AWS Fargate es informática sin servidor para los contenedores o se puede utilizar Amazon EC2 si necesita controlar la instalación, la configuración y la administración de su entorno informático. Segundo, elija qué coordinador de contenedor utilizar: Amazon Elastic Container Service (ECS) o Amazon Elastic Kubernetes Service (EKS)

Amazon Elastic Container Service (Amazon ECS) es un servicio coordinador de contenedor totalmente administrado que le permite ejecutar y administrar contenedores de manera automática en un clúster de instancias EC2 o instancias sin servidor con AWS Fargate. Puede integrar Amazon ECS de manera nativa con otros servicios, como Amazon Route 53, Secrets Manager, AWS Identity and Access Management (IAM) y Amazon CloudWatch.

Amazon Elastic Kubernetes Service (Amazon EKS) es un servicio de Kubernetes totalmente administrado. Puede elegir ejecutar sus clústeres de EKS con AWS Fargate, lo que elimina la necesidad de aprovisionar y administrar los servidores. EKS está profundamente integrado con servicios como Amazon CloudWatch, Auto Scaling Groups, AWS Identity and Access Management (IAM) y Amazon Virtual Private Cloud (VPC).

Cuando utiliza contenedores, debe usar datos a fin de seleccionar el tipo óptimo para su carga de trabajo, tal como usa los datos para seleccionar sus tipos de instancia EC2 o AWS Fargate. Tenga en cuenta las opciones de configuración de contenedores, como la configuración de la memoria, CPU y tenencia. Para habilitar el acceso a la red entre servicios de contenedores, considere utilizar una malla de servicios, como AWS App Mesh, que estandariza el modo en que se comunican los servicios. La malla de servicios otorga visibilidad total y garantiza alta disponibilidad para las aplicaciones.

Funciones

Las funciones abstraen el entorno de ejecución del código que desea ejecutar. Por ejemplo, AWS Lambda le permite ejecutar el código sin operar una instancia.

Puede utilizar AWS Lambda para ejecutar el código de cualquier tipo de aplicación o servicio de backend sin administración. Simplemente, cargue su código y AWS Lambda administrará todo lo que sea necesario para ejecutar y escalar ese código. Puede configurar su código para que se active automáticamente desde otros servicios de AWS, llamarlo directamente o usarlo con Amazon API Gateway.

Amazon API Gateway es un servicio totalmente administrado que facilita a los desarrolladores crear, publicar, mantener, monitorear y asegurar las API en cualquier escala. Puede crear una API que actúe como “puerta principal” para la función de Lambda. API Gateway gestiona todas las tareas involucradas en la aceptación y procesamiento de hasta cientos de miles de llamadas a la API simultáneas, incluido la administración de tráfico, control de autorización y acceso, monitoreo y administración de la versión de API.

Para entregar un rendimiento óptimo con AWS Lambda, seleccione la cantidad de memoria que desea para su función. Se le asigna potencia de CPU proporcional y otros recursos. Por ejemplo, elegir 256 MB de memoria asigna aproximadamente el doble de potencia de CPU para la función de Lambda cuando solicita 128 MB de memoria. Puede controlar la cantidad de tiempo que se le permite ejecutar cada función (hasta un máximo de 300 segundos).

Comprenda las opciones de configuración informática disponibles: comprenda de qué manera distintas opciones complementan su carga de trabajo y qué opciones de configuración son mejores para su sistema. Los ejemplos de estas opciones incluyen familia de instancias, tamaños, características (GPU, E/S), tamaños de funciones, instancias de contenedor y tenencia única contra múltiple.

Cuando selecciona familias y tipos de instancias, también debe tener en cuenta las opciones de configuración disponibles que cumplan con las necesidades de la carga de trabajo:

- **Unidades de procesamiento de gráficos (GPU)**: con la informática de uso general en las GPU (GPGPU), puede crear aplicaciones que se beneficien del alto grado de paralelismo que ofrecen las GPU mediante el aprovechamiento de las plataformas (como CUDA) en el proceso de desarrollo. Si su carga de trabajo requiere renderizado 3D o compresión de video, las GPU permiten el cálculo y la codificación aceleradas por hardware, lo que hace que su carga de trabajo sea más eficiente.
- **Matrices de puertas programables en campo (FPGA)**: con las FPGA, puede optimizar sus cargas de trabajo mediante la ejecución acelerada por hardware personalizada para las cargas de trabajo más exigentes. Puede definir sus algoritmos mediante el aprovechamiento de los lenguajes de programación generales compatibles, como C o Go, o los lenguajes orientados por hardware, como Verilog o VHDL.
- **AWS Inferentia (Inf1)**: las instancias Inf1 se crean para dar soporte a las aplicaciones de inferencia de aprendizaje automático. Con las instancias Inf1, los clientes pueden ejecutar aplicaciones de inferencia de aprendizaje automático a gran escala, como el reconocimiento de imagen, el reconocimiento de voz, el procesamiento de lenguaje natural, la personalización y la detección de fraudes. Puede crear un modelo en uno de los marcos de aprendizaje automático popular, como TensorFlow, PyTorch o MXNet y utilizar instancias de GPU, como P3 o P3dn para entrenar el modelo. Después de entrenar los modelos de aprendizaje automático para cumplir con los requerimientos, puede implementar su modelo en instancias Inf1 con AWS Neuron, un kit de desarrollo de software (SDK) especializado que consta de un compilador, tiempo de ejecución y herramientas de perfilado que optimizan el rendimiento de inferencia de aprendizaje automático de chips de Inferentia.
- **Familias de instancias ampliables**: las instancias ampliables están diseñadas para ofrecer rendimiento de referencia moderado y la capacidad de ampliar a un rendimiento significativamente mayor cuando lo requiera su carga de trabajo. Estas instancias están dirigidas a cargas de trabajo que no utilizan CPU completa o de manera consistente, pero en ocasiones necesitan ampliarse. Son adecuadas para cargas de trabajo de uso general, como servidores web, entornos de desarrolladores y bases de datos pequeñas. Estas instancias ofrecen créditos de CPU que se pueden consumir cuando la instancia debe ofrecer rendimiento. Los créditos se acumulan cuando la instancia no los necesita.

- **Características informáticas avanzadas:** Amazon EC2 brinda acceso a las características informáticas avanzadas, como la administración de registros de estado C y estado P y el control del impulso turbo de los procesadores. El acceso a coprocesadores permite la descarga de operaciones de criptografía a través de AES-NI o la informática avanzada mediante extensiones AVX.

El Sistema nitro de AWS es una combinación de hardware específico y de hipervisor liviano que permite una innovación más rápida y seguridad mejorada. Utilice el Sistema nitro de AWS cuando esté disponible para permitir el consumo completo de los recursos informáticos y de memoria del hardware del host. Además, las tarjetas de Nitro dedicadas permiten redes de alta velocidad, EBS de alta velocidad y aceleración de E/S.

Recopile métricas relacionadas con la informática: una de las mejores formas de comprender cómo rinden los sistemas informáticos es registrar y realizar un seguimiento del verdadero uso de diversos recursos. Estos datos se pueden utilizar para realizar determinaciones más precisas sobre las solicitudes de recursos.

Las cargas de trabajo (como las que se ejecutan en arquitecturas de microservicios) pueden generar grandes volúmenes de datos en forma de métricas, registros y eventos. Determine si el monitoreo existente y el servicio de observabilidad pueden administrar los datos generados. Amazon CloudWatch puede utilizarse para recopilar, acceder y correlacionar estos datos en una única plataforma de todos sus recursos, aplicaciones y servicios de AWS que se ejecutan en AWS y servidores locales, para que pueda obtener visibilidad en todo el sistema fácilmente y resolver problemas rápidamente.

Determine la configuración solicitada mediante el redimensionamiento: analice las diversas características de rendimiento de su carga de trabajo y de qué manera se relacionan con el uso de la memoria, red y CPU. Utilice estos datos para elegir los recursos que mejor se adapten al perfil de su carga de trabajo. Por ejemplo, una carga de trabajo de memoria intensiva, como una base de datos, puede ser el mejor modo de alcanzar la familia de instancias r. Sin embargo, una carga de trabajo ampliada puede obtener mayores beneficios de un sistema de contenedor elástico.

Utilice la elasticidad de los recursos disponibles: la nube ofrece la flexibilidad de expandir o reducir sus recursos de manera dinámica mediante una variedad de mecanismos para cumplir con los cambios en la demanda. En combinación con métricas relacionadas con la informática, una carga de trabajo puede responder a cambios de manera automática y utilizar el conjunto de recursos óptimos para lograr este objetivo.

La adaptación óptima de la oferta a la demanda ofrece los costos más bajos para una carga de trabajo, pero también debe planificar una provisión suficiente que permita el tiempo de aprovisionamiento y fallas de recursos individuales. La demanda puede ser fija o variable, lo que requiere métricas y automatización a fin de asegurar que la administración no se convierta en un costo excesivo y desproporcionadamente grande.

Con AWS, puede utilizar una cantidad de distintos enfoques para que coincida la oferta con la demanda. El documento técnico del Pilar de eficiencia de rendimiento describe cómo utilizar los siguientes enfoques de costo:

- Enfoque en función de la demanda
- Enfoque en función del búfer
- Enfoque en función del tiempo

Debe garantizar que la implementación de esa carga de trabajo pueda gestionar eventos de escala ascendente y descendente. Cree escenarios de prueba para eventos de escala descendente a fin de asegurar que la carga de trabajo tenga el comportamiento esperado.

Reevalúe las necesidades informáticas en función de las métricas: utilice métricas a nivel del sistema para identificar las conductas y solicitudes de su carga de trabajo a lo largo del tiempo. Evalúe las necesidades de su carga de trabajo mediante la comparación de los recursos disponibles con estas solicitudes y realice cambios en su entorno informático para que coincidan mejor con el perfil de su carga de trabajo. Por ejemplo, con el tiempo se podría observar que un sistema consume más memoria de lo que se pensaba inicialmente, por lo que pasar a una familia o tamaño de instancias diferente podría mejorar tanto el rendimiento como la eficiencia.

Recursos

Consulte los siguientes recursos para obtener más información sobre las prácticas recomendadas de AWS para la informática.

Videos

- Amazon EC2 foundations (CMP211-R2)
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Deliver high performance ML inference with AWS Inferentia (CMP324-R1)
- Optimize performance and cost for your AWS compute (CMP323-R1)
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (CMP202-R1)

Documentación

- Instancias:
  - Tipos de instancias
  - Control de los estados del procesador de la instancia EC2
- Contenedores de EKS: nodos de trabajo de EKS
- Contenedores de ECS: instancias de contenedores de Amazon ECS
- Funciones: configuración de la función de Lambda

Selección de arquitectura de almacenamiento

La solución de almacenamiento óptimo para un sistema particular varía según el tipo de método de acceso (bloque, archivo u objeto), patrones de acceso (aleatorio o secuencial), rendimiento requerido, frecuencia de acceso (en línea, sin conexión, de archivo) frecuencia de actualización (WORM, dinámico) y restricciones de durabilidad y disponibilidad. Los sistemas de buena arquitectura utilizan múltiples soluciones de almacenamiento y permiten que diferentes características mejoren el rendimiento.

En AWS, el almacenamiento es virtualizado y está disponible en varios tipos diferentes. Esto facilita la coincidencia de los métodos de almacenamiento con las necesidades y ofrece opciones de almacenamiento que no son fáciles de lograr con infraestructura en las instalaciones. Por ejemplo, Amazon S3 está diseñada para 11 nueves de durabilidad. También puede cambiar el uso de unidades de disco duro (HDD) magnético a SSD y mover fácilmente las unidades virtuales de una instancia a otra en segundos.

El rendimiento se puede medir mediante la observación de la utilidad, las operaciones de entrada y de salida por segundo (IOPS) y latencia. La comprensión de la relación entre esas medidas lo ayudarán a seleccionar la solución de almacenamiento más adecuada.

Almacenamiento	Servicios	Latencia	Rendimiento	Compartible
Bloque	Amazon EBS, Almacén de instancias EC2	Más bajo, constante	Único	Montado en la instancia EC2, copias mediante instantáneas
Sistema de archivo	Amazon EFS, Amazon FSx	Bajo, constante	Múltiple	Muchos clientes
Objeto	Amazon S3	Latencia baja	Escala web	Muchos clientes
Archivo	Amazon S3 Glacier	Minutos a horas	Alto	No

Desde una perspectiva de latencia, si accede a sus datos solo por una instancia, entonces debe utilizar el almacenamiento en bloque, como Amazon EBS. Los sistemas de archivos distribuidos, como Amazon EFS suelen tener una pequeña latencia general para cada operación de archivo, por lo tanto deben utilizarse donde múltiples instancias necesitan acceso.

Amazon S3 posee características que pueden reducir la latencia y aumentar el rendimiento. Puede utilizar la replicación entre regiones (CRR) para ofrecer acceso a datos de baja latencia para diferentes regiones geográficas.

Desde una perspectiva de rendimiento, Amazon EFS admite cargas de trabajo altamente paralelas (por ejemplo, con operaciones simultáneas de múltiples hilos y múltiples instancias EC2), lo que permite altos niveles de rendimiento y operaciones agregadas por segundo. Para Amazon EFS, utilice un punto de referencia o prueba de carga para seleccionar el modo de rendimiento adecuado.

Comprenda las características y requisitos de almacenamiento: comprenda las diferentes características (por ejemplo, compartible, volumen de archivo, volumen de caché, patrones de acceso, latencia, rendimiento y persistencia de los datos) que se requieren para seleccionar los servicios que mejor se adapten a su carga de trabajo, como almacenamiento de objetos, de bloques, de archivos o de almacenamiento de la instancia.

Determine la tasa de crecimiento esperada para su carga de trabajo y elija una solución de almacenamiento que cumpla con esas tasas. Las soluciones de almacenamiento de objetos y archivos, como Amazon S3 y Amazon Elastic File System, permiten el almacenamiento ilimitado. Amazon EBS posee tamaños de almacenamiento predeterminados. Los volúmenes elásticos le permiten incrementar dinámicamente la capacidad, ajustar el rendimiento y cambiar el tipo de cualquier volumen de generación actual nuevo o existente sin horas de inactividad o impacto de rendimiento, pero requiere cambios del sistema de archivos del sistema operativo.

Evalúe las opciones de configuración disponibles: evalúe las diferentes características y opciones de configuración y de qué manera se relacionan con el almacenamiento. Comprenda dónde y cómo usar IOPS provisionadas, SSD, almacenamiento magnético, almacenamiento de objetos, almacenamiento de archivos o almacenamiento efímero para optimizar el espacio de almacenamiento y el rendimiento para su carga de trabajo.

Amazon EBS ofrece una gama de opciones que le permiten optimizar el rendimiento y costo de almacenamiento para la carga de trabajo. Estas opciones se dividen en dos categorías principales: almacenamiento con respaldo SSD para cargas de trabajo transaccional, como bases de datos y volúmenes de arranque (el rendimiento depende principalmente de IOPS) y almacenamiento con respaldo HDD para cargas de trabajo de rendimiento intensivo, como MapReduce y procesamiento de registros (el rendimiento depende principalmente de MB/s).

Los volúmenes con respaldo SSD incluyen las IOPS SSD provisionadas de mayor rendimiento para cargas de trabajo transaccionales sensibles a la latencia y SSD de propósito general que equilibran el precio y el rendimiento para una amplia variedad de datos transaccionales.

Amazon S3 transfer acceleration permite la transferencia rápida de archivos a través de grandes distancias entre el cliente y el bucket de S3. La aceleración de transferencia aprovecha las ubicaciones de borde distribuidas globalmente de Amazon CloudFront para enrutar datos a través de una ruta de red optimizada. Para una carga de trabajo en un bucket de S3 que tiene solicitudes GET intensivas, utilice Amazon S3 con CloudFront. Cuando cargue archivos grandes, utilice cargas de múltiples partes con cargas de múltiples partes al mismo tiempo para ayudar a maximizar el rendimiento de la red.

Amazon Elastic File System (Amazon EFS) ofrece un sistema de archivo NFS elástico totalmente administrado, escalable y simple para utilizar con los servicios en la nube de AWS y recursos en las instalaciones. Para admitir una amplia variedad de cargas de trabajo de almacenamiento en la nube, Amazon EFS ofrece dos modos de rendimiento: el modo de rendimiento de propósito general y el de rendimiento de E/S máxima. También existen dos modos de rendimiento para elegir para su sistema de archivos, el rendimiento ampliado y el rendimiento aprovisionado. Para determinar que ajustes utilizar para su carga de trabajo, consulte la guía del usuario de Amazon EFS.

Amazon FSx ofrece dos sistemas de archivo: Amazon FSx for Windows File Server para cargas de trabajo empresariales y Amazon FSx for Lustre para cargas de trabajo de alto rendimiento. FSx está respaldado por SSD y diseñado para entregar rendimiento rápido, predecible, escalable y constante. Los sistemas de archivos de Amazon FSx entregan altas velocidades de lectura y escritura y acceso a datos de baja latencia constantes. Puede elegir el nivel de rendimiento que necesita para satisfacer las necesidades de su carga de trabajo.

Tome decisiones en función de métricas y patrones de acceso: elija sistemas de almacenamiento en función de los patrones de acceso de su carga de trabajo y configúrelos para establecer de qué manera accede a los datos la carga de trabajo. Elija el almacenamiento de objetos en lugar del almacenamiento en bloque para aumentar la eficiencia del almacenamiento. Configure las opciones de almacenamiento que elija para que coincidan con sus patrones de acceso a datos.

La forma en que accede a los datos impacta en el rendimiento de la solución de almacenamiento. Seleccione la solución de almacenamiento que mejor se alinee con sus patrones de acceso o considere la posibilidad de cambiar sus patrones de acceso para alinearlos con la solución de almacenamiento y así maximizar el rendimiento.

La creación de una matriz RAID 0 (cero) permite alcanzar un nivel de rendimiento más alto para un sistema de archivos que lo que puede aprovisionar en un volumen único. Considere utilizar RAID 0 cuando el rendimiento de E/S es más importante que la tolerancia a fallas. Por ejemplo, puede utilizarlo con una base de datos muy usada en la que la replicación de datos ya se configuró de manera separada.

Seleccione las métricas de almacenamiento apropiadas para su carga de trabajo en todas las opciones de almacenamiento consumidas por la carga de trabajo. Cuando utiliza sistemas de archivos que usan créditos ampliados, cree alarmas que le avisen cuando se acerca a los límites de los créditos. Debe crear paneles de almacenamiento para mostrar el estado general del almacenamiento de la carga de trabajo.

Para los sistemas de almacenamiento que son de un tamaño fijo, como Amazon EBS o Amazon FSx, asegúrese de que monitorea la cantidad de almacenamiento que se utiliza contra el volumen de almacenamiento general y, si es posible, cree la automatización a fin de aumentar el volumen de almacenamiento cuando alcance el límite

Recursos

Consulte los siguientes recursos para obtener más información sobre las prácticas recomendadas de AWS para el almacenamiento.

Videos

- Deep dive on Amazon EBS (STG303-R1)
- Optimize your storage performance with Amazon S3 (STG343)

Documentación

- Amazon EBS:
  - Almacenamiento de Amazon EC2
  - Tipos de volumen de Amazon EBS
  - Características de E/S
- Amazon S3: solicite consideraciones de rendimiento y tasa
- Amazon Glacier: documentación de Amazon Glacier
- Amazon EFS: rendimiento de Amazon EFS
- Amazon FSx:
  - Amazon FSx for Lustre Performance
  - Amazon FSx for Windows File Server Performance

Selección de arquitectura de base de datos

La solución de base de datos óptima para un sistema varía según los requerimientos de disponibilidad, consistencia, tolerancia en las particiones, latencia, durabilidad, escalabilidad y capacidad de consulta. Muchos sistemas utilizan soluciones de bases de datos diferentes para varios subsistemas y permiten que distintas características mejoren el rendimiento. La selección de las características y soluciones de base de datos incorrectas puede dar como resultado una menor eficiencia de rendimiento.

Comprenda las características de los datos: comprenda las diferentes características de los datos en su carga de trabajo. Determine si la carga de trabajo necesita transacciones, cómo interactúa con los datos y cuáles son las demandas de rendimiento. Utilice estos datos para seleccionar el enfoque de base de datos de mejor rendimiento para su carga de trabajo (por ejemplo, bases de datos relacionales, valor clave de NoSQL, documento, columna ancha, gráfico, serie temporal o almacenamiento en la memoria).

Puede elegir entre muchos motores de bases de datos creadas específicamente, incluidas las bases de datos relacionales, valor clave, documento, en la memoria, gráfico, serie temporal y libro mayor. Cuando elige la mejor base de datos para resolver un problema específico (o un grupo de problemas), puede separarse de las bases de datos monolíticas, universales y restrictivas y centrarse en la creación de aplicaciones que satisfagan las necesidades de los clientes.

Las bases de datos relacionales almacenan datos con esquemas predefinidos y las relaciones entre ellos. Estas bases de datos están diseñadas para dar soporte a las transacciones ACID (atomicidad, consistencia, aislamiento, durabilidad) y mantener la integridad referencial y una fuerte consistencia de datos. Muchas aplicaciones tradicionales, planificación de recursos empresariales (ERP), administración de relaciones con clientes (CRM) y comercio electrónico utilizan bases de datos relacionales para almacenar sus datos. Puede ejecutar muchos de estos motores de bases de datos en Amazon EC2 o elegir uno de los servicios de bases de datos administrados de AWS: Amazon Aurora, Amazon RDS y Amazon Redshift.

Las bases de datos de valor clave se optimizan para los patrones de acceso comunes, usualmente para almacenar y recuperar grandes volúmenes de datos. Estas bases de datos entregan tiempos de respuesta rápida, incluso en volúmenes extremos de solicitudes simultáneas.

Las aplicaciones web de alto tránsito, los sistemas de comercio electrónico y las aplicaciones de videojuegos son casos de uso típicos para bases de datos de valor clave. En AWS, puede utilizar Amazon DynamoDB, una base de datos duradera, totalmente administrada, de múltiples regiones, de múltiple maestros con seguridad integrada, copia de seguridad y restauración y almacenamiento de caché en memoria para aplicaciones a escala de Internet.

Las bases de datos en memoria se utilizan para aplicaciones que requieren acceso a datos en tiempo real. Con el almacenamiento de los datos directamente en la memoria, estas bases de datos entregan latencia de microsegundos a aplicaciones para quienes la latencia de

milisegundos no es suficiente. Puede utilizar bases de datos en memoria para almacenamiento en caché de aplicaciones, administración de sesiones, tablas de clasificaciones para videojuegos y aplicaciones geoespaciales. Amazon ElastiCache es un almacén de datos en memoria totalmente administrado, compatible con Redis o Memcached.

Para almacenar datos semiestructurados, como documentos similares a JSON, se diseña una base de datos de documentos. Estas bases de datos ayudan a los desarrolladores a crear y actualizar rápidamente aplicaciones, como la administración de contenidos, catálogos y perfiles de usuarios. Amazon DocumentDB es un servicio de base de datos de documentos totalmente administrado, con alta disponibilidad, escalable y rápido que admite cargas de trabajo de MongoDB.

Un almacén de columna ancha es un tipo de base de datos NoSQL. Utiliza tablas, filas y columnas, pero a diferencia de una base de datos relacional, los nombres y formatos de las columnas pueden variar de fila a fila en la misma tabla. Generalmente, puede ver un almacén de columna ancha en aplicaciones industriales de gran escala para equipos de mantenimiento, administración de flotas y optimización de rutas. Amazon Managed Apache Cassandra Service es un servicio de base de datos de columna ancha compatible con Apache Cassandra, de alta disponibilidad y escalable.

Las bases de datos de gráficos son para aplicaciones que tienen que navegar y consultar millones de relaciones entre conjuntos de datos de gráfico altamente conectados con milisegundos de latencia a gran escala. Muchas empresas utilizan bases de datos de gráficos para la detección de fraudes, redes sociales y motores de recomendación. Amazon Neptune es un servicio de base de datos de gráficos totalmente administrado, fiable y rápido que facilita crear y ejecutar aplicaciones que funcionan con bases de datos altamente conectadas.

Las bases de datos de series temporales recopilan, sintetizan y derivan de manera eficiente información de los datos que cambia con el paso del tiempo. Las aplicaciones IoT, DevOps y de telemetría industrial pueden utilizar bases de datos de series temporales. Amazon Timestream es un servicio de base de datos de series temporales totalmente administrado, escalable y rápido para IoT y aplicaciones operativas que facilitan almacenar y analizar trillones de eventos por día.

Las bases de datos de libro mayor ofrecen una autoridad centralizada y confiable a fin de mantener un registro de transacciones escalable, inmutable y verificable criptográficamente para cada aplicación. Vemos bases de datos de libro mayor que se utilizan para sistemas de registro, cadena de suministro, inscripciones e incluso, transacciones bancarias. Amazon Quantum Ledger Database (QLDB) es una base de datos de libro mayor totalmente administrada que ofrece un registro de transacción transparente, inmutable y verificable criptográficamente a cargo de una autoridad central confiable. Amazon QLDB realiza un seguimiento de todos los cambios de los datos de aplicación y mantiene un historial de cambios completo y verificable a lo largo del tiempo.

Evalúe las opciones disponibles: evalúe los servicios y las opciones de almacenamiento que están disponibles como parte del proceso de selección para los mecanismos de almacenamiento de su carga de trabajo. Comprenda de qué manera y cuándo utilizar un servicio o sistema de almacenamiento de datos determinados. Aprenda sobre las opciones de configuración disponibles que pueden optimizar el rendimiento o la eficiencia de la base de datos, como las IOPS provisionadas, los recursos de memoria e informática y el almacenamiento de caché.

Generalmente, las soluciones de base de datos poseen opciones de configuración que le permiten optimizar el tipo de carga de trabajo. Con los puntos de referencia o la prueba de carga, identifique las métricas de bases de datos que son importantes para su carga de trabajo. Tenga en cuenta las opciones de configuración para el enfoque de su base de datos seleccionada, como optimización de almacenamiento, configuración del nivel de base de datos, memoria y caché.

Evalúe las opciones de almacenamiento de caché de base de datos para su carga de trabajo. Los tres tipos de caché de base de datos más comunes son los siguiente:

- **Cachés integrados de base de datos**: algunas bases de datos (como Amazon Aurora) ofrecen un caché integrado que se administra en el motor de la base de datos y tienen capacidades de escritura directa integradas.
- **Cachés locales**: un caché local frecuentemente almacena los datos utilizados en su aplicación. Esto acelera la recuperación de datos y elimina el tráfico de redes asociadas a esto, de esta manera la recuperación de datos es más rápida que otras arquitecturas de almacenamiento de caché.
- **Cachés remotos**: los cachés remotos se almacenan en servidores dedicados y, por lo general, se basan en almacenes NoSQL de clave/valor como Redis y Memcached. Ofrecen hasta un millón de solicitudes por segundo por nodo de caché.

Para las cargas de trabajo de Amazon DynamoDB, DynamoDB Accelerator (DAX) ofrece caché en memoria totalmente administrado. DAX es un caché en memoria que entrega un rendimiento de lectura rápida para las tablas a escala mediante el uso un caché en memoria totalmente administrado. Con DAX, puede mejorar el rendimiento de lectura de las tablas DynamoDB hasta 10 veces, toma el tiempo necesario para las lecturas de milisegundos a microsegundos, incluso a millones de solicitudes por segundo.

Recopile y registre métricas de rendimiento de la base de datos: utilice herramientas, bibliotecas y sistemas que registren mediciones de rendimiento relacionadas con el rendimiento de la base de datos. Por ejemplo, mida las transacciones por segundo, las consultas lentas o los sistemas de latencia introducidos cuando accede a la base de datos. Utilice estos datos para comprender el rendimiento de los sistemas de su base de datos.

Instrumente tantas métricas de actividades de base de datos como pueda reunir de su carga de trabajo. Estas métricas deben publicarse directamente desde la carga de trabajo o deben reunirse del servicio de administración de rendimiento de la aplicación. Puede utilizar AWS X-Ray a fin de analizar y depurar la producción, las aplicaciones distribuidas, como aquellas creadas con una arquitectura de microservicios. Un rastro de X-Ray puede incluir segmentos que encapsulen todos los puntos de datos para un componente único. Por ejemplo, cuando la aplicación realiza una llamada a una base de datos en respuesta a una solicitud, crea un segmento para esa solicitud con un subsegmento que representa la llamada a la base de datos y su resultado. El subsegmento puede contener datos, como la consulta, la tabla utilizada, la marca de tiempo y el estado de error. Una vez instrumentado, debe permitir que las alarmas de las métricas de la base de datos que señalen cuando se alcanzan los límites.

Elija el almacenamiento de datos en función de los patrones de acceso: utilice patrones de acceso de la carga de trabajo para decidir qué servicios y tecnologías utilizar. Por ejemplo, use una base de datos relacional para las cargas de trabajo que requieren transacciones o un almacén de valor clave que ofrece un rendimiento mayor, pero que finalmente sea constante donde se aplique.

Optimice el almacenamiento de datos en función de los patrones de acceso y las métricas: utilice características de rendimiento y patrones de acceso que optimicen la forma en que los datos se almacenan o se consultan para lograr el mejor rendimiento posible. Mida de qué manera las optimizaciones, como el indexado, la distribución clave, el diseño de almacén de datos o las estrategias de caché, impactan en el rendimiento del sistema o la eficiencia general.

Recursos

Consulte los siguientes recursos para obtener más información sobre las prácticas recomendadas de AWS para bases de datos.

Videos

- AWS purpose-built databases (DAT209-L)
- Amazon Aurora storage demystified: How it all works (DAT309-R)
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Documentación

- AWS Database Caching
- Bases de datos en la nube con AWS
- Prácticas recomendadas de Amazon Aurora
- Rendimiento de Amazon Redshift

- Amazon Athena top 10 performance tips
- Prácticas recomendadas de Amazon Redshift Spectrum
- Prácticas recomendadas de Amazon DynamoDB
- Amazon DynamoDB Accelerator

Selección de arquitectura de red

La solución de red óptima para una carga de trabajo varía según la latencia, los requisitos de rendimiento, la fluctuación y el ancho de banda. Las restricciones físicas, como el usuario o los recursos en las instalaciones, determinan las opciones de ubicación. Estas restricciones se pueden compensar con ubicaciones de borde o ubicación de recurso.

En AWS, la red es virtualizada y está disponible en varios tipos y configuraciones diferentes. Esto facilita la coincidencia entre los métodos de red con las necesidades. AWS ofrece características de productos (por ejemplo, redes mejoradas, instancias optimizadas de red de Amazon EC2, Amazon S3 transfer acceleration y Amazon CloudFront dinámico) para optimizar el tráfico de red. AWS también ofrece características de red (por ejemplo, direccionamiento de latencia de Amazon Route 53, puntos de enlace de Amazon VPC, AWS Direct Connect y AWS Global Accelerator) para reducir la distancia o fluctuación de la red.

Comprenda de qué manera las redes impactan en el rendimiento: analice y comprenda cómo las características relacionadas con las redes impactan en el rendimiento de la carga de trabajo. Por ejemplo, la latencia de la red a menudo impacta en la experiencia del usuario y por no proporcionar suficiente capacidad de red puede generar un cuello de botella en el rendimiento de la carga de trabajo.

Como la red se encuentra entre todos los componentes de la aplicación, puede tener grandes impactos positivos o negativos en el rendimiento y comportamiento de la aplicación. También existen aplicaciones que dependen fuertemente del rendimiento de la red, como la informática de alto rendimiento (HPC) donde la comprensión profunda de la red es importante para aumentar el rendimiento del clúster. Debe determinar los requisitos de la carga de trabajo para el ancho de banda, la latencia, la fluctuación y el rendimiento.

Evalúe las características de red disponibles: evalúe las características de red en la nube que pueden aumentar el rendimiento. Mida el impacto de estas características mediante pruebas, métricas y análisis. Por ejemplo, aproveche las características de nivel de red que están disponibles para reducir la latencia, la distancia de red o la fluctuación.

Muchos servicios a menudo ofrecen características para optimizar el rendimiento de la red. Tenga en cuenta las características de productos, como la capacidad de red de instancia EC2, tipos de instancias de red mejorada, instancias optimizadas de Amazon EBS, Amazon S3 transfer acceleration y CloudFront dinámico para optimizar el tráfico de red.

AWS Global Accelerator es un servicio que mejora la disponibilidad y el rendimiento de la aplicación con la red global de AWS. Optimiza la ruta de red con el aprovechamiento de la vasta red global de AWS sin congestión. Ofrece direcciones de IP estáticas que facilitan el movimiento de puntos de enlace entre las zonas de disponibilidad o las regiones de AWS sin necesidad de actualizar la configuración DNS o cambiar las aplicaciones orientadas al cliente

La aceleración de contenido de Amazon S3 es una característica que permite a los usuarios externos beneficiarse de las optimizaciones de la red de CloudFront para actualizar los datos en Amazon S3. Esto facilita la transferencia de grandes cantidades de datos desde ubicaciones remotas que no poseen conectividad dedicada en la nube de AWS.

Las instancias EC2 más nuevas pueden aprovechar la red mejorada. Las instancias EC2 de serie N, como M5n y M5dn, aprovechan la cuarta generación de tarjetas Nitro personalizadas y el dispositivo Elastic Network Adapter (ENA) para entregar hasta 100 Gbps de rendimiento de red a una instancia única. Estas instancias ofrecen 4 veces el ancho de banda de la red y el proceso de paquetes en comparación con las instancias base M5 y son ideales para aplicaciones intensivas de red. Los clientes también pueden habilitar Elastic Fabric Adapter (EFA) en ciertos tamaños de instancias de las instancias M5n y M5dn para latencia de red baja y constante.

Los Amazon Elastic Network Adapters (ENA) ofrecen una mayor optimización porque proporcionan 20 Gbps de capacidad de red para sus instancias dentro de un solo grupo de ubicación. Elastic Fabric Adapter (EFA) es una interfaz de red para las instancias EC2 de Amazon que le permiten ejecutar cargas de trabajo que requieren altos niveles de comunicaciones entre nodos en escala de AWS. Con EFA, las aplicaciones de informática de alto rendimiento (HPC) que utilizan Message Passing Interface (MPI) y las aplicaciones de aprendizaje automático (ML) con la biblioteca de comunicaciones colectivas de NVIDIA (NCCL) pueden escalar a miles de CPU o GPU.

Las instancias optimizadas de Amazon EBS utilizan una pila de configuración optimizada y proporcionan capacidad dedicada adicional para E/S de Amazon EBS. Esta optimización proporciona el mejor rendimiento para sus volúmenes de EBS, ya que minimizan la contención entre la E/S de Amazon EBS y otro tráfico de su instancia.

El direccionamiento basado en latencia (LBR) para Amazon Route 53 ayuda a mejorar el rendimiento de las cargas de trabajo para una audiencia global. LBR funciona mediante el direccionamiento de sus clientes al punto de enlace de AWS (para instancias EC2, direcciones de IP elásticas o balanceadores de carga de ELB) que ofrecen la experiencia más rápida en función de mediciones de rendimiento reales de las diferentes regiones de AWS donde se ejecuta la carga de trabajo.

Los puntos de enlace de Amazon VPC ofrecen conectividad fiable a los servicios de AWS (por ejemplo, Amazon S3) sin necesidad de una gateway de Internet o una instancia de traducción de direcciones de red (NAT).

Elija una conectividad dedicada de tamaño apropiado o una VPN para cargas de trabajo híbridas: cuando exista un requisito para la comunicación en las instalaciones, asegúrese de contar con el ancho de banda adecuado para el rendimiento de la carga de trabajo. Según los requisitos de ancho de banda, una sola conexión dedicada o una única VPN puede que no sea suficiente y deba habilitar el equilibrio de carga de tráfico en varias conexiones.

Debe calcular los requisitos de ancho de banda y latencia para su carga de trabajo híbrida. Estos números impulsarán los requisitos de tamaño para AWS Direct Connect o los puntos de enlace de VPN.

AWS Direct Connect ofrece conectividad dedicada para los entornos de AWS, desde 50 Mbps hasta 10 Gbps. Esto le da latencia administrada y controlada y ancho de banda aprovisionado, de esta manera la carga de trabajo se puede conectar fácilmente y de manera eficiente a otros entornos. Con uno de los socios de AWS Direct Connect, puede tener conectividad continua desde múltiples entornos, lo que proporciona una red extendida con rendimiento constante.

Site-to-Site VPN de AWS es un servicio administrado para las VPC. Cuando se crea una conexión de VPN, AWS ofrece túneles para dos puntos de enlace de VPN diferentes. Con AWS Transit Gateway, puede simplificar la conectividad entre múltiples VPC y también conectarse a cualquier VPC conectada a AWS Transit Gateway con una sola conexión de VPN. AWS Transit Gateway también le permite escalar más allá del límite de rendimiento de IPsec VPN de 1.25 Gbps con la habilitación del direccionamiento de rutas múltiples de igual costo (ECMP) a través de múltiples túneles de VPN.

Aproveche el equilibrio de carga y la descarga cifrada: distribuya el tráfico mediante múltiples recursos o servicios para permitir que la carga de trabajo aproveche la elasticidad que ofrece la nube. También puede utilizar el equilibrio de carga para descargar la terminación de cifrado a fin de mejorar el rendimiento y administrar y direccionar el tráfico de manera efectiva.

Cuando implementa una arquitectura escalable en la que desea utilizar múltiples instancias para el contenido del servicio, puede aprovechar los balanceadores de carga en su Amazon VPC. AWS ofrece múltiples modelos para sus aplicaciones en el servicio de ELB. El balanceador de carga de aplicaciones es el más adecuado para el equilibrio de carga del tráfico HTTP y HTTPS y proporciona un direccionamiento de solicitudes avanzado dirigido a la entrega de arquitecturas de aplicaciones modernas, incluidos microservicios y contenedores.

El balanceador de carga de red es el más adecuado para el equilibrio de carga del tráfico TCP donde se requiere un rendimiento extremo. Es capaz de gestionar millones de solicitudes por segundo al mismo tiempo que mantiene latencias ultrabajas y se optimiza para gestionar patrones de tráfico volátiles y repentinos.

Elastic Load Balancing ofrece administración de certificados integrada y descifrado SSL/TLS, lo que le permite la flexibilidad de administrar centralmente la configuración SSL del balanceador de carga y descargar el trabajo intensivo de la CPU de su carga de trabajo.

Elija protocolos de red para optimizar el tráfico de red: tome decisiones sobre los protocolos para la comunicación entre sistemas y redes en función del impacto en el rendimiento de la carga de trabajo.

Existe una relación entre la latencia y el ancho de banda para lograr el rendimiento. Si la transferencia de archivos utiliza TCP, las latencias más altas reducirán el rendimiento general. Hay enfoques para corregir esto con ajustes TCP y protocolos de transferencia optimizados, algunos enfoques utilizan UDP.

Elija la ubicación en función de los requisitos de redes: utilice las opciones de ubicación en la nube disponibles para reducir la latencia de la red o mejorar el rendimiento. Utilice las regiones de AWS, las zonas de disponibilidad, los grupos de ubicación y las ubicaciones de borde, como Outposts, Local Zones (zonas locales) y Wavelength, para reducir la latencia de la red o mejorar el rendimiento.

La infraestructura de la nube de AWS se basa en las regiones y en las zonas de disponibilidad. Una región es una ubicación física en el mundo con múltiples zonas de disponibilidad.

Las zonas de disponibilidad consisten en uno o más centros de datos discretos, cada uno con potencia, redes y conectividad redundantes, alojados en instalaciones separadas. Estas zonas de disponibilidad le ofrecen la capacidad de operar aplicaciones de producción y bases de datos que son más disponibles, tolerantes a fallas y escalables de lo que sería posible desde un solo centro de datos.

Elija la región o regiones adecuadas para su implementación en función de los siguientes elementos clave:

- **Dónde se encuentran los usuarios**: elegir una región cercana a los usuarios de su carga de trabajo, garantiza una latencia más baja cuando utilizan la carga de trabajo.
- **Dónde se encuentran los datos:** para aplicaciones de datos pesados, el cuello de botella más importante en la latencia es la transferencia de datos. El código de la aplicación se debe ejecutar lo más cerca posible de los datos.
- **Otras restricciones:** tenga en cuenta restricciones como la seguridad y la conformidad.

Amazon EC2 ofrece grupos de ubicación para las redes. Un grupo de ubicación es un agrupamiento de instancias lógico dentro de una única zona de disponibilidad. El uso de grupos de ubicación con tipos de instancias compatibles y un Elastic Network Adapter (ENA) permite cargas de trabajo para participar en una red de 25 Gbps de baja latencia. Los grupos de ubicación se recomiendan para cargas de trabajo que se benefician de baja latencia de red, alto rendimiento de red o ambos. El uso de grupos de ubicación tiene la ventaja de bajar la fluctuación en las comunicaciones de red.

Los servicios sensibles a la latencia se entregan en el borde con una red global de ubicaciones de borde. Estas ubicaciones de borde a menudo ofrecen servicios, como la red de entrega de contenido (CDN) y el sistema de nombres de dominio (DNS). Con estos servicios en el borde, las cargas de trabajo pueden responder con baja latencia para solicitar contenido o resolución del DNS. Estos servicios también ofrecen servicios geográficos, como la orientación geográfica de contenido (que ofrece contenido diferente en función de la ubicación de los usuarios finales) o direccionamiento basado en la latencia para orientar a los usuarios finales a la región más cercana (latencia mínima).

Amazon CloudFront es una CDN global que se puede utilizar para acelerar el contenido estático, como imágenes, scripts y videos, así como también contenido dinámico, como las API o aplicaciones web. Depende de una red global de ubicaciones de borde que almacenará en caché el contenido y ofrecerá conectividad de red de alto rendimiento para los usuarios. CloudFront también acelera otras características, como la carga de contenido y las aplicaciones dinámicas, lo que lo convierte en un complemento de rendimiento para todas las aplicaciones que brindan tráfico a través de Internet. Lambda@Edge es una característica de Amazon CloudFront que le permitirá ejecutar código más cerca de los usuarios de su carga de trabajo, lo que mejora el rendimiento y reduce la latencia.

Amazon Route 53 es un servicio web de DNS en la nube escalable y de alta disponibilidad. Está diseñado para brindarle a los desarrolladores y empresa una forma extremadamente fiable y rentable de direccionar a los usuarios finales a las aplicaciones de Internet mediante la traducción de nombres, como www.example.com, a direcciones IP numéricas, como 192.168.2.1, que utilizan las computadoras para conectarse unas con otras. Route 53 es totalmente compatible con IPv6.

AWS Outposts está diseñado para cargas de trabajo que necesitan permanecer en las instalaciones debido a los requisitos de latencia, donde desea que esa carga de trabajo se ejecute sin problemas con el resto de las demás cargas de trabajo en AWS. AWS Outposts es un soporte de cómputo y almacenamiento configurables y totalmente administrados creados con hardware diseñado por AWS que le permite ejecutar el cómputo y el almacenamiento en las instalaciones, mientras se conecta sin problemas a la amplia gama de servicios de AWS en la nube.

AWS Local Zones (zonas locales) es un nuevo tipo de infraestructura de AWS diseñada para ejecutar cargas de trabajo que requieren latencia de milisegundos de un solo dígito, como la representación de videos y las aplicaciones de escritorio virtual intensivas en gráficos. Las zonas locales le permiten obtener todos los beneficios de tener los recursos de cómputo y almacenamiento cercanos al usuario final.

AWS Wavelength está diseñado para entregar aplicaciones de latencia ultrabaja a dispositivos 5G mediante la extensión de la infraestructura, los servicios, las API y las herramientas de AWS para las redes 5G. Wavelength integra el almacenamiento y el cómputo en las redes 5G de los proveedores de telecomunicaciones para ayudar a su carga de trabajo 5G si requiere de

latencia de milisegundos de un solo dígito, como los dispositivos IoT, streaming de videojuegos, vehículos autónomos y la producción de contenido multimedia en directo.

Utilice los servicios de borde para reducir la latencia y habilitar el almacenamiento de caché de contenido. Asegúrese que configuró el control de caché de manera correcta para DNS y HTTP/HTTPS con el fin de obtener el mayor beneficio de estos enfoques.

Optimice la configuración de la red en función de las métricas: utilice los datos recopilados y analizados para tomar decisiones fundamentadas sobre la optimización de la configuración de la red. Mida el impacto de esos cambios y utilice esas mediciones para tomar decisiones futuras.

Habilite los registros de flujo de VPC para todas las redes de VPC utilizadas por la carga de trabajo. Los registros de flujo de VPC son una característica que le permite capturar información sobre el tráfico de IP que va y viene de las interfaces de red en su VPC. Los registros de flujo de VPC lo ayudan con una cantidad de tareas, como la solución de problemas por los que el tráfico específico no llega a una instancia, lo que a su vez lo ayuda a diagnosticar reglas de grupo de seguridad demasiado restrictivas. Puede utilizar los registros de flujo como una herramienta de seguridad para monitorear el tráfico que llega a su instancia, perfilar el tráfico de la red y buscar conductas de tráfico anormales.

Utilice las métricas de la red para hacer cambios en la configuración de la red a medida que evoluciona la carga de trabajo. Las redes basadas en la nube se pueden recrear rápidamente, por lo tanto es necesario que la arquitectura en la red evolucione con el tiempo para mantener la eficiencia del rendimiento.

Recursos

Consulte los siguientes recursos para obtener más información sobre las prácticas recomendadas de AWS para la red.

Videos

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)

Documentación

- Transición al direccionamiento basado en la latencia en Amazon Route 53
- Productos de la red con AWS
- EC2
    - Instancias optimizadas para Amazon EBS
    - Redes mejoradas de EC2 en Linux

  - Redes mejoradas de EC2 en Windows
  - Grupos de ubicación de EC2
  - Habilitar las redes mejoradas con Elastic Network Adapter (ENA) en las instancias de Linux
- VPC
  - Transit Gateway
  - Puntos de enlace de VPC
  - Registros de flujo de VPC
- Balanceador de carga elástico
  - Balanceador de carga de aplicaciones
  - Balanceador de carga de red

Revisión

Cuando diseña las cargas de trabajo, hay una cantidad limitada de opciones entre las que puede elegir. Sin embargo, con el tiempo, las nuevas tecnologías y enfoques estarán disponibles para que pueda mejorar el rendimiento de la carga de trabajo. En la nube, es mucho más fácil experimentar con las nuevas características y servicios porque su infraestructura es un código.

Para adoptar un enfoque basado en los datos para la arquitectura, debe implementar un proceso de revisión de rendimiento que tenga en cuenta lo siguiente:

- **La infraestructura como código**: defina su infraestructura como código mediante enfoques como las plantillas de AWS CloudFormation. El uso de las plantillas le permite colocar la infraestructura en el control de origen junto con el código de aplicación y las configuraciones. Esto le permite aplicar las mismas prácticas que utiliza para desarrollar el software en su infraestructura, de esta manera puede repetir el proceso rápidamente.
- **Implementación de canalización:** utilice una canalización de integración continua e implementación continua (CI/CD) (por ejemplo, repositorio del código de origen, sistemas de creación, implementación y automatización de pruebas) para implementar la infraestructura. Esto le permite implementar de manera repetitiva, constante y a bajo costo a medida que itera.
- **Métricas bien definidas**: configure las métricas y monitoree para capturar indicadores claves de rendimiento (KPI). Recomendamos que utilice tanto las métricas técnicas como las empresariales. Para los sitios web o las aplicaciones móviles, las métricas clave capturan el tiempo hasta el primer byte o representación. Otras métricas que se aplican generalmente incluyen el recuento de hilos, la tasa de recolección de basura y los estados de espera. Las métricas empresariales, como el costo acumulado agregado por solicitud, pueden alertarlo sobre la forma de reducir los costos. Considere cuidadosamente de qué manera planifica para interpretar las métricas. Por ejemplo, puede elegir el percentil máximo o 99 en lugar del promedio.
- **Pruebe el rendimiento de manera automática**: como parte del proceso de implementación, active automáticamente las pruebas de rendimiento después de que las pruebas de ejecución más rápidas hayan pasado con éxito. La automatización debe crear un nuevo entorno, configurar las condiciones iniciales, como los datos de prueba y luego ejecutar una serie de referencias y pruebas de carga. Los resultados de estas pruebas deben vincularse a la creación, de esta manera puede realizar un seguimiento de los cambios de rendimiento a través del tiempo. Para pruebas de larga duración, puede hacer que esta parte de la canalización sea asincrónica del resto de la creación. De manera alternativa, puede ejecutar pruebas de rendimiento durante la noche con las instancias de spot de Amazon EC2.

- **Generación de carga:** debe crear una serie de scripts de prueba que repliquen los traspasos de los usuario sintéticos o pregrabados. Estos scripts deben ser idempotentes y no estar acoplados y puede necesitar incluir scripts de "precalentamiento" para producir resultados válidos. Las pruebas de script deben replicar la conducta de uso en producción tanto como sea posible. Puede utilizar software o soluciones de software como servicio (SaaS) para generar la carga. Tenga en cuenta el uso de las soluciones de AWS Marketplace y las instancias de spot, pueden ser una manera rentable de generar la carga.
- **Visibilidad del rendimiento**: las métricas claves deben ser visibles para su equipo, especialmente las métricas para cada versión de creación. Esto le permite ver cualquier tendencia positiva o negativa a lo largo del tiempo. También debe mostrar métricas sobre la cantidad de errores o excepciones para asegurarse de que está probando un sistema en funcionamiento.
- **Visualización:** utilice técnicas de visualización que dejen en claro dónde ocurren los problemas de rendimiento, los puntos calientes, los estados de espera o la baja utilización. Superponga las métricas de rendimiento sobre los diagramas de arquitectura, los gráficos de llamada o el código pueden ayudarlo a identificar los problemas rápidamente.

Este proceso de revisión del rendimiento se puede implementar como una extensión simple de la canalización de implementación existente y luego evolucionar con el tiempo a medida que los requisitos de prueba se vuelven más sofisticados. Para las arquitecturas futuras, puede generalizar el enfoque y reutilizar el mismo proceso y artefactos.

Las arquitecturas con rendimiento deficiente son generalmente el resultado de un proceso de revisión del rendimiento inexistente o dañado. Si su arquitectura presenta un rendimiento deficiente, implementar un proceso de revisión de rendimiento le permitirá aplicar el ciclo de planificación, ejecución, verificación y reacción (PDCA) de Deming para impulsar una mejora iterativa.

Desarrolle su carga de trabajo para aprovechar los nuevos lanzamientos

Aproveche las innovaciones continuas en AWS impulsadas por la necesidad del cliente. Lanzamos nuevas regiones, ubicaciones de borde, servicios y características regularmente. Cualquiera de estos lanzamientos puede mejorar de manera positiva la eficiencia del rendimiento de la arquitectura.

Manténgase actualizado sobre los nuevos recursos y servicios: evalúe las formas de mejorar el rendimiento a medida que estén disponibles los nuevos servicios, patrones de diseño y ofertas de productos. Determine cuál de ellos puede mejorar el rendimiento o aumentar la

eficiencia de la carga de trabajo mediante la evaluación ad hoc, el debate interno o los análisis externos.

Defina un proceso para evaluar las actualizaciones, las nuevas características y servicios de AWS. Por ejemplo, la creación de pruebas de concepto que utilicen nuevas tecnologías o la consulta con un grupo interno. Cuando pruebe nuevas ideas o servicios, ejecute las pruebas de rendimiento para medir el impacto que tienen sobre la eficiencia o el rendimiento de la carga de trabajo. Aproveche la flexibilidad que tiene en AWS para probar nuevas ideas o tecnologías frecuentemente con un costo o riesgo mínimo.

Defina un proceso para mejorar el rendimiento de la carga de trabajo: defina un proceso para evaluar los servicios nuevos, los patrones de diseño, los tipos de recursos y las configuraciones a medida que estén disponibles. Por ejemplo, ejecute pruebas de rendimiento existentes en ofertas de instancias nuevas para determinar el potencial de mejorar la carga de trabajo.

El rendimiento de la carga de trabajo presenta algunas restricciones claves. Documéntelas, de esta manera puede saber qué tipo de innovaciones pueden mejorar el rendimiento de su carga de trabajo. Utilice esta información cuando aprenda sobre los servicios o tecnologías nuevas a medida que estén disponibles para identificar las formas de mitigar las restricciones o cuellos de botella.

Permita que el rendimiento de la carga de trabajo evolucione con el paso del tiempo: como una organización, utilice la información que se recopila mediante el proceso de evaluación para impulsar activamente la adopción de nuevos servicios o recursos a medida que estén disponibles.

Utilice la información que recopila cuando evalúe nuevos servicios o tecnologías para impulsar el cambio. A medida que su empresa o carga de trabajo cambia, el rendimiento también necesita cambiar. Utilice los datos recopilados de las métricas de la carga de trabajo para evaluar áreas donde puede conseguir los mayores beneficios en eficiencia o rendimiento y adoptar nuevos servicios y tecnologías de manera proactiva a fin de continuar con la demanda.

Recursos

Consulte los siguientes recursos para obtener más información sobre las prácticas recomendadas de AWS para los puntos de referencia.

Videos

- Canal de YouTube de Amazon Web Services
- Canal de YouTube de AWS Online Tech Talks
- Canal de YouTube de AWS Events

Monitoreo

Después de implementar la arquitectura, debe monitorear su rendimiento, de esta manera puede solucionar cualquier problema antes que impacte en los clientes. El monitoreo de las métricas se debe utilizar para encender las alarmas cuando se alcanzan los límites.

El monitoreo en AWS consta de cinco fases diferentes, que se explican de manera más detallada en el documento técnico sobre el pilar de fiabilidad:

1. **Generación**: alcance del monitoreo, métricas y límites
2. **Agregado**: creación de una vista completa para múltiples orígenes
3. **Procesamiento y alarmas en tiempo real**: reconocimiento y respuesta
4. **Almacenamiento**: administración de datos y políticas de retención
5. **Análisis**: paneles, reportes e información

CloudWatch es un servicio de monitoreo para los recursos en la nube de AWS y las cargas de trabajo que se ejecutan en AWS. Puede utilizar CloudWatch para recopilar y realizar un seguimiento de las métricas, recopilar y monitorear archivos de registro y establecer las alarmas. CloudWatch puede monitorear los recursos de AWS, como las instancias EC2 y las instancias de base de datos de RDS, así como también las métricas personalizadas generadas por las cargas de trabajo y los servicios y los archivos de registro que generen las aplicaciones. Puede utilizar CloudWatch para conseguir visibilidad de todo el sistema sobre la utilización de recursos, el rendimiento de la aplicación y el estado operativo. Puede utilizar esta información para reaccionar rápidamente y mantener la carga de trabajo en funcionamiento sin inconvenientes.

Los paneles de CloudWatch le permiten crear gráficos reutilizables de recursos de AWS y métricas personalizadas para que pueda monitorear el estado operativo e identificar problemas a simple vista.

Para una solución de monitoreo efectiva es clave asegurarse de no ver falsos positivos. Los desencadenadores automatizados evitan el error humano y pueden reducir el tiempo que toma solucionar los problemas. Planifique los días de juego, en los que se realizan simulaciones en el entorno de producción, para probar las soluciones de alarma y garantizar que reconozca los problemas de manera correcta.

Las soluciones de monitoreo se dividen en dos tipos: monitoreo activo (AM) y monitoreo pasivo (PM). El monitoreo activo y el monitoreo pasivo se complementan para darle una visión completa del rendimiento de las cargas de trabajo.

El **monitoreo activo** simula la actividad del usuario en traspasos de usuario con scripts mediante rutas críticas en el producto. El monitoreo activo se debe realizar continuamente para probar el rendimiento y la disponibilidad de una carga de trabajo. El monitoreo activo se complementa con el monitoreo pasivo por ser continuo, liviano y predecible. Se puede ejecutar

mediante todos los entornos (especialmente los entornos de preproducción) para identificar problemas o inconvenientes de rendimiento antes de que impacten en el usuario final.

El **monitoreo pasivo** se utiliza generalmente con una carga de trabajo basada en la Web. El monitoreo pasivo recopila las métricas de rendimiento del explorador (las cargas de trabajo que no están basadas en la Web pueden utilizar un enfoque similar). Puede recopilar métricas para todos los usuarios (o un subconjunto de usuarios), geografías, navegadores y tipos de dispositivos. Utilice el monitoreo pasivo para comprender los siguientes problemas:

- **Rendimiento de la experiencia del usuario**: el monitoreo pasivo le ofrece métricas sobre lo que los usuarios experimentan, lo que le da una visión permanente sobre cómo funciona la producción, así como también sobre el impacto de los cambios a través del tiempo.
- **Variabilidad de rendimiento geográfico:** si una carga de trabajo tiene presencia global y los usuarios acceden a las cargas de trabajo desde todo el mundo, con el monitoreo pasivo puede detectar un problema de rendimiento que impacta en los usuarios en una geografía específica.
- **El impacto del uso de las API**: las cargas de trabajo modernas utilizan las API internas y las de terceros. El monitoreo pasivo ofrece visibilidad en el uso de las API, por lo tanto puede identificar cuellos de botella en el rendimiento que no solo se originan en las API internas, sino también de proveedores de API de terceros.

CloudWatch ofrece la habilidad de monitorear y enviar alarmas de notificación. Puede utilizar la automatización para resolver problemas de rendimiento mediante la activación de acciones a través de Amazon Kinesis, Amazon Simple Queue Service (Amazon SQS) y AWS Lambda.

Monitoree sus recursos para garantizar que rinden como se esperaba

El rendimiento del sistema se puede degradar con el tiempo. Monitoree el rendimiento del sistema para identificar la degradación y solucionar los factores internos y externos, como el sistema operativo o la carga de la aplicación.

Registro de las métricas relacionadas con el rendimiento: utilice un servicio de monitoreo y observabilidad para registrar las métricas relacionadas con el rendimiento. Por ejemplo, el registro de las transacciones de bases de datos, consultas lentas, latencia de E/S, rendimiento de solicitud HTTP, latencia de servicio u otro dato clave.

Identifique las métricas de rendimiento importantes para las cargas de trabajo y regístrelas. Estos datos son una parte importante para poder identificar qué componentes afecta el rendimiento general o la eficiencia de la carga de trabajo.

Cuando trabaje desde la experiencia del cliente, identifique las métricas importantes. Para cada métrica, identifique el objetivo, el enfoque de medición y la prioridad. Utilícelos para crear alarmas y notificaciones para abordar de manera proactiva los problemas relacionados con el rendimiento.

Analice las métricas cuando ocurren eventos o incidentes: en respuesta a (o durante) un evento o incidente, utilice los paneles o reportes de monitoreo para comprender y diagnosticar el impacto. Estas visualizaciones ofrecen información sobre qué partes de la carga de trabajo no funcionan como se esperaba.

Cuando escriba historias de usuarios críticas para su arquitectura, incluya requisitos de rendimiento, como especificar con qué rapidez se debe ejecutar cada historia crítica. Para estas historias críticas, implemente traspasos de usuario con secuencias de comandos adicionales para asegurarse que conozca sobre cómo estas historias funcionan según sus requisitos.

Establezca indicadores clave de rendimiento (KPI) para medir el rendimiento de la carga de trabajo: identifique los KPI que indican si la carga de trabajo rinde como se previó. Por ejemplo, una carga de trabajo basada en las API puede utilizar latencia de respuesta general como una indicación del rendimiento general y un sitio de comercio electrónico podría elegir usar el número de compras como su KPI.

Documente la experiencia de rendimiento solicitada por los clientes, incluido de qué manera los clientes juzgarán el rendimiento de la carga de trabajo. Utilice estos requisitos para establecer los indicadores claves de rendimiento (KPI), los que indicarán cómo rinde el sistema en general.

Utilice el monitoreo para generar notificaciones basadas en las alarmas: con los indicadores claves de rendimiento (KPI) relacionados con el rendimiento que ha definido, utilice un sistema de monitoreo que genera alarmas automáticamente cuando estas medidas están fuera de los límites esperados.

Amazon CloudWatch puede recopilar métricas mediante los recursos en su arquitectura. También puede recopilar y publicar métricas personalizadas para los negocios de superficie o métricas derivadas. Utilice CloudWatch o un servicio de monitoreo de terceros para establecer las alarmas que indican cuándo se alcanzan los límites. Las alarmas indican que una métrica está fuera de los límites esperados.

Revise las métricas en intervalos regulares: como rutina de mantenimiento o en respuesta a eventos o incidentes, revise que métricas se recopilan. Utilice estas revisiones para identificar que métricas eran claves en abordar los problemas y qué métricas adicionales, si se estuviera realizando un seguimiento, ayudarían a identificar, abordar o prevenir problemas.

Como parte de la respuesta a incidentes o eventos, evalúe qué métricas fueron útiles para abordar el problema y qué métricas podrían haber ayudado que actualmente no se rastrean. Utilice esto para mejorar la calidad de las métricas que recopila, de esta manera puede evitar o resolver incidentes futuros más rápidamente.

Monitoree y active las alarmas de manera proactiva: utilice los indicadores clave de rendimiento (KPI), combinados con los sistemas de monitoreo y alerta, para abordar de manera proactiva problemas relacionados con el rendimiento. Utilice alarmas para desencadenar acciones automatizadas a fin de solucionar los problemas donde sea posible. Escale la alarma a aquellos que puedan responder si no es posible una respuesta automatizada. Por ejemplo, puede tener un sistema que puede predecir los valores esperados de los indicadores clave de rendimiento (KPI) y la alarma cuando alcanzan ciertos límites o una herramienta que automáticamente puede detener o revertir las implementaciones si los KPI están fuera de los valores esperados.

Implemente procesos que ofrecen visibilidad en el rendimiento a medida que la carga de trabajo se ejecuta. Cree paneles de monitoreo y establezca normas de referencia para las expectativas de rendimiento para determinar si la carga de trabajo funciona de manera óptima.

Recursos

Consulte los siguientes recursos para obtener más información sobre las prácticas recomendadas de AWS respecto al monitoreo para promover la eficiencia del rendimiento.

Videos

- Cut through the chaos: Gain operational visibility and insight (MGT301-R1)

Documentación

- Documentación de X-Ray
- Documentación de CloudWatch

Compensaciones

Cuando diseñe las soluciones de arquitectura, piense en las compensaciones para garantizar un enfoque óptimo. En función de su situación, puede intercambiar la consistencia, la durabilidad y el espacio por tiempo o latencia, para entregar un rendimiento mayor.

Con AWS, puede incorporarse al mercado global rápidamente e implementar recursos en múltiples ubicaciones en el mundo para acercarse a sus usuarios finales. También puede agregar de manera dinámica réplicas de solo lectura a los almacenes de información (como sistemas de bases de datos) para reducir la carga en la base de datos primaria.

AWS ofrece soluciones de almacenamiento de caché, como Amazon ElastiCache, que proporciona un almacén de datos en la memoria o caché y Amazon CloudFront, que almacena copias en caché de contenido estático más cerca de los usuarios finales. Amazon DynamoDB Accelerator (DAX) ofrece una capa de almacenamiento de caché distribuido de lectura y escritura frente a Amazon DynamoDB, que es compatible con la misma API, pero ofrece una latencia inferior a milisegundos para las entidades que están en la caché.

Uso de las compensaciones para mejorar el rendimiento

Cuando diseñe soluciones, considerar las compensaciones de manera activa le permite seleccionar un enfoque óptimo. A menudo, puede mejorar el rendimiento con el intercambio de la consistencia, la durabilidad y el espacio por tiempo y latencia. Las compensaciones pueden aumentar la complejidad de la arquitectura y requiere la prueba de carga para asegurar que se obtenga un beneficio cuantificable.

Comprenda las áreas donde el rendimiento es más crítico: comprenda e identifique las áreas donde el aumento del rendimiento de su carga de trabajo tendrá un impacto positivo en la eficiencia o la experiencia del cliente. Por ejemplo, un sitio web que tiene una gran interacción con los clientes puede beneficiarse de utilizar servicios de borde para acercar la entrega de contenidos a los clientes.

Aprenda sobre los servicios y patrones de diseño: investigue y comprenda los diferentes servicios y patrones de diseño que pueden ayudar a mejorar el rendimiento de la carga de trabajo. Como parte del análisis, identifique lo que podría intercambiar para lograr un mejor rendimiento. Por ejemplo, con un servicio de caché puede ayudar a reducir la carga en los sistemas de la base de datos; sin embargo, implementar un almacenamiento de caché seguro o una posible introducción de consistencia final en algunas áreas requiere de algo de ingeniería.

Conozca qué opciones de configuración de rendimiento están disponibles y cómo pueden impactar en la carga de trabajo. La optimización del rendimiento de la carga de trabajo depende de entender cómo estas opciones interactúan con su arquitectura y el impacto que tendrán tanto en el rendimiento medido como en el rendimiento percibido por los usuarios.

La Amazon Builders´ Library les ofrece a los lectores una descripción detallada de cómo Amazon crea y opera la tecnología. Estos artículos gratuitos están escritos por ingenieros sénior de Amazon y cubren temas de arquitectura, entrega de software y operaciones. Por ejemplo, puede ver de qué manera Amazon automatiza la entrega de software para alcanzar más de 150 millones de implementaciones al año o cómo los ingenieros de Amazon implementan principios, como la partición aleatoria para crear sistemas resistentes de alta disponibilidad y tolerantes a fallas.

Identifique cómo las compensaciones impactan en los clientes y en la eficiencia: cuando evalúe las mejoras relacionadas con el rendimiento, determine que opciones impactarán en sus clientes y en la eficiencia de la carga de trabajo. Por ejemplo, si el uso de un almacén de datos de valor clave aumenta el rendimiento del sistema, es importante evaluar de qué manera la naturaleza finalmente constante de esto impactará en los clientes.

Identifique áreas de rendimiento deficientes en el sistema mediante métricas y monitoreo. Determine de qué manera puede implementar mejoras, qué compensaciones brindan esas mejoras y cómo impactan en el sistema y la experiencia del usuario. Por ejemplo, la implementación de datos en caché puede ayudar a mejorar drásticamente el rendimiento, pero requiere una estrategia clara en cuanto a cómo y cuándo actualizar o invalidar los datos en caché para evitar conductas incorrectas del sistema.

Mida el impacto de mejoras de rendimiento: como los cambios se llevan a cabo para mejorar el rendimiento, evalúe la recopilación de métricas y datos. Utilice esta información para determinar el impacto que la mejora del rendimiento tuvo en la carga de trabajo, en los componentes de la carga de trabajo y en los clientes. Estas medidas ayudan a comprender las mejoras que resultan de las compensaciones y lo ayudan a determinar si se introdujo algún efecto secundario negativo.

Un sistema de buena arquitectura utiliza una combinación de estrategias relacionadas con el rendimiento. Determine qué estrategia tendrá el mayor impacto positivo en un punto de conflicto o cuello de botella determinado. Por ejemplo, la partición de datos mediante múltiples sistemas de bases de datos relacionales puede mejorar el rendimiento general al mismo tiempo que conserva el soporte para las transacciones y, en cada partición, el almacenamiento de caché puede ayudar a reducir la carga.

Utilice diversas estrategias relacionadas con el rendimiento: según corresponda, utilice múltiples estrategias para mejorar el rendimiento. Por ejemplo, el uso de estrategias, como el caché de datos para evitar demasiadas llamadas a la red o a la base de datos, el uso de réplicas de lectura para motores de bases de datos a fin de mejorar los índices de lectura, la partición o compresión de datos cuando sea posible para reducir volúmenes de datos y el almacenamiento en búfer y streaming de los resultados a medida que estén disponibles para evitar el bloqueo.

A medida que implementa cambios en la carga de trabajo, recopile y evalúe las métricas para determinar el impacto de esos cambios. Mida el impacto en los sistemas y en el usuario final

para comprender de qué manera las compensaciones impactan en la carga de trabajo. Utilice un enfoque sistemático, como la prueba de carga, para explorar si las compensaciones mejoran el rendimiento.

Recursos

Consulte los siguientes recursos para obtener más información sobre las prácticas recomendadas de AWS para el almacenamiento de caché.

Video

- Introducing The Amazon Builders' Library (DOP328)

Documentación

- Amazon Builders' Library
- Prácticas recomendadas para implementar Amazon ElastiCache

Conclusión

El logro y mantenimiento de la eficiencia de rendimiento requiere de un enfoque basado en datos. Debe considerar activamente los patrones de acceso y las compensaciones que le permitirán optimizar para un mayor rendimiento. Con un proceso de revisión en función de puntos de referencia y pruebas de carga, puede seleccionar los tipos de recursos y configuraciones adecuados. Tratar la infraestructura como código le permite desarrollar su arquitectura de manera rápida y segura, y al mismo tiempo utilizar los datos para tomar decisiones basadas en hechos sobre la arquitectura. Poner en marcha una combinación de monitoreo activo y pasivo garantiza que el rendimiento de su arquitectura no se degrade con el paso del tiempo.

AWS se esfuerza para ayudarlo a crear arquitecturas que funcionen de manera eficiente mientras brindan valor comercial. Utilice las herramientas y técnicas analizadas en este documento para garantizar el éxito.

Colaboradores

Las siguientes personas y organizaciones contribuyeron en este documento:

- Eric Pullen, líder en eficiencia de rendimiento de buena arquitectura, Amazon Web Services
- Philip Fitzsimons, director sénior de buena arquitectura, Amazon Web Services
- Julien Lépine, director especialista en SA, Amazon Web Services
- Ronnen Slasky, arquitecto de soluciones, Amazon Web Services

Documentación adicional

Para ayuda adicional, consulte las siguientes fuentes:

- Marco de Buena Arquitectura de AWS

Revisiones de documentos

Fecha	Descripción
Abril de 2020	Actualización importante a v2
Julio de 2018	Actualización menor por problemas gramaticales
Noviembre de 2017	Actualización del documento técnico para reflejar los cambios en AWS
Noviembre de 2016	Primera publicación