

Instituto Tecnológico de Costa Rica

Escuela de Ingeniería en Computación

Programa de Maestría en Computación

Modelado y simulación de funciones en la nube en plataformas *Function-as-a-Service*

Propuesta de Tesis sometida a consideración del Departamento de Computación, para optar por el grado de Magíster Scientiae en Computación, con énfasis en Ciencias de la Computación

Estudiante

Carlos Martín Flores González

Profesor Asesor

Ignacio Trejos Zelaya

Noviembre, 2018

Índice

1. Introducción	1
2. Antecedentes	4
2.1. Marco conceptual	4
2.1.1. Ingeniería de rendimiento de software (<i>Software Performance Engineering</i>)	4
Ingeniería de rendimiento basada en mediciones	5
Ingeniería de rendimiento por medio de modelado	6
Modelado de Rendimiento	7
2.1.2. Modelado y simulación de rendimiento basado en componentes	8
Rendimiento de componentes de software	9
Factores que influyen en el rendimiento de un componente	10

Enfoques de ingeniería de rendimiento para software basado en componentes propuestos	11
Métodos de evaluación de rendimiento	12
Enfoques principales	12
Enfoques Suplementarios	14
Modelado de Arquitecturas de Software con <i>Palladio Component Model</i>	15
Modelado de Arquitecturas de Software con <i>Descartes Modeling Language</i>	17
2.1.3. Computación en nube	19
Modelos de entrega de servicio	19
Software como Servicio (SaaS)	20
Plataforma como Servicio (PaaS)	21
Infraestructura como Servicio (IaaS)	21
<i>Serverless y Function-as-a-Service</i>	21
Proveedores de servicios de <i>Function-as-a-Service</i>	22
AWS Lambda	22
Google Functions	23
Microsoft Azure Functions	23

IBM Apache OpenWhisk Functions	23
Otros proveedores de servicios FaaS	24
2.2. Trabajos relacionados	25
Ingeniería de rendimiento de software en aplicaciones en la nube	25
<i>Serverless y Function-as-a-Service</i>	26
Posicionamiento de la investigación con respecto de la literatura consultada	28
3. Definición del problema	29
4. Justificación	30
4.1. Innovación	30
4.2. Impacto	31
4.3. Profundidad	33
5. Hipótesis	34
6. Objetivos	36
6.1. Objetivo general	36
6.2. Objetivos específicos	36

7. Alcance	38
8. Entregables	40
8.1. Revisión de literatura	40
8.2. Implementación de caso de uso de función en la nube	41
8.3. Pruebas sobre el caso de uso	41
8.3.1. Diseño experimental	41
8.3.2. Pruebas de carga	41
8.4. Modelado y análisis del rendimiento de la función	42
8.5. Modelo de rendimiento de la función en la nube	42
8.5.1. Simulaciones sobre el modelo propuesto	42
8.6. Guía metodológica	42
9. Metodología	43
9.1. Etapa 1	43
9.1.1. Selección de caso de uso	43
9.1.2. <i>Manejador de imágenes</i>	45
9.1.3. Manejador de imágenes para SPE	46
¿Por qué este caso de uso se considera relevante? . .	48
9.1.4. Experimentos propuestos	49

9.2. Etapa 2	49
9.2.1. Modelo de rendimiento a partir de la función	49
9.2.2. Enfoque de trabajo propuesto	51
9.3. Etapa 3: ejecutar simulaciones sobre el modelo obtenido	52
10. Cronograma de actividades	54
Bibliografía	57

Índice de figuras

2.1. Factores que influyen en el rendimiento de un componente	10
2.2. Instancia de un modelo PCM	16
2.3. Relación de los diferentes modelos de una instancia DML y el sistema	18
2.4. Niveles de servicio presentes en computación en la nube	20
9.1. Arquitectura del manejador de imágenes	44
9.2. Arquitectura del manejador de imágenes propuesto para el estudio	46
9.3. Carga de trabajo sugerida para el manejador de imágenes	47
9.4. Herramientas utilizadas para ingeniería de rendimiento declarativo	50
9.5. Enfoque de trabajo propuesto	53

Git: (HEAD -> v1.1.0-actualizaciones-propuesta, tag: v1.0.0, origin/revision, origin/master,
revision, master)
Branch: v1.1.0-actualizaciones-propuesta
Tag:
Release:
Commit: e1c8190
Date: 2018-12-15 09:02:47 -0600
Author: Martin Flores
Email: martin.flores@bodybuilding.com
Committer: Martin Flores
Committer email: martin.flores@bodybuilding.com

Capítulo 1

Introducción

Los servicios de funciones en la nube (*Function-as-a-Service, FaaS*) representan una nueva tendencia de la computación en la nube en donde se permite a los desarrolladores instalar código, en forma de función, en una plataforma de servicios en la nube y en donde la infraestructura de la plataforma es responsable de la ejecución, el aprovisionamiento de recursos, monitoreo y el escalamiento automático del entorno de ejecución. El uso de recursos generalmente se mide con una precisión de milisegundos y la facturación es por usualmente 100 ms de tiempo de CPU utilizado.

En este contexto, el “código en forma de función” es un código que es pequeño, sin estado, que trabaja bajo demanda y que tiene una sola responsabilidad funcional. Debido a que el desarrollador no necesita preocuparse de los aspectos operacionales de la instalación o el mantenimiento del código, la industria empezó a describir este código como uno que no necesitaba de un servidor para su ejecución, o al menos de una instalación de servidor como las utilizadas en esquemas tradicionales de desarrollo, y acuñó el término *serverless* (sin servidor) para referirse a ello.

Serverless se utiliza entonces para describir un modelo de programación y una arquitectura en donde fragmentos de código son ejecutados en la nube sin ningún control sobre los recursos de cómputo en donde el código se ejecuta. Esto de ninguna manera es una indicación de que no hay servidores, sino simplemente que el desarrollador delega la mayoría de aspectos operacionales al proveedor de servicios en la nube. A la versión de *serverless* que utiliza explícitamente funciones como unidad de instalación se le conoce como *Function-as-a-Service*[1].

Aunque el modelo FaaS brinda nuevas oportunidades, también introduce nuevos retos. Uno de ellos tiene que ver con el rendimiento de la función, puesto que en este modelo solamente se conoce una parte de la historia, la del código, pero se omiten los detalles de la infraestructura que lo ejecuta. La información de esta infraestructura, su configuración y capacidades es relevante para arquitectos y diseñadores de software para lograr estimar el comportamiento de una función en plataformas FaaS.

El problema de la estimación del rendimiento de aplicaciones en la nube, como lo son las que se ejecutan en plataformas FaaS y arquitecturas basadas en microservicios, es uno de los problemas que está recibiendo mayor atención especialmente dentro de la comunidad de investigación en ingeniería de rendimiento de software. Se argumenta que a pesar de la importancia de contar con niveles altos de rendimiento, todavía hay una falta de enfoques de ingeniería de rendimiento que consideren de forma explícita las particularidades de los microservicios[2].

Si bien, para FaaS, existen plataformas *open source* por medio de las cuales se pueden obtener los detalles de la infraestructura y de esta manera lograr un mejor entendimiento acerca del rendimiento esperado, estas plataformas cuen-

tan con arquitecturas grandes y complejas, lo cual hace que generar estimación se convierta en una tarea sumamente retadora.

En este trabajo se plantea explorar la aplicación de modelado de rendimiento de software basado en componentes para funciones que se ejecutan en ambientes FaaS. Para esto se propone utilizar una función de referencia y, a partir de esta, generar cargas de trabajo para recolectar datos de la bitácora(*logs*) de ejecución y extraer un modelo a partir de ellos. Una vez que se cuente con un modelo, se procederá con su análisis y simulación a fin de evaluar si el modelo generado logra explicar el comportamiento de la función bajo las cargas de trabajo utilizadas.

Esta propuesta está organizada de la siguiente manera: en el capítulo 2 se presenta un marco conceptual sobre ingeniería de rendimiento de software y trabajos de investigación relacionados con ingeniería de rendimiento de software en aplicaciones en la nube, microservicios y *serverless*. En el capítulo 3 se define el problema a resolver. En el capítulo 4 se proporciona una justificación del proyecto desde las perspectivas de innovación, impacto y profundidad. El objetivo general y los objetivos específicos se plantean en el capítulo 6. El alcance del proyecto se define en el capítulo 7. Los entregables que se generarán a partir de esta propuesta se listan en el capítulo 8. La metodología de trabajo se indica en el capítulo 9. La propuesta concluye en el capítulo 10, donde se presenta el cronograma de actividades.

Capítulo 2

Antecedentes

2.1. Marco conceptual

2.1.1. Ingeniería de rendimiento de software (*Software Performance Engineering*)

Una descripción comúnmente utilizada para definir ingeniería de rendimiento de software (*Software Performance Engineering* - SPE) es la brindada en Woodside et al. [3]: “*Ingeniería de rendimiento del software representa toda la colección de actividades de ingeniería de software y análisis relacionados utilizados a través del ciclo de desarrollo de software que están dirigidos a cumplir con los requerimientos de rendimiento*”.

De acuerdo con estos mismos autores, los enfoques para ingeniería de rendimiento puede ser divididos en dos categorías: basadas en mediciones y basadas en modelos. La primera es la más común y utiliza pruebas, diagnóstico y ajustes una vez que existe un sistema en ejecución que se puede medir; es por esto

que solamente puede ser utilizada conforme se va acercando el final del ciclo de desarrollo de software. En contraste con el enfoque basado en mediciones, el enfoque basado en modelos se centra en las etapas iniciales del desarrollo. Como el nombre lo indica, en este enfoque los modelos son clave para hacer predicciones cuantitativas de cuán bien una arquitectura puede cumplir sus expectativas de rendimiento.

Se han propuesto otras clasificaciones de enfoques para SPE pero, con respecto de la evaluación de sistemas basados en componentes, en [4] se deja la clasificación a un lado debido a que se argumenta que la mayoría de enfoques de modelaje toman alguna medición como entrada y a la mayoría de los métodos de medición los acompaña algún modelo.

Ingeniería de rendimiento basada en mediciones

Los enfoques basados en mediciones son los que prevalecen en la industria[5] y son típicamente utilizados para verificación(¿cumple el sistema con sus requerimientos de rendimiento?) o para localizar y reparar *hot-spots* (cuáles son las partes que tienen peor rendimiento en el sistema). La medición de rendimiento se remonta al inicio de la era de la computación, lo que ha generado una amplia gama de herramientas como generadores de carga y monitores para crear cargas de trabajo ficticias y llevar a cabo la medición de un sistema respectivamente.

Las pruebas de rendimiento aplican técnicas basadas en medición y usualmente esto es hecho luego de las pruebas funcionales o de carga. Las pruebas de carga validan el funcionamiento de un sistema bajo cargas de trabajo pesadas, mientras que las pruebas de rendimiento son usadas para obtener datos

cuantitativos de características de rendimiento, como tiempos de respuesta, *throughput* y utilización de hardware para una configuración de un sistema bajo una carga de trabajo definida.

Ingeniería de rendimiento por medio de modelado

La importancia del modelado del rendimiento está motivada por el riesgo de que se presenten problemas graves de rendimiento y la creciente complejidad de sistemas modernos, lo que hace difícil abordar los problemas de rendimiento en el nivel de código[6]. Cambios considerables en el diseño o en las arquitecturas pueden ser requeridos para mitigar los problemas de rendimiento. Por esta razón, la comunidad de investigación de modelado de rendimiento intenta luchar contra el enfoque de “arreglar las cosas luego” durante el proceso de desarrollo. Con la aplicación de un modelo del rendimiento de software se busca encontrar problemas de rendimiento y alternativas de diseño de manera temprana en el ciclo de desarrollo, evitando así el costo y la complejidad de un rediseño o cambios en los requerimientos.

Las herramientas de modelado de rendimiento ayudan a predecir la conducta del sistema antes que este sea construido, o bien, evaluar el resultado de un cambio antes de su implementación. El modelado del rendimiento puede ser usado como una herramienta de alerta temprana durante todo el ciclo de desarrollo con mayor precisión y modelos cada vez más detallados a lo largo del proceso. Al iniciar el desarrollo, un modelo no puede ser validado contra un sistema real, por esto el modelo representa el conocimiento incierto del diseñador. Como consecuencia de esto el modelo incluye suposiciones que no necesariamente se van a dar en el sistema real, pero que van a ser útiles para obtener una abstracción del comportamiento del sistema. En estas fases inicia-

les, la validación se obtiene mediante el uso del modelo, y existe el riesgo de conclusiones erróneas debido a su precisión limitada. Luego, el modelo puede ser validado contra mediciones en el sistema real (o parte de este) o prototipos de este, lo cual hace que la precisión del modelo se incremente.

En [7] se sugiere que los métodos actuales deben superar diversos retos antes de poder aplicarlos en sistemas existentes que enfrentan cambios en su arquitectura o requerimientos. Primero, debe quedar claro cómo se obtienen los valores para los parámetros del modelo y cómo se pueden validar los supuestos. Estimaciones basadas en la experiencia para estos parámetros no son suficientes, por lo que es necesario hacer mediciones en el sistema existente para formular predicciones precisas. Segundo, la caracterización de la carga del sistema en un entorno de producción es problemática debido a los recursos compartidos (bases de datos, hardware). Tercero, deben desarrollarse métodos para capturar parámetros del modelo dependientes de la carga. Por ejemplo un incremento en el tamaño de la base de datos probablemente incrementará las necesidades de procesador, memoria y disco en el servidor.

Técnicas comunes de modelado incluyen redes de colas y también extensiones de estas como redes de colas en capas y varios tipos de redes de Petri, y álgebras de procesos estocástica.

Modelado de Rendimiento

En SPE, la creación y evaluación de modelos de rendimiento es un concepto clave para evaluar cuantitativamente el rendimiento del diseño de un sistema y predecir el rendimiento de diversas alternativas de diseño. Un modelo de rendimiento captura el comportamiento relevante de diversos elementos (y sus

interacciones) para identificar el efecto de cambios en la configuración o en la carga de trabajo sobre el rendimiento de un sistema. Un buen modelo permite predecir los efectos de tales cambios sin necesidad de implementar y ejecutar un sistema en un ambiente de producción, que podrían ser no solamente tareas costosas sino también un desperdicio en caso que el hardware con el que se cuenta demuestre ser insuficiente para soportar la intensidad de la carga de trabajo.[8]

La forma del modelo de rendimiento puede comprender desde funciones matemáticas a formalismos de modelado estructural y modelos de simulación. Estos modelos varían en sus características clave; por ejemplo, los supuestos de modelado de los formalismos, el esfuerzo de modelado requerido y el nivel de abstracción.

En cuanto a técnicas de simulación, a pesar que estas permiten un estudio más detallado de los sistemas que modelos analíticos, la construcción de un modelo de simulación requiere de conocimiento detallado tanto de desarrollo de software como de estadística[5]. Los modelos de simulación también requieren usualmente de mayor tiempo de desarrollo que los modelos analíticos. En [3] se menciona que “la construcción de un modelo de simulación es cara, algunas veces comparable con el desarrollo de un sistema, y, los modelos de simulación detallados puede tardar casi tanto en ejecutarse como el sistema”.

2.1.2. Modelado y simulación de rendimiento basado en componentes

En este enfoque, modelos clásicos de rendimiento tales como redes de colas, redes de Petri estocásticas o álgebras de procesos estocásticas son aplicadas

para modelar y analizar el rendimiento de software basado en componentes. La ingeniería de software basada en componentes se considera un sucesor del desarrollo de software orientado a objetos, en esta los componentes de software son unidades de composición a las que se le define una especificación y una interfaz. A partir de estas, los arquitectos de software construyen e integran componentes de software entre sí, creando de esta manera sistemas más complejos.[4]

El reto en los modelos de rendimiento para componentes, es que el rendimiento de un componente de software en un sistema en ejecución depende del contexto en que está instalado y su perfil de uso, el cual es usualmente desconocido por el desarrollador del componente que creó el modelo de un componente individual.

Rendimiento de componentes de software

Szyperski[9] define un componente de software como: *“Un componente es una unidad de composición en aplicaciones de software, que posee un conjunto de interfaces y un conjunto de requerimientos (una especificación), y que ha de poder ser desarrollado, adquirido, incorporado al sistema y compuesto con otros componentes de forma independiente, en tiempo y espacio”*.

Los componentes de software presentan los principios de ocultación de la información y la separación de responsabilidades. Fomentan la reutilización y preparan el sistema para que se puedan efectuar cambios de partes individuales. Permiten además una división de trabajo entre los desarrolladores de componentes y los arquitectos de software, lo que reduce la complejidad de la tarea de desarrollo. Los componentes de caja negra (*black-box*) solamente re-

velan sus interfaces a los clientes, mientras que los componentes de caja blanca (*white-box*) permiten ver y modificar el código fuente de la implementación del componente como tal. Los componentes compuestos agrupan varios componentes en unidades más grandes.

Factores que influyen en el rendimiento de un componente Especificar el rendimiento de componentes reutilizables es difícil porque esto no solamente va a depender de la implementación del componente, sino también del contexto en donde este se encuentre instalado. Los factores que influyen en el rendimiento de los componentes son:

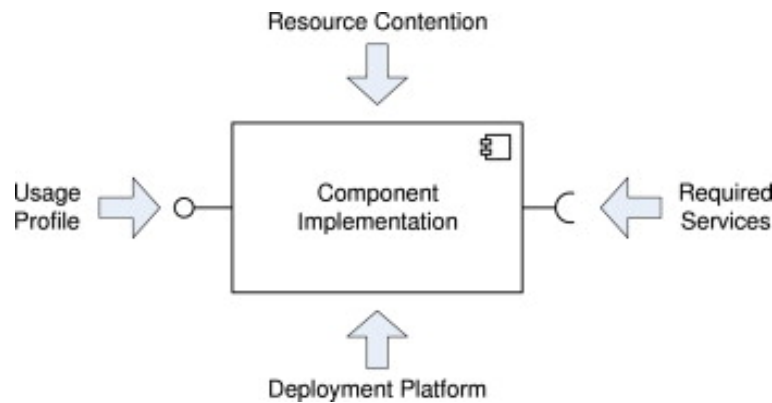


Figura 2.1: Factores que influyen en el rendimiento de un componente. Tomado de [4]

- *Implementación del componente:* los desarrolladores pueden implementar la funcionalidad especificada en una interfaz de diferentes formas. Dos componentes pueden proporcionar el mismo servicio pero presentar tiempos de ejecución diferentes cuando se ejecutan con los mismos recursos y con las mismas entradas.
- *Servicios requeridos:* cuando el componente *A* invoca el servicio del componente *B*, el tiempo de ejecución de *B* suma al tiempo de ejecución de *A*. Por lo tanto, el tiempo total de ejecución de un componente depende del tiempo de ejecución de los otros componentes/servicios que necesita.

- *Plataforma de instalación:* los arquitectos de software instalan un componente de software en diferentes plataformas. Una plataforma de instalación puede incluir varias capas de software (como por ejemplo contenedores de componentes o máquinas virtuales) y hardware (procesador, dispositivos de almacenamiento y red, etc)
- *Perfil de uso:* clientes pueden invocar los servicios del componente con diferentes parámetros de entrada. El tiempo de ejecución de un servicio puede cambiar dependiendo de los valores de los parámetros de entrada.
- *Contención de recursos:* un componente de software típicamente no se ejecuta como un solo proceso aislado en una plataforma determinada. Los tiempos de espera inducidos para acceder a recursos limitados se suman al tiempo de ejecución de un componente de software.

Enfoques de ingeniería de rendimiento para software basado en componentes propuestos

La encuesta llevada a cabo en [4] proporciona una clasificación de enfoques de medición y predicción de rendimiento para sistemas de software basados en componentes. Otra clasificación es la que se expone en [10], donde se presenta una revisión de métodos de predicción de rendimiento basado en modelos para sistemas en general, pero no analiza los requerimientos propios para sistemas basados en componentes.

De acuerdo con [4] durante los últimos diez años, los investigadores han propuesto muchos enfoques para evaluar el rendimiento (tiempos de respuesta, *throughput*, utilización de recursos) de sistemas de software basados en componentes. Estos enfoques abarcan tanto predicción como medición del rendi-

miento. Los primeros analizan el rendimiento esperado de un diseño de software basado en componentes para evitar problemas de rendimiento en la implementación del sistema, lo que podría llevar a costos substanciales para rediseñar la arquitectura. Los otros analizan el rendimiento observable de sistemas de software basados en componentes implementados para entender sus propiedades, determinar su capacidad máxima, identificar componentes críticos y para resolver cuellos de botella.

Métodos de evaluación de rendimiento Los enfoques se reunieron en dos grandes grupos: enfoques principales que proporcionan procesos de evaluación de rendimiento completo y enfoques suplementarios que se centran en aspectos específicos como medición de componentes individuales o modelaje de las propiedades de rendimiento de los conectores de un componente.

Enfoques principales

- **Enfoques de predicción basados en UML:** los enfoques en este grupo se enfocan en la predicción de rendimiento en tiempo de diseño para sistemas de software basado en componentes modelados con el Lenguaje de Modelado Unificado (UML por sus siglas en inglés). UML 2.0 tiene la noción de componente de software como una clase extendida. UML permite modelar el comportamiento de un componente con diagramas de secuencia, actividad y colaboración. La asignación de los componentes puede ser descrita mediante diagramas de despliegue(*deployment*).
- CB-SPE - *Component-Based Software Performance Engineering*
- **Enfoques de predicción basados en Meta-Modelos propietarios:** Los en-

foques en este grupo apuntan a las predicciones de rendimiento de tiempo de diseño. En lugar de usar UML como lenguaje de modelado para desarrolladores y arquitectos, estos enfoques tienen meta-modelos propietarios[4].

- CBML - *Component-Based Modeling Language*
 - PECT - *The Prediction Enabled Component Technology*
 - COMQUAD - *Components with Quantitative properties and Adaptivity*
 - KLAPPER
 - ROBOCOP
 - Palladio
- **Enfoques de predicción centrados en *middleware*:** hacen énfasis en la influencia del *middleware* en el rendimiento de un sistema basado en componentes. Por lo tanto miden y modelan el rendimiento de plataformas *middleware* como JavaEE o .Net. Se basan en suponer que la lógica de negocio de los componentes tiene poco impacto en el rendimiento general del sistema y por eso no requieren un modelado detallado.
- NICTA
- **Enfoques basados en especificaciones formales:** estos enfoques siguen teorías fundamentales de especificación de rendimiento y no toman en cuenta marcos de trabajo (*frameworks*) de medición y predicción.
- RESOLVE
 - HAMLET
- **Enfoques de monitoreo para sistemas implementados:** suponen que un sistema basado en componentes ha sido implementado y puede ser pro-

bado. El objetivo es encontrar problemas de rendimiento en un sistema en ejecución, identificar cuellos de botella y adaptar el sistema para que pueda lograr los requerimientos de rendimiento.

- COMPAS
- TESTEJB
- AQUA
- PAD

Enfoques Suplementarios

- **Enfoques de monitoreo para componentes implementados:** El objetivo de los enfoques de medición para implementaciones de componentes de software individuales es derivar especificaciones de rendimiento parametrizadas a través de mediciones múltiples. El objetivo es obtener el perfil de uso, dependencias y la plataforma de implementación a partir de la especificación de rendimiento, de modo que pueda usarse en diferentes contextos.
 - RefCAM
 - COMAERA
 - ByCounter
- **Enfoques de predicción con énfasis en conectores de componentes:** Estos enfoques suponen un lenguaje de descripción de componentes existente y se centran en modelar y medir el impacto en el rendimiento de las conexiones entre componentes. Estas conexiones pueden implementarse con diferentes técnicas *middleware*.

- Verdickt
- Grassi
- Becker
- Happe

Recientemente varios enfoques de predicción basados en meta-modelos propietarios han sido propuestos para la optimización del diseño de arquitecturas, modelado de calidad de servicio y escalabilidad. PerOpteryx[11] es un enfoque de optimización de diseño de arquitecturas que manipula modelos especificados en *Palladio Component Model*[6] y utiliza el algoritmo evolutivo multi-objetivo NSGA-II. Para análisis de rendimiento utiliza redes de colas en capas. *Descartes Modeling Language*[12] es un lenguaje de modelado de arquitecturas para modelar calidad de servicio y aspectos relacionados con la gestión de recursos de los sistemas, las infraestructuras y los servicios de tecnología de información dinámicos modernos. *CloudScale*¹[13] es un enfoque de diseño y evolución de aplicaciones y servicios escalables en la nube. En *CloudScale* se identifica y gradualmente se resuelven problemas de escalabilidad en aplicaciones existentes y también permite el modelado de alternativas de diseño y el análisis del efecto de la escalabilidad en el costo. Cabe mencionar que estos últimos enfoques han sido influenciados en gran medida por el trabajo llevado a cabo en *Palladio Component Model*.

Modelado de Arquitecturas de Software con *Palladio Component Model*

El *Palladio Component Model* es un enfoque de modelaje para arquitecturas de software basados en componentes que permite predicción de rendimien-

¹<http://www.cloudscale-project.eu/>

to basada en modelos. PCM contribuye al proceso de desarrollo de ingeniería basada en componentes y proporciona conceptos de modelaje para describir componentes de software, arquitectura de software, despliegue (*deployment*) de componentes y perfiles de uso de sistemas de software basados en componentes en diferentes submodelos (Figura 2.2).

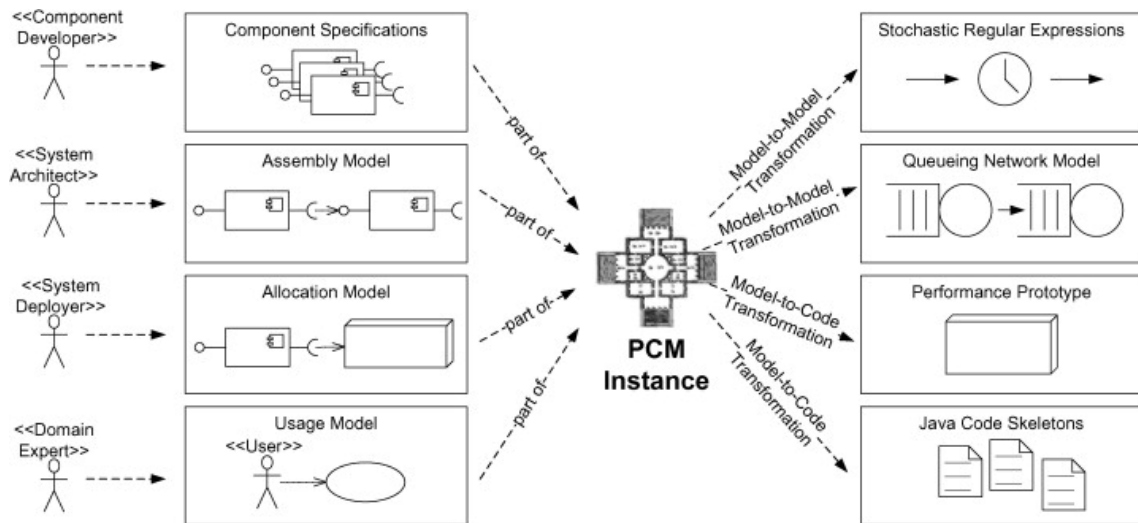


Figura 2.2: Instancia de un modelo PCM. Tomado de [14]

- **Especificaciones de componentes** son descripciones abstractas y paramétricas de los componentes de software. En las especificaciones de software se proporciona una descripción del comportamiento interno del componente así como las demandas de sus recursos en RDSEFFs (*Resource Demanding Service Effect specifications*) utilizando una sintaxis similar a los diagramas de actividad de UML.
- **Un modelo de ensamblaje** (*assembly model*) especifica qué tipo de componentes se utilizan en una instancia de aplicación modelada y si las instancias del componente se replican. Además, define cómo las instancias del componente se conectan, para representar la arquitectura de software.

- El entorno de ejecución y los recursos, así como la instalación (*deployment*) de instancias de componentes para dichos contenedores de recursos se definen en un **modelo de asignación** (*allocation model*).
- El **modelo del uso** especifica la interacción de los usuarios con el sistema utilizando una sintaxis similar al diagrama de actividades de UML, para proporcionar una descripción abstracta de la secuencia y la frecuencia con que los usuarios activan las operaciones disponibles en un sistema.

Un modelo PCM abstrae un sistema de software en el nivel de arquitectura y se anota con consumos de recursos que fueron medidos previamente y otros que son estimados. El modelo puede entonces ser usado en transformaciones de modelo-a-modelo o modelo-a-texto a un modelo de análisis en particular (redes de colas o simulación de código) que puede ser resuelto analíticamente o mediante simulación para obtener resultados sobre el rendimiento y predicciones del sistema modelado. Los resultados del rendimiento y las predicciones pueden ser utilizadas como retroalimentación para evaluar y mejorar el diseño inicial, permitiendo así una evaluación de calidad de los sistemas de software con base en un modelo[8].

Modelado de Arquitecturas de Software con *Descartes Modeling Language*

El *Descartes Modeling Language* (DML) es un lenguaje de modelado de nivel de arquitectura utilizado para describir calidad de servicio (QoS por sus siglas en inglés) y aspectos relacionados con la gestión de recursos de sistemas de información dinámicos, infraestructuras y servicios. DML distingue explícitamente diferentes tipos de modelos que describen el sistema y sus procesos de adaptación desde un punto de vista técnico y lógico. Juntos, estos diferentes

tipos de modelos forman una instancia DML. La idea detrás del uso de estos modelos es separar el conocimiento acerca de la arquitectura del sistema y el comportamiento de su rendimiento (aspectos técnicos), del conocimiento de los procesos de adaptación del sistema (aspectos lógicos)[12].

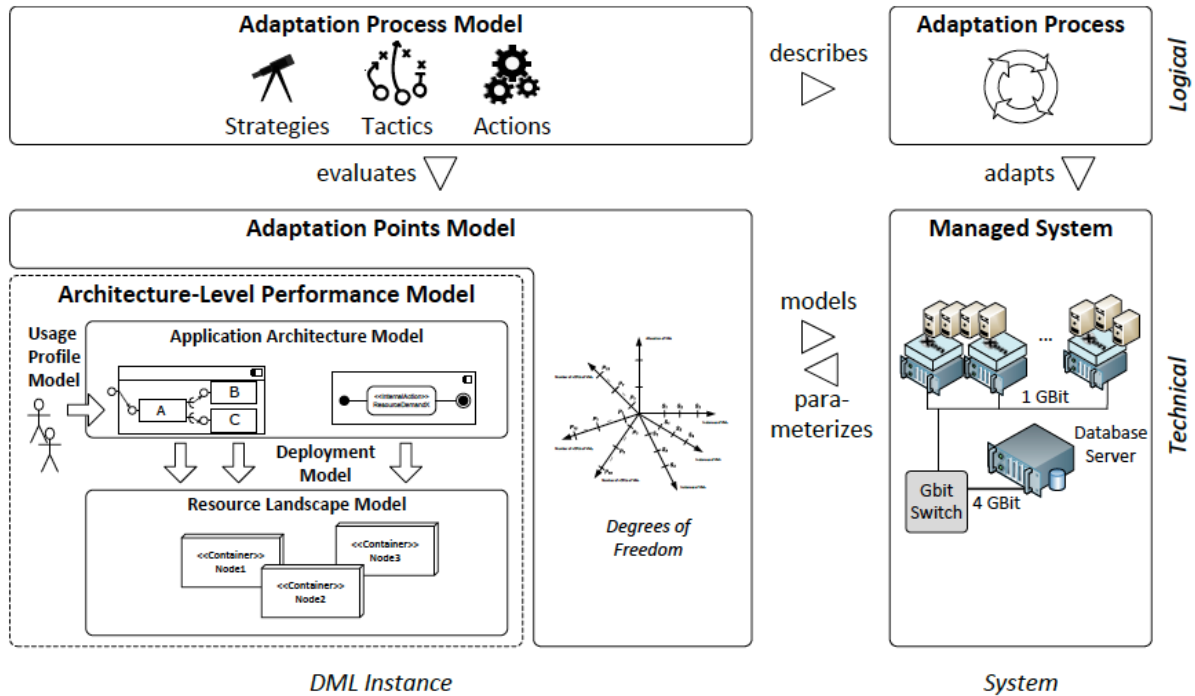


Figura 2.3: Relación de los diferentes modelos de una instancia DML y el sistema. Tomado de [12]

La figura 2.3 muestra la relación de los diferentes modelos que son parte de una instancia DML, el sistema gestionado (*managed system*) y el proceso de adaptación del sistema (*adaptation process*). En la esquina inferior derecha de la figura 2.3, se ve el sistema, el cual es gestionado por un proceso de adaptación, mostrado en la esquina superior derecha de la figura 2.3. En la esquina inferior izquierda se muestran modelos que reflejan los aspectos técnicos del sistema. Esos aspectos son los recursos de hardware y su distribución (*resource landscape model*), los componentes de software y el comportamiento relevante al rendimiento (*application architecture model*), la instalación de componentes de software en el hardware (*deployment model*), el comportamiendo del uso y

las cargas de trabajo de los usuarios del sistema (*usage profile model*) y los grados de libertad del sistema que pueden ser empleados para la adaptación del sistema en ejecución (*adaptation points model*). Por encima de estos modelos (esquina superior izquierda de la figura 2.3) se muestra el modelo de proceso de adaptación, que especifica un proceso de adaptación que describe cómo el sistema se adapta a cambios en su ambiente. El proceso de adaptación aprovecha las técnicas de predicción de rendimiento en línea para razonar sobre posibles estrategias, tácticas y acciones de adaptación.

2.1.3. Computación en nube

Según el NIST[15], el *Cloud Computing*, o la Computación en la Nube, es un modelo que, desde la perspectiva del consumidor, permite el acceso conveniente, por demanda, desde alguna red, a un conjunto de recursos computacionales específicos y configurables (por ejemplo, redes, servidores, almacenamiento, aplicaciones y servicios) que pueden ser rápidamente provisionados y entregados con un esfuerzo mínimo de administración o de interacción con el proveedor del servicio.

Modelos de entrega de servicio

La definición del NIST presenta 3 modelos básicos de servicio. Aunque los proveedores de servicios de Cloud han creado muchas variaciones de ofertas “*as-a-Service*”, estos 3 siguen siendo considerados los fundamentales. En la figura 2.4 se muestra lo que se conoce como infraestructura como servicio (IaaS), plataforma como servicio (PaaS) y software como servicio (SaaS) y las áreas que abarcan en términos de infraestructura, plataforma y software dentro de

un esquema de computación en la nube.

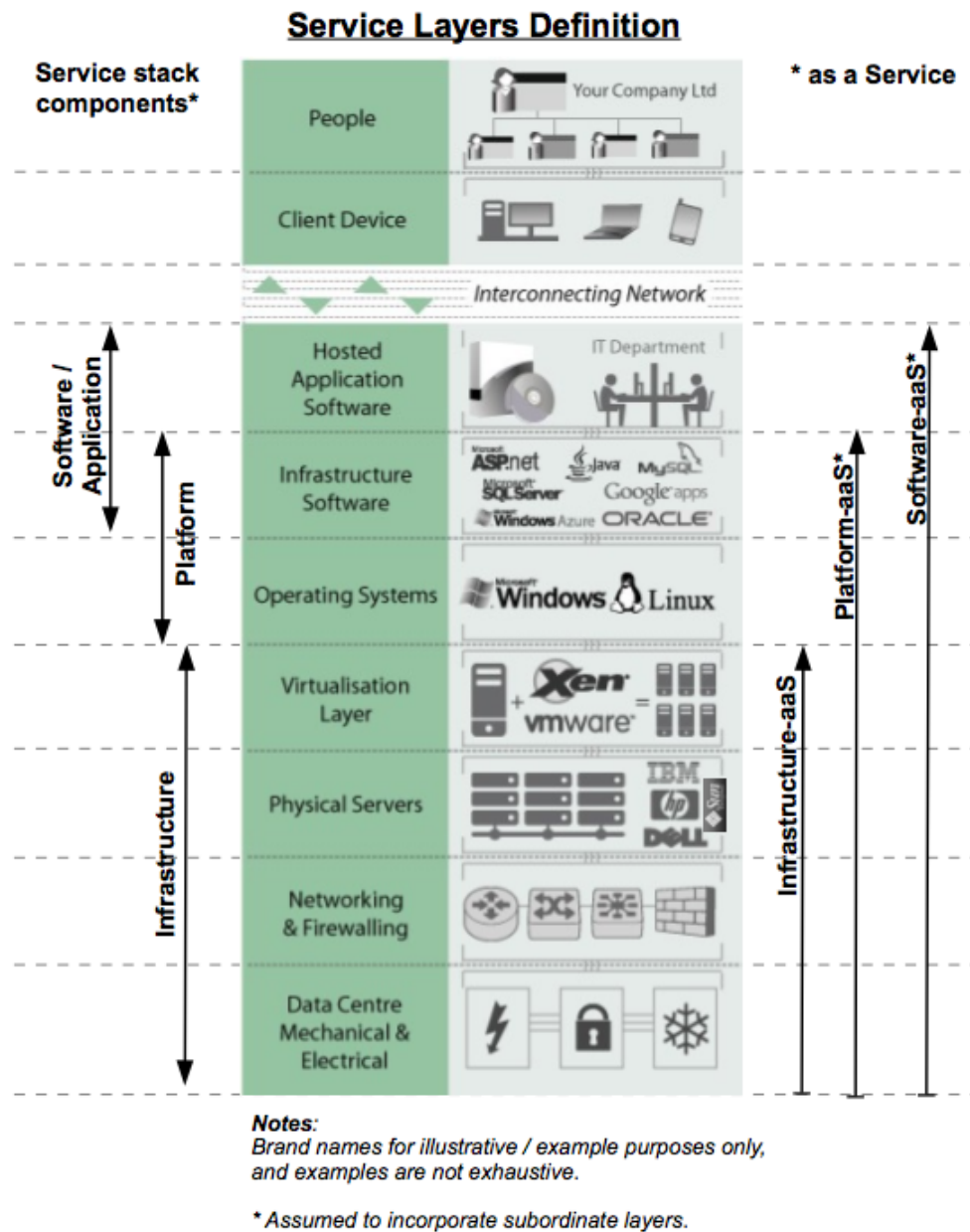


Figura 2.4: Niveles de servicio presentes en computación en la nube. Tomado de [16]

Software como Servicio (SaaS) El cliente usa una aplicación, que se ofrece como un servicio accedido remotamente con protocolos estandarizados, pero no controla el sistema operativo, los servidores de aplicación, el hardware o la infraestructura de red en la que opera. El proveedor instala, administra y mantiene

ne el software. El proveedor no necesariamente es dueño de la infraestructura física donde el software se está ejecutando. En general, se puede considerar un servicio para usuarios finales.

Plataforma como Servicio (PaaS) El cliente usa un sistema hospedado para sus aplicaciones. El cliente controla las aplicaciones que ejecuta en el ambiente (y posiblemente tenga algún control sobre el ambiente de hospedaje), pero no tiene acceso ni controla el sistema operativo, el hardware o la infraestructura de red donde ellas se ejecutan. La plataforma es típicamente un marco de ejecución aplicativo. El proveedor administra la infraestructura de software de la nube para la plataforma. En este caso, el cliente por lo general es un desarrollador de software.

Infraestructura como Servicio (IaaS) El cliente utiliza recursos computacionales fundamentales como poder de procesamiento, almacenamiento, comunicaciones y *middleware*. El cliente puede controlar el sistema operativo, el almacenamiento, implementar aplicaciones y posiblemente algunos componentes de red como cortafuegos (*firewalls*) y balanceadores de carga, pero la infraestructura física de red y de hardware no le pertenece. Los clientes habituales de estos servicios son administradores de sistemas.

Serverless y Function-as-a-Service

FaaS es un área de lo que hoy se conoce como computación sin servidores o *serverless computing*. Amazon.com, uno de los principales propulsores de esta tendencia define *serverless* como una tecnología la cual permite construir y ejecutar aplicaciones y servicios sin necesidad de pensar en los servidores. Las

aplicaciones *serverless* no requieren de aprovisionamiento, escalamiento y administración de ningún servidor. Se puede construir con ella casi cualquier tipo de aplicación o servicio, y todo lo que se requiere para ejecutar y escalar la aplicación con alta disponibilidad es gestionado por el proveedor del servicio[17].

FaaS se refiere a un tipo de aplicación *serverless* en donde la lógica del lado del servidor es escrita por un desarrollador pero, a diferencia de arquitecturas tradicionales, esta se ejecuta en contenedores de cómputo que no mantienen estado, son activados por medio de eventos, son efímeros (pueden durar solo una invocación) y son totalmente administrados por un tercero[18].

Cabe destacar que aunque el término *serverless* ha llegado a ser utilizado para referirse de forma directa a plataformas FaaS, hoy en día las aplicaciones de *serverless computing* abarcan otros tipos de servicios como almacenamiento, mensajería, análisis de datos, seguridad, entre otros.

Proveedores de servicios de *Function-as-a-Service*

Los principales proveedores de servicios de FaaS son: AWS Lambda, Google Functions, Microsoft Azure Functions e IBM Apache OpenWhisk Functions[19].

AWS Lambda (<https://aws.amazon.com/lambda/>) Proporcionado por Amazon Web Services. Soporta varios lenguajes como NodeJS, Java, C# y Python. Rango de asignación de memoria: mínimo 128Mb, máximo 1536Mb. Capacidad de disco (espacio en /tmp) 512Mb. La duración máxima de ejecución por solicitud es de 300 segundos.

El precio se establece en \$0.20 por millón de solicitudes y \$0.00001667 por

Gigabyte por segundo. 1 millón de solicitudes y 400.000 GBs por mes son gratuitos.

Google Functions (<https://cloud.google.com/functions>) Solamente soporta NodeJS. Está limitado a 1000 funciones con una duración máxima de 540 segundos por cuenta. El precio se establece en tres modelos diferentes. El primero es de \$ 0,40 por millón de invocaciones, pero considera que 2 millones de invocaciones son gratuitas. El segundo modelo de precio es de \$0.0000025 por GBs con 400,00 GBs por mes son gratis. El tercer modelo de precios, es de \$0.0000100 GHz por segundo con 200,000 GHz por segundo al mes que son gratuitos.

Microsoft Azure Functions (<https://azure.microsoft.com/en-us/services/functions>) Soporta lenguajes como NodeJS, C#, F#, Python, PHP y Java. Azure está limitado a solo diez ejecuciones simultáneas por una función. No hay límites en el límite máximo de tiempo de ejecución. Utiliza dos modelos de precios diferentes. El primero es de \$0.000016 GBs, con 400,000 GBs por mes que son gratuitos. El segundo modelo es de \$0.20 por millón de ejecuciones, con 1 millón de ejecuciones por mes de forma gratuita.

IBM Apache OpenWhisk Functions (<https://www.ibm.com/cloud/functions>) Soporta lenguajes como NodeJS, Swift, Java, PHP, Go y Python. Se puede utilizar cualquier otro lenguaje de programación siempre y cuando de proporcione un contenedor de Docker para esto. Tiene un costo básico que es de \$0.000017 por segundo de ejecución, por GB de memoria asignada.

Otros proveedores de servicios FaaS Se pueden encontrar otros proveedores y herramientas para utilizar el modelo de funciones en la nube. Entre ellos se encuentran:

- Apache OpenWhisk: <http://openwhisk.incubator.apache.org/>
- Webtask: <https://webtask.io/>
- OpenFaas: <https://github.com/openfaas/faas>
- IronFunctions: <https://github.com/iron-io/functions>
- Kubeless: <http://kubeless.io/>
- Fn Project: <http://fnproject.io/>
- Spotinst Functions: <https://spotinst.com/products/spotinst-functions/>

2.2. Trabajos relacionados

Ingeniería de rendimiento de software en aplicaciones en la nube

El estilo de arquitectura basado en microservicios es uno de los que ha logrado ganar mayor adopción y popularidad dentro de la comunidad de desarrolladores. Los microservicios, son una arquitectura de software que involucra la construcción y entrega de sistemas que se caracterizan por ser servicios pequeños, granulares, independientes y colaborativos [20]. Con respecto de SPE y microservicios se reporta que existen muchos retos en investigación que han sido poco o nada abordados.

En [2] se planean el monitoreo, pruebas y modelado del rendimiento como las tres áreas en donde se carece de investigación y desarrollo de SPE y microservicios. Se argumenta que aún hacen falta enfoques de SPE que tomen en cuenta las particularidades de los microservicios. Aderaldo et al.[21] señala una falta de investigación empírica repetible sobre diseño, desarrollo y evaluación de aplicaciones de microservicios y que esto dificulta la evaluación de este tipo de aplicaciones pues se cuenta con muy pocas aplicaciones y arquitecturas de referencia, así como de cargas de trabajo que contribuyan a caracterizar comportamiento.

El reporte de [22] también proporciona un listado de los retos asociados con microservicios y SPE, haciendo énfasis en actividades de SPE relacionadas con la integración de actividades de desarrollo y de operaciones de puesta en producción y mantenimiento del software. Se argumenta que a pesar del alto nivel de adopción de prácticas de integración continua, entrega continua y DevOps por la comunidad de ingeniería de software, ninguna toma en cuenta aspectos

relacionados con el rendimiento. Otros estudios, como el llevado a cabo en [23], indican que uno de los atributos de calidad que ha recibido mayor atención en la investigación en microservicios es el de la eficiencia del rendimiento pero vista mayoritariamente desde el punto de vista de la escalabilidad y mantenibilidad del código de los microservicios y su instalación.

Serverless y Function-as-a-Service

Los investigadores han empezado a describir y analizar FaaS a través de encuestas y experimentos[1, 24, 25], y también por análisis económicos[26, 27]. Sin embargo se reporta que aún no se conoce mucho acerca de SPE en FaaS. En [2] se menciona que al igual que con microservicios, FaaS también requiere de nuevas estrategias de modelado para capturar el comportamiento del código bajo estas infraestructuras. Los modelos de rendimiento tradicionales basados en la noción de máquinas independientes podría ser inadecuado.

En van Eyk et al.[28] se presenta un informe confeccionado por el *Standard Performance Evaluation Corporation RG Cloud Group*² (SPEC RG Cloud) sobre desafíos asociados a rendimiento en arquitecturas FaaS. Los principales temas son los que tienen que ver con evaluación y comparación de plataformas de FaaS, reducción del *overhead*, políticas de *scheduling*, la relación costo-rendimiento de una función y la predicción del rendimiento. El informe también señala que actualmente muchas de las tareas de evaluación y pruebas para FaaS se vuelven complicadas porque no se cuenta con aplicaciones, arquitecturas ni cargas de trabajo de referencia; este es un trabajo que pretende abordar el SPEC RG Cloud. Con respecto a predicción del rendimiento, se indica que la aplicación de modelos de rendimiento de sistemas de software tradicionales

²<https://research.spec.org/working-groups/rg-cloud.html>

en FaaS trae nuevos retos como la brecha de información (*information gap*) y el rendimiento de una función en particular. La brecha de información significa que el usuario de FaaS no está consciente de los recursos de hardware en los que las funciones son ejecutadas, mientras que, por otro lado, la plataforma de FaaS no tiene información acerca de los detalles de la implementación de la función. Tal y como en las aplicaciones tradicionales, la entrada (tamaño, estructura y contenido) influyen en el rendimiento de una función, el hecho de tener una infraestructura oculta hace necesario encontrar nuevos modelos que logren predecir de forma precisa el rendimiento de una función. Técnicas de modelado desarrolladas para sistemas de software se podrían aprovechar para FaaS, como por ejemplo modelado y simulación de arquitecturas de software basada en componentes.

La aplicación de *serverless computing* es una área activa de desarrollo. En trabajos previos [29, 30] se han estudiado arquitecturas alternativas de *serverless computing* con el fin de explotar aspectos de rendimiento y/o abordar retos técnicos que otras plataformas no han hecho. También se han investigado arquitecturas para recuperación de información[24] y *chatbots*[31] utilizando plataformas *serverless*. Muchas otras aplicaciones se han venido desarrollando en campos como *Machine Learning*, seguridad, Internet de las cosas, procesamiento de voz, sistemas de archivos, etc, y son solo una muestra del potencial de esta tecnología. Pese a este potencial, también se reporta que aún no se sabe mucho acerca de cuáles herramientas y arquitecturas tecnológicas se usan para producir, instalar y ejecutar funciones[32]. En [25] se examinan factores qué pueden influir en el rendimiento de plataformas de *serverless computing*. De acuerdo con esto, se logran identificar cuatro estados de una infraestructura *serverless*: *provider cold*, *VM cold*, *container cold* y *warm*. Además demuestra cómo el rendimiento de los servicios puede llegar a variar hasta 15 veces según

estos estados.

Posicionamiento de la investigación con respecto de la literatura consultada

La investigación que se propone en este documento, pretende dar un aporte al área de ingeniería de rendimiento de software en aplicaciones en la nube. De acuerdo con el material recolectado, se pudo conocer que existe una necesidad por aplicar enfoques de modelado de rendimiento en el desarrollo de sistemas modernos pero que, por otro lado, los esfuerzos que se han llevado a cabo para esto no han logrado ganar popularidad, o bien, no consideran las particularidades de la computación en la nube. Por ejemplo, se reporta que no existen enfoques de modelado de rendimiento para microservicios, el cual es hoy en día, un estilo de arquitectura de software sumamente popular.

Es por esto que se considera que la realización de un estudio exploratorio para determinar los factores que influyen en el rendimiento de una función en la nube, puede brindar nuevo conocimiento sobre cómo aplicar enfoques de ingeniería de rendimiento conocidos a estas aplicaciones y además podría representar un marco de referencia inicial por medio del cual se pueda evaluar la adopción de estas tecnologías *a priori* y fundamentar el planeamiento de capacidades al dimensionar sistemas intensivos en software que vayan a usar microservicios en ambientes en la nube.

Capítulo 3

Definición del problema

De acuerdo con la revisión de la literatura, se carece de modelos de rendimiento que contribuyan a caracterizar el comportamiento de funciones en la nube alojadas en plataformas FaaS bajo distintas cargas de trabajo. Contar con tales modelos permitiría validar si las funciones en la nube pueden cumplir criterios de calidad de servicio especificados.

En las plataformas FaaS en las que se ejecutan funciones en la nube, la infraestructura tecnológica subyacente se oculta por completo de los desarrolladores y diseñadores. El conocimiento de la influencia de esta infraestructura y su configuración es vital para que los arquitectos de software puedan obtener predicciones significativas del comportamiento de una función pues, al omitirse la influencia que esta tiene, puede conducir a la generación de predicciones erróneas con respecto del rendimiento de una función. Una función que reporte tiempos de respuesta sumamente prolongados o bien la utilización de significativas cantidades de recursos puede generar grandes costos económicos y hasta llegar a ser rechazada por la plataforma FaaS.

Capítulo 4

Justificación

4.1. Innovación

Los aspectos más novedosos que se aportarán en esta tesis serán:

1. Proponer un método mediante el cual se pueden obtener estimaciones del rendimiento de una función en la nube
2. Realizar modelado y simulación basados en componentes para caracterizar el rendimiento de funciones en la nube sobre plataformas FaaS
3. Proporcionar, a partir de lo anterior, un modelo del rendimiento de una función en la nube

Con respecto de (1), en [28] se reporta que la predicción del rendimiento es uno de los principales retos de investigación en plataformas FaaS y que no se cuenta con modelos de rendimiento que consideren las características de FaaS. Esto abre la posibilidad de explorar la aplicabilidad de técnicas conocidas de

modelaje de rendimiento en sistemas de software a nuestro dominio de problema.

(2) El modelado y simulación de arquitecturas de software basadas en componentes, representa una alternativa atractiva para abordar este problema, pues es un enfoque en donde hay una comunidad de investigación y desarrollo activa que ha logrado generar diferentes tipos de herramientas para la estimación del rendimiento de sistemas informáticos.

(3) La obtención de un modelo de rendimiento de una función en la nube (3), representaría un aporte relevante para SPE y FaaS porque significaría que existe una forma por medio de la cual evaluar el comportamiento de una función sin que esta esté necesariamente instalada en una plataforma FaaS y además permitiría que cambios futuros que necesite esa función puedan ser modelados *a priori*, simularlos y obtener predicciones del impacto de los cambios.

4.2. Impacto

En distintos reportes sobre el estado de tecnologías *serverless* en la industria [33, 34, 35, 36] se señala que la adopción de este tipo de tecnologías va en franco aumento. En [34] se reporta una tasa de crecimiento del 25 % en su adopción con respecto del 2017. Cloud Foundry Foundation [36] indica que el 2017 “la mayoría” de sus encuestados no estaban usando tecnologías *serverless* mientras que para el 2018, solamente 43 % **no** lo estaba haciendo. En el informe anual del estado de tecnologías en la nube de DZone[33] se notó un incremento del 14 % con respecto del 2017 en el uso de tecnologías *serverless*.

Pese a que los niveles de adopción de esta tecnología van en aumento, estos

mismos reportes señalan inconvenientes tales como:

1. Se tiene que depender de los niveles de servicio de un proveedor
2. Dificultad para monitorear y depurar
3. Preocupación por parte de los desarrolladores por el “cómo funciona” la función en la plataforma FaaS: ¿se estarán asignando los recursos adecuados para mi función en la plataforma FaaS? ¿Cómo y cuándo se hace?
4. Límites de tiempo de espera: dependiendo del tiempo de ejecución una función en la nube podría llegar a ser cancelada por la propia plataforma FaaS

Lo anterior refleja que aún existe una especie de “área gris” alrededor del uso de las tecnologías *serverless* y en particular funciones en la nube. Esto es lo que van Eyk et al.[28] llama brecha de la información: el usuario de FaaS no está consciente de los recursos de hardware en los que las funciones son ejecutadas, mientras que por otro lado, la plataforma de FaaS no tiene información acerca de los detalles de la implementación de la función.

Analizar el rendimiento de una función en la nube y obtener un modelo de este contribuiría a tener un mejor entendimiento de esta tecnología y de cómo esta es gestionada por la plataforma FaaS. Esto permitiría a los arquitectos y diseñadores tener mayor control sobre los cambios en una función para que esta no solamente pueda cumplir con los requerimientos de calidad de servicio sino que también, al ser *serverless* un servicio que se cobra por demanda, colaboraría a reducir gastos, ya que una función que tenga un tiempo de ejecución menor generará menores costos por el uso de la plataforma FaaS.

4.3. Profundidad

Para lograr obtener un modelo de rendimiento de una función en la nube se plantean las siguientes actividades:

- Diseño e implementación de un caso de uso de referencia de una función en la nube
- Obtención un modelo:
 - Realizar pruebas de carga sobre la función en la nube seleccionada
 - Obtener métricas asociadas al rendimiento a partir de bitácoras de ejecución de la función
 - Utilizar las métricas obtenidas para suministrarlas como entrada a una herramienta de extracción de modelos de rendimiento. La herramienta generará como resultado un modelo de rendimiento
- Diseñar experimentos y ejecutar simulaciones sobre el modelo obtenido con el fin de validar las estimaciones o determinar si es necesario calibrar el modelo.

Capítulo 5

Hipótesis

En el modelo FaaS, la infraestructura tecnológica subyacente se oculta por completo de los diseñadores y desarrolladores y, la duración de la ejecución de una función en la nube determina el costo del servicio: entre mayor sea la duración, mayor será el costo y viceversa. Si bien se han realizado estudios [25] para evaluar factores que influyen en el rendimiento de servicios basados en computación *serverless*, el modelado del rendimiento de este tipo de aplicaciones se sigue presentando como un reto de investigación [2]. El modelado del rendimiento ha ganado considerable atención en la comunidad de ingeniería de rendimiento en las pasadas dos décadas, principalmente en sistemas de software basados en componentes [4]. Al finalizar y alcanzar los objetivos de la investigación se podrá contar con un marco de referencia que permitirá:

- Tener con una función en la nube que servirá como prueba de concepto funcional para la evaluación del rendimiento
- Contar un método mediante el cual se pueda analizar el rendimiento de una función en la nube por medio de modelado y simulación.

Una vez alcanzados los objetivos, será posible responder a la pregunta:

¿Es posible predecir el rendimiento de una función en la nube por medio de modelado y simulación basado en componentes?

Capítulo 6

Objetivos

6.1. Objetivo general

Diseñar un método para modelar el rendimiento de funciones en la nube alojadas en plataformas *Function-as-a-Service* por medio del modelado y la simulación basados en componentes, con el fin de evaluar los factores que pueden influir en su comportamiento.

6.2. Objetivos específicos

1. Revisar el estado del arte de trabajos relacionados con enfoques de predicción y medición del rendimiento en sistemas de software como servicio.
2. Sintetizar un caso de uso de una función en la nube considerada como de referencia, con el propósito de analizar su comportamiento.
3. Elaborar, conforme a un diseño experimental, pruebas de rendimiento so-

bre el caso de uso seleccionado a fin de obtener datos base.

4. Analizar los datos experimentales mediante herramientas de extracción de modelos de rendimiento de software.
5. Proponer y validar modelos de rendimiento a partir de los experimentos realizados.
6. Formular una guía metodológica para dar a conocer aspectos de rendimiento en funciones en la nube a partir de la experiencia obtenida.

Capítulo 7

Alcance

El trabajo propuesto comprende:

- El estudio de los factores que pueden influir en el rendimiento de un caso de uso de una función en la nube bajo una plataforma FaaS en particular. Se espera que, a partir de la experiencia generada y los datos recolectados, este trabajo pueda ser replicado con otras funciones en la nube bajo plataformas FaaS alternativas.
- La propuesta de un modelo de rendimiento que permita la simulación de la ejecución de una función en la nube y obtener, a partir de esta, estimaciones del tiempo de respuesta de la ejecución de la función.

El trabajo propuesto **no** comprende:

- El modelado o simulación de toda una plataforma FaaS.
- La realización de un modelo o estudio de los aspectos económicos asociados a la ejecución de una función en la nube.

- Estudiar el rendimiento de funciones en la nube que mantengan alguna relación o dependencias entre sí (orquestración de funciones).
- Experimentar con múltiples plataformas FaaS ni con diversos lenguajes de programación.

Capítulo 8

Entregables

8.1. Revisión de literatura

Alineado con objetivo específico 1

Se pretende identificar los resultados de otros estudios relacionados con el modelado de rendimiento de software, así como retos y oportunidades de investigación que existan en esta área. Las preguntas de investigación inicialmente propuestas son las siguientes:

- PI1** ¿Cuáles enfoques de predicción y medición del rendimiento en sistemas de software basados en componente se han propuesto?
- PI2** ¿Cuáles enfoques de predicción y medición de rendimiento de software se han utilizado para aplicaciones en la nube?
- PI3** ¿Qué retos y oportunidades existen con estos enfoques en la actualidad?
- PI4** ¿Qué herramientas hay disponibles para el modelado de rendimiento de software?

8.2. Implementación de un caso de uso de función en la nube

Alineado con objetivo específico 2

Identificar un caso de uso de que sea considerado como de referencia o de utilidad común en donde las funciones en la nube hayan mostrado ser un buenas candidatas para su solución. Se procederá a implementar este caso de uso e instalarlo en una plataforma FaaS. Esta será la función que servirá como base para futuros análisis.

8.3. Pruebas sobre el caso de uso

Alineado con objetivo específico 3

8.3.1. Diseño experimental

Planeamiento y definición de las variables por observar de las pruebas de carga.

8.3.2. Pruebas de carga

Diseño, ejecución y análisis de los resultados de las pruebas.

8.4. Modelado y análisis del rendimiento de la función

Alineado con objetivo específico 4

Se tomarán los datos generados por las pruebas anteriores para utilizarlas como entrada y, a partir de ellos, realizar una labor de análisis y modelado.

8.5. Modelo de rendimiento de la función en la nube

Alineado con objetivo específico 5

A partir de la función propuesta, pruebas y análisis del rendimiento por medio de herramientas de modelado, se propondrá un modelo que logre caracterizar el comportamiento de la función.

8.5.1. Simulaciones sobre el modelo propuesto

Para validar el modelo propuesto, se ejecutarán simulaciones sobre él y de esta forma determinar si los resultados de las simulaciones se corresponden con los obtenidos en las pruebas de rendimiento de la función.

8.6. Guía metodológica

Alineado con objetivo específico 6

Una vez realizado el estudio, se estructurará una descripción del método utilizado para estimar el rendimiento de una función en la nube.

Capítulo 9

Metodología

9.1. Etapa 1

9.1.1. Selección de caso de uso

Uno de los principales problemas de hacer ingeniería de rendimiento para software en la nube es que no existen aplicaciones de referencia que hayan ganado popularidad o cuyo desarrollo se encuentre activo. A pesar de esto y de su reciente adopción, la industria ha empezado a reconocer casos de uso en donde las aplicaciones *serverless* encajan mejor. Amazon Web Services(AWS)[37] reconoce cinco patrones de uso predominantes en su servicio AWS Lambda:

1. Procesamiento de datos dirigidos por eventos.
2. Aplicaciones Web.
3. Aplicaciones móviles e Internet las cosas (IoT).
4. Ecosistemas de aplicaciones *serverless*.

5. Flujos de trabajo dirigidos por eventos.

Uno de las aplicaciones más comunes en *serverless* es desencadenar acciones luego de que ocurre un evento (1), por ejemplo luego de la modificación de un registro en una base de datos o bien luego de que se publica un mensaje en una cola de mensajería. Esto puede provocar que se active una función Lambda¹ que toma como entrada el evento recién publicado para su posterior procesamiento. Este estilo de caso de uso encaja bien en ambientes híbridos: ambientes en donde tecnologías *serverless* se aprovechan para realizar funciones específicas dentro de una aplicación (o aplicaciones) más grande.

AWS ha publicado una serie de arquitecturas de referencia[38] para su plataforma FaaS, AWS Lambda. Dentro de estas arquitecturas se destaca el caso de uso de un manejador de imágenes (*Image Handler*)[39].

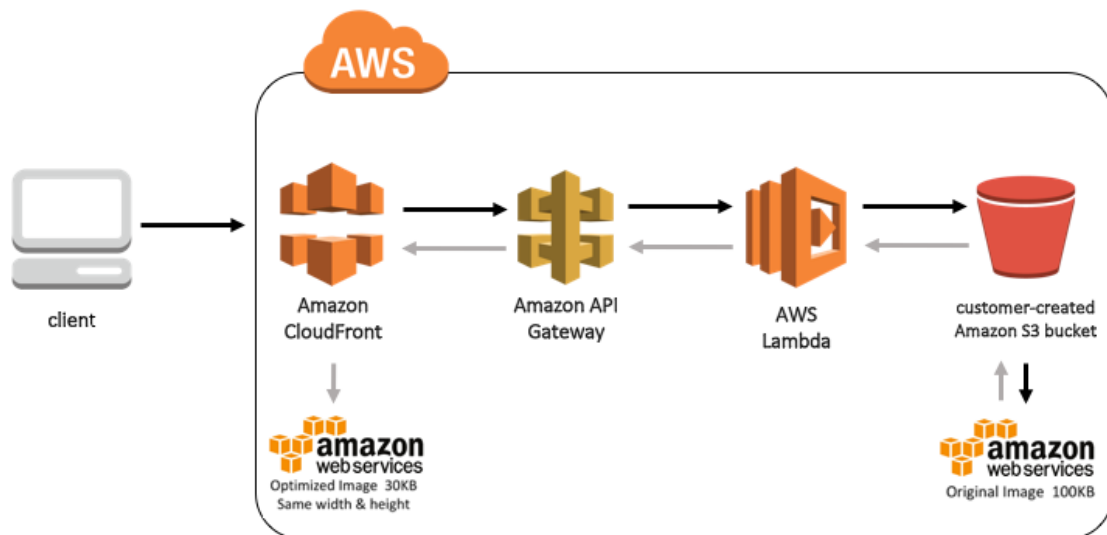


Figura 9.1: Arquitectura del manejador de imágenes. Tomado de [39]

¹En la plataforma AWS Lambda

9.1.2. *Manejador de imágenes*

Sitios Web con imágenes grandes pueden experimentar tiempos de carga prolongados, es por esto que los desarrolladores proporcionan diferentes versiones de cada imagen para que se acomoden a distintos anchos de banda o diseños de página. Para brindar tiempos de respuesta cortos y disminuir el costo de la optimización, manipulación y procesamiento de las imágenes, AWS propone un manejador de imágenes *serverless*, al cual se le pueda delegar tal trabajo como una función Lambda sobre la plataforma FaaS.

A continuación se describe la arquitectura de la figura 9.1:

1. Amazon CloudFront provee una capa de *cache* para reducir el costo del procesamiento de la imagen
2. Amazon API Gateway brinda acceso por medio de HTTP a las funciones Lambda
3. AWS Lambda obtiene la imagen de un repositorio de Amazon Simple Storage Service (Amazon S3) y por medio de la implementación de la función se retorna una versión modificada de la imagen al API Gateway
4. El API Gateway retorna una nueva imagen a CloudFront para su posterior entrega a los usuarios finales

Cabe mencionar que, en este contexto, una versión modificada de una imagen será cualquier imagen que haya presentado algún tipo de alteración con respecto de una imagen original como, por ejemplo, cambios de tamaño, color, metadatos, etc.

9.1.3. Manejador de imágenes para SPE

Para este estudio se proponemos implementar una variación del manejador de imágenes de la sección 9.1.2, que se muestra en la figura 9.2.

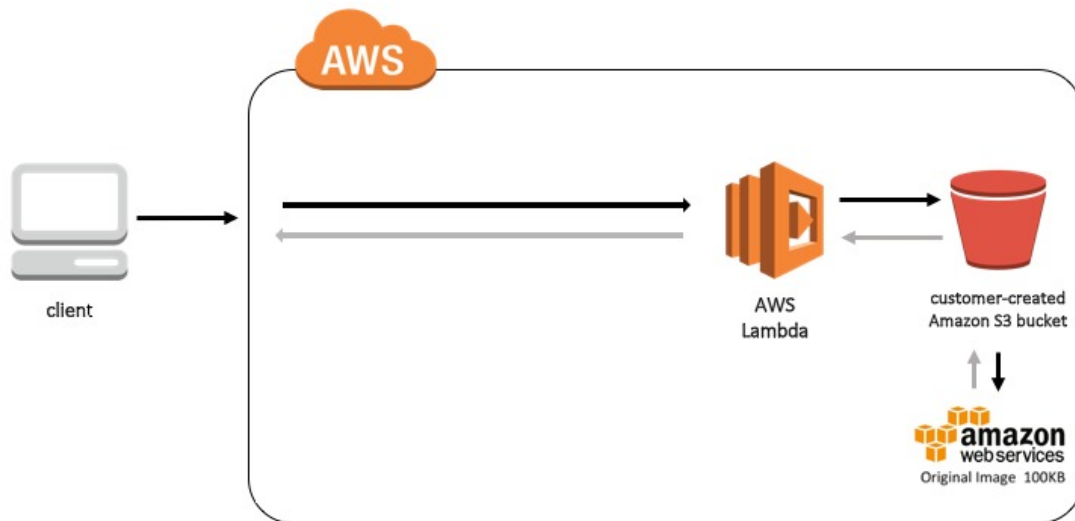


Figura 9.2: Arquitectura del manejador de imágenes propuesto para el estudio.

Se han dejado por fuera intencionalmente el AWS CloudFront y el AWS API Gateway. La razón de esto es porque se pretende ejercitar la función Lambda directamente. Se implementará una función Lambda que entregue a partir de una solicitud de redimensionamiento de una imagen almacenada, otra con dimensiones diferentes producida “al vuelo” como respuesta a la solicitud. Por ejemplo, si la imagen original mide 500 píxeles de ancho y alto, entregar una con dimensiones de 100 píxeles de ancho y alto.

Las actividades involucradas en el proceso de redimensionamientos de imágenes se muestran en la figura 9.3

1. Se envía una solicitud de redimensionamiento de imagen en formato JSON a la función Lambda con los datos acerca de la localización de la imagen y

su nuevo tamaño.

2. La solicitud de redimensionamiento llega a la función Lambda.
3. La función Lambda solicita al servicio de almacenamiento AWS S3 la imagen.
4. AWS S3 entrega a la función Lambda la imagen solicitada.
5. La función Lambda inicia el redimensionamiento de la imagen de acuerdo a los parámetros solicitados.
6. La nueva imagen modificada se entrega al cliente(s).

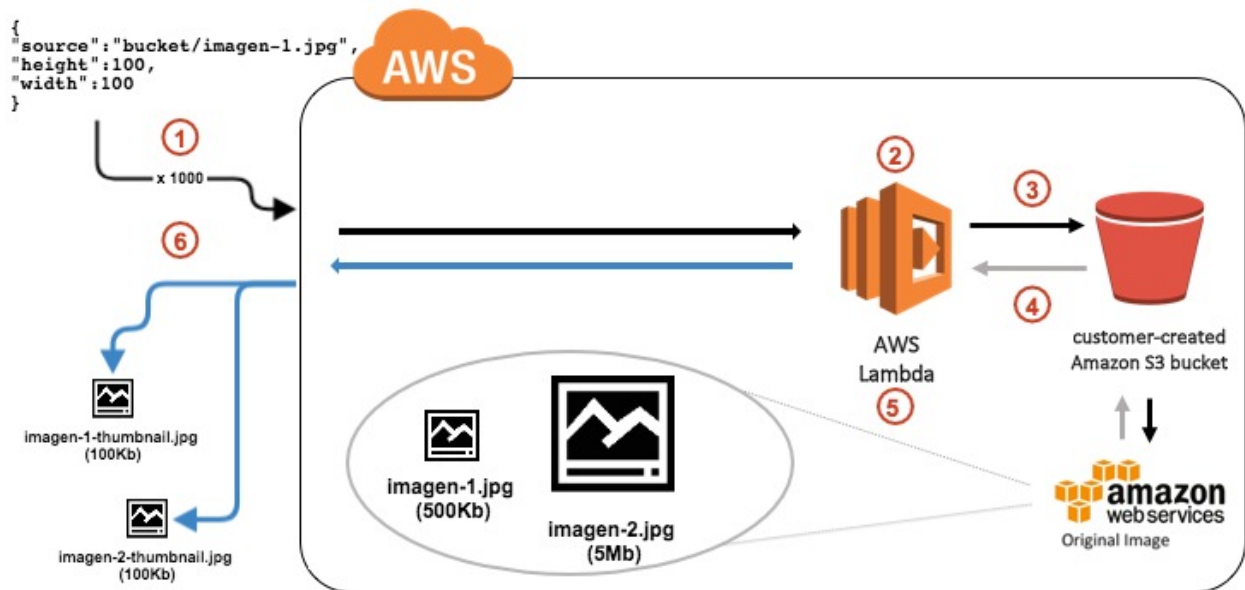


Figura 9.3: Carga de trabajo sugerida para el manejador de imágenes

A la función Lambda se le realizarán pruebas con imágenes de entrada de distinto tamaño para evaluar su comportamiento bajo estos escenarios, particularmente el tiempo de respuesta de la función. Los resultados obtenidos a partir de estas pruebas van a servir como un punto de referencia para experimentos futuros. La figura 9.3 muestra una sugerencia de dos posibles cargas de trabajo:

1. 1000 solicitudes de cambio de tamaño de una imagen grande. En la figura 9.3, imagen-2.jpg de tamaño de 5Mb, representa una imagen grande.
2. 1000 solicitudes de cambio de tamaño de una imagen pequeña. En la figura 9.3, imagen-1.jpg de tamaño menor o igual a 500Kb, representa una imagen pequeña.

En principio las cargas de trabajo generadas serían cerradas, lo que quiere decir que una solicitud se ejecuta solamente hasta que la anterior se termina. Esto ayudará en principio a tener mejor trazabilidad de lo que ocurre con la función.

¿Por qué este caso de uso se considera relevante? A continuación se listan las características que hacen este caso de uso representativo e interesante:

- Sencillo de entender e implementar: se cuenta únicamente con una función la cual lleva a cabo una tarea muy específica.
- Popular: sigue un patrón de procesamiento dirigido por eventos y, como se señala en [37], este es uno de los más populares que se ha empezado a adoptar para aplicaciones *serverless*. Otra de las razones de la popularidad de este caso de uso es que permite a los desarrolladores crear una unidad de instalación independiente y especializada para el manejo de imágenes, liberando así a sus servidores y aplicaciones del manejo de las peticiones y lógica asociadas a estas.
- Replicable en otros proveedores de servicios en la nube: varias de las arquitecturas de referencia para *serverless* propuestas por Amazon, están compuestas por herramientas y servicios muy propios de su plataforma,

lo cual hace muy difícil su reproducibilidad utilizando otros proveedores. Aunque en principio este trabajo plantea ser elaborado en la plataforma FaaS de Amazon Web Services, AWS Lambda, otros proveedores de servicios (ver sección 2.1.3) en la nube cuentan con sus propias plataformas de FaaS y de almacenamiento, lo cual permitiría replicar lo aquí propuesto en ellos.

- Replicable en los lenguajes de programación soportados por plataformas FaaS: actualmente JavaScript, Java (y lenguajes basados en la *Java Virtual Machine*), Python, C# y Go son los principales lenguajes de programación soportados por las plataformas FaaS. El caso de uso propuesto, no presenta ningún tipo de característica que lo ate a un lenguaje de programación en particular. En todos ellos se cuentan con bibliotecas para manejo de imágenes tanto de forma nativa como por medio de soluciones de terceros.

9.1.4. Experimentos propuestos

9.2. Etapa 2

9.2.1. Modelo de rendimiento a partir de la función

Las cargas de trabajo descritas en la sección 9.1.3 van a alimentar la bitácora de eventos de la función. En AWS el servicio encargado de llevar registro de los eventos que suceden en una función Lambda se llama Amazon CloudWatch².

Para obtener un modelo a partir del comportamiento de una función, se pro-

²<https://aws.amazon.com/cloudwatch>

bién incluye *Performance Analysis Visualization* (PAVO) para la visualización de indicadores de rendimiento.

- **Kieker**: una herramienta para el monitoreo del rendimiento de la aplicación (APM por sus siglas en inglés).
- **Performance Model Extractor (PMX)**: una herramienta que puede derivar modelos de rendimiento a partir de los datos del monitoreo del rendimiento de la aplicación. Puede tomar los datos obtenidos por Kieker y generar modelos de rendimiento en PCM y DML (ver sección 2.1.2)

9.2.2. Enfoque de trabajo propuesto

El enfoque que se propone para obtener un modelo de rendimiento a partir de una función, es similar al de Walter et al. [40] pero omitimos el uso de las herramientas para determinar y consultar niveles de servicio, y también de visualización. La figura 9.5 muestra los participantes involucrados para lograr esta tarea.

A partir de las cargas de trabajo a las que fue sometida la función Lambda, métricas de rendimiento serán introducidas dentro del servicio Amazon CloudWatch, este creará la bitácora de la función. Para migrar las métricas de CloudWatch en Kieker, se pretende implementar un adaptador con el fin de tomar una muestra de la bitácora y convertirla a un formato que pueda ser entendido por Kieker. Otra alternativa que se valora es la de publicar las métricas de rendimiento directamente en Kieker desde la función.

Una vez que se cuente con una muestra de la ejecución de la función en formato Kieker, esta va a servir como entrada para PMX con el fin de obtener un

modelo de rendimiento en PCM. El modelo será utilizado para la ejecución de simulaciones.

9.3. Etapa 3: ejecutar simulaciones sobre el modelo obtenido

*Palladio Workbench*³ es la herramienta que permite crear y simular modelos basados en PCM.

Cuando se haya derivado un modelo de rendimiento de la función, se procederá a cargar este modelo en *Palladio Workbench* para modificar y calibrar el modelo y, por medio de las herramientas de simulación que vienen con *Palladio Workbench*, realizar experimentos con el fin de evaluar el grado de coincidencia de los resultados de la simulación con los obtenidos al exponer la función a cargas de trabajo reales.

³<https://www.palladio-simulator.com/>

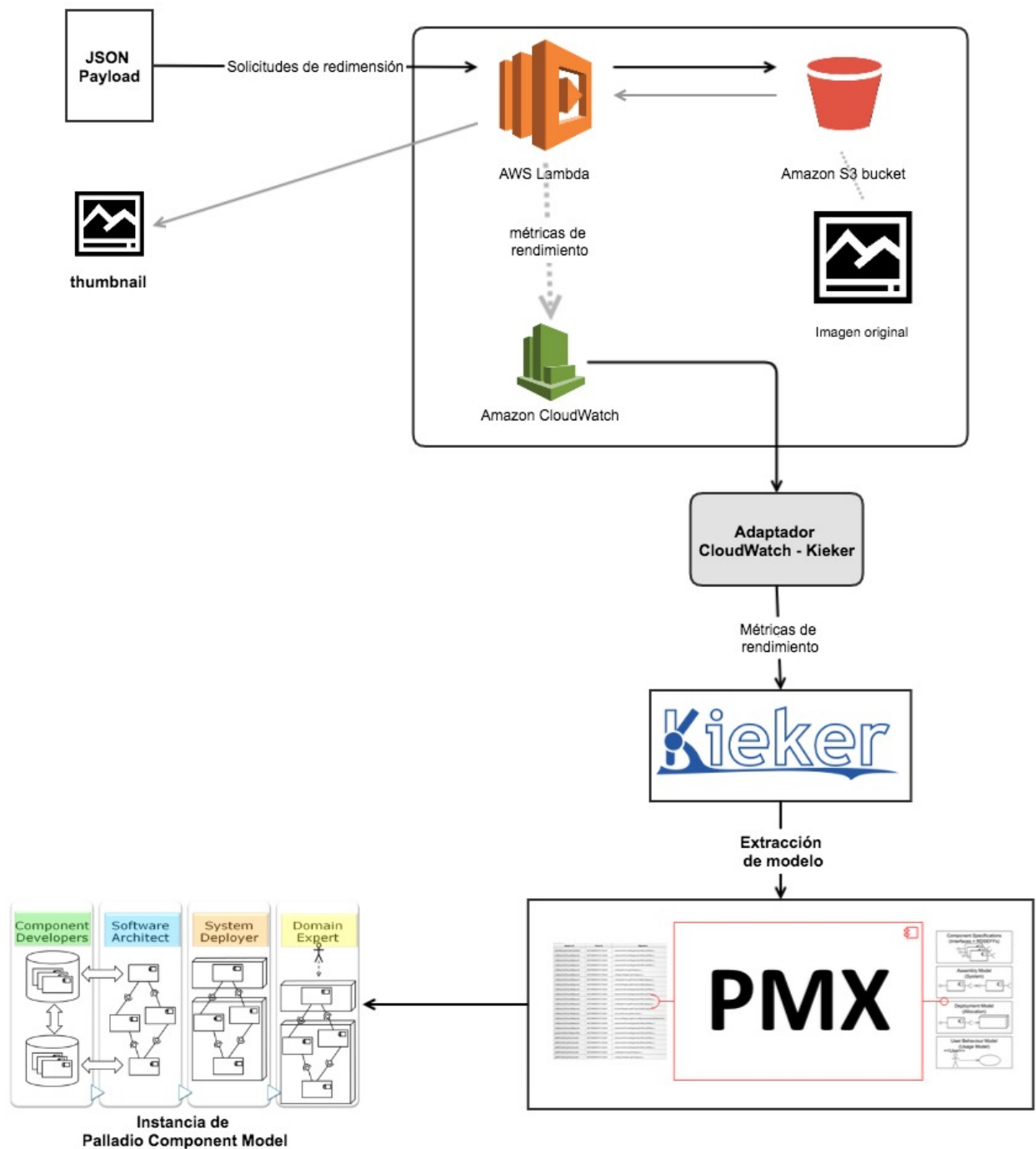


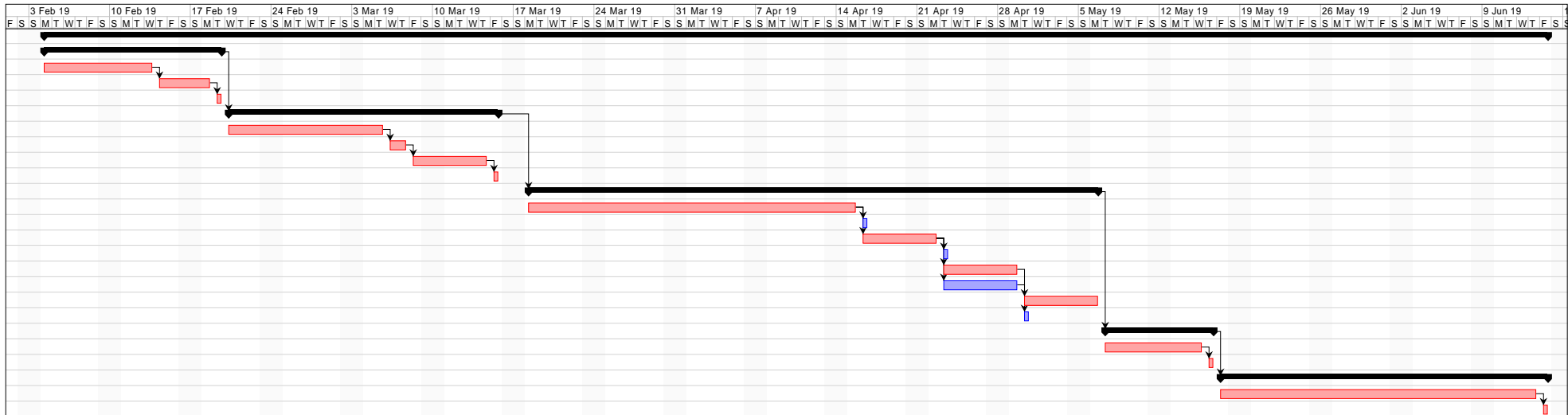


Figura 9.5: Enfoque de trabajo propuesto

Capítulo 10

Cronograma de actividades

		Name	Duration	Start	Finish	Predecessors
1		Modelado y simulación de FaaS	95 days	2/4/19 8:00 AM	6/14/19 5:00 PM	
2		Revisión de Literatura	12 days	2/4/19 8:00 AM	2/19/19 5:00 PM	
3		Selección de literatura relevante	8 days	2/4/19 8:00 AM	2/13/19 5:00 PM	
4		Preparación del documento	3 days	2/14/19 8:00 AM	2/18/19 5:00 PM	3
5		Revisión de avance con profesor asesor	1 day	2/19/19 8:00 AM	2/19/19 5:00 PM	4
6		Preparación del caso de uso	18 days	2/20/19 8:00 AM	3/15/19 5:00 PM	2
7		Implementar función en la nube	10 days	2/20/19 8:00 AM	3/5/19 5:00 PM	
8		Diseño de pruebas de carga	2 days	3/6/19 8:00 AM	3/7/19 5:00 PM	7
9		Ejecución de pruebas de carga	5 days	3/8/19 8:00 AM	3/14/19 5:00 PM	8
10		Revisión de avance con profesor asesor	1 day	3/15/19 8:00 AM	3/15/19 5:00 PM	9
11		Obtención de modelo de rendimiento	36 days	3/18/19 8:00 AM	5/6/19 5:00 PM	6
12		Integración AWS -> Kieker	21 days	3/18/19 8:00 AM	4/15/19 5:00 PM	
13		Revisión de avance con profesor asesor	1 day	4/16/19 8:00 AM	4/16/19 5:00 PM	12
14		Integración Kieker -> PMX	5 days	4/16/19 8:00 AM	4/22/19 5:00 PM	12
15		Revisión de avance con profesor asesor	1 day	4/23/19 8:00 AM	4/23/19 5:00 PM	14
16		Ejecución de simulaciones	5 days	4/23/19 8:00 AM	4/29/19 5:00 PM	14
17		Calibración del modelo	5 days	4/23/19 8:00 AM	4/29/19 5:00 PM	14
18		Validación de resultados	5 days	4/30/19 8:00 AM	5/6/19 5:00 PM	16
19		Revisión de avance con profesor asesor	1 day	4/30/19 8:00 AM	4/30/19 5:00 PM	17
20		Preparación de guía metodológica	8 days	5/7/19 8:00 AM	5/16/19 5:00 PM	11
21		Elaboración de la guía	7 days	5/7/19 8:00 AM	5/15/19 5:00 PM	
22		Revisión de avance con profesor asesor	1 day	5/16/19 8:00 AM	5/16/19 5:00 PM	21
23		Preparación de documento final	21 days	5/17/19 8:00 AM	6/14/19 5:00 PM	20
24		Elaboración del documento	20 days	5/17/19 8:00 AM	6/13/19 5:00 PM	
25		Revisión de avance con profesor asesor	1 day	6/14/19 8:00 AM	6/14/19 5:00 PM	24



Bibliografía

- [1] I. Baldini, P. C. Castro, K. S. Chang, P. Cheng, S. J. Fink, V. Ishakian, N. Mitchell, V. Muthusamy, R. M. Rabbah, A. Slominski, and P. Suter, “Serverless computing: Current trends and open problems,” *CoRR*, vol. abs/1706.03178, 2017.
- [2] R. Heinrich, A. van Hoorn, H. Knoche, F. Li, L. E. Lwakatare, C. Pahl, S. Schulte, and J. Wettinger, “Performance engineering for microservices: Research challenges and directions,” in *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion*, ICPE ’17 Companion, (New York, NY, USA), pp. 223–226, ACM, 2017.
- [3] M. Woodside, G. Franks, and D. C. Petriu, “The future of software performance engineering,” in *Future of Software Engineering (FOSE ’07)*, pp. 171–187, May 2007.
- [4] H. Koziolok, “Performance evaluation of component-based software systems: A survey,” *Perform. Eval.*, vol. 67, pp. 634–658, Aug. 2010.
- [5] T. de Gooijer, “Performance modeling of ASP .Net web service applications : an industrial case study,” 2011.

- [6] R. H. Reussner, S. Becker, J. Happe, R. Heinrich, A. Koziolk, H. Koziolk, M. Kramer, and K. Krogmann, *Modeling and Simulating Software Architectures: The Palladio Approach*. The MIT Press, 2016.
- [7] Y. Jin, A. Tang, J. Han, and Y. Liu, “Performance evaluation and prediction for legacy information systems,” in *Proceedings of the 29th International Conference on Software Engineering, ICSE ’07*, (Washington, DC, USA), pp. 540–549, IEEE Computer Society, 2007.
- [8] O.-Q. Noorshams, *Modeling and Prediction of I/O Performance in Virtualized Environments*. PhD thesis, 2015. Disponible en: <http://dx.doi.org/10.5445/KSP/1000046300>.
- [9] C. Szyperski, *Component Software: Beyond Object-Oriented Programming*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2nd ed., 2002.
- [10] S. Balsamo, A. D. Marco, P. Inverardi, and M. Simeoni, “Model-based performance prediction in software development: a survey,” *IEEE Transactions on Software Engineering*, vol. 30, pp. 295–310, May 2004.
- [11] A. Koziolk, H. Koziolk, and R. Reussner, “Peroptryx: Automated application of tactics in multi-objective software architecture optimization,” in *Proceedings of the Joint ACM SIGSOFT Conference – QoSA and ACM SIGSOFT Symposium – ISARCS on Quality of Software Architectures – QoSA and Architecting Critical Systems – ISARCS, QoSA-ISARCS ’11*, (New York, NY, USA), pp. 33–42, ACM, 2011.
- [12] S. Kounev, F. Brosig, and N. Huber, “The Descartes Modeling Language,” tech. rep., Department of Computer Science, University of Wuerzburg, October 2014.

- [13] G. Brataas, E. Stav, S. Lehrig, S. Becker, G. Kopčak, and D. Huljenic, “Cloudscale: Scalability management for cloud systems,” in *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering*, ICPE '13, (New York, NY, USA), pp. 335–338, ACM, 2013.
- [14] S. Becker, H. Koziolk, and R. Reussner, “The palladio component model for model-driven performance prediction,” *J. Syst. Softw.*, vol. 82, pp. 3–22, Jan. 2009.
- [15] P. M. Mell and T. Grance, “Sp 800-145. the nist definition of cloud computing,” tech. rep., Gaithersburg, MD, United States, 2011.
- [16] K. Craig-Wood, “IaaS vs. PaaS vs. SaaS definition,” May 2010. <https://www.katescomment.com/iaas-paas-saas-definition/>, obtenido el 27 de Octubre del 2018.
- [17] “Serverless computing - Amazon Web Services,” 2008. Obtenido el 26 Setiembre del 2018 de <https://aws.amazon.com/serverless>.
- [18] M. Roberts, “Serverless architectures,” 2018. Obtenido el 25 Setiembre del 2018 de <https://martinfowler.com/articles/serverless.html>.
- [19] N. Novkovic, “Top function as a service (FaaS) providers,” May 2018. <https://dashbird.io/blog/top-function-as-a-service-faas-providers/>, obtenido el 22 de Noviembre del 2018.
- [20] M. Cavallari and F. Tornieri, “Information systems architecture and organization in the era of microservices,” in *Network, Smart and Open* (R. Lamboglia, A. Cardoni, R. P. Dameri, and D. Mancini, eds.), (Cham), pp. 165–177, Springer International Publishing, 2018.
- [21] C. M. Aderaldo, N. C. Mendonça, C. Pahl, and P. Jamshidi, “Benchmark requirements for microservices architecture research,” in *2017 IEEE/ACM 1st*

- International Workshop on Establishing the Community-Wide Infrastructure for Architecture-Based Software Engineering (ECASE)*, pp. 8–13, May 2017.
- [22] A. Brunnert, A. van Hoorn, F. Willnecker, A. Danciu, W. Hasselbring, C. He-ger, N. R. Herbst, P. Jamshidi, R. Jung, J. von Kistowski, A. Koziol, J. Kroß, S. Spinner, C. Vögele, J. Walter, and A. Wert, “Performance-oriented devops: A research agenda,” *CoRR*, vol. abs/1508.04752, 2015.
- [23] P. D. Francesco, I. Malavolta, and P. Lago, “Research on architecting micro-services: Trends, focus, and potential for industrial adoption,” in *2017 IEEE International Conference on Software Architecture (ICSA)*, pp. 21–30, April 2017.
- [24] M. Crane and J. Lin, “An exploration of serverless architectures for information retrieval,” in *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR ’17*, (New York, NY, USA), pp. 241–244, ACM, 2017.
- [25] W. Lloyd, S. Ramesh, S. Chinthalapati, L. Ly, and S. Pallickara, “Serverless computing: An investigation of factors influencing microservice performance,” in *2018 IEEE International Conference on Cloud Engineering (IC2E)*, pp. 159–169, April 2018.
- [26] M. Villamizar, O. Garcés, L. Ochoa, H. Castro, L. Salamanca, M. Verano, R. Casallas, S. Gil, C. Valencia, A. Zambrano, and M. Lang, “Infrastructure cost comparison of running web applications in the cloud using aws lambda and monolithic and microservice architectures,” in *2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, pp. 179–182, May 2016.

- [27] E. F. Boza, C. L. Abad, M. Villavicencio, S. Quimba, and J. A. Plaza, “Reserved, on demand or serverless: Model-based simulations for cloud budget planning,” in *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–6, Oct 2017.
- [28] E. van Eyk, A. Iosup, C. L. Abad, J. Grohmann, and S. Eismann, “A specific cloud group’s vision on the performance challenges of faas cloud architectures,” in *Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, ICPE ’18*, (New York, NY, USA), pp. 21–24, ACM, 2018.
- [29] G. McGrath and P. R. Brenner, “Serverless computing: Design, implementation, and performance,” in *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pp. 405–410, June 2017.
- [30] S. Hendrickson, S. Sturdevant, T. Harter, V. Venkataramani, A. C. Arpaci-Dusseau, and R. H. Arpaci-Dusseau, “Serverless computation with open-lambda,” *Elastic*, vol. 60, p. 80, 2016.
- [31] M. Yan, P. Castro, P. Cheng, and V. Ishakian, “Building a chatbot with serverless computing,” in *Proceedings of the 1st International Workshop on Mashups of Things and APIs, MOTA ’16*, (New York, NY, USA), pp. 5:1–5:4, ACM, 2016.
- [32] J. Spillner, “Practical tooling for serverless computing,” in *Proceedings of the 10th International Conference on Utility and Cloud Computing, UCC ’17*, (New York, NY, USA), pp. 185–186, ACM, 2017.
- [33] DZone, “The 2018 dzone guide to cloud: serverless, functions, and multicloud,” 2018. <https://dzone.com/guides/>

cloud-serverless-functions-and-multi-cloud, obtenido el 17 de octubre del 2018.

- [34] RightScale, “State of the cloud report,” 2018. <https://www.rightscale.com/lp/state-of-the-cloud>, obtenido el 17 de octubre del 2018.
- [35] D. Ocean, “Currents: A quarterly report on developer trends in the cloud,” 2018. <https://www.digitalocean.com/currents/june-2018/>, obtenido el 17 de octubre del 2018.
- [36] C. F. Foundation, “Where PaaS, containers and serverless stand in a multi-platform world,” June 2018. <https://www.cloudfoundry.org/multi-platform-trend-report-2018/>.
- [37] M. Boyd, “Serverless architecture: Five design patterns,” March 2017. <https://thenewstack.io/serverless-architecture-five-design-patterns/>.
- [38] A. W. Services, “AWS lambda - resources: Reference architectures,” 2018. <https://aws.amazon.com/lambda/resources/reference-architectures/>.
- [39] A. W. Services, “Serverless image handler – AWS answers,” 2018. <https://aws.amazon.com/answers/web-applications/serverless-image-handler/>.
- [40] J. Walter, S. Eismann, J. Grohmann, D. Okanovic, and S. Kounev, “Tools for declarative performance engineering,” in *Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, ICPE ’18*, (New York, NY, USA), pp. 53–56, ACM, 2018.