Hacia un análisis centrado en el cliente de la disponibilidad en IaaS

Antonio Manuel Gutiérrez Fernández, Pablo Fernández, Manuel Resinas, Antonio Ruiz-Cortés
Escuela Técnica Superior de Ingeniería Informática, Universidad de Sevilla

<{amgutierrez, pablofm, resinas, aruiz@us.es}>

Este artículo fue seleccionado para su publicación en ***Novática*** entre las ponencias presentadas en las X Jornadas de Ciencia e Ingeniería de los Servicios (JCIS-2014) celebradas en Cádiz en septiembre de 2014 y de las que ATI fue entidad colaboradora.

Resumen: *La disponibilidad es una propiedad presente en los Acuerdos de Nivel de Servicios (ANSs) de la mayoría de servicios de infraestructura, tanto de computación (Amazon EC2, Windows Azure, Google Cloud, Joyent, Rackspace…) como de almacenamiento (Amazon S3, Google Cloud Storage, etc). Siendo una propiedad básica bien conocida y bien definida en infraestructuras tradicionales (*on-premise*), en el caso de IaaS existen importantes diferencias en relación a su alcance y la forma de compensar a las partes cuando se analiza el cumplimiento del ANS. Además, la disponibilidad se describe en lenguaje natural con frecuencia muy verboso y usando un vocabulario propio que ciertamente dificulta la comprensión por los potenciales clientes. Estas circunstancias hacen que el análisis comparativo y sistemático de la disponibilidad de un conjunto de proveedores de IaaS sea una actividad repetitiva, costosa y propensa a errores. En este artículo, describimos en detalle este problema e introducimos una primera aproximación para abordar el análisis de los ANSs basada en las técnicas de análisis de ANSs actuales.*

Palabras clave: *Acuerdos de Nivel de Servicio, almacenamiento,* cloud, *disponibilidad, IaaS, virtualización.*

1. Introducción

Hoy en día, los servicios en la nube se utilizan de forma masiva para proveer de infraestructura de computación (*Infrastructure as a Service*, IaaS) en el ámbito corporativo. Los clientes de estos servicios externalizan la gestión de la infraestructura para enfocarse en su modelo de negocio.

Los Acuerdos de Nivel de Servicio (ANSs) establecen niveles de calidad acordados entre cliente y proveedor de un servicio en el consumo del mismo. Para garantizar los niveles de calidad acordados, se incluyen responsabilidades sobre los mismos, normalmente en forma de penalizaciones en caso de incumplimiento.

Un ANS para servicios computacionales suele establecer valores acordados sobre periodos de disponibilidad (de 24 horas x 7 días, horario de oficina…), de rendimiento (peticiones por segundo, tiempo de respuesta…) y la penalización al proveedor en caso de incumplimiento, normalmente como compensación al cliente.

Los clientes de servicios de infraestructura en la nube actúan típicamente como proveedores de soluciones a terceras partes, por lo que deben revisar con cuidado las garantías y responsabilidades acordadas en el ANS de infraestructura que soporta sus propios servicios.

Sin embargo, a la hora de hacer un análisis de los diferentes proveedores de IaaS existen varios problemas. Los principales proveedores, tales como Amazon, Google, Rackspace o Joyent, proporcionan un ANS en el que la definición de los términos de garantía del servicio de infraestructura se basa en características tecnológicas propias del proveedor. Con lo que, por un lado, es difícil establecer un marco comparativo entre diferentes proveedores y, por otro, para el cliente de IaaS es difícil relacionar garantías en términos tecnológicos con las que él, como proveedor de servicios de más alto nivel (SaaS, PaaS, etc.) quiere garantizar (por ej., disponibilidad en términos tecnológicos específicos, como paradas de la máquinas por tareas de mantenimiento o errores de operaciones de lectura/escritura en disco, frente a una definición de disponibilidad de una aplicación de video *online)*.

En los ANSs, los niveles garantizados (o términos de garantía) se expresan habitualmente como restricciones sobre ciertas propiedades de calidad del servicio, tales como la latencia, el rendimiento o la disponibilidad. Estas propiedades dependen de la naturaleza del servicio (almacenamiento, computación, conectividad, bases de datos, etc.). En el caso de la disponibilidad, todos los proveedores ofrecen al menos un término de garantía relacionado con ella, aunque no existe una descripción comúnmente aceptada de su alcance y su modelo de compensación.

En este artículo, abordamos como asistir y automatizar el estudio comparativo de las garantías ofrecidas por los proveedores de IaaS, desde el dominio del negocio y el lenguaje del cliente final, y centrándonos en la disponibilidad. Para planificar el despliegue de la infraestructura, el cliente evalúa como se ajustan las garantías de los ANSs de los proveedores a los requisitos impuestos por los servicios que él provee sustentados en esta infraestructura. Sin embargo, a pesar de que las garantías en lenguaje natural son fácilmente entendibles por los clientes, la evaluación manual de cómo estas garantías se ajustan a las necesidades del cliente es tediosa, costosa y propensa a errores, por lo que automatizar dicho análisis tendrá un gran impacto en el plan de negocio del cliente [1].

Los principales servicios de infraestructura ofrecen computación y almacenamiento, por lo que nos centramos en ambos tipos de servicios. Particularmente, en el caso de los servicios de computación, tomamos ANSs donde se garantiza la disponibilidad de máquinas individuales. En cada tipo de servicio, la propiedad "disponibilidad" tiene diferentes definiciones, con semánticas vinculadas a la tecnología, con lo que las unidades métricas sobre esta propiedad son distintas y las preferencias del cliente se analizan considerando estas diferencias.

El primer paso para asistir en la comparación de garantías es modelar las garantías propuestas por los proveedores mediante un lenguaje formal. En segundo término, modelamos los requisitos del cliente como Preguntas Frecuentes (FAQ).

Con las FAQ definimos las preguntas como operaciones que un componente software pueda automatizar sobre el modelo formal de las garantías. En el dominio de los servicios computacionales, WS-Agreement es un esquema muy conocido y usado para definir ANSs, con soporte para términos de penalización y compensación y que usamos de soporte para nuestra aproximación [2]. En un segundo paso, al no existir una herramienta o solución para automatizar la evaluación de las garantías sobre la disponibilidad y de la posible aplicación de

> “ WS-Agreement es un esquema muy conocido y usado para definir Acuerdos de Nivel de Servicio (ANSs), con soporte para términos de penalización y compensación ”

penalizaciones o recompensas, describimos las operaciones necesarias para automatizar la respuesta a las cuestiones propuestas.

En concreto, introducimos tres cuestiones básicas, de interés para analizar las garantías de los proveedores de servicios de infraestructura. Estas cuestiones son:

- Q1: Dada la garantía de disponibilidad, ¿cuál es el máximo tiempo que puede estar el servicio no disponible de manera continuada sin que se apliquen penalizaciones?

- Q2: ¿Qué compensación recibe el cliente cuando el servicio ha estado no disponible durante N minutos consecutivos?

- Q3: ¿Cuánto tiempo ha de transcurrir con el servicio no disponible continuadamente para que el cliente reciba la máxima compensación establecida en el ANS?

En las siguientes secciones, describimos los ANSs para diferentes proveedores de computación (**sección 2**) y almacenamiento (**sección 3**). En la **sección 4** proponemos abordar el problema por medio del modelado de los ANSs en WS-Agreement y esbozamos las líneas principales para la automatización de estas cuestiones en forma de operaciones de análisis.

2. Disponibilidad en servicios de computación

2.1. Rackspace

La garantía sobre la disponibilidad de las máquinas del ANS de Rackspace establece que (traducido del original[1]):

“*Garantizamos el funcionamiento de todos las máquinas en la nube, incluyendo los servicios de computación, almacenamiento e hipervisor. Si una máquina en la nube falla, garantizamos que la restauración o reparación se completará en una hora desde la identificación del problema. Si fallamos en cumplir esa garantía, usted recibirá un crédito. Los créditos serán calculados como un porcentaje de las tarifas para los servidores en la nube afectados por el fallo del periodo actual de facturación mensual durante el que ocurrió el fallo (y será aplicado al final del ciclo de facturación), como sigue: Máquinas en la nube: 5% de las tarifas de la máquina por cada hora adicional de caída, hasta un 100% de la tarifa del servidor…”.*

En consecuencia, la garantía de disponibilidad excluye los primeros 60 minutos y ofrece un 5% de la facturación mensual por cada intervalo posterior de 60 minutos. De este modo, las respuestas a las cuestiones planteadas serían:

- Q1: El servicio puede estar no disponible hasta un máximo de 119 minutos sin que el cliente tenga derecho a compensación. Fíjese que los primeros 60 minutos tras la caída del servidor se consideran dedicados a la restauración de la misma, y que a partir de ahí, cada hora adicional sin restaurar da derecho al cliente a un crédito (penalización a Rackspace) del 5% de la factura. Es decir, que el cliente no recibe compensación alguna hasta que transcurran al menos 120 minutos tras la caída del servicio.

- Q2: El cliente recibirá una compensación del 5% de la factura mensual cuando el período de no disponibilidad esté entre 2 y 3 horas. Dicha compensación se incrementará en un 5% por cada hora o fracción adicional de no disponibilidad. La compensación máxima (100% de la facturación) se alcanza a las 21 horas (1.260 minutos) de no disponibilidad.

- Q3: La máxima penalización es alcanzada cuando las penalizaciones alcanzan el 100%. Como comentábamos en la pregunta anterior, ésta se alcanza a las 21 horas.

2.2. Joyent

En el ANS de Joyent la garantía sobre la disponibilidad de las máquinas se expresa en los siguientes términos (traducción del original[2]):

“*Objetivos: El objetivo de Joyent es conseguir el 100% de la disponibilidad de todos los clientes. Recurso: Sujeto a ciertas excepciones, si la disponibilidad del servicio al cliente es menor que el 100%, Joyent le dará al cliente un crédito del 5% de la factura mensual por cada 30 minutos de no disponibilidad (hasta el 100% de las tarifas mensuales de la máquina afectada).”.*

En consecuencia, Joyent garantiza cualquier máquina no disponible con un 5% por cada periodo de 30 minutos. Así, de manera análoga a Rackspace, la respuesta a las cuestiones planteadas es:

- Q1: El servicio no puede estar no disponible, por lo que no hay periodo de no disponibilidad sin penalización al proveedor. Nótese que el crédito del 5% de la factura mensual se otorga desde el primer minuto de no disponibilidad hasta el minuto 30; el 10% desde el minuto 31 al 60 y así sucesivamente.

- Q2: El cliente recibirá una compensación del 5% de la factura mensual por cada periodo de no disponibilidad de 30 minutos (comenzando desde el primer minuto), hasta un máximo de 571 minutos, en el que se compensa al cliente por el 100% de la factura mensual.

- Q3: En este caso, la penalización máxima del 100% se alcanza a los 571 minutos de no disponibilidad.

3. Disponibilidad en servicios de almacenamiento

Los servicios de almacenamiento son garantizados de manera similar a los servicios de computación, pero debido a la diferente operativa de la computación, la semántica de la disponibilidad es diferente.

En los servicios de almacenamiento, para evaluar que el servicio está disponible o no, se considera no solo el tiempo transcurrido sino las operaciones de lectura/escritura realizadas. De manera que las garantías se establecen en función de las operaciones fallidas a lo largo del tiempo.

Teniendo esto en cuenta, para responder a las preguntas el cliente debe indicar qué se considera “No disponible” en términos de almacenamiento en base al número de peticiones fallidas que se admiten. Esto depende de la naturaleza del negocio. Esto es, si el cliente utiliza el sistema de almacenamiento en un sistema crítico en el que no se pueden admitir ni un fallo de operación, la disponibilidad implica un 0% de fallo en las operaciones (ó 100% de éxito). En cambio, para aplicaciones no críticas, donde se pueden admitir operaciones de lectura/escritura incorrectas (por ejemplo, por cuestiones de rendimiento), un cliente puede considerar que el servicio está disponible cuando menos del 20% de las operaciones de lectura/escritura fallen (o dicho de otro modo, al menos un 80% de operaciones con éxito).

> “ En los servicios de almacenamiento, para evaluar que el servicio está disponible o no, se considera no solo el tiempo transcurrido sino las operaciones de lectura/escritura realizadas ”

Este valor umbral de fallos admitidos con éxito (UFA en adelante) será provisto como parámetro por el cliente para la evaluación de la disponibilidad en servicios de almacenamiento y tendrá que ser tenido en cuenta en el diseño de las operaciones. Teniendo en cuenta el valor UFA, que establece el mínimo de peticiones erróneas para considerar el periodo como no disponible, las respuestas a las preguntas tendrán la forma de intervalo temporal. Este intervalo tendrá como límite menor el correspondiente a un 0% de peticiones válidas y como límite mayor el correspondiente justo al valor de errores mínimos definidos en el UFA (para peticiones válidas mayores al 0%, tendremos, de manera general, menor penalización que para un 0% en el mismo periodo de tiempo, dicho de otra forma, se necesitará más tiempo para tener la misma penalización).

Como escenarios de ejemplo para las cuestiones planteadas, tomamos los servicios de almacenamiento Google Cloud Storage y Amazon S3.

3.1. Amazon S3

En Amazon S3[3], la garantía de disponibilidad del almacenamiento se define a partir de dos conceptos:

- *Error Rate*: Número de peticiones erróneas divididas por el número total de peticiones en un intervalo de 5 minutos.

- *Monthly Uptime Percentage* (MUP): 100% menos el promedio de *error rates* en un mes.

Así, el MUP, depende de la distribución de peticiones por periodos de 5 minutos y el *Error Rate* en dichos periodos. Las posibles penalizaciones dependen del MUP con la siguiente regla:

- Si el MUP es mayor que 99% y menor que 99,9%, la penalización es del 10% de la factura mensual.

- Si el MUP es menor que el 99%, la penalización es del 25% de la factura mensual.

Analizando este ANS, las respuestas a las cuestiones de referencia son:

- Q1: El máximo intervalo de no disponibilidad depende de la distribución de errores en las peticiones por intervalos de 5 minutos. El cálculo de los intervalos es el resultado de redondear los valores límite de la fórmula planteada abajo a partir de UFA, a partir del ANS. Siendo 8.640 el nº Total de Intervalos de 5 minutos en 1 mes de 30 días (por simplificar). Así, con un UFA de 0%, cualquier fallo se consideraría como servicio no disponible, así que asumiendo al menos una petición por intervalo, la respuesta a la pregunta es el rango que va desde el 100% de fallos hasta el 0,1% de fallos. Esto es, que el máximo periodo sin penalización va desde 45 minutos (9 intervalos x 5 minutos) hasta el mes completo (8.640 x 5 minutos), es decir podríamos tener entre 45 y 43.200 minutos sin penalización. En cambio, con un UFA del 20%, es decir, consideramos no disponible el servicio desde un 20% de errores hasta un 100% de errores, podrían pasar entre 9 intervalos (es decir, 45 minutos) hasta 43 intervalos (215 minutos) sin recibir compensaciones. En este segundo caso, si el *Error Rate* es menor al 20%, podríamos llegar a recibir compensación, pero el cliente ni siquiera consideraría que el servicio no está disponible (ver **figura 1**).

$$\frac{Tasa\ de\ fallos * Nº\ Intervalos}{8640} \geq 0{,}1\%$$

Figura 1. Fórmula para Amazon S3 que indica el momento a partir del cual el cliente tiene derecho a compensación.

En la **figura 2** se puede ver el efecto del *Error Rate* en los minutos sin penalización. El rango de éstos depende de UFA.

- Q2: De nuevo, la respuesta depende de la UFA. Con un UFA del 0% en las peticiones, la penalización depende únicamente del periodo temporal considerado. Por encima de 45 minutos y por debajo de 450, el cliente obtiene un 10% de compensación sobre su factura. Por encima de 450 minutos, obtiene un 25% de compensación.

- Q3: De manera análoga a la cuestión Q1, por debajo del 99% tenemos la máxima penalización (25% de devolución). Esto es, en el caso peor, con todas las peticiones erróneas e independientemente del UFA, con el 1% de periodos de 5 minutos sobre el total del mes, obtendríamos la máxima penalización. Así, con 450 minutos, se aplicaría la máxima penalización.

3.2. Google Cloud Storage

El ANS de Google Cloud Storage[4] describe sus garantías sobre la disponibilidad en el almacenamiento de manera similar a Amazon S3, apoyándose en una tasa de error de las peticiones. En este caso, los conceptos clave son:

- *Error Rate*: Número de peticiones erróneas, dividido por el número total de peticiones válidas.

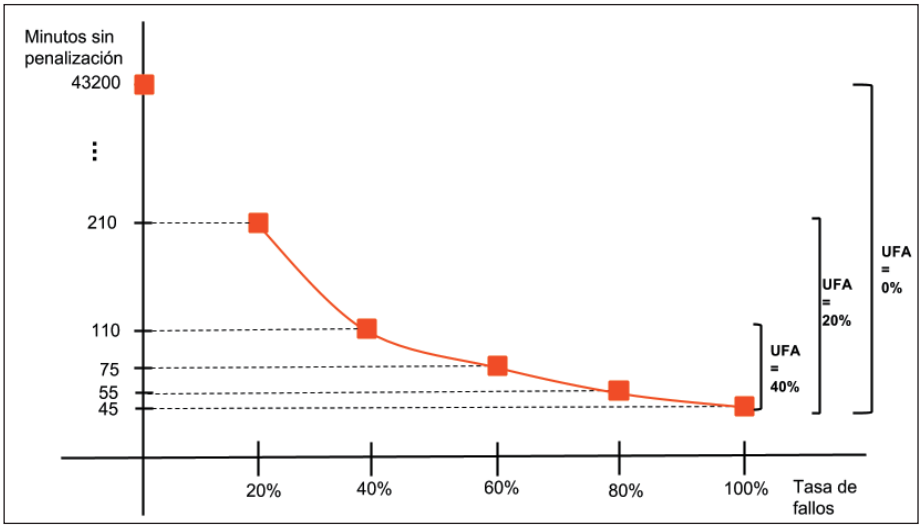

Figura 2. Rango de minutos sin penalización en función del umbral de fallos admitido (UFA) en Amazon S3.

> “El ANS de Google Cloud Storage describe sus garantías sobre la disponibilidad en el almacenamiento de manera similar a Amazon S3, apoyándose en una tasa de error de las peticiones”

- Periodo de no disponibilidad: Intervalos de 10 minutos consecutivos donde el *Error Rate* es mayor del 5%.

- *Monthly Uptime Percentage* (MUP): Minutos totales de un mes menos el número total de periodos de no disponibilidad dividido por el número total de minutos del mes.

Google establece dos ANSs diferentes para garantizar la disponibilidad. Consideramos el ANS estándar, ya que los conceptos son similares en ambos y la diferencia radica en una mayor disponibilidad a un mayor coste. El ANS estándar define la siguiente penalización:

- Si el MUP es mayor o igual que el 99% pero menor que el 99,9%, la penalización es el de 10% de descuento de la factura mensual.

- Si el MUP es mayor o igual que el 95% y menor que el 99%, la penalización es del 25% de descuento.

- Si el MUP es menor que el 95%, la penalización es del 50% del crédito.

Así, de manera similar a Amazon, las respuestas a las cuestiones de referencia dependen de la distribución de peticiones y fallos:

- Q1: De nuevo, la respuesta depende del UFA y del *Error Rate*. En este caso, dado que solo se tienen en cuenta *Error Rate* mayores al 5%, un UFA muy alto (mayor al 95%) haría que pudiéramos tener el servicio no disponible nunca, sin recibir compensación (con un porcentaje de errores menor a 100% - UFA). Considerando todas las peticiones fallidas hasta el minuto 50 (0,1% de los periodos de 10 minutos mensuales), no se aplicaría ninguna penalización.

- Q2: La respuesta a esta cuestión, como en el caso de Amazon S3, depende de la distribución de peticiones y fallos.

- Q3: En este caso, la penalización se calcula de manera similar a Amazon S3, aunque Google ofrece mayor garantía para escenarios pesimistas.

4. Nuestra propuesta

Hasta donde sabemos, las soluciones existentes se enfocan a la monitorización de la infraestructura y validar el ANS, pero ninguna propuesta se centra en el análisis en tiempo de diseño de la disponibilidad garantizada por los proveedores. El primer paso para automatizar el análisis de la disponibilidad es modelar las garantías propuestas mediante un lenguaje formal de manera que un componente software pueda resolver las cuestiones propuestas.

Para ello, basamos nuestra propuesta en WS-Agreement, que es un estándar ampliamente usado y que se utiliza con éxito en el ámbito computacional. WS-Agreement es la propuesta más destacada para modelar ANSs y existe un gran número de herramientas que soportan la edición y análisis de documentos WS-Agreement, como nuestro entorno de gestión de acuerdos, IDEAS[5].

4.1. Modelando ANSs con WS-Agreement

La especificación WS-Agreement define un metamodelo para Acuerdos de Nivel de Servicio. Este metamodelo propone un documento con varias secciones: El nombre, el contexto y los términos. El contexto provee información relativa a los participantes del acuerdo (es decir, proveedor y consumidor del servicio) o el periodo de validez del acuerdo. La sección de término describe el acuerdo en sí.

Se distinguen dos tipos de términos, llamados términos de descripción del servicio (SDT, del inglés) y términos de garantía (GT). Los términos del servicio son aquellos que identifican condiciones inherentes al servicio y que no pueden ser negociadas. Los términos de garantía establecen objetivos de nivel de servicio (SLOs) sobre las propiedades del servicio y definen el participante obligado a cumplirlos. Un SLO es una restricción sobre una propiedad que debe ser monitorizable (para comprobar la validez de la garantía).

Los términos de garantía pueden estar acompañados de una condición de cualificación (QC), que indica una precondición para aplicar la restricción del SLO. La valoración de una garantía se hace mediante la llamada Lista de Valores de Negocio [3][5]. Los valores de negocio incluyen la expresión de la importancia de la propiedad del SLO así como las posibles penalizaciones o recompensas. La expresión de los términos de garantía se utilizan para analizar los Acuerdos (por ejemplo, comprobar si se está violando para renegociarlo o el cálculo de las penalizaciones).

Usando la estructura de documento de WS-Agreement, los ANSs en lenguaje natural provistos por Joyent y Rackspace se definen de una manera directa (ver **figura 3** y **figura 4**).

La sintaxis de las figuras es iAgree. iAgree es una propuesta de C. Müller [4] que utiliza una sintaxis alternativa para documentos WS-Agreement más legible para humanos que el XML del estándar y, además, proporciona un lenguaje específico para definir SLOs.

Tal como las figuras muestran, la disponibilidad garantizada por Rackspace y Joyent se corresponden con Objetivos de Nivel de Servicio de la disponibilidad menores a 120 e iguales a 0 minutos, respectivamente. La métrica de la disponibilidad aparece como la variable MDT y la variable MDRTi para Joyent y Rackspace, respectivamente. MDT es el número de minutos acumulados (continuos o no) de no disponibilidad y MDRTi es el número de minutos de no disponibilidad por una caída *i* del servicio. Estas variables se definirán como las medidas de la no disponibilidad y dependen de la semántica que cada proveedor da a dichas métricas y que se ha descrito en la **sección 2**.

Así, Joyent considera que la suma de todo el tiempo en el que la máquina está inaccesible debe acumularse para el cálculo de las penalizaciones, mientras que en Rackspace hay que tener en cuenta cada periodo de no disponibilidad por separado y restar el margen de mantenimiento. Para simplificar el escenario, ciertas limitaciones en la garantía (como las tareas planificadas de mantenimiento), no se incluyen, pero podrían ser añadidas como condiciones de cualificación. Las penalizaciones son incluidas como Valores de Negocio para el intervalo mensual de pago, usando una expresión matemática acorde a las tareas planificadas de mantenimiento, no se incluyen, pero podrían ser añadidas

> “El reto en la automatización de dichas operaciones es que al tener las definiciones de no disponibilidad semánticas muy distintas, no es posible proveer soluciones genéricas, sino que es necesario adaptar la métrica de no disponibilidad al proveedor concreto”

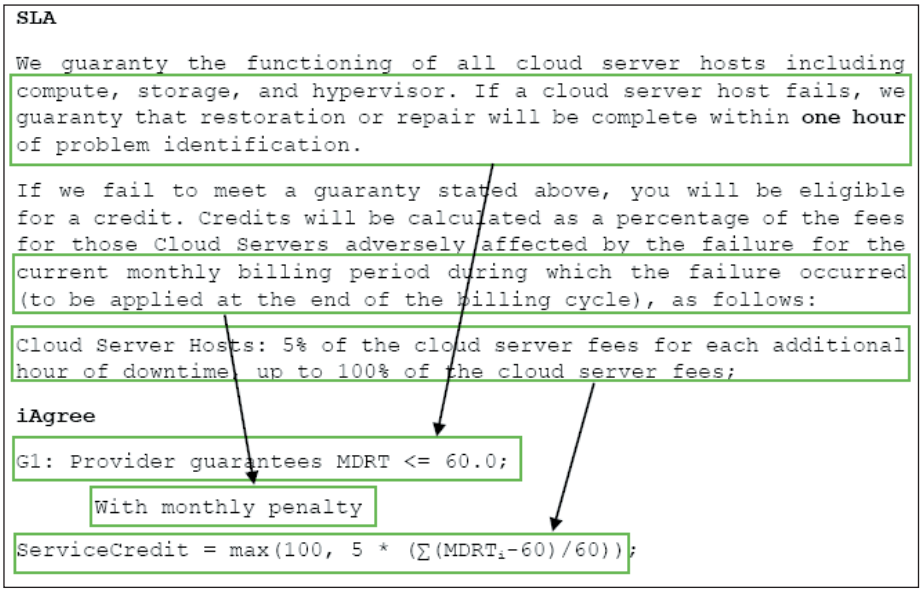

Figura 3. Modelado del ANS de Joyent a iAgree.

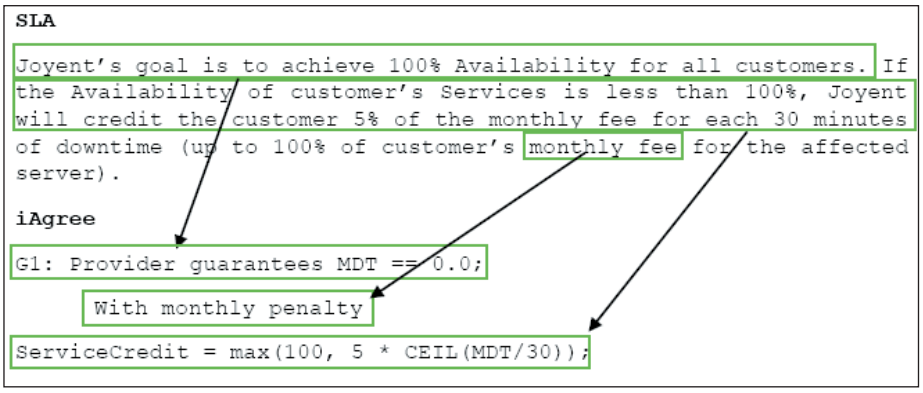

Figura 4. Modelado del ANS de Rackspace a iAgree.

como condiciones de cualificación. Las penalizaciones son incluidas como Valores de Negocio para el intervalo mensual de pago, usando una expresión matemática acorde a la definición expresada en lenguaje original en los acuerdos originales.

4.2. Automatizar el análisis de la disponibilidad

El análisis de los acuerdos implica extraer la información relevante de estos documentos, para lo que resulta útil describir estos análisis como operaciones que toman un conjunto de valores de entrada y devuelven el resultado del análisis [3].

Con los ANSs provistos en las secciones previas y el modelado en iAgree, analizamos como definir las operaciones que respondan a las operaciones propuestas. A pesar de que se han analizado para los proveedores presentados en el artículo, las operaciones se describen de forma genérica para cualquier ANS de un proveedor de la nube. Como hemos visto, el reto en la automatización de dichas operaciones es que al tener las definiciones de no disponibilidad semánticas muy distintas, no es posible proveer soluciones genéricas, sino que es necesario adaptar la métrica de no disponibilidad al proveedor concreto. Así pues, las operaciones se definen de manera abstracta mediante los parámetros de entrada y salida.

En todos los casos, el análisis de los ANSs de almacenamiento depende de la tasa de fallos permitida (referida anteriormente como UFA).

4.2.1. Operación de máximo fallo sin penalización

Considerando el escenario propuesto y la disponibilidad descrita según los términos de garantía, la operación de análisis correspondiente al máximo tiempo de no disponibilidad sin penalización (Q1) devuelve un valor de la unidad temporal de disponibilidad.

Para implementar esta operación, consideramos que la expresión del termino de garantía es equivalente a la expresión de la penalización (es decir, siempre que el SLO se viola, se aplica una penalización y viceversa). De esta forma, responder a esta operación no necesita evaluar la expresión de penalización, sino que se evalúa solo comprobando el cumplimiento de la expresión del SLO.

4.2.2. Operación de penalizaciones aplicadas

Los resultados de esta operación se obtienen considerando la propiedad relacionada con la no disponibilidad del servicio (MDT en el ANS de Joyent, MDRT en Rackspace o MUP en Amazon S3 y Google Cloud Storage).

Dado un valor temporal como parámetro de entrada, la solución de la restricción devolverá una expresión en términos de la penalización. A diferencia de la operación anterior, la solución a esta operación depende de si la garantía de disponibilidad se expresa sobre el tiempo acumulado o no acumulado (por ej.: Joyent ofrece una garantía sobre cualquier tiempo de no disponibilidad, pero Rackspace solo ofrece garantías sobre los periodos que excedan de los 60 minutos de interrupción del servicio). Así, considerado el caso más sencillo, es decir, donde todas las peticiones son erróneas, la solución en ambos casos es:

- Garantía sobre el tiempo acumulado: La solución se calcula sobre la suma de todos los tiempos de no disponibilidad.

- Garantía sobre periodos no acumulados: La solución es la suma de las penalizaciones de cada periodo de no disponibilidad individual.

4.2.3. Operación de mínimo tiempo con máxima penalización

Considerando el escenario propuesto y con los términos de garantías de la disponibilidad dados, la respuesta esperada a la cuestión Q3 es el valor mínimo en la unidad temporal cuando la máxima penalización aplicable se ha alcanzado.

De nuevo, esta operación recibe como parámetros los términos de garantía. Y para su cálculo se minimiza la métrica de disponibilidad para la máxima compensación contemplada en los términos de garantía. Esta función de optimización implica en los dos casos utilizar el peor escenario, es decir, un periodo de no disponibilidad continuado en el caso del servicio de computación, y todas las peticiones erróneas, en el caso del servicio de almacenamiento.

5. Conclusiones y trabajo futuro

La contribución de este artículo se centra en el análisis de la disponibilidad sobre servicios de infraestructura. En primer lugar, un análisis de los ANS de distintos proveedores IaaS nos lleva a concluir que las garantías de disponibilidad no se expresan con una semántica homogénea en los diferentes proveedores de infraestructura. Es más, la distancia semántica es aún mayor si comparamos proveedores de computación y de almacenamiento ya que usan enfoques diferentes para medir la disponibilidad.

Por otro lado, los términos de garantía relacionados con la disponibilidad normalmente expresan penalizaciones. Estas penalizaciones reflejan los objetivos de los proveedores, por lo que es recomendable extender los criterios de validación de los ANSs, así como los analizadores y compiladores para detectar errores relacionados.

En segundo lugar, hemos estudiado los ANS en relación a tres preguntas usuales de usuarios acerca de la disponibilidad. De este análisis se puede concluir que la respuesta a estas preguntas no resulta trivial en la mayoría de los casos. De hecho, para todos los casos, obtener esta respuesta resulta tedioso y fácilmente puede llevar a error por lo que automatizarlas resultaría muy útil desde un punto de vista práctico. En este artículo hemos dado un primer paso en esa dirección, expresando estas cuestiones como operaciones de análisis sobre ANS con el fin de simplificar el diseño de soluciones de infraestructura y acelera el desarrollo de pruebas de concepto. El siguiente reto es intentar ofrecer un mecanismo genérico para dar respuesta a estas preguntas a partir de una definición declarativa de la disponibilidad.

Por último, este trabajo se puede extender para definir criterios de análisis en proveedores donde la disponibilidad tenga un alcance más amplio y mayor complejidad o el dominio sea diferente, tales como proveedores de Plataforma como Servicios o Software como Servicio. Más aún, las operaciones de análisis sobre la disponibilidad se diseñan para comprobar las garantías en el diseño de servicios y la fase de planificación, pero no se ha analizado como estas operaciones se pueden aplicar a las etapas de ejecución y monitorización.

Referencias

[1] G. Copil, D. Moldovan, H.L. Truong, S. Dustdar. Sybl: An extensible language for controlling elasticity in cloud applications. *13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing* (CCGRID), pp. 112-119 (2013). <http://hydra.infosys.tuwien.ac.at/research/viecom/papers/SYBL_ccgrid2013.pdf>.

[2] Open Grid Forum. *Web Services Agreement Specification*. <https://www.ogf.org/documents/GFD.107.pdf>.

[3] H. Ludwig. W*s-agreement concepts and use agreement-based service-oriented architectures*. Technical Report (2006). <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.476&rep=rep1&type=pdf>.

[4] C. Müller. *On the Automated Analysis of WS-Agreement Documents. Applications to the Processes of Creating and Monitoring Agreements*. International dissertation, Universidad de Sevilla (2013).

[5] O. Rana, M. Warnier, T. Quillinan, F. Brazier, D. Cojocarasu. Managing violations in service level agreements. *Grid Middleware and Services*, pp. 349-358. Springer US (2008), <http://dx.doi.org/10.1007/978-0-387-78446-5_23>.

Notas

[1] <http://www.rackspace.com/es/information/legal/cloud/sla#cloud_sla6>.
[2] <https://www.joyent.com/company/policies/cloud-hosting-service-level-agreement>.
[3] <http://aws.amazon.com/es/s3/sla/>.
[4] <https://cloud.google.com/storage/sla>.
[5] <http://www.isa.us.es/IDEAS>.