Fernando de la Prieta Pintado

SISTEMAS ORGANIZATIVOS PARA LA ASIGNACIÓN DINÁMICA DE RECURSOS COMPUTACIONALES EN ENTORNOS DISTRIBUIDOS

DOI: https://doi.org/10.14201/0VI0449

COLECCIÓN VÍTOR

Ediciones Universidad Salamanca

FERNANDO DE LA PRIETA PINTADO

SISTEMAS ORGANIZATIVOS PARA LA ASIGNACIÓN DINÁMICA DE RECURSOS COMPUTACIONALES EN ENTORNOS DISTRIBUIDOS

Ediciones Universidad
Salamanca

COLECCIÓN VÍTOR

449

1.ª edición: diciembre, 2021
I.S.B.N.: 978-84-1311-611-2
DOI: https://doi.org/10.14201/0VI0449

Ediciones Universidad de Salamanca
Plaza San Benito s/n
E-37002 Salamanca (España)
httsp://www.eusal.es
eus@usal.es

Hecho en UE-Made in EU

Realizado por:
Cícero, S.L.U.
Tel. +34 923 12 32 26
37007 Salamanca (España)

Ediciones Universidad de Salamanca es miembro de la UNE
Unión de Editoriales Universitarias Españolas www.une.es

♠

Accesible en:
https://eusal.es/index.php/eusal/catalog/book/978-84-1311-184-1

♠

Catalogación de editor en ONIX diponible en https://www.dilve.es/

RESUMEN

Cloud Computing, el conocido paradigma computacional, está emergiendo en los últimos años con gran fuerza. Este paradigma incluye un novedoso modelo de comercialización basado en el pago por uso que ha cambiado radicalmente el modelo de negocio en Internet. Este nuevo modelo computacional también ha derivado en que el modelo de producción de estos recursos computacionales evolucione hasta una aproximación cercana al modelo de producción *just-in-time*, en el que sólo se consumen los recursos necesarios para la producción de los servicios en función de la demanda existente en cada momento, hablándose dentro de este ámbito de elasticidad en los servicios ofertados.

Pese a los indudables avances que se han producido a nivel tecnológico, todavía hoy existe una gran capacidad de mejora de estos sistemas. En este sentido, en el marco de esta Tesis Doctoral se propone el uso de los sistemas multiagente y, especialmente, aquellos basados en modelos organizativos para el control y monitorización de un sistema Cloud Computing. Gracias a esta aproximación, una de las primeras en este campo de investigación, será posible incluir en las plataformas Cloud una nueva generación características derivadas de la Inteligencia Artificial, como son la autonomía, la proactividad y, también, la capacidad de aprendizaje.

PALABRAS CLAVE: Computación en nube, sistemas multiagente, distribución de recursos, aprendizaje, razonamiento basado en casos, programación lineal, sistemas distribuidos inteligentes, sistemas adaptativos, organizaciones virtuales, ingeniería del software.

ABSTRACT

Cloud Computing, a known computing paradigm, has been emerging in recent years with great strength. This paradigm includes a new marketing model based on the pay per use revenue model, which has radically changed the business model on the Internet. This new computational model has also allowed the production model for these computational resources to evolve very closely along the lines of the *just-in-time* production model, in which only the resources needed for the production of services are consumed according to the demand that exists at a given time. Within this context, the term *elasticity* is used in reference to the services being offered.

Despite the undeniable advances that have been produced at a technological level, there is still much room for improvement in these systems. In this regard, the framework of this doctoral thesis proposes the use of multiagent systems, particularly those based on organizational models to control and monitor a Cloud Computing system. Due to this approach, one of the first in this field of research, new generation Cloud platforms will be able to include characteristics derived from Artificial Intelligence, such as autonomy, proactivity, and learning capabilities.

KEY WORDS: Cloud Computing, Multiagent Systems, Resource allocation, learning, case-based reasoning, lineal programming, distributed intelligent systems, adaptive systems, virtual organizations, software engineering.

ÍNDICE DE CONTENIDOS

Índice de figuras

Índice de tablas

Capítulo 1. Introducción

Durante los últimos años la industria tecnológica y la comunidad científica están realizando un gran esfuerzo de investigación e implantación del paradigma tecnológico *Cloud Computing* (CC). Así, el número de plataformas tanto privadas, como públicas está creciendo rápidamente [1-4], lo que ha motivado en gran medida su desarrollo debido al interés económico de las grandes empresas tecnológicas que subyace frente a los aspectos puramente técnicos [5, 6].

Las plataformas CC desde un punto de vista externo, ofrecen tres tipos de servicios ampliamente conocidos (software, plataforma e infraestructura) [7], donde la principal novedad con respecto a las tecnologías precedentes radica en el hecho de ofrecer cualquier tipo de capacidad computacional como servicio a través de Internet. El modelo de comercialización también es innovador ya que está basado en el pago por uso (*pay-as-you-go*) [6], al igual que cualquier otro servicio público tradicional (luz, agua, gas, etc.). En este sentido, CC sigue el modelo propuesto por el paradigma abstracto *Utility computing* [8] en el que los usuarios para acceder a los servicios deben negociar y establecer previamente un acuerdo de uso —*Service Level Agreement (SLA)*— [9]. Una vez establecido este contrato de uso de bienes computacionales, tanto los usuarios (habitualmente pagando), como el sistema CC (manteniendo el servicio) están obligados a cumplir dicho contrato.

Este novedoso modelo de comercialización obliga a las plataformas CC a mantener la calidad de los servicios según se ha acordado previamente, lo que hace necesario analizar cómo su arquitectura interna hace posible producir este tipo de servicios. Así pues, las novedades vienen determinadas por el innovador espectro de tecnologías subyacentes (virtualización, granjas de servidores, servicios web, etc.), cuya madurez recientemente alcanzada, hace posible que los servicios sean ofertados con el mismo nivel de calidad con independencia de la demanda existente por parte de los usuarios [10-12].

Las nuevas posibilidades a nivel tecnológico conllevan el nacimiento de un nuevo concepto, el de la elasticidad [13]. Este nuevo concepto está basado en el método de producción *just-in-time* [14] que hace referencia a la forma de producción de los servicios (computacionales) y los recursos que son necesarios para ello. Así pues, los servicios producidos reciben sólo la cantidad de recursos necesarios para mantener un nivel calidad constante, atendiendo a la demanda instantánea [3, 4]. En la práctica, se traduce en que el usuario final hace uso de un conjunto de servicios (a través de internet) que siempre están disponibles y que a priori son ilimitados.

Para los usuarios del paradigma, este modelo de comercialización es muy ventajoso, ya que les permite no tener la obligación de aprovisionarse con recursos computacionales adicionales para contener los picos en la demanda de sus productos o servicios, ni tampoco disponer de la infraestructura necesaria para proporcionarlos. Lo que permite transformar los gastos de capital en gastos operacionales [6]. En contraposición, los proveedores deben proporcionar servicios CC elásticos que deben ser gestionados a través de entornos de tipo clúster HPC (*High Performance Computing*) cuyo control y monitorización debe realizarse de forma automática, proporcionando un acceso transparente a los usuarios finales.

Pese a los indudables avances que se han conseguido en los últimos años en el marco de los sistemas CC, no hay que olvidar que es una tecnología relativamente reciente que comienza en el año 2007 [15]. Aún, hoy en día, en algunos aspectos es un paradigma inmaduro [16-18] y, por lo tanto, todavía existen retos y necesidades que son del interés de una gran parte de la comunidad científica [19, 20]. En este sentido, uno de los grandes retos es la gestión automática de los recursos computacionales de un entorno CC [21, 22] y su correcta asignación a los servicios proporcionados por este tipo de plataformas, ya que deben asignarse en función de la demanda y los acuerdos SLA alcanzados.

Este trabajo de Tesis Doctoral se plantea como una de las primeras aproximaciones de los Sistemas Multiagente (SMA) basados en Organizaciones Virtuales (OV) en el marco del paradigma CC. El propósito del sistema organizativo propuesto no es sólo el control y monitorización de la infraestructura *hardware* (real o virtual) de un entorno CC, sino también la asignación de recursos computacionales disponibles entre los diferentes servicios ofertados a los usuarios finales, en función de la demanda existente. Este modelo de distribución o reparto de recursos que se propone se sustentará sobre las capacidades auto-adaptativas en entornos abiertos, dinámicos y heterogéneos que proporcionan los SMA con capacidades organizacionales [23-25].

Este primer capítulo de la memoria del trabajo de investigación comienza con la identificación de los problemas abiertos en el marco del paradigma CC y la formulación de la hipótesis inicial de esta Tesis Doctoral (apartado 1.1). A continuación, los apartados 1.2 y 1.3 presentan, respectivamente, los objetivos propuestos en el inicio del trabajo y la metodología de investigación que se ha empleado. Finalmente, en el último apartado del capítulo (1.4) se detalla la estructura de la presente memoria.

1.1.Problemas abiertos, motivación e hipótesis

Actualmente, las plataformas CC son principalmente utilizadas por su gran capacidad computacional, tanto en términos de potencia de computación, como por su facilidad para el almacenamiento de grandes volúmenes de datos. En este sentido, la mayoría de las plataformas CC conocidas [3, 4, 26] hacen un mayor énfasis en proporcionar servicios de infraestructura (a través de la virtualización del hardware), sin tener en cuenta los servicios de nivel superior (plataforma y software).

Cualquier sistema CC se caracteriza por las propiedades de su infraestructura subyacente y los servicios (o capacidades computacionales) que ofrecen al usuario a través de Internet. Mientras la infraestructura aporta recursos computacionales al sistema, los servicios consumen estos recursos. El principal reto en este contexto, consiste en mantener el nivel de calidad de los servicios (QoS, *Quality of Service*) con independencia de su demanda, lo que implica que sea necesario gestionar dinámicamente los recursos computacionales asignados a cada servicio ofertado al usuario. Para ello estas plataformas, en la mayoría de los casos, utilizan un modelo de gestión (monitorización y control) de la infraestructura propia que es centralizado, lo que constituye en sí mismo una limitación, ya que choca frontalmente con el principio de alta disponibilidad que se les presupone a los sistemas CC. Además, dada la complejidad tecnológica que implica proporcionar servicios de infraestructura elásticos, éstas plataformas tampoco hacen un énfasis especial en la optimización de los recursos *hardware*, los cuáles consumen ingentes cantidades de energía y tienen enormes necesidades de refrigeración [27].

Formalmente, el problema se plantea desde dos puntos de vista. Por un lado, una vista micro, en el que la reasignación de recursos se realiza en cada uno de los servicios ofertados, o dentro de cada servidor físico. Por otro lado, a nivel macro, la reasignación de recursos se produce a nivel global del entorno CC, lo que puede implicar varios servicios o nodos computacionales. En el estado del arte se observan diversos trabajos en este ámbito, que pueden ser clasificados en tres grandes grupos: (i) los que toman las decisiones de forma centralizada (que son la mayoría) [22, 28]; (ii) los que estiman o planifican el uso sin tener en cuenta la demanda instantánea [29, 30]; y, finalmente, (iii) los que optimizan los recursos en función de la demanda instantánea sin tener en cuenta la demanda histórica [31].

Si se realiza una investigación más profunda y detallada, en primer lugar, desde el punto de vista micro, aunque no es posible encontrar una gran cantidad de estudios en el estado de arte, si que existen modelos que satisfacen el problema; ya sea en cuanto la reasignación de recursos entre las máquinas virtuales que aloja un servidor [32], como en el balanceo entre las diferentes máquinas virtuales asignadas a un recurso [33]. En segundo lugar, desde el punto de vista macro las referencias existentes siguen modelos principalmente orientados al mercado [34, 35], modelos matemáticos [31, 36] y, recientemente, se empiezan a observar trabajos que tienen en cuenta el consumo energético en la distribución de los recursos computacionales [28].

Estos modelos no sólo deberían proporcionar un modelo de elasticidad (creciente y decreciente), sino que también deben proporcionar herramientas para reducir los costes operacionales del entorno CC; tanto desde el punto de vista del usuario final, como también desde el punto de vista de proveedor, ya que como se ha indicado previamente, los entornos HPC requieren una gran cantidad de recursos energéticos y de refrigeración [27]. Sin embargo, los trabajos existentes en el estado del arte están basados en métodos que utilizan algoritmos centralizados, basados en modelos matemáticos y heurísticos, que no aseguran la eficiencia del sistema y menos su disponibilidad

frente a fallos. Este tipo de algoritmos deberían evolucionar hacia modelos en el que los diferentes componentes dispongan de una representación parcial del entorno, que les permita interaccionar y compartir información con sus iguales, de forma que los algoritmos de gestión de los recursos se distribuyan a lo largo de la infraestructura, lo que por ende facilitará su implantación con independencia del tamaño del centro de datos.

En este contexto de inmadurez tecnológica, la teoría de agentes y SMA [37, 38] puede aportar un nuevo modelo de gestión de sistemas CC basado en la distribución de responsabilidades, la flexibilidad y la autonomía. La gestión de las funcionalidades del núcleo de un sistema CC mediante un modelo basado en SMA permitiría a las plataformas resultantes ser mucho más eficientes, escalables y adaptables que las existentes actualmente. Sin embargo, la unión de ambos modelos computacionales (SMA y CC) es un gran reto, pero un reto asumible en el marco de este trabajo de investigación, debido a que un entorno CC es considerado un sistema abierto debido a su heterogeneidad, dinamicidad e incertidumbre:

- Es **heterogéneo**, ya que puede incluir diversos elementos y componentes muy diversos entre sí, tanto a nivel hardware, como a nivel software.
- Es **dinámico** ya que la demanda de sus servicios varía a lo largo del tiempo y, además, el funcionamiento de sus componentes tiene que adaptarse a los nuevos estados que puedan darse, incluyendo la admisión y salida de elementos del sistema en pleno funcionamiento.
- Tiene **incertidumbre** ya que cada uno de sus componentes conoce tan sólo una parte del sistema; teniendo que tomar decisiones en base a la información parcial disponible en cada momento.

La aplicación de SMA en sistemas abiertos es un reto conocido. En el estado del arte existe una gran variedad de trabajos que ha sido desarrollados y evaluados en diversos contextos abiertos, obtenido un notable éxito en su aplicación [39-43]. En el marco de este trabajo, se consideran especialmente efectivos los SMA diseñados mediante modelos organizativos ya que pueden aportar soluciones avanzadas e innovadoras [44] que permitan explotar estrategias diferenciadoras que aporten flexibilidad, capacidad y rapidez de respuesta, dentro de una estrategia dirigida a satisfacer al consumidor [45]. El uso de SMA permite continuar con la investigación en técnicas, herramientas y metodologías que permitan incorporar a las plataformas CC características inteligentes como pueden ser la autonomía o el aprendizaje, entre otras.

Así pues, la hipótesis de partida es que *es posible modelar el sistema de control y monitorización de una plataforma CC mediante el uso de OV de agentes inteligentes que se auto-adapten y reorganicen en función de las necesidades del contexto. Así pues, los agentes individuales de forma autónoma gestionarán los recursos computacionales y los servicios disponibles, coordinándose con el resto de agentes para proporcionar un modelo de adaptación dinámico y distribuido en función de los intereses del usuario final.*

Aprovechando las ventajas que nos proporcionan las particularidades del desarrollo de SMA desde el punto de vista organizacional y considerando las carencias existentes actualmente en el marco del paradigma CC, se formalizará un modelo arquitectónico dinámico y adaptativo que hará uso de técnicas de optimización basada en la investigación operativa y de aprendizaje mediante sistemas de razonamiento basados en casos [46, 47].

1.2. Objetivos

Considerando las características existentes en el paradigma CC en tanto en cuanto a la redistribución de recursos y dependencia del entorno tecnológico subyacente. Como hipótesis de partida se plantea la necesidad de incorporar modelos asociados a la Inteligencia Artificial (IA) para la gestión de entornos computacionales de última generación. Las particularidades del desarrollo de SMA desde el punto de vista organizacional posibilitan la formalización de una arquitectura que satisfaga las demandas del entorno CC.

Por lo tanto, el objetivo principal que se pretende conseguir es *modelar una arquitectura multiagente basada en OV para la monitorización y control de un entorno CC, así como la investigación y desarrollo de un modelo de asignación dinámica y adaptativa de recursos computacionales en entornos distribuidos*.

Para satisfacer este objetivo se plantean un conjunto de objetivos más específicos que haga posible este objetivo principal:

- Estudiar el paradigma CC a nivel global, incluyendo tipos de servicios, modelos de despliegue, arquitecturas existentes y posibles debilidades en el paradigma.
- Analizar la relación existente entre los SMA y los sistemas CC, identificando las principales áreas de aprovechamiento de los SMA en el marco del paradigma CC.
- Estudiar las tendencias en el marco del análisis, diseño y posterior desarrollo de los SMA organizativos, así como de los modelos de adaptación y coordinación.
- Realizar un estudio de las tecnologías complementarias, ya sean aquellas que tienen un perfil claramente tecnológico (virtualización, balanceo de carga, computación de alto rendimiento, etc.). Como aquellas más cercanas a la investigación como son los sistemas de razonamiento basados en casos.
- Diseñar un modelo de integración basado en la teoría de agentes, prestando especial atención a los sistemas organizativos.
- Aplicar mecanismos generales de optimización de recursos en el marco de la distribución de recursos computacionales en entornos CC.
- Diseñar un modelo adaptativo y de aprendizaje sobre la base que proporciona los sistemas de razonamiento basados en casos, es decir, que permita apoyarse en experiencias pasadas, para generar la base del conocimiento actual.
- Evaluar empíricamente el modelo de arquitectura y los algoritmos de distribución de recursos propuestos en un entorno de aplicación real.

1.3.Metodología de investigación

En una tesis donde el principal componente de trabajo está relacionado con la investigación, parece adecuado plantear una metodología que permita establecer objetivos alcanzables en un tiempo fijado. Posteriormente, evaluar estos los resultados de forma conjunta con los directores y, discutir el siguiente conjunto de objetivos atendiendo a los resultados obtenidos.

El método de investigación que se ha seguido es conocido como *action-research* [48]. Se caracteriza por estar orientado a la acción y el cambio, lo que permite focalizarse en un problema para generar conocimiento de investigación relevante en un periodo de tiempo establecido. Este método es uno de los más habituales en el ámbito la investigación de los sistemas de información, ya que hace posible la investigación empírica.

A continuación, se presentan los hitos que se plantearon al inicio del presente trabajo de tesis:

- Revisión detallada del estado del arte e identificación de novedades.
- Diseño de un entorno de pruebas.
- Inclusión de los principios de los SMA en los entornos CC.
- Diseño de un sistema multiagente basado en OV.
- Propuesta, investigación y mejora de algoritmos de distribución de recursos computacionales.
- Evaluación de los algoritmos propuestos.

Estos hitos, se han alcanzado a través de un modelo de indagación original y planificada, que se ha basado en tres grandes etapas. La primera etapa se centró en el análisis del estado del arte incluyendo tanto los trabajos relacionados con el paradigma CC, como con los SMA y las OV. Posteriormente, durante la segunda etapa se ha diseñado y desarrollado de forma iterativa e incremental tanto el modelo arquitectónico basado en SMA, como los algoritmos distribución de recursos computacionales. En este sentido, dado que se esperaba obtener una arquitectura software, se ha seguido un modelo de desarrollo basado en prototipos, de forma que los objetivos perseguidos se han ido incorporando gradualmente al resultado final. Gracias a este modelo ha sido posible alcanzar los objetivos en el tiempo estimado. Finalmente, la última etapa ha consistido en la evaluación del trabajo desarrollado, lo que ha dado lugar a la validación tanto de la plataforma multiagente, como la de los algoritmos propuestos.

De forma paralela a estas tres grandes etapas, se ha realizado un trabajo constante de diseminación del conocimiento, resultados y experiencias obtenidas, siendo posible su explotación por parte de la comunidad científica. Esta actividad ha dado lugar a la elaboración y presentación de diversas publicaciones en revistas, congresos, talleres, generación de informes técnicos, etc. que han permitido validar y dar a conocer los avances y resultados parciales de los distintos hitos de la investigación en diferentes ámbitos con la consiguiente realimentación al trabajo.

No cabe duda de que, para la investigación de un entorno CC, es necesario disponer de un entorno de pruebas adecuado. Así, en este sentido, el grupo de investigación BISITE[1] de la Universidad de Salamanca ha desplegado una completa granja 15 de servidores de última generación, con capacidad de virtualización por hardware y 16Gb de memoria RAM por equipo. Todos ellos están conectados punto a punto mediante una red 10Gbps. Gracias a esta infraestructura ha sido posible validar todas las innovaciones que se han obtenido en el marco de esta tesis doctoral.

[1] http://bisite.usal.es

1.4.Estructura del documento

El presente documento se organiza en 6 capítulos que siguen la estructura secuencial, pero iterativa, del trabajo realizado en el marco de la presente tesis doctoral. A continuación, se detalla cada una de estas piezas, las cuáles describen detalladamente el proceso de investigación que se ha seguido.

El Capítulo 2 y el Capítulo 3 abordan el estudio del estado del arte de las tecnologías, métodos, técnicas y herramientas directamente relacionadas con la temática de este trabajo de investigación. En primer lugar, en el Capítulo 2 se presenta un análisis detallado del paradigma de computación CC. En este sentido, dado su reciente nacimiento, inicialmente se analizan las definiciones existentes con el objetivo de acotar el término, el modelo tecnológico y su ámbito de aplicación (modelos, servicios, roles, etc.). En este segundo capítulo, también se realiza un análisis de las plataformas existentes, así como los riesgos y vulnerabilidades asociadas. El capítulo concluye con el análisis de las debilidades que hoy en día presenta la tecnología CC. A partir de este análisis, el Capítulo 3 se centra en el estudio del estado del arte de los SMA y la incipiente relación que se empieza a observar entre estos sistemas y el paradigma CC. Así, este capítulo aporta una extensa relación de los trabajos existentes en el estado de arte sobre la aplicación de SMA en diferentes áreas del paradigma CC y cómo estas aportaciones contribuyen a la mejora del modelo tecnológico en su conjunto. Finalmente, el Capítulo 3 aborda el análisis de los modelos organizativos de SMA con el objetivo de facilitar su aplicación en el ámbito del paradigma CC.

El Capítulo 4 está centrado en la descripción del modelo arquitectónico propuesto para la monitorización y control de un entorno CC mediante SMA basados en OV, así como los algoritmos y técnicas desarrolladas para la distribución de recursos en entornos CC. El capítulo comienza con una descripción detallada del entorno computacional, con el objetivo de caracterizar y formalizar la infraestructura computacional subyacente, incluyendo el modelo, así como la relación formal existente entre una plataforma CC y los usuarios de la misma. Posteriormente, a partir de esta representación se describe en detalle la arquitectura multiagente propuesta, denominada +Cloud. Esta arquitectura se ha modelado mediante el uso de la metodología GORMAS [49, 50] la cuál facilita la especificación de SMA abiertos siguiendo la perspectiva de las organizaciones humanas. Finalmente, este capítulo también describe los algoritmos propuestos para la distribución dinámica de recursos en función de la demanda en los servicios. Así pues, dada la naturaleza distribuida de estos algoritmos, el modelo de adaptación dinámica se describe atendiendo a los componentes que realizan la distribución o reparto de recursos computacionales existentes en el sistema CC entre los diferentes servicios de la plataforma, así como las interacciones que se producen entre ellos.

A continuación, el Capítulo 5 recoge la evaluación de la arquitectura y los algoritmos de distribución de recursos propuestos. En este sentido, el entorno de pruebas ha sido una plataforma CC desarrollada en el marco de esta tesis doctoral y que integra la arquitectura de agentes (+Cloud) propuesta. Gracias a esta plataforma ha sido posible evaluar el sistema y algoritmos propuestos en un entorno de explotación realista, lo que por tanto asegura su validez. El capítulo, en primer lugar, presenta una breve descripción de la plataforma CC y como se ha realizado la integración con el SMA, posteriormente, se presenta un caso de estudio que muestra como el sistema se adapta dinámicamente a los cambios en la demanda, así como el proceso de distribución de recursos computacionales. Finalmente, el capítulo concluye con un análisis crítico del modelo propuesto y su comparación con otros trabajos existentes en el estado del arte y plataformas.

Finalmente, el Capítulo 6 presenta las conclusiones obtenidas una vez finalizado el trabajo de investigación, aportando las contribuciones al estado del arte de esta investigación y proponiendo una serie de líneas de trabajo a realizar a medio plazo.

Capítulo 2. Los sistemas Cloud Computing

CC [7], entendido como paradigma computacional, está emergiendo en los últimos años con gran fuerza, hasta el punto de que está empezando a ser muy habitual en cualquier ámbito relacionado con la computación, más allá del contexto social de Internet. Muchas de las iniciativas de la Unión Europea, en el marco del programa Horizonte 2020 (H2020)[2] se centran en el paradigma tecnológico que proporciona los sistemas CC. Además, empresas punteras en el ámbito tecnológico tales como IBM, Amazon, Microsoft, etc. apuestan decididamente por este modelo computacional en su modelo de negocio y oferta de servicios. Aunque a primera vista pueda ser considerado como un paradigma meramente tecnológico, la realidad demuestra que su rápida progresión está principalmente motivada por el interés económico que subyace alrededor de las características puramente computacionales [5, 6]. Hoy en día, parece evidente que, aunque todavía existen muchas barreras a superar a nivel técnico [16, 51], el ritmo de innovación viene determinado por los intereses macroeconómicos impuestos por las grandes compañías tecnológicas [5].

Como ya se ha anticipado en el Capítulo 1, es posible identificar en las actuales plataformas CC ciertas debilidades a la hora de gestionar de forma eficiente sus recursos, responder de forma autónoma a los picos en la demanda, automatizar el proceso de toma de decisiones, responder a las caídas en los sistemas, etc. [3, 4]. Para abordar estos problemas existentes hoy en día, se identifican a los SMA [37, 38] según la hipótesis de este trabajo, como la clave para proporcionar una solución eficiente a estos problemas, permitiendo la evolución del paradigma CC mediante la incorporación de nuevas características y capacidades a las plataformas CC existentes [52, 53].

Dada la complejidad que presenta los sistemas CC donde existen una gran variedad de tecnologías, componentes, sistemas, etc. que tienen que trabajar de forma conjunta, se revisará en detalle las características de las plataformas CC, identificando fortalezas y debilidades, así como necesidades concretas. Este análisis servirá como punto de partida para la incorporación de los SMA en este contexto tecnológico.

El siguiente apartado del capítulo (2.1) contiene una revisión histórica del paradigma CC, revisión que se enlaza con las tecnologías relacionadas, las cuáles se presentan en el apartado 2.2. A continuación en los apartados 2.3 y 2.4 se analizan de forma crítica las diferentes definiciones existentes en el estado del arte, los tipos de servicios ofertados, roles del paradigma, los modelos de despliegue, así como los patrones de interconexión entre plataformas CC. Posteriormente, el apartado 2.5 analiza algunas de las plataformas más extendidas y, también, presenta un modelo de arquitectura de referencia. Finalmente, el apartado 2.6 muestra las vulnerabilidades y riesgos documentados en el estado del arte, que se analizarán en el apartado final (2.7) a modo de conclusión del capítulo.

[2] http://ec.europa.eu/programmes/horizon2020/

2.1.El paradigma Cloud Computing

Históricamente, el término CC fue utilizado por primera vez por el Profesor Chellappa [54], quien ya avanzaba que el modelo de computación en el futuro estaría mucho más ligado a los intereses económicos, que a las limitaciones impuestas por la tecnología. Aunque hace más de una década, en un contexto con continuos avances tecnológicos esta afirmación resultaba utópica, hoy en día, la realidad demuestra que el desarrollo de la tecnológica está motivado por los intereses del mercado [34].

El concepto fue haciéndose popular de la mano de Salesforce.com Inc., empresa que centraba su estrategia comercial en ofrecer *software* como servicio a grandes empresas [55]. Su modelo de negocio era radicalmente diferente al del resto de empresas de la época. Así, Salesforce, en lugar de ofrecer paquetes software a medida, lo que comercializaba era acceso a software genérico desplegado en sus propios servidores y accesible a través de Internet [56]. Este nuevo modelo de negocio, pese a la innovación en su concepción, ya había sido enunciado previamente por Carr [57].

El primer producto de Saleforce fue un CRM (*Customer Relationship Management*), que posteriormente daría lugar a la plataforma Force.com, lanzada en el año 2007. Esta plataforma puede considerarse el origen conceptual de otras plataformas ampliamente conocidas actualmente. Este nuevo mercado tecnológico se completó, en primer lugar, con el nacimiento de los servicios web al inicio de la década de los 2000 [58], lo que permitió el desarrollo de una nueva arquitectura software basada en servicios [59]. Así mismo, unido a esta nueva arquitectura *software*, Amazon desarrolló en torno al año 2006 su plataforma *Amazon Web Services* (AWS)[60] que incluía un novedoso servicio de infraestructura, conocido como Amazon S3, basado en la tecnología de virtualización; así como un servicio de computación distribuida basados en el concepto de *Grid Computing* (Amazon EC2) [61].

Muchos años atrás, desde un punto de vista de investigación, fue IBM quien avanzó, a través del documento *Autonomic Computing Manifesto* [62] las directrices que guiarían el modelo computacional (auto-configuración, auto-monitorización, auto-optimización, etc.). Partiendo de estas directrices, diferentes empresas (Yahoo, Amazon, eBay y Microsoft), también investigaron ampliamente en este concepto durante los años 2003 y 2007, con resultados muy dispares, donde la mayoría de aproximaciones estuvieron centradas en la capa de infraestructura [63].

El desarrollo del paradigma CC tal y como hoy lo conocemos, tanto a nivel tecnológico como de comercialización, fue desarrollado por el consorcio formado 2007 por Google e IBM junto con 6 prestigiosas universidades americanas: Massachusetts Institute of Technology (MIT), Standford University, University of California, University of Berkeley, University of Meryland y University of Washington [15]. Este esfuerzo, se debe a la ventaja competitiva que implica dominar este marco tecnológico; poseer una plataforma propia permite por un lado, (i) al proveedor CC ofrecer sus servicios siguiendo un modelo de pago por uso [34, 64] y los principios propuestos por el concepto *Utility Computing* [8]. Y por otro lado, (ii) al usuario CC, le permite no tener la obligación de aprovisionarse con recursos computacionales adicionales para contener los picos en la demanda de sus productos o servicios, ni tampoco disponer de la infraestructura para proporcionarlos; transformando los gastos de capital en gastos operacionales [6].

2.2.Tecnologías relacionadas

No cabe duda de que el desarrollo de este paradigma computacional CC ha sido motivado por el rápido desarrollo de un conjunto de tecnologías, sobre las que se sustenta, para ofrecer el catalogo de servicios computacionales bajo demanda. A continuación, se presentarán de forma resumida el conjunto de tecnologías precedentes, así como su aportación individual al paradigma CC [5, 11, 12]:

- **Grid Computing.** Es un paradigma de computación distribuida cuyo objetivo es coordinar recursos computacionales y de red para satisfacer un objetivo [12]. El desarrollo de este paradigma distribuido fue motivado para satisfacer las necesidades computacionales del software científico. Las características principales [61] de este paradigma son la heterogeneidad, reparto de recursos, múltiples administradores, acceso consistente y transparente y, finalmente, distribución geográfica a gran escala.
  Todos estos rasgos identificadores están orientados a satisfacer los objetivos a nivel de aplicación, y también son compartidos por el paradigma CC. Sin embargo, CC avanza sobre estos principios para proporcionar un entorno hardware virtualizado a diferentes niveles (infraestructura y plataforma) gracias al cuál se puede realizar un aprovisionamiento automático de recursos (*pooling*). Dentro de CC, se habla de *Big Data* [65] en lugar de *Grid Computing* cuando se refiere al análisis de grandes cantidades de información.
- **Utility computing.** Es un modelo de aprovisionamiento de recursos computacionales (procesamiento y almacenamiento) en el que el proveedor del servicio suministra los recursos a medida que los usuarios hacen uso de los mismos [8].
  Dentro del concepto es mucho más relevante los aspectos micro y macroeconómicos, que los tecnológicos en sí. Así, la literatura existente [8, 66, 67] trata de justificar los beneficios, en términos de ahorro de costes, que obtiene una empresa cuando externaliza sus necesidades computacionales, refiriéndose habitualmente al centro de proceso de datos.
  Este modelo desde el punto de vista del mercado sigue una dirección paralela al paradigma CC, salvando la distancia temporal entre ambos. Pero, a nivel tecnológico, *Utility Computing* centra sus esfuerzos en el sector empresarial; mientras que CC abarca también al gran público debido a que no centra sus esfuerzos en la externalización del centro de proceso de datos, sino en cualquier tipo de servicio computacional, con independencia de la escala a la que vaya a ser suministrado.
  No se encuentran, en el estado del arte, plataformas que provean servicios computacionales siguiendo un modelo de utilidad a gran escala, por lo que se puede afirmar, que CC es la implementación a gran escala de los principios propuestos por *Utility Computing*.
- **Autonomic computing** [62, 68, 69]. Aunque no es un término tan conocido como pueden ser los anteriores, resulta clave a nivel computacional dentro del paradigma que nos incumbe. Las directrices de esta tecnología determinaban que los sistemas computacionales fueran capaces de autogestionarse y reaccionar al entorno (interno o externo) sin la necesidad de la intervención humana. Aunque esta tecnología sólo fue desarrolla a nivel teórico, en la actualidad no cabe duda de que ha sentado las bases sobre las que se han construido el paradigma computacional CC.
- **Virtualization Technology** [70]. Es la tecnología que más ha favorecido al rápido desarrollo del paradigma CC. Permite crear abstracciones hardware de forma dinámica y automática en forma de máquinas virtuales sobre el infraestructura disponible [71]. Además, los últimos avances permiten variar los recursos asignados a cada uno de los nodos virtualizados e incluso migrarlos entre diferentes equipos físicos sin tener que detenerlos [72]. Así mismo también permite crear y destruir nodos de una red.

Gracias a estas nuevas y potentes características, la virtualización facilita la gestión de los recursos *hardware*, tarea que en el pasado era altamente compleja, siendo un proceso manual y lento, que habitualmente requería personal altamente especializado. Hay que destacar, en último lugar que gracias a que el usuario final, no tiene acceso a los recursos físicos, sino a la capa superior virtualizada, es posible presentarle una visión ilimitada de los recursos disponibles.

Como única desventaja de esta tecnología, se observa que la gestión del entorno virtualizado requiere de un *Hypervisor*, que es el módulo que gestiona las abstracciones *hardware* en cada nodo físico para lo cuál consume recursos [73]. No obstante, este consumo computacional depende del modelo de virtualización elegido. En los últimos modelos, este sobrecoste no supera el 2% de la potencial computacional del entorno físico [74, 75], aunque para ello requieren hardware dedicado, algo que también está ampliamente implantado gracias a las tecnologías como INTEL-VT y AMD-V [76].

En la actualidad existen varios sistemas de virtualización que utilizan diferentes técnicas para minimizar el consumo de recursos, destacando Xen [71], KVM [77], VMWARE [78], OpenVZ [79], etc.

- **High Availability Computing.** El nacimiento de la tecnología de virtualización [72] y la organización de los centros de procesos de datos en clúster de tipo HPC [80], ha propiciado el desarrollo de nuevas técnicas de alta disponibilidad [81], que van mucho más allá de las vetustas técnicas basadas en la replicación de elementos *hardware* [82].

  CC toma los principios de esta tecnología en tanto en cuanto los servicios computacionales ofrecidos se tienen que ofrecer sin interrupciones. Sin embargo, CC avanza en este sentido, ya que los parámetros de calidad de los servicios también deben ser constantes con independencia en la demanda de los mismos [83], asociados a un SLA.
- **Service Oriented Architecture**, **Service Flow** y **Workflow**. Como ya se ha comentado, el nacimiento de los servicios web [58] y las Arquitecturas Orientadas a Servicios (SOA) [20] han contribuido al desarrollo del paradigma CC, especialmente en la capa de los servicios de plataforma, en la que se exponen un conjunto de APIs (*Application Programming Interface*) a los programadores cuya funcionalidad es muy heterogénea.

  La investigación en CC en este sentido, es la misma que la que se venía desarrollando en el paradigma anterior (SOA) y está relacionada con la composición de servicios de forma automatizada [84], tema en el que se ha venido trabajando recurrentemente en los últimos años [85].

Finalmente, el desarrollo de la Web 2.0, Internet del Futuro [86] y las nuevas técnicas de programación Web enriquecida [87] han contribuido también en gran medida al notable crecimiento del paradigma CC. Estas tecnologías han evolucionado enormemente en los últimos años [88], permitiendo realizar tareas que antes estaban limitadas al *software* tradicional de escritorio. Esto ha facilitado que el uso de los productos *software* ofertados como servicio lleguen al gran público, lo que sin duda contribuye en gran medida a que las grandes compañías apuesten por el desarrollo del paradigma.

En la Tabla 1 se presenta un cuadro resumen de cada una de las tecnologías y su aportación (a nivel hardware, software o Negocio), así como una descripción de la influencia en el desarrollo de paradigma Cloud.

	Hard.	**Soft.**	**Neg.**	**Influencia**
Grid Computing	X			Computación de altas prestaciones al servicio de los usuarios finales
Utility Computing			X	Origen del modelo de negocio del paradigma CC, basado en la provisión de servicio computacionales bajo demanda
Autonomic Computing		X		Define las bases de la auto-monitorización y autocontrol del entorno computacional
Virtualization	X			Entorno *software* subyacente que hace posible la provisión rápida de servicios.
High Availability	X	X		Agrupamiento de centros de datos en clúster y técnicas de alta disponibilidad
SOA, Services flow		X		Modelo de provisión de servicios computacionales a nivel *software/hardware*.

Tabla 1.- Relación Cloud Computing y tecnologías asociadas

2.3.Cloud Computing hoy en día: Definición

El concepto CC se ha ido afianzando tanto en el plano empresarial, como en el de la investigación. Por lo que han ido surgiendo una gran variedad de definiciones [7, 11, 89, 90] que se presentarán a continuación. En cada una de ellas, los autores tratan de resaltar aquellas características que a su juicio son más relevantes. Al analizar un amplio conjunto de trabajos realizados sobre el tema, se puede distinguir dos grandes grupos:

a) Aquellos que centran sus intereses en la definición de los aspectos tecnológicos, diferenciando a su vez a los autores que enfocan la definición a las características de la infraestructura subyacente y los que se enfocan en los servicios que se ofrecen. Es decir, a las características a nivel interno y externo, respectivamente.
b) Aquellos cuyo interés es destacar los aspectos relacionados con el modelo de negocio que intrínsecamente está asociado con el entorno CC.

A lo largo de este apartado, se presentarán de forma crítica las definiciones que más interés han despertado por parte de la comunidad científica y empresarial, con el objetivo de acotar claramente el término, analizando y discerniendo qué se entiende hoy en día por CC.

En el año 2007, junto con el nacimiento de esta tecnología, muchos autores y directivos de importantes compañías tecnológicas dieron su particular visión sobre CC. La mayoría de estas definiciones no eran precisas, es más, muchas de ellas ni siquiera se acercaban a lo que hoy en día conocemos como modelo de computación en nube. Había los que indicaban que (i) no era una tecnología novedosa [91], o incluso (ii) confundían este paradigma con tecnologías anteriores (*Grid Computing*, *Utility Computing*, *Autonomic Computing*, etc.)[91], o simplemente en la mayoría de los casos ofrecían una definición parcial, tal como se presentará a continuación. Tal era el desconcierto sobre esta nueva tecnológica, que algunos como Irving Wladawsky-Berger (IBM) indicaban en el año 2010 que "*there is a clear consensus that there is no real consensus on what cloud computing is*". Teniendo en cuenta que CC nació en el año 2007 y los grandes esfuerzos que se pusieron en su desarrollo, resulta destacable que, en el año 2010, todavía no hubiera un acuerdo sobre esta tecnología, debido sobre todo a que era difícil observar el cambio en el modelo computacional, y más aún en el modelo de prestación de servicios que conlleva la adopción de este paradigma tecnológico.

Más allá, de las definiciones dadas a nivel empresarial, a nivel académico las primeras definiciones formales sobre CC empiezan a surgir en el año 2008. Así, una de las primera definiciones fue propuesta por Want *et al.* [11], la cuál estaba claramente enfocada a las características tecnológicas del paradigma:

> *A computing Cloud is a set of networks enabled services, providing scalable, QoS guaranteed, normally personalized, inexpensive computing platform on demand, which could be accessed in a simple and pervarsive way.*

De la definición dada por Wang *et al.* cabe resaltar que los autores ya enfatizaban la idea conceptual de ofrecer servicios escalables y personalizados a través de Internet, asegurando un nivel de calidad mínimo en los mismos. Lo que denota tres características clave del paradigma CC (escalabilidad, personalización y nivel en la calidad del servicio). Estas características, en ningún caso, suponían un reto tecnológico para la época. Sin embargo, la particularidad viene dada por el hecho de que ya se cita que los servicios se exponen bajo demanda siguiendo los principios del paradigma *Utility computing* [8]. Por el contrario, como crítica a la definición destacar que, en primer lugar, (i) aunque intrínsecamente se habla de escalabilidad, lo que pueden asociarse a un concepto más novedoso

como es la elasticidad [13], no se habla de que ésta se gestione de forma automática. Tampoco, se menciona (ii) nada acerca de los posibles tipos de servicios ofertados, es decir, todavía no se mencionaba la posibilidad de ofrecer servicios de infraestructura a nivel *hardware* algo que diferencia este paradigma de las tecnologías anteriores. Además, tampoco se destaca que (iii) el nivel de calidad puede ser establecido mediante un acuerdo SLA, lo que puede implicar implícitamente la escalabilidad automática del sistema para asegurar que se satisfacen los diferentes SLAs acordados. Y, finalmente, y como crítica más importante, (iv) no se hace ninguna mención al modelo de distribución ni comercialización del *hardware* o *software* ofrecido a través de servicios CC, que es una de las bases del paradigma.

Al mismo tiempo, otros autores introducen definiciones más completas, como es el caso de Foster *et al.* [90] que presenta la siguiente definición sobre el paradigma de computación en nube:

> *A large-scalle distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dinamically-scalable, managed computing power, storage, platforms, and service are delivered on demand to external customers over the Internet.*

En este caso se trata de una definición adecuada y precisa a nivel tecnológico. Se enfatiza la importancia de los servicios ofertados a través de Internet dentro del paradigma. También se relaciona explícitamente estos servicios con el entorno *hardware* subyacente, destacando que es un entorno distribuido a gran escala, e introduciendo tecnologías muy novedosas como en su momento era la virtualización, que permite la abstracción de recursos *hardware*, la gestión de la capacidad de cómputo, etc. Se destaca que se habla de gestión dinámica (y, quizás, automática) de los propios servicios. Además, se puede sobreentender que se habla de servicios de plataforma e infraestructura, pero no así de servicios *software*. Finalmente, se hace referencia a un modelo computacional cuyo crecimiento está motivado por la economía de escala pero, sin embargo, no dan ninguna información más sobre su modelo de comercialización.

Por su parte, Vaquero *et al.* [89] introducen la que quizás es una de las definiciones más completas, ya que incluye aspectos tecnológicos y de negocio:

> *Cloud are a large pool of easily usable and accesible virtualized resources (such as hardware, development platforms and/or services). These resources can be dynamically reconfigured to adjust to a variable load (scalable), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized SLA.*

La primera parte de la definición se centra en los aspectos tecnológicos, presentando características que también consideran otros autores (virtualización, escalabilidad, etc.), y adicionalmente destaca que los recursos computacionales que se ofrecen en forma de servicio pueden ser de cualquier tipo y que, además, tienen que ser fácilmente accesibles a través de Internet. Esta definición es de las primeras que introduce la necesidad de reajuste automático del sistema en función de la demanda, lo que asegura un óptimo uso de los recursos *hardware*. Es decir, estos autores presentan las características que diferencian al paradigma CC a nivel tecnológico de los modelos anteriores, la elasticidad de los servicios, y por consiguiente, la necesidad de reajustar los recursos en función de la demanda de los mismos consiguiendo, por tanto, un uso óptimo de la infraestructura *hardware*. En la segunda parte de la definición, se indica que los recursos son explotados a través de un modelo de comercialización basado en el establecimiento de un acuerdo personalizado de la calidad del mismo (SLA) y que a su vez permite fijar el precio del servicio. Es decir, debido a este acuerdo, los proveedores deben asegurar un nivel de calidad mínimo en el servicio [34], o QoS. Este hecho,

ha provocado, que los servicios, no se ofrezcan de forma tradicional, sino siguiendo un modelo de pago por uso, conducido por el SLA alcanzado previamente, que establece formalmente el precio del producto y la calidad mínima del servicio. En esta definición, pese a que habla de diferentes tipos recursos/servicios ofertados, parece que los autores enfocan la definición tan sólo a los aspectos relacionados con la infraestructura *hardware* y plataforma en parte, debido al auge que la tecnología *Grid computing* [61, 90] tenía en aquel momento. Además, también parece que no se le atribuye una gran importancia los servicios en las capas superiores del paradigma, es decir, al *software* ofrecido como servicio.

En contraposición, para otros autores las capas superiores constituyen la clave que ha permitido que este paradigma haya adquirido la popularidad que hoy en día atesora [64, 92]. Además, la realidad hoy en día muestra que una gran cantidad de plataformas CC se orientan únicamente a ofrecer servicios dentro de las capas superiores del paradigma. En este sentido, existen incluso autores que centran su definición en este tipo de servicio. Así, por ejemplo Bragg [93] indica que *the key concept behind the Cloud is the Web application - software stored on some company's server and accessed by users through their browsers. That means everything is done through the same window used to surf Web sites. Most of the software used and the data produced remains in the Cloud, i.e., on the main servers of the company.*

Finalmente, a finales del año 2011, el NIST (*National Institute of Standards and Technology*) americano [7], propone la definición que a nuestro juicio es la más acertada desde un punto de vista técnico y funcional. Aunque, en ella no se presta especial atención a la economía de escala que ha promovido el desarrollo incesante de este modelo computacional.

> *Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.*

En este caso los autores, además de referirse a cualquier tipo de recurso computacional, indican un concepto clave: la reducción al mínimo el esfuerzo de obtener los servicios proporcionados; algo que, sin duda, ha sido clave en el gran auge que ha tenido está tecnología.

Como complemento, esta definición incluye 5 características obligatorias que cualquier entorno CC debería poseer:

- **Servicios automáticos bajo demanda**. Se refiriere a que los servicios, con independencia de su tipo, se tienen que proporcionar automáticamente y sin interacción humana en función de la demanda de los usuarios.
  Estas características de simplicidad y automatización, ya la habían apuntado numerosos autores de forma previa, como se ha visto a lo largo de este trabajo. Como crítica, no se habla de la necesidad de alcanzar un acuerdo en la calidad del servicio, lo que quizás constituye una de las mayores debilidades en la definición.
- **Disponibilidad de servicios a través de la red.** Los clientes deben acceder a los servicios a través de la red, y por lo tanto, los proveedores tienen que utilizar este medio para proporcionar sus servicios.
- **Disponibilidad de recursos.** El proveedor debe disponer de la capacidad de ofrecer los servicios con independencia de la demanda de éstos, utilizando recursos *hardware* físicos o virtuales asignados dinámicamente a cada servicio y reasignados en función de la demanda. Esta característica implica que el usuario no conoce dónde está su información en cada

momento, ya que existe un principio de no localidad (o deslocalización), donde los datos se trasladan entre los recursos virtuales (o no) que tienen asignados en cada momento.
En este sentido, existen autores como [5, 12] que hablan directamente de servicios de alta disponibilidad, siguiendo un modelo de computación de alta disponibilidad.

- **Elasticidad.** Los diferentes recursos se deben proporcionar de forma elástica e incluso automática en función de la demanda.
  Esta característica, está directamente ligada con la anterior, la alta disponibilidad de los servicios, lo que implica que el proveedor debe asignar los recursos en función de la demanda de los servicios, algo que ya se venía haciendo en tecnologías anteriores. Sin embargo, CC ofrece servicios elásticos, esto significa que los recursos asignados a cada servicio varían automáticamente en función de la demanda, lo que permite proporcionar más recursos cuando más demanda existe, pero también reducir recursos cuando disminuye la demanda. Todo este proceso se debería realizar de forma automática sin que el cliente de los servicios tenga conocimiento de la readaptación interna. Gracias a esta característica, los usuarios tienen la sensación de interactuar ante un sistema con recursos ilimitados, aunque en realidad no sea así.
  La existencia de niveles de acuerdo en cuanto a la calidad del servicio, introduce una complejidad mayor en la gestión de la elasticidad, ya que no sólo es necesario proporcionar mayores recursos a más demanda en los servicios, sino que el tiempo necesario para la readaptación está limitado en base a la calidad acordada.
- **Servicios a medida.** Los servicios proporcionados por un sistema CC deben estar totalmente monitorizados, y su control tiene que realizarse de forma automática.
  Gracias a ello, se puede optimizar el uso de recursos; en este sentido, sólo teniendo un conocimiento total sobre el propio sistema, es posible ofrecer un servicio de pago por uso eficiente y competitivo. Permitiendo ofrecer una transparencia total de la información entre el proveedor del servicio y el usuario del mismo.

Estas características, junto a la definición previa, acotan en gran medida el concepto que hoy se tiene de un entorno de CC y de los servicios que éste ofrece, los cuáles se conocen habitualmente como *Cloud Services* [94]. No obstante, casi parece lógico, teniendo en cuenta que la última versión de la definición es relativamente reciente (septiembre del año 2011). Lo que ha provocado que se hayan ido eliminando los vestigios de tecnologías predecesoras a CC; aunque en algunos casos, estén relacionadas con éste nuevo paradigma (*Utility computing*, *Service Computing*, *Data Centers*, *Market-oriented computing*, etc.), en otros casos el concepto se ha ido alejando (*Grid Computing*, *P2P Computing*, etc.).

Sin embargo, como debilidades en la definición, en cuanto a las características *hardware*, destacar que por un lado, el NIST menciona la elasticidad dinámica, incluyendo como rasgo opcional la automatización de esta característica. Desde nuestro punto de vista, y el de algunos autores [12], esta es una de las características clave del entorno CC, siendo objetivo de estudio a lo largo de este trabajo. Y por otro lado, según [12, 95] también hablan del principio de "multi-tenencia" (*multitenant*), el cuál se refiere a que pueden existir varios entornos CC en un mismo centro de proceso de datos, lo que obliga a una clara división de responsabilidad en la gestión del entorno.

La definición del NIST sin duda es la más acertada y extensa, aunque podría complementarse incluyendo explícitamente el concepto de SLA, algo que sin duda ayudaría a completar todos los aspectos comerciales y negocios del paradigma [34]; dado que además existe una gran cantidad de estudios centrados en este campo. Finalmente, la definición y características de CC propuestas por el NIST, pueden complementarse mediante el conjunto de características propuestas por [12]: *Multi-*

tenancy, s*hared resource pooling*, *geo-distribution and ubiquitous network access*, *service oriented*, *dynamic resource provisioning*, *self-organizing* y *utility-based pricing*. Dentro de estas características, cabe destacar la última, que es una propiedad estrechamente relacionada con los aspectos económicos que envuelven al paradigma. El precio del servicio está basado en su utilidad, es decir, dado que CC emplea un modelo de pago por uso, el precio concreto de un servicio, depende del propio servicio y del uso que el cliente hace de este, basándose en la teoría clásica de la regulación (oferta y demanda) [96]. Lo que a priori logra reducir el coste operacional debido a:

- No existe inversión inicial, ya que al implementar un modelo de negocio de pago por uso, desde el punto de vista del usuario de servicios CC no es necesario realizar una inversión previa en infraestructura, sino que se alquilan un conjunto de servicios acordes a las necesidades en cada momento.
- Reducción de los riesgos comerciales y los costes de mantenimiento. Dado que siguiendo, el modelo CC, los recursos computacionales no se encuentran ligados a la empresa, los costes derivados de su mantenimiento, actualización, fallos, desconexión por actualización, etc. no constituyen un riesgo empresarial para la empresa.

A lo largo de este apartado se han presentado diferentes definiciones sobre CC que denotan la evolución del paradigma tecnológico desde su nacimiento en 2007. En primer lugar destaca el interés que existe tanto en la comunidad científica, como en el entorno empresarial acerca de este paradigma computacional. Así mismo, también es reseñable la estrecha relación que existe con tecnologías previas, así como que la unión de todas ellas en una plataforma única, lo que da como resultado un entorno tecnológico superior. Finalmente, la clave en el desarrollo y la rápida introducción en el mercado ha sido el modelo de comercialización basados en acuerdos específicos de uso, que resultan favorables tanto para los proveedores, como para los consumidores del servicio.

En el modelo de negocio, las tecnologías implicadas y los servicios ofertados, hacen que un sistema CC sea complejo y en algunos aspectos difícil de comprender. Por lo que en los siguientes apartados se presentará con más detalle los tipos de servicios ofertados, los modelos de despliegue posibles y roles de usuario implicados en el modelo de comercialización.

2.4.Capacidades ofertadas, Something as a Service

Como se indicó anteriormente, en la definición del NIST [7] se proponen tres tipos de servicios y cuatro modelos de despliegue, ampliamente conocidos. Se comenzará describiendo los diferentes servicios, en este sentido dado que se ofrece casi cualquier tipo de servicio computacional, ya sea *hardware* o *software* y siempre a través de Internet, es habitual hablar de *Something as a Service* [97, 98]. No obstante, antes de enumerarlos, resulta clave indicar que en la propia definición se habla de *capabilities* o capacidades, en lugar de servicios. Este concepto es quizás mucho más intuitivo, alejando estos servicios computacionales del concepto tradicional de servicio *web* en las arquitecturas anteriores, como por ejemplo SOA [99].

Según el NIST, los principales modelos de servicios ofrecidos son los siguientes:

- **SaaS. Software as a Service.** Este tipo de capacidad permite al proveedor, ofrecer al consumidor sus aplicaciones. De forma que éstas se ejecutan directamente en la infraestructura en nube, lo que conlleva una gran cantidad de ventajas como son la ubicuidad de las aplicaciones o el uso de clientes ligeros. Sin embargo, también lleva asociadas un conjunto de dificultades directamente relacionadas con que el hecho de que el consumidor pierde el control sobre la infraestructura (red, almacenamiento, sistema operativo, dificultad de configuración, etc.), aunque quizás desde un punto de vista más moderno, pueden ser incluso consideradas como fortalezas.
  Dentro de este tipo de servicio, algunos autores como Lenk *et al.* [100] proponen una división en *Applications* y *Applications Services*, dado que existen numerosos trabajos relacionados con la composición de servicios de aplicación en entornos CC [101, 102].
- **PaaS. Platform as a Service.** Este tipo de capacidades proporcionadas por los proveedores, permiten al consumidor disponer de las herramientas necesarias para crear sus propias aplicaciones. Entre estos servicios se encuentran entornos de programación, librerías, herramientas, etc. De la misma manera que los servicios de la capa anterior, el programador no controla la infraestructura subyacente, ni el entorno de despliegue de sus aplicaciones.
  Este nivel puede subdividirse según Lenk *et al.* [100] en *Programming environment* y *Execution environment*, separando de este modo los servicios ofrecidos dentro del entorno de despliegue del paradigma de computación sobre el que son desarrollados.
- **IaaS. Infrastructure as a Service.** También denominado *Hardware as a Service* (HaaS) [11]. Es el tipo de capacidad que se proporciona al consumidor es *hardware*, es decir, procesamiento, almacenamiento, red, etc. Al igual que en servicios anteriores, el consumidor o usuario no tiene el control sobre la infraestructura subyacente, en este caso, sistemas operativos nativos, características de red, etc. Aunque sí que puede tener la configuración sobre el entorno *hardware* virtualizado que se le presenta.
  Dentro de este tipo de servicio, Lenk *et al.* [100] separa el entorno hardware subyacente de los servicios de infraestructura ofertados:
  - *Resource Set*, representan los recursos computacionales que posee el entorno CC, es decir, el conjunto de computadores físicos. Constituye la visión interna del entorno de computación, encargado de ofrecer potencia computacional a las capacidades que ofrece el entorno CC a nivel externo. Su objetivo es facilitar la autogestión de los recursos, proporcionando una visión externa de recursos ilimitados. La gestión de esta capa debe realizarse de forma dinámica (y preferiblemente automática), optimizando el uso de los recursos, tal y como indicaba Vaquero *et al.* [89]. En este

sentido, en la actualidad, existen diferentes investigaciones centradas en maximizar la eficiencia de este capa subyacente a todo el entorno CC [103, 104].
Esta capa, a su vez es dividida en dos subgrupos:

- *Physical Resource Set* (o simplemente *Hardware* para [12]), que es el propio *hardware* físico que proporciona los recursos computacionales al sistema.
- *Virtual Resource Set*, habitualmente construida sobre una tecnología de virtualización subyacente. También es posible encontrar algún entorno CC que no utiliza virtualización, pero son casos aislados.

Esta división entre recursos físicos y virtuales tiene como principal ventaja el incremento en la capacidad de la gestión del entorno. Teniendo en cuenta que esta gestión tiene que ser realizada de forma dinámica y en la mayoría de los casos automática, resulta fundamental, además de ser uno de los aspectos diferenciadores de la tecnología CC. Sin embargo, también existen debilidades, la gestión de un entorno virtualizado requiere el uso de *Hypervisores* que gestionen y encapsulen su complejidad intrínseca, lo que induce una carga adicional de recursos computacionales [73].

- *Infraestructure Services*, divididos a su vez en (i) *Basic infrastructure services* donde se engloban los servicios computacionales (almacenamiento, procesamiento, red, etc.); y (ii) *Higher infrastructure service* (o *Data as a Service* (DaaS) según Want *et al.* [11]), que agrupa servicios estructurados de almacenamiento de información. Este segundo grupo es denominado
  - *Basic infrastructure services*, englobando en este grupo servicios computacionales (almacenamiento, procesamiento, red, etc.).
  - *Higher infrastructure Service*, englobando en este grupo servicios estructurados de almacenamiento de información. Es decir, en la práctica, las bases de datos.

Finalmente, con respecto a las capacidades proporcionadas, añadir que algunos autores [100] introducen el concepto de *Human as a Service* (HuaaS), dentro de esta taxonomía de modelos de servicios que ofrecen las plataformas CC. Dentro esta categoría se enmarcan los servicios de recolección de información de forma masiva procedente de las bien conocidas plataformas sociales (Facebook, Twitter, etc.). Esta información es utilizada, principalmente, como complemento de otros servicios.

2.4.1. Tipos de usuarios y modelos de despliegue

La definición propuesta en el NIST, presentada en el apartado 2.3, se complementa con 3 tipos de servicios y 4 modelos de despliegue que se describirán a continuación. De forma previa a enumerar los entornos de despliegue, resulta conveniente hacer referencia a los actores y roles que intervienen en el modelo de negocio CC. Así pues, en este apartado se presentarán en primer lugar los diferentes roles que intervienen en el paradigma CC, teniendo en cuenta las diferentes perspectivas existentes. Y, a continuación, se presentarán los modelos de despliegue propuestos por el NIST.

En la literatura es posible encontrar varios trabajos sobre los roles o tipos de usuarios de un entorno CC [5, 6, 12, 105, 106]. Al analizar estos estudios, se observan dos vertientes que se presentan a continuación.

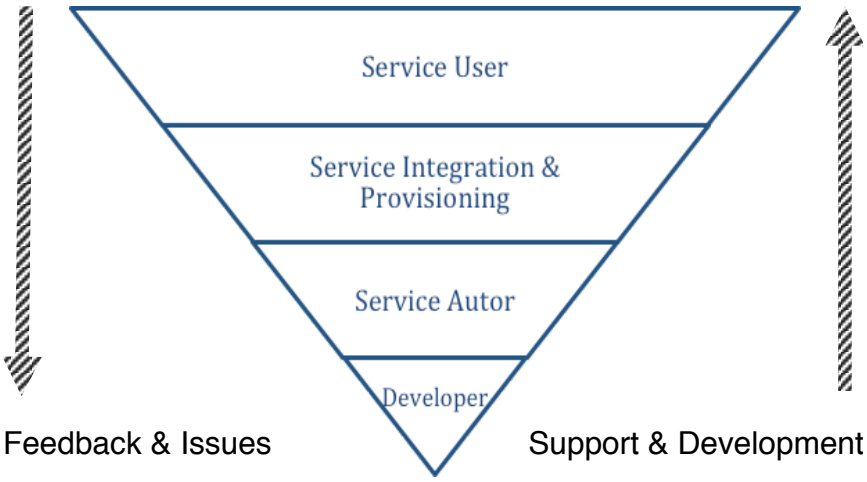

Figura 1.- Usuarios en Cloud Computing [(adaptada de Vouk [105])

Por un lado, a nivel técnico, siguiendo a Vouk [105] se definen cuatro categorías de actores no excluyentes (Figura 1):

- **Service User.** Son los consumidores de servicios del entorno CC, su rol es el más importante y del que dependen el resto de los usuarios.
- **Service Integration & Provisioning.** Expertos en composición de servicios, cuyo objetivo es satisfacer la demanda de un cliente concreto.
- **Service Author.** Responsables del desarrollo de servicios que puedan ser utilizados de forma individual, o como parte de otros.
- **Cyberinfrastructure developer.** Responsables del desarrollo y mantenimiento de la infraestructura.

Por otro lado, existe otra clasificación de roles que está más aceptada y constituye un estándar de facto. En ella, no sólo se tiene en cuenta la perspectiva técnica del despliegue *hardware* y *software*, sino también los intereses económicos que motivan el desarrollo del paradigma. Dentro de estos roles, se distinguen los siguientes, no siendo excluyentes ninguno de ellos:

- **Cloud provider.** Es el rol que proporciona capacidades CC. Normalmente, dentro de esta categoría están los servicios de tipo IaaS y PaaS. Fue propuesto inicialmente por Armbrust *et al.* [6]. A su vez puede subdividirse en:
  - *Infrastructure Providers.* Identifica al rol que proporciona servicios hardware, es decir, IaaS. Propuesto también por Vaquero [89].
  - *Service Provider.* Rol que proporciona servicios de tipo software en el modelo CC. Puede convertirse de forma simultánea en *Cloud User* si a su vez utiliza otros servicios de tipo SaaS, mediante la composición de servicios a este nivel. Propuesto por [12], aunque Armbrust *et al.*, [6] lo denomina *SaaS Provider.*
- **Cloud User.** Este rol hace uso de las capacidades ofrecidas por el rol *Cloud provider*, con referencia a los niveles PaaS e IaaS. Fue propuesto por Armbrust *et al.* [6], denominado User/Broker en Buyya *et al.* [5].
- **SaaS User/End User.** Es el usuario final de las aplicaciones SaaS.

Además, [5] propone la existencia de un actor adicional, *SLA Resource Allocator*, entidad intermedia entre el entorno CC y los usuarios externos, con las funciones de monitorización de peticiones, control de admisión, gestión de precios, gestión de cuentas y monitor de infraestructura.

En el marco de investigación del NIST [7], se propone una división mucho más sencilla, en 6 actores que modelan las interacciones tanto a nivel de mercado, como técnico, integrando por tanto las dos perspectivas anteriores. Así pues, por un lado, se proponen tres actores desde un punto de vista de mercado: (i) *Customer*, que es un usuario u organización que utiliza las capacidades que ofrece un entorno CC; (ii) *Client*, que es una agente artificial que accede al entorno CC para utilizar y desplegar sus servicios y ofrecer productos al anterior rol; y finalmente (iii) *Provider*, que es una organización que proporciona las capacidades CC. Por otro lado, otros tres actores que modelan la relación entre el consumidor y el proveedor desde un punto de vista técnico: (iv) *Broker* que es el intermediario que gestiona la relación comercial entre proveedor y consumidor; (v) *Carrier*, que proporciona la conexión desde el proveedor y el consumidor; (vi) *Auditor*, que evalúa de forma independiente los servicios ofrecidos en cuanto a rendimiento y seguridad.

Una vez que se han presentado los diferentes roles, el último paso consiste en identificar los diferentes modelos de despliegue que existen, comenzando por los públicos y privados:

- **Private Cloud.** La infraestructura CC es utilizada por una única organización, que a su vez puede incluir diferentes consumidores. Esta infraestructura puede pertenecer a la propia organización, o ser ofrecida por un tercero.
- **Public Cloud.** Este tipo de infraestructuras se proporcionan para el uso abierto del público en general. Están alojadas por organizaciones, universidades, gobiernos o una combinación de ellos. Obligatoriamente debe existir el actor proveedor.

A partir de estos dos modelos de despliegue propuestos por el NIST, [12, 107] proponen la existencia de un modelo híbrido denominado *Virtual Private Cloud*, también denominado *Outsources Private Cloud* por Liu *et al.* [10] en el que se combinan entornos CC públicos y privados. De forma que se construye un entorno CC privado sobre las capacidades de infraestructura a modo de entorno subyacente que proporciona un entorno *public Cloud*. [12] define este tipo de despliegue como una plataforma situada por encima de los sistemas CC públicos, siendo la principal diferencia el hecho de que se permite el diseño de una topología específica y un modelo de seguridad específico, aunque en contrapartida el entorno CC esté deslocalizado.

Como consecuencia de la existencia de *Virtual Private Cloud*, se puede dar la situación de que varios entornos CC, que aún perteneciendo a organizaciones diferentes tengan que compartir la misma infraestructura, es decir CC público. En este caso se estaría hablando de *multitenencia* [7, 12, 108]. La investigación en los últimos tiempos también se ha dirigido al desarrollo de modelos que hagan posible esta convivencia, principalmente a nivel SaaS [109, 110], y a nivel PaaS [111, 112].

Finalmente, el NIST, también define otros dos modelos de despliegue, el de comunidad y el híbrido:

- **Community cloud.** La infraestructura del sistema CC es utilizada por un grupo de consumidores u organizaciones específicos que comparten intereses relacionados. La infraestructura puede estar alojada por una o varias organizaciones dentro de los grupos de interés (on-*site Comunity Cloud*), o incluso, puede pertenecer a un tercero (*out-source comunity cloud*).
- **Hybrid Cloud.** Este tipo de infraestructuras es la combinación de alguna de las anteriores, cada una de las cuáles tiene una entidad propia, pero trabajando de forma coordinada. La Figura 2 muestra un ejemplo de un sistema CC híbrido que se ha construido mediante la agregación de 5 variantes.

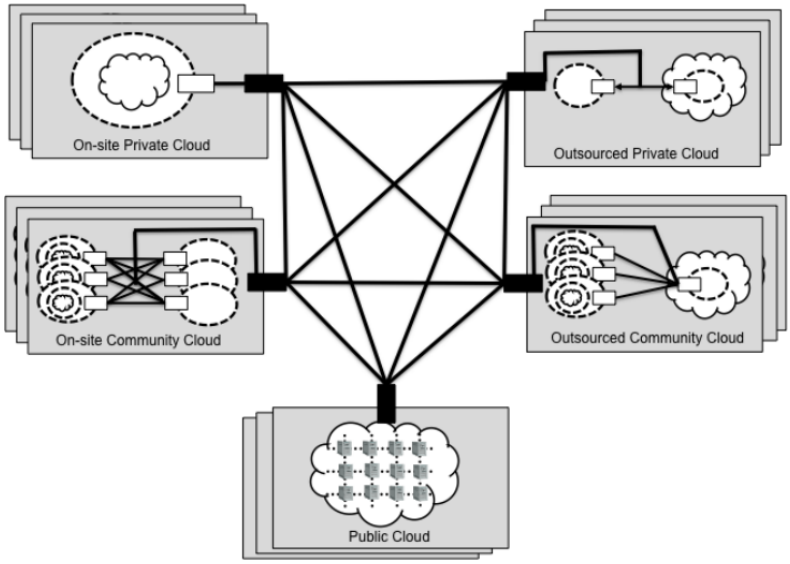

Figura 2.- Cloud híbrido [10]

Cuando se habla de *Hybrid clouds* es necesario hablar de interoperabilidad entre plataformas [113], algo que también puede hacerse extensible al resto de despliegues CC. Un usuario de este tipo de entornos habitualmente trabaja con diferentes plataformas al mismo tiempo, cada una de ellas tiene su propia interfaz de comunicación, habitualmente en forma de API, lo que las hace cerradas y dificulta su interoperabilidad con el resto.

La dificultad en la interoperabilidad en cierto modo está basada en la falta de madurez de la tecnología (recordemos que ésta nació en 2007). Sin embargo, también está motivada por los intereses que tienen las grandes empresas tecnológicas en imponer su plataforma CC sobre sus competidores.

En cuanto a la interoperabilidad, diferentes autores han propuesto dos clasificaciones diferentes atendiendo al nivel de portabilidad, aunque como se puede observar ambas tienen un gran paralelismo:

- Celesti *et al.* [114], habla de tres fases de ejecución para la consecución del objetivo de federación: (i) la fase monolítica que es aquella en la que los servicios son independientes y propietarios; (ii) la fase de distribución vertical, en la que los proveedores de servicios utilizan los servicios proporcionados por proveedores terceros, para proporcionar sus propios servicios a los consumidores; y, por último, (iii) la federación horizontal, en el que todos los proveedores con independencia de su tamaño disponen de servicios federados que ofrecen al consumidor siguiendo un modelo de economía en escala.
  Hoy en día nos encontramos a medio camino entre la fase 1 y la fase 2, donde la mayoría de sistema CC, sobre todo aquellos proporcionados por las grandes compañías, son monolíticos (fase 1); pero al mismo tiempo, también se observa una gran cantidad de *Virtual Private Clouds* [10] que están directamente relacionados con la fase 2.
- Machado *et al.* [115], considera tres tipos diferentes de escenarios. Aquel en el que (i) un usuario debe portar la configuración de sus servicios entre dos entornos CC diferentes. Otro sería en el que (ii) los usuarios tienen contratados servicios con diferentes servicios proveedores y desean que todos ellos trabajen de forma conjunta. Y finalmente, el último es el denominado (iii) CC federado, donde se ofrecen servicios al usuario final haciendo uso de los recursos de varios entornos CC, para lo que es necesario utilizar herramientas de

coordinación entre ellos. La Figura 3 muestra una contextualización de este último escenario.

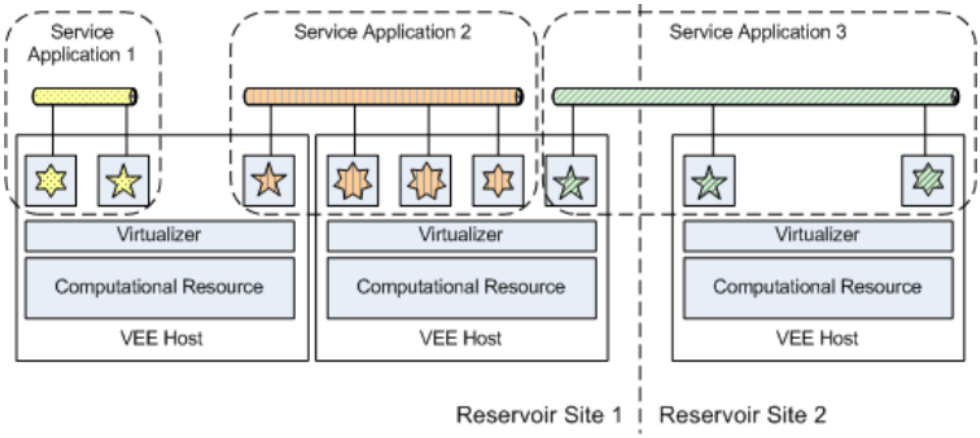

Figura 3.- Arquitectura en el proyecto Reservoir [116]

En la actualidad, para coordinar la investigación en el ámbito de la interoperabilidad, especialmente en la capa IaaS y la interoperabilidad entre sistemas CC [115, 117, 118], comienzan aparecer las principales iniciativas y organizaciones que pretenden desarrollar y facilitar la interoperabilidad entre plataforma, donde se destaca:

- *DMTF's Open Cloud Standard Incubator (OCSI)*[3], que se centra en el desarrollo de estándares de interacción entre proveedores y usuarios.
- *Open Cloud Computing Interface Working Group (OCCI-WG)*[4], que desarrolla especificaciones para la interoperabilidad de servicios de infraestructura.
- *Cloud Data Management Interface (CDMI)*[5], que trata de especificar funcionalmente como las aplicaciones crean, recuperan, actualizan y eliminan datos desde un entorno CC.
- *Global Inter-Cloud Technology Forum (GICTF)*[6], centro en la interoperabilidad entre plataformas.

[3] http://dmtf.org/standards/cloud
[4] http://occi-wg.org/
[5] http://www.snia.org/tech_activities/standards/curr_standards/cdmi
[6] http://www.gictf.jp/

2.5.Modelos de referencia: plataformas y arquitecturas

No cabe duda de que un sistema CC es complejo y forma parte de un entorno abierto en el que se conjugan diferentes tecnologías, usuarios e intereses económicos para dar lugar a un nuevo modelo computacional que ha revolucionado la forma de ofertar servicios a través de Internet. Para que todos estos aspectos trabajen de forma coordinada con el objetivo de lograr unos objetivos comunes se han desarrollado arquitecturas complejas, que además han teniendo que hacer frente a las limitaciones técnicas existentes.

Hasta hace no mucho tiempo las grandes compañías como IBM, Intel, Google, Amazon, etc. no mostraban información relevante sobre las arquitecturas de sus respectivos modelos CC. De hecho, hoy en día el ocultismo en este aspecto sigue siendo la tónica general ya que, aunque muchas de ellas han hecho públicas las directrices generales, en ningún caso se han liberado algoritmos u otros aspectos técnicos o computacionales que permitan el avance de la tecnología sobre la base existente. Así por ejemplo, la Figura 4 y la Figura 5 presentan la pila tecnológica de sistemas CC propuestas por Intel [119] y Cisco [120], respectivamente. En ellas se pueden observar los componentes básicos de cualquier sistema CC como infraestructura subyacente (servidores, equipamiento de red, almacenamiento, etc.), así como componentes software (entorno de virtualización, gestión, monitorización, seguridad, etc.). Sin embargo, ninguna de estas compañías es un gran proveedor de servicios CC, por lo que ambas utilizan esta información para ofrecer sus productos a los verdaderos proveedores CC, los cuáles no publican sus arquitecturas CC.

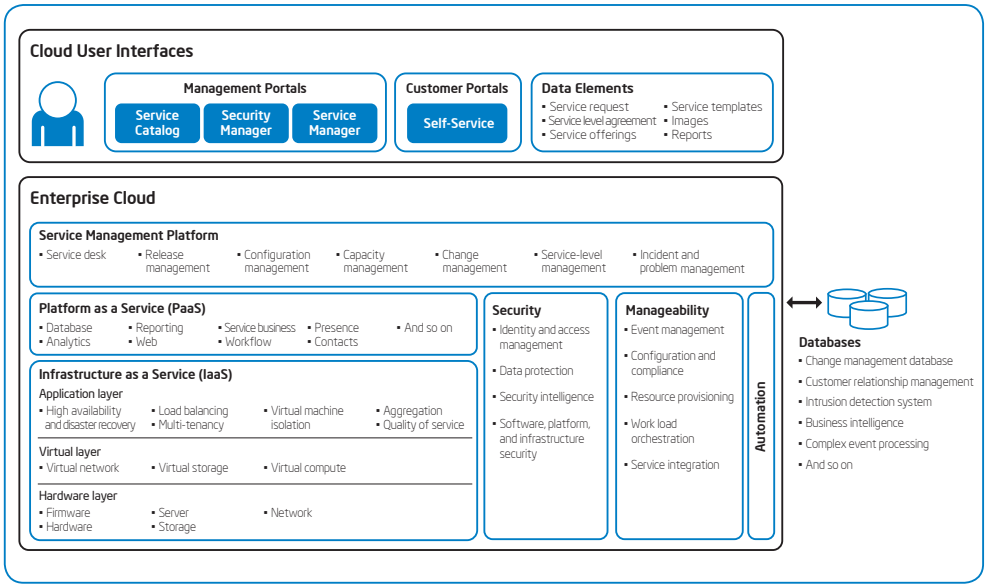

Figura 4.- Arquitectura Intel [119]

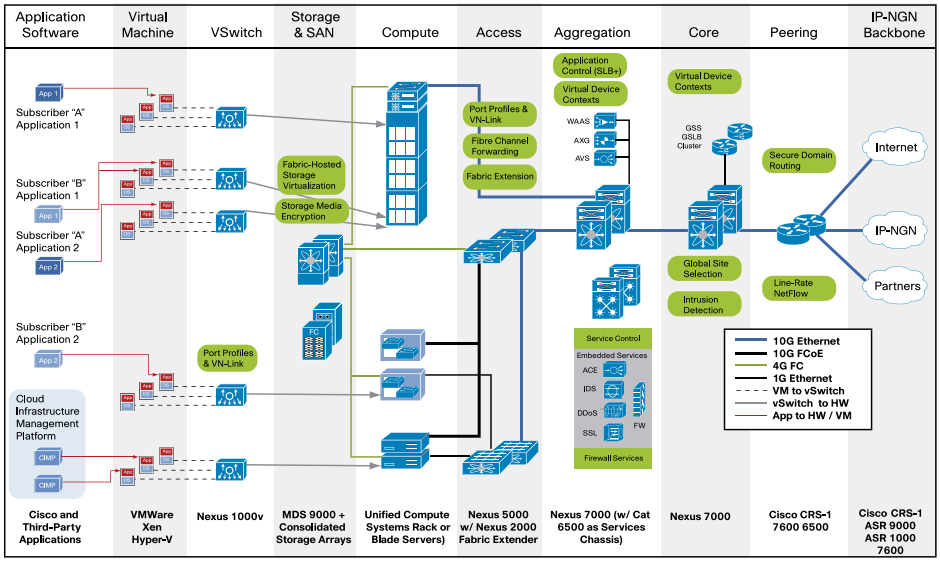

Figura 5.- Modelo de Arquitectura para Cisco [120]

Siguiendo el trabajo realizado por el NIST, este organismo también propone una arquitectura de referencia [10], que trata de describir y acotar correctamente todos los aspectos a tener en cuenta en un entorno CC. Esta arquitectura se presentará a continuación. No obstante, resulta interesante presentar de forma previa la clasificación que propone Tianfield [121] respecto a las arquitecturas CC:

- La **arquitectura de plataforma** es aquella que tiene en cuenta el entorno tecnológico subyacente al paradigma CC
- La **arquitectura de aplicación** es aquella que encapsula la gestión de la calidad de los servicios que se le ofrecen a los usuarios finales.

El modelo de referencia de arquitectura CC propuesto por el NIST [10] aúna ambos modelos (según Tianfield [121]), es decir, tecnología de plataforma y nivel de calidad hacia el usuario final, tal y como se muestra en la Figura 6.

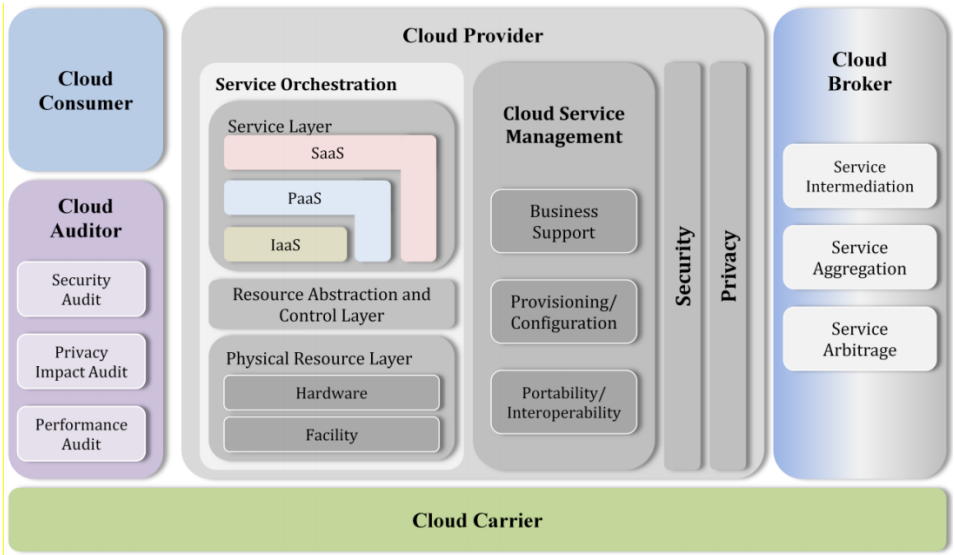

Figura 6.- Arquitectura Cloud de referencia para el NIST [10]

La arquitectura que se presenta en la Figura 6 tiene en cuenta no sólo los componentes tecnológicos, sino también los roles que participan dentro del paradigma CC. Así, en función de cada uno de los roles del sistema, tenemos los siguientes módulos o componentes:

- **Cloud Provider.** Es el rol encargado de la (i) Coordinación de Servicios, proporcionando éstos a terceros, con independencia de su tipo (IaaS, PaaS o SaaS), por lo que necesita gestionar la infraestructura subyacente (física y virtual) a través de una capa de control. Así mismo, también debe (ii) facilitar la gestión de esos servicios ofertados a través de una capa de soporte a la comercialización y negocio que facilite la monitorización acerca del uso de esos servicios (características de calidad y acuerdos en el nivel del servicio) y su cobro en función del uso que se realice (facturación, consumos, etc.), así mismo también deberá proporcionar características de portabilidad e interoperabilidad entre plataformas, y un modelo que permita la rápida configuración y la gestión del cambio de aprovisionamiento de los recursos. Finalmente, también tendrá que hacerse cargo de la (iii) seguridad y (iv) privacidad, que son características clave para que un *Cloud Provider* sea aceptado por los consumidores de sus servicios.
- **Cloud Auditor.** Es un agente externo que capaz de monitorizar el servicio con el objetivo de validar si se están cumpliendo los requisitos acordados mediante SLA de seguridad, privacidad y rendimiento.
- **Cloud Broker.** Es un agente externo que actúa como intermediario entre consumidores y proveedores con el objetivo de buscar y proporcionar los servicios más adecuados a los objetivos de los consumidores con independencia del proveedor. Sus labores principales son la (i) intermediación, encargada de buscar y encontrar los mejores servicios y validar que efectivamente se están cumpliendo los valores de calidad acordados; (ii) agregación, proporcionando la capacidad de integrar servicios de diferentes entornos CC; y, finalmente (iii) arbitraje, proporcionando un entorno similar a la agregación de servicios, pero en el que estos servicios no están predefinidos y cambian dinámicamente.
- **Cloud Carrier.** Que es el agente que proporciona conectividad entre el proveedor y el consumidor. Resulta importante, ya que muchos de los objetivos acordados en el SLA dependen de que el Cloud Carrier sea capaz de proveer con la tecnología para transportar la información con la suficiente rapidez.

No obstante, más allá de esta arquitectura de referencia también es posible estudiar las arquitecturas de las plataformas abiertas CC, cuya información es pública. Así pues, hoy en día existe una gran variedad de plataformas CC abiertas, distribuidas bajo la licencia *Open Source* y como no podía ser de otro modo también existen diferentes trabajos que comparan estas plataformas [3, 26, 36, 122, 123].

A continuación, se presentarán aquellas plataformas CC, que a nuestro juicio son las más importantes y que además son las más utilizadas, es decir, OpenStack[7], OpenNebula[8] y, en menor medida, Eucalyptus[9]. No cabe duda que existen otras plataformas, a las que a priori se les asigna el apellido cloud, como por ejemplo OpenQRM [124], Nimbus [125], CloudStack [126], Cloud Fondry [127], etc. Sin embargo, después de analizarlas estas plataformas no están diseñadas para cubrir un espectro generalista, o bien su único interés es el de facilitar la gestión de la capa hardware física y virtual, tarea que la mayoría de los sistemas de virtualización modernos ya incorpora como tarea por defecto.

[7] http://www.openstack.org/
[8] http://opennebula.org/
[9] https://www.eucalyptus.com/

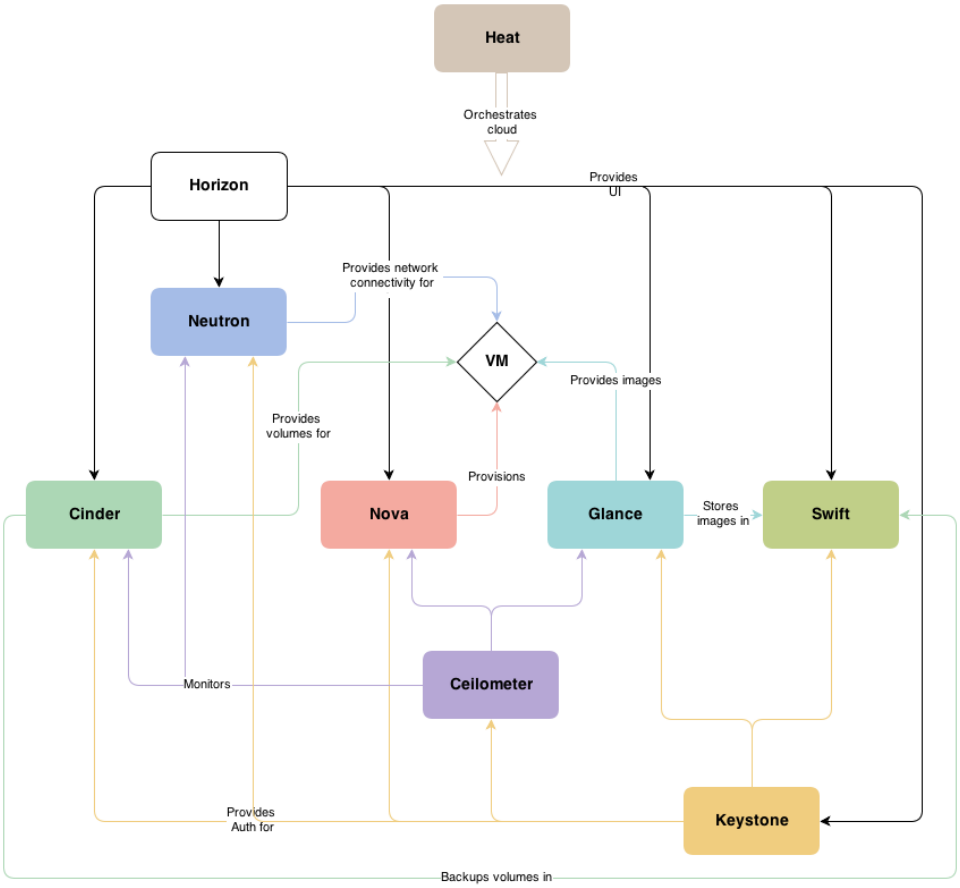

Figura 7.- Arquitectura conceptual de OpenStack

En primer lugar, OpenStack apareció en escena en torno a Julio de 2010, y fue convirtiéndose en popular debido a que grandes empresas fueron aceptando su modelo (Citrix, Dell, Suse, Telefónica, etc.), hoy en día agrupa a más de 13.500 personas en 132 países. Es un conjunto de proyectos de código abierto que pueden ser utilizados para configurar un entorno CC a medida.

En la Figura 7 se presentan los principales componentes de esta arquitectura [128]. Éstos son el servicio de identidad (*Keystone*), computación que incluye diferentes subservicios de almacenamiento, red, planificador y comunicaciones (*Nova*), almacenamiento (*Swift*), imágenes (*Glance*) y el servicio de configuración (*Dashboard*). Openstack es abierta y cada uno de estos servicios puede desplegarse en cualquier servidor, ya que están desacoplados entre sí, comunicándose a través de una cola de mensajes basada en el protocolo *Advanced Message Queuing Protocol* (AMQP), no obstante, el componente NOVA que gestiona los recursos, debe estar disponible en todos servidores, o al menos en sus componentes básicos.

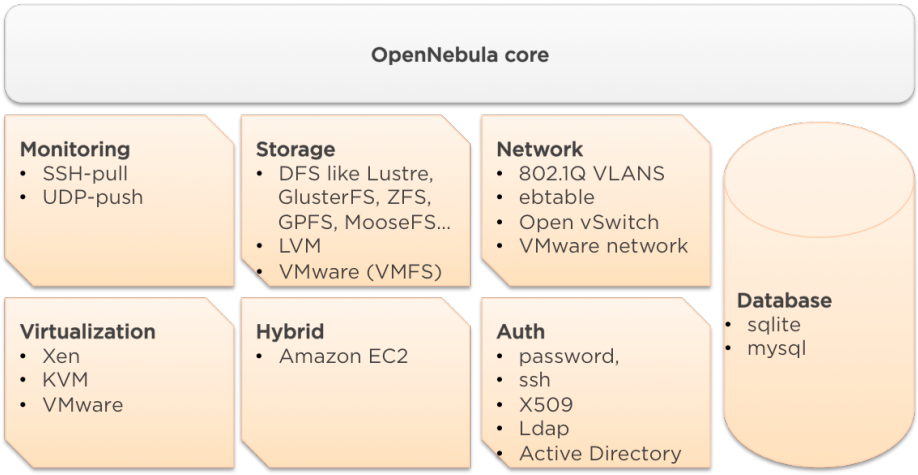

Figura 8.- Componentes OpenNebula

Por su parte, OpenNebula [129] fue el primer sistema CC abierto, su desarrollo empezó en 2005 y su primera versión estuvo disponible en 2008. Es un proyecto mucho más extendido que Openstack y es utilizado por una gran cantidad de instituciones y empresas para construir su propio entorno CC. OpenNebula está pensado para su uso en grandes centros de datos de forma que pueda transformar esos recursos en un entorno CC para ofrecer servicios de infraestructura. Su arquitectura (Figura 8) es flexible y modular, permitiendo integrar una gran cantidad de almacenamiento, red y entornos de virtualización.

Finalmente, Eucalyptus (Ela*stic Utility Computing Architecture for Linking Your Programs To Useful Systems*) fue lanzada inicialmente en el año 2008. Al igual que las anteriores, su objetivo es la transformación de un centro de proceso de datos tradicional en otro basado en CC con la capacidad de ofrecer servicios de tipo infraestructura. Sus componentes principales son el NC (*Node controller*), situado en cada máquina física, el SC (*Storage controller*) y el CLC (*Cloud Controller*) que es único en cada entorno de despliegue, adicionalmente dispone del módulo *Walrus* que actúa de puerta de entrada de las peticiones del entorno CC. Su principal característica es la compatibilidad con Amazon EC2 y S3, permitiendo incluso un modo de funcionamiento mixto, con gestión simultánea de máquinas virtuales en la instalación local y en EC2.

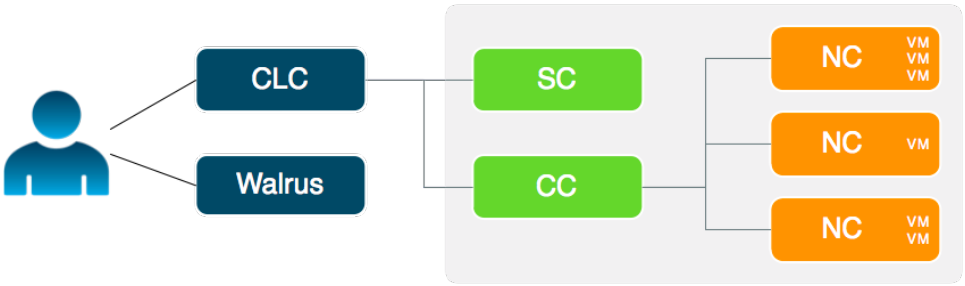

Figura 9.- Arquitectura de referencia en Eucalyptus

Aunque Eucalyptus, OpenNebula y OpenStack (así como el resto de las plataformas abiertas) comienzan a ser muy utilizadas, la realidad demuestra que su desarrollo todavía está lejos de las características deseables de una plataforma CC, ya que en su amplia mayoría se centran en la gestión de la infraestructura, sin tener en cuenta al usuario final. Una arquitectura moderna debería tener en cuenta la calidad de los servicios que ofrece (QoS) y los acuerdos (SLA) alcanzados con los usuarios. Sin embargo, estas arquitecturas abiertas están más centradas en la gestión de la compleja infraestructura subyacente, es decir, en las capas inferiores, que en los servicios que se les ofrecen a los usuarios finales.

Finalmente, en la Tabla 2 se presenta una relación de otras plataformas CC existentes en la actualidad.

Proveedor	Nivel	Empresa	Lic.	Plataforma	Descripción
vSphere	IaaS	VMWare	Comer.	N/A	Sistema operativo orientado a la formación de infraestructuras CC a través de virtualización.
Cloud Foundry	PaaS	VMWare	Libre	vSphere EC2	Capa PaaS para *vSphere*, que puede ser desplegada en infraestructuras de otros proveedores.
CloudStack	IaaS	Citrix	Libre	VMWare, Oracle VM, KVM, Xen	Capa de gestión de infraestructura que abstrae la plataforma tecnológica de virtualización subyacente.
AppScale	PaaS	Comun.	Libre	KVM, XEN, EC2, Eucalyptus	Ejecución en CC privado de aplicaciones de *Google App Engine*.
Nimbus	IaaS	Comun.	Libre	KVM, Xen	Gestión de infraestructura.
OpenQRM	IaaS	Comun.	Libre	KVM, Xen, VMWare	Gestión de infraestructura de centros de proceso de datos.

Tabla 2.- Plataformas Cloud Computing

2.6. Riesgos y vulnerabilidades

Tal y como se ha presentado, el paradigma computacional CC ha crecido con fuerza en los últimos años y su desarrollo ha motivado el avance de una gran cantidad de plataformas, tanto públicas, como privadas. También es relevante el cambio que ha producido en el modelo de negocio tradicional, así como la externalización de la información empresarial a los grandes centros de datos.

Sin embargo, todavía existen una serie de retos que la tecnología tiene que superar para que el uso de este paradigma computacional se extienda a todos los niveles. Estos retos existentes son principalmente los riesgos relacionados con la seguridad, entre las que se destacan las siguientes vulnerabilidades [17]: problemas de acceso, riesgos de seguridad en la tecnología de virtualización, vulnerabilidades en la tecnología web (SQL injection, cross-site scripting, seguridad física, IP spoofing, etc.) Así mismo, también existen dificultades debido a la pérdida del control de la información que se produce al alojar las bases de datos y el software de forma externa en grandes centros de datos, que son una caja negra para la empresa que hace uso de servicios. Entre estas vulnerabilidades destaca [17, 130]: la privacidad y confidencialidad de los datos, la pérdida y robo de información, problemas de autenticación, cumplimiento con las leyes de protección de datos y alta disponibilidad.

Diferentes autores han tratado de clasificar y acotar los problemas abiertos [16, 17, 51]. Estos estudios han propuesto diferentes taxonomías para organizar las vulnerabilidades, principalmente, atendiendo respectivamente (i) al tipo final de servicios en el que se produce (SaaS, PaaS o IaaS); (ii) la tecnología a la que se asocia el problema de seguridad; y (iii) finalmente, una clasificación atendiendo al tipo de despliegue (Público, Privado y/o de Comunidad).

Para analizar los problemas de seguridad, siguiendo el trabajo de Grobauer *et al.* [51], en primer lugar hay que destacar que CC está basado en la integración de diferentes tecnologías subyacentes que trabajan de forma coordinada. Así, resulta necesario distinguir entre los problemas de seguridad intrínsecos a cada una de esas tecnologías y los problemas derivados de su uso de forma conjunta en un entorno CC. En este sentido es posible referirse a las vulnerabilidades de las tecnologías asociadas a un entorno CC:

- **Aplicaciones web y servicios.** Donde se enmarcan las vulnerabilidades relativas a las capas PaaS y SaaS de la pila de servicios del paradigma.
  Destacan vulnerabilidades relacionadas con la gestión de los datos [17] (localidad, integridad, segregación, control de acceso, confidencialidad, disponibilidad, violación de la información y la gestión de copia de seguridad), así como las vulnerabilidades clásicas de la tecnología Web [17] (*Cross-site scripting*, inyecciones OS, SQL o LDAP, manipulación de cookies, problemas de configuración, vulnerabilidades en la capa TCP, ataques *Captcha*, etc.). No obstante, la mayoría de estas vulnerabilidades no están relacionadas con la capa de aplicación, sino que en torno al 60% están relacionadas con los niveles inferiores de la pila de tecnologías (sistema operativo y vulnerabilidades conocidas y no conocidas) [16, 131].
- **Protocolos de Comunicaciones.** Dado que la información es accedida de forma ubicua a través de diferentes redes, habitualmente Internet, utilizando para ello protocolos estándares de redes. Todas las vulnerabilidades asociadas a estos protocolos también constituyen riesgos de seguridad para el entorno CC. Entre estos riesgos destacan [16, 17] el análisis de paquetes y la violación de redes, la gestión débil de sesiones, las peticiones SSL inseguras, ataques DNS, *sniffers*, reutilización de direcciones IP y *Prefix hijacking*.

Todas estas vulnerabilidades pueden ser utilizadas para realizar ataques de tipo *Man-In-The-Middle* lo que puede llegar a comprometer la seguridad de los datos y de la infraestructura que subyace.

- **Persistencia de la información.** Como no podía ser de otro modo, también existen problemas relacionados con las bases de datos utilizadas en el entornos CC, que habitualmente son bases de datos orientadas a documentos [30]. Los problemas habituales están realizados por el uso de protocolo de encriptación de la información obsoletos [51], los cuáles están ampliamente estudiados y las soluciones propuestas ofrecen un nivel de seguridad aceptable [132].

  También existen problemas de confidencialidad, en cuanto a la verificación de la identidad [17] y, fundamentalmente, debido a que los datos de diferentes empresas (*Cloud users*) son almacenados en el mismo centro de proceso de datos, lo que puede provocar que puedan tener acceso a información confidencial de otros usuarios. En este sentido, el rol *Cloud provider* debe asegurar la confidencialidad de los datos entre sus clientes, pero además implementar medidas de seguridad adicionales para impedir el acceso de sus propios empleados a datos confidenciales de terceros.

  Finalmente, se observan problemas relativos a la necesidad de disponibilidad de Internet para acceder a la información o los problemas de latencia derivados del uso de redes como medio de transmisión.
- **Gestión de la infraestructura.** El uso de una capa de virtualización tiene innumerables ventajas a la hora de gestionar de forma dinámica la infraestructura. Sin embargo, también tiene más problemas de seguridad debido a que los usuarios potencialmente malintencionados comparten un mismo medio, e incluso un mismo servidor físico. Aunque la virtualización en sí misma puede ser considerada como una técnica de seguridad debido al aislamiento que proporciona. Los principales riesgos se enmarcan en la posibilidad de romper este aislamiento y conceder acceso a la máquina física, lo que por ende, da acceso al núcleo del entorno CC [17]. En este sentido, existen amplios estudio sobre las posibles vulnerabilidades [133] y las soluciones a las mismas [134].

  Finalmente, en cuanto a los riesgos en la capa de virtualización, teniendo en cuenta que un entorno CC ofrece servidores virtualizados en el nivel IaaS, los proveedores de servicios tienen un nuevo reto, ya que deben implementar un modelo de seguridad compartida, donde el *Cloud Provider* debe asegurar la seguridad del entorno CC a nivel físico y el Cloud User debe ocuparse de la seguridad de la máquina virtual que adquiere como servicio [51]. El reto reside en que el *Cloud provider* debe asumir problemas de seguridad dentro de su propia infraestructura debido a una gestión incorrecta de la configuración por parte del Cloud user.

En cuanto a los problemas de seguridad inherentes propios al paradigma, cabe destacar, en primer lugar las (i) Vulnerabilidades en la monitorización, ya que un entorno CC es aquel que tiene que autogestionarse, constituyendo un paradigma a medio camino entre *Utility Computing* [8] y *Autonomic Computing* [62], es indudable que debe recoger datos sobre el uso a lo largo de toda su infraestructura (hardware y software). Con esta información, se evalúa el uso que cada usuario realiza del entorno CC y por lo tanto cuál es el coste asociado. Un problema de seguridad en esta monitorización, puede provocar la pérdida de los rendimientos derivados del alquiler de infraestructura o servicios a terceros [51]. En segundo lugar, la (ii) dificultad para la gestión de la identidad, que debe seguir un modelo de tipo *Single sing-on* [135] y que es uno de los retos de este tipo de plataformas, no sólo por el hecho de asegurar la autenticación de usuarios en diferentes aplicaciones desplegadas en entornos geográficamente distribuidos, sino por el hecho de un

almacenamiento compartido de las credenciales de seguridad de cada usuario. Las soluciones a estos problemas de seguridad son [17] sistemas de gestión de la identidad independientes entre los diferentes *Cloud users*, sincronización de credenciales de seguridad a nivel SaaS, y la federación del modelo de gestión de la identidad. Y, finalmente, la (iii) Interoperabilidad entre plataformas, que como ya se ha comentado es un campo de investigación en la actualidad, pero que además dificulta en gran medida su implantación a gran escala. Para una empresa deslocalizar su información en el centro de procesos de datos de un tercero es un reto en sí mismo, pero además existe el problema de *lock-in*, que consiste en que una empresa no está dispuesta a depender únicamente de un solo proveedor debido a la dificultad de conexión entre plataformas.

Tal y como se ha presentado, los entornos CC como cualquier sistema de información y especialmente aquellos que son distribuidos, tiene un conjunto de riesgos y vulnerabilidades asociadas no sólo a las tecnologías subyacentes, sino al propio paradigma en sí mismo. Existen investigaciones activas y estudios sobre como asegurar la seguridad e interoperabilidad de las diferentes plataformas CC. La organización *Cloud Security Alliance* (CSA) publica una guía sobre buenas prácticas en materia de seguridad en el marco de los entornos Cloud [136]. Aunque actualmente existen trabajos en todas las áreas vulnerables anteriormente señaladas, cuyo estudio se escapa de este trabajo, destacar que la gestión de seguridad depende en gran medida del modelo de despliegue del entorno CC (público o privado) [16] ya que el tipo de usuarios y servicios que implementa son totalmente opuestos.

2.7.Conclusiones

A lo largo de este segundo capítulo se ha revisado de forma detallada todo el contexto tecnológico y social que envuelve al paradigma CC, así como su caracterización interna y marco de aplicación. No cabe duda de que este modelo tecnológico ha experimentado una notable evolución desde sus inicios en el 2007, donde ni si quiera IBM, que hoy en día es uno de los grandes proveedores CC, tenía una idea clara de cuál iba a ser su evolución.

Según Gartner la gran aceptación por parte del tejido empresarial [137], así como su fácil y rápida integración con las arquitecturas tecnológicas tradicionales [138] ha motivado el rápido desarrollo hasta nuestros días. Así mismo, el modelo de comercialización (*pay-as-you-go*) [6], similar al de los productos de utilidad tradicionales, también ha sido otras de las claves de su rápida evolución. Ya que, las tecnologías subyacentes del paradigma CC han permitido producir los recursos computacionales que se ofertan como servicio atendiendo a las necesidades de los usuarios y permitiéndoles utilizar (y pagar) sólo la cantidad y calidad de recursos que realmente necesitan en cada momento.

A lo largo de este capítulo se ha analizado detalladamente como las tecnologías relacionadas y precedentes han hecho posible la explosión de este modelo computacional. En este sentido, también se ha revisado como ha sido la evolución de las definiciones propuestas por los diferentes grupos y centros de investigación que centran su trabajo en el desarrollo del paradigma CC. En este sentido, se ha destacado sobre el resto la definición propuesta por el NIST [7] en la que, no sólo se define el propio paradigma y lo que representa, sino que también se describen las características que debe cumplir (principalmente, la disponibilidad de servicios a medida de forma automática y la provisión de recursos computacionales mediante la elasticidad). Así mismo, la definición propone los tres tipos de servicios ofertados ampliamente conocidos, pero que denomina capacidades (*software*, plataforma e infraestructura) y cuatro modelos de despliegue (público, privado, híbrido y comunitario). Estos modelos de despliegue, como también se ha detallado a lo largo del capítulo, dan lugar a un amplio debate sobre la interoperabilidad entre plataformas, el denominado InterCloud [139, 140].

Sin embargo, pese a la calidad y amplitud de esta definición, no son pocas las voces [141] que indican que la definición no es lo suficientemente precisa y que, en cierta medida, está influenciada por el interés de los gobiernos, por la implantación de este paradigma computacional en las diferentes administraciones con el objetivo de incrementar la eficiencia y reducir de costes operacionales. En este sentido, tampoco son pocas las nuevas innovaciones tecnológicas que han surgido al amparo de este paradigma [142, 143]. Sin embargo, después de realizar un estudio acerca de las plataformas existentes a lo largo del capítulo; se observa una pobre implantación de estas novedades. Así pues, la mayoría de estas plataformas simplemente basan sus esfuerzos en proporcionar servicios de infraestructura *hardware* mediante el uso de la tecnología de virtualización subyacente, sin tener en cuenta las capacidades de los niveles superiores como plataforma y *software*.

El capítulo, para terminar, también revisa las vulnerabilidades del paradigma CC, determinando que la mayoría de las debilidades están asociadas a las tecnologías sobre las que se sustenta. Sin embargo, si que ha sido posible encontrar debilidades directamente asociadas al paradigma CC, las cuales pueden ser agrupadas en las tres siguientes categorías: (i) servicios y aplicaciones, (ii) persistencia de la información y, finalmente, (iii) las debilidades relacionadas con los protocolos de comunicaciones.

Partiendo de estas debilidades y vulnerabilidades descubiertas, así como de las limitaciones en las plataformas CC y siguiendo la hipótesis propuesta en el presente trabajo de investigación, se seleccionan los SMA y sobre todo aquellos basados en OV para combatir los obstáculos hallados. Este modelo arquitectónico permite la autoadaptación dinámica mediante la integración de innovaciones basadas en métodos, técnicas, herramientas y modelos derivados de la IA y de la distribución de responsabilidades entre los diferentes componentes del entorno CC. Por tanto, en el siguiente capítulo se analizará en detalle la relación existente actualmente entre SMA y el paradigma CC con el objetivo de determinar un ámbito claro de aplicación del modelo arquitectónico de los SMA basados en sociedades artificiales en el marco tecnológico propuesto por CC.

Capítulo 3. Sistemas multiagente y Cloud computing

Una vez que se ha acotado y caracterizado correctamente el término CC, a lo largo de este apartado se presentará una revisión del estado del arte de los SMA, prestando especial atención a la relación existente con el paradigma CC, de forma que sea posible la identificación de cómo los SMA pueden contribuir a la mejora del paradigma CC.

En primer lugar, los agentes y SMA han sido, desde su nacimiento, un campo de gran interés por parte de la comunidad científica. Este interés ha sido motivado, sin duda alguna, por el conjunto de capacidades diferenciales que presentan este tipo de sistemas con respecto a los sistemas distribuidos tradicionales, entre las que cabe destacar la autonomía, la capacidad de interacción, la racionalidad, así como sus características inteligentes y de aprendizaje [38]. La creciente complejidad de las aplicaciones software y los sistemas de información está motivada por el incesante avance de la infraestructura hardware y de los sistemas de comunicación. Éste es el caso de los entornos CC, los cuáles aúnan la gestión de grandes cantidades información, junto con interfaces web de última generación, todo ello, gestionado a través de un entorno masivamente distribuido. No cabe duda que el diseño de este tipo de entornos complejos requiere de nuevas técnicas de Ingeniería del Software [144, 145] que sean capaces de afrontar las posibilidades que hoy en día ofrece la técnica, teniendo en cuenta la perspectiva de la IA. La propia teoría de agentes ha tenido que evolucionar en gran medida desde sus inicios, allá en la década de los 90, para facilitar la creciente complejidad software, desde el concepto inicial de entidad con comportamiento autónomo, capaz de actuar en un medio para satisfacer unos requisitos individuales. Hoy en día, estos sistemas no pueden ser entendidos sin su capacidad para trabajar de forma coordinada, comunicándose y cooperando entre sí, formando SMA. La estructura de estos sistemas también se ha desarrollado utilizando la base que proporcionan las organizaciones en las sociedades humanas, en el que se tiene un fin común y están sujetos a un conjunto preestablecido de normas, acuerdos y protocolos de comunicación, sin perder, el dinamismo y la pro-actividad así como la capacidad de adaptarse al medio de forma autónoma.

Aunque en un primer momento, se puede pensar que ambos sistemas distribuidos (SMA y CC) son incompatibles, un análisis más detallado demuestra que son complementarios, siendo posible identificar un buen número de sinergias entre ellos. Mientras que, por un lado, los entornos CC puede cubrir las necesidades computacionales de persistencia de información y el potencial de cómputo que requieren los SMA en diferentes aplicaciones como la minería de datos, gestión de servicios complejos, etc. Por otro lado, los SMA pueden ser utilizados para el diseño de entornos CC mucho más eficientes, escalables y adaptables que los existentes actualmente. Además, el uso de SMA en el marco del diseño de sistemas CC aporta a este paradigma nuevas características, como el aprendizaje o la inteligencia, lo que hace posible desarrollar entornos de computación mucho más avanzados en todas sus facetas (servicios inteligentes, interoperabilidad entre plataformas, distribución de recursos más eficiente, etc.). La relación simbiótica entre ambos paradigmas no es muy amplia y como se verá a lo largo del capítulo, por la naturaleza de los trabajos existentes, puede caracterizarse incluso como incipiente.

Este Capítulo 3 se estructura tal y como sigue, el apartado 3.1 contiene una breve revisión histórica sobre la teoría de agentes y los SMA. Posteriormente, el apartado 3.2 se realiza un análisis exhaustivo de los trabajos que aúan el paradigma CC con los SMA. A continuación, dado que se propone el uso de SMA basados en organizaciones virtuales para la formalización de la arquitectura, en el apartado 3.3 se realiza una revisión de este modelo de construcción de SMA, así como de las metodologías que hacen posible el análisis y diseño de sistemas complejos con el que se propone. Finalmente, el apartado 3.4 contiene las conclusiones del capítulo.

3.1. Teoría de agentes y los sistemas multiagente

La teoría, y por ende, el paradigma computacional que conforman los agentes y los SMA, están basados en el concepto de agente, inicialmente propuesto por Russel *et* Norving [37]. Esta teoría de agentes no tiene que ser vista únicamente desde la perspectiva técnica (Ingeniería del Software, Sistemas Distribuidos, IA, etc.) sino que, como se verá más adelante a lo largo de esta sección, es necesario ir más allá, incluyendo otras ciencias como pueden ser la Psicología y la Sociología [23, 146, 147].

En primer lugar, cuando se estudia la teoría de agentes, es necesario enunciar una definición del concepto de agente. Esta definición, aunque en el pasado fue un foco de discusión en la comunidad científica [37, 38, 148]; hoy en día, el concepto de agente está suficientemente acotado. Entre todas las definiciones que han surgido, cabe destacar por su simplicidad, precisión y claridad la definición que enunció Wooldridge [149]:

> *An agent is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives.*

Posteriormente se refina esta definición, en colaboración con Jennings [38] en la que atribuía al agente las capacidades de autonomía, habilidades sociales, reactividad y pro-actividad. El paradigma de agentes surgió para satisfacer las deficiencias que la Ingeniería del Software clásica tenía para modelar sistemas software complejos. En este sentido, según [44, 150], el paradigma de agentes puede ser aplicado en tres tipos de sistemas principalmente:

- **Sistemas abiertos.** Estos sistemas son capaces de cambiar dinámicamente, debido a que sus componentes no se conocen a priori y son altamente heterogéneos (diferentes entidades, implementaciones y técnicas, etc.).
  Para dar solución a este tipo de sistemas son necesarias técnicas de negociación y cooperación, cuya base forma parte de los sistemas multiagente [151].
- **Sistemas complejos.** Se relacionan con sistemas grandes e impredecibles, cuya forma de abordarlos son el uso de técnicas de abstracción y modularidad. Ambas características forman parte intrínseca de los SMA, ya que la noción de agente autónomo es en sí mismo una abstracción de un módulo que encapsula procedimientos y datos (objetos) que permiten resolver de forma autónoma un problema [152].
- **Sistemas ubicuos.** Este tipo de sistemas mejoran el uso de un sistema informático mediante la utilización de la potencia computacional de un entorno físico, normalmente distribuido, pero que de algún modo se abstrae su complejidad de cara al usuario final. El sistema debe cooperar con el usuario para alcanzar los objetivos perseguidos. Existen diferentes ejemplos dentro de este tipo de sistemas, destacando el sector del control de procesos [153] y la manufacturación [154].

El otro gran foco de interés ha sido el de clasificar a los agentes atendiendo a diferentes criterios [37, 149, 155-158]. Así pues, hoy en día existen varias clasificaciones de agentes, destacando principalmente las siguientes taxonomías:

- **Clasificación atendiendo al tipo de implementación del agente** [37]. Según esta clasificación se distinguen los siguientes cuatro tipos de agentes que se presentan ordenados en función de su complejidad: agente reflejo simple, agente reflejo con estado interno, agente basado en metas y agentes basados en utilidad.

- **Clasificación atendiendo al tipo de atributos de los agentes** [155]. Nwana realiza una compleja clasificación de los agentes, distinguiendo a los agentes en función de diferentes características. Así, en función de la (i) movilidad existen agentes estáticos y móviles; en función del (ii) modelo de razonamiento interno existentes agentes deliberativos y reactivos; y finalmente, atendiendo a (iii) otros atributos como la autonomía, la cooperación y el aprendizaje, se distinguen agentes colaborativos, de interfaz, de aprendizaje colaborativo y agentes Smart.
  Finalmente, la unión de características de estos tipos de agentes, da lugar, a un nuevo tipo de agente, denominado híbrido.
- **Clasificación atendiendo al diseño conceptual** [158]. Teniendo en cuenta el diseño conceptual de los agentes, existen agentes interfaz, búsqueda e información.
- **Clasificación atendiendo a la arquitectura** [149]. La última clasificación que se presentará en este documento atiende a la arquitectura interna del agente. Siguiendo este criterio se distinguen arquitecturas reactivas, deliberativas e híbridas.

Dentro de estas clasificaciones de agentes destaca la que clasifica atendiendo a la arquitectura [149], ya que ésta determina cuáles son sus componentes principales, y como interactúan entre sí para alcanzar la misión final. Si se considera un agente como un sistema complejo, la arquitectura deberá describir la estructura interna del agente, explicando como se descompone el agente en módulos independientes que interactúan para conseguir la funcionalidad requerida. En este sentido Wooldridge [149] propone tres arquitecturas clásicas. La (i) arquitectura reactiva en el que el agente carece tanto de razonamiento como de capacidad para representar su entorno, y sus acciones se modelan mediante reglas básicas [159]. La (ii) arquitectura deliberativa donde el agente es capaz de mantener una representación simbólica del conocimiento y planificar el conjunto de acciones a tomar para conseguir sus metas. Y, finalmente, (iii) la arquitectura híbrida que es un modelo de arquitectura que aúna características de las dos anteriores.

La arquitectura deliberativa más ampliamente estudiada y extendida es el modelo BDI [160], que ha sido el más difundido dentro de los modelos de razonamiento, ya que combina un modelo filosófico asociado al razonamiento humano y un número considerable de implementaciones [161]. Para el desarrollo del modelo se plantean un conjunto actividades mentales de los agentes inteligentes [162, 163]:

- **Creencias.** Relacionado con el conjunto de proposiciones que el agente acepta como verdaderas. Es decir, la visión que el agente tiene del entorno y el estado del resto de agentes.
- **Deseos o metas u objetivos.** Que hacen referencia al conjunto de propiedades que el agente trata de hacer verdaderas.
- **Intenciones.** Asociadas al conjunto de acciones planificadas que le permiten llegar a un estado deseado.

Estos agentes proporcionan soluciones en entornos dinámicos e inciertos. Además, son capaces de enfrentarse a problemas del mundo real, incluso cuando sólo cuentan con una visión parcial del problema y con un número limitado de recursos.

3.1.1.Los sistemas multiagente

Los agentes son entendidos por definición como entidades que interaccionan con el medio y también con otros agentes. En este sentido, cuando dos o más agentes son capaces de trabajar de forma conjunta con el objetivo de resolver un problema se habla de SMA [164]. Además, la mayoría

de sistemas software existentes hoy en día son altamente complejos (como los entornos CC), ya que suelen ser sistemas concurrentes, que interaccionan entre sí y con otros sistemas externos [25]. Como consecuencia, los sistemas software tienen que ser abiertos, es decir, existe la posibilidad de que nuevos componentes se incorporen y deban trabajar con el resto de los componentes ya existentes. Éste es el entorno ideal de trabajo de los SMA.

Los SMA extienden la idea de agente individual mediante la descripción de la infraestructura, que proporciona la comunicación e interacción, adaptándose a la complejidad de un sistema abierto y a los retos existentes en el entorno gracias a sus características particulares [147]:

- **Previsibilidad.** Los agentes individuales tienen diferentes niveles de libertad, de forma que pueden tomar acciones dentro de sus respectivos dominios. Los SMA tienen que ser lo suficientemente predecibles como para ser capaces de anticipar las situaciones futuras y elaborar planes o secuencias de tareas que les permita asegurar los objetivos preestablecidos. Entre las técnicas asociadas a la planificación cabe destacar la (i) teoría de grafos [165], la (ii) planificación basada en la satisfactibilidad [166], el (iii) razonamiento basado en modelos [167, 168], la (iv) utilización de heurísticas [169], la (vi) planificación con tiempo y recursos [170] y, finalmente, la (vii) planificación basada en casos [171].
- **Seguridad.** Es una medida del grado en el que el sistema puede protegerse de usos indebidos, así como de la capacidad para asegurar la integridad de los datos y los recursos [172].
- **Adaptabilidad.** Un SMA debe de ser capaz de adaptarse a los cambios que se produzcan en el entorno. Esta adaptación depende de la capacidad de los agentes individuales para aprender y predecir los cambios [173], e incluso de su capacidad para evaluar el estado del propio entorno [174].
- **Comunicación.** La coordinación es una de las características de los SMA [175]. Es la capacidad de distribuir la experiencia, recursos e información entre las diferentes entidades del sistema [176].

La comunicación puede ser vista desde dos puntos de vista. En primer lugar, (i) la coordinación y negociación donde los agentes son capaces de armonizar sus acciones con otros agentes para satisfacer objetivos comunes; y (ii) la competitividad en los que existe una disputa entre los diferentes miembros del sistema. En este tipo de sistemas, el éxito de un subconjunto de agentes, implica el fallo de otro subconjunto.

- **Disponibilidad.** Aunque la disponibilidad puede ser considerada un subobjetivo de seguridad, la necesidad de cooperación en un SMA requiere que sea necesario considerar la disponibilidad en el diseño de los SMA, ya sea desde un punto de vista implícito o explícito.
- **Tolerancia a fallos.** En un SMA, el fallo de un agente no tiene que provocar el fallo de todo el sistema. En este sentido, un fallo de un individuo debe provocar la coordinación del sistema para determinar quien debe asumir las tareas del agente erróneo.
- **Modularidad.** La modularidad de un SMA incrementa la eficiencia en la definición de las tareas, reduce la comunicación y, finalmente, incrementa la flexibilidad del sistema.
- **Agregación.** Es una medida del grado en que los agentes pueden convertirse en partes de otros agentes, o formar parte de un sistema de nivel superior. La composición de agentes, implica la pérdida del control por cada uno de los componentes para, en contrapartida,

tener un control global de los agentes agregados, lo que proporciona mayor flexibilidad [177].

En definitiva, los sistemas multiagente proporcionan herramientas para modelar sistemas complejos y altamente distribuidos [178]. Además, los SMA definen formalmente las tareas de comunicación, coordinación y negociación, lo que habilita el intercambio de conocimiento y, por lo tanto, coordinar el desarrollo de las actividades, pudiendo ser posible la negociación en caso de conflictos [164].

3.2.Los sistemas multiagente en el marco del paradigma Cloud Computing

El número de trabajos que se pueden encontrar en el estado del arte que relacionan CC y la tecnología de agentes es muy reducido, aunque sin embargo esta tendencia está disminuyendo y cada vez es más fácil encontrar estudios y aplicaciones centradas en este campo. Pese al limitado número de estudios sobre la materia, empieza a ser habitual el concepto *Agent-based cloud computing* o *Agent-based cloud platform*, enunciado por diferentes autores en los últimos años [14, 52, 97, 179, 180]. Así pues, a lo largo de este apartado, se hará un estudio de los principales trabajos existentes en el estado del arte que relacionan ambos paradigmas computacionales.

El estado del arte no presenta un gran número de estudios acerca de la relación entre agentes y CC. Únicamente es posible encontrar la clasificación que propone Talia [52, 97], aunque con un limitado número de referencias. Esta taxonomía se realiza desde un punto de vista agentivo, diferenciando dos grandes grupos. Por un lado, (i) las aplicaciones de MAS que utilizan los servicios CC o están directamente desplegados sobre él (*Agents using cloud*). Y, por otro lado, (ii) los entornos CC que utilizan la tecnología de agentes para la gestión de sus recursos (*Cloud using agents*).

Sin embargo, después de realizar un análisis detallado del estado del arte, no resulta fácil enmarcar los trabajos en uno u otro grupo. Por ello, en este trabajo, se propone una nueva clasificación desde el punto de vista CC, en base a la arquitectura de referencia propuesta por el NIST [10] y las diferentes responsabilidades de cada uno de los roles participantes en el paradigma CC según la citada arquitectura: *Provider*, *Consumer*, *Broker*, *Carrier* y *Auditor*. Si comparamos ambas clasificaciones, el grupo de Agents using cloud, estaría totalmente alineado con rol *Consumer*, en la clasificación que se propone; pero el grupo *Cloud Using Agents* tendría tres tipos de roles implicados (*Provider*, *Broker* y *Auditor*) por lo que esta propuesta aporta una mayor capacidad semántica. Dentro de esta clasificación, no se incluye al rol *Cloud Carrier*, ya que es la entidad encargada de proporcionar transporte a la información y no aportaría ninguna funcionalidad posible entre SMA y CC. Los siguientes apartados presentarán los trabajos existentes en el estado del arte según esta clasificación propuesta.

3.2.1.Cloud Consumer

El rol *Cloud consumer* es aquel que hace uso de las capacidades que proporciona un proveedor de servicios computacionales de tipo CC. Existen números trabajos en el que los sistemas SMA utilizan los servicios computacionales proporcionados por los sistemas CC. A continuación, se presentan los trabajos en los que los SMA utilizan servicios proporcionados por los entornos CC; organizados en función de los servicios más habituales:

- **Servicios de persistencia y computación.** La persistencia de la información es uno de los grandes servicios en el marco de los entornos CC, permitiendo el almacenamiento de grandes cantidades de datos; o liberando a dispositivos móviles ligeros de la necesidad de almacenar información. Existen autores que incluso llaman a estos servicios *Data as a Service* (DaaS) [11]. En el estado del arte se pueden encontrar diferentes aplicaciones en ámbitos muy diversos, como la gestión del transporte [181], la persistencia de objetos de aprendizaje [182], almacenamiento de vulnerabilidades de seguridad [183], etc.
- El almacenamiento de grandes cantidades de información en entornos CC suele denominarse *Big Data* [184] y su análisis se denomina *Big Data Analysis* [185], para lo cuál además de utilizarse las capacidades de persistencia de las plataformas CC, también se requiere su potencia computacional. En este campo puede encontrarse una variedad amplia

de trabajos aplicados en diferentes áreas como el análisis de mercado [186-188], fusión de información [189, 190], ámbito socio-sanitario [191], etc.

Aunque estas aplicaciones están basadas en plataformas CC, muchas de ellas siguen utilizando el concepto del paradigma *Grid Computing* [12], como por ejemplo la herramienta CASE de simulación propuesta por Decraene *et al.* [192, 193], el uso del clúster Hadoop de Apache [194], la simulación sobre población [195], las redes de sensores [196], etc.

No obstante, también existe otro tipo de aplicación de CC y agentes en las cuáles el entorno CC absorbe las necesidades computacionales que de otro modo no podrían realizarse por el SMA, como en dispositivos ligeros [183, 197, 198], o el análisis de grandes cantidades de información utilizando técnicas de minería de datos [199] en diferentes campos como la bioinformática [200], la gestión de tráfico rodado [201], recolección y análisis de información [202], etc.

Finalmente, otra gran aplicación es el uso de las plataformas CC para paralelizar secuencias de tareas, así pues, es posible encontrar herramientas para la ejecución de tareas utilizando algoritmos de planificación mediante agentes [203]; o el que, quizás, sea uno de los trabajos más prometedores en el campo ya que se trata de una pionera modificación sobre la metodología Moise+ [204], denominada ParaMoise [205, 206] y que trata de facilitar y mejorar la ejecución paralela de los diferentes agentes que forman una sociedad u organización de agentes inteligentes a través de las características computacionales de un entorno CC.

- **Servicios de infraestructura.** Además de los servicios anteriormente identificados de computación y persistencia, el otro de gran capacidad que proporciona un entorno CC es la infraestructura, es decir, la capacidad de ofertar máquinas virtuales de múltiples características bajo demanda. Estas capacidades pueden utilizarse por el SMA para su despliegue en entornos de supercomputación y especialmente han sido utilizados para la simulación basada en agentes en un entorno de supercomputación [207, 208], o mediante el uso de una plataforma de simulación específica como Repast HPC [209-211].

  Finalmente, también hay que destacar que es creciente el interés de la comunidad científica en ofertar no sólo una infraestructura tradicional en el marco del paradigma CC, sino también otro tipo de infraestructura como redes inalámbricas de sensores en el que los sistemas multiagente cobran especial importancia [212-216].

Tal y como se observa en los trabajos presentados, cuando el SMA juegue el rol de *Cloud user* es posible mejorar sustancialmente las capacidades de los SMA ya que les permite disponer de mayores recursos para realizar las tareas. Por ende, el marco de aplicación de los SMA crece debido a que es posible realizar razonamientos más complejos, se puede modelar el entorno con un mayor nivel de detalle.

3.2.2.Cloud broker

El rol del *Cloud Broker* es el intermediario entre proveedores y consumidores, constituyendo un rol fundamental en el paradigma CC desde un punto de vista de la comercialización de servicios. El objetivo del *Cloud Broker* es proporcionar al consumidor servicios computacionales, con independencia del proveedor, adecuados a sus necesidades computacionales en función de su presupuesto. Para ello, el *Cloud Broker* está en contacto con varios proveedores de forma simultánea, manteniendo un archivo con las características de los servicios que ofrece cada uno y el coste asociado. Así el consumidor, en este caso, delega en el *Cloud Broker* la capacidad de contratación, lo que le permite obtener la mejor calidad/precio en los servicios que utilizan los consumidores. En

definitiva, las tareas principales del *Cloud Broker* son la intermediación y gestión de acuerdos, la agregación de servicios y el arbitraje.

Un bróker tiene que tener como cualidades principales la negociación, extracción de información compleja, así como la gestión de información y servicios; características que están perfectamente alineadas con las características de los SMA. Así pues, en el estado del arte se encuentran un buen número de trabajos, que además tienen la suficiente madurez como para ser utilizados en entornos reales. Lo que en parte se argumenta debido a la relación existente de los trabajos existentes con campos ampliamente estudiados en los últimos años (búsqueda, composición de servicios, negociación, etc.).

K.M Sim en unos de sus trabajos más conocidos [179], hace una disertación acerca de la aplicación de agentes en el marco CC, enunciando conceptos como *agent-based computing*, *agent-based cloud computing*, *agent-based cloud search engine*, *negociation agents and agent-based cloud commerc*e y, finalmente, *agent-based cloud service composition*; siguiendo sus trabajos en este contexto. Sin embargo, de forma más generalista en el presente trabajo únicamente se propone tres subgrupos, tal y como sigue:

- **Buscador de proveedor Cloud.** Este tipo de herramientas se basan en la centralización de la búsqueda de proveedores de servicios CC en función de las características de calidad/precio requeridas por los consumidores.
  El propio Sim propone el buscador Cloudle [14, 179] que hace uso de un SMA que utiliza un *web crawler* [217] para recoger la información sobre servicios y precios que disponen de los proveedores, para posteriormente relacionar estas características a través de una ontología que permita la interoperabilidad entre proveedores que usen terminología distinta para referirse al mismo servicio o característica. Dentro de esta línea, existen otros trabajos como el propuesto por Alhamad *et al.* [9, 218] en el que proponen un modelo de confianza y reputación, basado en diferentes pesos sobre parámetros del SLA, de forma que pueda servir como referencia a los autores a la hora de elegir un proveedor de servicios. Para ello, en función de los parámetros requeridos por el consumidor y los proporcionados por el proveedor valida que, efectivamente, no se viola el acuerdo SLA. More *et al.* proponen un modelo similar mucho menos avanzado [219]. En esta línea, otro enfoque es el propuesto por Habib *et al.* [220] en el que se presenta un modelo de confianza multifaceta [221] más complejo. Finalmente, Mouratidis *et al.* proponen un modelo de selección de proveedores CC en base a criterios de seguridad [222].
- **Negociación de acuerdos SLA.** En este grupo se enmarcan a los SMA que permiten la gestión automática o semiautomática de los acuerdos SLA entre el consumidor y el proveedor, liberando al consumidor de esta tarea.
  Uno de los primeros modelos de negociación basado en agentes, fue propuesto por An *et al.* que utiliza una máquina de estados que tenía en cuenta las fases de negociación en el que únicamente incluía tiempo y costes [223]. K.M. Sim también propone un modelo propio de negociación [179, 224, 225] bastante más complejo que el anterior, ya que se realiza entre varios consumidores y proveedores, se basa en una estrategia de negocio y tiene en cuenta aspectos como el tiempo, regateo, información incompleta y la estimación de costes. Otros trabajos que destacan en este sentido son el Cloud Agency [180, 197], desarrollado en el marco del proyecto FP7 mOSAIC. El modelo de este *framework* utiliza, al igual que Cloudle, una ontología para facilitar la interoperabilidad semántica entre plataformas, así como un SMA que realiza la negociación, monitorizando los servicios proporcionados y renegociando en el caso de que cambien las necesidades del consumidor, o los servicios del

proveedor. Finalmente, también existen otros modelos de negociación específicamente diseñados para dispositivos ligeros [226].

- **Composición de servicios.** Se refiere a la composición de servicios CC para da lugar a un único servicio más complejo proporcionado por uno o varios proveedores de forma simultánea. Diversos autores se refieren a la composición de servicios como *Cloud Manufacturing* [227-229].

  La composición de servicios en CC puede verse desde dos vertientes [102]: (i) horizontal que se refiere a servicios heterogéneos de múltiples proveedores; y, (ii) vertical, que se refiere a la composición de servicios homogéneos orientados a incrementar la capacidad del servicio en un determinado momento.

  En cuanto a las aplicaciones existentes, [102, 179] propone un servicio automático de composición vertical basado en el FSCNP (*Focused Selection Contract Net Protocol*) para la selección automática de servicios y una tabla de compatibilidad de servicios, previamente establecida; de forma, que si un plataforma CC no es capaz de asumir el SLA acordado, se puede seleccionar otro servicio de las mismas características. Estos mismos autores, disponen de otro modelo [102] en el que desarrollan un SMA para la composición de servicios verticales en una ontología que proporcione interoperabilidad y un modelo de composición basado en redes de Petri, redes que han sido ampliamente utilizadas en este campo [230].

La combinación de estos tres tipos de componentes (además de la autenticación multicloud [231]) habilita uno de los grandes objetivos perseguidos en el marco de la investigación de los sistemas CC, como es la federación [114, 139, 140, 231]. Esto es, tal y como se ha estudiado (apartado 2.4.1), la posibilidad de que un mismo consumidor pueda trabajar con varios proveedores de forma simultánea para una misma tarea, pudiendo alterar su relación contractual con cada proveedor de servicio de forma dinámica. Entre los trabajos en el marco de los sistemas CC federados destacan, los propuestos por Singh *et* Malhotra [232] que proponen un modelo basado en agentes inteligentes para la provisión de servicios de infraestructura. Chen *et al.* propone un modelo de interacción entre CCs basado en agentes móviles [233], abstrayendo de la complejidad a los usuarios. Un enfoque mucho más avanzado que incluye experimentación en un entorno real es el presentado por [234] y basado en el *framework Cloud Agency* [197] previamente presentado; en él, se utiliza un SMA para la provisión, gestión, ejecución y reconfiguración de servicios CC de tipo infraestructura proporcionados por terceros. Otros dos modelos similares, pero adaptados a cualquier tipo de servicio, no sólo de infraestructura, son los desarrollados de forma paralela por [235, 236] en el que los servicios migran entre entornos CC con el objetivo de satisfacer los SLA acordados con los consumidores. Otra perspectiva es la posibilidad de distribuir las tareas entre varios entornos de forma simultánea tal y como propone Palmieri *et al.* [237], configurando un modelo altamente escalable. Finalmente, una aproximación diferente es de SERA [238] en la que el modelo distribución de servicios en múltiples entornos CC se basa en la negociación con varios entornos CC utilizando un motor de inferencia ontológico.

Aunque el número de trabajos existentes en este campo es muy amplio, en todos ellos se observa un problema de fondo, y es la dependencia de entornos CC. La búsqueda y composición de servicios no es posible si los entornos CC no proporcionan las herramientas que lo hagan posible. Así mismo, en cuanto al aseguramiento de la calidad de los servicios, si que es posible medir el grado de cumplimiento de la calidad acordada mediante SLA. Sin embargo, sin un control de la infraestructura subyacente no es posible asegurar el cumplimiento el nivel de calidad en los servicios.

3.2.3.Cloud Provider

El rol *Cloud provider* es aquel encargado de proporcionar servicios a terceros, y el más interesante desde el punto de vista del trabajo de tesis que se propone. Las tareas asignadas a este rol son fundamentalmente el control y organización de las plataformas CC a nivel interno, lo que implican tanto el despliegue y orquestación de servicios, como la gestión de la infraestructura tecnológica subyacente. En este sentido, es necesario asegurar al mismo tiempo la privacidad y seguridad tanto del entorno, como de la información que se gestiona; además de satisfacer los acuerdos SLA acordados con los usuarios.

A continuación, se presentan los trabajos existentes más importantes agrupados tal y como sigue:

- **Seguridad y privacidad.** Un entorno computacional de altas prestaciones, como es el caso de un entorno CC está sometido a una gran variedad de ataques [130]. En estos entornos, la seguridad puede verse a nivel interno, como medio para asegurar que toda la infraestructura subyacente funciona correctamente; y también a nivel externo, que hace énfasis en las comunicaciones con el exterior, validando que no comprometan al propio entorno.

  Como no podía ser de otro modo, los SMA cubren ambos puntos de vista. A nivel externo, los trabajos están principalmente relacionados con la autenticación de los usuarios en el entorno CC, como el proyecto ABAC [239], el modelo de autenticación en el lado del cliente [240], o el trabajo propuesto por [241] en el que un SMA se basa en políticas de acceso para conceder privilegios sobre la información. A nivel interno, por un lado se observan trabajos iniciales en diferentes aspectos como la privacidad de los datos [242], la monitorización de la plataforma [134] y la seguridad de la infraestructura hardware (real o virtual) [243, 244]. Por otro lado, los trabajos que consideramos más maduros están relacionados con el almacenamiento seguro de los datos, destacando el proyecto de almacenamiento CloudZone [245-250] que está basado en un SMA para asegurar la privacidad, acceso a la información y otros parámetros de seguridad en el almacenamiento. Unido a este proyecto, los mismos autores han diseñado la metodología Prometheus [251] que pretende garantizar la seguridad en el almacenamiento de datos en entornos CC. Existen otros trabajos, como el presentado por Govinda *et* Sathiyamoorthy [252] en el que se trata de garantizar la seguridad de información mediante la ofuscación de los datos. Finalmente, el último trabajo a destacar es el modelo de seguridad multicapa para el acceso a la información persistente que también esta basado en agentes [253, 254].
- **Oferta de Servicios.** Todo el marco tecnológico y comercial que envuelve al paradigma CC está orientado a ofrecer servicios computacionales, con independencia de su tipo o complejidad. En este sentido en el estado del arte se pueden encontrar diferentes tipos de servicios gestionados por un SMA aplicados a entornos de oficina [255], gestión eficiente de energía [256, 257], seguridad [183], e-learning [182, 258, 259], etc. Incluso, existen perspectivas diferentes en la que se proponen ofrecer los propios agentes como servicio AaaS [260].

  Sin embargo, cuando se habla de servicios en el marco del paradigma CC, es necesario cumplir las restricciones y objetivos acordados con los consumidores en el SLA, y por lo tanto los servicios tienen que monitorizarse para cumplir los niveles de calidad, en este sentido se ha acuñado el concepto de servicios sensibles a la calidad (QoS-aware)[261]. En este campo los SMA tienen una gran aplicación y por ello han surgido en los últimos años diferentes trabajos relacionados [262-264], aunque todos ellos son trabajos iniciales, que abordan el problema de forma parcial y que tratan de satisfacer las demandas en la calidad

del servicio. Más allá de la monitorización, se hace necesario disponer de la capacidad de gestión de la infraestructura subyacente que es la que aporta potencia computacional y permite satisfacer la calidad demandada por un servicio dado en un tiempo establecido.

- **Gestión de la infraestructura subyacente.** La gestión de la infraestructura subyacente es una de las tareas más complejas debido a la incertidumbre, dinamicidad y heterogeneidad de un entorno en el que existen diferentes máquinas físicas, sistemas de virtualización, topologías de red, modelos de almacenamiento, etc.; además de la complejidad en el hardware y software existente.
  Los SMA tienen un gran campo de aplicación en este contexto, sin embargo, no existen muchos ejemplos en el estado del arte. No obstante, es posible encontrar trabajos preliminares en el que un SMA gestiona recursos en función de los acuerdos SLA de forma centralizada [36], el trabajo propuesto por Wei *et* Blake [265] en el que se asignan recursos a servicios sigue un modelo de flujos de trabajo, y finalmente, la arquitectura MASCloud que utiliza un SMA para optimizar costes en el marco del paradigma CC [266].

A excepción de los relacionados con la seguridad de sistemas, el número de trabajos existentes en el que los SMA toman el rol de *Cloud Provider* son mucho menores que en el caso de roles anteriores, además, su grado de madurez es menor. Así es muy difícil encontrar SMA para el control de infraestructura subyacente, la orquestación de servicios y recursos computacionales y la coordinación del entorno, en general. Esto sin duda es debido a la novedad de las tecnologías subyacentes empleadas como es el caso de la virtualización hardware, la monitorización de sistemas de alta disponibilidad en clúster, etc.

3.2.4.Cloud auditor

Finalmente, el Cloud auditor es el rol cuyas responsabilidades consisten en asegurar, desde un punto de vista externo y objetivo, que los acuerdos SLA se cumplen satisfactoriamente. Este es el grupo donde existen menos trabajos en lo que se pueden ver la relación entre SMA y entornos CC, observándose únicamente dos referencias destacables, la arquitectura FOSII [267] en el que una arquitectura multiagente detecta violaciones en los SLA de forma automática; y el trabajo de Ramaswamy *et al.* [268] en el que un SMA, en el lado del cliente detecta posibles violaciones en los acuerdos SLA.

3.2.5.Conclusión

Después de analizar la relación existente entre ambos sistemas computacionales, es posible concluir que hoy en día el uso común de SMA en entornos CC continúa siendo incipiente. Aunque por la variedad de trabajos presentados se puede afirmar que existe un gran interés en la comunidad científica en el desarrollo de técnicas y herramientas que aúnen las ventajas de ambos entornos tecnológicos.

Así, cuando los SMA toman el rol de *Cloud Consumer*, el entorno CC ofrece tecnología de altas prestaciones, con un alto rendimiento, disponibilidad y escalabilidad [52]. Permiten y facilitan la aplicación de SMA en una gran variedad de aplicaciones complejas. El uso de las capacidades computacionales permite expandir el modelo de razonamiento y conocimiento de los SMA tradicionales, ya que reduce las restricciones temporales y ofrece diferentes modelos de almacenamiento de alto rendimiento. En la Tabla 3 se presenta un resumen con los principales usos de los servicios CC en el marco de los SMA, incluyendo el tipo de servicios que se utiliza y si estos servicios sirven para proporcionar otros servicios a terceros. Como se aprecia, los principales usos están realizados con los servicios de infraestructura, computación y persistencia. Existiendo un

buen número de trabajos que utilizan los servicios CC para proveer servicios más complejos a los usuarios finales.

Usos principales	Tipo de servicio utilizado			Proporciona servicios terceros
	SaaS	PaaS	IaaS	
Persistencia y Computación				
Servicios BigData			X	
Análisis de información			X	
Servicios Grid Computing			X	
Uso con dispositivos móviles		X	X	X
Ejecución de tareas			X	X
Servicio de infrastructura				
Entorno de simulación		X	X	X
Uso con redes de sensores		X	X	X

Tabla 3.- Resumen SMA y CC, rol Cloud Consumer

En segundo lugar, cuando los SMA toman el rol de *Cloud Auditor*, o *Cloud Broker*, actúan como terceras partes (intermediarios) que intervienen en la relación comercial existente entre usuarios y consumidores, también disponen de una gran aplicación, principalmente como *Cloud Broker*. En este sentido, destacan las facilidades que ofrecen los SMA para la búsqueda, selección de servicios, así como para la negociación automatizada de los acuerdos en la calidad del servicio con diferentes proveedores de forma simultánea. Incipientemente se observa un rápido crecimiento del uso de SMA en la composición en servicios en nube (*Cloud manufacturing*). El uso de agentes facilita la tarea de composición de servicios ofertados por diferentes proveedores CC, lo que en sí mismo constituye una aportación relevante hacia el objetivo de federación e interoperabilidad de entornos CC, el denominado InterCloud [139].

Aplicaciones principales	Nivel externo			
	Negociac. SLA	Oferta servicios	Composic.	Evaluación Servicios
Búsqueda de proveedores				
Buscador de servicios		X	X	
Interoperabilidad	X	X		
Modelo de confianza en SLA	X		X	
Contratación de servicios	X		X	
Negociación de acuerdos SLA				
Negociación de condiciones		X		X
Monitorización de acuerdos	X		X	X
Composición de servicios				
Búsqueda de servicios compatibles		X	X	
Interoperabilidad		X		
Composición Horizontal y Vertical		X		

Tabla 4.- Resumen SMA y CC, rol Cloud Broker

Finalmente, cuando actúan como *Cloud Provider*, la principal aportación de los SMA se relaciona con la seguridad y privacidad de los datos, gracias a la capacidad de los agentes para monitorizar, razonar y responder proactivamente a los cambios en el entorno. Sin embargo, pese a que existen diferentes trabajos relacionados que ofertan servicios de tipo CC, tan sólo es posible encontrar en el estado del arte prometedores e incipientes trabajos relacionados con la calidad en el servicio y la provisión de recursos computacionales. Teniendo en cuenta que el entorno tecnológico que envuelve a los proveedores de servicios CC es un entorno abierto por su dinamicidad, heterogeneidad e incertidumbre; no cabe duda que la aplicación de SMA en este contexto añade características relevantes como la detección automatizada de fallos, monitorización de servicios y operaciones avanzadas, negociación automatizada en la calidad del servicio, interoperabilidad, planificación dinámica, gestión eficiente de recursos, aprendizaje, etc.

Usos principales	**Nivel interno**			
	Monitoriz.	**Control**	**Seguridad Privacidad**	**Gestión Infraest.**
Seguridad y privacidad				
Monitorización de infraestructura	X			
Modelos de autenticación			X	
Privalegios de acceso			X	
Seguridad de la información			X	
Almacenamiento seguro			X	
Oferta de servicios				
Servicios sensibles a la calidad	X	X		
Oferta de servicios				
Gestión de nfrastructura				
Gestión de recursos	X			
Optimización de costes		X		X
Gestión de flujo de trabajo				X

Tabla 5.- Resumen SMA y CC, rol Cloud Provider

Más allá de los trabajos en los que los SMA hacen uso de servicios CC generalistas (rol *Cloud User*); los trabajos que son más interesantes en el marco de esta tesis doctoral son aquellos en los que los SMA se aplican a problemas o debilidades existentes en los sistemas CC. Estos trabajos están basados en la experiencia en tecnologías relacionadas o precedentes y su aplicación en entornos CC es directa. Sin embargo, lo que se observa es que todos ellos solucionan problemas parciales o específicos, pero que no abordan el control de un entorno CC de forma integral, haciendo frente a todos los retos existentes de forma conjunta (gestión de infraestructura, acuerdos SLA, orquestación de servicios, etc.). Las tablas anteriormente presentadas (Tabla 4 y Tabla 5) dan una idea clara de este problema, ya que las soluciones existentes en el marco del rol *Cloud Broker* abordan los problemas abiertos desde una perspectiva externa, mientras que aquellos trabajos enmarcados como *Cloud Provider* se centran más en las cuestiones internas de un entorno Cloud. En definitiva, no existen soluciones integrales que hagan frente a los problemas desde ambos puntos de vista. En este sentido, los SMA que pretendan gestionar un entorno CC deberán abordar los problemas existentes como un todo, de forma que sea posible abarcar todos los aspectos que estos complejos sistemas requieren.

Para llevar a cabo esta tarea, un entorno CC de última generación basado en un SMA debe hacer uso de las últimas innovaciones existentes en el marco de los SMA, es decir, las OV [24, 269] y las modernas metodologías de desarrollo de software alineadas con modelos de ingeniería de desarrollo de software complejos que aseguran el desarrollo de productos de alta calidad, estabilidad y escalabilidad. Este tipo de sistemas se consideran adecuados en este contexto, ya que permiten dividir las diferentes responsabilidades entre grupos de agentes con objetivos específicos y modelos de razonamiento adecuados a cada problema. No obstante, pese a poder modelar el sistema como grupos de trabajo con responsabilidades y objetivos independientes, todos trabajan en conjunto para conseguir los objetivos globales del sistema visto como un todo.

En el estado del arte no se han encontrado referencias en las que se apliquen modelos sociales en el contexto CC, por lo que en el siguiente apartado se estudiarán las OV de agentes, y las metodologías de desarrollo de software con el objetivo de determinar cuál es el mejor modelo de desarrollo de una plataforma CC basada en SMA organizativos (*Agent-based Cloud platform*).

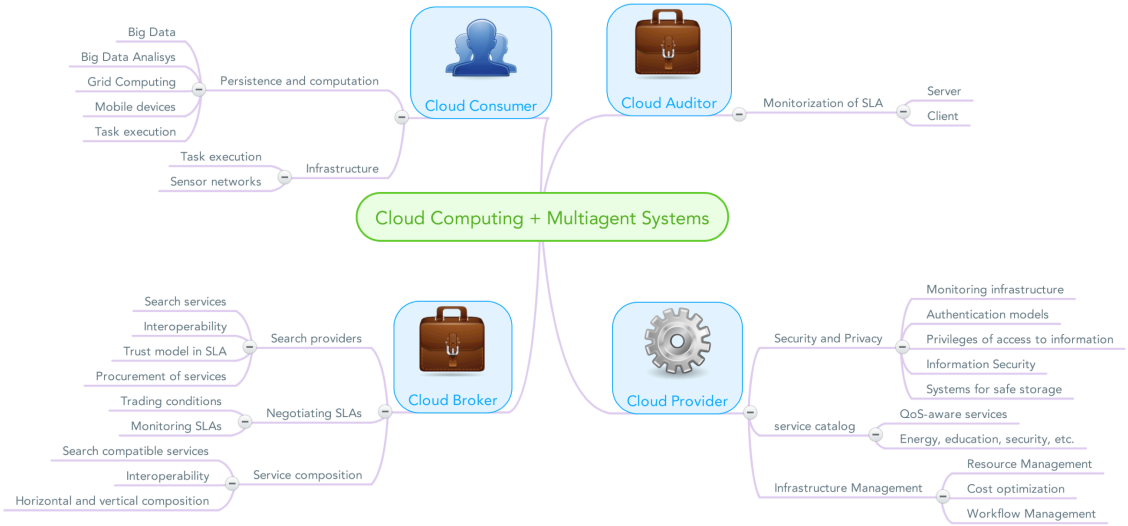

Figura 10.- Cloud Computing y sistemas multiagente

3.3.Organizaciones de agentes

En los anteriores apartados de este capítulo se ha realizado una revisión histórica que facilita al lector la comprensión de los sistemas CC, así como la incipiente relación existente con los SMA. Así, después de analizar las SMA en el marco de los entornos CC durante el apartado anterior se concluyó que los SMA basados en sistemas organizativos constituyen la solución más adecuada para facilitar su implantación en un sistema CC. Por ello, a lo largo de este apartado se hará una revisión de los SMA que siguen un modelo organizativo para el diseño de sistemas software.

En la literatura se distinguen dos aproximaciones en cuanto al diseño de SMA [204, 270, 271], estas son las que se centran en el propio agente o en la organización. En el primero, son los propios agentes los que incluyen sus objetivos individuales, relaciones, etc.; mientras que en la segunda aproximación, existe un nivel de abstracción mayor, el de la organización, que es el que define las tareas, objetivos globales y relacionales durante la fase de diseño.

La primera aproximación, quizás mucho más tradicional y ligada a los SMA clásicos, sólo tiene en cuenta la perspectiva funcional en el diseño, por lo que se hace mayor énfasis en los modelos y algoritmos que permiten abordar problemas complejos [272, 273]. Siguiendo esta perspectiva, resulta difícil predecir el comportamiento individual, y por ende, el comportamiento agregado del sistema en su conjunto [274]. Este hecho, unido a la capacidad de situarse en un sistema abierto, que es intrínsecamente impredecible, convierte a los SMA en entornos difíciles de aplicar en entornos reales por la dificultad de coordinación [275].

Sin embargo, la segunda aproximación, trata de resolver el problema de diseñar sistemas software de gran envergadura [25], añadiendo al paradigma tradicional nuevos capacidades como roles, normas, organizaciones, etc. que maximizan sus cualidades intrínsecas (autonomía, habilidades sociales, reactividad y pro-actividad). Este conjunto de aproximaciones se puede dividir en los siguientes grupos en función de los mecanismos utilizados [269]:

- **Mecanismos basados en interacciones directas.** Proporcionan principios básicos de comunicación y cálculos locales de los agentes para proporcionar un estado global del sistema [39].
- **Mecanismos basados en interacciones indirectas.** Consiguen comportamientos complejos a partir de interacciones entre los agentes a través del entorno [43].
- **Mecanismos basados en refuerzo.** Basados en la capacidad de los agentes para modificar su comportamiento en función de las entradas (positivas o negativas) [41].
- **Mecanismos basados en cooperación.** Basados en la descomposición de agentes para llevar a cabo tareas complejas; y la composición de agentes para llevar a cabo tareas simples [42].
- **Mecanismos basados en arquitecturas genéricas.** Basadas en meta-modelos y arquitecturas de referencias que son modificadas (automática o manualmente) en función del contexto de aplicación [43].

Dentro de este conjunto de propuestas, el modelo que más ha sido estudiado en los últimos años es, sin duda, el de las OV de agentes [147]. Este tipo de mecanismos modelan los SMA siguiendo un esquema semejante al de las organizaciones humanas, acotando la impredecibilidad del sistema dentro de un conjunto de normas institucionales, haciendo posible la evolución del SMA gracias al uso de sociedades artificiales como abstracción de nivel superior [276]. Las OV se enmarcan dentro del mecanismo de comunicación directa, sin embargo, las últimas referencias en la literatura apoyan

la definición de estas organizaciones mediante arquitecturas genéricas (meta-modelos) [277], así como en el uso del entorno no sólo como medio de interacción indirecta sino también como parte del propio SMA [278].

El nacimiento de esta nueva metodología para diseñar SMA está motivado por la autonomía de los propios agentes, lo que les convierte en entidades que, por su naturaleza son impredecibles. Este hecho conlleva a que el propio SMA se convierta en un entorno cuyo nivel de variabilidad está condicionada por el medio en el que se sitúa [25], esta característica se acentúa aún más si cabe en un sistema CC donde el entorno es altamente complejo e impredecible, dependiente en gran medida de las condiciones externas y las necesidades de los usuarios finales.

El uso de un nivel mayor de abstracción para el diseño de SMA en el marco de estos entornos complejos, mediante el uso de las OV de agentes inteligentes, acota la autonomía de los agentes a los intereses comunes de la organización de la que forman parte, más allá de sus intereses individuales. De este modo, un agente dentro de una sociedad necesita considerar no sólo su propio entorno, sino también el comportamiento del sistema como un todo, de forma que unos agentes influyen en el comportamiento del resto.

Hoy en día, como no podría ser de otro modo, el diseño de SMA se orienta al uso de modelos organizativos que facilitan el diseño e implementación de los sistemas abiertos [279]. Tal y como ya se ha indicado, los sistemas abiertos se caracterizan por la heterogeneidad de sus participantes, la limitación en cuanto a la confianza debido a la existencia de objetivos individuales en conflicto, y finalmente debido a la gran probabilidad de disconformidad con las especificaciones [280]. El dinamismo es otra de las características clave de este tipo de sistemas [24]. Estas dos propiedades (dinamismo y heterogeneidad) provocan que modelar un sistema abierto constituya hoy en día un reto [39, 41, 43, 281-283]. Para abordar el diseño de este tipo de sistemas es necesario disponer de nuevos modelos y mecanismos, que permitan abordar el diseño de sistemas con capacidad de adaptación dinámica a cambios en la organización y en el propio entorno [284]. Además, estas nuevas teorías, deben estar apoyadas por herramientas que guíen el proceso de diseño.

Antes de detallar en profundidad todos los aspectos que envuelven a esta aproximación de diseño y desarrollo de SMA, se presentará la Teoría de Organizaciones humanas que sustenta el diseño SMA modernos.

3.3.1.Organizaciones artificiales, una perspectiva sociológica

Las estructuras organizativas han sido estudiadas por la Teoría de la Organización [285, 286], con el objetivo de comprender la estructura y el diseño de una organización, así como las Alianzas Estratégicas [287, 288] entre los diferentes integrantes que permitan alcanzar los objetivos perseguidos. Estos estudios son relativamente recientes, ya que no han existido grandes organizaciones hasta los inicios del siglo XX.

Antes de continuar, con el análisis es necesario definir el concepto principal, la organización [286]:

> *Una organización es una entidad social conscientemente coordinada, con un límite relativamente identificable, que funciona de forma continua para lograr una o un conjunto de metas comunes.*

El estudio de las organizaciones trae consigo innumerables ventajas, entre las que cabe destacar [289]:

- Permite producir bienes y servicios de forma eficientemente.

- Reúne los recursos para alcanzar las metas deseadas.
- Facilita la innovación.
- Permite considerar la influencia que conlleva los cambios en el entorno.
- Crea un valor para los clientes, empleados y propietarios.
- Permite gestionar los retos dentro de un entorno social diverso, ético, coordinado y motivado.

El estudio o diseño de una organización es un proceso complejo, para el cuál es necesario tener en cuenta una gran cantidad de puntos de vista que se organizan en las siguientes dimensiones [289]:

- **Dimensiones estructurales.** Proporciona un método para describir las características internas de la organización. Entre las principales subdimensiones estructurales, cabe destacar: la formalización, la centralización, la especialización, la jerarquía de autoridad, el profesionalismo y las proporciones de personal.
- **Dimensiones contextuales.** Caracteriza a la organización como un todo, incluyendo tamaño, tecnologías, entorno y objetivos. Entre las dimensiones contextuales cabe destacar la cultura, el entorno, los objetivos y estrategia, el tamaño y la tecnología.

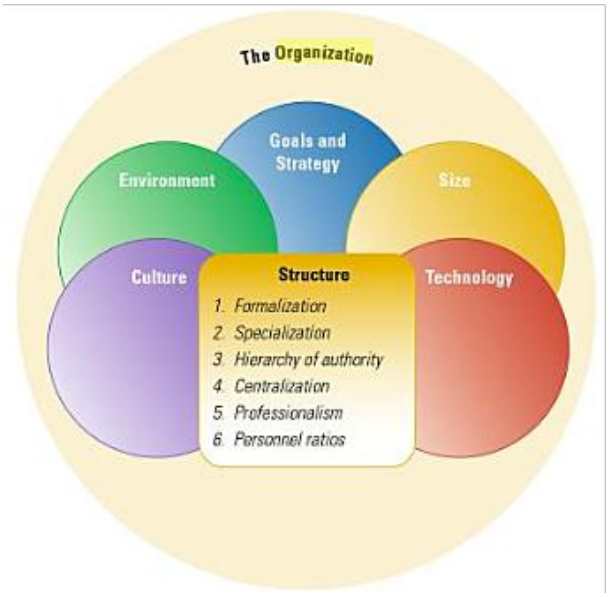

Figura 11.- Teoría de la organización [289]

Si se extrapola el concepto de organización (según la teoría de organizaciones) al contexto de los SMA, una organización puede verse como un conjunto de entidades reguladas por mecanismos de orden social, que persiguen objetivos comunes. Además, la teoría moderna de organizaciones, y la evolución de los SMA convergen basando su desarrollo en los mismos principios directores [289]:

- Se evoluciona el control formal de los sistemas y la información, hasta el desarrollo de organizaciones con información compartida y distribuida.
- La toma de decisiones y el control que gobierna la organización sigue una estrategia colaborativa, en detrimento de las estrategias competitivas previas.
- La organización es en sí misma una estructura que evoluciona y que se adapta a los cambios que se producen en el entorno, promoviendo la cultura de la adaptación en detrimento de las estructuras rígidas.

La funcionalidad principal de una organización, es permitir a sus miembros coexistir en un entorno compartido y llevar a cabo sus respectivos objetivos, cooperando (o no) con el resto de miembros. Este entorno común de consecución de los objetivos individuales permite que el diseño de SMA basado en el concepto de organización facilite su comprensión, diseño y por lo tanto su posterior implementación [290-292]. No obstante, aunque existe una gran cantidad de elementos convergentes en ambas teorías, también se pueden encontrar una gran variedad de diferencias, tal y como sigue [293, 294]:

- En las sociedades artificiales los agentes emulan comportamientos que tratan de satisfacer un conjunto de requisitos, mientras que en la Teoría de Organizaciones está basada en un conjunto de modelos matemáticos.
- La Teoría de Organizaciones modela funciones de utilidad asumiendo el comportamiento racional de los actores. Sin embargo, los SMA utilizan un modelo de representación donde los actores pueden tomar decisiones racionales, pero también no racionales, dependiendo de la representación que tengan los agentes sobre su entorno.
- La Teoría de Organizaciones representa a los actores de forma homogénea, en cambio, los SMA se sitúan en entornos abiertos, permitiendo la agregación de una gran cantidad de sistemas con modelos arquitectónicos diferentes.

Todas estas diferencias, invitan a pensar que aunque las sociedades artificiales y las OV están sustentadas inicialmente sobre la Teoría de Organizaciones, en realidad conforman una nueva teoría, acompañada de un conjunto de modelos y metodologías para el desarrollo de sistemas software complejos.

3.3.2.Sociedades artificiales y organizaciones virtuales

Las sociedades artificiales y las OV son términos estrechamente relacionados, a lo largo de este apartado se clarificará la relación entre los conceptos tomando como base la literatura existente. En primer lugar, se presentará el concepto de Sociedad Artificial, presente en los resultados de una gran variedad de autores [23, 276, 290, 295]. Así, por ejemplo, Artikis *et* Pitt [290] caracterizan formalmente una sociedad artificial en base a un conjunto de agentes, restricciones en la sociedad, un lenguaje de comunicación, los roles que los agentes pueden desarrollar y el conjunto de propietarios (*owners*) de los agentes. No obstante, la definición más certera desde nuestro punto de vista es el que proponen Annunziato *et* Pierucci [276]:

> *Una sociedad artificial está definida como un conjunto de entidades artificiales interrelacionadas e interactuantes, que se rigen bajo determinadas reglas y condiciones*

Dentro de la literatura, Davidsson *et* Johansson [23, 295] propone una clasificación de las sociedades artificiales basándose en la (i) apertura, que es la posibilidad de que un agente pueda unirse a la sociedad; la (ii) flexibilidad, que indica el grado de restricción que tiene el agente en su comportamiento debido a las normas de la sociedad; la (iii) estabilidad, que es una medida de la previsibilidad de las acciones; y, finalmente, la (iv) confianza, que mide el grado en el que el agente puede confiar en la sociedad.

A partir de estas características Davidsson *et* Johansson propone la existencia de cuatro tipos de sociedades. Los dos primeros tipos de sociedades son los siguientes:

- **Sociedades abiertas.** Que son flexibles lo que las convierten en sociedades poco estables y en las que es difícil confiar. En este tipo de sociedades los agentes entran en la organización sin ningún tipo de restricción, para ello tan sólo tienen que interaccionar con

alguno de los agentes que la forman. Para que la comunicación entre las diferentes entidades se pueda llevar a cabo deben compartir el mismo lenguaje de comunicación y existir un conjunto mínimo de roles. En el caso de no tener este mínimo conjunto de características (comunicación y roles) serían sociedades anárquicas.

- **Sociedades cerradas**. Al contrario que las sociedades abiertas son estables y confiables. Este tipo de sociedades pueden denominarse también sociedad fija o inalterable.
  Es el tipo de sociedad más utilizada para el diseño de SMA clásicos, ya que en ellas existe un conjunto de agentes que se conocen a priori, cada agente tiene sus propios objetivos (que no tienen que ser comunes) y que interaccionan de forma distribuida para resolver problemas, teniendo en cuenta que pueden confiar entre sí.

Para el desarrollo de sistemas organizativos estas sociedades no son las más apropiadas, debido a sus restricciones intrínsecas, ya sea por la flexibilidad, o la falta de ella. Las OV se basan en otros dos tipos de sociedades más flexibles, permitiendo la entrada de nuevos miembros, donde la estabilidad de la misma se regula a través de algún tipo de mecanismo o institución:

- **Sociedades semi-abiertas.** En la que se introduce el concepto de portero, las entidades externas que pretendan acceder a la sociedad deberán contactar con dicho portero y demostrar que son entidades seguras, o al menos que su grado de seguridad no van alterar la confiabilidad y estabilidad del sistema.
  Adicionalmente, estas entidades también pueden monitorizar a los agentes internos, lo que permite acotar a la propia sociedad y además auditar el estado de la misma
- **Sociedades semi-cerradas.** En este tipo de sociedades, los agentes no están autorizados a entrar. Sin embargo, si que puede existir un agente de la propia sociedad que realice las tareas del agente externo. Igualmente, existe una entidad que interactúa con los agentes externos, y que crea los agentes internos mediadores dentro de la sociedad para realizar las tareas del agente externo.

Una vez que se ha acotado el concepto de sociedad artificial, el otro concepto, el de organización puede verse como un conjunto de entidades reguladas por mecanismos de orden social, de forma que habiliten la consecución de objetivos comunes. Al igual que con las sociedades, existe una gran variedad de literatura acerca de las organizaciones de agentes [24, 271, 296, 297], así por ejemplo, Wooldridge *et al.*, [297] enuncian la siguiente definición:

> *Una organización es un conjunto de roles con cierta relación entre sí, que interaccionan con otros roles a través de unos patrones institucionales sistematizados.*

No obstante, desde nuestro punto de vista, la mejor forma de definir una organización de agentes es atendiendo a sus características[271]:

- Una organización está formada por agentes (individuales) que manifiestan su comportamiento.
- La organización puede ser dividida en suborganizaciones que pueden estar solapadas (suborganizaciones o grupos).
- El comportamiento de los agentes está relacionado a nivel funcional con la organización como un todo, definiendo su funcionalidad como un rol a tener dentro de la organización.
- Los agentes se relacionan de forma dinámica.
- Los tipos de comportamiento se relacionan mediante enlaces entre roles, tareas y protocolos.

Dentro de estas características, el concepto clave es el de rol, que es una descripción abstracta del comportamiento de los agentes, incluyendo obligaciones, restricciones y habilidades. Así mismo, el rol debe incluir la descripción de los patrones de interacción entre los diferentes agentes que conforman el sistema. No obstante, tal y como se presentará a continuación los modelos organizativos van mucho más allá, introduciendo una gran variedad de capacidades que permiten el uso de SMA en una gran variedad de contextos.

Este enfoque parece adecuado al problema a modelar en el marco de este trabajo de investigación ya que es posible introducir nuevas técnicas para el control y monitorización de sistemas CC donde la colaboración y las estrategias competitivas sustituyen al modelo jerárquico en la toma de decisiones. Gracias a lo cuál es posible que los agentes desarrollen modelos de información compartida y distribuida, lo que permite reducir el espacio de datos necesario para llevar a cabo el proceso de decisión, reduciendo, por tanto, la complejidad en la toma de decisiones. Además, no sólo se reduce su complejidad, sino que el modelado basado en roles y objetivos permite hacer frente a la autonomía de los individuos que forman parte de la organización, ya que las decisiones de los agentes individuales no sólo atenderán a las intenciones individuales que les permitan satisfacer sus deseos personales, sino que también estarán orientadas a satisfacer los objetivos globales de la organización u organizaciones de las que forman parte. Finalmente, la abstracción e independencia del entorno de la organización y los agentes permite no solo abstraer su complejidad, sino también reducir la dependencia entre los órganos de control y actuación.

3.3.3.Metodologías y modelos organizativos

En los últimos años, el auge de los conceptos organizativos ha experimentado un proceso de rápido desarrollo. Así en el estado del arte se pueden encontrar una gran variedad de modelos y metodologías para su desarrollo [298], algunos de ellos toman como base las sociedades artificiales. El desarrollo de estas metodologías, cuyo objetivo es facilitar las tareas de diseño software basado en agentes, también ha propiciado la evolución de la propia Teoría de Agentes [38].

El proceso de desarrollo de un sistema CC mediante el uso de estas novedosas técnicas requiere que, inicialmente, se realice un estudio de los principales modelos existentes basados en OV de SMA. Esto es así debido a que cada metodología incluye herramientas, técnicas y modelos propios para el modelado, pero en la mayoría de los casos estas metodologías tienen una gran cantidad de rasgos comunes, en lo que diferentes autores han definido como meta-modelo del paradigma multiagente [299, 300].

Así, gracias a este estudio previo será posible seleccionar aquella metodología y su metamodelo asociado que mejor se adapte al problema a resolver. Por tanto, a continuación se presenta una breve descripción de las corrientes metodológicas principales:

- **MAS-CommondKADS** [301] se basa en la metodología de propósito general CommondKADS [302]. Fue una de las primeras metodologías de desarrollo de software en el marco de los SMA. La metodología se fundamenta en el desarrollo de siete modelos principales: agente, tareas, experiencia, organización, coordinación, comunicación y diseño.
- **MESSAGE/UML** [303] fue una de las primeras metodologías que incluyó el ciclo de vida del desarrollo software y el modelado mediante UML.
- **GAIA** [297] es, sin lugar a dudas, una de las principales metodologías de desarrollo de SMA y que ha sido ampliamente utilizada en diferentes trabajos y contextos [304-307]. Posteriormente a su nacimiento, la metodología fue revisada para incluir conceptos organizativos, la noción de entorno y un conjunto de herramientas y técnicas que faciliten

su uso en entornos abiertos; a esta revisión se la ha denominado **GAIA II** o **GAIAexOA** [25]. Finalmente, existe una nueva actualización más reciente que se ha centrado en el diseño de sistemas adaptativos [308].

- **ROADMAP** [309] que partiendo de la metodología GAIA original sigue su misma evolución de forma paralela, ya que añade los conceptos y técnicas necesarias para su uso en entornos abiertos. Posteriormente, se refina esta versión inicial para tener en cuenta la adaptación dentro de estos entornos abiertos [310].
- **PROMETHEUS** [311] es una metodología que centra sus intereses en cubrir todas las fases del diseño y desarrollo desde el punto de vista de la Ingeniería del Software, siguiendo un esquema similar al software tradicional. El objetivo es acercar el modelado de SMA a su posterior implementación. No obstante, el modelo propuesto en PROMETHEUS es limitado, aunque trabajos más modernos han demostrado que puede ser utilizado favorablemente si se combina con metodologías más modernas [312] como INGENIAS [313] que se presentará a continuación.
- **AALADIN** [314] fue una de las primeras metodologías. Su principal característica es su simplicidad, ya que sólo incluye tres conceptos principales: Agente, Grupo y Rol. Precisamente estos tres conceptos son la base de la segunda versión de la metodología, denominada AGR (*Agent-Group-Role*)[271] que se centra en el diseño basado en modelos organizativos. La última revisión de esta metodología es el modelo AGRE (*Agent-Group-Role-Environment*) [315] que incluye el concepto de entorno como novedad.
- **SODA** [316] propone una metodología de desarrollo de sistemas software orientados a Internet, en la cuál se tiene en cuenta tanto los aspectos del propio agente, como de los de la sociedad, y en el que también está presente el concepto de entorno. Para ello, la metodología se centra en las tareas que un agente o conjunto de agentes tienen que realizar. Recientemente esta metodología ha sido adaptada [317] para su uso con SPEM 2.0 [318] que es una metodología de desarrollo software.
- **MOISE** también puede ser considerada una de las principales metodologías, además de ser una de las pioneras. Se distingue del resto en que no sólo se centra en los conceptos sociales, sino que también tiene en cuenta las características del agente y las tareas a realizar (misiones) [270]. Posteriormente a su nacimiento, también ha sido refinada para cubrir las debilidades encontradas en la versión inicial, como la falta de planes globales y la dependencia entre estructura y funcionalidad. Esta nueva versión se ha denominado MOISE+ [204]. Recientemente también esta segunda versión se ha complementado [319] incluyendo la Teoría de Agentes y Artefactos (*Agents and Artifacts –A&A–*) [278, 320]. Esta teoría (o metodología) facilita la tarea de modelar el entorno, permitiendo encapsular las interacciones del SMA con el exterior a través de un nuevo componente denominado artefacto. Esta aproximación ha dado lugar a la metodología ORA4MAS [319] que trata de articular la adaptación de la propia organización.
- **TROPOS** [321, 322] es una metodología que difiere de las anteriores en que propone el desarrollo de organizaciones multiagente basándose en la teoría de organizaciones humanas [285]. En este sentido, aboga por el uso de una de las siguientes topologías *Flat-Structure*, Cadena de valores, Pirámide, Matriz, *Structure-in-Five*, *Co-optation*, *Joint Venture*, *Bidding*, *Arm's Length* y *Hierarchical Contracting*. La metodología se basa en el entorno de modelado i* [323]. Al igual que las anteriores ha sido refinada utilizando para ello las últimas referencias en el marco de la teoría de organización [147].
- **OperA** [44] es una metodología desarrollada desde su primera versión teniendo en cuenta el concepto de sociedad, objetivos, estructuras y normas que rigen esta familia de

metodologías de desarrollo software. Su principal aportación consiste en el soporte que ofrece a la dinamicidad y heterogeneidad del tiempo de vida de los agentes en el marco de los sistemas abiertos. La metodología fue extendida en el modelo OMNI [324] para integrar el marco normativo de HARMONIA [325] el cuál aporta una mayor potencia semántica a la definición de las normas de una sociedad de agentes.

- **ADELFE** [326, 327] sigue el Proceso Unificado [328] para el desarrollo de SMA incluyendo las fases de requisitos, análisis y diseño. También utiliza la notación UML [329] y AUML [178] para la descripción de los sistemas construidos.
- **PASSI** [330] es una metodología para el desarrollo software que incluye todas las fases en el desarrollo software y que, además, es iterativa. Sin embargo, no está adaptada a los conceptos de sociedad artificial, entorno, normas, etc. Su segunda versión [331] se centra en la inclusión de los principios de las metodologías ágiles [332].
- **OMACS** [299, 333] es la que se puede denominar una metodología de nueva generación, ya que no sólo está desarrollada sobre la base que proporcionan las sociedades de agentes, sino que modela la propia reorganización de la sociedad incluyendo cambios estructurales en la misma.
- **MASE** [334, 335] fue una de las primeras metodologías. La principal innovación en el momento de su nacimiento fue la inclusión de las fases habituales del desarrollo software. No obstante, pese a ello, el modelo desde un punto de vista de agentes era limitado. Para superar estos problemas, recientemente ha sido refinada mediante la inclusión del metamodelo de una de la metodología más avanzadas como es OMACS [299] y mejorando la metodología de desarrollo software inherente, ya que se basa en SPEM 2.0. El nuevo modelo resultante se denomina O-MASE [283].
- **INGENIAS** [312, 336], basada inicialmente en la metodología MESSAGE/UML [303]. Propone un modelo basado en ingeniería y que incluye todo el ciclo del software, así como un conjunto de herramientas que dan soporte al proceso. Esta metodología se ha ido refinando progresivamente [336-338], gracias a lo cuál, en las últimas versiones ya es posible hacer frente a la necesidad de readaptación del modelo de agentes [338].
  A partir de esta metodología ha nacido también el modelo ANEMONA [339] que basa el diseño de sistemas multiagente mediante Sistemas Holónicos de Fabricación (HMS) [340].
- **GORMAS** [49, 50] es una metodología de última generación basada en las OV para el diseño de SMA, incluyendo las fases de análisis y diseño de sistemas software, así como una fase de diseño de la dinámica que facilita la descripción de las interacciones de los agentes individuales. Estas fases permiten describir seis modelos de especificación (organización, actividad, interacción, entorno, agente y normativo) que abarcan las características del metamodelo básico de los SMA sociales.
  El modelo descrito por la guía metodológica GORMAS extiende a los modelos propuestos en la metodología ANEMONA [339], que a su vez es una extensión de INGENIAS [312]. La metodología también está alineada con SPEM.
- **e-Institutions** [341] que asocia el modelado de SMA basándose en instituciones electrónicas que a su vez se basan en instituciones reales. El modelo ha sido refinado para incluir normas [342] y su aplicación en entornos abiertos [343].

3.3.4.Análisis de modelos organizativos

Una vez se han presentado las principales metodologías para el diseño y (en algunos casos) desarrollo de SMA, a simple vista se puede observar que todas ellas tienen un conjunto de rasgos comunes, rasgos que han ido evolucionando a medida que se han ido revisando a lo largo del

tiempo. Así, las metodologías clásicas tienen un claro enfoque de diseño centrado en el propio agente (GAIA, MOISE, ALAADIN, MASE, entre otras). Mientras que revisiones de estas metodologías han ido evolucionando hasta un modelo de diseño organizativo (GAIAexOA, MOISE+, e-Institutions, AGRE, GORMAS, etc.).

Dentro de este último grupo de metodologías, antes de comenzar el diseño y posterior desarrollo de un SMA basado en OV especialmente orientado a modelar el problema propuesto en este trabajo de investigación (monitorización y control de un sistema CC) resulta necesario realizar un análisis de los conceptos, herramientas, técnicas y mecanismos que incorporan estas metodologías para el desarrollo de SMA. Este análisis, descrito a lo largo de este apartado, permite estudiar los metamodelos propios de las metodologías presentadas. Gracias a lo cuál es posible identificar rasgos comunes y opuestos, lo que por tanto, permite seleccionar la metodología, y por lo tanto la corriente científica, que se adecúa mejor al problema a resolver.

Todas las metodologías están orientadas al desarrollo de sistemas software complejos, por lo que incluyen diferentes modelos de especificación y descripción de requisitos. Sin embargo, resulta contradictorio que muchas de ellas no incluyan un proceso de ingeniería al uso (requisitos, análisis, diseño, implementación y pruebas). Así en este sentido, tres grandes grupos claramente diferenciados en tanto en cuanto a la aplicación de modelos de ingeniería iterativos:

- No incluye ningún modelo de ingeniería fundamentado en ingeniería, como TROPOS, MOISE+, AGRE, etc.
- Incluyen algunas fases de desarrollo, normalmente sólo abarcan el modelo de dominio y de la solución. Es decir, describen como debería ser el análisis y el diseño, pero no tienen en cuenta la implementación. Entre las metodologías que se encuentran en este grupo GAIAexOA o GORMAS.
- Incluyen un proceso de ingeniería completo como PASSI, ADELFE, INGENIAS, etc. Por su parte, metodologías como SODA, O-MASE, GORMAS, recientemente han evolucionado para incluir el proceso de desarrollo de ingeniería SPEM 2.0 [318].

Dentro de estos modelos organizativos, como ya se ha indicado, muchos autores hablan del meta-modelo del paradigma de agentes [299, 300, 310, 314]. El metamodelo incluye todos aquellos conceptos, artefactos y herramientas que permiten diseñar un SMA siguiendo una perspectiva organizativa. A continuación en la Figura 12 se presenta el meta-modelo de la metodología OMACS [299].

En el metamodelo de los sistemas organizativos, existen dos conceptos clave, rol y organización (o grupo), además del propio concepto de agente. Estos tres conceptos organizativos, además de definir una metodología por sí solos (AGR[271]), permiten modelar en líneas generales un SMA. Es habitual dividir el estudio de los sistemas organizativos en dos niveles de abstracción [271, 297], el estructural (o macro) que tiene en cuenta los aspectos dinámicos y de organización (roles y grupos). Y, el nivel concreto (o micro) que conforma la definición de bajo nivel de los agentes (tareas, planes, etc.).

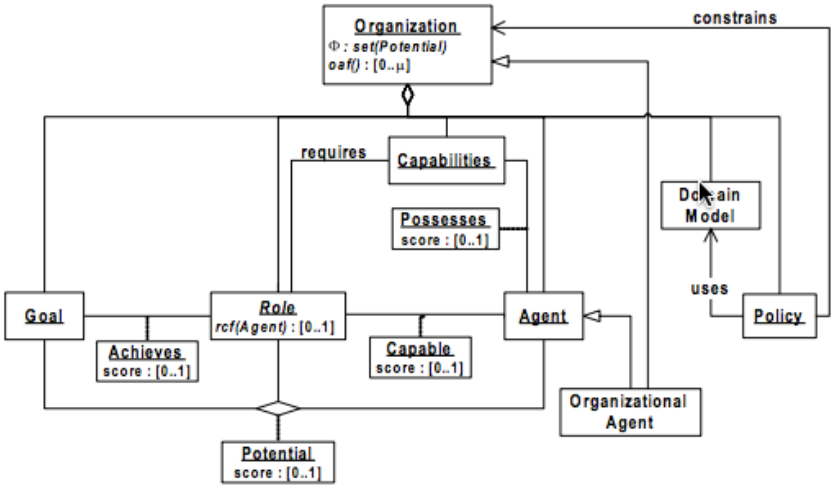

Figura 12.- Meta-modelo Sistemas Organizativos [299]

Por un lado, el concepto de rol, presente en la Teoría de Agentes desde sus inicios, define de forma abstracta el comportamiento de las entidades que forman parte del sistema (agentes). Este comportamiento se define por medio de las obligaciones/responsabilidades y requisitos que son necesarios satisfacer por el rol, pero también mediante las capacidades y derechos que se le presuponen al individuo o agente que adquiriera (*playing/enacting*) dicho rol en el sistema, de forma que estas habilidades le permitan satisfacer las obligaciones y requisitos asociados [271, 297]. Este concepto de adquisición del rol por parte de un agente individual resulta de gran importancia, ya que representa la entidad abstracta (no instanciable), mientras que el agente representa al individuo concreto y con una arquitectura determinada (reactiva, deliberativa, modelo BDI, etc.) [149]. En sentido, los agentes también pueden tener una planificación [161] definida mediante una representación actual del mundo, un objetivo y la secuencia de acciones (o misiones [204]) para conseguirlo.

Además, en el marco de los sistemas abiertos, los agentes son entidades que pueden entrar o salir de la organización, pero que para poder jugar un rol específico deben tener unas características concretas definidas por el propio rol, así como seguir un protocolo de entrada en la organización (sociedades semi-abiertas o semi-cerradas) como ya se ha presentado anteriormente [23, 295]. En este sentido, existen metodologías que hablan de roles institucionales y roles externos [44] en función del tipo de agentes que existan en cada momento de vida del sistema.

Por otro lado, la definición del concepto de organización resulta mucho más compleja, en primer lugar, la organización tiene que describir los objetivos para lo que ha sido diseñada, teniendo en cuenta que estos objetivos tienen que estar alineados con los de sus miembros (roles o suborganizaciones) y que en cierta medida aportan racionalidad al sistema en su conjunto. Este diseño, aunque con diferentes nomenclaturas según la metodología, tiende a incluir aspectos sociales, de comunicación, de interacción y normativos [44]. A continuación, se describe cada uno de ellos:

- Los **aspectos sociales** (o estructurales [271]) se refieren a la descripción del conjunto de roles, grupos (asociaciones de roles) y la relación entre ellos. En cuanto a las relaciones existentes entre roles y grupos (recursivamente) destacar que hay autores que han definido un conjunto de estructuras sociales que permiten modelar las interacciones entre los miembros [298]. Entre las principales estructuras destacan la siguientes: jerarquías, holarquías, coaliciones, equipos, congregaciones, sociedades, federaciones, mercados, matrices y organizaciones compuestas.

Sin embargo, también existen trabajos en el que simplemente se definen posibles relaciones entre miembros [44] como dependencia, jerarquía, uso, etc.

- Los **aspectos de comunicación** hacen referencia a los medios o el lenguaje que hace posible el intercambio de información. Es decir, un lenguaje de representación de conocimiento (habitualmente representado mediante una ontología) y un lenguaje de comunicación. La secuencia de comunicación entre dos agentes se denomina ilocución [341], acto de comunicación [44] o enlace [204], y puede tener diferentes fines según la filosofía del mensaje [44]: representación, prohibición, permisos, declaración, expresivos, compromisos o directivos.
- Los **aspectos de interacción** se refieren a como los roles colaboran para alcanzar objetivos comunes. Es decir, dado que pueden existir objetivos que no se pueden alcanzar individualmente, y que requieren de la combinación de varios agentes para su consecución, es necesario describir una estructura de interacción que permita articular o regular la consecución de los sub-objetivos individuales que a su vez hagan posible la consecución de objetivos de más alto nivel. Hay que destacar que estas secuencias de interacción evolucionan en la medida que avanzan las relaciones entre los roles y grupos. A este conjunto articulado y reglado de ilocuciones entre varios agentes para conseguir un objetivo común se le denomina habitualmente escena [44].
- Finalmente, en cuanto a los **aspectos normativos** cabe destacar que es uno de los grandes pilares de los SMA organizativos, e incluso existe una metodología que se basa en este concepto (HARMONIA [325]). Las normas (o patrones institucionales [25, 271]) permiten establecer una relación de confianza entre los miembros de una organización, ya que permiten acotar el libre albedrío de los agentes individuales. Las normas deben ser acatadas por cualquier agente externo que quiera formar parte de una organización. En resumen, definen formalmente las obligaciones, prohibiciones y permisos de los miembros y de las comunicaciones entre estos.

Además de los conceptos que se acaban de presentar (rol, organización, normas y estructuras sociales), los SMA organizativos incluyen de forma habitual otro concepto clave, el de entorno. La Teoría de Agentes, tradicionalmente concibe al agente como una entidad que planifica sus acciones en función de las percepciones del medio. Sin embargo, la complejidad creciente del propio medio en el marco de los sistemas abiertos (dinámicos, heterogéneos e impredecibles) no sólo puede llegar a convertir en impredecible al SMA, sino que las interacciones con él resultan complejas.

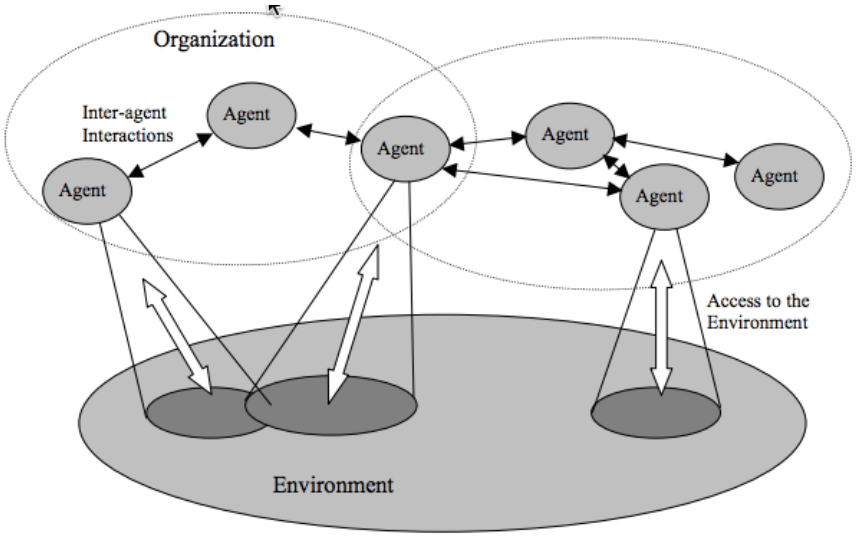

Figura 13.- Concepto de entorno en sistemas organizativos [25]

Por ello, las metodologías recientes modelan al entorno dentro del propio SMA, de forma que es posible diseñar el sistema teniendo en cuenta la complejidad de éste. En este sentido cabe destacar la aproximación de la metodología *Agents and Artifacts* (A&A) [278] que modelan las interacciones con el entorno mediante un mecanismo denominado artefactos organizativos que encapsula la observación de las propiedades y eventos del entorno y articula las acciones de control y uso de los recursos que proporciona el propio entorno.

Todo el paradigma anteriormente presentado supone una evolución de los SMA, de cara a facilitar su aplicación en entornos abiertos, así como los nuevos requisitos que requieren el desarrollo de software complejo existente hoy en día. Recientemente se observa que este proceso evolutivo tiene un nuevo hito, el de la readaptación. La readaptación puede definirse como el proceso de cambio en la propia organización para dar lugar a una nueva [344]. Normalmente, esta evolución viene motivada por un cambio en el entorno, o la introducción de nuevos requisitos en el dominio del problema. El proceso de readaptación puede incluir roles, objetivos, servicios, normas, etc. El proceso de readaptación según So *et* Durfee incluye las siguientes fases [345]:

- **Monitorización.** En esta fase se define el momento y el motivo en el que la organización necesita evolucionar. El modelo de monitorización puede ser centralizado si se realiza por una autoridad o rol específico, o puede ser distribuido si se delega entre un conjunto de agentes. La dos decisiones (cuándo y por qué) pueden ser tomadas según Dignum *et al.*,[346] siguiendo un modelo reactivo o proactivo, en función de si existe un proceso de razonamiento.
- **Diseño.** En esta fase se determina el cómo, es decir, los cambios que deben producirse durante la fase de reorganización para dar soporte a los cambios en el entorno o a los requisitos del problema. El diseño de la reorganización puede ser realizado de forma centralizada o descentralizada, en función de donde se lleve a cabo. Así mismo, en función de cómo se decida la reorganización, las decisiones pueden ser tomadas de forma autónoma o por acuerdo.
  El modelo de diseño puede incluir decisiones acerca de [344]: (i) la dinámica de los agentes y su comportamiento, (ii) modelo de interacción y comunicación, (iii) cambios en el modelo social, (iv) evolución del contexto normativo, (v) el soporte a situaciones de emergencia, y finalmente, (vi) los cambios de población en sistemas abiertos.
- **Selección.** Es la fase en el que se elige uno de los diseños especificados en la fase anterior. Esta decisión también puede ser tomada de forma centralizada o distribuida en función de si la decisión es tomada por una autoridad individual de forma autónoma, o mediante acuerdo, de forma distribuida. El proceso de selección puede tener en cuenta los siguientes criterios [344]: consecución de objetivos, grado de utilidad, coste global y coste individual.
- **Evaluación.** Finalmente, en esta última fase se evalúa el modelo de diseño que se ha seguido, y si se han alcanzado los objetivos pretendidos con la readaptación. Gracias a este modelo de evaluación es posible tomar mejores decisiones en un futuro.

Los modelos de re-adaptación han sido estudiados por diferentes autores [344, 345], en cuanto a la evolución de normas [347], estructuras [346, 348], así como la inclusión de modelos de readaptación en diferentes metodologías MOISE+ [349], e-Institutions [350], MASE/OMACS/O-MASE [351], etc. Sin embargo, hoy en día sigue siendo uno de los campos abiertos en el marco del paradigma de los SMA.

3.4.Conclusiones

Después de analizar el paradigma tecnológico CC en el Capítulo 2, este capítulo ha tenido como principal interés el estudio de la teoría de agentes y los SMA. Este modelo arquitectónico se basa en la autonomía, habilidades sociales, reactividad y pro-actividad de sus componentes individuales, denominados agentes [38]. Su marco ideal de aplicación son los sistemas complejos, abiertos o ubicuos. El entorno tecnológico del paradigma CC se asocia, precisamente, con estos entornos de aplicación debido a su heterogeneidad, dinamicidad e incertidumbre. Por ello, a lo largo del capítulo se ha analizado la relación existente entre ambos paradigmas distribuidos, obteniendo diferentes conclusiones tal y como se detallará a continuación.

Existe una relación incipiente, aunque muy prometedora debido a las sinergias encontradas entre ambos sistemas distribuidos. Se determina que la mejor forma de clasificar los trabajos existentes en el estado del arte es mediante la referencia a los roles del paradigma CC especificados en el modelo arquitectónico propuesto por el NIST [10].

A partir de esta clasificación, se observa una gran cantidad de trabajos asociados a los roles *Cloud Consumer* debido a que los sistemas CC proporcionan unas capacidades computacionales que permiten a los SMA maximizar sus cualidades intrínsecas. Así mismo, también se observa un abundante número trabajos en los que los SMA actúan como *Cloud Broker*, y en menor medida como Cloud Auditor; debido, principalmente, a su similitud con tecnologías precedentes como la búsqueda, composición y monitorización externa de servicios web. Los modelos, técnicas y herramientas propuestas en estos trabajos solucionan principalmente el problema de la interoperabilidad entre plataformas, siguiente la aproximación de InterCloud [139, 140].

Sin embargo, pese a la incipiente presencia de los SMA en estos tres roles, se observa un limitado número de trabajos en el marco del rol *Cloud Provider* del paradigma CC. Es decir, aplicaciones de los SMA en el núcleo de una plataforma CC para permitir su control y monitorización mediante IA distribuida. No cabe duda de que este limitado número de trabajos está relacionado con la dificultad para gobernar un entorno CC de última generación, así como su complejidad y heterogeneidad inherente. Por ello, más si cabe, la hipótesis de este trabajo de investigación se considera innovadora.

Para que esta aproximación de aplicación de SMA en entornos CC sea posible, es necesario utilizar las últimas innovaciones existentes en este campo, especialmente aquellos modelos de SMA que se diseñan desde una perspectiva social [23], destacando sobre todos ellos, el modelo de las OV[24].

Este modelo de diseño de SMA permite dividir las diferentes responsabilidades de un sistema CC entre grupos de agentes con objetivos específicos y modelos de razonamiento adecuados al problema a resolver. No obstante, pese a modelar el sistema como grupos de trabajo con responsabilidades y objetivos independientes, todos trabajan de forma conjunta para conseguir los objetivos globales del sistema. Así mismo, modelar un entorno CC, mediante el uso de OV de agentes inteligentes permite la autoadaptación del sistema en función de las condiciones del entorno lo que facilita la adaptación dinámica de la plataforma y no sólo la de sus individuos. Así mismo, también permite la inclusión de agentes especializados, con capacidades de razonamiento avanzadas que permiten modelar el sistema utilizando la base que proporciona la IA. Dentro de las metodologías estudiadas se ha determinado que GORMAS [49, 50] es la metodología de diseño que se considera más adecuada al problema a resolver. En este sentido, se considera que tanto las vistas como metamodelos propuestos en el marco del trabajo cubren perfectamente todas las fases de diseño y los conceptos necesarios para la descripción del sistema.

Partiendo de la hipótesis inicial y de los trabajos analizados, en el capítulo siguiente se presentará la arquitectura que se ha propuesto, incluyendo tanto el modelo de la arquitectura, como los modelos de razonamiento avanzado que hacen posible su adaptación. Posteriormente, en el Capítulo 5 se evaluará empíricamente a través de un caso de estudio tanto la arquitectura propuesta como los algoritmos de adaptación que se han diseñado en el marco de este trabajo de investigación.

Capítulo 4. Modelo de arquitectura

En este capítulo se presenta el modelo arquitectura basado en OV de agentes inteligentes, para la distribución de recursos computacionales en entornos distribuidos, concretamente en sistemas CC. Este modelo arquitectónico se considera el más apropiado para modelar un sistema CC utilizando SMA. Así, el uso de OV permite aplicar en el diseño de un sistema CC características avanzadas derivadas de la IA distribuida, no sólo en tanto en cuanto a la autonomía, pro-actividad, aprendizaje y razonamiento avanzado de sus componentes [38]; sino que también será posible la incorporación de métodos y modelos para la auto-adaptación dinámica del propio SMA, lo que facilita la adaptación del sistema en su conjunto a los cambios que se puedan producir en el entorno, principalmente, aquellos motivados por la variabilidad en la demanda de los servicios que ofertados por el entorno CC.

Como ya se han indicado en capítulos precedentes, un sistema CC se asocia a un sistema abierto siguiendo un modelo de diseño basado en SMA. Es un sistema abierto por ser dinámico y heterogéneo. En primer lugar, es un sistema dinámico ya que la población de la organización (los agentes) pueden salir o entrar en la citada organización. Además se puede dar el caso, en que agentes externos ofrezcan servicios dentro de la propia organización, jugando roles específicos dentro de la misma [295]. En segundo lugar, es heterogéneo ya que el diseño de los agentes puede llegar a ser muy diverso, teniendo que interactuar con un entorno también heterogéneo donde las diferentes tecnologías subyacentes complican en gran medida las acciones de los agentes individuales. Finalmente, como se detallará a lo largo de este capítulo, un sistema CC modelado mediante SMA organizativos, también es un sistema con cierta incertidumbre, sobre todo en la toma de decisiones, debido a que los agentes individuales no tienen un conocimiento total del sistema en su conjunto, sino que tan sólo disponen de un modelo de la realidad parcial, centrado en aquellas áreas que tienen a su cargo y, por lo tanto, tienen que recurrir a la coordinación y negociación con sus semejantes para llevar a cabo las tareas que tiene asignadas.

A lo largo de este capítulo también se detallarán los algoritmos propuestos para la autoadaptación de la organización en función de la demanda de los servicios por parte de los usuarios. Para ello, como se detallará posteriormente, se han diseñado modelos de razonamiento avanzado que se implementan por los agentes individuales que participan en la organización, según el rol que jueguen dentro de la misma. Estos agentes serán capaces de repartir dinámicamente la cantidad de recursos en función de la demanda en los servicios y los contratos de QoS acordados a través de los correspondientes SLA.

El capítulo se organiza tal y como sigue, a continuación en el apartado 4.1 se formaliza el entorno de aplicación del modelo arquitectónico propuesto. Posteriormente, el apartado 4.2 contiene una descripción detallada del SMA propuesto, el cuál se ha diseñado a partir de la metodología GORMAS [49]. A partir de esta arquitectura, el apartado 4.3 detalla el modelo de razonamiento de los agentes especializados en la distribución de recursos, detallando su modelo de representación del entorno y sus algoritmos razonamiento. Finalmente, el último apartado (4.4) del capítulo se presentan las conclusiones del mismo.

4.1.Caracterización de un entorno Cloud Computing

De forma previa a formalizar la arquitectura multiagente basada en OV que se propone en el marco de este trabajo de tesis doctoral, resulta necesario formalizar el contexto y el entorno en el que se ejecutará dicha arquitectura. Dada la complejidad asociada a un entorno CC, así como los diferentes componentes artificiales y humanos que están involucrados en este contexto, resulta necesario realizar una caracterización y formalización del entorno, que permita acotar el ámbito de aplicación del modelo arquitectónico que se ha propuesto y que se detallará posteriormente en las secciones 4.2 y 4.3. Esta caracterización que se presenta a continuación, puede abstraer a la mayoría de entornos CC que existen en la actualidad en el ámbito científico y en el mercado.

4.1.1.Modelo de comercialización de servicios

En líneas generales, tal y como se ha descrito en los capítulos anteriores, un sistema CC ofrece a los usuarios un conjunto de servicios computacionales de cualquier tipo, es decir, *software*, plataforma e infraestructura. En cierta manera, este modelo no varía en gran medida con respecto a las tecnologías anteriores; sin embargo, la característica clave de este tipo de sistemas es que los servicios se ofertan siguiendo un modelo de pago por uso. Para ello, tomando las bases de la tecnología *Utility Computing [8]*, el usuario contrata un conjunto de servicios computacionales de forma similar a como se contratarían otros servicios de utilidad en la vida real como el gas, el agua, la electricidad, etc. Cada contrato de prestación de servicios que se realiza con los usuarios deberá incluir información acerca de los recursos contratados, el grado de calidad que se asocia a los mismos, el tiempo de duración del contrato y el coste, el cuál vendrá determinado por la cantidad y calidad de los recursos utilizados.

Una vez que se ha establecido el acuerdo, el usuario es libre de utilizar, o no, los servicios contratados. A diferencia de otros servicios de utilidad tradicionales, el tiempo de vida de los acuerdos SLA establecidos suele ser más corto [180], lo que determina que el coste pueda variar a lo largo del tiempo, en función de la carga y la demanda de usuarios existente en cada momento [9].

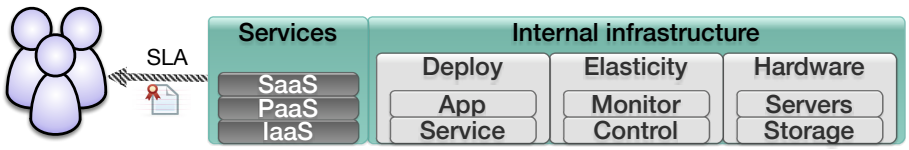

Figura 14.- Modelo de comercialización de servicios en entorno Cloud Computing

Así, de forma general, se ha formalizado el acuerdo de servicio para un usuario cualquiera ($ServA^j$) como la unión del conjunto de acuerdos de uso establecidos sobre cada uno de los servicios de forma individual (SLA_i^j) [352, 353]. El acuerdo de uso por servicio concreto ($ID_{i,}$) se determina en función del tiempo de duración del acuerdo (t), las partes implicadas en el mismo (u), el coste asociado (c) y los niveles de calidad acordados (QM_i).

$$ServA^j = \bigcup_{i=0}^{i=m} SLA_i^j \; donde \; SLA_i^j = \{ID_{i,}, u, t, c, QM_i\}$$

A partir de esta sencilla expresión, el reto para conseguir un modelado adecuado del contexto, consiste en determinar la medida de calidad de los servicios que se ofertan. Este proceso de medición es complejo, más si cabe cuándo se ofertan servicios computacionales a diferentes niveles: *software*, plataforma e infraestructura. En el estado del arte es posible encontrar varios trabajos relacionados [9, 354, 355]. Sin embargo, es necesario obtener una medida adecuada al problema a

resolver dentro de este trabajo de tesis doctoral, que no es más que la distribución dinámica de recursos computacionales en sistemas CC. Por ello, se ha estudiado el conjunto de métricas existentes, descartando directamente aquellas que no resultaban de interés en el marco de este trabajo de investigación (seguridad, mantenimiento, políticas de internacionalización, gestión del ciclo *software*, etc.) por no estar directamente relacionadas con el problema a resolver.

A continuación, se detallarán las métricas relevantes, para ello en primer lugar, se enuncian los productos que se ofertan en un entorno CC, las cuáles se pueden agrupar en tres tipos de productos: (i) **productos hardware**, (ii) **productos software** que siguen la estructura típica cliente-servidor, y finalmente, (iii) **productos combinados** donde se accede a recursos *hardware* mediante el uso de recursos *software*, como por ejemplo, bases de datos o sistemas de almacenamiento con interfaz web.

En primer lugar, se presentan las métricas que permitan medir la calidad de los productos de tipo *hardware*. Principalmente, en este grupo los servicios a tener en cuenta son de almacenamiento, computación y servidores dedicados. En este tipo de servicios es necesario garantizar la disponibilidad de los mismos, de acuerdo con las características que se hayan acordado a través de un SLA específico:

- Los servicios de almacenamiento según el modelo CC se miden en términos del espacio de almacenamiento (s), número de operaciones (n_r) del tipo añadir, eliminar, recuperar, etc. y el ancho de banda de conexión (BW):

$$SLA_{storage} = \{s, n_r, BW\}$$

- Los servicios de computación y servidores dedicados, en el marco de los sistemas CC, se tienden a enmarcar en la misma categoría, ya que la computación de servicios *grid* se realiza en servidores dedicados en el marco de la tecnología CC [356, 357]. Indicar, que en este trabajo de investigación no se contempla la computación de alto rendimiento siguiendo un esquema de paralelización tipo *MapReduce* [358].
  En definitiva, en este tipo de servicios lo que se determina como métricas son las características del servidor dedicado que es contratado por parte del usuario y del cuál, el proveedor debe garantizar su disponibilidad. Así, las principales características son el número de núcleos de computación (n_{cpu}), la cantidad de memoria asignada (s_{ram}), la capacidad del disco duro (s_{hdd}), el ancho de banda (BW) y el sistema operativo ($ssoo$).

$$SLA_{virtual-server} = \{n_{cpu}, s_{ram}, s_{hdd}, BW, ssoo\}$$

Una vez que se han definido las métricas que se utilizan en este trabajo para los servicios *hardware*, a continuación se procede a definir las métricas para los productos de tipo *software*. Las métricas que permiten medir estos servicios resultan más complejas de definir, ya que no sólo hay que asegurar la disponibilidad de los servicios, sino también es necesario evaluar parámetros, que en la mayoría de los casos son dependientes del dominio de uso. Por ejemplo, en productos *software* es necesario evaluar la disponibilidad o no de una funcionalidad específica. Por tanto, una vez más, resulta necesario determinar qué métricas son de adecuadas para los productos *software* teniendo en cuenta el ámbito de esta investigación, la cuál se centra en proporcionar mecanismos que aseguren la disponibilidad y la calidad de los servicios que un usuario haya contratado previamente. Por ello, es necesario establecer un acuerdo de calidad del servicio que atienda a parámetros de calidad directamente relacionados con la forma en que se distribuyen los recursos computacionales subyacentes.

Así pues, en general, las métricas para servicios software están basadas en parámetros como el rendimiento, la confiabilidad, integridad, accesibilidad, robustez, interoperabilidad, seguridad y gestión de errores del servicio [359]. En la práctica [354], la calidad de los servicios, y por lo tanto los valores de estas métricas están asociados a aspectos relacionados con la ingeniería del *software*, la algoritmia y la infraestructura tecnológica donde se despliegan. Por ejemplo, una correcta arquitectura *software* mejora métricas como la interoperabilidad, la accesibilidad, etc. pero en ningún caso garantiza la eficiencia del *software*. Del mismo modo, unos algoritmos eficientes en la mayoría de los casos permiten incrementar el rendimiento y la robustez del producto *software*, pero no afectan a otras características como la gestión de errores. Por su parte, cuanta más potencia disponga la infraestructura subyacente donde se despliegue la oferta de servicios, mejores valores se obtendrán en cuanto al rendimiento y disponibilidad del *software* o *hardware* ofertado como servicio. Otro problema que existe a la hora de medir los servicios *software* es que, hoy en día, la aplicaciones web modernas y los servicios web son diseñadas siguiendo una arquitectura orientada a servicios [360]. Este modelo es muy flexible y permite que una aplicación pueda dividirse en varios niveles, de forma que cada uno intercambie información con el resto, siguiendo el principio de alta cohesión y bajo acoplamiento [361]. En la Figura 15 se observa un esquema de este tipo de aplicaciones. En este modelo, orientado a la reutilización e integración de componentes puede darse el caso de que haya componentes cuyo rendimiento sea muy alto, pero al mismo tiempo, haya otros componentes cuyo rendimiento no sea tal, lo que degrada el tiempo de respuesta del servicio en su conjunto, desde el punto de vista del usuario final. En conclusión, en este tipo de aplicaciones es necesario que la evaluación de la calidad de los servicios se realice sobre cada elemento atómico, cuya dependencia con terceros no influya en el rendimiento final del usuario.

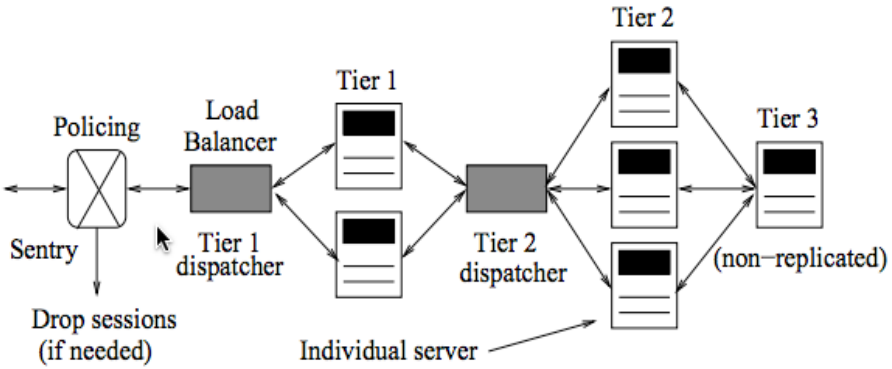

Figura 15.- Varios niveles en una aplicación o servicio web [361]

Así, dentro de este subconjunto de métricas que determinan la calidad de los productos *software* y que, a la vez, están directamente relacionadas con los recursos computacionales subyacentes, también existen diferentes parámetros a valorar como puede ser el tiempo de respuesta, el número de respuestas en un tiempo determinado, la latencia, etc. De este conjunto de métricas, en el marco de este trabajo de tesis doctoral interesan aquellas que permiten evaluar la calidad del servicio y en este sentido se selecciona el tiempo de respuesta medio de las peticiones. El tiempo de respuesta en las peticiones, viene determinado por el tiempo que tarda una petición en ser atendida. Así, dado un servicio k con un conjunto de métodos (r_i^k) que forman el API del servicio. La calidad de cada una de las peticiones que forman el API del servicio viene determinada por la siguiente expresión de tamaño de la respuesta (s_i^k) y tiempo de la transmisión (t_i^k) de la misma:

$$QoS^k = \left\{\overline{r_1^k} \cdots \overline{r_i^k} \cdots \overline{r_n^k}\right\} donde\ \overline{r_i^k} = \frac{1}{m}\sum_{i=1}^{m} s_i^k \Big/ t_i^k$$

Junto a la calidad del servicio, también es necesario denotar la disponibilidad (A) del mismo, que en los sistemas computacionales de alta disponibilidad suele ser mayor al 99.5% [362]. Este nivel de calidad suele venir determinado por el nivel de redundancia de los componentes del entorno *hardware* y de los sistemas de alimentación ininterrumpida. En este sentido, los niveles TIER[10] determinan el grado de redundancia de cada uno de estos componentes. En definitiva, el nivel de calidad de un servicio *software* (SLA_{sw}), se determina en función de su calidad según la métrica que se acaba de definir, el número de peticiones que se pueden realizar (n_r^k), el ancho de banda (BW) y la disponibilidad del mismo (A).

$$SLA_{sw} = \{QoS^k, n_r^k, BW, A\}$$

Finalmente, cuando existen productos computacionales combinados, su caracterización viene determinada por una combinación de las métricas que se acaban de presentar, tal y como se muestra en la siguiente expresión

$$SLA_{sw-hw} = \{SLA_{sw}, SLA_{storage}, SLA_{virtual-server}\}$$

En definitiva, la calidad de los servicios en el marco de este trabajo de investigación viene determinado por la cantidad de recursos computacionales asignados a cada servicio particular. Dada la importancia de la infraestructura en este trabajo, en el siguiente apartado se modelará y caracterizará también la infraestructura subyacente de un entorno CC, que como ya se ha avanzado está compuesta de *hardware* real y *hardware* virtual o abstracto.

4.1.2.Caracterización del entorno computacional

La infraestructura *hardware* de un entorno CC es uno de los sistemas tecnológicos más complejos e impredecibles que existen. En él conviven equipamiento variado como servidores, módulos de red, unidades de alimentación ininterrumpida y un largo etcétera de infraestructura informática de alto rendimiento. El *hardware* existente en estos centros de computación suele ser altamente heterogéneo debido al incesante avance de la electrónica, según dicta la *Ley de Moore* [363] en el que existe un proceso constante de miniaturización en los componentes, dando lugar a una nueva generación cada 12 ó 18 meses aproximadamente [364].

Los recursos computacionales a considerar en un entorno CC son principalmente de ejecución y almacenamiento, aunque también pueden existir recursos combinados:

- Los **recursos de ejecución** son aquellos que tienen una alta capacidad de procesamiento y memoria para la ejecución de *software*. Aunque también disponen de almacenamiento persistente, éste no es relevante dentro del entorno CC, ya que simplemente es utilizado por el sistema operativo nativo del recurso *hardware*, o el de las máquinas virtuales que puedan contener para la computación de los servicios a los que estén asociados.
  Hoy en día, la administración, monitorización y control de estos recursos puede realizarse de forma dinámica gracias a la tecnología de virtualización [71].
- Los **recursos de persistencia** son aquellos que proporcionan al entorno CC capacidad para el almacenamiento persistente y, aunque, también disponen de recursos de procesamiento, éstos no son utilizados como recursos activos en el entorno CC. Su función es la de almacenar de forma duradera y segura la información de los usuarios del entorno CC.

[10]ANSI/TIA-942 Telecommunications Infrastructure Standard for Data Centers
http://www.tiaonline.org/news-media/press-releases/tia-issues-new-telecommunications-infrastructure-standard-data-center

Los recursos de persistencia suelen agruparse para formar un entorno SAN (*Storage Area Network*) que proporciona un acceso homogéneo a los diferentes componentes de la capa de persistencia distribuidos en la mayoría de los casos en los diferentes nodos *hardware* (reales o virtuales) del entorno CC.

- Finalmente, también pueden existir **recursos híbridos** que aporten a un entorno CC recursos computacionales y de persistencia. En la práctica este tipo de recursos no suele ser tan habitual.

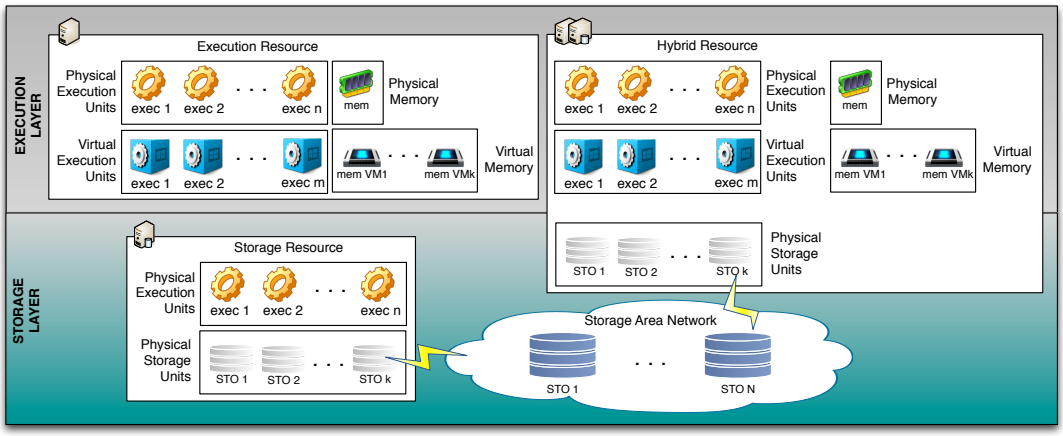

Figura 16.- Recursos computacionales a nivel físico en un entorno Cloud Computing

Dentro de los recursos de un entorno CC también es necesario considerar el equipamiento de red, que hace posible las comunicaciones entre los diferentes elementos del sistema, así como, entre éstos con el exterior. Hoy en día, gracias a los sistemas de interconexión cableados basados en fibra óptica, este medio de comunicación no constituye un problema a la hora de gestionar la infraestructura. Así mismo, en la actualidad, también es posible virtualizar elementos de red (enrutadores, subredes, balanceadores, etc.) [365]. Sin embargo, dentro del alcance de este trabajo de investigación no se ha considerado esta capacidad debido a la falta de madurez de la tecnología de virtualización subyacente para el control dinámico de este tipo de recursos computacionales [366] ya que, en la práctica sigue necesitando la presencia de *hardware* real para que no se limite el rendimiento.

4.1.2.1.El sistema de virtualización

La heterogeneidad y complejidad intrínseca a este tipo de entornos computacionales hacen que sean complicados de gestionar y mantener a lo largo del tiempo [367]. Por lo que, para ello, se requiere personal altamente especializado. Sin embargo, hoy en día la tecnología de virtualización [72] facilita en gran medida esta tarea dentro de este tipo de entornos computacionales. La virtualización permite abstraer la complejidad del *hardware* subyacente en forma de máquinas virtuales. Una máquina virtual es un recurso computacional abstracto que emula una máquina física con unas características *software* y *hardware* concretas.

La gran ventaja de esta tecnología es que la asignación de recursos a cada instancia de ejecución puede reconfigurarse de forma dinámica. Así, es posible, mostrar una vista abstracta o virtual, pero homogénea y controlable del entorno *hardware* real [71]. Es decir, a partir del *hardware* existente en un determinado equipo físico (servidor), la virtualización permite repartir y encapsular los recursos físicos entre un conjunto de máquinas virtuales, siguiendo un esquema similar al modelo *master-slave* [368], en el que un servidor aloja diferentes máquinas virtuales. Así, cada servidor virtualizado, o simplemente máquina virtual, tendrá un conjunto de recursos *hardware* dedicados y será totalmente independiente, siendo posible incluso la ejecución de sistemas operativos diferentes y, por supuesto,

software totalmente independiente. Esta capacidad que ofrece la tecnología de virtualización permite que los recursos computacionales en la capa de virtualización sean considerados como ilimitados, mientras que en contraposición los recursos físicos o infraestructura *hardware* estén limitados a la infraestructura real existente (número y velocidad de los procesadores, cantidad de memoria RAM, espacio de almacenamiento en disco, elementos de red, etc.).

Los sistemas de virtualización modernos se agrupan en dos categorías, tal y como sigue [369]:

- Los **sistemas de virtualización completa** que emulan un entorno *hardware* entero, de manera que el sistema operativo huésped no tiene conocimiento de estar ejecutándose en una plataforma virtual. Como desventaja, la total encapsulación de las máquinas implica la emulación de una gran cantidad de *hardware* y componentes, lo que reduce en gran medida el rendimiento.
- Los **sistemas de para-virtualización** requieren que el sistema operativo anfitrión y los sistemas huésped sean el mismo o, al menos, compartan ciertas semejanzas. Los sistemas huéspedes deben ser modificados y, habitualmente, ejecutan un *kernel* o núcleo del sistema operativo especial. La pérdida de rendimiento que provoca la necesidad de emulación del *hardware* se reduce en gran medida [370, 371].

Como única desventaja de esta tecnología, se observa que la gestión del entorno virtualizado requiere de un *software* altamente especializado, conocido como *Hypervisor* [372]. Este módulo es el encargado de gestionar las abstracciones *hardware* en cada nodo físico y para ello consume también recursos [73]. No obstante, este consumo computacional depende del modelo de virtualización elegido. Actualmente, este sobrecoste no constituye un gran problema ya que no supera el 2% de la potencia computacional del entorno físico [74, 75], aunque para ello requieren *hardware* dedicado, algo que también está ampliamente implantado gracias a las tecnologías como INTEL-VT y AMD-V [76].

La virtualización no sólo permite crear esta abstracción en cuanto a la gestión de recursos, sino que hace posible una gestión dinámica de los recursos con las siguientes características:

- **Despliegue basado en plantillas.** Las máquinas virtuales pueden ser instanciadas a partir de una serie de plantillas pre-configuradas con los servicios que se desean incluir.
- **Facilidad para la escalabilidad del *hardware*.** El nuevo equipamiento, con independencia de sus características, sólo debe ser configurado inicialmente con el *software* de virtualización asociado. Los servicios a desplegar se añaden a partir de plantillas de máquina virtual.
- **Reasignación de recursos entre máquinas virtuales.** Dentro de un único servidor físico es posible configurar de forma dinámica los recursos asignados a cada máquina virtual, lo que permite realizar un ajuste detallado en cuanto a los recursos computacionales necesarios para la ejecución de un determinado servicio.
- **Migración dinámica de las máquinas virtuales entre servidores.** Además de reubicar los recursos dentro de un único servidor físico, también es posible migrar máquinas virtuales de un servidor a otro, lo que hace posible realizar la distribución de recursos no sólo a nivel micro en cada servidor físico, sino también a nivel macro en todo el entorno CC.
- **Aislamiento entre servicios.** Con frecuencia se busca desplegar cada servicio en máquinas separadas, ya que de este modo se facilita la gestión de la seguridad y se dificulta las interferencias con otros servicios.

Para formalizar un entorno de ejecución basado en virtualización, en primer lugar, es necesario caracterizar cada uno de los servidores físicos, es decir, el *hardware* real que alojan las máquinas

físicas. Así, cada servidor físico (PR) estará caracterizado por un conjunto de parámetros tal y como sigue:

$$PR = \{hostname, IP, mac, state, M_{max}, vcpu_{max}, M_{min}, vcpu_{min}, benchmark\}$$

Donde existe un conjunto de parámetros, la mayoría de ellos estáticos:

- $hostname$, se corresponde con el nombre del equipo.
- IP, dirección de internet del equipo, habitualmente este tipo de dirección también suele ser estática dentro de la red del sistema CC.
- mac, se refiere a la dirección MAC (*Media Access Control*) del dispositivo de comunicación del servidor, es estática y se necesita para hacer operaciones de arranque del equipo a través del protocolo *Wake on Lan* (WoL) [373].
- $state$, que determina el estado del servidor ("en ejecución", "reiniciándose", "apagado" y "error").
- M_{max}, se refiere a la memoria física del servidor, en *megabytes* o *kilobytes*.
- $vcpu_{max}$, número máximo de CPUs virtuales del servidor. Normalmente este parámetro no está directamente relacionado con el número de procesadores, o el número de núcleos de cada procesador, ya que suelen intervenir también otros parámetros relacionados con la tecnología de virtualización que determinan el número máximo de procesadores virtuales que se le asignarán a la máquina virtual.
- M_{min}y $vcpu_{min}$, se refiere a la asignación mínima de memoria y CPU, respectivamente, que debe estar disponible para el servidor físico (anfitrión) y que en ningún caso podrá ser asignado a las instancias virtuales en ejecución. Estos recursos se utilizan para el correcto funcionamiento del software de control del equipo físico y el *software* de gestión de la capa de virtualización.
- $benchmark$, que es una medida del rendimiento de la máquina física, con respecto a otros equipos que formen parte del sistema CC.

Por su parte, cada instancia virtual (VM) en ejecución vendrá determinada por un segundo vector con los recursos virtuales que tiene asignados en cada momento.

$$VM = \{IP, state, M, vcpu, M_i, \overline{p_{cpu}}\}$$

Donde existe un conjunto de parámetros, la mayoría de ellos dinámicos:

- IP, será la dirección IP (*Internet Protocol*) asignada al servidor virtualizado.
- $state$, que determina el estado del servidor ("sin estado", "ejecutándose", "reiniciándose", "apagado" y "error").
- M y $vcpu$, se refiere a la memoria y número de procesadores virtuales que tiene asignados el servicio en cada momento.
- M_i, se refiere a la media de memoria utilizada por la instancia virtual en un periodo de tiempo.
- $\overline{p_{cpu}}$, que determina si el uso de la CPU en un periodo de tiempo determinado ha sido ascendente o descendente, se determina a través del análisis estadístico de la regresión lineal de los valores tomados a lo largo del periodo de tiempo.

Cada servicio se construye a partir de una o varias máquinas virtuales, por lo que las características intrínsecas del servicio vienen determinadas por la plantilla de la máquina virtual que ofrece el servicio. En tiempo de ejecución, a partir de esta plantilla será posible instanciar y modificar los

recursos asociados a un determinado servicio. Así cada plantilla de nodo virtualizado asociado a un servicio cualquier *k* (VM_t^k), se describirá a través de un conjunto de propiedades tal y como sigue:

$$VM_t^k = \{ID^k, M_{min}, vcpu_{min}, type, state\}$$

Donde:

- ID^k, se refiere al identificador del servicio correspondiente.
- M_{min}, se refiere a la memoria mínima que tiene que estar asignada a una instancia en ejecución de la máquina virtual.
- $vcpu_{min}$, se refiere al número de procesadores virtuales que tiene que tener asignada una instancia en ejecución de la máquina virtual, como mínimo para que el servicio funcione correctamente.
- $type$, que determina el tipo de servicio al que está asociado, es decir, si es un producto *hardware* o *software*. Este parámetro lo que determina, por ende, es si se pueden modificar los recursos asignados a cada instancia virtual en tiempo de ejecución.
- $state$, que determina si el servicio es balanceable, es decir, si aloja una aplicación con o sin estado. Más tarde en el apartado 4.1.3 se explicará la diferencia.

Cada servidor físico será el anfitrión de un conjunto de máquinas virtuales. Por lo tanto, en cualquier momento, cada máquina física del sistema tendrá asociada una matriz con información relativa a la máquina física, así como a las diferentes máquinas virtuales que aloja en un determinado momento. Esta matriz es una instantánea de cada servidor físico en ejecución, a través del cuál se puede determinar los servicios a los que presta recursos y la cantidad de recursos que presenta a cada uno.

$$exec^e = \begin{Bmatrix} & PR^e & \\ VM_1 & \dots & VR_m \end{Bmatrix}$$

4.1.2.2.El sistema de persistencia

Los sistemas de ficheros tradicionales no pueden cumplir con los requerimientos de los nuevos y grandes sistemas de almacenamiento, que incorporan grandes cantidades de datos. Estos nuevos tipos de sistemas precisan de nuevos mecanismos de almacenamiento que permitan características como la tolerancia a fallos, la alta disponibilidad, la escalabilidad del espacio de almacenamiento y la velocidad de acceso [374]. Para conseguir estas capacidades es necesario diseñar un sistema de ficheros distribuido que permita la replicación y la partición de datos. La replicación de datos, no sólo en diferentes discos duros, sino también en diferentes nodos *hardware*. Gracias a lo cuál es posible proporcionar un acceso a la información más rápido, que el servicio no se interrumpa por fallos en un nodo y, finalmente, que no se pierda información en el caso de que exista algún fallo en alguno de los componentes. La partición de datos aumenta la velocidad de lectura y escritura de información, y permite aumentar la capacidad del sistema mediante la adicción de nuevos nodos *hardware*.

El espacio de almacenamiento del sistema de ficheros distribuido, tal como se aprecia en la Figura 17, se utiliza habitualmente para tres grandes necesidades de almacenamiento en los sistemas CC [375]: (i) la persistencia de los datos de usuarios y aplicaciones, (ii) los discos duros virtuales de la máquinas virtuales en ejecución y, finalmente, (iii) para el almacenamiento de las plantillas de los servicios.

Hoy en día, los mecanismos de gestión de sistemas SAN permiten una gran cantidad de acciones sobre el entorno de persistencia, que van desde la modificación dinámica de los diferentes espacios

de almacenamiento y la asignación de cuotas, hasta la configuración del número de volúmenes distribuidos existente, pasando por la configuración de algoritmos de sincronización e intercambio de información [376].

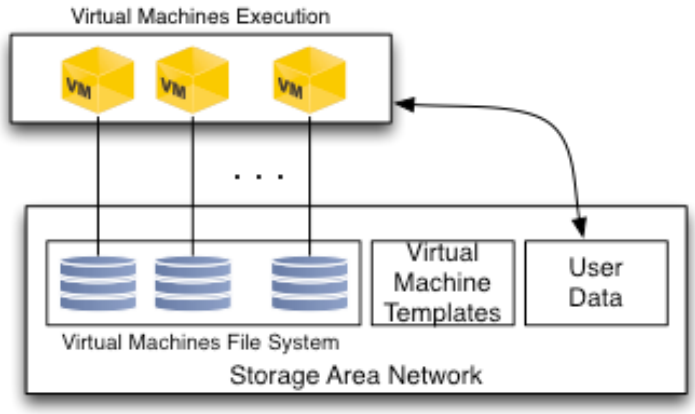

Figura 17.- Despliegue de las máquinas virtuales en el entorno *hardware*

A diferencia de lo que sucedía con los entornos de virtualización, la readaptación de este tipo de sistemas no puede ser tan dinámica debido a las restricciones tecnológicas existentes. Por ejemplo, la modificación de las diferentes réplicas implica la sincronización de datos cada vez que se crease una nueva réplica, lo que redundaría en una pérdida de rendimiento y un incremento de la actividad de red. Así pues, en el marco de este trabajo se limita su aplicación al establecimiento de cuotas de uso por usuario.

4.1.3.Modelo de oferta de servicios

La combinación de las capas que se acaban de presentar (ejecución y persistencia) permite abstraer los recursos haciendo posible una gestión dinámica de los mismos lo que, por ende, hace posible ofrecer un salto cualitativo y cuantitativo a la hora de ofertar servicios computacionales. En este sentido, siguiendo un modelo CC como se presenta en la Figura 18, cada servicio *software* de la plataforma, a nivel PaaS o SaaS, puede estar desplegado de forma simultánea en varias máquinas virtuales (nodos, o *workers*). Gracias a esta capacidad es posible configurar de forma elástica los recursos asignados a cada servicio. La demanda en términos de peticiones de un determinado servicio es balanceada entre las diferentes máquinas virtuales que están asociadas al servicio. Además, de forma dinámica se puede variar en tiempo de ejecución el peso que cada nodo virtual tiene en el balanceo. Así, la elasticidad se basa en modificar los recursos (virtuales) asignados a cada servicio de forma dinámica en función de la demanda, siguiendo las siguientes estrategias:

- Variar los recursos computacionales asociados a una determinada máquina virtual (procesamiento y memoria).
- Modificar el número de máquinas virtuales asociadas a un determinado servicio, incrementando o decrementando por consiguiente la cantidad de recursos que se asignan a un determinado servicio.
- Incluso es posible variar el servidor físico en el que se hospeda una determinada máquina virtual. Este proceso se conoce como migración y se puede realizar sin la necesidad de detener la ejecución de la máquina virtual que ofrece el servicio, pero para ello tiene el inconveniente de sobrecargar la red durante el proceso de migración.

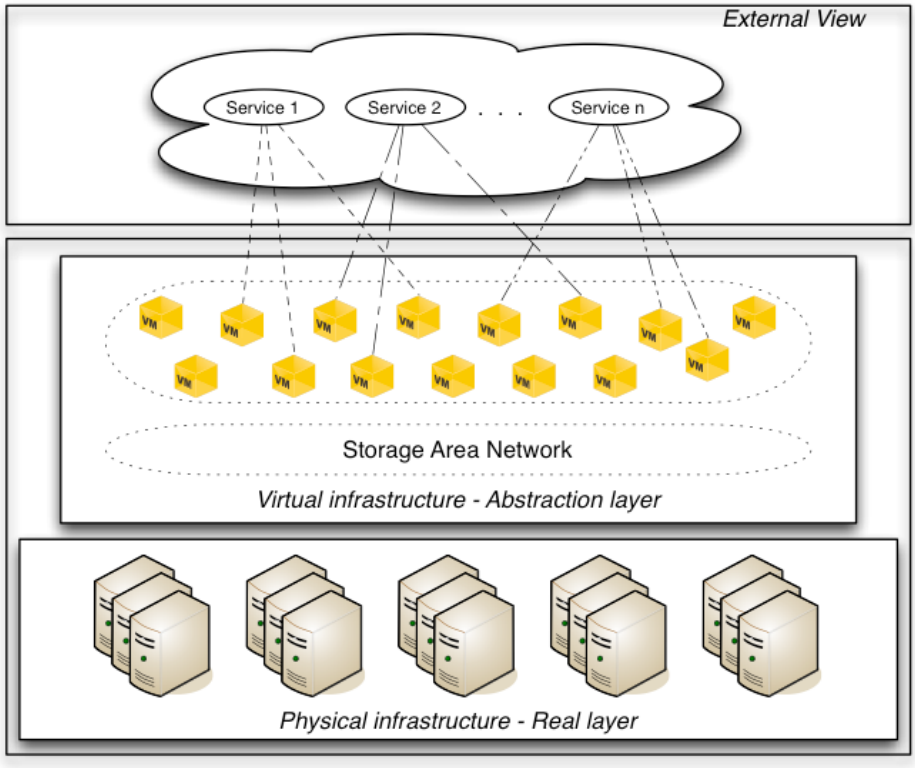

Figura 18.- Modelo de despliegue en CC

Cada nodo trabajador asociado a un servicio concreto tiene unas características particulares que dependen del tipo de servicio y que se almacenan en la plantilla de la máquina virtual que proporciona el servicio, como ya se detalló en el apartado 4.1.2.1. En este sentido, estos nodos pueden tener estado o no, tal y como sigue:

- **Stateless**. Son aquellos servicios que no tienen estado y que por lo tanto pueden existir diferentes nodos atendiendo peticiones de forma simultánea, lo que hace posible su elasticidad. El estado de estas máquinas reside en la capa de persistencia del entorno CC.
  Los nodos de tipo *stateless* son aquellos que de forma conjunta ofertan un servicio concreto de la plataforma. No sólo se tiene que garantizar la disponibilidad del servicio, sino también la calidad del mismo, para lo que es necesario distribuir sus recursos de forma dinámica en función de la demanda haciendo uso de todas las posibilidades que proporciona la tecnología de virtualización.
- **Statefull**. Son aquellos servicios que tienen estado y, por lo tanto, no pueden ser replicados de la misma forma que si no tuvieran estado, ya que obligatoriamente también se tendría que replicar el estado.
  Los servicios *statefull*, suelen asociarse a productos de infraestructura *hardware* donde los usuarios contratan servicios concretos de infraestructura con unas características *hardware* (virtual) determinadas. En este sentido, sólo es necesario garantizar su disponibilidad. También pueden asociarse a productos *software* en los que la persistencia de información no es realizada en una base de datos de tipo CC y, por lo tanto, la información reside en cada nodo individual lo que impide su replicación debido a que no es posible la sincronización efectiva de los discos duros virtuales asociados las copias redundantes. En este caso, la cantidad de recursos asociados al servicio sólo puede alterarse mediante la modificación de los recursos del único nodo en ejecución.

4.2.Formalización de la arquitectura multiagente

El diseño de un sistema de monitorización y control de un entorno tecnológico como el que se acaba de presentar en la sección 4.1 requiere del uso de técnicas de IA para poder incorporar tareas que hagan posible la adaptación dinámica a los cambios y alteraciones en la demanda de los servicios que se oferta. La adaptación dinámica a cambios que ocurren en el entorno requiere de capacidades de aprendizaje, representación distribuida de conocimiento y razonamiento avanzado. En este sentido, se seleccionan los SMA y especialmente aquellos basados en OV para dar soporte al diseño de un *software* de control que sea robusto y fiable. Dichos SMA basados en organizaciones disponen de teorías, modelos, mecanismos, métodos y herramientas que permiten desarrollar sistemas con capacidad de reorganización y que pueden adaptarse de esta forma a futuros cambios en su entorno [24]. Un diseño basado en OV permite el diseño mediante el uso de aspectos organizativos y sociales, lo que supone una representación más cercana al comportamiento humano. Además, gracias a este modelo basado en este tipo de arquitecturas distribuidas es posible no sólo repartir las tareas de control y monitorización a lo largo de todo el entorno, sino que también es posible el diseño, y posterior implementación, de agentes especializados que tomen decisiones autónomas a partir de un modelo de representación del entorno y un sistema de razonamiento que puede ser más o menos complejo.

Este modelo de SMA basado en OV también permite que agentes externos realicen servicios dentro de la organización lo que facilita la incorporación de nuevas funcionalidades, que no están directamente desarrolladas por el sistema. En este sentido, en los últimos años, los investigadores han llevado a cabo diversos estudios para ofrecer procedimientos y metodologías que permitan diseñar SMA abiertos que admitan la entrada y salida dinámica de sus entidades. Dentro de estos nuevos modelos tienen cabida agentes heterogéneos, con arquitecturas e incluso lenguajes distintos [25]. Este tipo de modelos de diseño detallan la estructura del SMA en términos de roles y la participación de los agentes en los roles del sistema, junto con normas, protocolos de interacción y objetivos asociados a las tareas. La utilización de este tipo de modelos en entornos CC supone un valor añadido ya que proporciona grandes capacidades de aprendizaje y adaptación. Por otro lado, incorporan aspectos organizativos en la gestión de entornos CC, permite diseñar sistemas abiertos, basados en aspectos sociales, con un comportamiento más cercano a las organizaciones humanas, y que por lo tanto, facilitan la interacción con los humanos.

La arquitectura propuesta en el marco de este trabajo de tesis doctoral se denominada **+Cloud** (*Multiagent System Cloud*) y está basada en OV de agentes inteligentes, gracias a lo cuál es posible hacer frente a las necesidades de adaptación, cambio, entrada y salida de componentes que requieren las plataformas de tipo CC. El principal objetivo de +Cloud consiste en la monitorización de un entorno CC y el control del mismo para que pueda adaptarse de forma automática y dinámica a las necesidades existentes en cada momento. +Cloud tomará datos a lo largo del entorno CC en su conjunto, incluyendo tanto la infraestructura subyacente, como la demanda de los servicios que se proporcionan. Así, gracias a este modelo de monitorización distribuida es posible adaptar los recursos existentes en el entorno CC en función de la demanda de cada servicio en cada momento, haciendo posible satisfacer el doble objetivo de cumplir con los acuerdos SLA establecidos y la reducción del consumo energético. +Cloud propone una arquitectura innovadora para la gestión de sistemas CC, proporcionando una nueva visión basada en aspectos organizativos. Entre los aspectos innovadores de +Cloud cabe destacar el diseño de agentes con capacidades avanzadas de razonamiento, las cuáles se detallarán en el apartado 4.3, para la gestión elástica de entornos CC.

A lo largo de este apartado, se presenta la arquitectura +Cloud en su conjunto. Para modelar el innovador SMA que se propone se ha utilizado la metodología de diseño GORMAS (*Guidelines for Organization-based MultiAgent Systems*) [49] que extiende los modelos propuestos en la metodología ANEMONA [339], que a su vez son una extensión de INGENIAS [312]. Esta metodología se basa en seis metamodelos (agente, actividad, interacción, entorno, organización y normativo) gracias a los cuáles es posible describir cualquier SMA organizativo a través de cuatro vistas (estructural, funcional, social y dinámica) que permiten especificar de forma detallada cualquier sistema *software* organizativo. Para facilitar esta tarea, la propia metodología propone una secuencia-guía para definir procedimentalmente el análisis y diseño del SMA.

Esta guía metodológica parte de las guías de diseño de organizaciones humanas [285, 377] y se divide en tres etapas principales: el análisis, diseño del sistema y diseño de la dinámica. Etapas que a su vez se llevan a cabo en las siguientes ocho fases:

- **Fase A. Misión.** Se realiza un análisis de la motivación que se persigue al definir la organización o sistema, es decir, el porqué de la existencia de dicha organización o para qué se crea. Así como los resultados que, en conjunto, se esperan conseguir. También se determina el entorno en el SMA existe, detallando los productos y/o servicios a ofrecer, cuáles son los grupos de interés y su localización.
- **Fase B. Tareas y procesos.** Se analizan con mayor detalle los servicios a ofrecer en el sistema, sus requisitos y los procesos que conllevan. Se detallan también las tareas y objetivos asociados a dichos servicios.
- **Fase C. Dimensiones organizativas.** Donde se analizan las dimensiones de la organización (departamentalización, especialización, sistema decisor, formalización, coordinación), que imponen ciertos requisitos sobre los tipos de trabajo, así como la diversidad e interdependencia de las tareas a realizar.
- **Fase D. Estructura organizativa.** Se determina y selecciona la estructura organizativa más adecuada para la organización, en función de sus dimensiones. Se hace uso de modelos organizativos para especificar los roles, interacciones y normas relacionados con la propia estructura.
- **Fase E. Procesos de información-decisión.** Para cada servicio identificado, se detallan las interacciones (flujos de información y de toma de decisiones) necesarias para llevar a cabo el servicio. Además, se definen los contratos de calidad de servicio a los que se comprometen los proveedores y consumidores cuando éste sea llevado a cabo.
- **Fase F. Dinamicidad del sistema abierto.** Se establece la funcionalidad ofrecida como sistema abierto, que incluye los servicios que se deben publicitar y las políticas de adquisición y liberación de roles. Además, se diseñan los agentes propios del sistema.
- **Fase G. Sistemas de medición, evaluación y control.** Se cuantifican o evalúan las tareas y actividades; y se establecen mecanismos para determinar si los objetivos del sistema se cumplen. Así mismo, se revisan las normas de la organización para especificar quiénes se encargan de ellas y las supervisan.
- **Fase H. Sistemas de recompensas.** Se determina el sistema de incentivos, para recompensar a los miembros que avancen en dirección de los intereses de la organización. También se analizan los sistemas de sanción para aquellos miembros que no cumplan con las normas dadas.

Según la metodología, las fases A y B se corresponden con el análisis de la arquitectura, mientras que el diseño se corresponde las fases C y D. Finalmente, el diseño de la dinámica de la arquitectura se realiza en las fases finales (E, F, G y H).

A continuación, en las siguientes secciones, se realizará una descripción de los principales componentes de la arquitectura propuesta, identificando los aspectos innovadores de cada uno de ellos.

4.2.1. Identificación de los componentes de la arquitectura

La arquitectura que se propone se basa en aspectos organizativos y, por tanto, es necesario identificar la estructura organizativa a utilizar. Para ello, el primer paso ha consistido en la identificación de sus componentes, lo que permite establecer el modelo de interacción partiendo del análisis de las necesidades de los usuarios potenciales del sistema. Posteriormente, a partir de este análisis ha sido posible deducir los roles de los usuarios y componentes que participan en el sistema y la forma en la que van a intercambiar información.

En este sentido, el desarrollo de un sistema de monitorización y gobierno de un entorno CC siguiendo un modelo de diseño basado en SMA difiere respecto a los modelos tradicionales de control en este tipo de plataformas, donde habitualmente la toma de decisiones se realiza de forma centralizada [21]. En el marco de esta tesis doctoral se sigue un modelo alternativo, basado en la teoría de agentes y los SMA. En este sentido, se han distribuido las responsabilidades, principalmente de monitorización y toma de decisiones, a lo largo de todos los componentes de la plataforma. Gracias a este modelo es posible que la toma de decisiones se realice allí donde se recoge la información, sobre la base que proporciona el conocimiento local, lo que ha permitido el diseño de procesos de control ágiles basados en información con incertidumbre y la interacción entre semejantes. Esta peculiaridad provoca, incluso, que a medida que el sistema se adapta a la demanda siguiendo el principio de elasticidad de los sistemas CC, parte de los agentes entren y salgan del sistema en función del ciclo de vida de los componentes físicos donde se sitúan. En la Figura 19 se puede apreciar como cada uno de los agentes/roles que participan en la organización están situados a lo largo de todo el entorno computacional.

Siguiendo este modelo de distribución que se indica, en cada servidor físico del entorno CC existe un agente encargado de la monitorización, y otro del control a nivel local, entre ambos tienen la autoridad para gobernar totalmente el servidor físico donde se sitúan, lo que a su vez implica el reparto de recursos en las máquinas virtuales que aloja. Pero, cuando se tenga que realizar una distribución de recursos que conlleve la asignación o retirada de nodos a un determinado servicio entonces, es otro agente especializado y que también está situado en cada uno de los nodos físicos de la infraestructura, el encargado de la toma de este tipo de decisiones que implican a más de un nodo físico de la plataforma CC.

Siguiendo un modelo similar, a cada servicio ofertado a los usuarios se le asocian dos agentes, uno para la monitorización y otro para el control que son los encargados de asegurar el cumplimiento de los acuerdos SLA previamente establecidos. Físicamente están situados en el nodo que realiza el balanceo de carga entre los diferentes nodos trabajadores lo cuál les permite disponer de información precisa para la correcta toma de decisiones en este nivel. En este sentido, las tareas a este nivel están relacionadas con el balanceo de la carga entre los diferentes nodos, la detección de errores y, principalmente, la monitorización de los parámetros de calidad del servicio.

Finalmente, existen otros agentes situados en el punto de entrada del sistema CC con tareas muy dispares. En primer lugar, dos agentes de control, el primer de ellos que se encarga de controlar la

infraestructura *hardware*, su estado, así como el arranque o parada de los nodos físicos en función de la demanda. Por otro lado, otro agente supervisor que es un controlador global que verifica que el resto de los componentes y agentes funcionan correctamente de acuerdo con su especificación. Finalmente, también existe un agente encargado de establecer acuerdos de servicio con los usuarios de la plataforma.

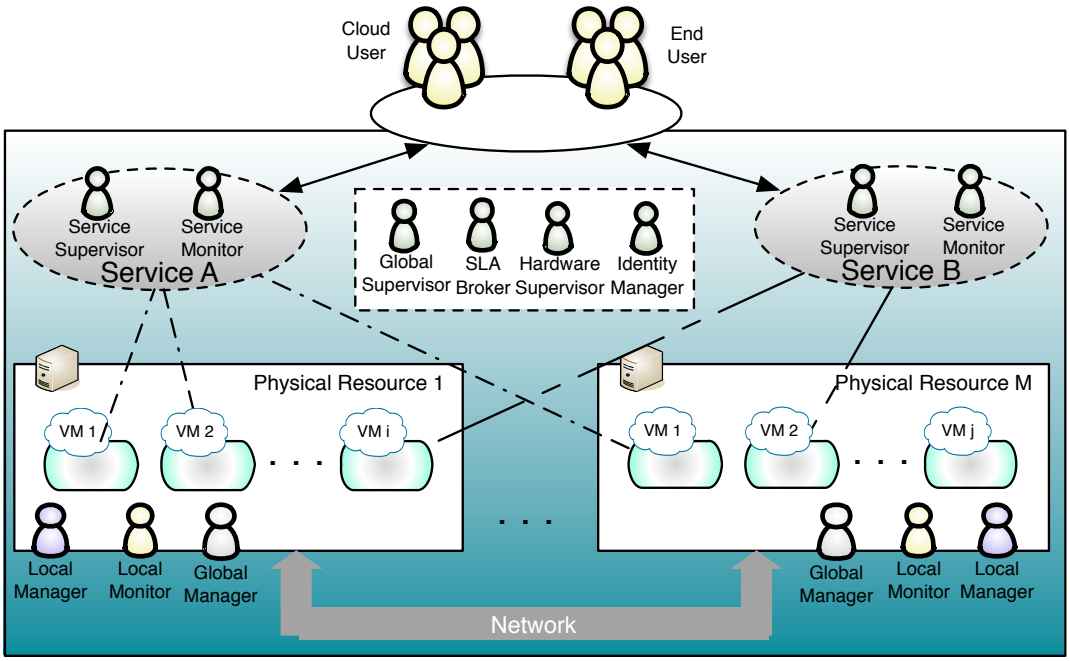

Figura 19.- Distribución roles en la infraestructura

A continuación, se describen en detalle los roles que se han identificado después de analizar de la interacción. Cada uno de los roles tiene unas funcionalidades que son capaces de ejecutar y unas responsabilidades dentro de la arquitectura, así como capacidades de interacción con otros roles. Cada rol normalmente será instanciado y ejecutado por un agente individual y dedicado, aunque es posible que un agente pueda ejecutar varios roles.

En primer lugar, se comenzará por los roles que están directamente relacionados con los usuarios que hacen uso del sistema. Estos agentes humanos tienen asociados agentes inteligentes virtuales, que facilitan su interacción con el resto de la arquitectura multiagente:

- **Cloud User.** Representa a usuarios físicos, es decir, personas que utilizan el sistema. Este rol se asocia al tipo de usuario *Cloud user* según el modelo de arquitectura descrito por el NIST [10], ya descrito en el apartado 2.5. En este trabajo, dentro de este tipo de usuarios se han definido dos subgrupos:
  - *Desarrollador*. Es un usuario externo que tiene la capacidad de introducir nuevas aplicaciones en el sistema, principalmente aplicaciones en la capa SaaS. Por otro lado, también puede consumir servicios y ofrecer funcionalidades a los usuarios finales.
  - *Manager*. Este usuario puede administrar aplicaciones de la capa SaaS dentro del sistema, configurando parámetros sobre las mismas y negociando con el sistema acerca del uso de los servicios que ofrece la plataforma CC.
- **End User.** Representa al usuario final del sistema, siendo el verdadero consumidor de los servicios y, por lo tanto, también de los recursos de la plataforma a través del uso de las aplicaciones que están desplegadas en el entorno CC. Destacar que estas aplicaciones pueden

ser tanto las aplicaciones nativas del propio entorno CC, como las aplicaciones desplegadas por terceros, es decir, usuarios de tipo *Cloud User*, anteriormente descrito.

- **Identity Manager.** Este rol, que no se asocia a una entidad humana, sino a un ente artificial, se encarga de gestionar la entrada y salida de usuarios en la unidad organizativa que representa el sistema. Para ello, dispone de un repositorio con los usuarios del sistema y encapsula los diferentes procedimientos de autenticación que disponga la plataforma.
  Su labor consiste en iniciar el procedimiento de autenticación cuando otros usuarios (*End User, Cloud User* o *Cloud Admin*) traten de acceder al sistema, de este modo es posible conocer cuál es su tipo de usuario, el nivel de privilegios y el conjunto de acuerdos alcanzados de servicios del sistema en el caso de que existan.

Una vez que se han presentado los agentes humanos y los roles encargados de su entrada en el sistema. A continuación, se prestará atención a los roles encargados de la gestión de la infraestructura subyacente, que como ya se ha indicado se corresponde con una de las características clave de este tipo de plataformas. Así pues, se identifican los siguientes roles directamente relacionados con la monitorización y control del *hardware*:

- **Local Monitor.** Es el encargado de recoger datos acerca del estado de los recursos locales de cada servidor físico, incluyendo tanto a la propia máquina física, como a las diferentes máquinas virtuales que pueda hospedar. Su principal objetivo es recoger del entorno la información que otros componentes de la plataforma CC puedan necesitar, realizando el tratamiento previo de estos datos hasta convertirlos en información relevante para la toma de decisiones.
  Para realizar esta monitorización debe existir al menos un agente asociado a este rol en cada servidor físico, el cuál debe tener un conocimiento detallado de cada nodo individual, pero al mismo tiempo tiene un desconocimiento total acerca del resto de nodos de la plataforma CC. Dado que existe la posibilidad de que se paren y arranquen nodos de ejecución, entonces también existe la posibilidad de que agentes de este tipo entren y salgan del sistema de forma dinámica, ligando su ciclo de vida al del servidor que monitorizan.
- **Local Manager.** Encargado del control y distribución de los recursos computacionales de la máquina física, pudiendo redistribuir recursos entre las instancias en ejecución, lanzar o apagar máquinas virtuales a partir de plantillas de servicio previamente configuradas.
  Su objetivo es, por tanto, la gestión eficiente de los recursos del servidor físico entre los diferentes nodos que hospeda, maximizando su uso, pero sin disminuir la calidad de los servicios que se están proporcionando; y, garantizando que la máquina física siempre dispone de los recursos mínimos indispensables para poder realizar las tareas de coordinación y control, así como para realizar el procesamiento necesario de monitorización.
  Al igual que sucedía con el rol Local Monitor, este rol debe ser adquiridos por un agente en cada uno de los servidores físicos del entorno CC, el cuál trabaja en estrecha colaboración con el agente anterior, es decir el agente *Local Monitor* del mismo equipo, para realizar una monitorización completa y un control y adaptación automática de cada servicio individual. Por lo tanto, ambos roles tienen un conocimiento y control total sobre el estado y reparto de recursos en cada equipo individual.
- **Global Manager.** Es el rol encargado de la toma de decisiones acerca de cómo distribuir los recursos computacionales entre varios nodos de la plataforma CC. El objetivo es la distribución de recursos a nivel global, y no sólo a nivel local como sucedía con el rol *Local Manager*. Como ayuda a la toma decisiones acerca de cómo distribuir los recursos, utilizan

una base de conocimiento parcial (proporcionada por el rol *Local Monitor*) y las experiencias pasadas almacenadas en el repositorio de histórico de uso.
Son los encargados de determinar cómo y con qué características se instanciarán nuevas máquinas virtuales que se asocien a servicios. Así como el proceso de compactación de las máquinas virtuales existentes en un número menor de servidores, para que de este modo sea posible apagar servidores físicos y reducir el consumo energético.

- ***Hardware* Manager.** Es el rol encargado de la gestión de los nodos *hardware* existentes en el entorno CC. Tiene acceso un repositorio con los nodos existentes y el estado de cada uno, siguiendo la caracterización de cada nodo, la cuál ya se ha presentado en el apartado 4.1.2.1.
  Sus tareas son la gestión del estado de cada nodo del sistema, manteniendo el repositorio actualizado, así como el arranque y parada de los equipos físicos según determinen los algoritmos de elasticidad desarrollados.

Otro de los pilares de un entorno CC es la capacidad de ofertar servicios computacionales. La gran diferencia de un entorno CC respecto a otras tecnologías anteriores es la capacidad de ofertar los servicios bajo demanda siguiendo un modelo de pago por uso [6]. Para lo cuál es necesario establecer un modelo de acuerdo de uso en los servicios a través de SLAs específicos. Posteriormente, una vez se ha establecido este acuerdo, la plataforma debe realizar una monitorización precisa del servicio que permita asegurar que se están cumpliendo los niveles de calidad acordados, o por el contrario es necesario tomar medidas que aseguren su cumplimiento. Así en este sentido, se observan los siguientes roles implicados en esta tarea:

- **Service Monitor.** Es el rol encargo de monitorizar cada uno de los servicios ofertados por el sistema. Para ello recoge información acerca las peticiones que estén realizando los usuarios, midiendo parámetros de calidad sobre las mismas, rendimiento, errores, etc. Al igual que sucedía con el rol *Local Monitor*, deberá pre-procesar la información recibida para que pueda estar disponible en tiempo y forma para el proceso de toma de decisiones. Así mismo, este agente también dispone de acceso a un repositorio histórico donde guarda la información relativa a la demanda histórica del servicio que monitoriza.
  Físicamente se sitúa en el nodo que balancea las peticiones entre las diferentes instancias que atienden el servicio, de forma que pueda monitorizar toda la información referente al intercambio de peticiones. Por lo que existe un agente asociado a este rol por cada uno de los servicios que oferta la plataforma.
- **Service Supervisor.** Es el encargado de tomar decisiones acerca de cada servicio individual, tomando como base la información que le proporciona el rol *Service Monitor*.
  Su principal labor es la de asegurar los acuerdos acerca del uso del servicio que controla, tomando las acciones adecuadas en el caso de que no se estén cumpliendo los acuerdos SLA previamente establecidos. Las acciones posibles cuando se detecta una pérdida del rendimiento que puede incurrir en una violación de los acuerdos establecidos pasan por la solicitud de recursos adicionales al rol *Local Manager* (primero) y *Global Manager* (después). Finalmente, también se asegura de que al menos existen un número determinado de nodos trabajando en equipos físicos independientes, lo que permite asegurar de este modo la disponibilidad del servicio monitorizado.
- **SLA Broker.** Representa al conjunto de entidades encargadas de establecer acuerdos de uso entre la plataforma CC y los usuarios externos que pretenden hacer uso de los servicios ofertados. Esta tarea es compleja ya que implica un proceso de negociación con los usuarios finales, principalmente, con el rol *Cloud User*; así como la recogida de información a lo largo

de todo el entorno CC para establecer un modelo de coste de los servicios en función de su calidad y la aceptación o rechazo de nuevos usuarios y/o acuerdos SLA.
En este sentido, este rol deberá recoger información de los siguientes componentes de la plataforma:

- Interacciona con el rol *Service Monitor* para recuperar información acerca del uso y calidad de los servicios, lo que permite elaborar un modelo de coste en función de la calidad existente.
- Interacciona con los roles *Hardware Manager* y *Global Manager* para obtener información sobre el grado de utilización de la infraestructura que permita conocer si existen recursos disponibles para asumir el establecimiento de nuevos acuerdos a largo plazo.
- Finalmente, interacciona con el *repositorio de histórico de uso* para obtener información relativa al consumo medio del conjunto de servicios en el pasado, de forma que esta variable también influya en el modelo de coste y previsión de acuerdos SLA

A partir de esta información y del modelo de coste establecido es posible realizar acuerdos con los usuarios de la plataforma a través de la negociación en cuanto al uso y calidad de servicios. En este sentido, es necesario reseñar que esta faceta del paradigma CC se escapa de este trabajo de investigación. No obstante, en el estado del arte se observan una gran variedad de técnicas y algoritmos, algunos de ellos basados en SMA, con la suficiente madurez en cuanto a su desarrollo para establecer acuerdos con los usuarios [9, 180, 223, 224, 378].

Finalmente, una vez que se han presentado los roles que fundamentalmente hacen posible gobernar un entorno CC, a continuación se presenta el último rol que se encarga de la supervisión y control de todo el sistema:

- **Global Supervisor.** Este rol se encarga de supervisar que el resto de los roles y componentes de la plataforma funcionan correctamente según sus especificaciones. En el caso de que algo falle, o alguno de los agentes no responda a sus mensajes, es el encargado de tomar las acciones que sean necesarias para devolver el sistema a un estado de funcionamiento óptimo.
  Hay que destacar que este rol puede ser adquirido por un agente que interaccione con un supervisor humano que represente al administrador sistema y cuya labor es la de monitorizar el correcto funcionamiento del mismo. Para ello tiene acceso a los datos de monitorización del sistema, así como a las herramientas que permiten reconfigurar la redistribución de los recursos de forma supervisada.

A partir de esta identificación de roles que participan en el sistema, es posible diseñar la organización de forma unificada, intuitiva y con un alto nivel de abstracción [379, 380]. Mediante esta aproximación se facilita y se simplifica el proceso de diseño de SMA orientado a organizaciones obteniendo modelos de OV que pueden ser implementados en diferentes plataformas. En el siguiente subapartado se presenta en detalle el diseño de la arquitectura que se propone en el marco de este trabajo de investigación.

4.2.2.Diseño de la arquitectura

+Cloud es un SMA basado en organizaciones para cuyo diseño se han seguido las pautas propuestas por la guía metodológica GORMAS [49] desarrollada en el marco de los proyectos de investigación THOMAS[11] [381, 382], OVAMAH[12] [269] e iHAS[13]. Esta metodología permite

[11] http://thomas-tin.usal.es/
[12] http://gti-ia.dsic.upv.es:8080/ovamah

diseñar una organización de forma unificada, intuitiva y con un alto nivel de abstracción. Mediante esta aproximación se facilita y simplifica el proceso de diseño de SMA orientado a organizaciones obteniendo modelos de OV que pueden ser implementados en diferentes plataformas. En esta sección se presentarán los aspectos principales del diseño. Dada la extensión de los artefactos obtenidos durante el proceso diseño, en este apartado tan sólo se incluye un resumen de los productos principales que permiten describir la unidad organizativa de +Cloud.

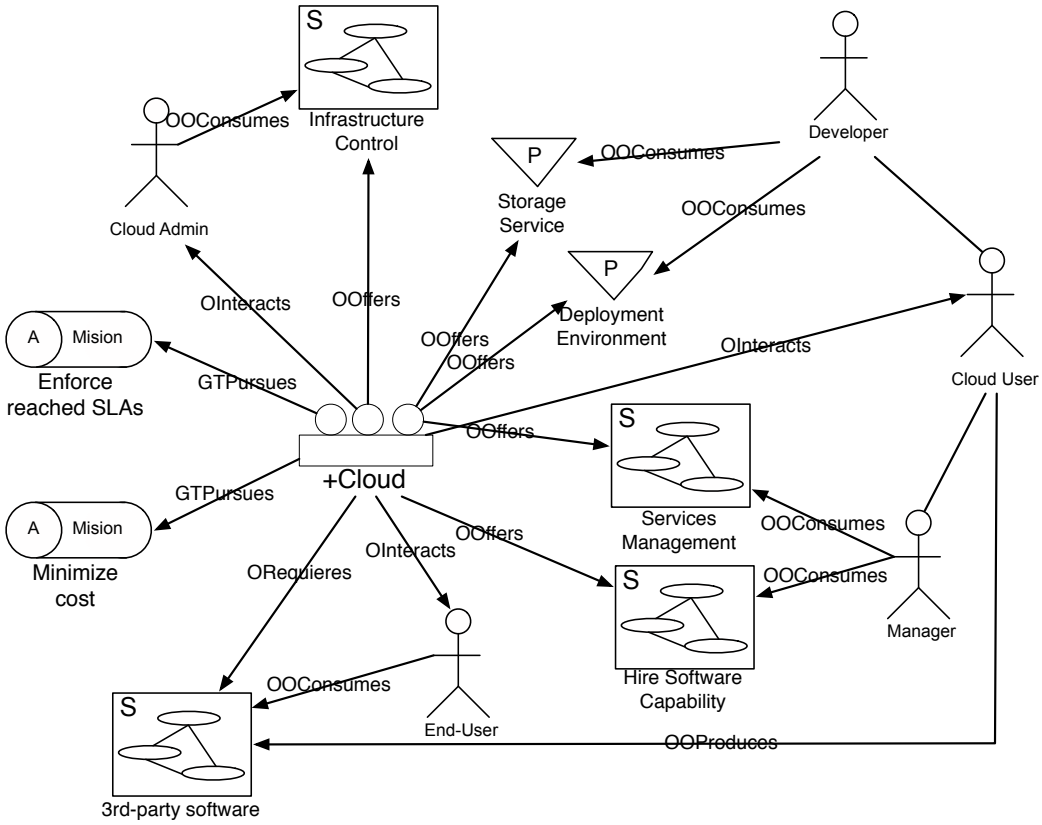

Figura 20.- Vista funcional (misión) del modelo de organización de +Cloud

Siguiendo las pautas indicadas en la guía metodológica GORMAS, una de las primeras tareas es instanciar la vista funcional (*misión*) del modelo de organización, que se presenta en la Figura 20. Esta vista presenta los productos y servicios que ofrece el sistema, los objetivos globales que se persigue (misión y justificación) y los grupos de interés a los que afecta.

Así, en primer lugar, la misión y razón de existencia de la organización será el cumplimiento de los acuerdos de servicios alcanzados con el rol *Cloud User*, pero minimizando para ello los costes asociados a esta misión. En el diagrama mostrado en Figura 20 también se aprecian los tipos de usuarios que hacen uso del sistema (*Cloud Admin*, *Cloud User* y *End User*) y los productos que se ofertan (almacenamiento y despliegue *software*). Para facilitar la interacción de la plataforma también se ofertan los siguientes servicios intrínsecos: gestión de *software*, contratación de *software* y control de la infraestructura. Dentro de los servicios ofertados, hay que destacar que la plataforma también ofrece como servicio las aplicaciones que son desplegadas en el sistema por terceros (*Cloud User*), es decir, este tipo de aplicaciones son requeridas por la plataforma para justificar la necesidad de ofertar productos de almacenamiento e infraestructura, sin embargo, dado que una plataforma CC es un simple medio (y no un fin), estas aplicaciones también son servicios que se ofertan a los usuarios finales (*End User*).

[13] http://ihas.usal.es/

Según la metodología GORMAS, en la Fase C se especifican las cuatro dimensiones organizativas (departamentalización, especialización, coordinación y normalización), según el modelo propuesto por Mintzberg [285]. Las dimensiones del sistema +Cloud se presentan en la Tabla 6 correspondiente al Documento C propuesto en la metodología. En primer lugar, en cuanto a la departamentalización se propone un modelo *funcional* agrupando las tareas según la especialización de los agentes que las realizan. La especialización es *horizontal* precisamente debido a este alto grado de especialización, pero también se desea que cada agente individual tenga poder de decisión y responsabilidad sobre las tareas que realiza, lo que conlleva a una especialización donde existe *ampliación vertical*. Por su parte, la coordinación en el sistema se produce de forma *descentralizada,* existiendo *adaptación mutua*, ya que cada agente es el responsable de sus propias tareas pero, para conseguir los objetivos perseguidos con la realización de las mismas, debe interaccionar con otros agentes de la unidad organizativa. Finalmente, en cuanto a la última dimensión, la de formalización, se dice que hay *normalización de resultados* ya que importa la calidad de los mismos, porque los productos tienen que estar dentro de los SLA acordados con los usuarios finales.

A partir de este análisis de las dimensiones organizativas, se analiza la departamentalización y los mecanismos de coordinación existentes dentro del sistema. Para ello, en la metodología se propone un árbol de decisión. Sin embargo, a partir del análisis dimensional realizado, si se siguen las diferentes ramas de este árbol de decisión no se obtiene una estructura ideal. No obstante, después de realizar un análisis más detallado, en el que se tiene en cuenta el propio árbol de decisión, el tamaño del sistema y la necesidad de coordinación existente entre tareas, se selecciona la estructura de *coalización* como la más adecuada para modelar el sistema. Se ha seleccionado este patrón de organización ya que las diferentes entidades que forman parte de una estructura de este tipo forman grupos en función de la similitud en cuanto a la funcionalidad, coordinándose entre sí para ofrecer una funcionalidad global más compleja y elaborada, descripción que se ajusta a la perfección con el modelo de coordinación que se perseguía en +Cloud.

C. Dimensiones organizativas
Departamentalización: *Funcional* • Se agrupan las tareas en función de los conocimientos y habilidades de quién las realiza (control de servicios, gestión de infraestructura, etc.). Esto es así, debido al grado de complejidad y especialización de las tareas a realizar, en este sentido se tiene en cuenta que la producción de los resultados es intensiva.
Especialización + Coordinación: *Especialización horizontal con ampliación vertical. Descentralización* • Los roles que realizan las diferentes tareas están altamente especializados debido a la complejidad del entorno con el que se interacciona para satisfacer la funcionalidad que se busca. Este tipo de organización se asocia a la especialización horizontal. Además, tienen el control sobre el mecanismo a utilizar para la consecución de los objetivos perseguidos. Sin embargo, también se desea que los roles tengan flexibilidad y capacidad de decisión sobre las tareas a realizar, lo que se asocia a un modelo de organización con ampliación vertical. • La coordinación se realiza de forma descentralizada por lo que se hace necesaria negociación entre los roles para llevar a cabo la toma de decisiones.
Coordinación y Formalización: *Adaptación mutua y normalización de resultados* • Existe adaptación mutua ya que el modo de realización de las tareas es flexible y requiere coordinación entre los diferentes roles implicados. • Existe normalización de resultados ya que importa la calidad de los resultados obtenidos, teniendo en cuenta que se está trabajando en un sistema de alta disponibilidad. El sistema es libre de tomar las acciones que considere oportuno para llegar a alcanzar los resultados previstos.

Tabla 6.- Documento C.- Dimensiones organizativa

A partir del patrón de organización seleccionado, coalización en este caso, el siguiente paso ha consistido en el análisis del proceso de departamentalización, lo que ha dado lugar a la segmentación de la unidad organizativa principal (+Cloud) en las siguientes (sub-)unidades organizativas especializadas:

- **SLA Negotiation.** Es la unidad organizativa encargada de agrupar las tareas de establecimiento de acuerdos y la negociación de los mismos con el rol externo Cloud User. Este proceso tiene que estar totalmente automatizado, así mismo se debe determinar si el sistema tiene capacidad para asumir nuevos acuerdos sin violar los existentes.
- **Service Management.** Es la unidad que tiene asignadas las tareas de monitorización y supervisión de productos que se ofrecen y si éstos están siendo ofertados atendiendo a los acuerdos de calidad estipulados en el acuerdo de uso específico con cada usuario. Las tareas de esta unidad organizativa, al igual que las anteriores también tienen que estar automatizadas.
- **Infraestructure Management.** Es la unidad organizativa que controla la infraestructura hardware subyacente, es decir, los recursos computacionales (reales o virtuales) del sistema mediante la gestión y monitorización de los mismos. Aunque las tareas también deben estar automatizadas, esta unidad organizativa puede estar supervisada por el A-Agente Cloud Admin, el cuál es externo y se asocia al rol interno Global Supervisor, cuya funcionalidad ya se ha descrito en el apartado 4.2.1

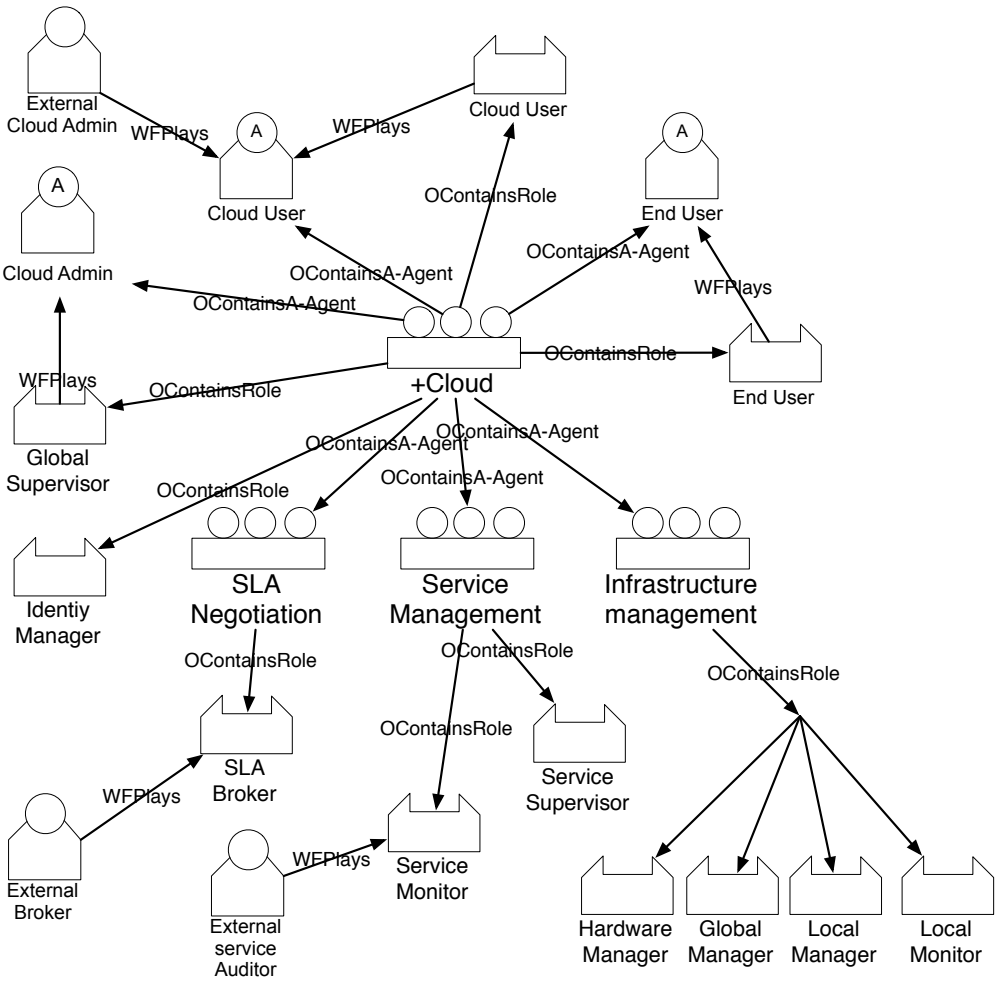

Figura 21.- Vista estructural del modelo de organización actualizado

Siguiendo con el modelo metodológico GORMAS utilizado para el diseño de +Cloud, en la Figura 21 se presenta la vista estructural del modelo de organización donde se puede apreciar cómo la unidad organizativa principal +Cloud se ha segmentado en las tres unidades que se acaban de explicar. Así mismo, en esta Figura 21 también se aprecia cómo +Cloud se ha diseñado como un sistema abierto, en el que se permite la existencia de agentes externos que ofrezcan dentro del sistema servicios y aporten funcionalidad al sistema. En este sentido, esta vista estructural presenta los diferentes tipos de agentes, ya sean internos o externos, que forman parte del sistema. En este sentido cabe destacar, en primer lugar, los roles que están directamente relacionados con agentes humanos los cuáles, han sido claramente identificados previamente. Sin embargo, también existen otros dos roles dentro del sistema +Cloud cuyas funcionalidades pueden ser desarrollados por agentes externos. Éstos son el rol *SLA Manager* y *Service Supervisor*, que pueden asociarse a los roles *Broker Cloud* y *Auditor Cloud*, respectivamente, siguiendo el modelo de roles propuestos por el NIST en su arquitectura de referencia [10].

Finalmente, como último producto a destacar del diseño de la arquitectura se describirá el entorno en el que se sitúa el SMA organizativo propuesto. En GORMAS durante la fase de análisis se realiza una caracterización del entorno, recogido en el Documento A.3. (Tabla 7). En este sentido, el entorno es dinámico, complejo, con incertidumbre y hostil. Estas características definen al entorno complejo donde los SMA, y especialmente aquellos basados en OV consiguen un mayor grado de efectividad.

A.3- Condiciones de entorno		
Condición	**Valor**	**Justificación**
Tasa de cambio	Dinámico	Ya que constantemente pueden cambiar los acuerdos con los usuarios y el *software* desplegado en el sistema +Cloud.
Complejidad	Complejo	Existe un gran número de componentes en el sistema con múltiples relaciones entre ellos.
Incertidumbre	Alto	Es un entorno dinámico y complejo.
Receptividad	Hostil	Ya que el *software* requerido es desarrollado por terceros y podría contener errores.
Diversidad	Uniforme	Ya que los servicios que se ofertan y se proporcionan son homogéneos (servicios computacionales).

Tabla 7.- Documento A.3.- Condiciones de entorno

El entorno, además de ser el lugar donde se sitúan los diferentes agentes que forman parte del sistema, debe proporcionar los puertos de acceso a la reconfiguración del mismo. Así tal y como se puede aprecia en la Figura 22, donde se muestra una vista completa del modelo de entorno, los diferentes roles del sistema interaccionan con el entorno para conseguir sus objetivos individuales a través de los puertos que este proporciona, ya sean para acciones de lectura o de modificación:

- Dos aplicaciones que actúan con interfaces con los roles externos del sistema. Es decir, la aplicación que permite monitorizar la infraestructura subyacente del entorno CC, específicamente diseñada para el rol *Cloud Admin* y el escritorio web que dispone de las aplicaciones específicas para el rol *Cloud User*. Los usuarios finales acceden directamente a las aplicaciones de terceros que están desplegadas en el sistema.
- Tres repositorios asociados a la información que debe ser persistente, el almacén de usuario, los acuerdos establecidos con estos usuarios, el *hardware* real disponible y el histórico de uso.
- También se puede apreciar la existencia de diferentes puertos para el acceso a cada uno de los recursos del entorno. En este sentido, por un lado se contemplan los puertos para el

acceso a los recursos que ya se han identificado previamente (repositorios e interfaces). Y, por otro lado, destacan otros tres puertos que permiten la reconfiguración de la infraestructura subyacente (control del entorno *hardware*, entorno virtual y sistemas de balanceo).

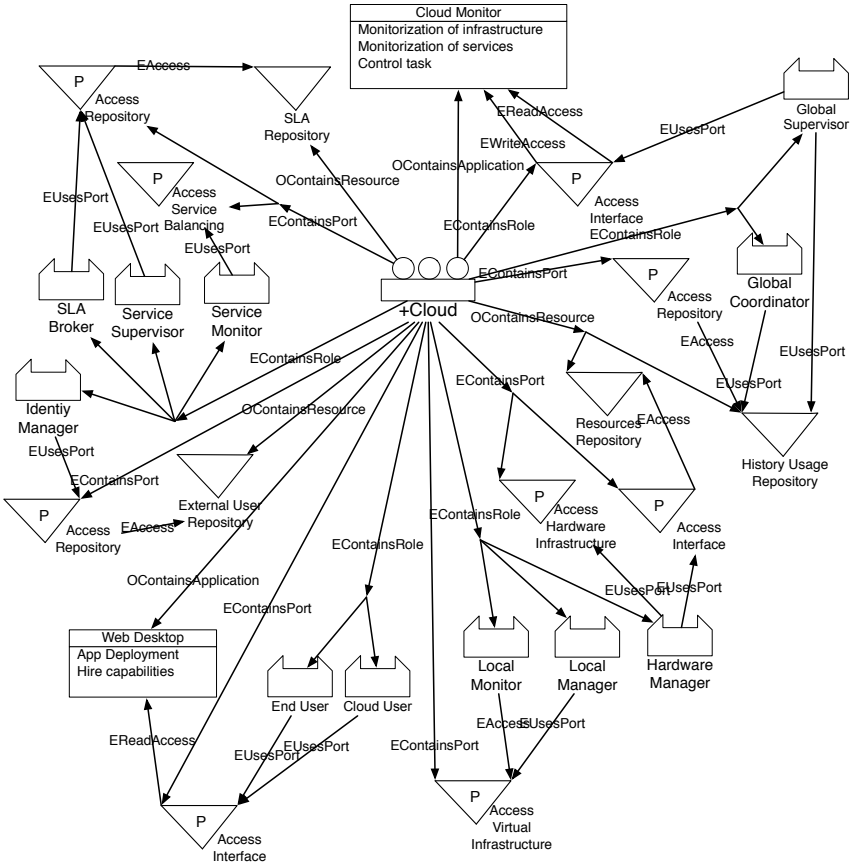

Figura 22.- Modelo de entorno actualizado. Acceso a los puertos del entorno

Con la descripción del entorno, termina la descripción de los principales artefactos que se han generado durante el análisis y diseño de SMA basado en OV +Cloud. La conclusión después de presentar la solución propuesta es que este sistema pretende ser una arquitectura integradora, en la que una organización de agentes pueda gobernar y ejecutar diferentes acciones a lo largo de todo el entorno CC. La novedad de dicha arquitectura radica en el hecho de utilizar un modelo organizativo para llevar a cabo sus objetivos, proporcionando una innovadora solución en entornos altamente dinámicos. En la arquitectura propuesta se plantea una distribución de funciones analizada y diseñada a través de una guía metodológica, quedando abierta la posibilidad de añadir y modificar nuevas funcionalidades de manera sencilla.

Una vez que se ha descrito la arquitectura que se ha diseñado, el siguiente apartado presenta el modelo de razonamiento que siguen los agentes individuales que participan en el sistema para llevar a cabo la distribución de recursos computacionales. Estos algoritmos, siguen la filosofía de los SMA, basada en la autonomía de los componentes, así como en la toma de decisiones distribuida. Gracias a este modelo, se posibilita no sólo una optimización en el proceso de distribución de recursos, sino también mejorar la disponibilidad del sistema. Este modelo de distribución de recursos es en sí mismo innovador ya que, como se presentará en el apartado siguiente, los modelos actuales son centralizados y, por su puesto, no tienen en cuenta la autoadaptación dinámica del sistema en función de los cambios que se produzcan en el entorno.

4.3.Modelo de distribución de recursos en un entorno Cloud Computing

En el apartado anterior se han descrito los diferentes componentes del modelo de arquitectura propuesto para el control y monitorización de un sistema CC. Este modelo arquitectónico está basado en OV de agentes inteligentes lo que hace posible la adaptación dinámica del sistema a los cambios que se producen en el entorno [24]. Para ello, los agentes artificiales, que forman parte de la organización, ejercen el papel de alguno o varios de los roles (y sus responsabilidades asociadas) descritos en el apartado 4.2.1. Estos agentes tienen la capacidad de integrar mecanismos y métodos de razonamiento avanzado, que permiten tomar decisiones de forma individual (pero coordinada) que hace posible no sólo interactuar sobre el entorno, sino también adaptar dinámicamente la propia organización y sus componentes para responder a los cambios que se producen en el propio entorno. El presente apartado tiene como objetivo describir los tipos de agentes especializados en la distribución de recursos computacionales según el modelo propuesto. Algunos de los cuáles, dado que su ciclo de vida está ligado al de las máquinas en la que residen, también permite describir un modelo auto-adaptativo de la arquitectura +Cloud.

Este modelo permite hacer frente a la demanda en los servicios que se ofertan por la plataforma CC, sin violar los acuerdos SLA establecidos. Es decir, gracias a los diferentes agentes existentes con capacidades avanzadas de razonamiento y adaptación que se integra en la arquitectura, es posible distribuir los recursos físicos disponibles entre los diferentes recursos virtuales, y por ende, entre los diferentes servicios que son proporcionados por estos recursos computacionales. En este sentido, el problema reside en cómo distribuir eficientemente los recursos disponibles, haciendo uso únicamente de aquellos recursos necesarios para el correcto funcionamiento del entorno, lo que permite seguir una aproximación que se acerca al modelo de producción tradicional *just-in-time* [14]. Con relación a este modelo de distribución, que se presenta a lo largo de esta sección, es necesario destacar que está altamente influenciado por las tecnológicas de virtualización existentes en la actualidad [1].

Tal y como se ha descrito al principio del capítulo (apartado 4.1.2), a nivel general, un sistema CC está compuesto por un conjunto de máquinas físicas que alojan máquinas virtuales, cada una de ellas asociada a un servicio determinado. Cada máquina virtual individual, o nodo (*worker*) atiende un servicio concreto de la plataforma. Uno o varios nodos pueden atender un único servicio, estos nodos para proporcionar el servicio hacen uso de recursos computacionales. Estos recursos computacionales son ofrecidos al entorno CC por las máquinas físicas que dispone el sistema. La Figura 23 muestra un pequeño ejemplo del funcionamiento de este modelo en un entorno CC, se puede observar como un servicio tiene asociado tres nodos que se encargan de satisfacer la demanda de peticiones. Estos nodos se encuentran hospedados en diferentes servidores físicos de la plataforma CC. Para que los usuarios finales (rol *End User*) sean liberados de esta complejidad inherente, es necesario que exista un sistema capaz de balancear la demanda de peticiones entre los diferentes nodos que en un momento determinado atienden el servicio.

En cuanto a estas máquinas físicas, es necesario hacer notar que no todas ellas tienen por qué estar en funcionamiento en todo momento. En este sentido, en los sistemas CC modernos los servicios deben activarse o desactivarse en función de la demanda existente. De este modo, se contribuye al ahorro de costes, siguiendo el paradigma computacional *Green Computing* [383, 384].

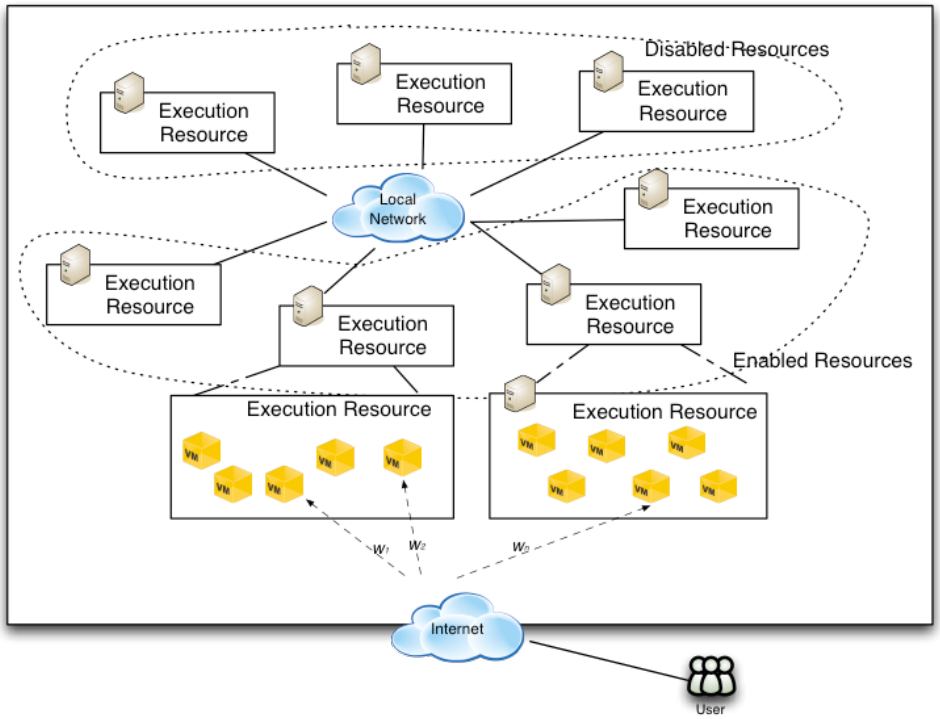

Figura 23.- Distribución de recursos en sistemas Cloud Computing

La distribución de recursos dentro de estos sistemas computacionales es una tarea compleja, donde la infraestructura subyacente ofrece una gran cantidad de posibilidades. Así, es posible distribuir recursos tanto a nivel local, en cada servicio individual o nodo físico, como a nivel global, entre diferentes nodos físicos. En este sentido, el correcto reparto de recursos computacionales en un entorno CC es una de las decisiones clave, debido a que, en función de esta distribución, es posible maximizar la eficiencia del entorno y la disminución de los costes asociados. Sin embargo, esta tarea es compleja, ya que requiere tanto del análisis de la demanda actual y futura de los servicios que se alojan, como de los recursos disponibles en cada momento en el entorno CC. Así pues, atendiendo a las capacidades que permite la tecnología subyacente, la distribución de recursos se presenta desde diferentes puntos de vista, tal y como sigue:

- **Distribución a nivel servicios.** A este nivel, el más alto posible, la distribución de recursos consiste en el simple balanceo de carga entre los nodos trabajadores que atienden un determinado servicio. Este tipo de balanceo está altamente extendido desde el nacimiento de los sistemas computacionales de tipo clúster en entornos HPC [385, 386].
  El objetivo habitual en este tipo de reparto de carga computacional consiste en equilibrar la carga de todos los nodos que están ofertando el servicio [21, 33].
- **Distribución a nivel infraestructura.** La distribución a nivel infraestructura es un método más complejo y novedoso, que es posible gracias al nacimiento de la tecnología de virtualización [72]. De forma previa al nacimiento de esta novedosa tecnología, la distribución de recursos a nivel infraestructura estaba basada en la incorporación de nodos (servidores físicos) al sistema de balanceo, lo cuál en la mayoría de los casos era un proceso manual, por lo que habitualmente la asignación terminaba siendo estática [387].
  Con el nacimiento de la virtualización, la distribución es mucho más potente, dinámica y, además, puede ser automatizada, ya que los nodos que se incorporan son virtuales y pueden ser situados en cualquier servidor físico gracias al uso de plantillas de servicio para la instanciación, es decir, el arranque de máquinas virtuales de forma dinámica. Incluso, recientemente, con el avance de las técnicas de virtualización y el nacimiento del nuevo *kernel*

de Linux es posible modificar los recursos computacionales asignados a una determinada máquina virtual que esté en ejecución.

En el marco de este trabajo, se hablará de distribución de recursos a nivel micro cuando la reasignación de recursos se realice dentro de un servidor físico, y a nivel macro, cuando la reasignación incluya diferentes servidores físicos.

4.3.1. Trabajos relacionados

Como se ha venido comentando desde el inicio del capítulo, los recursos de un entorno CC son compartidos entre los numerosos clientes que atienden, lo que implica que sea necesario un ajuste dinámico en función de la demanda. Los modelos existentes de distribución de recursos tienen en cuenta habitualmente tres conceptos básicos [34]: los acuerdos SLA establecidos con los usuarios, el consumo de energía del entorno CC y el propio estado del entorno CC. Este problema, por su complejidad y contemporaneidad está siendo de gran interés por la comunidad científica, la cuál está proponiendo diferentes aproximaciones para hacer frente a este problema abierto en los sistemas CC.

El problema puede observarse desde dos puntos de vista [21]. Por un lado, la búsqueda del proveedor más barato y eficiente en cuanto al uso de recursos, para lo cuál se sigue un modelo en el que un bróker o gestor de recursos, habitualmente externo, está en permanente contacto con varios proveedores, seleccionando el más adecuado para cada cliente en cada momento. Por otro lado, otro enfoque de mayor interés en el marco de este trabajo, consiste la búsqueda de la eficiencia de recursos dentro de los centros de proceso de datos de un gran número de proveedores de servicios CC. Dentro de este segundo grupo, también se observa el problema desde dos perspectivas, a pequeña y gran escala [388, 389]:

- Un proveedor de recursos CC a gran escala, que disponga de diferentes centros de datos distribuidos a lo largo de todo el mundo, debe dirigir las peticiones de los clientes según dos variables, la distancia al centro de datos y la carga instantánea de cada centro. Así será posible reducir la latencia en la respuesta a los clientes. A este modelo se le conoce como modelo de distribución basado en tiempo (*time-drive adaptive mechanism*)[389].
- Nos referimos a pequeña escala, cuando la distribución de recursos se realiza dentro de un centro de datos individual, con independencia de su tamaño. La tarea de distribución de recursos dentro de este medio se realiza por un componente que habitualmente se denomina *Resource Allocation System* (RAS). Dentro de esta perspectiva del problema, en el estado del arte existen principalmente dos aproximaciones [21]:
    - **QoS-awared based**, u orientados al mercado [34]. Este primer grupo se asocia a un modelo de distribución de recursos orientados al cliente en el que se trata de minimizar los riesgos computacionales con el objetivo de distribuir los recursos computacionales en función de los acuerdos SLA alcanzados, siguiendo el modelo económico de pago por uso.

      Así, según este modelo, las técnicas de gestión de recursos computacionales tienen como objetivo no violar nunca estos acuerdos, proporcionando al usuario final la calidad del servicio que se ha contratado y por lo tanto que el usuario espera recibir. En este sentido, en el estado de arte existen trabajos dentro de esta aproximación como RAS-M [35] que propone un modelo basado en la economía de mercado donde se busca una redistribución de recursos que de lugar a un precio de mercado justo. Otros trabajos utilizan base matemática para la distribución de recursos como la teoría de juegos [31] o la maximización de una función de optimización, basada en técnicas de

programación lineal, ya que al fin y al cabo el problema se reduce a la optimización de recursos [36, 390].

- **Energy-awared based** [21]. En segundo lugar, la distribución de recursos puede realizarse teniendo en cuenta no sólo los acuerdos SLA preestablecidos, sino también el consumo energético que implica el cumplimiento de éstos. El número de trabajos existente en el estado del arte es menor que en el anterior grupo, y los existentes son más novedosos. Así pues, existe una gran variedad de técnicas como la aplicación de políticas de ahorro de energía en las máquinas físicas o virtuales [391], redistribución eficiente de las máquinas virtuales en función del consumo de cada equipo físico [28] o modelos que se basan en técnicas de optimización [392]. Todos ellos, se encuentran en un estado de desarrollo inicial que dificultan su aplicación real en grandes centros de cómputo.

El problema, aunque complejo, es simple en su declaración, ya que tan sólo se basa en la redistribución de recursos físicos entre los diferentes nodos virtuales. A partir de estos trabajos que se pueden encontrar en el estado del arte, es necesario destacar que este trabajo de investigación ha centrado el esfuerzo en el desarrollo de algoritmos que tienen en cuenta el consumo energético. En este sentido, el objetivo es reducir el consumo energético necesario para satisfacer los acuerdos SLA que se establezcan con los usuarios de la plataforma. Así mismo, los algoritmos que se han presentado a lo largo de esta sección siguen aproximaciones centralizadas, que utilizan modelos matemáticos y teoría económicas para dar como resultado algoritmos adecuados al problema de resolver.

En este trabajo se sigue una aproximación diametralmente opuesta basada en técnicas de optimización e IA que permiten repartir los recursos siguiendo un modelo distribuido y escalable que hace posible que el sistema aprenda a lo largo del tiempo. Para ello, se toma como base la arquitectura multiagente basada en OV que habilita la interacción dentro los diferentes componentes de la plataforma, y permite introducir algoritmos de razonamiento avanzados que facilitan el desarrollo de modelos autónomos y distribuidos basados en la autoadaptación dinámica a los cambios que se puedan introducir en el entorno.

4.3.2.Propuesta de agentes especializados en la distribución de recursos

La hipótesis de este trabajo de tesis doctoral es que es posible utilizar técnicas de IA, como son los SMA basados en OV, así como los sistemas de aprendizaje automático para elaborar un modelo que permita reducir el consumo energético y, que al mismo tiempo, permita asegurar el cumplimiento de los acuerdos SLA que se hayan alcanzado con los usuarios.

Hasta el momento, se ha presentado el SMA propuesto en el apartado 4.2, sin embargo, esta arquitectura basada en OV es tan sólo una parte de la solución que se propone en el marco de este trabajo de investigación. Esta arquitectura, por ser integradora, hace posible la inclusión de componentes individuales y autónomos (agentes) que incluyen diferentes modelos de razonamiento (agentes especializados) para la toma de decisiones en este entorno complejo y abierto que es un sistema CC. Por lo tanto, a lo largo de este apartado se presentará la segunda parte de la solución. Es decir, estos modelos de razonamiento avanzado y algoritmos que hacen posible la distribución de recursos computacionales dentro de este entorno complejo y abierto.

Los modelos de razonamiento integrados en los diferentes agentes especializados que forman parte de la arquitectura multiagente se describen partiendo de las dos posibilidades de redistribución de recursos que ofrece la tecnología actualmente (servicios e infraestructura) presentados en la introducción del apartado 4.3. Así, como se aprecia en la Figura 24, el algoritmo de distribución

sigue un proceso jerárquico de distribución de recursos. En primer lugar, la distribución de recursos se realiza a nivel de servicio, balanceando las diferentes peticiones entre los nodos que se asocian al servicio. Cuando este balanceo de peticiones no es suficiente, y por lo tanto los nodos que atienen el servicio no son capaces de hacer frente a la demanda de peticiones, se realiza una distribución de infraestructura desde una perspectiva micro; en la cuál, se solicita a los diferentes equipos físicos que alojan los nodos de un determinado servicio, que proporcionen más recursos computacionales a cada uno de estos nodos, lo que permite incrementar su rendimiento y, por ende, la cantidad y calidad de peticiones que puede atender por unidad de tiempo. Finalmente, en una tercera fase, cuando tampoco este segundo proceso de redistribución no ha sido suficiente para hacer frente a la demanda de un determinado servicio, se realiza una distribución de infraestructura a nivel macro. Este último paso, da lugar a la instanciación de un nuevo nodo que puede ser hospedado por una máquina física activa del entorno CC, o por una de las máquinas apagadas que permanecen en estado latente, esperando para ser reactivadas en el momento que se necesiten más recursos.

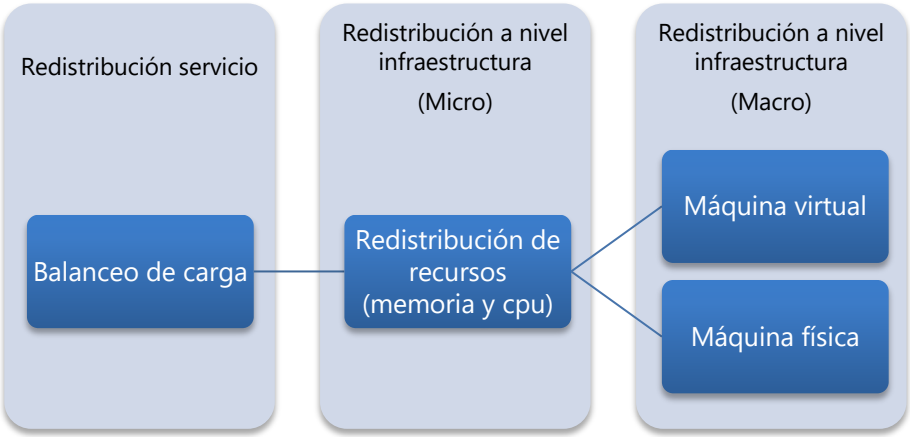

Figura 24.- Secuencia en la redistribución de recursos

En los siguientes apartados se presenta este modelo de distribución de recursos. En primer lugar, en el apartado 4.3.2.1 se presenta la distribución de recursos a nivel servicio. Posteriormente, la distribución de recursos a nivel infraestructura micro y macro se presenta en los apartados 4.3.2.2 y 4.3.2.3, respectivamente. Finalmente, el apartado 4.3.2.4 contiene la descripción del algoritmo que hace posible la compactación de los diferentes recursos computacionales para el ahorro de energía.

4.3.2.1.Redistribución a nivel servicio

Este tipo de balanceo es el más simple, pero su importancia es clave ya que es el encargado de asegurar que se están cumpliendo los acuerdos SLA relativos a servicios *software* que se hayan alcanzados con los usuarios de la plataforma CC. Dentro de esta redistribución intervienen dos de los agentes presentados en la etapa anterior, que son los agentes de *Service Monitor* y el agente *Service Supervisor*. Tal y como ya se indicó, existe una pareja de estos agentes por cada uno de los servicios y se encuentran situados físicamente en el nodo que realiza el balanceo de carga. A continuación, se describe la funcionalidad, tareas y modelo de razonamiento de cada uno de ellos:

- El agente *Service Monitor* es el encargado de controlar el estado del sistema de balanceo. Para ello, implementa algoritmos que permiten calcular de forma continuada la QoS de cada uno de los nodos que se está balanceando en cada momento
  El balanceo de recursos se realiza en base a los pesos de cada uno de los nodos, el objetivo es que todos los nodos tengan un nivel de QoS similar. De este modo, los usuarios perciben que el entorno atiende a sus peticiones de forma homogénea. Para ello, este rol calcula periódicamente el peso de cada nodo dentro del servicio siguiendo el siguiente algoritmo:

- Se mantiene un conjunto de grupos de enrutamiento, a cada uno de estos grupos se le asocian diferentes nodos trabajadores. Todos los nodos que pertenecen a un mismo grupo tienen el mismo peso dentro del sistema de balanceo y, por lo tanto, deben tener unos niveles de QoS similares.
- Para determinar los grupos de QoS se utiliza el algoritmo de agrupamiento del vecino más próximo, basando el cálculo de la distancia entre dos nodos a partir de la distancia euclídea entre los vectores de calidad del servicio en cada nodo $QoS_{n_i}^k = \left\{\overline{r_1^k} \cdots \overline{r_i^k} \cdots \overline{r_n^k}\right\}$. A partir de los valores obtenidos, se forman grupos a los cuáles se les asignan el mismo peso con el objetivo de que todos los grupos tengan una tasa de respuesta similar.
- Para asignar el peso a cada grupo, dado que no puede determinarse mediante un proceso exacto, se realiza siguiendo un proceso iterativo. Así, inicialmente todos los grupos tienen un mismo peso, el cuál se ajusta progresivamente en función de la demanda que sea capaz de admitir cada grupo. Así pues, los grupos con mejor QoS reciben un peso mayor, mientras que los grupos con menor QoS reciben un peso menor. Después de un periodo de estabilización del proceso de balanceo en el que se recalcula el QoS con el nuevo reparto de pesos, el proceso se repite desde el punto 2, volviendo a calcular los grupos afines teniendo en cuenta el nuevo reparto de pesos.

Finalmente, este agente de monitorización del servicio también será el encargado de almacenar periódicamente la calidad del servicio en el repositorio de histórico de la demanda del servicio. Así, para un tiempo t y un servicio k, se almacenará un vector $U_t^k = \{QoS^k, n, timestamp\}$ donde n es el número de nodos que están atendiendo el servicio en el momento $timestamp$.

- El agente *Service Supervisor* del servicio tiene un papel fundamental, ya que a partir de los datos generados por el agente *Service Monitor* debe determinar si es necesario asignar más o menos recursos al servicio.
  - Por un lado, tienen que decidir cuándo retirar un subconjunto de recursos asociados al servicio. En este proceso se elimina uno de los nodos que están asociados al servicio. Al retirar un nodo de ejecución, el resto de los nodos asociados al servicio tendrán que atender las peticiones del nodo que se elimina. Este incremento de carga en cada uno de los nodos lleva asociado un periodo de estabilización en el que el agente *Service Monitor* recalculará los nuevos parámetros de QoS para cada uno de los nodos que todavía atienden al servicio y, posteriormente, se ejecutará el *algoritmo de determinación de los grupos de enrutamiento* y los nodos asociados a estos grupos.
  - Por otro lado, asociar más recursos a un servicio se realiza cuando los nodos existentes con los recursos que tienen asignados, no son capaces de atender la demanda de peticiones según los acuerdos SLA establecidos con los usuarios, lo que conlleva que se violen los acuerdos que se han alcanzado con los usuarios de los servicios.

En ambos casos, el problema consiste en determinar cuándo es el momento adecuado para iniciar un proceso de asignación o retirada de recursos a un servicio:

- La retirada de recursos se realiza cuando el grupo de enrutamiento cuya calidad de servicio es más alta, tiene 2 o más nodos asignados, después de que se haya ejecutado el algoritmo determinación de los grupos de enrutamiento n. En este caso, se elimina uno de los nodos, debido a que sino se hiciera se estarían infrautilizando recursos activos. Este proceso de parada de nodos puede dar lugar a la dispersión de los nodos en varios servidores físicos del entorno CC. Este estado de dispersión se produce cuando existen varios nodos físicos con recursos libres, es decir, disponen de recursos que no están asociados a máquinas

virtuales. Para asegurar la eficiencia del sistema se realiza un proceso de compactación a través del algoritmo descrito en el apartado 4.3.2.4.

- La solicitud de nuevos recursos es realizada por el agente *Service Supervisor* cuando un incremento en la demanda de peticiones no hace posible que las peticiones se devuelvan en un tiempo adecuado según los acuerdos SLA acordados. Este proceso implica una redistribución de recursos a nivel infraestructura. Para ello se ejecuta una distribución de recursos a nivel micro en primer lugar, la cuál se describen en el apartado 4.3.2.2. Cuando el problema no se elimina en este primer paso, se procede a realizar una distribución de recursos a nivel macro, cuyo procedimiento se describen en el apartado 4.3.2.3.

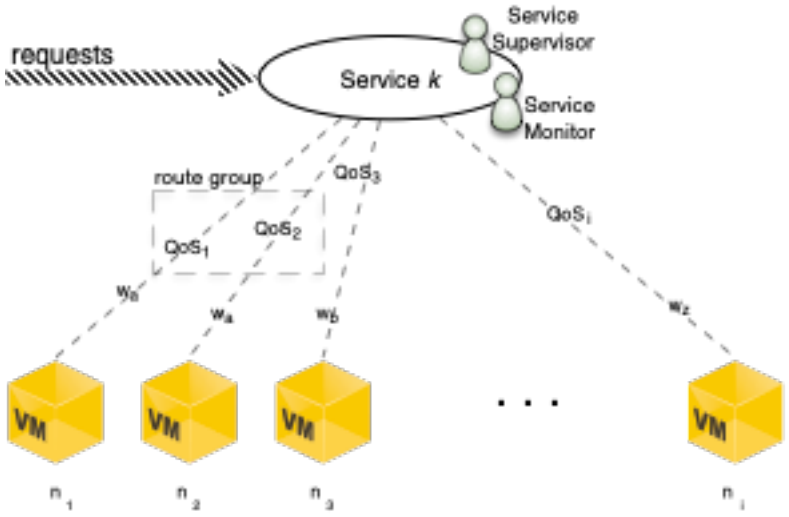

Figura 25.- Redistribución de recursos a nivel servicio

4.3.2.2. Redistribución a nivel infraestructura (micro)

La redistribución desde un punto de vista de infraestructura a nivel micro se asocia a la redistribución de recursos que se realiza dentro de una máquina física asignando diferentes recursos entre las diferentes máquinas virtuales que aloja y reservando suficientes recursos para el servidor principal. Este proceso está gobernado por dos agentes especializados de la arquitectura multiagente basada en OV, el agente *Local Monitor* y el agente *Local Manager*. Existirá una instancia de estos agentes en cada uno de los servicios físicos de la máquina. El procedimiento es el siguiente:

- El agente *Local Monitor* es el encargado de recuperar información sobre el estado de cada máquina virtual y del servidor donde están alojadas. Sin embargo, no podrá en ningún caso modificar las características de asignación.
- Dependiendo de la tecnología de virtualización subyacente (OpenVZ, KVM, Xen Server, etc.), este agente está formado por un único componente que se aloja en la máquina principal o tendrá subcomponentes en cada una de las máquinas virtuales que hospeda el servicio físico las cuáles periódicamente envían al agente *Local Monitor* el estado de cada máquina virtual.

  En definitiva, el agente monitor mantendrá actualizada una matriz con la siguiente información:

$$exec^{e_1} = \begin{Bmatrix} VM_1 \\ \vdots \\ VM_m \\ PR^{e_1} \end{Bmatrix} donde \begin{cases} PR^j = \{hostname, IP, mac, state, M_{max}, vcpu_{max}, M_{min}, vcpu_{min}\} \\ VM_i = \{IP, state, M, vcpu, M_i\, \overline{p_{cpu}}\} \end{cases}$$

- El agente *Local Manager* es el que tiene el poder para gobernar la asignación de recursos dentro de la máquina incluyendo las siguientes funcionalidades: (i) actualizar recursos

asignados a cada una de las máquinas virtuales, (ii) instanciar una nueva máquina virtual a partir de una plantilla y, finalmente, (iii) detener la ejecución de una máquina virtual. A partir de estas tres funcionalidades es posible redistribuir los recursos dentro de un servidor físico. Los recursos que se pueden redistribuir son el número de procesadores virtuales (*vcpu*) y la memoria física (M).

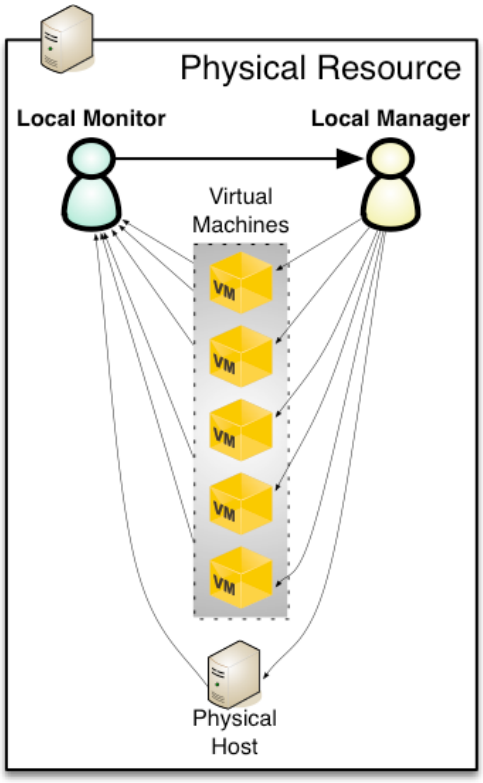

Figura 26.- Redistribución de recursos de infraestructura a nivel micro

El proceso de distribución de recursos tiene en cuenta el tipo de servicio al que pertenecen las máquinas virtuales que existen en ejecución, cuya información se recoge de la plantilla que permite su instanciación, la cuál también determina los recursos mínimos que se le deben asignar. El proceso de reasignación a nivel interno lo inicia un agente de tipo *Service Supervisor* cuando detecta que los recursos mínimos que se le han asignado a su subconjunto de máquinas virtuales no son suficientes. Este agente envía un mensaje, indicando la necesidad de recursos, a cada uno de los agentes *Local Manager* de las máquinas físicas que hospedan los nodos para ese determinado servicio. Cada una de estas máquinas inician un proceso reasignación de recursos en paralelo, que da como resultado un incremento global de los recursos asignados a cada servicio.

Cuando se inicia un proceso de distribución de recursos, el agente *Local Manager* de la máquina configura una estructura de información de asignación de recursos como la que se muestra en la Tabla 8, a partir de la información suministrada por el agente *Local Monitor* de la misma máquina. Donde se recoge información instantánea I_{ei}^{t} del estado completo de la asignación de recursos en un tiempo t, el grado de uso actual de éstos, el servicio y la prioridad que tiene el nodo en el sistema de balanceo del servicio según el cálculo realizado por el agente *Service Monitor* asociado. Estos tres últimos parámetros resultan de importancia debido a:

- ID_s, asociado al identificador del servicio que permite determinar si es una máquina virtual correspondiente a un nodo de servicio de tipo infraestructura o *software*. En el caso de los servicios de infraestructura, el usuario establece un acuerdo SLA para unas determinas

condiciones *hardware* que no pueden modificarse. En el caso de corresponderse con un servicio de *software*, donde la máquina virtual es un nodo asociado a un servicio, sí que se podrían modificar sus propiedades.

- *Type*, que determina si son servicios de tipo *stateless* o *statefull*, recordar que como ya se indicó en el apartado 4.1.3 las aplicaciones *stateless* pueden tener varias instancias en ejecución de forma simultánea, ya que cada una de ellas no almacena el estado del servicio. Mientras que las máquinas virtuales *statefull* únicamente tienen una instancia en ejecución. Por lo tanto, si este servicio es el que necesita más recursos es prioritario respecto al resto.
- *Priority*, que determina la prioridad de la máquina virtual en el proceso de balanceo de carga. La cuál da una medida acerca del grado de rendimiento de la máquina virtual con respecto al resto de nodos del mismo servicio. Se corresponde con un valor dentro del rango 0 a 1 y se calcula mediante la expresión simple $Priority_i = {}^{w_i}/_{w_{max}}$. Este valor es calculado por el agente *Local Manager*, mediante la interrogación al agente *Service Monitor* del servicio
  Así pues, los valores próximos a 0 tienen una prioridad baja y, por lo tanto, no están ofreciendo la QoS necesaria, y los valores próximos a 1, tienen un nivel de QoS alta en comparación con otras instancias que atienden el servicio.

	processing			**Memory**				**General**		
	$vcpu_{max}$	**$vcpu_{min}$**	**$vcpu_{used}$**	**M_{max}**	**M_{min}**	**M_{used}**	**$M_{asigned}$**	**ID_s**	**type**	**priority**
$\boldsymbol{VM_1}$										
$\boldsymbol{VM_2}$										
$\boldsymbol{VM_m}$										
$\boldsymbol{PR^{e_1}}$										

Tabla 8.- Instancia de servidor físico en la redistribución de recursos de infraestructura a nivel micro

Antes de describir el procedimiento de redistribución de recursos que realiza el agente especialido *Local Manager*. En primer lugar, destacar que éste sólo afecta a las máquinas que atiendan servicios *software*. Aquellas máquinas con características estáticas no se tienen en cuenta dentro del procedimiento. Así, de las n máquinas que hospeda un servidor físico, sólo computan en los algoritmos que se presentan a continuación las m máquinas que atienden servicios *software* (*stateless* o *statefull*) cuyas características pueden ser dinámicas.

El proceso de distribución de recursos es diferente para la asignación de *vcpus* virtuales y de memoria. En primer lugar, para la asignación de *vcpus* se utiliza la máquina de estados que se presenta en la Figura 27. El objetivo de este procedimiento es no disponer de recursos de ejecución infrautilizados y, si éstos existiesen, tan sólo podrían estar asignados a la máquina física. Así el procedimiento de asignación de *vcpus* se inicia verificando si la máquina física tiene *vcpus* que no estén asignadas, es decir:

$$PR^{e_1}(vcpu_{max}) \neq \sum_1^m VM_i\left(vcpu_{used)}\right) + PR^{e_1}(vcpu_{used})$$

Por lo tanto, puedan asignarse a la máquina o máquinas que necesita recursos. Si no existe ninguna *vcpu* que no esté asignada, se itera sobre cada máquina virtual para comprobar si la cantidad de procesamiento asignado está siendo infrautilizado. Si en algún caso, el valor del parámetro $vcpu_{used}$ es menor que una constante definida por el administrador, entonces se reasigna uno de los núcleos de esta instancia a aquella máquina virtual que necesite un mayor número de recursos.

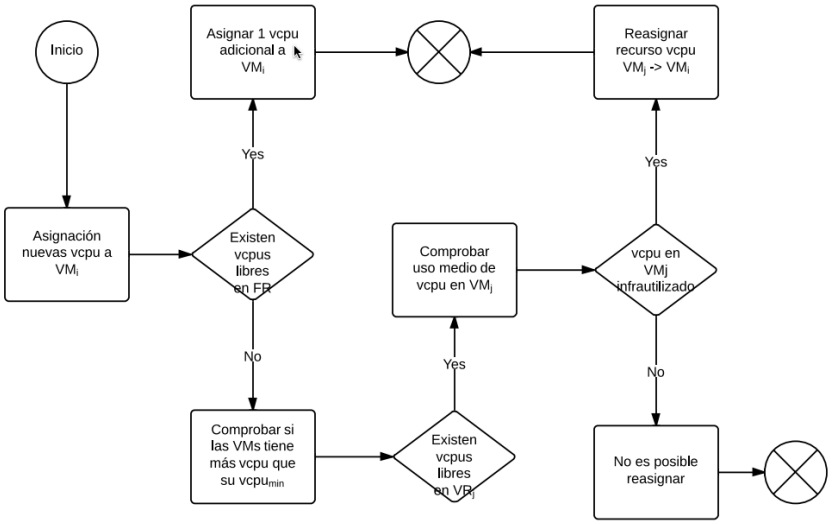

Figura 27.- Reasignación de infraestructura (*vcpu*) a nivel micro

Por otro lado, el proceso de reasignación de memoria es más complejo y para ello se sigue un modelo de programación lineal que determina la asignación de recursos óptima. No obstante, el paso inicial que se realiza es similar al proceso de asignación de unidades de procesamiento virtual (*vcpus*). En este sentido, si el servidor físico tiene memoria libre que puede asignarse, se asigna directamente a la máquina virtual demandante, cumpliendo la restricción de memoria máxima asignable a la plantilla

$$VM_i(M_{asigned}) \leq VM_i(M_{max}).$$

Este proceso de asignación directa se realiza en base a la siguiente expresión donde $VM_i(M'_{asigned})$ se corresponde a la memoria disponible previamente en la máquina virtual y se basa en la prioridad que tiene la máquina virtual en el balanceo de carga. Se ha modelado este procedimiento de este modo para no modificar en gran medida este reparto de prioridades en el balanceo previamente asignado y, además, no asignar todos los recursos libres a una única instancia virtual.

$$\begin{cases} VM_i(M_{asigned}) = VM_i(M'_{asigned}) + VM_i(M_{max}) * priority_i \;\; si \;\; (VM_i(M_{max}) * priority_i) < \; PR^{e_i}(M_i) \\ VM_i(M_{asigned}) = VM_i(M'_{asigned}) + PR^{e_i}(M_i) \;\; si \;\; (VM_i(M_{max}) * priority_i) > \; PR^{e_i}(M_i) \end{cases}$$

En caso de no existir memoria libre, será necesario reasignar memoria, es decir, retirar memoria de un conjunto de instancias de ejecución y asignársela a la máquina demandante. A nivel técnico este proceso es posible siempre y cuando no se elimine memoria que esté siendo utilizada por las instancias en ejecución. Es decir, es posible redistribuir la memoria que está siendo infrautilizada siempre dejando un margen de seguridad (M_l) al retirar memoria de un nodo activo.

Así pues, la memoria infrautilizada de cada máquina virtual es la siguiente:

$$VM(M_i) = VM(M_{asigned}) - VM(M_{used})$$

Y, por lo tanto, la memoria infrautilizada de todo el servidor físico vendrá dada por la siguiente expresión:

$$PR'^{e_1}(M_i) = \sum_1^m VM_i(M_{assigned}) - VM_i(M_{used})$$

El problema consiste en cómo redistribuir la memoria infrautilizada entre las diferentes máquinas virtuales y el servidor físico, de forma que se reduzca la cantidad de memoria infrautilizada. El problema es de optimización y se resuelve utilizando programación lineal, de forma que se tienen m variables independientes que se corresponden con la nueva asignación de memoria. De forma que la función que hay que minimizar está compuesta por el sumatorio de la memoria infrautilizada que se asignará a cada máquina virtual por el porcentaje de infrautilización actual, lo que permitirá ponderar cuál es la ganancia en la asignación de memoria en cada caso.

$$min\ PR^{e_1}(M_i) = \sum_1^m x_i \left({VM_i(M_i)} \Big/ {PR^{e_1}(M_i)} \right)$$

Este modelo de minimización tiene las dos restricciones siguientes:

- La suma de las variables objetivo será igual a la suma de la memoria infrautilizada inicial.

$$\sum_1^m x_i \leq PR'^{e_1}(M_i)$$

- También será necesario que cada variable x_i sea mayor que el margen de seguridad.

$$x_i, \dots, x_m > M_l$$

Una vez resuelto el problema obtendremos que el valor de memoria asignada para cada máquina virtual está determinada por la siguiente expresión:

$$VM_i\left(M_{asigned}\right) = VM_i(M'_{used}) + x_i$$

Una vez que se han determinado el nuevo reparto de memoria, si los resultados obtenidos en el proceso de cálculo de la nueva asignación son superiores a los recursos que tuviera la máquina del servicio que inicia el cambio, entonces esta nueva asignación se aplica por el agente *Local Manager* y se informa al agente *Service Supervisor* asociado a el servicio que inició el proceso que se han asignado recursos al nodo.

Finalmente, es importante destacar que este proceso puede ser iniciado por varios servicios a la vez, y en este caso es necesario tener en cuenta las siguientes consideraciones para el agente *Local Manager*:

- El proceso de cálculo de nuevos recursos no depende del tipo de servicio, sino de la prioridad de las máquinas dentro de cada servicio y de los recursos que esté utilizando en un determinado momento.
- Por lo tanto, si dos o más servicios solicitan un incremento de recursos a nivel local sólo se aplicarán los cambios propuestos si:
    - Si, la maquina tiene recursos libres que pueda asignar a los recursos.
    - Si la máquina no tiene recursos libres, sólo se realizará el cambio de recursos si para todos los servicios que hayan solicitado un incremento de los recursos, el cálculo de la nueva asignación incrementa los recursos que tiene asignados en un determinado momento cada uno de los nodos asociados a los servicios que solicitan más recursos.
    - Si la máquina no tiene recursos libres, y en el paso dos no se ha encontrado una solución adecuada para todos los demandantes de recursos, entonces se procederá

a informar al agente que inicia el proceso, es decir, al agente *Service Supervisor*, que no hay recursos suficientes para que tome las medidas que estime oportunas.

En este sentido, la solución es aplicar un procedimiento de redistribución de infraestructura a nivel macro, el cuál se describe en el apartado 4.3.2.3.

- En cualquier caso, una vez que se haya aplicado un cambio en los recursos, no se podrá realizar un nuevo reajuste de servicios hasta que se estabilice cada una de las máquinas virtuales y servicios atendiendo a los recursos asignados. El tiempo de espera entre dos reasignaciones será una constante supervisada por el administrador del entorno CC.

4.3.2.3. Redistribución a nivel infraestructura (macro)

La redistribución de recursos a nivel macro se inicia por un agente de tipo *Service Supervisor* cuando detecta que los recursos asignados al servicio al que se asocia no son suficientes. Solo se podrá iniciar una redistribución a nivel macro, cuando previamente se haya iniciado una redistribución a nivel micro que como resultado de la misma, no se hayan satisfecho las necesidades computacionales del servicio.

La redistribución de recursos a nivel macro se realiza por los agentes *Global Manager* los cuáles se encuentran situados en cada una de las máquinas físicas y tiene acceso a la información que proporciona el agente de *Local Monitor* en la misma máquina en la que se sitúan. Así mismo, también tiene autoridad sobre el agente *Local Manager* pudiendo indicarle la necesidad de arrancar una nueva máquina de un determinado servicio con unas características determinadas. No obstante, el agente *Global Manager* no podrá realizar ninguna acción si el agente *Local Manager* está realizando una adaptación a nivel local.

Este agente *Global Manager* es un agente altamente especializado que implementa una arquitectura deliberativa de tipo CBR-BDI [307, 393-396]. Antes de detallar el procedimiento, se hará una breve revisión del modelo de razonamiento basado en caso (CBR, *Case-based Reasoning*). Este modelo utiliza la base de razonamiento del pensamiento humano, en el que se recurre a experiencias pasadas para resolver nuevos problemas [397]. Así pues, si en un tiempo pasado se decidió resolver un problema utilizando una determinada solución y, una vez aplicada esa solución se obtuvo un determinado resultado, entonces parece lógico que si se presenta un nuevo problema con características similares al que se resolvió previamente en el pasado, se recurra a esta experiencia adquirida para dar solución el nuevo problema. Por tanto, el modelo se basa en la idea de que problemas similares tienen soluciones similares. Sin embargo, carecer de problemas similares no supone que el sistema no sea capaz de proponer buenos resultados, sino que la reutilización de memorias pasadas se convierte entonces en un proceso creativo. Sea cual sea el resultado de este proceso creativo, el individuo aprende una nueva experiencia, ya sea positiva o negativa.

En este modelo de razonamiento, el concepto de caso es fundamental. Así pues, un caso es un fragmento de conocimiento contextualizado que representa una experiencia [398] que no es más que una terna que incluye el modelo de problema, el modelo de la solución y el resultado de su aplicación.

```
Case: <Problem, Solution, Result>
Problem: initial_state
Solution: sequence of <action, [intermediate_state]>
Result: final_state
```

Figura 28 - Modelo de caso en un sistema de razonamiento CBR

El sistema debe estar formado por dos componentes fundamentales: la memoria de casos y el mecanismo de razonamiento. En primer lugar, la memoria de casos no es más que un sistema de

almacenamiento dónde se extraen las soluciones anteriores y en ella se almacena lo aprendido, por ello es la encargada de mantener la representación y organización de los casos. Por su parte, el ciclo de un sistema CBR está formado por cuatro procesos secuenciales [46, 47] tal y como sigue:

- **Etapa de recuperación.** Es la primera etapa que realiza un sistema CBR. En ella se realiza la recuperación de casos, esto es, el acceso a los casos almacenados que cuentan con una descripción de problema más similar a la del problema actual. Para ello, lo habitual es utilizar un algoritmo que garantice el acceso rápido y eficiente, aplicando una métrica de similitud que determina cuál o cuáles de ellos son los mejores casos [399].
- **Etapa reutilización.** A partir de los casos más similares, la reutilización o adaptación consiste en trabajar con las soluciones correspondientes a los casos más similares recuperados en la etapa anterior para poder obtener una solución al problema actual. Trabajar con las soluciones significa modificarlas y combinarlas, o simplemente decidir cuál de ellas es la óptima y, por lo tanto, reutilizarla.
- **Etapa de revisión.** Fase en la que se comprueba la bondad de la solución finalmente aplicada para resolver el problema actual [400]. Se comprueba si la solución propuesta en la etapa anterior es apropiada para el caso actual. Para ello se utiliza un sistema experto de conocimiento, o bien una persona experta. En ocasiones, en esta etapa se puede realizar una reparación de los fallos o errores detectados.
- **Etapa de retención y aprendizaje.** En esta última etapa se aprende a partir de la nueva experiencia adquirida. Para ello se almacena el caso actual y la solución aplicada para resolverlo. Además, se tiene en cuenta el resultado obtenido en la etapa de revisión para asignar una eficiencia al caso. De esta forma el caso puede ser indexado en la memoria de casos.

Como ya se ha indicado el proceso de distribución de infraestructura a nivel macro es iniciado por el agente *Service Supervisor* asociado al servicio con dificultades para atender la demanda. Este agente alerta a todas las máquinas físicas que hospedan los nodos del servicio que es necesario una redistribución de recursos a nivel global para hacer frente a la demanda de peticiones. El mensaje es recibido por el agente *Global Manager* de cada servidor físico que a su vez alerta el resto de los agentes *Global manager* del entorno CC. Existe un agente *Global Manager* por cada máquina física activa existente en el entorno CC. Estos agentes, solicitarán al agente *Local Monitor* de la máquina una instantánea del estado del uso de los recursos (I_{ei}^{t}). A partir de esta instantánea, este agente especializado evalúa la cantidad de recursos disponibles, es decir, aquellos que no están asignados a ninguna máquina virtual. Si esta cantidad de recursos fueran mayores que los mínimos indispensables para instanciar un nodo asociado al servicio (información que se especifica en la plantilla del servicio), entonces iniciaría el proceso de razonamiento que se explica a continuación para determinar la cantidad de recursos que se pueden reservar al nuevo nodo de ejecución que demanda el agente *Service Supervisor* del servicio. Si no tuviera recursos libres o estos fueran menores que los demandados como mínimo por el servicio, entonces el agente *Global Manager* determina que el nodo físico no formará parte del proceso de asignación global. En el caso de que ninguna máquina atienda a la necesidad de incremento de servicio, entonces se solicitará al agente *Hardware Manager* el arranque de un nuevo servidor físico donde se instanciará una máquina virtual en base a las características mínimas definidas en la plantilla del servicio.

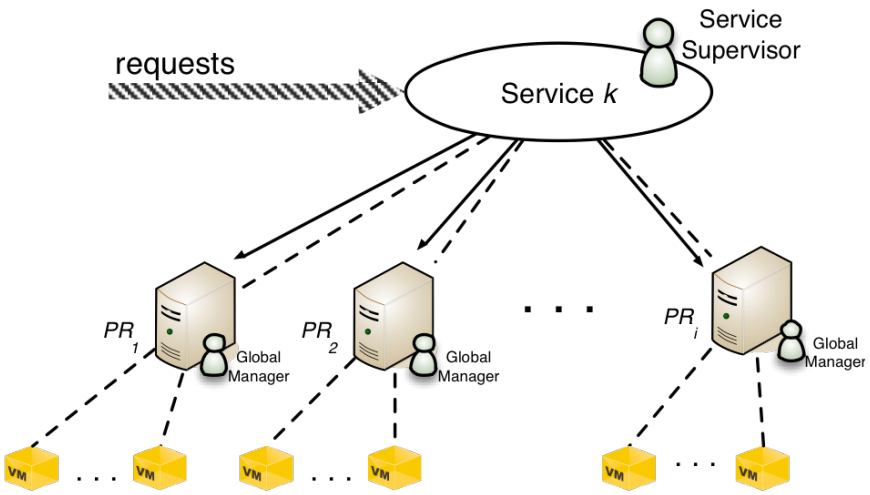

Figura 29.- Inicio distribución de infraestructura a nivel macro

Cada agente de tipo *Global Manager* que atiende la petición de recursos implementa un modelo de razonamiento avanzado basado en un sistema CBR-BDI. Este conjunto de agentes especializados, como están situados en las diferentes máquinas físicas del sistema, ejecutan de forma paralela el algoritmo de razonamiento, que se explica a continuación, determinando de forma individual una solución al problema de demanda del servicio. Estas soluciones individuales se envían al agente *Service Supervisor* que ha iniciado la demanda de recursos, para que seleccione aquella que incluye más recursos para la máquina virtual que se instanciará, priorizando la memoria en detrimento de la asignación propuesta de unidades de procesamiento virtuales (*vcpu*).

El proceso de razonamiento en cada nodo físico, desarrollado por el agente especializado *Global Manager* está basado en un sistema CBR como ya se ha mencionado. Este modelo se basa en la experiencia adquirida en el caso mediante el almacenamiento de casos similares. La memoria de casos es central a todo el sistema CC, de forma que el conocimiento global del sistema pueda ser compartido por cada uno de sus miembros, es decir, el agente *Global Manager*. Dado que esta memoria puede crecer exponencialmente como estrategia de mantenimiento se utiliza una base de datos sin esquema de alta velocidad que permite un acceso rápido a los datos almacenados.

En primer lugar, resulta necesario definir el concepto de caso dentro del modelo de razonamiento $C = \{P, S(P), E\}$ donde:

- P se corresponde con la descripción del problema, esta descripción tiene una representación matricial asociada a la instantánea de uso de recursos I_{ei}^{t}, junto con la descripción de la plantilla de servicio que se pretende instanciar y la marca temporal que denota el instante en el que se detecta el problema.

$$P = \{I_{ei}^{t} \quad VM_{t}^{k} \quad timestamp\}$$

- $S(P)$ se asocia a la solución del problema P, que vendrá determinada por una instantánea similar a la del problema donde se incluyan los recursos asociados

$$S(P) = \{M, vcpu\}$$

- Finalmente, la eficiencia (E) se mide desde ambas perspectivas (micro y macro) de la distribución de infraestructura:
  - Por una lado, la eficiencia a nivel micro (E_m) se asocia al grado de eficiencia de la solución propuesta dentro del servidor físico donde se instancia la máquina virtual. Este

grado de eficiencia es planteada por el agente *Local Monitor* en función el porcentaje de uso del procesador y de la memoria asignada. Ambos porcentajes están entre 0 y 1, de forma que la solución es más eficiente a medida que los resultados se acerquen a 1.

$$E_m = \left\{\overline{p_{cpu}}, {M_{used}}/{M_{assigned}}\right\}$$

- Por otro lado, la eficiencia a nivel macro (E_M) se asocia al grado de eficiencia desde el punto de vista del servicio, determinando si una vez que se ha aplicado la solución propuesta ha sido necesario iniciar un proceso de distribución de recursos de infraestructura a nivel macro. En este sentido, este grado de eficiencia mide el número de nodos adicionales que ha necesitado el servicio.

$$E_M = \{\mathrm{n}\}$$

Por lo tanto, la eficiencia viene dada según la expresión $E = \{E_m \quad E_M\}$

El proceso CBR se inicia recuperando los casos anteriores que son similares de la memoria, para ello se seleccionan aquellos casos más similares atendiendo a las siguientes reglas:

- Se seleccionan los casos de las máquinas físicas con características similares, con un grado de eficiencia superior al 90%. Las máquinas físicas similares se evalúan en función del parámetro $benchmark$ que caracteriza cada máquina física. Se recuerda, que este parámetro se almacena en el vector que identifica a cada una de las máquinas físicas (PR) que se detalló en el apartado 4.1.2.1.
- A partir de este subconjunto de casos recuperados se configura un vector por cada caso que incluye el número de máquinas virtuales existentes en el caso y los recursos libres disponibles $C_i = \{n, M, vcpu\}$. A partir de este vector se evalúa la similitud con respecto al vector que representa el caso actual utilizando para ello la distancia euclídea, lo que permite seleccionar los casos más similares.
- Posteriormente, de este conjunto de casos se seleccionan aquellos en los que en las experiencias pasadas estuvieran implicado el servicio que está demandando recursos en este caso durante un periodo de tiempo similar al que se produce el caso, utilizando para ello un patrón de uso de la web en una semana [401].

A partir de los casos recuperados, durante la fase de reutilización se elabora una solución al problema:

- Si la base de casos no tuviera ningún caso previo similar, entonces la solución al problema se asocia a los recursos mínimos determinados en la plantilla de instanciación del servicio:

$$S(P) = \{M_{min}, vcpu_{min}\}$$

- Si en cambio, si se recuperan casos similares, la solución al problema sería la del caso más cercano multiplicado por la eficiencia del caso:

$$S(P) = \{\mathrm{M}' * (E_m(1) + E_M(1)), vcpu'(E_m(2))\}$$

- Si los valores de asignación de la solución anterior son mayores que los valores asumibles por el equipo, debido a que no tienen tantos recursos disponibles. El resultado del caso serán los máximos recursos disponibles por la máquina.

$$S(P) = \{PR(M_{\max}), PR(vcpu_{min})\}$$

Una vez que se ha calculado la solución al caso, este es enviado por cada uno de los agentes especializados *Global Manager* que se hospedan en servidores físicos con recursos suficientes al agente *Service Supervisor*. Este agente selecciona reactivamente el nodo que ofrezca más recursos a la nueva máquina virtual. Posteriormente durante la etapa de revisión una vez instanciado el nuevo nodo asociado se evaluará su uso desde una perspectiva micro y macro obteniendo el valor de la eficiencia de la solución. Finalmente, durante la última etapa del ciclo CBR propuesto, se guarda el caso junto con su eficiencia para que pueda ser utilizado en posteriores readaptaciones.

El modelo de adaptación propuesto es distribuido lo que permite mejorar la alta disponibilidad del sistema, ya que la toma de decisiones se realiza a lo largo de todo el sistema CC. Además, este modelo permite distribuir la potencia de cálculo que requiere la obtención de la solución, y por tanto, reduce el impacto que la búsqueda de una solución pueda tener en el entorno CC en su conjunto.

4.3.2.4. Redistribución a nivel infraestructura (Compactación)

A lo largo del modelo de distribución de recursos propuestos no se ha mencionado el uso de una de las características claves de la tecnología de virtualización, que es la migración de máquinas entre servidores físicos. En el proceso de incremento de recursos no se ha utilizado esta característica, ya que la práctica demuestra que exige un coste adicional de recursos y ancho de banda que afectan negativamente al servicio y a los equipos físicos que alojan la máquina a migrar (emisor y receptor).

No obstante, esta característica resulta muy efectiva a la hora de compactar las máquinas virtuales en el menor número de servidores posibles, de forma que se pueda apagar o hibernar el conjunto de servidores físicos que no tengan asignada ninguna máquina con lo que se consigue un incremento de la eficiencia energética. Según el modelo de distribución que presenta el sistema, el procedimiento es simple.

El problema se origina cuando las máquinas tienen una gran dispersión, debido a que cuando un agente de tipo *Service Supervisor* asociado a un determinado servicio, detecta que tiene varios nodos que están en el nivel más alto de prioridad (tienen una alta calidad de resultados), solicita la eliminación de uno de ellos de forma aleatoria. Este procedimiento puede derivar en varios servidores físicos con un número muy bajo de máquinas virtuales y por lo tanto muchos recursos no asignados. Sin embargo, el coste energético asociado a estos equipos es similar a que si alojaran una gran cantidad de nodos virtuales. En este caso, se dice que el sistema está *disperso*.

El procedimiento de compactación es simple y se realiza por el agente especializado *Global Manager* tal y como sigue:

- Aquellos servidores virtuales con un número de máquinas virtuales muy bajo (el número es determinado de forma supervisada por el administrador del sistema), solicita al resto de servidores que acojan sus máquinas virtuales para que de este modo pueda pasar a un estado de hibernación que no consuma recursos.
- Los nodos con recursos disponibles al recibir la petición evalúan la instantánea del equipo (I_{ei}^{t}) proporcionada por al agente Local Monitor del equipo, para determinar si pueden hospedar una máquina con unas características como la que se pretende migrar.
- En el caso de que existan recursos disponibles se envía la configuración al *Global Manager* solicitante y se inicia el proceso de migración. Si este agente recibiera varias confirmaciones simultáneas iniciaría el proceso de migración a una de las máquinas de forma aleatoria ya que desconoce el conjunto de detalles internos de la máquina que acogerá el nuevo nodo virtual.

Gracias a este procedimiento simple, es posible compactar el conjunto de máquinas virtuales sin alterar la calidad de los servicios, ya que no se modifican los recursos individuales de cada máquina. Con lo que el sistema pasaría a estar en estado *compacto*.

4.4.Conclusiones preliminares

A lo largo de este capítulo se ha detallado el modelo arquitectónico propuesto en este trabajo de investigación. En primer lugar, se destaca que el modelo se sustenta sobre la caracterización del entorno presentada en el apartado 4.1 y que permite formalizar un sistema CC de propósito general. A partir de ahí, se propone un modelo arquitectónico, denominado +Cloud, para el control y monitorización de una plataforma CC mediante el uso de OV de agentes inteligentes. Gracias a este modelo es posible optimizar la eficiencia del sistema, permitiendo que los agentes alcancen sus objetivos, adaptándose de una manera autónoma pero coordinada a los cambios que se producen en el entorno.

El capítulo finaliza, precisamente, con la descripción de los algoritmos avanzados que se integran en los agentes especializados de la OV. Este modelo de distribución basado en algoritmos sociales hace posible la distribución de responsabilidades entre las diferentes individuos y organizaciones que forman parte de +Cloud, lo que por tanto permite la adaptación dinámica de la sociedad en función de los cambios que se produzcan en el entorno, es decir, teniendo en cuenta la demanda de los servicios ofertados en tiempo de ejecución.

Cabe destacar que es un modelo único ya que incorpora un mecanismo propio de reorganización y adaptación social. Los SMA siguen un paradigma organizacional y el modelo propuesto añade las funcionalidades de reorganización y adaptación necesarios en este contexto de aplicación.

Capítulo 5. Caso de Estudio y resultados experimentales

En el capítulo anterior se ha presentado un novedoso enfoque para distribuir recursos computacionales de forma adaptativa en entornos CC. El modelo dinámico que se propone hace uso de una arquitectura basada en OV de agentes inteligentes, que permite la integración de algoritmos distribuidos que hacen uso de técnicas de optimización y de razonamiento basado en casos para el reparto de recursos computacionales entre los diferentes servicios de una plataforma CC, construyendo de este modo un modelo adaptativo con la capacidad de aprender a partir de las experiencias pasadas.

La evaluación del modelo de distribución dinámica que se propone es una tarea compleja, que requiere disponer de un entorno *hardware* y *software*. Para llevar a cabo una evaluación de estas características en el estado del arte existen simuladores de entorno CC que pueden ser utilizados como marco de evaluación (Cloudsim[14] [402], CloudAnalyst[15] [403], MDCSim[16] [404], GreenCloud [405], CDOSim [406], etc.).

Sin embargo, en el marco de este trabajo la evaluación se ha realizado en una plataforma diseñada y desarrollada en paralelo al modelo CC propuesto en el capítulo anterior. Esta plataforma se ha desarrollado, en el marco de este trabajo de tesis doctoral, por el grupo de investigación BISITE e integra el SMA +Cloud. La plataforma expone diferentes servicios computacionales, a nivel *hardware* y *software*. Desde su nacimiento ha sido concebida para integrar el modelo de monitorización y control detallado en el capítulo anterior. Por tanto, las funcionalidades de reorganización y adaptación de los agentes en su comportamiento son necesarias para el correcto funcionamiento de la plataforma.

Dada la experiencia acumulada por el grupo en esta plataforma, así como los recursos necesarios para la implantación del sistema +Cloud en otras plataformas alternativas, se ha optado por la utilización de esta plataforma para la evaluación de la propuesta, dejando como trabajo futuro la comparación de los resultados obtenidos en dicha plataforma con los que se puedan obtener en otras plataformas existentes. Además, el uso de este entorno de evaluación permite una aproximación más realista de las tareas de evaluación, lo que ha permitido realizar una evaluación precisa y realista tanto de la arquitectura multiagente propuesta, como de los algoritmos de distribución de recursos computacionales.

Este capítulo se organiza como sigue: en el apartado siguiente (5.1) se presenta brevemente la plataforma CC que se ha utilizado para la evaluación del modelo propuesto. Posteriormente, en el apartado 5.2 se presenta la evaluación empírica del modelo de adaptación propuesto mediante la realización de diferentes pruebas de rendimiento. Finalmente, en el último apartado (5.3) se

[14] A Framework For Modeling And Simulation Of Cloud Computing Infrastructures And Services. http://www.cloudbus.org/cloudsim/
[15] A cloudsim-based visual modeller for analysing cloud computing environments and applications
[16] A Multi-tier Data Center Simulation Platform

presenta una discusión comparativa entre el sistema propuesto y los resultados obtenidos frente al estado del arte.

5.1.El entorno de evaluación, la plataforma +Cloud

Este apartado tiene como objetivo describir brevemente los detalles de la plataforma CC que integra el SMA denominado +Cloud descrito en el Capítulo 4. Esta plataforma ha sido el entorno de pruebas que ha permitido evaluar tanto la arquitectura multiagente, como el modelo de razonamiento dinámico y adaptativo para el reparto de recursos computacionales que se ha propuesto en el marco de este trabajo de esta investigación.

La plataforma CC con la que se integra el SMA +Cloud a nivel *externo*, aúna un conjunto de servicios que hacen posible el desarrollo y despliegue de aplicaciones web. Para ello dispone de un conjunto de servicios PaaS que hacen posible la persistencia de los datos que manejan las aplicaciones desarrolladas por terceros. Estas aplicaciones pueden estar desplegadas en la propia plataforma CC haciendo uso del servicio de despliegue a nivel IaaS, que proporciona servidores privados virtuales a petición de los desarrolladores/gestores, es decir, el agente *Cloud User*. Finalmente, la plataforma también proporciona un punto de acceso unificado y aplicaciones nativas de tipo *software* que se enmarcan en el nivel SaaS del paradigma tecnológico CC.

Por otro lado, a nivel *interno*, la plataforma está gobernada por el SMA +Cloud que se propone en el marco de este trabajo de tesis doctoral. Este SMA tiene fundamentalmente dos ventajas respecto a otros modelos existentes. En primer lugar, dado que se trata de un SMA abierto, permite que agentes externos (tanto humanos, como artificiales) puedan acceder a la organización jugando un rol concreto, de forma que puedan proporcionar servicios y funcionalidades dentro de la propia organización. Y, en segundo lugar, +Cloud es un sistema integrador que hace posible la aplicación de algoritmos avanzados para la distribución de los recursos entre los servicios de la capa superior e interactuar con el entorno, es decir, los sistemas de virtualización, balanceo de carga y persistencia.

A continuación, en el subapartado 5.1.1 se revisará brevemente los servicios externos que ofrece la plataforma (SaaS, PaaS e IaaS). Posteriormente, en el subapartado 5.1.2 se presentará brevemente el contexto tecnológico de la plataforma y la interacción con el SMA. El diseño interno de la misma ya se ha presentado, en detalle, en el Capítulo 4 del presente documento.

5.1.1.Servicios externos

La plataforma CC diseñada propone un modelo de computación que abarca los tres niveles de servicios (capacidades) que propone el NIST en su definición (*software*, plataforma e infraestructura) [7]. En la Figura 30 se presenta una vista global de los servicios ofertados, los cuáles se presentarán en los siguientes subapartados.

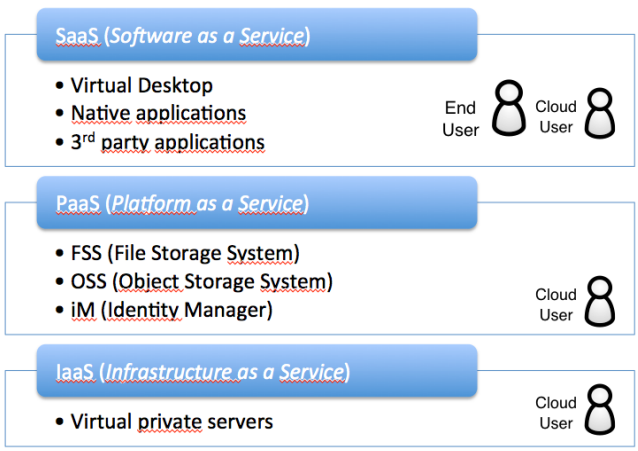

Figura 30.- Servicios externos de la plataforma +Cloud

5.1.1.1.El nivel SaaS

Las capacidades que proporciona la plataforma a nivel *software* están formadas por las aplicaciones web a las que los usuarios finales tienen acceso. Este nivel se ha diseñado para utilizar dos tipos de aplicaciones *software*, las aplicaciones nativas de la plataforma y las aplicaciones desplegadas por terceros.

En primer lugar, se presentan las aplicaciones web nativas que proporciona la plataforma para la administración del propio entorno CC, así como para facilitar el trabajo de los usuarios que hacen uso de la plataforma (*Cloud User*, *End User* y *Administrator User*). En el desarrollo de estas aplicaciones se han empleado las últimas tecnologías en el marco del desarrollo web. De este modo es posible ofrecer al usuario un entorno de uso avanzado. Así, se han utilizado las últimas novedades ofrecidas por HTML5[17] y CSS3[18], además del novedoso protocolo de comunicación WebSockets[19] que hace posible actualizar los contenidos visualizados por el cliente sin necesidad de recargar páginas o realizar sondeos periódicos. También se utiliza intensivamente bibliotecas como jQuery[20] que facilitan la manipulación del árbol del documento estático de las páginas web, y finalmente, también se integra la tecnología AJAX[21] que garantiza la compatibilidad del código entre navegadores.

A continuación, se introducen las principales aplicaciones nativas que se han desarrollado en el marco de la plataforma CC:

- Un **escritorio virtual**, que es una aplicación web que emula un escritorio similar al de cualquier sistema operativo, así es posible interaccionar con la plataforma de forma semejante a como se haría a través de un escritorio tradicional. Su utilidad es la de servir como punto de acceso unificado a las aplicaciones y funcionalidades disponibles en +Cloud de forma individualizada para cada usuario.
- Una aplicación de **gestión personal,** que incluye herramientas para gestionar la información relativa a cada usuario de forma individualizada incluyendo aplicaciones, gestión de servicios contratados, perfil de acceso, datos personales, etc. Así, en función del rol que cada usuario tiene en la plataforma, tendrá acceso a diferentes funcionalidades. En este sentido, se contempla la existencia de tres roles (Administrador, Desarrollador/Manager y Usuario), de forma que un usuario (real) puede tener asignados varios roles al mismo tiempo dentro de la plataforma. Como se puede apreciar, estos roles están alineados con los roles externos *Cloud Admin*, *Cloud User* y *End User* de la organización virtual principal +Coud.
- Un **catálogo de aplicaciones,** que básicamente es un *market* según la analogía con los mercados tradicionales ampliamente utilizada en las plataformas móviles [407]. A través de esta aplicación es posible consultar información sobre las aplicaciones que se ofertan por terceros, así como adquirirlas para su uso. En función de la configuración propuesta por los desarrolladores de cada aplicación, el usuario para adquirirla deberá aceptar una licencia de uso, después de lo cuál se le asignará un rol específico dentro de la aplicación. El conjunto de roles de cada aplicación puede ser configurado dinámicamente por el *manager*/desarrollador de la misma.
- El **panel de control de la infraestructura** que permite monitorizar y gestionar la infraestructura subyacente, así como los servicios desplegados. Está disponible para el rol *Cloud Admin* principalmente. Incluye dos vistas principales:

[17] http://www.w3.org/html/
[18] http://www.w3.org/Style/CSS/
[19] https://tools.ietf.org/html/rfc6455
[20] http://jquery.com/
[21] http://www.w3.org/TR/XMLHttpRequest/

- La *monitorización y control de infraestructura* que hace posible gestionar el estado de la infraestructura interna, distribuyendo sus recursos entre los servicios ofertados a los usuarios.
- La *vista de servicios* que permite visualizar el rendimiento de cada servicio desplegado en la plataforma.

Además, también existen las aplicaciones desarrolladas por terceros, las cuáles se encuentran desplegadas en el entorno CC y que, por tanto, también se ofrecen como servicio a los usuarios finales, en este caso el rol *End User*. Este segundo grupo de aplicaciones, constituye la razón de existencia de la plataforma CC, y por tanto del SMA +Cloud, según se ha descrito en la vista funcional del modelo de organización (Capítulo 4, apartado 4.2) de la metodología GORMAS, ya que esta plataforma CC es de propósito general y, por lo tanto, está orientada a proporcionar servicios a terceros.

Para que todas las aplicaciones (nativas y de terceros) sean realmente elásticas, deben utilizar los servicios de la capa inferior de la plataforma para satisfacer sus necesidades de persistencia de información. Gracias a este modelo arquitectónico es posible asegurar la elasticidad de las aplicaciones desplegadas en la plataforma, ya que hace posible efectuar el balanceo de las peticiones y, por lo tanto, el reparto de la carga entre las distintas réplicas de un servicio sin que se pierda información o se produzcan incongruencias en la información disponible. En el siguiente apartado se detallan los servicios de persistencia de información que se enmarcan en la capa PaaS de la plataforma CC.

5.1.1.2.El nivel PaaS

Los componentes de esta capa están orientados a hacer posible la persistencia de la información de las aplicaciones *software* desplegadas en la plataforma, lo que hace posible la elasticidad de las mismas. Gracias al modelo arquitectónico, las aplicaciones pueden desplegarse en máquinas virtuales sin estado, las cuáles pueden ser instanciadas y paradas en función de la demanda de los usuarios.

En esta capa se dispone de un conjunto de componentes cuya funcionalidad se expone en términos de un conjunto de APIs basadas en servicios web sin estado, implementadas mediante el protocolo REST[22] de servicios web. A continuación, se presenta una descripción de los principales componentes:

- El sistema de **gestión de la identidad** que permite validar las credenciales de usuarios y aplicaciones que hacen uso del sistema CC. Para ello, se incluyen dos servicios principalmente:
  - *Autenticación de usuarios* que permite una gestión de la identidad, unificando a todas las aplicaciones y servicios del sistema CC. Para ello, implementa un módulo que proporciona una autenticación de tipo *Single Sign On* [408]. De este modo, el usuario puede autentificarse una única vez con independencia de la aplicación (nativa o no) que esté utilizando.
  - *Autenticación de aplicaciones* que valida las credenciales de las aplicaciones que han hecho uso de servicios desplegados dentro del sistema CC. Para ello, a los desarrolladores de las aplicaciones se les proporciona un identificador y una clave para cada aplicación. Posteriormente, estas credenciales deben adjuntarse a cada petición lo que permite validar su autenticidad.

[22] http://www.w3.org/TR/ws-arch/

- El sistema de **gestión de persistencia de información** hace posible la persistencia de los datos en cada una de las aplicaciones. Para ello, expone dos servicios en función del tipo de información que sea necesario almacenar:
  - S*ervicio de almacenamiento de ficheros,* que proporciona una interfaz en términos de servicios web a un contenedor de ficheros que emula una estructura de directorio de un sistema de ficheros tradicional. Gracias a este servicio, las aplicaciones pueden guardar y recuperar ficheros sin conocimiento acerca de dónde se encuentran físicamente almacenados.
    Este servicio también incluye otras características como un mecanismo de control de versiones, el almacenamiento de metadatos asociados al fichero y la gestión eficiente de subidas y descargas de ficheros grandes.
  - *Servicios de almacenamiento de información,* proporciona un conjunto de operaciones básicas (creación, eliminación, actualización y recuperación) sobre una base de datos orientada a documentos [30] que es simple y flexible. Así las aplicaciones tienen acceso a un conjunto de servicios web que abstraen la tecnología concreta de persistencia, es decir, la base de datos que se utiliza finalmente para la persistencia.

Para la implementación de estos componentes se ha utilizado el lenguaje de programación Python[23] debido a su potencia, facilidad de mantenimiento y la flexibilidad para el manejo de estructuras de datos. En este sentido, el formato de intercambio de datos está basado en JSON[24] (*JavaScript Object Notation*). La capa de servicios web se ha implementado mediante el uso del *framework* de desarrollo web Tornado[25], por su capacidad de gestionar un gran número de conexiones de clientes.

5.1.1.3.El nivel IaaS

La plataforma también ofrece un servicio de infraestructura, aunque mucho más limitado que los servicios proporcionados en las capas anteriormente descritas (SaaS y PaaS). Así la plataforma proporciona un servicio de infraestructura de tipo VPS (*Virtual Private Server*) [409] orientado al despliegue de las aplicaciones de terceros que se ofrecen como servicio en el nivel SaaS. Sin embargo, los VPS proporcionados en este nivel no podrán ser utilizados en ningún caso para otros fines como puede ser la computación de alto rendimiento o la persistencia de grandes volúmenes de información.

Aunque las capacidades de este nivel son bastante limitadas, el usuario puede seleccionar el tipo de máquina virtual que necesita. Estas máquinas virtuales posteriormente serán instanciadas por el sistema de virtualización subyacente a partir de una plantilla previamente creada. Gracias a este modelo del servicio de infraestructura, la plataforma permite desplegar las aplicaciones del usuario en la propia infraestructura de la plataforma lo que facilita la integración en su ecosistema CC desarrollado (aplicaciones nativas, escritorio virtual, etc.) y, también, reduce el tiempo de latencia en el acceso a los servicios de la capa PaaS.

5.1.2.El entorno tecnológico de la plataforma

En el apartado 4.1 del Capítulo 4 ya se ha caracterizado en profundidad un entorno CC de forma general. Este apartado, por tanto, se centra en describir en líneas generales el entorno tecnológico de la plataforma CC que integra el sistema +Cloud. En la Figura 31 se presenta a muy alto nivel los estratos de cualquier plataforma CC que, como se puede apreciar se subdividen en dos grandes grupos, la capa de despliegue *software* y la capa de infraestructura interna *hardware*. En primer lugar, la capa de despliegue *software* incluye los servicios en los niveles SaaS y PaaS, así mismo, esta capa

[23] https://www.python.org/
[24] http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf
[25] http://www.tornadoweb.org/en/stable/

también incluye las bases de datos que hacen posible la persistencia de la información del entorno tecnológico. Por su parte, la infraestructura *hardware* interna incluye tanto los componentes físicos, como la capa de virtualización que hace posible la gestión dinámica de los recursos disponibles en este estrato.

Este complejo entorno tecnológico dentro de la plataforma es monitorizado y controlado por el SMA +Cloud que se ha presentado en el 4.2 del Capítulo 4. Esta plataforma que está basada en organizaciones virtuales permite la integración de modelos de razonamiento avanzado que hace posible la distribución de recursos (físicos y virtuales) entre los servicios ofertados al usuario por la plataforma CC. Para interaccionar con el entorno hace uso de puertos, según se describe en la metodología GORMAS. Gracias a estos puertos es posible abstraer el complejo entorno subyacente de los agentes individuales que realizan la toma de decisiones, lo que permite independizar estos algoritmos y modelos desarrollados de la tecnología concreta que se utilice en el entorno.

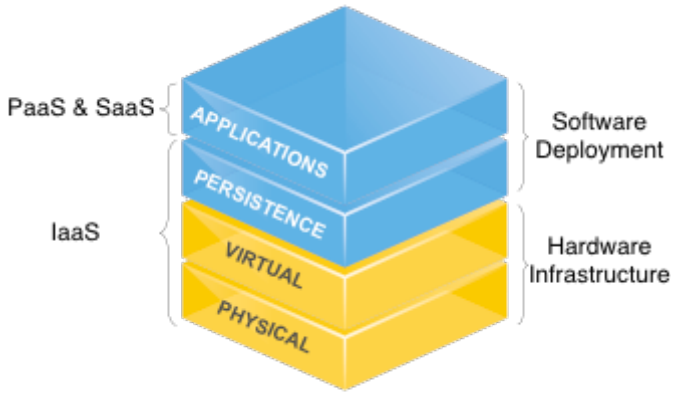

Figura 31.- Estratos a nivel tecnológico de un entorno Cloud Computing

A lo largo de los siguientes subapartados se presentarán los principales rasgos tecnológicos de los componentes internos que hacen posible ofrecer los servicios externos elásticos a la plataforma CC. Así, en el apartado 5.1.2.1 se presentará el modelo y tecnología empleada para la persistencia de la información, tanto de objetos como de ficheros. A continuación, en el apartado 5.1.2.2 se presentará los detalles acerca del sistema de balanceo de carga. Posteriormente, el apartado 5.1.2.3 incluye una descripción del sistema de comunicación que hace posible la interacción entre los agentes del sistema +Cloud, integrado en la plataforma. Finalmente, el último apartado de la sección (5.1.2.4) se presenta la tecnología de virtualización que emplea la plataforma CC, así como una breve descripción de cómo se realiza la interacción con el mismo.

5.1.2.1.La capa de persistencia de información

En esta capa se enmarcan tanto la persistencia de información en bases de datos, como el almacenamiento de ficheros. Su cometido es hacer frente a la demanda de almacenamiento en un entorno CC de altas prestaciones. A continuación, se presentan detalles sobre estos dos componentes que hacen posible la persistencia en la plataforma CC:

- **Persistencia de información de bases de datos**. Las limitaciones y rigidez del modelo relacional hacen que sea complicado su escalabilidad en entornos distribuidos y, además, dificulta el mantenimiento evolutivo de las bases de datos construidas sobre este modelo de almacenamiento. Para solucionar estos problemas, en los últimos años han aparecido una gran cantidad de propuestas de modelos alternativos al relacional [30], de los cuáles, el que más éxito ha tenido es el modelo de almacenamiento orientado a documentos.

La persistencia de objetos en la plataforma CC se basa en la utilización de las bases de datos orientadas a documentos denominada MongoDB[26], que es un sistema gestor de bases de datos de código abierto, distribuido y escalable que se adapta perfectamente a las necesidades perseguidas en el marco de la plataforma desarrollada.

- **Persistencia de ficheros.** En un sistema CC, en general, la persistencia de ficheros debe realizarse en entornos distribuidos que permitan la replicación y partición de los datos [374], habitualmente se utilizan sistemas SAN (*Storage Area Network*).
  El sistema SAN de la plataforma es altamente escalable, ya que se ha probado y validado su correcto funcionamiento con dos sistemas diferentes. Así, se ha utilizado tanto el sistema tradicional de ficheros distribuido NFS (*Network File System*)[27], como del moderno sistema GlusterFS[28] que ofrece mejores características en cuanto a escalabilidad y mayores posibilidades de configuración y readaptación dinámicas.

5.1.2.2. Balanceo de carga en los servicios

Un *proxy* inverso es un componente *hardware* o *software* que recibe peticiones de clientes y las delega a otros nodos que son los proveedores reales del servicio. Por su parte, los balanceadores de carga pueden ser considerados un tipo especial de *proxy* con la habilidad de repartir las peticiones de forma homogénea entre diferentes nodos en tiempo real y, además, permitiendo otras funcionalidades como la modificación de las peticiones, la encriptación y desencriptación, etc.

En la plataforma, como implementación del *proxy* y balanceador de carga se ha optado por Nginx[29] que es un proyecto de *software* libre con un amplio espectro de utilización, que se adapta perfectamente a las necesidades que se perseguían al inicio del trabajo de investigación. La interacción con el sistema de balanceo de carga es realizada por los agentes especializados *Service Monitor* y *Service Supervisor*, los cuáles utilizan un puerto del entorno para encapsular los mensajes correspondientes a la interacción. Este puerto es capaz de procesar en tiempo real el archivo de registro para determinar el estado del balanceado. Al mismo tiempo, también es capaz de variar dinámicamente en tiempo de ejecución la configuración y los pesos entre los diferentes nodos a balancear. Para el balanceo de carga se ha optado por la variación del algoritmo *Round-Robin* que se ha descrito en el 4.3.2.1 del Capítulo 4.

5.1.2.3. Comunicación interna de la plataforma

Dentro de la plataforma CC, el sistema de comunicación está formado por una plataforma de mensajería de alta disponibilidad basada en colas de mensajes e implementada a través del sistema RabbitMQ[30], que está basado en el protocolo AMQP[31] (*Advanced Message Queuing Protocol*). Gracias a este modelo es posible la comunicación entre los agentes el SMA +Cloud. Este componente ofrece el soporte necesario para que los diferentes agentes que forman parte de las organizaciones del sistema ofrezcan y descubran servicios. En este sentido, este componente proporciona un mecanismo mediante el cual las entidades autónomas pueden registrar la descripción de servicios como entradas en un directorio y suscribirse a los canales de comunicación que ofrecen el resto de los integrantes del SMA.

Dentro del sistema de paso de mensajes basado en colas, se destaca que se ha hecho posible la alta disponibilidad del mismo ya que está replicado en varios servidores de la propia infraestructura del

[26] https://www.mongodb.org/
[27] http://nfs.sourceforge.net/
[28] http://www.gluster.org/
[29] http://nginx.org/
[30] https://www.rabbitmq.com/
[31] http://www.amqp.org/

entorno CC. Para ello se ha hecho uso de las bondades que ofrece RabbitMQ. El rol *Global Supervisor* del SMA será el encargado de comprobar que este módulo está funcionando correctamente en todo momento. Si no lo estuviera, se instancia una réplica del servicio en un nuevo servidor físico y se determinarán las causas que produjo la no disponibilidad del mismo.

5.1.2.4.El entorno de virtualización

La virtualización [71] es uno de los denominadores comunes de las plataformas CC. Consiste, a grandes rasgos, en la emulación de un entorno *hardware* para un sistema operativo, en este caso virtualizado. De forma que el sistema operativo, en general, no conoce que está siendo ejecutado en una instancia virtual, denominada máquina virtual. Gracias a la virtualización, es posible que en un único servidor físico (anfitrión), pueden existir múltiples servidores virtuales (huéspedes), cada uno de ellos con su propia identidad, compartiendo de manera transparente los recursos del servidor físico.

Ya se describió en detalle en el apartado 4.1.2.1 del Capítulo 4 el modelo de virtualización empleado en este trabajo de tesis doctoral. En cuanto al entorno tecnológico utilizado en la plataforma CC, se ha optado por el uso de entornos con licencias libres, utilizando dos plataformas como son OpenVZ y KVM. En una primera fase del desarrollo se utilizó el sistema de virtualización OpenVZ que está basado en contenedores y, por lo tanto, restringe su funcionamiento a sistemas Linux. Sin embargo, debido al débil aislamiento de OpenVZ, en una segunda fase se migró a KVM que es un sistema de virtualización basado en Qemu [410]. KVM también permite la migración en vivo de máquinas virtuales entre sistemas anfitrión. La asignación de memoria RAM es menos rígida que en otros entornos de virtualización y funciona correctamente con versiones estables (2.6) del *kernel* de Linux. Se ha desarrollado un puerto del entorno, basado en la librería Libvirt[32], que es utilizado por los agentes especializados *Local Monitor* y *Local Manager* para la monitorización y asignación dinámica de recursos en este tipo de entorno distribuido.

5.1.3.Conclusión

A lo largo de este apartado se ha descrito brevemente la plataforma CC que se ha desarrollado en el marco de este trabajo de investigación. La plataforma ofrece un conjunto de servicios externos a nivel *software*, plataforma e infraestructura. Para proporcionar este conjunto de servicios, internamente dispone de un conjunto de componentes que hace posible la virtualización del *hardware*, el balanceo de carga y la comunicación entre componentes.

Esta plataforma es gobernada por el SMA +Cloud, que se integra con la propia plataforma CC. El sistema +Cloud dispone de un conjunto de agentes que se sitúan estratégicamente en diferentes puntos de la plataforma CC para, en primer lugar, monitorizar el estado de la plataforma y de los servicios que se ofrecen. Posteriormente, a partir de estos datos de monitorización el sistema +Cloud es el encargado de realizar las tareas que sean necesarias para hacer frente a las necesidades computacionales de los servicios que se ofertan, haciendo posible que estos servicios sean elásticos. Estas tareas pasan por la ejecución de los algoritmos avanzados que disponen los agentes especializados de +Cloud y que dan lugar a la autoadaptación del sistema y, por tanto, de la propia plataforma CC con la que se integra.

Para evaluar el correcto funcionamiento del sistema y de los algoritmos de adaptación se ha diseñado un caso de estudio, en el que se incluyen diferentes experimentos sobre la plataforma CC y el sistema +Cloud. El siguiente apartado incluye los resultados obtenidos durante la realización de este caso de estudio específico.

[32] http://libvirt.org/

5.2.Caso de Estudio

Para evaluar la arquitectura multiagente propuesta en este trabajo de tesis doctoral se han realizado una serie de experimentos que se describen en esta sección. En los casos de evaluación llevados a cabo se ha hecho uso de la plataforma CC descrita en el apartado anterior, que integra el sistema adaptativo +Cloud propuesto en el Capítulo 4 y que es el centro de los experimentos de evaluación. Las capacidades de reorganización y adaptación de los agentes que forman parte de +Cloud son necesarias en esta plataforma CC para hacer frente a la demanda de peticiones en los servicios que se ofertan a los usuarios. A continuación, se detalla el caso de estudio que se ha diseñado para validar el modelo arquitectónico propuesto, así como el proceso seguido por los algoritmos de razonamiento, que integran los agentes especializados en la distribución de recursos computacionales en entornos distribuidos.

En primer lugar, se enuncia el objetivo principal del caso de estudio, que en este caso es doble: en primer lugar, la simulación del comportamiento de los miembros de la unidad organizativa principal y subunidades que forman el SMA, en un caso de adaptación real desplegados sobre la plataforma CC. Posteriormente, a partir de la validación del correcto funcionamiento de la organización dentro de esta simulación en el que se incluyen los principales elementos del SMA +Cloud, se procederá a evaluar el comportamiento de los modelos de razonamiento que hacen posible la adaptación dinámica del SMA organizativo mediante la distribución de los recursos de la infraestructura del plataforma CC entre los diferentes servicios ofertados en función de la demanda observada.

Para este caso de estudio, la plataforma CC, junto con el SMA +Cloud que la gobierna, se ha desplegado en el entorno HPC del grupo de investigación BISITE, formado por 15 equipos de última generación que permiten la virtualización por hardware a través de la tecnología Intel-VT y del sistema de virtualización KVM. El sistema operativo tanto de las máquinas físicas, como de las instancias virtuales es la distribución de Linux CentOS[33] (*Community ENTerprise Operating System*) que es la variante libre de la conocida distribución Red Hat[34].

La descripción de cada una de las subcapas según la caracterización del entorno propuesto en el apartado 4.1.2 del Capítulo 4 es la siguiente:

- La *subcapa de ejecución*, que en el caso de estudio ha estado formada por los procesadores y memoria de 10 equipos físicos, cada uno de ellos con 16Gb de memoria RAM y un procesador de cuatro núcleos que permite la ejecución de 8 hilos de forma simultánea, es decir, se disponen de 8 procesadores virtuales (*vcpu*). Por lo tanto, cada máquina admite un máximo de 8 máquinas virtuales, cada una de ellas con un núcleo de ejecución. Si alguna máquina necesita un número de *vcpus* mayor según su plantilla de servicio, el número de máquinas que podría hospedar el equipo disminuiría en proporción.
- La *subcapa de persistencia* está formada por 4 equipos que conforman un entorno SAN construido mediante la plataforma GlusterFS. Esta capa tiene como objetivo almacenar los discos duros de las máquinas virtuales en ejecución, las plantillas de las máquinas virtuales de los servicios y, finalmente, los datos de las aplicaciones/usuarios gestionados mediante cuotas.

El equipamiento restante se integra de forma única y no replicada (por simplicidad y dado que no se evalúa la disponibilidad de la plataforma) los repositorios que utilizan los diferentes agentes de +Cloud y el sistema de comunicación basado en RabbitMQ. Todos los nodos de ambas capas están

[33] https://www.centos.org/
[34] http://www.redhat.com/

unidos mediante un equipamiento de red (*switch*) que permite la comunicación a alta velocidad (*Gigabit*).

5.2.1.Descripción del estado inicial

El caso de estudio se basa en la realización de un ataque de tipo denegación de servicio (*Denial of Service*, DoS)[411] sobre una de las APIs que expone la plataforma CC en la capa PaaS, concretamente, sobre el servicio de persistencia de ficheros. De todos los servicios web de tipo REST que expone este componente de cara a los usuarios del sistema CC, se seleccionaron los dos siguientes:

- **GetSize**. Obtiene el tamaño en *bytes* de un fichero que se le suministra como parámetro, o de la suma de tamaños del contenido de un directorio.
- **GetFolderContent**. Recupera el contenido de un directorio que se le suministra como parámetro.

El método *GetSize* es una función compleja que utiliza recursividad para calcular la suma de los tamaños de los ficheros contenidos en un directorio. Por su parte, el método *GetFolderContent*, es una función mucho más sencilla que tan sólo devuelve los identificadores de los ficheros o directorios contenidos en la ruta suministrada como parámetro.

En el transcurso de este caso de estudio se procedió a evaluar, a través de una serie de experimentos, en primer lugar, la redistribución de la infraestructura a nivel micro, y posteriormente, la distribución de infraestructura a nivel macro. En cuanto a la distribución de recursos a nivel servicio, ésta queda implícitamente evaluada en cada uno de los experimentos que se van a presentar a continuación debido a que este algoritmo es ejecutado por el agente *Service Monitor* asociado al servicio de forma continuada para mantener el nivel de QoS asociado a cada uno de los servicios ofertados por la plataforma. Así, como se pudo observar en los experimentos que se presentan a continuación, una vez que se produce una ejecución de los algoritmos de readaptación en el nivel de infraestructura, el modelo de adaptación de recursos a nivel servicio tiende a equilibrar los pesos de los nodos que proporcionan los servicios para que el nivel de QoS permanezca constante de cara al usuario final. De este modo, el agente *Cloud User* que accede a los servicios tiene una visión de uniformidad en el acceso a cada uno de los productos que se ofertan a través de la plataforma CC, pese a que la adaptación se produce de forma interna en la plataforma.

Todos los experimentos que se presentan a continuación parten, de un mismo estado inicial. En este sentido, se opta por un punto de partida en el que los experimentos que se han realizado durante el caso de estudio se aproximen en gran medida a las condiciones de un despliegue real, en el que la plataforma CC estuviera en producción sometida a peticiones de usuarios reales. Por tanto, este punto de partida, además de ser didáctico y fácil de comprender, refleja un despliegue habitual de cualquier servicio en un entorno CC.

En cuanto a la distribución de recursos inicial, el servicio de almacenamiento de ficheros está desplegado en dos nodos diferentes (VM1 y VM2), cada uno de ellos hospedado por una máquina física diferente (PR1 y PR2, respectivamente). Gracias a este despliegue el servicio tiene alta disponibilidad (está desplegado en dos servidores), pero al mismo tiempo está situado en máquinas físicas con diferente carga computacional, tal y como sucede en un entorno real, ya que ambas máquinas físicas hospedan otras máquinas virtuales correspondientes a otros servicios de la plataforma CC. En este sentido, el servidor físico PR1 tiene recursos libres, sin asignar. Mientras que el servidor PR2 no tiene recursos libres y las máquinas que aloja, están sometidas a carga computacional alta. En la Figura 32 se representa gráficamente el estado de partida de los casos de

estudio que se han realizado. Así mismo, también se puede observar los principales agentes que intervienen en el proceso de readaptación dentro del caso de estudio.

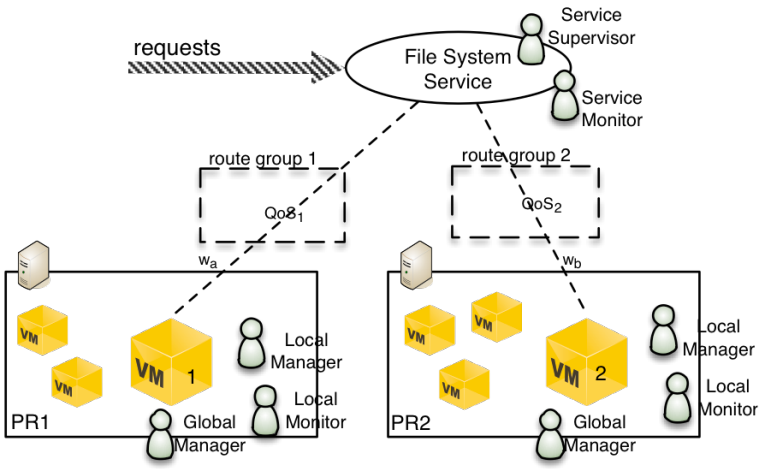

Figura 32.- Estado inicial del caso de estudio de evaluación

Los experimentos que se han diseñado comienzan, realizando peticiones a los dos métodos anteriormente indicados (*GetSize* y *GetFolderContent*) del componente que permite la persistencia de ficheros. El número de peticiones se aumenta progresivamente, intensificando por tanto paulativamente la demanda del servicio y las necesidades computacionales. Debido a este aumento, llega un momento en el que se hace necesaria la readaptación del sistema para hacer frente al creciente número de peticiones.

Destacar en último lugar, que los algoritmos de readaptación en los experimentos que se presentan a continuación, se han ejecutado en el momento en que se detecta un bajo nivel de QoS. En los experimentos se sigue este modelo ya que de cara a la presentación de los resultados se obtiene una vista más clara, simple y ordenada del funcionamiento de los algoritmos. No obstante, en caso de que la plataforma CC estuviera en una fase de explotación con usuarios reales, el proceso de readaptación se debería ejecutar cuando se observa un estado continuado de bajo QoS en alguno de los servicios.

5.2.2.Evaluación del modelo de adaptación de infraestructura a nivel micro

Para evaluar la readaptación de infraestructura a nivel micro, se realiza el primer tipo de experimentos en el que se lanza progresivamente, cada segundo, un hilo de ejecución hasta un máximo de 10 segundos (10 hilos). En este caso, cada hilo continuamente consulta el servicio web *GetSize* del servicio de almacenamiento de ficheros. Este método, como se ha indicado previamente es complejo, ya que realiza consultas internas recursivas para conocer el tamaño del objeto (fichero o carpeta). En el experimento se considera que la QoS del servicio es adecuada cuando el tiempo de respuesta de las peticiones no supera los 2,5 segundos. Por lo tanto, en caso de que se supere este tiempo de respuesta, sería necesaria una readaptación del sistema.

El resultado del experimentado se presenta en la Figura 33 que muestra una gráfica con los tiempos de respuesta del método *GetSize* durante el tiempo que ha durado la prueba. Así pues, el sistema una vez que detecta que se ha reducido el nivel de QoS en el servicio, es decir, cuando los tiempos de respuesta medios son superiores a 2,5 segundos, inicia de forma automática un proceso de adaptación de infraestructura a nivel micro. Este proceso es lanzado por el agente *Service Supervisor*,

en función de la información que le proporciona el agente *Service Monitor* acerca del estado del servicio. Estos dos agentes especializados están asociados únicamente al servicio que se considera.

Como se puede observar en la Figura 33, una vez completado este proceso de autoadaptación y ajustados los pesos de balanceo, el tiempo de respuesta del servicio vuelven a situarse por debajo de los de los niveles de QoS considerados como aceptables (por debajo del límite de los 2,5 segundos).

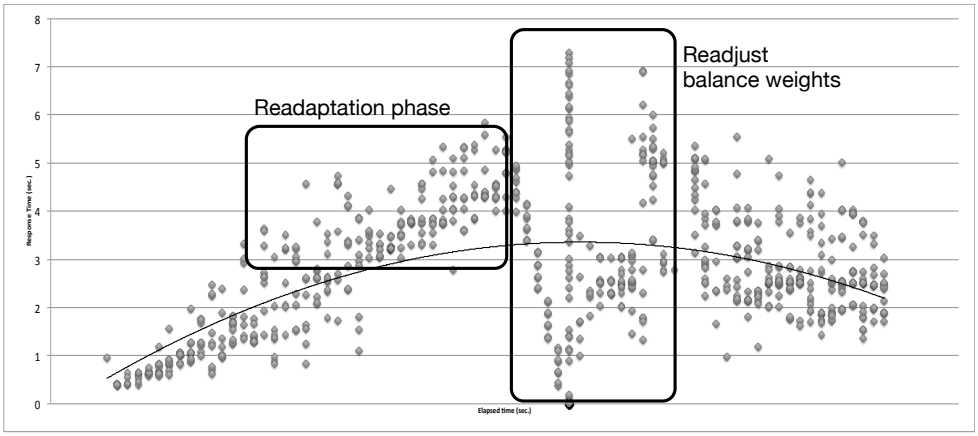

Figura 33.- Experimento 1: Reajuste de recursos de infraestructura a nivel micro (Método: *GetSize*)

A continuación, se detalla el proceso que se ha llevado a cabo durante la distribución de recursos de infraestructura a nivel micro. El algoritmo comienza cuando el agente *Service Monitor* detecta que el nivel de QoS asociado al servicio se ha degradado progresivamente superando el mínimo tiempo de respuesta que se considera aceptable (2,5 segundos).

El agente *Service Supervisor*, a partir de estos datos sobre la QoS del servicio, decide iniciar un proceso de redistribución de recursos de infraestructura a nivel micro, el proceso completo se presenta en la Figura 34. En primer lugar, este agente *Service Supervisor* envía un mensaje (Paso 1, Figura 34) a cada uno de los agentes *Local Manager* de las máquinas físicas (PR1 y PR2) que alojan los nodos trabajadores del servicio para indicarles que el servicio de persistencia de ficheros necesita un mayor número de recursos. Junto a este mensaje también se le indica el peso que tiene cada nodo en el proceso de balanceo de peticiones. A partir de esta información cada agente *Local Manager*, de forma independiente, determina la cantidad de recursos adicionales que puede dedicarle al servicio. Para ello, el agente *Local Manager* de cada máquina solicita a su agente homólogo *Local Monitor*, la información acerca del estado de cada máquina, es decir, la instantánea I_{FR}^{t} que caracteriza al equipo físico.

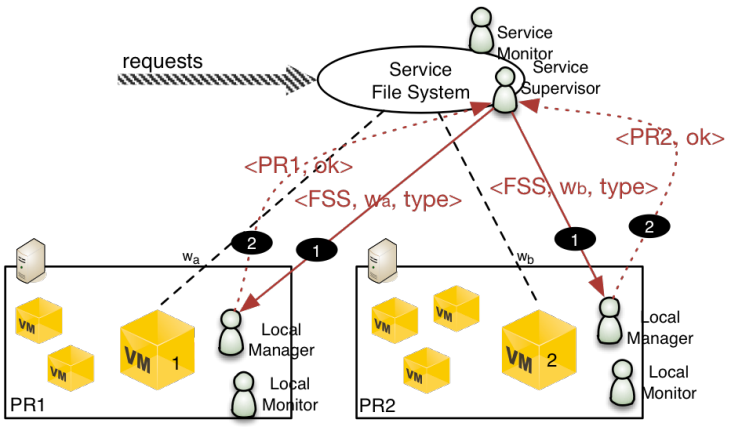

Figura 34.- Experimento 1: Intercambio de mensajes en la distribución de infraestructura a nivel micro

Una vez que el agente *Local Manager* ha recuperado toda la información local que utiliza para llevar a cabo el proceso de toma de decisión, el siguiente paso consiste en determinar mediante su proceso de razonamiento interno la cantidad de recursos (*vcpu* y memoria) que puede provisionar para el nodo que hospeda. Concretamente, en el experimento se han dado las dos opciones posibles atendiendo a la cantidad de recursos libres de cada uno de los servidores que hospedan los nodos virtuales:

- En el caso del servidor PR1, como dispone de recursos libres que no están asignados a ningún otro nodo, la asignación se produce de forma directa. Así pues, se le asigna una *vcpu* adicional al nodo que demanda recursos, y más memoria en función de su prioridad dentro del sistema de balanceo, según la expresión que se detalló en el apartado 4.3.2.2 del Capítulo 4.
- En el caso del servidor PR2 hospeda un número mayor de máquinas virtuales todas ellas con una mayor carga computacional, las cuáles están asociadas a otros servicios de la plataforma. Por lo tanto, no dispone de recursos libres, por lo que evalúa si el conjunto de nodos que hospeda tiene recursos infrautilizados:
  - En cuanto a la evaluación de unidades de cómputo (*vcpu*), el servidor PR2 no dispone de ninguno libre. Ya que de sus 8 *vcpus*, todos ellos están asignados a otras máquinas virtuales que hospeda en el momento en el que se ha realizado experimento, así como al servidor físico. Sin embargo, detecta que una de las máquinas virtuales que tiene asignados 2 *vcpus* está infrautilizando uno de ellos, por lo que decide retirar este recurso de computación a esta máquina y asignárselo al nodo virtual del servicio de almacenamiento de ficheros que ha demandado recursos.
  - En cuanto a la provisión de nueva memoria tampoco dispone de recursos libres, por lo que determina la memoria infrautilizada por las máquinas existentes mediante el proceso de optimización basado en programación lineal. Después de realizar esta evaluación el modelo de optimización determina que no es conveniente reasignar nuevos recursos, ya que no existen recursos infrautilizados, por lo que no se provisiona memoria para el nodo virtual que demanda recursos.

  Después del proceso de razonamiento, el agente *Local Manager* en la máquina PR2 le asigna al nodo del servicio que demanda mayor cantidad de recursos en términos de unidades de cómputo (*vcpus*), pero no así en términos de memoria.

Una vez finalizado este proceso de razonamiento individualizado en cada servidor físico, ambos agentes *Local Manager* situados en PR1 y PR2 notifican (Paso 2, Figura 34) de forma individualizada al agente *Service Supervisor* el resultado del proceso de aprovisionamiento de servicios. Una vez se reciben las notificaciones, el agente *Service Supervisor* no realiza ninguna acción adicional, ya que al menos uno de los nodos físicos ha provisionado más recursos para el servicio en su conjunto (en este caso ambos nodos físicos). Por su parte, el agente *Service Monitor*, realiza el equilibrado de pesos para ajustar la demanda a las capacidades de los nodos atendiendo a los nuevos recursos. En función de este ajuste en los pesos de cada nodo, si se reduce el tiempo de respuesta medio del servicio (situándolo por debajo del umbral de los 2,5 segundos, como así ha sucedido), el agente *Service Supevisor* no realizará ninguna acción más. Sin embargo, en el caso, de que el problema persistiera se realizaría un proceso de adaptación a nivel macro.

5.2.3.Evaluación del modelo de adaptación de infraestructura a nivel macro

Una vez que se ha presentado la distribución de recursos de infraestructura de recursos a nivel micro, a continuación en el segundo tipo de experimentos se procede a evaluar el nivel macro. En

este segundo tipo de experimentos, por claridad sólo se evalúa la adaptación de infraestructura a nivel macro, sin considerar de forma previa el nivel micro que se acaba de evaluar en el apartado anterior. Es decir, cuando el agente *Service Supervisor* una vez que detecta una pérdida de rendimiento ejecuta directamente el proceso de adaptación de infraestructura a nivel macro.

Este tipo de adaptación se produce cuando la demanda en el servicio es mucho mayor, por lo que en este tipo de experimentos, el incremento en la carga computacional del servicio no se realiza tan progresivamente como en el experimento de adaptación anterior, sino que se realiza un incremento más agresivo de la carga. Así, en este caso se lanzan 10 hilos cada tres segundos, hasta un máximo de 40 hilos. Estos hilos al igual que en el caso anterior realizan consultas concurrentes al servicio de almacenamiento de ficheros. Concretamente los hilos consultan continuamente a los dos métodos ya mencionados: *GetSize* (método complejo) y *GetFolderContent* (simple).

La Figura 35 y la Figura 36 presentan el resultado de la ejecución del experimento cuando no se ha producido ningún tipo de adaptación. En ambas gráficas, en primer lugar, se puede observar un mayor número de peticiones, ya que al existir un mayor número de hilos en ejecución, el número de peticiones es mucho mayor. Como también se puede apreciar, en el caso de que no se produzca adaptación, el tiempo de respuesta se incrementa en gran medida, por lo que se reduce proporcionalmente el nivel de QoS aceptable. En este experimento, el nivel de QoS aceptable para la función *GetSize* sigue siendo 2,5 segundos, mientras que el nivel umbral de QoS para *GetFolderContent* se ha establecido en 0,5 segundos.

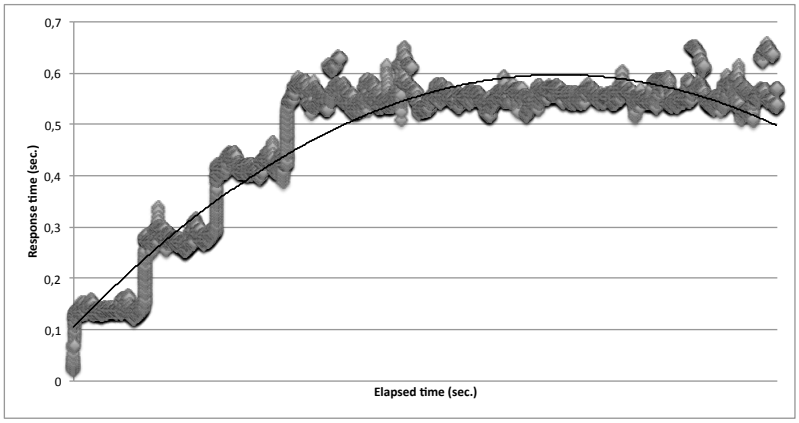

Figura 35.- Experimento 2: Reajuste de recursos de infraestructura a nivel macro, sin adaptación (Método: *GetFolderContent*).

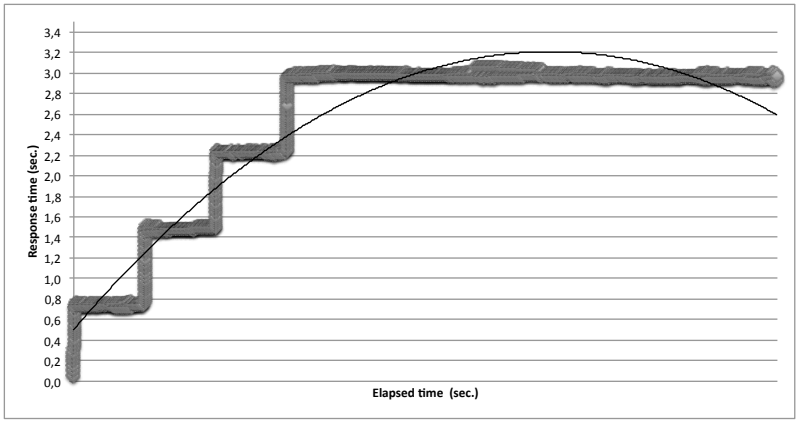

Figura 36.- Experimento 2: Reajuste de recursos de infraestructura a nivel macro, sin adaptación (Método: *GetSize*).

El proceso de interacción entre agentes para la adaptación de infraestructura a nivel macro se presenta en la Figura 37. Este proceso de adaptación se inicia de forma similar al proceso de distribución de recursos de infraestructura a nivel micro, pero en este caso, el agente especializado *Service Supervisor*, que inicia el servicio, envía la alerta (paso 1, Figura 37) al agente *Global Manager* de cada una de las máquinas físicas que alojan los nodos del servicio. Se recuerda, que este agente *Global Manager* es un agente especializado que utiliza un proceso de razonamiento de tipo CBR-BDI [393, 394] encargado de la distribución de recursos a nivel macro. Estos agentes, una vez que reciben la alerta inicial, reenvían este mensaje de alerta al resto de agentes *Global Manager* del sistema CC (paso 2, Figura 37).

El proceso que sigue se realiza de forma paralela en cada uno de los nodos físicos del sistema CC. Cada uno de los agentes *Global Manager*, que reciben el mensaje de alerta, solicita al agente *Local Monitor,* de la máquina en la que residen, la instantánea de estado del sistema (I_{FR}^{t}). A partir de esta información, determinan si existen recursos libres para instanciar un nuevo nodo asociado al servicio que demanda recursos. Si la máquina a partir de esta instantánea detecta que el servidor físico no tiene recursos libres, entonces no realiza ninguna acción. Pero, en el caso de que sí tuviera recursos libres inicia el proceso de razonamiento que se detalló en el apartado 4.3.2.3 del Capítulo 4. Este proceso es realizado de forma paralela en cada una de las máquinas que dispone de recursos libres.

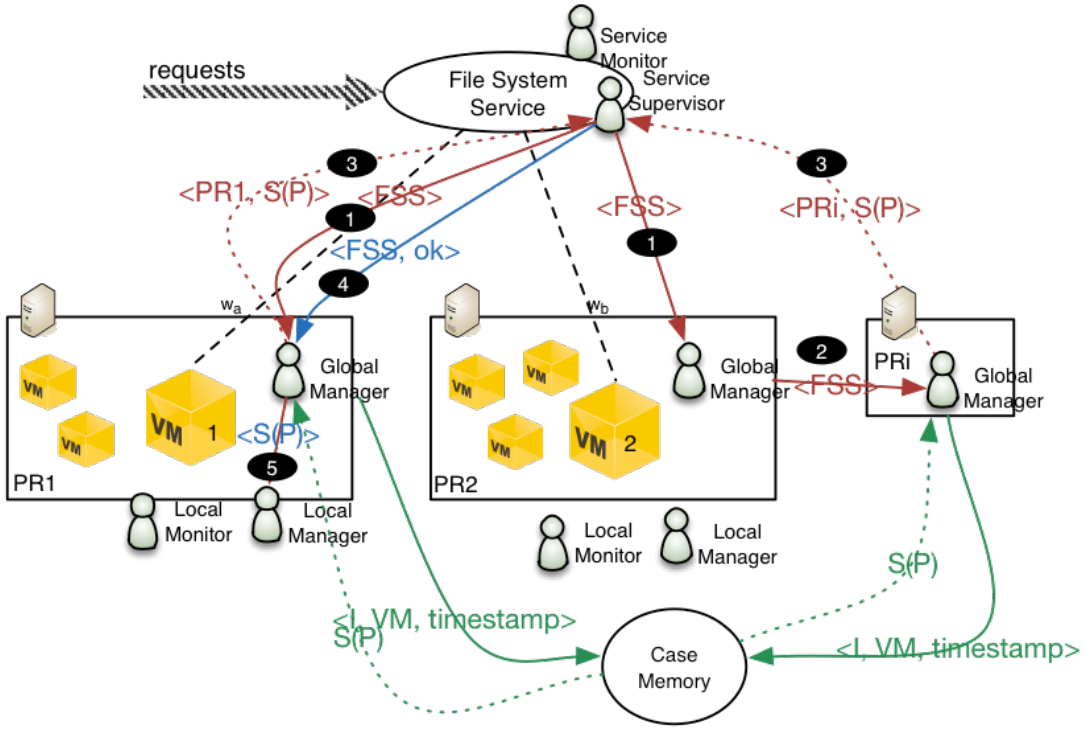

Figura 37.- Experimento 2: Intercambio de mensajes en la distribución de infraestructura a nivel macro

El proceso de razonamiento se inicia definiendo el problema, P, en cada máquina física. La caracterización del problema se construye a partir de la instantánea de uso que proporciona el agente *Local Monitor*, los recursos mínimos del nuevo nodo que se desea instanciar y la marca de tiempo actual. A partir de esta descripción del problema se recuperan los casos similares de la memoria de casos. Cada agente *Global Manager* de forma independiente determina los recursos que puede conceder al nuevo nodo que se acaba de instanciar. Esta nueva provisión de recursos es, en sí misma, la solución del problema para el caso actual, $S(P)$.

El conjunto de soluciones que proponen cada uno de los agentes CBR-BDI de cada servidor físico con recursos libres se envía al agente *Service Supervisor* que ha iniciado el proceso debido a sus problemas con el rendimiento (paso 3, Figura 37). Este agente *Service Supervisor* encargado de controlar el servicio, una vez que haya recibido el conjunto de soluciones propuestas por los diferentes agentes CBR-BDI de la organización, envía un mensaje de aceptación a aquel agente *Global Manager* que haya ofertado un mayor número de recursos para el nuevo nodo que se tiene que instanciar (paso 4, Figura 37). Finalmente, el agente *Global Manager* que recibe la solicitud, solicita al agente *Local Manager* (paso 5, Figura 37) de su misma máquina, que instancie un nuevo nodo virtual a partir de la plantilla de la máquina virtual asociada al servicio según la solución del problema que se ha propuesto.

Posteriormente, una vez que se lanza el nuevo nodo de ejecución, resulta necesario evaluar la solución propuesta, esta evaluación se realiza por el agente *Local Manager*, en función del grado de recursos infrautilizados del nodo que se acaba de instanciar. En el caso de que sea necesario un nuevo proceso de distribución de recursos a nivel macro, será el agente *Service Supervisor* el que complete la evaluación realizada por el agente *Local Manager* para así penalizar dicha solución. En ambos casos, se evalúa la eficiencia de la solución propuesta en función de la cantidad de recursos infrautilizados del nuevo nodo instanciado.

El caso de estudio se ha repetido en numerosas ocasiones lo que ha permitido almacenar en la memoria de casos un buen número de experiencias pasadas. A partir de este proceso se detecta que el proceso de razonamiento está altamente influenciado por el contenido de la memoria de casos, así en función de la cantidad y calidad de los casos existentes, correspondientes a experiencias pasadas, se obtienen mejores resultados en cuanto a la adaptación de infraestructura a nivel macro. Por ejemplo, en la Figura 38, y posteriormente, en la Figura 39 se presenta una distribución de recursos de infraestructura a nivel macro donde la memoria de casos estaba vacía y, por lo tanto, se instanciaron máquinas con una cantidad de recursos igual a la cantidad de recursos mínimos que puede tener el servicio, según su plantilla. Como se aprecia, en ambos casos, la adaptación de recursos soluciona el problema.

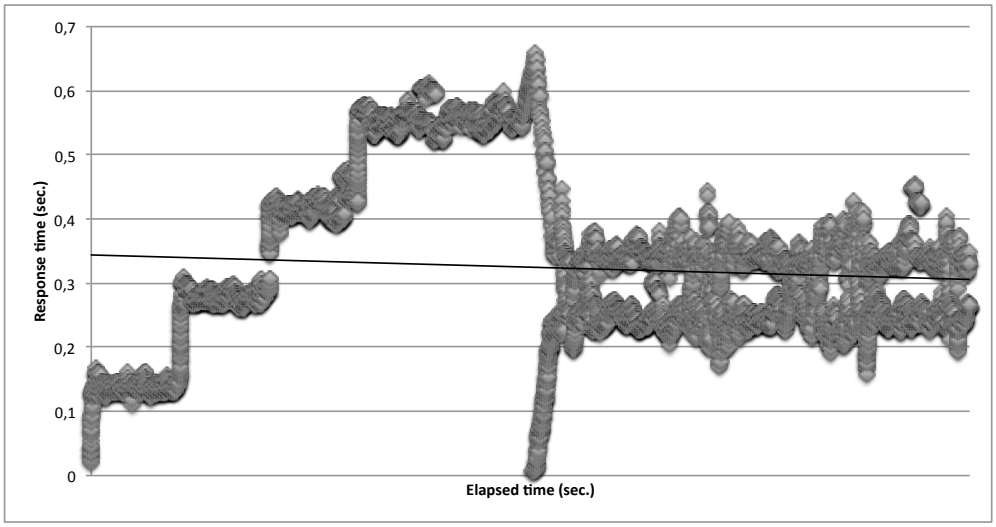

Figura 38.- Experimento 2: Reajuste de recursos de infraestructura a nivel macro, adaptación 1 (Método: *GetFolderContent*)

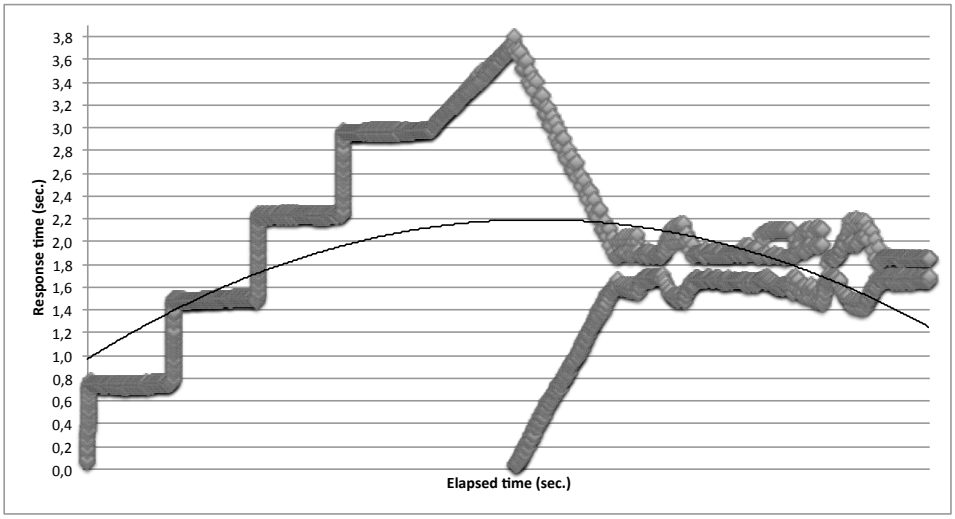

Figura 39.- Experimento 2: Reajuste de recursos de infraestructura a nivel macro, adaptación 1 (Método: *GetSize*)

Aunque en los dos casos anteriores el problema se soluciona, el nivel de QoS se aproxima al margen establecido como de baja calidad (0,5 en el caso de *GetFolderContent* y 2,5 en el caso de *GetSize*). En cambio, como se presentan en la Figura 40 y en la Figura 41, cuando se dispone de una gran cantidad de casos en la memoria, de forma que exista un buen número de experiencias pasadas similares a la actual, se observa que los resultados de adaptación son mejores debido a que el nivel de QoS es inferior. Sin embargo, en este caso, también se observa como desventaja, que el tiempo de razonamiento es mayor, debido a que existe un mayor coste computacional asociado a la recuperación de los casos y la búsqueda de la mejor solución a partir de estos.

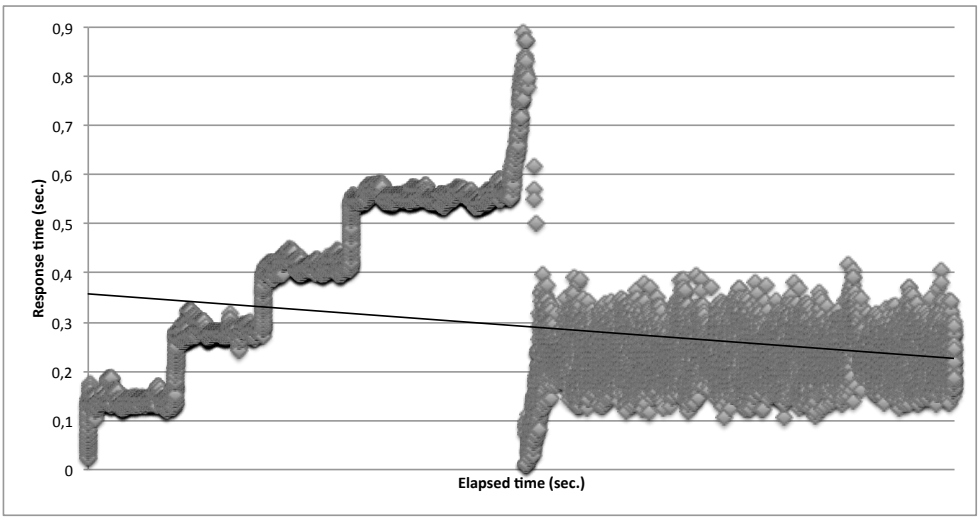

Figura 40.- Experimento 2: Reajuste de recursos de infraestructura a nivel macro, adaptación 2 (Método: *GetFolderContent*).

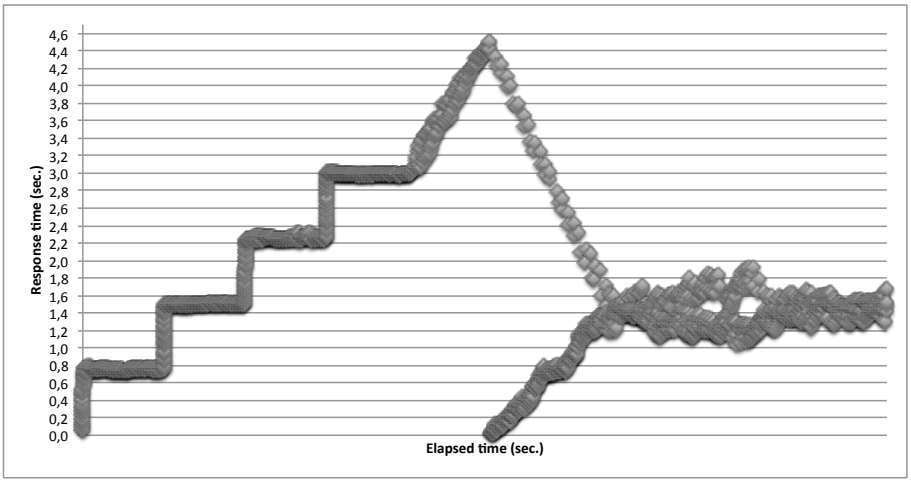

Figura 41.- Experimento 2: Reajuste de recursos de infraestructura a nivel macro, adaptación informada (Método: *GetSize*).

Posteriormente a estos experimentos, también se realizaron otros conjuntos de experimentos de distribución de recursos de infraestructura a nivel macro en los que se buscaban varias ejecuciones consecutivas del proceso de adaptación, de forma que, con una única ejecución del algoritmo de adaptación no fuera posible satisfacer la demanda en los servicios. Este tipo de pruebas, desde el punto de vista del modelo de adaptación fue positiva, ya que éste funcionó perfectamente dentro de los límites del caso de estudio. Sin embargo, tal y como se presenta en la gráfica de la Figura 42 el tiempo de respuesta después de 2 ó 3 procesos de readaptación, dependiendo del método que se esté evaluando, la QoS no se reduce como en las primeras readaptaciones, y el tiempo de respuesta sufre una gran dispersión. En este sentido, se determina que esta dispersión no está derivada de los algoritmos de distribución de recursos propuestos, sino de la saturación debido a la demanda de recursos en las capas inferiores que contienen la persistencia de los datos (bases de datos y sistema de almacenamiento distribuido).

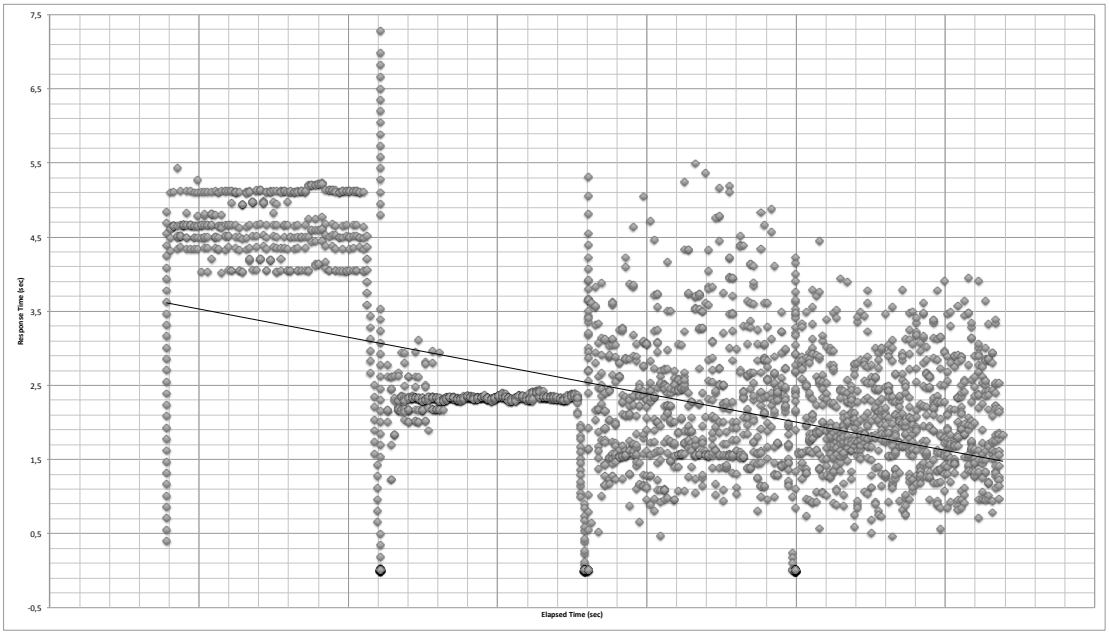

Figura 42.- Experimento 3: Reajuste de recursos de infraestructura a nivel macro, adaptaciones consecutivas (Método: *GetSize*).

5.2.4.Conclusión del caso de estudio

El modelo de adaptación dinámico propuesto se enmarca dentro de las propuestas que buscan incrementar la eficiencia energética, según la clasificación descrita en el apartado 4.3.1 del Capítulo 4. Este tipo de modelos hoy en día están todavía en un estado de desarrollo e inmadurez, debido a que las grandes plataformas utilizan la otra gran aproximación para la distribución de recursos computacionales, la cuál es predominante y está guiada por los intereses económicos [34, 35], en detrimento de los intereses energéticos y de eficiencia. Sin embargo, en el marco de la investigación, los trabajos analizados señalan que los esfuerzos se dirigen a la búsqueda de soluciones a la distribución de recursos computacionales en entornos CC que sean eficientes, es decir, en línea con el modelo que se ha propuesto en este trabajo de investigación. De estas forma es posible asegurar los niveles de QoS acordados mediante acuerdas SLA con los usuarios, mientras que al mismo tiempo sean capaces de ser eficaces energéticamente asegurando una consumo de recursos energéticos adecuado a la demanda de los recursos [21].

La comparación empírica del modelo propuesto con respecto a otras aproximaciones existentes en el estado del arte no es posible, ya que resulta difícil recrear los entornos de computación o/y simulación en el que han sido evaluados. Sin embargo, si que se puede realizar una comparación teórica de la aproximación propuesta con respecto a otros trabajos existentes en el estado del arte. En primer lugar, se observa que el modelo propuesto sigue una aproximación distribuida para resolver el problema lo que lo diferencia totalmente de los trabajos existentes en el estado del arte [28, 391]. Esta aproximación, la cuál se ha demostrado válida para la distribución de recursos computacionales en este tipo de entornos CC, tiene ventajas con respecto a su disponibilidad, ya que no existe un único componente encargado de la distribución de recursos, sino que es el propio sistema (sociedad) el que se reorganiza en su conjunto mediante la adaptación individual de sus componentes (agentes).

En las aproximaciones existentes en el estado del arte, tal y como se detalla especialmente [412], la ejecución de los algoritmos de asignación es una tarea compleja que requiere una gran cantidad de potencia y tiempo computacional. Sin embargo, en el modelo propuesto se simplifica la búsqueda de una solución adecuada al problema, ya que (i) se distribuyen las necesidades computacionales entre diferentes nodos; (ii) el espacio de valores a considerar es menor, ya que cada nodo sólo debe considerar los datos sobre sus recursos propios y, además, no se necesita un conocimiento global de la plataforma; y, finalmente, (iii) cada nodo puede aplicar de forma autónoma una solución parcial al problema, eliminando las necesidades de coordinación a nivel global de la plataforma. Por otro lado, en cuanto a los algoritmos concretos de adaptación, en la propuesta se utilizan técnicas de optimización, las cuáles han sido utilizadas previamente [392], pero nunca siguiendo una aproximación distribuida.

La otra gran diferencia de este trabajo con respecto a otras aproximaciones existentes en el estado del arte es la unidad mínima de distribución. Mientras que habitualmente en el estado del arte esta unidad mínima es la máquina virtual [21, 28], en este trabajo se considera como unidades mínimas el número de *vcpus* y la memoria virtual asignada a cada máquina virtual de la infraestructura. Gracias a esta aproximación es posible distinguir entre nivel micro y macro en la distribución de recursos de infraestructura lo que permite solucionar problemas de demanda sin la necesidad de instanciar máquinas virtuales, lo que constituye en sí mismo una solución energéticamente eficaz y maximiza el uso de los recursos computacionales.

Finalmente, la clara ventaja que se observa del modelo propuesto con respecto a otros trabajos, es su capacidad para el aprendizaje. En el modelo de distribución de infraestructura a nivel macro se

ha utilizado como algoritmo de adaptación un sistema de razonamiento basado en casos que se integra en un agente especializado de la organización +Cloud. Gracias a este enfoque, tal y como se ha demostrado, es posible que el sistema aprenda a partir de las experiencias pasadas, consiguiendo una mayor eficiencia en las adaptaciones a medida que aprende, memorizando tanto las experiencias positivas, como negativas. En el estado del arte no existe ninguna aproximación similar en la que la distribución de recursos computacionales se realice en base a los resultados obtenidos en procesos de adaptación pasados.

5.3.Discusión

En el apartado anterior, se ha evaluado de forma empírica a través de un caso de estudio la arquitectura multiagente y los modelos de adaptación propuestos en el marco de esta investigación. Gracias al conjunto de experimentos que se han realizado se ha podido validar varios aspectos tal y como se detallará a continuación.

En primer lugar, después de evaluar el modelo propuesto se puede afirmar que una arquitectura multiagente basada en OV resulta adecuada al problema a resolver (monitorización y control de un entorno CC) dado que puede ser incorporada en un entorno masivamente distribuido y que, además, permite la integración de diferentes modelos de razonamiento avanzado que hacen posible satisfacer las necesidades de distribución dinámica de recursos computacionales entre los diferentes recursos ofertados.

Por otro lado, el diseño basado en OV permite realizar un modelado de alto nivel, de forma que es posible abstraer completamente el diseño, de la implementación posterior. Por tanto, esta aproximación permite asegurar la independencia entre los estratos software donde se realiza la toma de decisiones y aquellos en los que se realiza la ejecución de acciones en función de las decisiones tomadas. En un entorno CC, esta separación de responsabilidades es particularmente importante ya que, como se ha determinado durante las primeras fases del trabajo de investigación, en las plataformas existentes hoy en día se observa una alta dependencia del entorno tecnológico (herramientas de virtualización, balanceadores de carga, sistemas de archivos distribuidos, etc.). Esta dependencia, constituye una gran limitación y dificulta su evolución, ya que un cambio en alguno de los componentes *hardware* o *software* obliga también a alterar los algoritmos y técnicas que hacen posible la elasticidad del sistema.

En el caso de la arquitectura propuesta +Cloud, dado que utiliza puertos para la comunicación con el entorno esta dependencia se limita a la implementación del propio puerto, es decir, a la interfaz de comunicaciones con el entorno. No cabe duda de que un cambio en las capacidades que ofrece la tecnología subyacente también obligaría a modificar los modelos de razonamiento propuestos, como en una aproximación de diseño tradicional. Sin embargo, para estos casos, los modelos organizativos también ofrecen una respuesta adecuada a esta dificultad. En un sistema como el que se propone, los agentes que forman parte de la sociedad, juegan uno o varios roles concretos. Cada rol define de forma abstracta los objetivos, responsabilidades y privilegios del individuo que lo adquiere. A partir de esta definición de alto nivel, en el caso de que la tecnología proponga nuevas capacidades, el trabajo de adaptación consistiría en modificar al individuo o individuos que realizan las tareas concretas, los cuáles juegan roles en la organización. Por lo tanto, dado que no sería necesario alterar la sociedad en su conjunto y sólo las entidades individuales; la arquitectura propuesta también está por encima de otras plataformas existentes en el mercado.

En segundo lugar, se discuten las capacidades de +Cloud como sistema abierto. Gracias a este modelo de construcción de sistemas software es posible permitir que agentes externos a la organización accedan a la misma para proporcionar cierta funcionalidad, a cambio de que asuman y cumplan un conjunto de normas que garanticen la estabilidad y supervivencia de la sociedad. En el marco del trabajo propuesto se ha diseñado el sistema para que los agentes sean o puedan ser externos al sistema inicialmente diseñado. Dentro de este trabajo, por supuesto, los agentes humanos que participan en la organización (*Cloud User*, *End user* y *Cloud Admin*) son externos, pero también el agente es el *SLA broker* se ha diseñado de este modo. Debido a que el contexto de comercialización de un sistema CC pueda cambiar en el futuro. El sistema +Cloud está adaptado a

estos nuevos modelos desde su concepción, ya que el diseño de este rol como externo permite que agentes accedan a la organización para proporcionar la funcionalidad de negociación y contratación de servicios, con independiencia del modelo de comercialización en el que se basen. Gracias a este modelo de diseño se contribuye a la interoperabilidad entre plataformas, ya que se construye el sistema facilitando un modelo abierto de conexión entre plataformas (Intercloud [140]). Por otro lado, los SMA que siguen un proceso de diseño basado en modelos organizativos, como es el caso del modelo arquitectónico propuesto tiene claras ventajas en el dominio de aplicación de los sistemas CC, ya que permiten la modificación en el diseño mediante la introducción de nuevos roles dentro de la sociedad, para lo cuál tan sólo es necesario definir sus responsabilidades, objetivos y normas.

Finalmente, en el apartado 3.2 del Capítulo 3 se presentó un estudio detallado acerca de las sinergias entre el paradigma CC y los SMA, determinado que en el estado del arte existe un incipiente número de trabajos en los que los SMA aportaban ventajas a los sistemas CC, dentro del grupo denominado por Talia como *Cloud using agents* [97] que posteriormente ha dado lugar a un concepto más amplio: *Agent-based Cloud Computing* [52, 179, 180]. Este conjunto de trabajos analizados se determinó que podían organizarse según los roles de uso de los sistemas CC en la arquitectura de referencia del NIST [10]. Pero además de esta estructuración inicial, también se identificó un conjunto de ámbitos de aplicación de los SMA en el marco de los sistemas CC (seguridad, búsqueda de proveedores, negociación, monitorización, etc.).

En este trabajo, el modelo propuesto es la primera aproximación conocida de los SMA basados en organizaciones virtuales dentro del paradigma CC. Este enfoque, junto con el diseño de la organización como un sistema abierto, permite considerar el entorno computacional CC, el cuál es complejo, tanto desde una perspectiva interna, como externa. Es decir, teniendo en cuenta tanto las capacidades y funcionalidades que se le ofrecen a los usuarios (negociación, acuerdos SLA, capacidades, etc.); como las propias características internas del entorno de computación (monitorización, gestión de infraestructura, etc.). En este sentido, en la Tabla 9 se presentan los ámbitos de aplicación principales del los SMA en el marco del rol *Cloud Broker* del paradigma CC. Como se puede apreciar, siguiendo un modelo de diseño como el del sistema +Cloud, en el paradigma CC no sólo se obtienen ventajas de la aplicación de los SMA desde una perspectiva externa, sino que también se obtienen ventajas debido al control de las características internas de plataforma.

Usos principales	**Nivel interno**				**Nivel externo**			
	Moni.	**Cont.**	**Seg. Priv.**	**Gest. Infr.**	**Neg. SLA**	**Ofer. serv.**	**Com.**	**Eval. Serv.**
Búsqueda de proveedores								
Buscador de servicios	X	X	X			X	X	
Interoperabilidad					X	X		
Modelo de confianza en SLA	X	X			X		X	
Contratación de servicios		X	X		X		X	
Negociación de acuerdos SLA								
Negociación de condiciones		X				X		
Monitorización de acuerdos	X	X	X		X		X	
Composición de servicios								
Búsqueda de servicios compatibles						X	X	
Interoperabilidad	X		X			X		
Composición Horizontal y Vertical	X	X	X			X		

Tabla 9.-Tabla resumen SMA y CC, rol *Cloud Broker* en el entorno +Cloud

Hay que destacar, finalmente, que tras revisar el estado del arte existente en el momento de la realización de este trabajo de tesis doctoral también se observó un limitado número de trabajos en los que los SMA jugaban el rol *Cloud Provider* del paradigma CC. El trabajo constituye una de las primeras aproximaciones como aplicación de los SMA dentro este rol. Finalmente, del mismo modo, en la Tabla 10 se presentan las sinergias entre el nivel externo e interno del uso de SMA organizativos para el rol *Cloud Provider* del paradigma CC. Este rol en el caso de seguir una aproximación tradicional tendría tan sólo ventajas a nivel interno.

Usos principales	Nivel interno				Nivel externo			
	Mon.	**Contr.**	**Seg. Priv.**	**Gest. Infr.**	**Neg. SLA**	**Ofer. serv.**	**Com.**	**Eval. Serv.**
Seguridad y privacidad								
Monitorización de infraestructura	X				X			X
Modelos de autenticación			X				X	
Privalegios de acceso			X				X	
Seguridad de la información			X			X		X
Almacenamiento seguro			X			X		
Oferta de servicios								
Servicios sensibles a la calidad	X	X			X			X
Oferta de servicios						X		
Gestión de infrastructura								
Gestión de recursos	X				X			X
Optimización de costes		X		X	X			X
Gestión de flujo de trabajo				X		X		

Tabla 10.- Tabla resumen SMA y CC, rol *Cloud Provider* en el entorno +Cloud

Capítulo 6. Conclusiones y líneas de trabajo futuras

A lo largo de este trabajo de tesis doctoral se ha presentado un innovador enfoque basado en OV de agentes inteligentes para el control y monitorización de plataformas CC. Gracias a este modelo arquitectónico, implementado en el SMA +Cloud, ha sido posible modelar la elasticidad de los servicios que ofrece una plataforma CC siguiendo un enfoque distribuido y haciendo uso de técnicas derivadas de la IA. Gracias a lo cuál se ha podido introducir en los sistemas CC características como la *autonomía*, *habilidades sociales*, *reactividad* y *pro-actividad*; que tradicionalmente estaban asociados a los SMA. Posteriormente, esta arquitectura propuesta (+Cloud) se ha sometido a un intenso proceso de evaluación a través de un conjunto de experimentos realizados en el marco de un caso de estudio diseñado específicamente, lo que ha permitido validar tanto el modelo, como las diferentes técnicas y algoritmos propuestos, en el marco del proceso de investigación, para proporcionar elasticidad al sistema CC.

Una vez que el proceso de investigación ha terminado, el presente capítulo tiene como objetivo enunciar las conclusiones alcanzadas, así como la contribución de esta investigación al estado del arte (apartado 6.1). Posteriormente, a partir de este hito en el proceso de investigación, en el apartado 6.2 se enunciarán las líneas de trabajo futuro que guiarán el proceso de investigación a medio y largo plazo.

6.1.Conclusiones

Este trabajo se planteaba, en sus inicios, como una de las primeras aproximaciones de los SMA, y más concretamente aquellos que siguen una orientación basada en OV, en el marco de los sistemas de control y monitorización de un entorno CC. Como se ha analizado ampliamente, tanto en el Capítulo 2, como posteriormente en el Capítulo 3 de esta memoria; hoy en día el desarrollo de este tipo de plataformas CC está fundamentalmente dirigido por los intereses empresariales de las grandes compañías tecnológicas. Sin embargo, el trabajo realizado viene a ratificar que en este ámbito de investigación no tiene por qué estar únicamente ligado a la innovación dirigida a la creación de productos comerciales, sino que también hay cabida para la investigación fundamental, a través de la búsqueda de nuevos modelos y algoritmos que permitan hacer más eficiente el ecosistema de los sistemas CC y mejorar la interacción con los usuarios finales.

En este marco de investigación, se ha propuesto un nuevo modelo arquitectónico, basado en OV de SMA, que tiene un carácter claramente integrador. Dentro de esta arquitectura se ha desarrollado, evaluado y validado un conjunto de algoritmos para la distribución de recursos computacionales en entornos CC. Su principal novedad se basa en la capacidad de autoadaptación dinámica del sistema propuesto en función de la demanda. Así, no sólo se realiza una distribución de recursos en función de la demanda de los servicios, sino que es el propio sistema el que se adapta dinámicamente en función de los cambios que se produzcan en el entorno. Así mismo, este esquema de adaptación dinámica se basa en modelos distribuidos lo que reduce la cantidad de información en la toma de decisiones, permitiendo el reparto de responsabilidades y subobjetivos entre los diferentes miembros del sistema.

En conclusión, la plataforma CC que integra la arquitectura propuesta, denominada +Cloud, a tenor de los resultados (empíricos y teóricos) presentados a lo largo de este capítulo permiten validar que la hipótesis de partida de este trabajo de investigación: "*es posible modelar el sistema de control y monitorización de una plataforma CC mediante el uso de OV de agentes inteligentes que se auto-adapten y reorganicen en función de las necesidades del contexto. Así pues, los agentes individuales de forma autónoma gestionarán los recursos computacionales y los servicios disponibles, coordinándose con el resto de agentes para proporcionar un modelo de adaptación dinámico y distribuido en función de los intereses del usuario final*".

Por lo tanto, una vez que se ha validado la hipótesis de partida, se concluye que también se ha satisfecho el objetivo principal, y los subobjetivos asociados. Así, en el siguiente subapartado (6.1.1) se presentarán las principales contribuciones de este trabajo de investigación. Posteriormente, en el subapartado 6.1.2 se presentarán las contribuciones de este trabajo a los proyectos de investigación en el que se enmarca, así como las tareas de difusión realizadas.

6.1.1.Contribuciones a la investigación

El trabajo de investigación que se ha detallado a lo largo de la presente memoria, ha permitido contribuir al estado del arte en los ámbitos de la teoría de agentes, las organizaciones virtuales y los sistemas distribuidos inteligentes. A continuación, se describen las principales contribuciones de esta investigación:

- **Se ha estudiado las necesidades de un entorno CC.** A partir del análisis del estado del arte, incluyendo tanto estudios y trabajos previos, como las plataformas existentes; se ha elaborado un marco de aplicación del trabajo de investigación propuesto, lo que ha permitido detectar las debilidades de la tecnología y la identificación de las áreas susceptibles de mejoras.

- **Se ha analizado la relación entre la teoría de agentes y el paradigma CC.** Durante el trabajo de investigación que se ha llevado a cabo se ha analizado exhaustivamente el estado del arte con el objetivo de descubrir sinergias entre ambos sistemas distribuidos (SMA y CC). De este modo, se ha podido demostrar que en el estado del arte existe suficiente trabajo previo como para sostener la hipótesis inicial de este trabajo de investigación.
- **Se han analizado los modelos organizativos de agentes inteligentes para el diseño de sistemas *software* complejos.** Durante el análisis del estado del arte se ha determinado que el modelo más adecuado para diseñar una arquitectura orientada a la gestión, control y monitorización de un sistema CC es mediante el uso de modelos organizativos de agentes inteligentes.

  Partiendo de esta premisa se ha procedido a analizar los diferentes modelos para el diseño de sistemas abiertos y complejos, como es el caso de un sistema CC. Así pues, como resultado de este análisis, se ha determinado que la metodología GORMAS [49, 50] es la que más se ajusta a los intereses perseguidos, debido a que su amplio meta-modelo permite caracterizar correctamente un sistema *software* complejo. Además, esta metodología está basada en un proceso de diseño iterativo, alineado con SPEM [318].
- **Se ha realizado un amplio estudio de las tecnologías complementarias**. Este trabajo hace uso de un buen número de tecnológicas que completan el trabajo de investigación realizado. En las fases iniciales, se ha realizado un estudio de todas ellas. Por un lado, se ha hecho un *análisis del contexto tecnológico en el entorno CC*, lo que ha permitido seleccionar el conjunto de tecnologías que han formado parte del entorno CC desarrollado (balanceo, virtualización, persistencia, etc.). Así mismo, también se ha realizado un *estudio de técnicas, algoritmos, métodos y modelos de IA* que han constituido la base de los modelos de razonamiento avanzado y aprendizaje desarrollados.
- **Se ha caracterizado el entorno de computación.** A partir del análisis del estado del arte de los diferentes ámbitos que abarca este trabajo de investigación, se ha procedido a formalizar el entorno de computación distribuido sobre el que posteriormente se ha aplicado el modelo de adaptación propuesto. Esta caracterización del entorno no sólo ha constituido el punto de partida para la formalización de la arquitectura que se ha propuesto, sino que también forma parte de los resultados del trabajo realizado, sirviendo como base para diferentes líneas de investigación que se realicen en el futuro.
- **Se ha diseñado una arquitectura multiagente autoorganizativa adecuada al problema a resolver.** Se ha diseñado un modelo arquitectónico específico para el control y monitorización de un sistema CC a través de un modelo organizativo de agentes inteligentes. Este modelo se ha evaluado y validado como apropiado para este contexto. Esta arquitectura integradora permite incluir diferentes agentes especializados con capacidades de razonamiento avanzadas para la distribución de recursos computacionales en un entorno distribuido.

  Para el diseño de esta arquitectura se ha seguido una aproximación basada en sistemas abiertos, lo que hace posible que en el futuro se puedan incluir nuevos modelos, técnicas y algoritmos como parte de la organización diseñada. Así, este trabajo constituya un resultado extrapolable a futuras investigaciones.
- **Se ha diseñado un modelo de adaptación dinámica mediante un sistema de razonamiento basado en casos y algoritmos de optimización.** La distribución de recursos en entornos computacionales es, básicamente, un problema de optimización. Este punto de partida inicial, derivado del estudio del estado del arte, se ha complementado mediante la inclusión de un motor de razonamiento basado en casos [394] que permite que los algoritmos propuestos aprendan de las experiencias basadas, obteniendo mejores resultados con cada nuevo proceso de adaptación.

- **Validación del modelo arquitectónico propuesto en un entorno real de aplicación.** En el contexto de los proyectos de investigación asociados a este trabajo de tesis doctoral se ha desarrollado una plataforma CC, la cuál ha integrado el modelo arquitectónico propuesto basado en OV (+Cloud). Gracias a lo cuál, se ha podido evaluar el modelo propuesto en un entorno de aplicación real y un contexto de explotación. Gracias a esta posibilidad se ha podido evaluar extensivamente el modelo propuesto.

6.1.2.Contribución a proyectos y difusión de resultados

Además de las contribuciones al estado del arte que se acaban de presentar en el apartado anterior, este trabajo de tesis doctoral también ha permitido sentar las bases del desarrollo de un importante proyecto de investigación fundamental no orientada, a nivel nacional:

- **iHAS: Intelligent Social Computing for Human-Agent Societies** (TIN2012-36586-C03-0), otorgado por el Ministerio de Ciencia e Innovación y los fondos FEDER. http://ihas.usal.es

Por otro lado, este trabajo de tesis doctoral también ha sido fundamental para el desarrollo de otros proyectos de investigación y desarrollo, que se han llevado a cabo en colaboración con diferentes empresas a nivel nacional y regional:

- **Cloud-IO: Plataforma Cloud Computing para la Integración y Despliegue Rápido de Servicios sobre Redes Inalámbricas de Sensores** (IDI-20111471), otorgado por el Centro para el Desarrollo Tecnológico e Industrial (CDTI) del Ministerio de Economía y Competitividad y los fondos FEDER.
  http://cloud-io-project.com/
- **SmartMedical: Plataforma Cloud Computing para gestión de historiales clínicos en aseguradoras privadas** (IDI-20120794), otorgado por el Centro para el Desarrollo Tecnológico e Industrial (CDTI) del Ministerio de Economía y Competitividad y los fondos FEDER.
  http://smartmedical.es/
- **DoyFE.ES: Verificación y prevención de fraude en contenidos digitales** (IDI-20120798) otorgado por el Centro para el Desarrollo Tecnológico e Industrial (CDTI) del Ministerio de Economía y Competitividad y los fondos FEDER.
  http://www.doyfe.es/

Finalmente, resaltar que también se ha realizado un gran esfuerzo para intercambiar conocimiento con distintos investigadores y expertos en los ámbitos de investigación de este trabajo de tesis doctoral. De esta forma se ha podido obtener una opinión crítica y constructiva, que ha constituido una pieza clave, en términos de realimentación en las diferentes etapas iterativas del desarrollo del modelo propuesto.

Así mismo, también se ha realizado un intenso esfuerzo en la difusión y diseminación del trabajo mediante la publicación de los resultados en revistas internacionales y la asistencia a conferencias, congresos y foros especializados. Parte de estas contribuciones y actividades de difusión se han realizado en colaboración con grupos interesados en el trabajo de investigación realizado, como es el caso del GTI-IA[35] - *Grupo de Tecnología Informática - Inteligencia Artificial* (UPV, Universidad Politécnica de Valencia) y el *Laboratoire d'Informatique de Grenoble* (CNRS, *Centre National de la Recherche Scientifique*). Finalmente, en cuanto al intercambio de conocimiento y difusión con expertos en el ámbito de la investigación, también se ha buscado la colaboración con el tejido empresarial en el

[35] http://www.gti-ia.upv.es/

marco de la TIC, prueba de ello son las publicaciones en congresos internacionales realizadas conjuntamente con expertos de empresas como INSA – Ingeniería del Software Avanzado S.A.[36], Flag Solutions S.L.[37] y Nebusens S.L.[38]

[36] http://www.insags.com/
[37] http://www.flagsolutions.net/
[38] http://www.nebusens.com/

6.2.Líneas de trabajo futuras

El trabajo de investigación que se ha detallado a lo largo de esta memoria constituye un hito plausible en un largo proceso de búsqueda e indagación de nuevas soluciones, técnicas, herramientas y modelos. Sin embargo, no cabe duda de que este hito constituye el punto de partida de una investigación con un recorrido en el tiempo mucho mayor, debido a que se centra en un ámbito que hoy en día está en clara expansión. Por ello, se han establecido un conjunto de líneas de trabajo a desarrollar que permiten continuar fortaleciendo la base que ya se ha alcanzado con el desarrollo de este trabajo de investigación. Gracias a esta planificación en la investigación futura, será posible inferir nuevo conocimiento orientado solventar problemas y corregir debilidades en el marco de este contexto tecnológico.

Este marco de búsqueda e indagación de nuevas soluciones y modelos, indudablemente también estará altamente influenciado por los objetivos e intereses de los proyectos que sirvan como base a este trabajo de investigación. Concretamente, este trabajo está estrechamente enlazado con el proyecto iHAS, el cuál centra sus intereses en la investigación en mecanismos, algoritmos, herramientas y modelos que permitan la creación de máquinas sociales en los que convivan e interactúen agentes virtuales y humanos de forma transparente en un entorno totalmente integrado. Denominamos este tipo de sistemas sociedades humano-agente (*Human-Agent Societies* HAS).

Dentro este proyecto, el modelo arquitectónico propuesto en este trabajo de investigación está alineado con el primer demostrador del proyecto, que pretende desarrollar un modelo de computación social en el marco de los entornos CC. Por lo tanto, el trabajo desarrollado en el marco de esta tesis doctoral satisface los objetivos marcados para el desarrollo del citado demostrador, ya que se han (i) estudiado las necesidades de un entorno CC; se (ii) ha diseñado un sistema CC utilizando la base que proporcionan las OV de agentes inteligentes; a partir de este diseño, se (iii) ha implementado un prototipo funcional que, posteriormente, se (iv) ha evaluado pormenorizadamente en un entorno de explotación real.

Aunque, como se indicado, ya se han alcanzado los objetivos propuestos, no sólo en el marco de la tesis doctoral, sino también dentro del proyecto al que se asocia este trabajo de investigación; partiendo de esta base, es posible proponer las siguientes líneas de trabajo que se acometerán a corto y medio plazo como complemento a los objetivos perseguidos inicialmente:

- Comparación del modelo propuesto con plataformas existentes, evaluando la posibilidad de integración del sistema +Cloud con otras plataformas CC y evaluando el rendimiento de los modelos de monitorización, control y adaptación en diferentes sistemas CC libres.
- Definición y construcción de un modelo de negociación de acuerdos SLA, que integre diferentes aproximaciones de razonamiento avanzados derivados de la IA para guiar el proceso de negociación, lo que permitirá integrar en la plataforma desarrollada, el servicio de contratación automatizada de los productos que se oferten.
- Definición de un modelo de interoperabilidad de servicios en función de su tipo y ámbito (*software*, plataforma e infraestructura) lo que permitirá definir un protocolo que facilite su composición con independencia del proveedor de los mismos.
- Definición de un modelo de coste que permita determinar en función de la demanda, el modelo de producción y el gasto energético, un modelo de coste que permita establecer precios variables en función del estado concreto del sistema CC.
- La combinación de las tres líneas de trabajo anteriores, permitirá dar soporte a la interoperabilidad entre plataformas a nivel horizontal y vertical de los productos ofertados en una plataforma CC. Para ello, se pretende contribuir al desarrollo del concepto denominado

como Intercloud [139, 140] mediante el desarrollo de nuevas técnicas, herramientas y modelos que hagan uso de las capacidades los SMA para participar en este modelo de interoperabilidad

- Extensión de los algoritmos de distribución de recursos computacionales con dos objetivos principales. Por un lado, se pretende adaptar el modelo de autoadaptación dinámico propuesto para así poder incluir todas las capas *software* de un sistema CC, incluyendo el estrato de persistencia. Para ello, será necesario actualizar el modelo propuesto de forma que también integre las peculiaridades referentes a la elasticidad de las bases de datos y los sistemas de almacenamiento distribuido en red que se han analizado y caracterizado a lo largo del trabajo de investigación ya realizado. Por otro lado, también se pretende ampliar el modelo de adaptación propuesto para que pueda abarcar otros productos de infraestructura, especialmente aquellos que hagan posible la computación de altas prestaciones orientadas al análisis masivo de datos.
- Finalmente, como última línea de trabajo se pretende investigar en el desarrollo de nuevos modelos de interacción de los usuarios con el sistema CC, a través de los últimos dispositivos y modelos disponibles. De esta forma, es posible transformar las plataformas del paradigma CC en máquinas sociales que se adapten a las necesidades de los usuarios no sólo en términos de rendimiento de los servicios, sino también en función de las interacciones con los usuarios.

Bibliografía

1. Fisher, P., R. Pant, and J. Edberg, *Cloud computing: assessing Azure, Amazon Ec2, google App Engine and Hadoop for it decision making and developer career growth*. 2010: Apress.
2. Luo, J.-Z., et al., *Cloud computing: architecture and key technologies*. Journal on Communications, 2011. 7(7): p. 3-21.
3. Wen, X., et al. *Comparison of open-source cloud management platforms: OpenStack and OpenNebula*. in *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*. 2012. IEEE.
4. Von Laszewski, G., et al. *Comparison of multiple cloud frameworks*. in *2012 IEEE Fifth International Conference on Cloud Computing*. 2012. IEEE.
5. Buyya, R., et al., *Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility*. Future Generation computer systems, 2009. 25(6): p. 599-616.
6. Armbrust, M., et al., *A view of cloud computing*. Communications of the ACM, 2010. 53(4): p. 50-58.
7. Mell, P. and T. Grance, *The NIST definition of cloud computing.(2011)*. NIST special publication, 2011. 800: p. 145.
8. Ross, J.W. and G. Westerman, *Preparing for utility computing: The role of IT architecture and relationship management*. IBM systems journal, 2004. 43(1): p. 5-19.
9. Alhamad, M., T. Dillon, and E. Chang. *Conceptual SLA framework for cloud computing*. in *4th IEEE International Conference on Digital Ecosystems and Technologies*. 2010. IEEE.
10. Liu, F., et al., *NIST cloud computing reference architecture*. NIST special publication, 2011. 500(2011): p. 292.
11. Wang, L., et al., *Cloud computing: a perspective study*. New Generation Computing, 2010. 28(2): p. 137-146.
12. Zhang, Q., L. Cheng, and R. Boutaba, *Cloud computing: state-of-the-art and research challenges*. Journal of internet services and applications, 2010. 1(1): p. 7-18.
13. Chiu, D., *Elasticity in the cloud*. XRDS: Crossroads, The ACM Magazine for Students, 2010. 16(3): p. 3-4.
14. Hutchins, D.C., *Just in time*. 1999: Gower Publishing, Ltd.
15. Lohr, S., *Google and IBM join in 'cloud computing'research*. New York Times, 2007. 8.
16. Bhadauria, R. and S. Sanyal, *Survey on security issues in cloud computing and associated mitigation techniques*. arXiv preprint arXiv:1204.0764, 2012.
17. Subashini, S. and V. Kavitha, *A survey on security issues in service delivery models of cloud computing*. Journal of network and computer applications, 2011. 34(1): p. 1-11.
18. Tripathi, S. and N. Jigeesh, *A review of factors that affect cloud computing adoption*. IUP Journal of Computer Sciences, 2013. 7(4): p. 48-59.
19. Ullah, S. and Z. Xuefeng, *Cloud Computing Research Challenges*. arXiv preprint arXiv:1304.3203, 2013.
20. Moreno-Vozmediano, R., R.S. Montero, and I.M. Llorente, *Key challenges in cloud computing: Enabling the future internet of services*. IEEE Internet Computing, 2012. 17(4): p. 18-25.
21. Buyya, R., A. Beloglazov, and J. Abawajy, *Energy-efficient management of data center resources for cloud computing: A vision*. Architectural Elements, and Open Challenges, 2010.
22. Buyya, R., S.K. Garg, and R.N. Calheiros. *SLA-oriented resource provisioning for cloud computing: Challenges, architecture, and solutions*. in *2011 international conference on cloud and service computing*. 2011. IEEE.
23. Davidsson, P. *Categories of artificial societies*. in *International Workshop on Engineering Societies in the Agents World*. 2001. Springer.
24. Rodríguez González, S., *Modelo adaptativo para Organizaciones virtuales de agentes*. 2010.

25. Zambonelli, F., N.R. Jennings, and M. Wooldridge, *Developing multiagent systems: The Gaia methodology*. ACM Transactions on Software Engineering and Methodology (TOSEM), 2003. 12(3): p. 317-370.
26. Mahjoub, M., et al. *A comparative study of the current cloud computing technologies and offers*. in *2011 First International Symposium on Network Cloud Computing and Applications*. 2011. IEEE.
27. Song, S., et al., *Energy profiling and analysis of the hpc challenge benchmarks*. The International Journal of High Performance Computing Applications, 2009. 23(3): p. 265-276.
28. Beloglazov, A., J. Abawajy, and R. Buyya, *Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing*. Future generation computer systems, 2012. 28(5): p. 755-768.
29. Gong, Z., X. Gu, and J. Wilkes. *Press: Predictive elastic resource scaling for cloud systems*. in *2010 International Conference on Network and Service Management*. 2010. Ieee.
30. Han, J., et al. *Survey on NoSQL database*. in *2011 6th international conference on pervasive computing and applications*. 2011. IEEE.
31. Wei, G., et al., *A game-theoretic method of fair resource allocation for cloud computing services*. The journal of supercomputing, 2010. 54(2): p. 252-269.
32. Miao, X. and J. Han. *The Design of a private cloud infrastructure based on Xen*. in *2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science*. 2011. IEEE.
33. Cardellini, V., M. Colajanni, and P.S. Yu, *Dynamic load balancing on web-server systems*. IEEE Internet computing, 1999. 3(3): p. 28-39.
34. Buyya, R., C.S. Yeo, and S. Venugopal. *Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities*. in *2008 10th IEEE international conference on high performance computing and communications*. 2008. Ieee.
35. You, X., et al. *RAS-M: resource allocation strategy based on market mechanism in cloud computing*. in *2009 Fourth ChinaGrid Annual Conference*. 2009. IEEE.
36. Van, H.N., F.D. Tran, and J.-M. Menaud. *SLA-aware virtual resource management for cloud infrastructures*. in *2009 Ninth IEEE International Conference on Computer and Information Technology*. 2009. IEEE.
37. Russell, S.J. and P. Norvig, *Artificial intelligence: a modern approach*. 2016: Malaysia; Pearson Education Limited.
38. Wooldridge, M. and N.R. Jennings, *Intelligent agents: Theory and practice*. The knowledge engineering review, 1995. 10(2): p. 115-152.
39. Zambonelli, F., et al. *Spray computers: Frontiers of self-organization for pervasive computing*. in *13th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*. 2004. IEEE.
40. Reitbauer, A., et al. *The MaBE middleware: extending multi-agent systems to enable open business collaboration*. in *6th IFIP International Conference on Information Technology for Balanced Automation Systems in Manufacturing and Services (BASYS'04)*. 2004.
41. Weyns, D., et al. *Role based model for adaptive agents*. in *Fourth Symposium on Adaptive Agents and Multiagent Systems at the AISB'04 Convention*. 2004.
42. Capera, D., et al. *Emergence of organisations, emergence of functions*. in *AISB'03 symposium on Adaptive Agents and Multi-Agent Systems*. 2003.
43. Razavi, R., J.-F. Perrot, and N. Guelfi. *Adaptive modeling: an approach and a method for implementing adaptive agents*. in *International Workshop on Massively Multiagent Systems*. 2004. Springer.
44. Dignum, M., *A model for organizational interaction: based on agents, founded in logic*. 2004: SIKS.
45. Schertler, W., *Virtual Enterprises in Tourism: Folklore and Facts—Conceptual Challenges for Academic Research—*, in *Information and Communication Technologies in Tourism 1998*. 1998, Springer. p. 278-288.
46. De Mantaras, R.L., et al., *Retrieval, reuse, revision and retention in case-based reasoning*. The Knowledge Engineering Review, 2005. 20(3): p. 215-240.
47. Aamodt, A. and E. Plaza, *Case-based reasoning: Foundational issues, methodological variations, and system approaches*. AI communications, 1994. 7(1): p. 39-59.
48. Reason, P. and H. Bradbury, *Handbook of action research: Participative inquiry and practice*. 2001: Sage.

49. Argente Villaplana, E., *Gormas: Guías para el desarrollo de sistemas multiagente abiertos basados en organizaciones*. 2008.
50. Argente, E., V. Botti, and V. Julian. *Gormas: An organizational-oriented methodological guideline for open mas*. in *International Workshop on Agent-Oriented Software Engineering*. 2009. Springer.
51. Grobauer, B., T. Walloschek, and E. Stocker, *Understanding cloud computing vulnerabilities.* IEEE Security & privacy, 2010. 9(2): p. 50-57.
52. Talia, D., *Clouds meet agents: Toward intelligent cloud services.* IEEE Internet Computing, 2012. 16(2): p. 78-81.
53. Kang, J. and K.M. Sim. *Ontology and search engine for cloud computing system*. in *Proceedings 2011 International Conference on System Science and Engineering*. 2011. IEEE.
54. Chellappa, R. *Intermediaries in cloud-computing: A new computing paradigm*. in *INFORMS Annual Meeting, Dallas*. 1997.
55. Weissman, C.D. and S. Bobrowski. *The design of the force. com multitenant internet application development platform*. in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. 2009.
56. Cusumano, M., *Cloud computing and SaaS as new computing platforms.* Communications of the ACM, 2010. 53(4): p. 27-29.
57. Carr, N.G., *The end of corporate computing.* MIT Sloan Management Review, 2005. 46(3): p. 67.
58. Staab, S., et al., *Web services: been there, done that?* IEEE Intelligent Systems, 2003. 18(1): p. 72-85.
59. Newcom, E. and G. Lomow, *Understanding SOA with Web Services (Independent Technology Guides).* 2004, Addison-Wesley Professional, White Plains.
60. Sotomayor, B., et al., *Virtual infrastructure management in private and hybrid clouds.* IEEE Internet computing, 2009. 13(5): p. 14-22.
61. Bote-Lorenzo, M.L., Y.A. Dimitriadis, and E. Gómez-Sánchez. *Grid characteristics and uses: a grid definition*. in *European Across Grids Conference*. 2003. Springer.
62. Horn, P., *Autonomic computing: IBM's perspective on the state of information technology.* 2001.
63. Ogrizović, D., B. Sviličić, and E. Tijan. *Open source science clouds*. in *The 33rd International Convention MIPRO*. 2010. IEEE.
64. Erdogmus, H., *Cloud computing: Does nirvana hide behind the nebula?* IEEE software, 2009. 26(2): p. 4-6.
65. Agrawal, D., S. Das, and A. El Abbadi. *Big data and cloud computing: current state and future opportunities*. in *Proceedings of the 14th International Conference on Extending Database Technology*. 2011.
66. Rappa, M.A., *The utility business model and the future of computing services.* IBM systems journal, 2004. 43(1): p. 32-42.
67. Broberg, J., S. Venugopal, and R. Buyya, *Market-oriented grids and utility computing: The state-of-the-art and future directions.* Journal of Grid Computing, 2008. 6(3): p. 255-276.
68. Kephart, J.O. and D.M. Chess, *The vision of autonomic computing.* Computer, 2003. 36(1): p. 41-50.
69. Parashar, M. and S. Hariri. *Autonomic computing: An overview*. in *International workshop on unconventional programming paradigms*. 2004. Springer.
70. Anderson, T., et al., *Overcoming the Internet impasse through virtualization.* Computer, 2005. 38(4): p. 34-41.
71. Barham, P., et al., *Xen and the art of virtualization.* ACM SIGOPS operating systems review, 2003. 37(5): p. 164-177.
72. Yamamoto, V.Y.O.V.T., *Server virtualization technology and its latest trends.* Fujitsu Sci. Tech. J, 2008. 44(1): p. 46-52.
73. McDougall, R. and J. Anderson, *Virtualization performance: perspectives and challenges ahead.* ACM SIGOPS Operating Systems Review, 2010. 44(4): p. 40-56.
74. Che, J., et al. *Performance measuring and comparing of virtual machine monitors*. in *2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing*. 2008. IEEE.
75. Che, J., et al. *A synthetical performance evaluation of openvz, xen and kvm*. in *2010 IEEE Asia-Pacific Services Computing Conference*. 2010. IEEE.
76. Chen, W., et al. *A novel hardware assisted full virtualization technique*. in *2008 The 9th International Conference for Young Computer Scientists*. 2008. IEEE.
77. Kivity, A. *KVM: The kernel-based virtual machine for Linux*. in *Proc. Linux Symposium, Ottawa, Canada, June 2007*. 2007.

78. Rosenblum, M. *Vmwares virtual platform*. in *Proceedings of hot chips*. 1999.
79. Mirkin, A., A. Kuznetsov, and K. Kolyshkin. *Containers checkpointing and live migration*. in *Proceedings of the Linux Symposium*. 2008.
80. Bianchini, R. and R. Rajamony, *Power and energy management for server systems*. Computer, 2004. 37(11): p. 68-76.
81. Petrucci, V., O. Loques, and D. Mossé. *Dynamic optimization of power and performance for virtualized server clusters*. in *Proceedings of the 2010 ACM Symposium on Applied Computing*. 2010.
82. Gray, J. and D.P. Siewiorek, *High-availability computer systems*. Computer, 1991. 24(9): p. 39-48.
83. Andrzejak, A., D. Kondo, and S. Yi. *Decision model for cloud computing under SLA constraints*. in *2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*. 2010. IEEE.
84. Tai, S., et al. *Cloud service engineering*. in *2010 ACM/IEEE 32nd International Conference on Software Engineering*. 2010. IEEE.
85. Dustdar, S. and W. Schreiner, *A survey on web services composition*. International journal of web and grid services, 2005. 1(1): p. 1-30.
86. Shenker, S., *Fundamental design issues for the future Internet*. IEEE Journal on selected areas in communications, 1995. 13(7): p. 1176-1188.
87. Fraternali, P., G. Rossi, and F. Sánchez-Figueroa, *Rich internet applications*. IEEE Internet Computing, 2010. 14(3): p. 9-12.
88. Tan, L. and N. Wang. *Future internet: The internet of things*. in *2010 3rd international conference on advanced computer theory and engineering (ICACTE)*. 2010. IEEE.
89. Vaquero-Gonzalez, L., et al., *A break in the clouds: Towards a cloud definition*. Computer Communication Rev, 2009. 39(1): p. 50-55.
90. Foster, I., et al. *Cloud computing and grid computing 360-degree compared*. in *2008 grid computing environments workshop*. 2008. Ieee.
91. Rapport, M. *Some Experts Say Cloud Computing is Nothing New*. 2010.
92. Miller, M., *Cloud computing: Web-based applications that change the way you work and collaborate online*. 2008: Que publishing.
93. Bragg, R., *Cloud computing: When computers really rule*. Tech News World, 2008. 12(12): p. 2009.
94. Sosinsky, B., *Cloud computing bible*. Vol. 762. 2010: John Wiley & Sons.
95. Zhang, F., et al. *Cloudvisor: retrofitting protection of virtual machines in multi-tenant cloud with nested virtualization*. in *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*. 2011.
96. Peltzman, S., *Toward a more general theory of regulation*. The Journal of Law and Economics, 1976. 19(2): p. 211-240.
97. Talia, D. *Cloud Computing and Software Agents: Towards Cloud Intelligent Services*. in *WOA*. 2011. Citeseer.
98. Rimal, B.P., E. Choi, and I. Lumb. *A taxonomy and survey of cloud computing systems*. in *2009 Fifth International Joint Conference on INC, IMS and IDC*. 2009. Ieee.
99. Jones, S., *Toward an acceptable definition of service [service-oriented architecture]*. IEEE software, 2005. 22(3): p. 87-93.
100. Lenk, A., et al. *What's inside the Cloud? An architectural map of the Cloud landscape*. in *2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing*. 2009. IEEE.
101. Wei, Y. and M.B. Blake, *Service-oriented computing and cloud computing: Challenges and opportunities*. IEEE Internet Computing, 2010. 14(6): p. 72-75.
102. Gutierrez-Garcia, J.O. and K.-M. Sim. *Self-organizing agents for service composition in cloud computing*. in *2010 IEEE Second International Conference on Cloud Computing Technology and Science*. 2010. IEEE.
103. Liu, L., et al. *GreenCloud: a new architecture for green data center*. in *Proceedings of the 6th international conference industry session on Autonomic computing and communications industry session*. 2009.
104. Heras, S., et al., *Agreement technologies and their use in cloud computing environments*. Progress in Artificial Intelligence, 2012. 1(4): p. 277-290.
105. A Vouk, M., *Cloud computing–issues, research and implementations*. Journal of computing and information technology, 2008. 16(4): p. 235-246.
106. Badger, M.L., et al., *Cloud computing synopsis and recommendations*. 2012: National Institute of Standards & Technology.

107. Krautheim, F.J. *Private Virtual Infrastructure for Cloud Computing*. in *HotCloud*. 2009.
108. Bezemer, C.-P., et al. *Enabling multi-tenancy: An industrial experience report*. in *2010 IEEE International Conference on Software Maintenance*. 2010. IEEE.
109. Pervez, Z., S. Lee, and Y.-K. Lee. *Multi-tenant, secure, load disseminated SaaS architecture*. in *2010 The 12th International Conference on Advanced Communication Technology (ICACT)*. 2010. IEEE.
110. Kang, S., S. Kang, and S. Hur. *A design of the conceptual architecture for a multitenant saas application platform*. in *2011 First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering*. 2011. IEEE.
111. Azeez, A., et al. *Multi-tenant SOA middleware for cloud computing*. in *2010 IEEE 3rd international conference on cloud computing*. 2010. IEEE.
112. Mietzner, R., F. Leymann, and T. Unger, *Horizontal and vertical combination of multi-tenancy patterns in service-oriented applications*. Enterprise Information Systems, 2011. 5(1): p. 59-77.
113. Ortiz Jr, S., *The problem with cloud-computing standardization*. Computer, 2011(7): p. 13-16.
114. Celesti, A., et al. *How to enhance cloud architectures to enable cross-federation*. in *2010 IEEE 3rd international conference on cloud computing*. 2010. IEEE.
115. Machado, G.S., D. Hausheer, and B. Stiller. *Considerations on the Interoperability of and between Cloud Computing Standards*. in *27th open grid forum (OGF27), G2C-Net workshop: from grid to cloud networks*. 2009.
116. Rochwerger, B., et al., *The reservoir model and architecture for open federated cloud computing*. IBM Journal of Research and Development, 2009. 53(4): p. 4: 1-4: 11.
117. Loutas, N., et al. *Cloud computing interoperability: the state of play*. in *2011 IEEE Third International Conference on Cloud Computing Technology and Science*. 2011. IEEE.
118. Petcu, D. *Portability and interoperability between clouds: challenges and case study*. in *European Conference on a Service-Based Internet*. 2011. Springer.
119. Chahal, S., et al., *An enterprise private cloud architecture and implementation roadmap*. IT@ Intel White Paper, USA, 2010.
120. Bakshi, K., *Cisco cloud computing-data center strategy, architecture, and solutions*. CISCO White Paper. Retrieved October, 2009. 13: p. 2010.
121. Tianfield, H. *Cloud computing architectures*. in *2011 IEEE International Conference on Systems, Man, and Cybernetics*. 2011. IEEE.
122. Voras, I., et al. *Evaluating open-source cloud computing solutions*. in *2011 Proceedings of the 34th International Convention MIPRO*. 2011. IEEE.
123. Sempolinski, P. and D. Thain. *A comparison and critique of eucalyptus, opennebula and nimbus*. in *2010 IEEE Second International Conference on Cloud Computing Technology and Science*. 2010. Ieee.
124. Halinka, T., *The openQRM users guide*. 2008.
125. Keahey, K., *Cloud computing for science*. 2010.
126. Sabharwal, N. and R. Shankar, *Apache CloudStack cloud computing*. 2013: Packt Publishing.
127. Cohen, B., *PaaS: New opportunities for cloud application development*. Computer, 2013(9): p. 97-100.
128. Sefraoui, O., M. Aissaoui, and M. Eleuldj, *OpenStack: toward an open-source solution for cloud computing*. International Journal of Computer Applications, 2012. 55(3): p. 38-42.
129. Milojičić, D., I.M. Llorente, and R.S. Montero, *Opennebula: A cloud management tool*. IEEE Internet Computing, 2011. 15(2): p. 11-14.
130. Hamdi, M. *Security of cloud computing, storage, and networking*. in *2012 International Conference on Collaboration Technologies and Systems (CTS)*. 2012. IEEE.
131. Baker, W., et al., *2011 data breach investigations report*. Verizon RISK Team, Available: www. verizonbusiness. com/resources/reports/rp_databreach-investigationsreport-2011_en_xg. pdf, 2011: p. 1-72.
132. Kamara, S. and K. Lauter. *Cryptographic cloud storage*. in *International Conference on Financial Cryptography and Data Security*. 2010. Springer.
133. Reuben, J.S., *A survey on virtual machine security*. Helsinki University of Technology, 2007. 2(36).
134. Pal, S., et al., *A new trusted and collaborative agent based approach for ensuring cloud security*. arXiv preprint arXiv:1108.4100, 2011.
135. De Clercq, J. *Single sign-on architectures*. in *International Conference on Infrastructure Security*. 2002. Springer.

136. Alliance, C., *Security guidance for critical areas of focus in cloud computing v3. 0*. Cloud Security Alliance, 2011. 15.
137. Smith, D., et al., *Predicts 2013: cloud computing becomes an integral part of IT*. Gartner, ID G, 2012. 230929.
138. Desisto, R., *The Top Three Impacts of Cloud Computing on Sales and Business Applications*. 2013.
139. Buyya, R., R. Ranjan, and R.N. Calheiros. *Intercloud: Utility-oriented federation of cloud computing environments for scaling of application services*. in *International Conference on Algorithms and Architectures for Parallel Processing*. 2010. Springer.
140. Sim, K.M., *Cloud intelligence: Agents within an intercloud*. Awareness Magazine. The Official Magazine for Future and Emerging Technologies Proactive Initiative, 2013.
141. McKendrick, J., *NIST definition of cloud computing doesn't go far enough*. ZDNET. http://www.zdnet.com/blog/service-oriented/nist-definition-of-cloudcomputing-doesnt-go-far-enough/8634, 2012.
142. Low, C., Y. Chen, and M. Wu, *Understanding the determinants of cloud computing adoption*. Industrial management & data systems, 2011.
143. Azodolmolky, S., P. Wieder, and R. Yahyapour, *Cloud computing networking: Challenges and opportunities for innovations*. IEEE Communications Magazine, 2013. 51(7): p. 54-62.
144. Boehm, B.W., *Software engineering economics*. IEEE transactions on Software Engineering, 1984(1): p. 4-21.
145. Pressman, R.S., *Software engineering: a practitioner's approach*. 2005: Palgrave macmillan.
146. d'Inverno, M., M. Luck, and M.M. Luck, *Understanding agent systems*. 2004: Springer Science & Business Media.
147. Kolp, M., P. Giorgini, and J. Mylopoulos, *Multi-agent architectures as organizational structures*. Autonomous Agents and Multi-Agent Systems, 2006. 13(1): p. 3-25.
148. Labidi, S. and W. Lejouad, *De l'intelligence artificielle distribuée aux systèmes multi-agents*. 1993.
149. Wooldridge, M., *An introduction to multiagent systems*. 2009: John Wiley & Sons.
150. Jennings, N.R. and M. Wooldridge, *Applications of intelligent agents*, in *Agent technology*. 1998, Springer. p. 3-28.
151. Bond, A.H. and L. Gasser, *Readings in distributed artificial intelligence*. 2014: Morgan Kaufmann.
152. Jennings, N.R., *An agent-based approach for building complex software systems*. Communications of the ACM, 2001. 44(4): p. 35-41.
153. Jennings, N.R., J.M. Corera, and I. Laresgoiti. *Developing Industrial Multi-Agent Systems*. in *ICMAS*. 1995.
154. Wooldridge, M., S. Bussmann, and M. Klosterberg. *Production sequencing as negotiation*. in *Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM-96)*. 1996.
155. Nwana, H.S., *Software agents: An overview*. The knowledge engineering review, 1996. 11(3): p. 205-244.
156. Brenner, W., R. Zarnekow, and H. Wittig, *Intelligent software agents: foundations and applications*. 2012: Springer Science & Business Media.
157. Franklin, S. and A. Graesser. *Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents*. in *International Workshop on Agent Theories, Architectures, and Languages*. 1996. Springer.
158. Cvetković, D. and I. Parmee, *Agent-based support within an interactive evolutionary design system*. AI EDAM, 2002. 16(5): p. 331-342.
159. Maes, P., *Designing autonomous agents: Theory and practice from biology to engineering and back*. 1990: MIT press.
160. Rao, A.S. and M.P. Georgeff, *An abstract architecture for rational agents*. KR, 1992. 92: p. 439-449.
161. Georgeff, M.P. and A.L. Lansky. *Reactive reasoning and planning*. in *AAAI*. 1987.
162. Bratman, M.E., D.J. Israel, and M.E. Pollack, *Plans and resource-bounded practical reasoning*. Computational intelligence, 1988. 4(3): p. 349-355.
163. Bratman, M., *Intention, plans, and practical reason*. Vol. 10. 1987: Harvard University Press Cambridge, MA.
164. Mas, A., *Agentes software y sistemas multiagente: conceptos, arquitecturas y aplicaciones*. 2005: Prentice Hall.

165. Rahmani, A., et al., *Controllability of multi-agent systems from a graph-theoretic perspective.* SIAM Journal on Control and Optimization, 2009. 48(1): p. 162-186.
166. Kautz, H.A. *Deconstructing planning as satisfiability.* in *Proceedings of the National Conference on Artificial Intelligence.* 2006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
167. Jensen, R.M. and M.M. Veloso, *OBDD-based universal planning for synchronized agents in non-deterministic domains.* Journal of Artificial Intelligence Research, 2000. 13: p. 189-226.
168. Cimatti, A. and M. Roveri, *Conformant planning via symbolic model checking.* Journal of Artificial Intelligence Research, 2000. 13: p. 305-338.
169. Szer, D., F. Charpillet, and S. Zilberstein, *MAA*: A heuristic search algorithm for solving decentralized POMDPs.* arXiv preprint arXiv:1207.1359, 2012.
170. Shen, W., L. Wang, and Q. Hao, *Agent-based distributed manufacturing process planning and scheduling: a state-of-the-art survey.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2006. 36(4): p. 563-577.
171. De Paz, J.F., et al., *Mathematical model for dynamic case-based planning.* International Journal of Computer Mathematics, 2009. 86(10-11): p. 1719-1730.
172. Tesauro, G., et al. *A multi-agent systems approach to autonomic computing.* in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1.* 2004. IEEE Computer Society.
173. Bousquet, F., et al., *Multi-agent systems and role games: collective learning processes for ecosystem management.* Complexity and ecosystem management: The theory and practice of multi-agent systems, 2002: p. 248-285.
174. Horling, B., et al. *Diagnosis as an integral part of multi-agent adaptability.* in *Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00.* 2000. IEEE.
175. Omicini, A., et al., *Coordination of Internet agents: Models, technologies, and applications.* 2013: Springer Science & Business Media.
176. Ramchurn, S.D., D. Huynh, and N.R. Jennings, *Trust in multi-agent systems.* The Knowledge Engineering Review, 2004. 19(1): p. 1-25.
177. Sabater, J., et al., *Engineering Executable Agents using Multi-context Systems.* Journal of Logic and Computation, 2002. 12(3): p. 413-442.
178. Bauer, B., J.P. Müller, and J. Odell, *Agent UML: A formalism for specifying multiagent software systems.* International journal of software engineering and knowledge engineering, 2001. 11(03): p. 207-230.
179. Sim, K.M., *Agent-based cloud computing.* IEEE transactions on services computing, 2011. 5(4): p. 564-577.
180. Venticinque, S., et al. *A cloud agency for SLA negotiation and management.* in *European Conference on Parallel Processing.* 2010. Springer.
181. Li, Z., C. Chen, and K. Wang, *Cloud computing for agent-based urban transportation systems.* IEEE Intelligent Systems, 2011. 26(1): p. 73-79.
182. De la Prieta, F., et al. *An enhanced approach to retrieve learning resources over the cloud.* in *The 2nd International Workshop on Learning Technology for Education in Cloud.* 2014. Springer.
183. Oberheide, J., et al. *Virtualized in-cloud security services for mobile devices.* in *Proceedings of the first workshop on virtualization in mobile computing.* 2008.
184. Manyika, J., *Big data: The next frontier for innovation, competition, and productivity.* http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation, 2011.
185. Russom, P., *Big data analytics.* TDWI best practices report, fourth quarter, 2011. 19(4): p. 1-34.
186. Amato, A., B. Di Martino, and S. Venticinque. *Agent-based decision support for smart market using big data.* in *International Conference on Algorithms and Architectures for Parallel Processing.* 2013. Springer.
187. Cristelli, M., A. Tacchella, and L. Pietronero, *An overview of the new frontiers of economic complexity*, in *Econophysics of Agent-Based Models.* 2014, Springer. p. 147-159.

188. Surendran, R. and B.P. Varthini. *Intelligent based large scale multi agent's resource management on shopping service with security*. in *2012 Nirma University International Conference on Engineering (NUiCONE)*. 2012. IEEE.
189. Blasch, E., et al., *Information fusion in a cloud-enabled environment*, in *High Performance Cloud Auditing and Applications*. 2014, Springer. p. 91-115.
190. Chamorro, J.M.C. and R.S. Montes, *An agent-based approach for data fusion in homeland security*. IJIMAI, 2013. 2(3): p. 44-49.
191. Taboada, M., et al., *Using an agent-based simulation for predicting the effects of patients derivation policies in emergency departments*. Procedia Computer Science, 2013. 18: p. 641-650.
192. Decraene, J., et al. *Evolving agent-based simulations in the clouds*. in *Third international workshop on advanced computational intelligence*. 2010. IEEE.
193. Decraene, J., et al. *Evolutionary design of agent-based simulation experiments*. in *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. 2011. Citeseer.
194. Sethia, P. and K. Karlapalem, *A multi-agent simulation framework on small Hadoop cluster*. Engineering Applications of Artificial Intelligence, 2011. 24(7): p. 1120-1127.
195. Chen, D., et al., *Hybrid modelling and simulation of huge crowd over a hierarchical grid architecture*. Future Generation Computer Systems, 2013. 29(5): p. 1309-1317.
196. Haghighi, M. *An agent-based multi-model tool for simulating multiple concurrent applications in WSNs*. in *Journal of Advances in Computer Networks (JACN), 5th International Conference on Communication Software and Networks*. 2013.
197. Aversa, R., et al. *Cloud agency: A mobile agent based cloud system*. in *2010 international conference on complex, intelligent and software intensive systems*. 2010. IEEE.
198. Yun, Y., Y. Li, and H. Han. *A Mobile Agent-Based Secure and Efficient Task Allocation Algorithm for Cloud&Client Computing*. in *2012 Fourth International Conference on Multimedia Information Networking and Security*. 2012. IEEE.
199. Othmane, B. and R.S.A. Hebri. *Cloud computing & multi-agent systems: A new promising approach for distributed data mining*. in *Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces*. 2012. IEEE.
200. Bajo, J., et al., *Cloud computing in bioinformatics*, in *Distributed Computing and Artificial Intelligence*. 2010, Springer. p. 147-155.
201. Wang, K. and Z. Shen, *Artificial societies and GPU-based cloud computing for intelligent transportation management*. IEEE Intelligent Systems, 2011. 26(4): p. 22-28.
202. Krol, D., et al., *Elastic infrastructure for interactive data farming experiments*. Procedia Computer Science, 2012. 9: p. 206-215.
203. Gutierrez-Garcia, J.O. and K.M. Sim, *A family of heuristics for agent-based elastic cloud bag-of-tasks concurrent scheduling*. Future Generation Computer Systems, 2013. 29(7): p. 1682-1699.
204. Hübner, J.F., J.S. Sichman, and O. Boissier. *A model for the structural, functional, and deontic specification of organizations in multiagent systems*. in *Brazilian Symposium on Artificial Intelligence*. 2002. Springer.
205. Guzek, M., G. Danoy, and P. Bouvry. *Paramoise: Increasing capabilities of parallel execution and reorganization in an organizational model*. in *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems, AAMAS'13*. 2013.
206. Guzek, M., G. Danoy, and P. Bouvry. *System design and implementation decisions for paramoise organisational model*. in *2013 Federated Conference on Computer Science and Information Systems*. 2013. IEEE.
207. Wittek, P. and X. Rubio-Campillo. *Scalable agent-based modelling with cloud HPC resources for social simulations*. in *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*. 2012. IEEE.
208. Leitão, P., U. Inden, and C.-P. Rückemann. *Parallelising multi-agent systems for high performance computing*. in *Third International Conference on Advanced Communications and Computation (INFOCOMP'13)*. 2013. IARIA.
209. Collier, N. and M. North, *Repast HPC: A platform for large-scale agent-based modeling*. Large-Scale Computing, 2012: p. 81-109.

210. Nouman, A., A. Anagnostou, and S.J. Taylor. *Developing a distributed agent-based and des simulation using portico and repast.* in *2013 IEEE/ACM 17th International Symposium on Distributed Simulation and Real Time Applications*. 2013. IEEE.
211. Minson, R. and G.K. Theodoropoulos, *Distributing RePast agent-based simulations with HLA.* Concurrency and Computation: Practice and Experience, 2008. 20(10): p. 1225-1256.
212. Tapia, D.I., et al. *Cloud-IO: cloud computing platform for the fast deployment of services over wireless sensor networks*. in *7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing*. 2013. Springer.
213. Yuriyama, M. and T. Kushida. *Sensor-cloud infrastructure-physical sensor management with virtualized sensors on cloud computing*. in *2010 13th International Conference on Network-Based Information Systems*. 2010. IEEE.
214. Hassan, M.M., B. Song, and E.-N. Huh. *A framework of sensor-cloud integration opportunities and challenges*. in *Proceedings of the 3rd international conference on Ubiquitous information management and communication*. 2009.
215. Cuzzocrea, A., G. Fortino, and O. Rana. *Managing data and processes in cloud-enabled large-scale sensor networks: state-of-the-art and future research directions*. in *2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*. 2013. IEEE.
216. Baktashmotlagh, M., A. Bigdeli, and B.C. Lovell. *Dynamic resource aware sensor networks: Integration of sensor cloud and ERPs*. in *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2011. IEEE.
217. Heydon, A. and M. Najork, *Mercator: A scalable, extensible web crawler.* World Wide Web, 1999. 2(4): p. 219-229.
218. Alhamad, M., T. Dillon, and E. Chang. *Sla-based trust model for cloud computing*. in *2010 13th international conference on network-based information systems*. 2010. IEEE.
219. More, A., S. Vij, and D. Mukhopadhyay. *Agent based negotiation using cloud–an approach in E-commerce*. in *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol I*. 2014. Springer.
220. Habib, S.M., S. Ries, and M. Muhlhauser. *Towards a trust management system for cloud computing*. in *2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*. 2011. IEEE.
221. Dasgupta, P. and I. Serageldin, *Social capital: a multifaceted perspective.* 1999: The World Bank.
222. Mouratidis, H., et al., *A framework to support selection of cloud providers based on security and privacy requirements.* Journal of Systems and Software, 2013. 86(9): p. 2276-2293.
223. An, B., et al. *Automated negotiation with decommitment for dynamic resource allocation in cloud computing.* in *AAMAS*. 2010.
224. Sim, K.M. *Towards complex negotiation for cloud economy*. in *International Conference on Grid and Pervasive Computing*. 2010. Springer.
225. Sim, K.M., *Complex and concurrent negotiations for multiple interrelated e-markets.* IEEE transactions on cybernetics, 2012. 43(1): p. 230-245.
226. Lai, Y.-L., *Analyzing strategies of mobile agents on malicious cloud platform with Agent-Based Computational Economic Approach.* Expert systems with applications, 2013. 40(7): p. 2615-2620.
227. Li, B.-H., et al., *Further discussion on cloud manufacturing.* Computer integrated manufacturing systems, 2011. 17(3): p. 449-457.
228. Tao, F., et al., *Cloud manufacturing: a computing and service-oriented manufacturing model.* Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 2011. 225(10): p. 1969-1976.
229. Xu, X., *From cloud computing to cloud manufacturing.* Robotics and computer-integrated manufacturing, 2012. 28(1): p. 75-86.
230. Hamadi, R. and B. Benatallah. *A petri net-based model for web service composition*. in *Proceedings of the 14th Australasian database conference-Volume 17*. 2003. Australian Computer Society, Inc.
231. Celesti, A., et al. *Three-phase cross-cloud federation model: The cloud sso authentication*. in *2010 Second International Conference on Advances in Future Internet*. 2010. IEEE.
232. Singh, A. and M. Malhotra, *Agent based framework for scalability in cloud computing.* International Journal of Computer Science & Engineering Technology (IJCSET), 2012. 3(4): p. 41-45.

233. Chen, G., et al. *Saaas-the mobile agent based service for cloud computing in internet environment*. in *2010 Sixth International Conference on Natural Computation*. 2010. IEEE.
234. Venticinque, S., L. Tasquier, and B. Di Martino. *Agents based cloud computing interface for resource provisioning and management*. in *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*. 2012. IEEE.
235. Calheiros, R.N., et al., *A coordinator for scaling elastic applications across multiple clouds*. Future Generation Computer Systems, 2012. 28(8): p. 1350-1362.
236. Fan, C.-T., W.-J. Wang, and Y.-S. Chang. *Agent-based service migration framework in hybrid cloud*. in *2011 IEEE International Conference on High Performance Computing and Communications*. 2011. IEEE.
237. Palmieri, F., et al., *A distributed scheduling framework based on selfish autonomous agents for federated cloud environments*. Future Generation Computer Systems, 2013. 29(6): p. 1461-1472.
238. Ejarque, J., R. Sirvent, and R.M. Badia. *A multi-agent approach for semantic resource allocation*. in *2010 IEEE Second International Conference on Cloud Computing Technology and Science*. 2010. IEEE.
239. Venkataramana, K. and M. Padmavathamma, *Agent based approach for authentication in cloud*. IRACST-International Journal of Computer Science and Information Technology & Security, 2012. 2(3): p. 598-603.
240. Hajivali, M., et al. *Applying an agent-based user authentication and access control model for cloud servers*. in *2013 International Conference on ICT Convergence (ICTC)*. 2013. IEEE.
241. Habiba, M., M.R. Islam, and A.S. Ali. *Access control management for cloud*. in *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. 2013. IEEE.
242. Damiani, E. and F. Pagano, *Handling confidential data on the untrusted cloud: an agent-based approach*. arXiv preprint arXiv:1012.0759, 2010.
243. Li, J., et al., *CyberGuarder: A virtualization security assurance architecture for green cloud computing*. Future Generation Computer Systems, 2012. 28(2): p. 379-390.
244. Meera, A. and S. Swamynathan, *Agent based resource monitoring system in IaaS cloud environment*. Procedia Technology, 2013. 10: p. 200-207.
245. Talib, A.M., et al. *Security framework of cloud data storage based on Multi Agent system architecture-A pilot study*. in *2012 International Conference on Information Retrieval & Knowledge Management*. 2012. IEEE.
246. Talib, A.M., et al., *Towards a comprehensive security framework of cloud data storage based on multi agent system architecture*. Journal of Information Security, 2012. 3(04): p. 295.
247. Talib, A.M., et al., *Security facilitation in collaborative cloud data storage implementation environment based on multi agent system architecture*. J. Software Eng, 2012. 6: p. 49-64.
248. Talib, A.M., et al. *CloudZone: Towards an integrity layer of cloud data storage based on multi agent system architecture*. in *2011 IEEE Conference on Open Systems*. 2011. IEEE.
249. Talib, A.M., et al. *Towards new data access control technique based on multi agent system architecture for cloud computing*. in *International Conference on Digital Information Processing and Communications*. 2011. Springer.
250. Talib, A.M., et al., *Security framework of cloud data storage based on multi agent system architecture: Semantic literature review*. Computer and Information Science, 2010. 3(4): p. 175.
251. Talib, A.M., et al., *Multi agent system architecture oriented prometheus methodology design to facilitate security of cloud data storage*. Journal of Software Engineering, 2011. 5(3): p. 78-90.
252. Govinda, K. and E. Sathiyamoorthy, *Agent based security for cloud computing using obfuscation*. Procedia Engineering, 2012. 38: p. 125-129.
253. Islam, M.R. and M. Habiba. *Collaborative swarm intelligence based Trusted Computing*. in *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*. 2012. IEEE.
254. Islam, M.R. and M. Habiba. *Agent based framework for providing security to data storage in cloud*. in *2012 15th International Conference on Computer and Information Technology (ICCIT)*. 2012. IEEE.
255. Chao, P.-C. and H.-M. Sun, *Multi-agent-based cloud utilization for the IT office-aid asset distribution chain: An empirical case study*. Information Sciences, 2013. 245: p. 255-275.
256. Yang, S.-Y., *A novel cloud information agent system with Web service techniques: Example of an energy-saving multi-agent system*. Expert Systems with Applications, 2013. 40(5): p. 1758-1785.
257. Yang, S.-Y., et al. *Energy-saving information multi-agent system with web services for cloud computing*. in *International Conference on Security-Enriched Urban Computing and Smart Grid*. 2011. Springer.

258. Babu, S.R., K.G. Kulkarni, and K.C. Sekaran. *A generic agent based cloud computing architecture for e-learning*. in *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol I*. 2014. Springer.
259. Kalagiakos, P. and P. Karampelas. *Cloud computing learning*. in *2011 5th international conference on application of information and communication technologies (AICT)*. 2011. Ieee.
260. Lopez-Rodriguez, I. and M. Hernandez-Tejera, *Software agents as cloud computing services*, in *Advances on practical applications of agents and multiagent systems*. 2011, Springer. p. 271-276.
261. Zeng, L., et al., *QoS-aware middleware for web services composition*. IEEE Transactions on software engineering, 2004. 30(5): p. 311-327.
262. Wei, L., et al. *Multi-agent based QoS-aware service composition*. in *2010 IEEE International Conference on Systems, Man and Cybernetics*. 2010. IEEE.
263. Wei, Y. and M.B. Blake. *An agent-based services framework with adaptive monitoring in cloud environments*. in *2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*. 2012. IEEE.
264. Li, W., et al., *Resource virtualization and service selection in cloud logistics*. Journal of Network and Computer Applications, 2013. 36(6): p. 1696-1704.
265. Wei, Y. and M.B. Blake. *Adaptive service workflow configuration and agent-based virtual resource management in the cloud*. in *2013 IEEE International Conference on Cloud Engineering (IC2E)*. 2013. IEEE.
266. Núñez, A., C. Andrés, and M.G. Merayo. *Mascloud: a framework based on multi-agent systems for optimizing cost in cloud computing*. in *International Conference on Computational Collective Intelligence*. 2012. Springer.
267. Emeakaroha, V.C., et al., *Towards autonomic detection of SLA violations in Cloud infrastructures*. Future Generation Computer Systems, 2012. 28(7): p. 1017-1029.
268. Ramaswamy, A., et al. *A Mobile Agent based Approach of ensuring Trustworthiness in the Cloud*. in *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*. 2011. IEEE.
269. Rodríguez, S., et al. *Ovamah: Multiagent-based adaptive virtual organizations*. in *2009 12th International Conference on Information Fusion*. 2009. IEEE.
270. Hannoun, M., et al., *MOISE: An organizational model for multi-agent systems*, in *Advances in Artificial Intelligence*. 2000, Springer. p. 156-165.
271. Ferber, J., O. Gutknecht, and F. Michel. *From agents to organizations: an organizational view of multi-agent systems*. in *International workshop on agent-oriented software engineering*. 2003. Springer.
272. Claus, C. and C. Boutilier, *The dynamics of reinforcement learning in cooperative multiagent systems*. AAAI/IAAI, 1998. 1998(746-752): p. 2.
273. Stone, P. and M. Veloso, *Multiagent systems: A survey from a machine learning perspective*. Autonomous Robots, 2000. 8(3): p. 345-383.
274. Woods, S.G. and M.R. Barbacci, *Architectural evaluation of collaborative agent-based systems*. 1999, CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST.
275. Ciancarini, P., A. Omicini, and F. Zambonelli. *Multiagent system engineering: The coordination viewpoint*. in *International Workshop on Agent Theories, Architectures, and Languages*. 1999. Springer.
276. Annunziato, M. and P. Pierucci. *The emergence of social learning in artificial societies*. in *Workshops on Applications of Evolutionary Computation*. 2003. Springer.
277. Garcia-Ojeda, J.C., et al. *O-MaSE: a customizable approach to developing multiagent development processes*. in *International Workshop on Agent-Oriented Software Engineering*. 2007. Springer.
278. Ricci, A., M. Viroli, and A. Omicini. *Give agents their artifacts: the A&A approach for engineering working environments in MAS*. in *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. 2007.
279. Dignum, V., et al. *An organizational-oriented model for agent societies*. in *Proc. Int. Workshop on Regulated Agent-Based Social Systems: Theories and Applications (RASTA'02), at AAMAS, Bologna, Italy*. 2002.
280. Giret, A., V. Botti, and S. Valero. *MAS methodology for HMS*. in *International Conference on Industrial Applications of Holonic and Multi-Agent Systems*. 2005. Springer.
281. Sims, M., D. Corkill, and V. Lesser, *Automated organization design for multi-agent systems*. Autonomous agents and multi-agent systems, 2008. 16(2): p. 151-185.

282. Omicini, A., A. Ricci, and M. Viroli, *Artifacts in the A&A meta-model for multi-agent systems.* Autonomous agents and multi-agent systems, 2008. 17(3): p. 432-456.
283. DeLoach, S.A. and J.C. Garcia-Ojeda, *O-MaSE: a customisable approach to designing and building complex, adaptive multi-agent systems.* International Journal of Agent-Oriented Software Engineering, 2010. 4(3): p. 244-280.
284. Dignum, M., *A landscape of agent systems in the real world.* 2006, UU WINFI Informatica en Informatiekunde.
285. Mintzberg, H., *Structure in fives: Designing effective organizations*. 1993: Prentice-Hall, Inc.
286. Morabito, J., I. Sack, and A. Bhate, *Organization modeling: innovative architectures for the 21st century*. 1999: Prentice Hall.
287. Gomes-Casseres, B., *The alliance revolution: The new shape of business rivalry*. 1996: Harvard University Press.
288. Dussauge, P., B. Garrette, and C. Prahalad, *Cooperative strategy: Competing successfully through strategic alliances*. 1999: John Wiley Chichester.
289. Daft, R.L., J. Murphy, and H. Willmott, *Organization theory and design*. 2010: Cengage learning EMEA.
290. Artikis, A., L. Kamara, and J. Pitt. *Towards an open agent society model and animation*. in *Proceedings of the Agent-Based Simulation II workshop, Passau*. 2001.
291. Castelfranchi, C. *Engineering social order*. in *International Workshop on Engineering Societies in the Agents World*. 2000. Springer.
292. Zambonelli, F., N.R. Jennings, and M. Wooldridge. *Organisational abstractions for the analysis and design of multi-agent systems*. in *International Workshop on Agent-Oriented Software Engineering*. 2000. Springer.
293. Parunak, H.V.D., R. Savit, and R.L. Riolo. *Agent-based modeling vs. equation-based modeling: A case study and users' guide*. in *International Workshop on Multi-Agent Systems and Agent-Based Simulation*. 1998. Springer.
294. Gilbert, N., *Simulation: A new way of doing social science.* American Behavioral Scientist, 1999. 42(10): p. 1485-1487.
295. Davidsson, P. and S. Johansson, *On the potential of norm-governed behavior in different categories of artificial societies.* Computational & Mathematical Organization Theory, 2006. 12(2-3): p. 169-180.
296. Gasser, L., *An overview of DAI.* Distributed Artificial Intelligence: Theory and Praxis, 1992. 9(9-29): p. 28.
297. Wooldridge, M., N.R. Jennings, and D. Kinny, *The Gaia methodology for agent-oriented analysis and design.* Autonomous Agents and multi-agent systems, 2000. 3(3): p. 285-312.
298. Horling, B. and V. Lesser, *A survey of multi-agent organizational paradigms.* The Knowledge engineering review, 2004. 19(4): p. 281-316.
299. DeLoach, S.A., *OMACS: A framework for adaptive, complex systems*, in *Handbook of research on multi-agent systems: Semantics and dynamics of organizational models.* 2009, IGI Global. p. 76-104.
300. Bonjean, N., et al. *Metamodel-based metrics for agent-oriented methodologies*. in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 2012.
301. Iglesias, C.A., et al. *Analysis and design of multiagent systems using MAS-CommonKADS*. in *International Workshop on Agent Theories, Architectures, and Languages*. 1997. Springer.
302. Schreiber, A.T., et al., *Knowledge engineering and management: the CommonKADS methodology*. 2000: MIT press.
303. Caire, G., et al. *Agent oriented analysis using MESSAGE/UML*. in *International Workshop on Agent-Oriented Software Engineering*. 2001. Springer.
304. Moraitis, P. and N. Spanoudakis, *The Gaia2Jade process for multi-agent systems development.* Applied Artificial Intelligence, 2006. 20(2-4): p. 251-273.
305. Moraitis, P. and N. Spanoudakis, *Combining gaia and jade for multi-agent systems development*. 2004: na.
306. Spanoudakis, N. and P. Moraitis. *Gaia agents implementation through models transformation*. in *International Conference on Principles and Practice of Multi-Agent Systems*. 2009. Springer.
307. Carrascosa, C., et al., *Hybrid multi-agent architecture as a real-time problem-solving model.* Expert Systems with Applications, 2008. 34(1): p. 2-17.

308. Cernuzzi, L. and F. Zambonelli. *Dealing with adaptive multi-agent organizations in the gaia methodology*. in *International Workshop on Agent-Oriented Software Engineering*. 2005. Springer.
309. Juan, T., A. Pearce, and L. Sterling. *ROADMAP: extending the gaia methodology for complex open systems*. in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*. 2002.
310. Juan, T. and L. Sterling. *The ROADMAP meta-model for intelligent adaptive multi-agent systems in open environments*. in *International Workshop on Agent-Oriented Software Engineering*. 2003. Springer.
311. Padgham, L. and M. Winikoff. *Prometheus: A methodology for developing intelligent agents*. in *International Workshop on Agent-Oriented Software Engineering*. 2002. Springer.
312. Pavón, J. and J. Gómez-Sanz. *Agent oriented software engineering with INGENIAS*. in *International Central and Eastern European Conference on Multi-Agent Systems*. 2003. Springer.
313. Gascuena, J.M. and A. Fernández-Caballero. *Prometheus and INGENIAS agent methodologies: A complementary approach*. in *International Workshop on Agent-Oriented Software Engineering*. 2008. Springer.
314. Ferber, J. and O. Gutknecht. *A meta-model for the analysis and design of organizations in multi-agent systems*. in *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*. 1998. IEEE.
315. Ferber, J., F. Michel, and J. Báez. *AGRE: Integrating environments with organizations*. in *International Workshop on Environments for Multi-Agent Systems*. 2004. Springer.
316. Omicini, A. *SODA: Societies and infrastructures in the analysis and design of agent-based systems*. in *International Workshop on Agent-Oriented Software Engineering*. 2000. Springer.
317. Molesini, A., et al. *Advancing object-oriented standards toward agent-oriented methodologies: SPEM 2.0 on SODA*. in *9th Workshop "From Objects to Agents"(WOA 2008)–Evolution of Agent Development: Methodologies, Tools, Platforms and Languages*. 2008.
318. OMG, S. and O. Notation, *Software & systems process engineering meta-model specification*. OMG Std., Rev, 2008. 2: p. 18-71.
319. Hübner, J.F., et al., *Instrumenting multi-agent organisations with organisational artifacts and agents*. Autonomous agents and multi-agent systems, 2010. 20(3): p. 369-400.
320. Ricci, A., M. Viroli, and A. Omicini. *Programming MAS with artifacts*. in *International Workshop on Programming Multi-Agent Systems*. 2005. Springer.
321. Giunchiglia, F., J. Mylopoulos, and A. Perini. *The tropos software development methodology: processes, models and diagrams*. in *International Workshop on Agent-Oriented Software Engineering*. 2002. Springer.
322. Bresciani, P., et al., *Tropos: An agent-oriented software development methodology*. Autonomous Agents and Multi-Agent Systems, 2004. 8(3): p. 203-236.
323. Yu, E., *Modelling strategic relationships for process reengineering*. Social Modeling for Requirements Engineering, 2011. 11: p. 2011.
324. Vázquez-Salceda, J., V. Dignum, and F. Dignum, *Organizing multiagent systems*. Autonomous Agents and Multi-Agent Systems, 2005. 11(3): p. 307-360.
325. Vázquez-Salceda, J., *The role of norms and electronic institutions in multi-agent systems applied to complex domains. The HARMONIA framework*. Ai Communications, 2003. 16(3): p. 209-212.
326. Bernon, C., et al., *The Adelfe Methodology For an Intranet System Design*. AOIS@ CAiSE, 2002. 57.
327. Bernon, C., et al. *ADELFE: a methodology for adaptive multi-agent systems engineering*. in *International Workshop on Engineering Societies in the Agents World*. 2002. Springer.
328. Kruchten, P., *The rational unified process: an introduction*. 2004: Addison-Wesley Professional.
329. Booch, G., *The unified modeling language user guide*. 2005: Pearson Education India.
330. Cossentino, M., *From Requirements to Code with the PASSI Methodology, Agent-Oriented Methodologies, B. Henderson-Sellers and P. Giorgini*. Idea Group Inc., Hershey, PA, USA, 2005.
331. Chella, A., et al., *Agile passi: An agile process for designing agents*. International Journal of Computer Systems Science & Engineering, 2006. 21(2): p. 133-144.
332. Martin, R.C., *Agile software development: principles, patterns, and practices*. 2002: Prentice Hall.
333. Deloach, S.A., W.H. Oyenan, and E.T. Matson, *A capabilities-based model for adaptive organizations*. Autonomous Agents and Multi-Agent Systems, 2008. 16(1): p. 13-56.
334. DeLoach, S.A., *Multiagent systems engineering: a methodology and language for designing agent systems*. 1999, AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH DEPT OF ELECTRICAL AND ….

335. DeLoach, S.A., M.F. Wood, and C.H. Sparkman, *Multiagent systems engineering.* International Journal of Software Engineering and Knowledge Engineering, 2001. 11(03): p. 231-258.
336. Pavón, J., J.J. Gómez-Sanz, and R. Fuentes, *The INGENIAS methodology and tools*, in *Agent-oriented methodologies*. 2005, IGI Global. p. 236-276.
337. Fernández-Caballero, A. and J.M. Gascueña. *Developing multi-agent systems through integrating Prometheus, INGENIAS and ICARO-T.* in *International Conference on Agents and Artificial Intelligence*. 2009. Springer.
338. García-Magariño, I., J.J. Gómez-Sanz, and R. Fuentes-Fernández. *Model transformations for improving multi-agent system development in INGENIAS*. in *International Workshop on Agent-Oriented Software Engineering*. 2009. Springer.
339. Botti, V. and A. Giret, *Holonic manufacturing systems*. 2008: Springer.
340. Höpf, M. and C.F. Schaeffer. *Holonic manufacturing systems*. in *International Working Conference on the Design of Information Infrastructure Systems for Manufacturing*. 1996. Springer.
341. Esteva, M., et al., *On the formal specification of electronic institutions*, in *Agent mediated electronic commerce*. 2001, Springer. p. 126-147.
342. García-Camino, A., P. Noriega, and J.A. Rodríguez-Aguilar. *Implementing norms in electronic institutions*. in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. 2005.
343. Arcos, J.L., et al., *Engineering open environments with electronic institutions.* Engineering applications of artificial intelligence, 2005. 18(2): p. 191-204.
344. Alberola, J.M., V. Julian, and A. Garcia-Fornes, *Open issues in multiagent system reorganization*, in *Highlights in Practical Applications of Agents and Multiagent Systems*. 2011, Springer. p. 151-158.
345. So, Y.-p. and E.H. Durfee. *An organizational self-design model for organizational change*. in *Proceedings of AAAI93 Workshop on AI and Theories of Groups and Organizations*. 1993.
346. Dignum, M., E. Sonenberg, and F. Dignum. *Dynamic reorganization of agent societies*. in *Proceedings of workshop on coordination in emergent agent societies*. 2004.
347. Carvalho, G., et al. *Dynamic law evolution in governance mechanisms for open multi-agent systems*. in *Second workshop on software engineering for agent-oriented systems*. 2006.
348. Hoogendoorn, M. and J. Treur, *An adaptive multi-agent organization model based on dynamic role allocation.* International journal of knowledge-based and intelligent engineering systems, 2009. 13(3-4): p. 119-139.
349. Hübner, J.F., J.S. Sichman, and O. Boissier. *Using the $\mathcal {M}oise+ for a Cooperative Framework of MAS Reorganisation*. in *Brazilian Symposium on Artificial Intelligence*. 2004. Springer.
350. Bou, E., M. López-Sánchez, and J.A. Rodríguez-Aguilar. *Towards self-configuration in autonomic electronic institutions*. in *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*. 2006. Springer.
351. DeLoach, S.A. and E. Matson. *An organizational model for designing adaptive multiagent systems*. in *The AAAI-04 Workshop on Agent Organizations: Theory and Practice (AOTP 2004)*. 2004.
352. Patel, P., A.H. Ranabahu, and A.P. Sheth, *Service level agreement in cloud computing.* 2009.
353. Emeakaroha, V.C., et al. *Low level metrics to high level SLAs-LoM2HiS framework: Bridging the gap between monitored metrics and SLA parameters in cloud environments*. in *2010 International Conference on High Performance Computing & Simulation*. 2010. IEEE.
354. Goiri, Í., et al. *Resource-level QoS metric for CPU-based guarantees in cloud providers*. in *International Workshop on Grid Economics and Business Models*. 2010. Springer.
355. Li, A., et al. *CloudCmp: comparing public cloud providers*. in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 2010.
356. Kibe, S., M. Yamagiwa, and M. Uehara. *Grid on cloud.* in *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*. 2013. IEEE.
357. Rings, T., et al., *Grid and cloud computing: opportunities for integration with the next generation network.* Journal of Grid Computing, 2009. 7(3): p. 375.
358. Dean, J. and S. Ghemawat, *MapReduce: simplified data processing on large clusters.* Communications of the ACM, 2008. 51(1): p. 107-113.
359. Ladan, M.I., *Web services metrics: A survey and a classification.* Journal of Communication and Computer, 2012. 9(7): p. 824-829.
360. Erl, T., *SOA Principles of Service Design (paperback)*. 2016: Prentice Hall Press.

361. Urgaonkar, B., et al., *An analytical model for multi-tier internet services and its applications.* ACM SIGMETRICS Performance Evaluation Review, 2005. 33(1): p. 291-302.
362. Turner IV, W.P., et al., *Tier classification define site infrastructure performance.* Uptime Institute, 2006. 17.
363. Schaller, R.R., *Moore's law: past, present and future.* IEEE spectrum, 1997. 34(6): p. 52-59.
364. Barroso, L.A., *The price of performance.* Queue, 2005. 3(7): p. 48-53.
365. Niebert, N., et al., *Network virtualization: A viable path towards the future internet.* Wireless Personal Communications, 2008. 45(4): p. 511-520.
366. Chowdhury, N.M.K. and R. Boutaba, *Network virtualization: state of the art and research challenges.* IEEE Communications magazine, 2009. 47(7): p. 20-26.
367. Barroso, L.A. and U. Hölzle, *The datacenter as a computer: An introduction to the design of warehouse-scale machines.* Synthesis lectures on computer architecture, 2009. 4(1): p. 1-108.
368. Sullivan, M. and D. Anderson, *Marionette: A system for parallel distributed programming using a master/slave model.* 1988, CALIFORNIA UNIV BERKELEY COMPUTER SCIENCE DIV.
369. Crosby, S. and D. Brown, *The virtualization reality.* Queue, 2006. 4(10): p. 34-41.
370. Youseff, L., et al. *Paravirtualization for HPC systems.* in *International Symposium on Parallel and Distributed Processing and Applications*. 2006. Springer.
371. Habib, I., *Virtualization with kvm.* Linux Journal, 2008. 2008(166): p. 8.
372. Sahoo, J., S. Mohapatra, and R. Lath. *Virtualization: A survey on concepts, taxonomy and associated security issues.* in *2010 Second International Conference on Computer and Network Technology*. 2010. IEEE.
373. Lieberman, P., *White paper: Wake on lan technology.* Lieberman Software Corporation: Los Angeles, CA, USA, 2006.
374. Rabe, B.R., M. Clifford, and N. Miles, *Storage area network (SAN) management system for discovering SAN components using a SAN management server.* 2007, Google Patents.
375. Grossman, R.L., et al., *Compute and storage clouds using wide area high performance networks.* Future Generation Computer Systems, 2009. 25(2): p. 179-183.
376. Lubbers, C., et al., *Storage area network, data replication and storage controller, and method for replicating data using virtualized volumes*. 2005, Google Patents.
377. Moreno-Luzón, M.D., F.J. Peris Bonet, and T. González Cruz, *Gestión dela calidad y diseño de organizaciones: teoría y estudio de casos.* 2001: Prentice Hall.
378. Son, S. and K.M. Sim, *A price-and-time-slot-negotiation mechanism for cloud service reservations.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2011. 42(3): p. 713-728.
379. Agüero, J., et al. *Agent design using model driven development*. in *7th International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS 2009)*. 2009. Springer.
380. Agüero, J., et al., *MDD for Virtual Organization design*, in *Trends in Practical Applications of Agents and Multiagent Systems*. 2010, Springer. p. 9-17.
381. Carrascosa, C., et al. *Service oriented MAS: an open architecture*. in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 2009.
382. Giret, A., et al. *An open architecture for service-oriented virtual organizations*. in *International Workshop on Programming Multi-Agent Systems*. 2009. Springer.
383. Hooper, A., *Green computing.* Communication of the ACM, 2008. 51(10): p. 11-13.
384. Guo, B., Y. Shen, and Z. Shao, *The redefinition and some discussion of green computing.* Chinese Journal of Computers, 2009. 32(12): p. 2311-2319.
385. Barroso, L.A., J. Dean, and U. Holzle, *Web search for a planet: The Google cluster architecture.* IEEE micro, 2003. 23(2): p. 22-28.
386. Darling, C.L., et al., *Dynamic monitor and controller of availability of a load-balancing cluster.* 2007, Google Patents.
387. Beloglazov, A. and R. Buyya, *Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers.* MGC@ Middleware, 2010. 4.
388. Anuradha, V. and D. Sumathi. *A survey on resource allocation strategies in cloud computing.* in *International Conference on Information Communication and Embedded Systems (ICICES2014)*. 2014. IEEE.
389. Jung, G., et al. *A time-driven adaptive mecahnism for cloud resource allocation*. in *2011 4th IEEE International Conference on Broadband Network and Multimedia Technology*. 2011. IEEE.

390. Van, H.N., F.D. Tran, and J.-M. Menaud. *Autonomic virtual resource management for service hosting platforms*. in *2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing*. 2009. IEEE.
391. Raghavendra, R., et al. *No" power" struggles: coordinated multi-level power management for the data center*. in *Proceedings of the 13th international conference on Architectural support for programming languages and operating systems*. 2008.
392. Kusic, D., et al., *Power and performance management of virtualized computing environments via lookahead control*. Cluster computing, 2009. 12(1): p. 1-15.
393. Corchado, J.M. and R. Laza, *Constructing deliberative agents with case-based reasoning technology*. International Journal of Intelligent Systems, 2003. 18(12): p. 1227-1241.
394. Corchado, J.M., et al., *Replanning mechanism for deliberative agents in dynamic changing environments*. Computational Intelligence, 2008. 24(2): p. 77-107.
395. González-Bedia, M., *Fundamentos cognitivos para el diseño de arquitecturas de agentes planificadores en contextos dinámicos de acción*. 2004, Tesis Doctoral. Salamanca, Salamanca, España: Universidad de Salamanca.
396. Bajo, J., et al., *An execution time planner for the ARTIS agent architecture*. Engineering Applications of Artificial Intelligence, 2008. 21(5): p. 769-784.
397. De Mantaras, R.L. and E. Plaza, *Case-Based Reasoning: an overview*. AI communications, 1997. 10(1): p. 21-29.
398. Schank, R.C., *Dynamic memory: A theory of learning in people and computers*. 1982, Cambridge: Cambridge University Press.
399. Golding, A.R. and P.S. Rosenbloom. *Combining analytical and similarity-based CBR*. in *Proc. of the 2 nd Workshop on Case-Based Reasoning*. 1989.
400. Leake, D.B., et al. *Capture, storage and reuse of lessons about information resources: Supporting task-based information search*. in *Proceedings of the AAAI-2000 Workshop on Intelligent Lessons Learned Systems*. 2000. Menlo Park, CA.
401. Abraham, A., *Business intelligence from web usage mining*. Journal of Information & Knowledge Management, 2003. 2(04): p. 375-390.
402. Calheiros, R.N., et al., *CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms*. Software: Practice and experience, 2011. 41(1): p. 23-50.
403. Wickremasinghe, B., R.N. Calheiros, and R. Buyya. *Cloudanalyst: A cloudsim-based visual modeller for analysing cloud computing environments and applications*. in *2010 24th IEEE international conference on advanced information networking and applications*. 2010. IEEE.
404. Lim, S.-H., et al. *MDCSim: A multi-tier data center simulation, platform*. in *2009 IEEE International Conference on Cluster Computing and Workshops*. 2009. IEEE.
405. Kliazovich, D., P. Bouvry, and S.U. Khan, *GreenCloud: a packet-level simulator of energy-aware cloud computing data centers*. The Journal of Supercomputing, 2012. 62(3): p. 1263-1283.
406. Fittkau, F., S. Frey, and W. Hasselbring. *CDOSim: Simulating cloud deployment options for software migration support*. in *2012 IEEE 6th International Workshop on the Maintenance and Evolution of Service-Oriented and Cloud-Based Systems (MESOCA)*. 2012. IEEE.
407. Holzer, A. and J. Ondrus, *Mobile application market: A developer's perspective*. Telematics and informatics, 2011. 28(1): p. 22-31.
408. Jani, H. *Single sign-on*. in *Proc. Helsinki Uiniversity of Technology Seminar on Network Security*. 1997.
409. Tormasov, A.G., S.S. Protassov, and S.M. Beloussov, *Management of virtual and physical servers using graphic control panels*. 2008, Google Patents.
410. Bellard, F. *QEMU, a fast and portable dynamic translator*. in *USENIX Annual Technical Conference, FREENIX Track*. 2005.
411. Needham, R.M. *Denial of service*. in *Proceedings of the 1st ACM Conference on Computer and Communications Security*. 1993.
412. Goudarzi, H. and M. Pedram. *Multi-dimensional SLA-based resource allocation for multi-tier cloud computing systems*. in *2011 IEEE 4th International Conference on Cloud Computing*. 2011. IEEE.

COLECCIÓN VÍTOR, 449

Ediciones Universidad
Salamanca

ISBN 978-84-1311-611-2