Universidad Zaragoza

Proyecto Fin de Carrera
Ingeniería en Informática

Diseño e implementación de un sistema dinámico de gestión de trabajos distribuidos en un entorno de máquinas virtuales.

David Ceresuela Palomera

Director: Javier Celaya

Departamento de Informática e Ingeniería de Sistemas
Escuela de Ingeniería y Arquitectura
Universidad de Zaragoza

Curso 2011/2012
Septiembre 2012

Resumen

A la hora de ejecutar trabajos en un entorno distribuido, la aproximación clásica ha sido bien el uso de un *cluster* de ordenadores o bien el uso de la computación en malla o *grid*. Con la proliferación de entornos *cloud* durante estos últimos años y su facilidad de uso, una nueva opción se abre para la ejecución de este tipo de trabajos.

De hecho, la ejecución de trabajos distribuidos es uno de los principales usos dentro del ámbito de los sistemas *cloud*. Sin embargo, la administración de este tipo de sistemas dista de ser sencilla: cuestiones como la puesta en marcha del sistema, el aprovisionamiento de nodos, las modificaciones del sistema y la evolución y actualización del mismo suponen una tarea intensa y pesada.

En vista de lo cual, en este proyecto se ha diseñado una solución capaz de automatizar la administración de sistemas *cloud*, y en particular de un sistema de ejecución de trabajos distribuidos. Para ello se han estudiado entornos clásicos de ejecución de trabajos como Torque y entornos de ejecución de trabajos en *cloud* como AppScale. Además, se han estudiado herramientas clásicas de configuración automática de sistemas como Puppet y CFEngine. El objetivo principal de estas herramientas de configuración de sistemas es la gestión del nodo. En este proyecto se ha extendido la funcionalidad de una de estas herramientas — Puppet — añadiéndole la capacidad de gestión de sistemas *cloud*.

Como resultado de este proyecto se presenta una solución capaz de administrar de forma automática sistemas de ejecución de trabajos distribuidos. La validación de esta solución se ha llevado a cabo sobre los entornos de ejecución de trabajos Torque y AppScale y también, para mostrar su carácter genérico, sobre una arquitectura de servicios web de tres niveles.

Índice general

Capítulo 1
Introducción

[Revisar]

La computación en la nube es un nuevo paradigma que pretende transformar la computación en un servicio. Durante estos últimos años la computación en la nube ha ido ganando importancia de manera progresiva, ya que la posibilidad de usar la computación como un servicio permite a los usuarios de una aplicación acceder a ésta a través de un navegador web, una aplicación móvil o un cliente de escritorio, mientras que la lógica de la aplicación y los datos se encuentran en servidores situados en una localización remota. Esta facilidad de acceso a la aplicación sin necesitar de un profundo conocimento de la infraestructura es la que, por ejemplo, brinda a las empresas la posibilidad de ofrecer servicios web sin tener que hacer una gran inversión inicial en infraestructura propia. Las aplicaciones alojadas en la nube tratan de proporcionar al usuario el mismo servicio y rendimiento que las aplicaciones instaladas localmente en su ordenador.

A lo largo de los últimos años las herramientas de gestión de configuración (o herramientas de administración de sistemas) también han experimentado un considerable avance: con entornos cada vez más heterogéneos y complejos la administración de estos sistemas de forma manual ya no es una opción. Entre todo el conjunto de herramientas de gestión de configuración destacan de manera especial Puppet y CFEngine. Puppet es una herramienta basada en un lenguaje declarativo: el usuario especifica qué estado debe alcanzarse y Puppet se encarga de hacerlo. CFEngine, también con un lenguaje declarativo, permite al usuario un control más detallado de cómo se hacen las cosas, mejorando el rendimiento a costa de perder abstracciones de más alto nivel.

Sin embargo, estas herramientas de gestión de la configuración carecen de la funcionalidad requerida para administrar infraestructuras distribuidas. Son capaces de asegurar que cada uno de los nodos se comporta de acuerdo a la configuración que le ha sido asignada pero no son capaces de administrar una infraestructura distribuida como una entidad propia. Si tomamos la administración de un *cloud* como la administración de las máquinas virtuales que forman los nodos del mismo nos damos cuenta de que la administración es puramente *software*. Únicamente tenemos que asegurarnos de que para cada nodo de la infraestructura distribuida hay una máquina virtual que está cumpliendo con su función.

Teniendo en cuenta el considerable avance de la computación en la nube, parece claro que el siguiente paso de las herramientas de gestión de configuración debería ir encaminado a la gestión de la nube. Para demostrar una posible manera en la que esto se podría lograr, en este proyecto se ha tomado una de esas herramientas de gestión de la configuración y se ha modificado añadiéndole la posibilidad de gestionar infraestructuras distribuidas. La modificación realizada se

ha validado usando tres ejemplos de infraestructuras distribuidas, que se explican a continuación.

La primera de ellas es AppScale [2], una implementación *open source* del App Engine de Google [7]. App Engine permite alojar aplicaciones web en la infraestructura que Google posee. Además del alojamiento de aplicaciones web, AppScale también ofrece las APIs [1] de EC2 [3], MapReduce [4] y Neptune [5]. La API de EC2 añade a las aplicaciones la capacidad de interactuar con máquinas alojadas en Amazon EC2 [6]. La API de MapReduce permite escribir aplicaciones que hagan uso del *framework* MapReduce. La última API, Neptune, añade a App Engine la capacidad de usar los nodos de la infraestructura para ejecutar trabajos. Los trabajos más representativos que puede ejecutar son: de entrada, de salida y MPI [2]. El trabajo de entrada sirve para subir ficheros (generalmente el código que se ejecutará) a la infraestructura, el de salida para traer ficheros (generalmente los resultados obtenidos después de la ejecución) y el de MPI para ejecutar un trabajo MPI.

La segunda infraestructura es Torque, una infraestructura de ejecución de trabajos. Este tipo de infraestructuras está especializada en la ejecución de grandes cargas de trabajo paralelizable e intensivo en computación. Son por lo tanto idóneas para ser usadas en la computación de altas prestaciones.

La tercera y última es la de servicios web en tres capas. Este tipo de infraestructura tiene tres niveles claramente diferenciados: balanceo o distribución de carga, servidor web y base de datos. El balanceador de carga es el encargado de distribuir las peticiones web a los servidores web que se encuentran en el segundo nivel de la infraestructura. Éstos procesarán las peticiones web y para responder a los clientes puede que tengan que consultar o modificar ciertos datos. Los datos de la aplicación se encuentran en la base de datos, el tercer nivel de la estructura, y por consiguiente, cada vez que uno de los elementos del segundo nivel necesite leer información o modificarla, accederá a este nivel. Para esta infraestructura no se puede elegir un ejemplo que destaque sobre los demás porque es tan común que cualquier página web profesional de hoy en día se sustenta en una infraestructura similar a ésta.

1.1 Contexto del proyecto

Para la realización de este proyecto de fin de carrera se ha hecho uso del laboratorio 1.03b de investigación que el Departamento de Informática e Ingeniería de Sistemas posee en la Escuela de Ingeniería y Arquitectura de la Universidad de Zaragoza. Los ordenadores que forman este laboratorio poseen procesadores con soporte de virtualización, lo que permite la creación de diversas máquinas virtuales. La creación de los distintos tipos de *cloud* que representan cada una de las infraestructuras distribuidas se ha llevado a cabo a través de máquinas virtuales alojadas en distintos ordenadores del laboratorio.

En este laboratorio se ha comprobado la validez de la extensión introducida en la herramienta de gestión de configuraciones Puppet para administración de infraestructuras distribuidas que se ha desarrollado a lo largo de este proyecto de fin de carrera.

[1] API (del inglés *Application Programming Interface*, Interfaz de programación de aplicaciones) es el conjunto de funciones y procedimientos que ofrece una biblioteca para ser utilizado por otro *software* como una capa de abstracción.

[2] MPI (del inglés *Message Passing Interface*, Interfaz de Paso de Mensajes) es un estándar que define la sintaxis y la semántica de las funciones de una biblioteca de paso de mensajes diseñada para ser usada en programas que exploten la existencia de múltiples procesadores.

1.2 Objetivos

El objetivo de este proyecto es proporcionar una herramienta que facilite la puesta en marcha de infraestructuras distribuidas y su posterior mantenimiento. Las tareas principales en las que se puede dividir este proyecto son:

1. Análisis de las herramientas de administración de virtualización *hardware*.
2. Estudio de algunas de las infraestructuras distribuidas existentes profundizando en la parte relativa a la ejecución de trabajos distribuidos.
3. Investigación de las herramientas de gestión de configuración existentes más relevantes y elección de aquella que mayor facilidad de integración y uso proporcione.
4. Extensión de la herramienta de gestión de configuración para que soporte la puesta en marcha y el mantenimiento de un sistema de ejecución de trabajos distribuidos.

1.3 Trabajos previos

Desde un primer momento se decidió trabajar con la herramienta de configuración Puppet para la realización de este proyecto. La otra alternativa posible era CFEngine, pero a diferencia de ésta, Puppet posee un nivel mayor de abstracción que permite un mejor modelado de los recursos de un sistema. Además, el hecho de que Puppet esté programado en Ruby hace que sea más fácil trabajar y realizar abstracciones de alto nivel en él que en CFEngine, que está programado en el lenguaje C.

También se decidió desde el principio trabajar con la infraestructura de ejecución de trabajos que proporciona AppScale. AppScale combina la capacidad de ejecutar trabajos con el alojamiento de aplicaciones web. Esta dualidad la convierte en una infraestructura muy interesante para trabajar con ella.

La infraestructura de ejecución de trabajos Torque también se eligió desde el inicio.

1.4 Tecnología utilizada

Para la elaboración de este proyecto se ha hecho uso de las siguientes tecnologías:

- KVM, QEMU, libvirt y virsh para el soporte y la gestión de las máquinas virtuales.
- Puppet como herramienta de configuración automática.
- Ruby como lenguaje de programación para la extensión de Puppet.
- AppScale y Torque como infraestructuras de ejecución de trabajos distribuidos en las que validar la extensión.
- Nginx, WEBrick y MySQL como balanceador de carga, servidor web y base de datos para la infraestructura web de tres niveles en la que se valida la extensión.
- Shell como lenguaje de programación de los *scripts* de configuración de las máquinas virtuales.
- Sistema operativo Debian para las máquinas del laboratorio y Ubuntu para las máquinas virtuales.

- LaTeX [10] para la redacción de esta memoria.
- Dia para la elaboración de los diagramas que aparecen en esta memoria.

1.5 Herramientas utilizadas

[Revisar]

Una de las herramientas sobre las que se ha basado este proyecto ha sido la virtualización *hardware* o virtualización de plataforma, que permite la simulación de un ordenador completo (llamado huésped) dentro de otro ordenador (llamado anfitrión). A la hora de hacer una virtualización *hardware* hay varias opciones entre las que elegir, siendo las más ampliamente usadas Xen y KVM. La principal diferencia entre ellas es que Xen ofrece paravirtualización mientras que KVM ofrece virtualización nativa.

La virtualización nativa permite hacer una virtualización *hardware* completa de manera eficiente. Para ello, y a diferencia de la paravirtualizacón, no requiere de ninguna modificación en el sistema operativo de la máquina virtual, pero a cambio necesita un procesador con soporte para virtualización. KVM, que proporciona virtualización *hardware*, está incluido como un módulo del núcleo de Linux desde su versión 2.6.20, así que viene incluido por defecto en cualquier sistema operativo con núcleo Linux.

Como los ordenadores del laboratorio poseen procesadores con extensiones de soporte para virtualización y sistema operativo Debian, se eligió KVM para dar soporte a las máquinas virtuales. Esto significa que se puede usar cualquier sistema operativo para las máquinas virtuales, sin necesidad de hacer ninguna modificación en el mismo.

El resto de herramientas utilizadas se explican en detalle en sus respectivas secciones.

1.6 Organización de la memoria

El resto de este documento queda organizado de la siguiente manera:

Capítulo 2 Análisis de las herramientas e infraestructuras utilizadas.

Capítulo 3 Extensión de Puppet para gestión de infraestructuras distribuidas.

Capítulo 4 Diseño de recursos distribuidos específicos.

Capítulo 5 Validación de la solución planteada.

Capítulo 6 Conclusiones.

Además consta de una serie de anexos organizados de esta manera:

Anexo 1 Anexo 1.

Anexo 2 Anexo 2.

1.7 Agradecimientos

Agradecimientos

Capítulo 2
Herramientas e infraestructuras utilizadas

[Revisar]

En este capítulo se realiza un breve análisis de las distintas herramientas e infraestructuras usadas a lo largo del proyecto.

2.1 Análisis de la herramienta de gestión de configuración

Las herramientas de gestión de configuración tienen como objetivo llevar a un nodo a un cierto estado. Normalmente esto suele incluir la especificación de los recursos (ficheros, usuarios, paquetes instalados, etc.) que dicho nodo debería tener. Para cumplir esta misión se apoyan en dos conceptos básicos: convergencia e iteración. Esto quiere decir que iteración tras iteración tratan de acercar al nodo lo máximo posible al estado deseado. Es posible, por tanto, que el estado final no se alcance en una única ejecución, sino que sean necesarias varias ejecuciones. Aunque esto pueda contrastar con el funcionamiento habitual de los programas (sería sorprendente que sólo se abriera medio editor de texto), no es tan excepcional en este entorno: si tenemos que poner en marcha dos servicios, de los cuales uno de ellos depende del otro, hasta que el primero no esté funcionando no podrá hacerlo el segundo. Una sola ejecución de la herramienta no serviría para poner ambos servicios en marcha, sino que serían necesarias dos iteraciones como mínimo.

Puppet es una herramienta de gestión de configuración basada en un lenguaje declarativo. A través de este lenguaje se modelan los distintos elementos de configuración, que en la terminología de Puppet se llaman recursos. Mediante el uso de este lenguaje se indica en qué estado se quiere mantener el recurso y será tarea de Puppet el encargarse de que así sea. Cada recurso está compuesto de un tipo (el tipo de recurso que estamos gestionando), un título (el nombre del recurso) y una serie de atributos (los valores que especifican el estado del recurso).

La agrupación de uno o más recursos en un fichero de texto da lugar a un manifiesto. En general, un manifiesto contiene la información necesaria para realizar la configuración de un nodo. Un usuario normal de Puppet creará manifiestos en los que especifique los recursos y su estado. Por ejemplo, para crear un fichero un usuario escribirá en un manifiesto algo similar a lo siguiente:

```
file {'testfile':
  path    => '/tmp/testfile',
  ensure  => present,
```

```
  mode    => 0640,
  content => "I'm a test file.",
}
```

Los usuarios más avanzados de Puppet pueden crear sus propios tipos de recursos. Para que Puppet sepa cómo tratar con ese recurso deberá crear un proveedor. Esencialmente, el tipo se encarga de definir qué se puede hacer con un recurso y el proveedor se encarga de definir cómo hacerlo.

Cuando se tiene listo el manifiesto a Puppet se le da la orden de aplicarlo. Los pasos que sigue para aplicarlo son:

- Interpretar y compilar la configuración.
- Comunicar la configuración compilada al nodo.
- Aplicar la configuración en el nodo.
- Enviar un informe con los resultados.

Normalmente Puppet se ejecuta de manera periódica mediante un planificador de trabajos (por ejemplo, cron). Cada cierto tiempo contactará con el nodo que debe ser administrado y volverá a repetir los pasos anteriores. Es decir, Puppet está continuamente intentando llevar al nodo al estado especificado en el manifiesto. Si entre una ejecución y otra algo cambiara en el nodo, Puppet se daría cuenta e intentaría llevar al nodo al estado que le corresponde.

Aunque Puppet tiene la capacidad de actuar sobre un nodo distinto al nodo desde el que se aplica el manifiesto, a lo largo de este proyecto las ejecuciones de Puppet siempre serán sobre el nodo que aplica el manifiesto, siempre serán locales.

Además de permitir la creación de tipos y proveedores para los usuarios más avanzados, Puppet tiene otros puntos de extensión. Uno de ellos es la API *Faces* [11], que permite crear subcomandos y acciones dentro de Puppet. Después de analizar esta API a fondo se vio que las opciones que ofrecía no permitían la integración del recurso distribuido dentro del modelo de Puppet. Como lo que interesaba era crear una abstracción del recurso distribuido esta opción se acabó descartando en favor de la creación de un tipo y un proveedor, que soluciona el problema de una manera más elegante.

2.2 Análisis de las infraestructuras de ejecución de trabajos distribuidos

En esta sección se analizarán en profundidad las dos infraestructuras de ejecución de trabajos distribuidos usadas en este proyecto: AppScale y Torque

2.2.1 Análisis de la infraestructura AppScale

AppScale es una implementación *open source* del App Engine de Google. Al igual que App Engine, AppScale permite alojar aplicaciones web; a diferencia de App Engine, las aplicaciones no serán alojadas en la infraestructura que Google posee, sino que serán alojadas en una infraestructura que el usuario posea. Además de permitir alojar aplicaciones web, AppScale también ofrece las APIs de MapReduce y Neptune. La API de MapReduce permite escribir aplicaciones que hagan uso del *framework* MapReduce. La API de Neptune añade a App Engine la capacidad

de usar los nodos de la infraestructura para ejecutar trabajos. Los trabajos más representativos que puede ejecutar son: de entrada, de salida y MPI, aunque también se pueden ejecutar trabajos de otro tipo.

AppScale trata de imitar de la manera más fiel los servicios que ofrece el App Engine de Google, tratatando de alcanzar el mayor grado de compatibilidad posible. Como la tecnología que Google usa para hacer esto posible permanece oculta al público, AppScale tiene que hacer uso de otras tecnologías existentes para ofrecer las mismas funciones. En la Figura 2.1 se puede ver toda la pila de tecnologías usadas en AppScale.

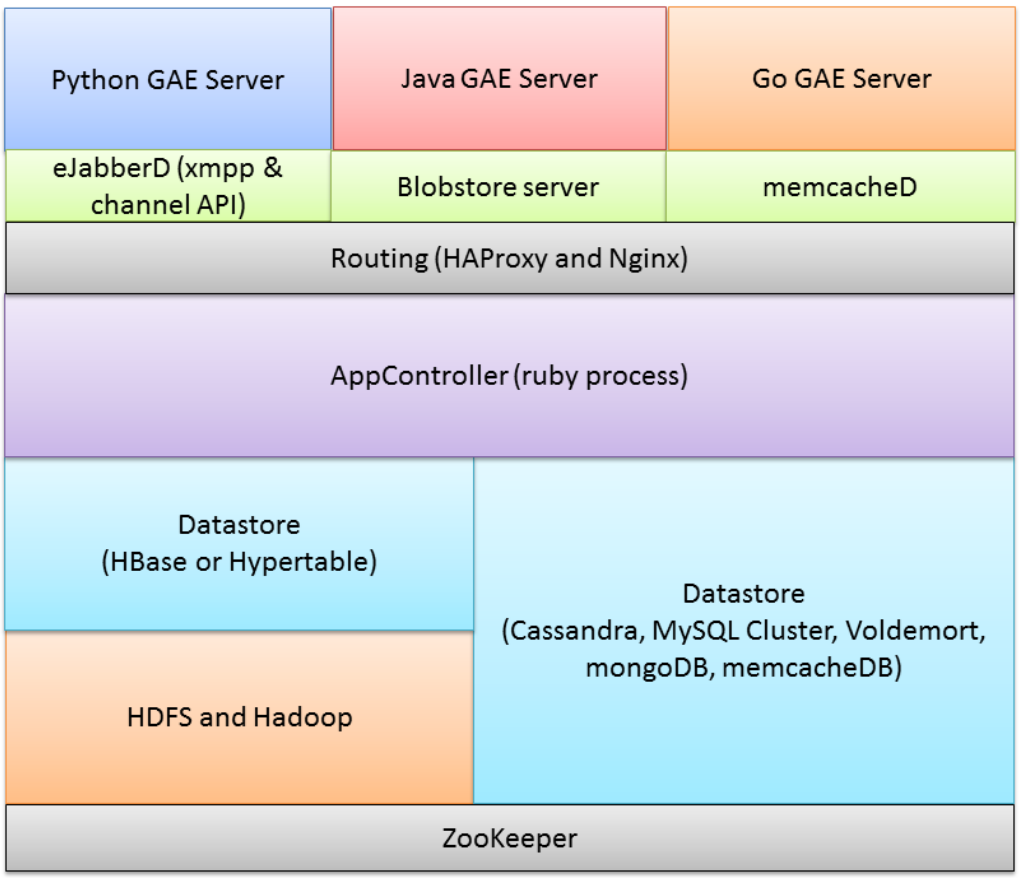

Figura 2.1: Tecnologías usadas por AppScale.

Para permitir el alojamiento de aplicaciones web y la ejecución de trabajos una compleja infraestructura debe ponerse en marcha. Los distintos roles que un nodo puede desempeñar en una infraestructura de este tipo son los siguientes:

`Shadow` : Comprueba el estado en el que se encuentran el resto de nodos y se asegura de que están ejecutando los servicios que deben.

`Load balancer` : Servicio web que lleva a los usuarios a las aplicaciones. Posee también una página en la que informa del estado de todas las máquinas desplegadas.

`AppEngine` : Versión modificada de los SDKs de Google App Engine. Además de alojar las aplicaciones web añaden la capacidad de almacenar y recuperar datos de bases de datos que soporten la API de Google Datastore.

`Database` : Ejecuta los servicios necesarios para alojar a la base de datos elegida.

`Login` : La máquina principal que lleva a los usuarios a las aplicaciones App Engine. Difiere del Load Balancer en que esta es la única máquina desde la que se pueden hacer funciones administrativas. Puede haber muchas máquinas que hagan la función de Load Balancer pero sólo habrá una que haga función de Login.

`Memcache` : Proporciona soporte para almacenamiento en caché para las aplicaciones App Engine.

`Zookeeper` : Aloja los metadatos necesarios para hacer transacciones en las bases de datos.

`Open` : No ejecuta nada por defecto, pero está disponible por si acaso. Estas máquinas son las utilizadas para ejecutar trabajos MPI.

Como muchos de estos roles suele desempeñarlos una máquina, se han creado unos roles agregados para facilitar el despliegue de la infraestructura:

`Controller` : Shadow, Load Balancer, Database, Login y Zookeeper.

`Servers` : App Engine, Database y Load Balancer.

`Master` : Shadow, Load Balancer y Zookeeper.

No todos estos roles serán usados a la vez al hacer un despliegue de AppScale. AppScale contempla la posibilidad de dos tipos de despliegue: por defecto y personalizado. En el despliegue por defecto un nodo sólo puede ser o bien `Controller` o bien `Server`. En el despliegue personalizado el usuario es libre de especificar los roles de un nodo como más le convenga.

Una vez puesta en marcha la infraestructura de AppScale, servirá tanto para alojar las aplicaciones web que el usuario despliegue como para ejecutar trabajos. Para desplegar las aplicaciones web hay que hacer uso de las AppScale Tools, un conjunto de herramientas que permiten, entre otras cosas, iniciar y terminar instancias, desplegar aplicaciones y eliminar aplicaciones. Para ejecutar trabajos hay que servirse de la API de Neptune, que no es tan sencilla. En general esto se consigue mediante tres pasos: en el primero se le indica el código fuente que se quiere subir a la infraestructura; en el segundo se le da la orden de ejecutar el trabajo; en el tercero se le piden los resultados de la ejecución. Cada uno de estos pasos debe indicarse, mediante un lenguaje específico de dominio, en un fichero que luego se interpretará con el programa neptune. En el caso de un trabajo MPI, podemos definir, además del código a ejecutar, el número de máquinas sobre las que ejecutar el código y el número de procesos que se usarán para el trabajo.

2.2.2 Análisis de la infraestructura Torque

La otra infraestructura de ejecución de trabajos distribuidos que se ha elegido ha sido Torque. Torque es una de las infraestructuras clásicas en lo que a ejecución de trabajos se refiere. Una infraestructura de este tipo está compuesta de un nodo maestro y tantos nodos de computación como se desee. Una vez puesta en marcha la infraestructura, los usuarios que tengan los permisos correspondientes pueden mandar sus trabajos al nodo maestro. El nodo maestro, valiéndose de un planificador (en su versión más simple es una cola FIFO), decidirá a cuál de los nodos de computación le enviará el trabajo. El nodo de computación que reciba el trabajo será el encargado de ejecutarlo y de enviar los resultados de vuelta al nodo maestro una vez terminada la ejecución.

Capítulo 3
Modelado de recursos distribuidos con Puppet

[Revisar]

Las herramientas de gestión de la configuración se han centrado en la gestión de recursos de manera local a un nodo. Por otra parte la automatización existente en la administración de infraestructuras distribuidas es de bajo nivel, no yendo mucho más allá de la gestión de máquinas virtuales. En este capítulo veremos cómo se pueden crear recursos distribuidos en la herramienta de gestión de la configuración Puppet y cómo ésta puede ser usada para automatizar una administración de más alto nivel en infraestructuras distribuidas que tenga en cuenta conceptos como el de disponibilidad o el de prestaciones.

3.1 Configuración de recursos distribuidos

En Puppet, la definición clásica de recursos se presupone dentro del ámbito local del nodo. Es decir, para cada nodo especificamos qué recursos debe contener y cuál debe ser su estado. Dentro de este tipo de recursos se encuentran, entre otros, el recurso usuario y el recurso fichero. Sin embargo, el modelado de un recurso distribuido plantea ciertos desafíos al ejemplo anterior, ya que no está pensado teniendo en cuenta la problemática asociada a los sistemas distribuidos.

Al modelar un recurso distribuido, deben tenerse en cuenta las características propias de este tipo de recursos, como la disponibilidad, las prestaciones y las dependencias. La disponibilidad contempla los fallos que se pueden dar en una infraestructura distribuida y en ella estarían incluidos los fallos de procesos y los fallos de máquinas. Las prestaciones contemplan los servicios ofrecidos y dentro de ellas tendríamos la creación de máquinas para repartir la carga. Las dependencias contemplan la necesidad de que un servicio esté funcionando para que otro pueda hacerlo. Asimismo, un recurso distribuido puede presentar elementos comunes con otros recursos distribuidos, tales como una monitorización básica. La presencia de elementos comunes entre los recursos clásicos de Puppet, por ejemplo un fichero y un usuario, no es tan corriente.

A la hora de definir un recurso distribuido tenemos que presentarlo como un único sistema coherente, es decir, como una única abstracción, y por lo tanto no vale con describir un recurso distribuido como una colección de recursos clásicos de Puppet. Afortunadamente, Puppet proporciona los medios para crear nuevos tipos de recurso, y se puede crear un nuevo tipo de recurso con sus parámetros correspondientes para definir los recursos distribuidos.

3.2 Modelización en Puppet

Puppet puede ser extendido para incluir la definición de nuevos recursos. Para ello hay que proporcinarle como mínimo dos ficheros: uno en el que se define el recurso y otro en el que se define cómo gestionar ese recurso. Al fichero en el que se define el recurso se le llama tipo y al fichero en el que se define cómo gestionarlo se le llama proveedor. Es decir, el tipo se encarga del "qué" y el proveedor se encarga del "cómo".

Para definir un recurso distribuido, o recurso *cloud*, se han considerado como fundamentales los siguientes parámetros:

- Nombre: Para identificar al recurso de manera única.
- Fichero de dominio: Para definir una plantilla de creación de máquinas virtuales especificando sus características *hardware*.
- Conjunto de máquinas físicas: Para indicar qué máquinas físicas pueden ejecutar las máquinas virtuales definidas.

Estos parámetros se obtienen mediante la observación de los elementos comunes a todo recurso distribuido. Además de estos parámetros cada tipo de recurso distribuido puede añadir los que considere necesarios para una completa especificación del recurso.

3.2.1 Patrón de diseño del proveedor

En Puppet, el proveedor es el encargado de llevar al recurso al estado que se le indique en el manifiesto. Típicamente el proveedor posee las funciones necesarias para crear un nuevo recurso y para destruirlo. Para llevar a cabo estas funciones en un recurso de tipo *cloud* el proveedor se apoya en cuatro grupos de funciones: puesta en marcha de un *cloud*, monitorización de un *cloud*, elección de líder y parada de un *cloud*. Las funciones de los tres primeros grupos se usan a la hora de crear un nuevo recurso de tipo *cloud* mientras que las del último grupo se usan a la hora de parar un *cloud* ya existente.

Las operaciones de puesta en marcha son las encargadas de poner en funcionamiento el *cloud* especificado en el manifiesto. Las más importantes son:

- Inicio como líder: Función de puesta en marcha que realizará el nodo líder del *cloud*. Ésta es la función más importante dentro de las funciones de inicio del proveedor ya que es la que se encarga de iniciar el cloud. A grandes rasgos, los pasos que realiza son:

  1. Comprobación de la existencia del *cloud*: si no existe se creará.
  2. Comprobación del estado del conjunto de máquinas físicas.
  3. Obtención de las direcciones IP de los nodos y los roles que les han sido asignados.
  4. Comprobación del estado de las máquinas virtuales: si están funcionando se monitorizan, mientras que si no están funcionando hay que definir una nueva máquina virtual y ponerla en funcionamiento. Las funciones de monitorización incluyen el envío de un fichero mediante el cual cada nodo se autoadministre la mayor parte posible.
  5. Cuando todas las máquinas virtuales estén funcionando se procede a inicializar el *cloud*.
  6. Operaciones de puesta en marcha particulares dependiendo de cada tipo de *cloud*.

- Inicio como nodo común: Función de puesta en marcha que realizarán los nodos comunes del *cloud*.

- Inicio como nodo externo: Función de puesta en marcha que realizará un nodo no perteneciente al *cloud*.

La monitorización del *cloud* únicamente la llevará a cabo el nodo líder ya que sería redundate que más de un nodo se encargara de comprobar el estado global del *cloud*. Por tanto, sólo hay una función importante:

- Monitorización como líder: Función de monitorización que realizará el nodo líder del *cloud*.

Si sólo el nodo líder se encarga de comprobar el estado global del *cloud* deberá haber siempre un nodo que cumpla este papel. Para elegir un líder de entre todos los nodos del *cloud* se utiliza el algoritmo peleón (en inglés *Bully algorithm*). En este algoritmo todos los nodos tratan periódicamente de convertirse en el líder; si hay un líder impedirá que otro nodo le quite el liderazgo y si no lo hay uno de los restantes nodos se convertirá en líder. Para ayudar a la implementación de este algoritmo se proporcionan las funciones de elección de líder, de las cuales las más importantes son:

- Lectura y escritura de identificador: Funciones de lectura y escritura del identificador del nodo actual.

- Lectura y escritura de identificador de líder: Funciones de lectura y escritura del identificador del nodo líder.

- Escritura de identificador e identificador de líder remoto: Funciones de escritura del identificador y del identifiacor del líder en un nodo distinto del actual.

Por último, es posible que en algún momento se desee parar por completo el funcionamiento del *cloud*. Las operaciones de parada de *cloud* más importantes son:

- Apagado de máquinas virtuales: Función de apagado de las máquinas virtuales que forman el *cloud*.

- Borrado de ficheros: Función de eliminación de todos los ficheros internos de gestión del *cloud*.

Aunque no se proporcionan como funciones, hay que tener en cuenta que cada tipo de cloud puede tener sus propias funciones de parada. Estas funciones de parada deben realizarse antes de apagar las máquinas virtuales. De forma general, los pasos que hay que hacer a la hora de parar un cloud son:

1. Comprobación de la existencia del *cloud*: si existe se procederá a su parada.

2. Operaciones de parada particulares a cada tipo de *cloud*.

3. Apagado de las máquinas virtuales creadas explícitamente para este *cloud*.

4. Parada de las funciones de automantenimiento de los nodos.

5. Eliminación de los ficheros internos de gestión del *cloud*.

3.2.2 *Framework* de implementación

Las operaciones de puesta que se proporcionan son:

- Inicio como líder: `leader_start`.
- Inicio como nodo común: `common_start`.
- Inicio como nodo externo: `not_cloud_start`.

La función de monitorización que se proporciona es:

- Monitorización como líder: `leader_monitoring`.

Las funciones de elección de líder que se proporcionan son:

- Lectura y escritura de identificador: `get_id` y `set_id`.
- Lectura y escritura de identificador de líder: `get_leader` y `set_leader`.
- Escritura de identificador e identificador de líder remoto: `vm_set_id` y `vm_set_leader`.

Las funciones de parada de *cloud* que se proporcionan son:

- Apagado de máquinas virtuales: `shutdown_vms`.
- Borrado de ficheros: `delete_files`.

Capítulo 4
Metodología de diseño e implementación de recursos distribuidos

4.1 Especificación del tipo

El primer paso que hay que hacer a la hora de crear un nuevo recurso en Puppet es definir el tipo de recurso. Para ello en un fichero habrá que especificar el nombre del nuevo tipo y los argumentos que éste tiene. La creación del tipo `mitipo` se haría en un fichero `mitipo.rb` con un contenido similar a éste:

```
1  Puppet::Type.newtype(:mitipo) do
2
3     @doc = "Tipo 'mitipo'."
4
5     ensurable do
6        desc "El campo ensure puede tomar uno de los siguientes valores:
7     'running': El cloud esta en marcha.
8     'stopped': El cloud esta parado.\n"
9
10       newvalue(:stopped) do
11          provider.stop
12       end
13
14       newvalue(:running) do
15          provider.start
16       end
17
18    end
19
20    # Parametros
21
22    newparam(:name) do
23       desc "El nombre del recurso"
24       isnamevar
25    end
26
27    newparam(:param1) do
28       desc "El parametro 1 es un parametro muy simple"
```

```
29      end
30
31      newproperty(:param2, :array_matching => :all) do
32          desc "El parametro 2 es un vector de opciones"
33      end
34
35  end
```

4.2 Diseño e implementación del proveedor

Una vez que se ha definido el tipo del recurso, queda definir el proveedor para ese tipo. En el caso de los recursos distribuidos, y teniendo en cuenta las funciones especificadas en la parte de `ensurable`, el proveedor deberá contener obligatoriamente las funciones `start` y `stop`. Un proveedor simplificado sería similar a éste:

```
1   Puppet::Type.type(:mitipo).provide(:mitipo_proveedor) do
2       desc "Proveedor para el tipo 'mitipo'."
3
4       ...
5
6       # Poner el cloud en marcha.
7       def start
8           ...
9       end
10
11      # Parar el cloud.
12      def stop
13          ...
14      end
15
16      ...
17
18  end
```

Dentro de las funciones `start` y `stop` se puede hacer uso de las funciones especificadas en la sección 3.2.2.

4.3 Ejemplos de diseño e implementación

En esta sección se verá cómo aplicar los pasos anteriores en ejemplos prácticos de arquitecturas distribuidas. Veremos como se aplica a AppScale, a Torque y a una infraestructura web de servicios en tres niveles.

4.3.1 Diseño e implementación de un recurso distribuido para una infraestructura AppScale

Una infraestructura AppScale puede ser definida de dos maneras: mediante un despliegue por defecto o uno personalizado. En un despliegue por defecto un nodo es el encargado de controlar

la infraestructura y el resto de nodos se encargan de hacer el resto del trabajo. En un despliegue personalizado podemos especificar con mayor grado de precisión qué tipo de trabajo debe hacer cada nodo. Por ejemplo, podemos indicar qué nodos se encargarán de alojar las aplicaciones de los usuarios, qué nodos alojarán la base de datos o qué nodos serán los encargados de ejecutar los trabajos de computación. Para administrar una infraestructura AppScale, sin importar el tipo de despliegue, necesitaremos una cuenta de correo y una contraseña. Este usuario y contraseña son necesarios para poder administrar las aplicaciones alojadas y observar el estado de la infraestructura.

La sintaxis del manifiesto del recurso AppScale no se verá afectada por los dos tipos de despliegue posibles, pero sí que tendrá que reflejar los parámetros necesarios para realizar las tareas de administración de la infraestructura. Éste podría ser un ejemplo de un manifiesto para la puesta en marcha de una infraestructura de tipo AppScale:

```
appscale {'mycloud':
   ip_file      => "/etc/puppet/modules/appscale/files/appscale-ip.yaml
      ",
   img_file     => "/etc/puppet/modules/appscale/files/appscale-img.yaml
      ",
   domain       => "/etc/puppet/modules/appscale/files/mycloud-template.
      xml",
   pool         => ["155.210.155.70"],
   app_email    => "user@mail.com"
   app_password => "password"
   ensure       => running,
}
```

La parada de una infraestructura AppScale no requiere un manifiesto tan complejo:

```
appscale {'mycloud':
   pool   => ["155.210.155.70"],
   ensure => stopped,
}
```

El fichero de roles sí que debe reflejar los dos posibles tipos de despliegue. En un despliegue por defecto los posibles roles que puede tomar un nodo son:

`controller` : La máquina que desempeñará el rol de nodo controlador.

`servers` : La lista de máquinas que desempeñarán el rol de nodos de trabajo.

Un fichero de roles para este despliegue sería de esta forma:

```
---
:controller: 155.210.155.73
:servers:
- 155.210.155.177
- 155.210.155.178
```

Por otra parte, los posibles roles que puede desempeñar un nodo en un despliegue personalizado y que resultan interesantes desde nuestro punto de vista son:

`master` : La máquina que desempeñará el rol de nodo maestro.

`appengine` : Los servidores para alojar las aplicaciones.

`database` : Las máquinas que contienen la base de datos.

`login` : La máquina encargada de redirigir a los usuarios a sus servidores. Es también la que se le facilita al administrador de la infraestructura para que realice las tareas administrativas.

`open` : Las máquinas de ejecución de trabajos. También pueden ser usadas como nodos de reserva por si falla algún otro nodo.

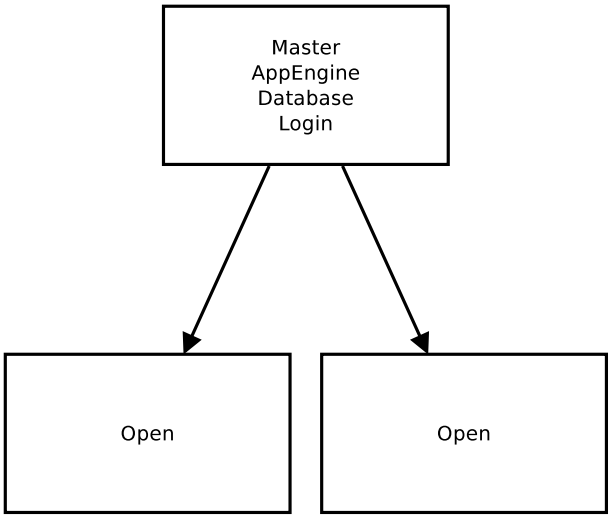

Figura 4.1: Infraestructura AppScale en despliegue personalizado.

Hay multitud de despliegues posibles combinando estos roles, pero será de especial interés para nosotros el que permite ejecutar trabajos de computación en AppScale (Figura 4.1). Un despliegue de este tipo podría conseguirse con un fichero similar a éste:

```
---
:master:    155.210.155.73
:appengine: 155.210.155.73
:database:  155.210.155.73
:login:     155.210.155.73
:open:
- 155.210.155.177
- 155.210.155.178
```

4.3.2 Diseño e implementación de un recurso distribuido para una infraestructura Torque

[Cambiará mucho]

Una infraestructura Torque está formada por un nodo maestro y un conjunto de nodos de computación (Figura 4.2). El nodo maestro es el encargado de recibir los trabajos a ejecutar y de asegurar una correcta planificación para esos trabajos; en su versión más simple el planificador es una cola FIFO. Los nodos de computación son los encargados de ejecutar los trabajos enviados por el nodo maestro y, una vez terminados, enviarle los resultados de vuelta.

La sintaxis del manifiesto distribuido es similar a la usada en el ejemplo de AppScale, sólo que aquí no aparecen los parámetros de administración que aparecían en aquél, ya que Torque

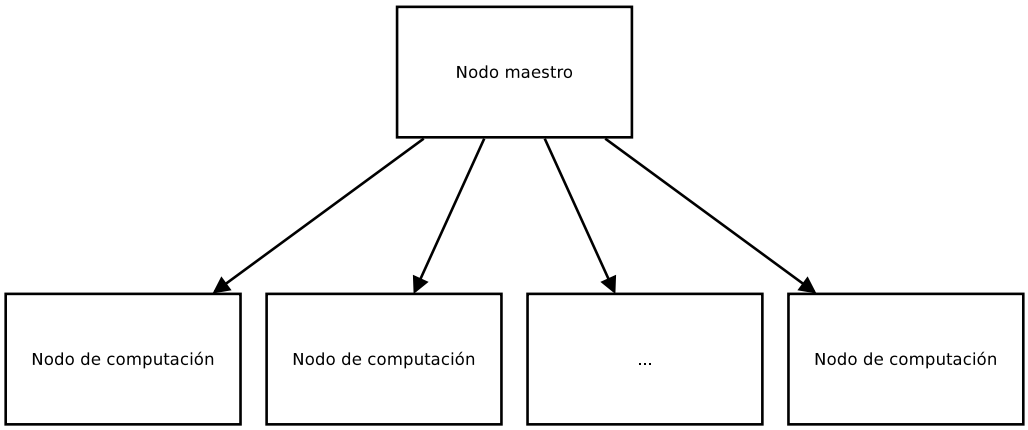

Figura 4.2: Infraestructura Torque.

no requiere su uso. Un ejemplo para la puesta en marcha de una infraestructura Torque sería similar a éste:

```
torque {'mycloud':
   ip_file  => "/etc/puppet/modules/torque/files/jobs-ip.yaml",
   img_file => "/etc/puppet/modules/torque/files/jobs-img.yaml",
   domain   => "/etc/puppet/modules/torque/files/mycloud-template.xml",
   pool     => ["155.210.155.70"],
   ensure   => running,
}
```

En este caso, la parada de la infraestructura es algo más compleja que en el ejemplo de AppScale. Un posible manifiesto de parada sería similar a éste:

```
torque {'mycloud':
   ip_file  => "/etc/puppet/modules/torque/files/jobs-ip.yaml",
   img_file => "/etc/puppet/modules/torque/files/jobs-img.yaml",
   pool     => ["155.210.155.70"],
   ensure   => stopped,
}
```

4.3.3 Diseño e implementación de un recurso distribuido para una infraestructura web de tres niveles

[Cambiará mucho]

Una típica arquitectura de servicios web consta de al menos tres niveles: balanceo de carga, servidores web y base de datos. Cada uno de estos niveles está compuesto por al menos un elemento clave: el balanceador de carga, el servidor web y el servidor de base de datos, respectivamente. El balanceador de carga es el punto de entrada al sistema y el que se encarga, como su nombre indica, de repartir las peticiones de los clientes a los distintos servidores web. Los servidores web se encargan de servir las páginas web a los clientes y para ello, dependiendo de las peticiones que hagan los clientes, podrán leer o almacenar información en la base de datos. Para manipular dicha información los servidores web tendrán que comunicarse con el servidor de base de datos, que es el que hará efectiva la lectura y modificación de la información.

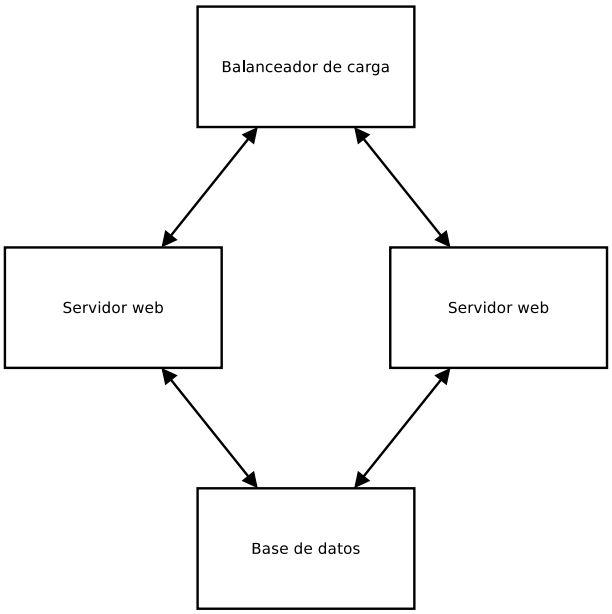

Figura 4.3: Infraestructura web de tres niveles.

Para demostrar la validez del modelo desarrollado se verá como, además de sobre las infraestructuras AppScale y Torque, también se puede aplicar dicho modelo sobre una infraestructura que no tiene nada que ver con la ejecución de trabajos: una infraestructura de servicios web. En el ejemplo se ha validado una infraestructura que consta de un balanceador de carga, dos servidores web y un servidor de bases de datos (Figura 4.3).

La sintaxis del manifiesto de puesta en marcha es fundamentalmente similar a la utilizada en el ejemplo de Torque, ya que tampoco en este caso tenemos parámetros de administración de la infraestructura. Así pues, si se quiere poner en marcha una infraestructura web de tres niveles podemos usar un manifiesto similar a éste:

```
web {'mycloud':
    ip_file  => "/etc/puppet/modules/web/files/web-ip.yaml",
    img_file => "/etc/puppet/modules/web/files/web-img.yaml",
    domain   => "/etc/puppet/modules/web/files/mycloud-template.xml",
    pool     => ["155.210.155.70"],
    ensure   => running,
}
```

En este caso, el manifiesto de parada vuelve a ser sencillo. Un posible ejemplo es éste:

```
web {'mycloud':
    pool   => ["155.210.155.70"],
    ensure => stopped,
}
```

Capítulo 5
Validación de la solución planteada

x Máquinas físicas
y Máquinas virtuales
AppScale: 1 maestro, dos esclavos

Para validar la solución desarrollada, se ha hecho uso del laboratorio 1.03b que el Departamento de Informática e Ingeniería de Sistemas posee en la Escuela de Ingeniería y Arquitectura de la Universidad de Zaragoza. Los ordenadores de este laboratorio poseen procesadores con soporte para virtualización, lo que hace posible la creación de máquinas virtuales para simular los nodos que forman cada una de las infraestructuras distribuidas.

Antes de empezar con las pruebas hay que configurar el entorno en el que se realizarán. En particular, y dado que las máquinas virtuales necesitan conectarse a internet para la descarga e instalación de paquetes, el uso de un servidor de DNS es bastante recomendable. De este modo, podemos usar direcciones IP públicas (que no se estén usando en ese momento) para nuestras máquinas virtuales. El servidor DNS se usa también para hacer la resolución de nombres, tanto normal como inversa, que requieren AppScale y Torque para su correcto funcionamiento.

5.1 Pruebas comunes a todas las infraestructuras

En todas y cada una de las infraestructuras se han realizado las siguientes pruebas para comprobar el correcto funcionamiento del proveedor distribuido:

- Apagado de una máquina virtual que había empezado encendida y no era líder.
- Apagado de una máquina virtual que no había empezado encendida y no era líder.
- Apagado de una máquina virtual que había empezado encendida y era líder.
- Puesta en marcha de la infraestructura desde una máquina que no pertenece al *cloud*.
- Puesta en marcha de la infraestructura desde una máquina que pertenece al *cloud*.

5.2 Prueba de infraestructura AppScale

5.3 Prueba de infraestructura Torque

Para probar la infraestructura Torque se han usado cuatro (de momento) máquinas virtuales alojadas en X máquinas físicas. Una de las máquinas virtuales actúa como nodo maestro y las

otras tres (de momento) actúan como nodos de computación. Además de las pruebas comunes, para comprobar el provedor de la infraestructura Torque se han realizado las siguientes pruebas:

- Parada del proceso de autenticación (trqauthd) en el nodo maestro.
- Parada del proceso servidor (pbs_server) en el nodo maestro.
- Parada del proceso planificador (pbs_sched) en el nodo maestro.
- Parada del proceso de ejecución de trabajos (pbs_mom) en un nodo de computación.
- Parada del proceso que monitoriza al proceso de autenticación en el nodo maestro.
- Parada del proceso que monitoriza al proceso servidor en el nodo maestro.
- Parada del proceso que monitoriza al proceso planificador en el nodo maestro.
- Parada del proceso que monitoriza al proceso de ejecución de trabajos en un nodo de computación.

5.4 Prueba de infraestructura web de tres niveles

Para probar la infraestructura web se han usado cuatro máquinas virtuales repartidas entre X máquinas físicas. Una máquina virtual actúa como balanceador de carga, dos actúan como servidores web y la última actúa como base de datos. Las pruebas que se han realizado para comprobar el correcto funcinamiento del proveedor de la infraestructura web han sido:

- Parada del proceso balanceador de carga.
- Parada del proceso servidor web.
- Parada del proceso base de datos.
- Parada del proceso que monitoriza al proceso balanceador de carga.
- Parada del proceso que monitoriza al proceso servidor web.
- Parada del proceso que monitoriza al base de datos.

Capítulo 6
Conclusiones

En este proyecto fin de carrera se ha creado un modelo de recursos distribuidos para facilitar la puesta en marcha y la automatización de infraestructuras distribuidas. Para comprobar la validez del modelo se ha extendido una herramienta de gestión de la configuración a la que se le ha añadido el recurso cloud. Posteriormente se ha comprobado la extensión realizada haciendo que la herramienta de gestión de configuración sea capaz de administrar infraestructuras distribuidas de ejecución de trabajos como AppScale y Torque. Además se ha visto que el modelo también es válido para infraestructuras más generales como la de servicios web en tres niveles.

Bibliografía

[1] Puppet labs: The leading open source data center automation solution. `http://www.puppetlabs.com/`.

[2] AppScale: An open-source implementation of the Google AppEngine (GAE) cloud computing interface. `http://appscale.cs.ucsb.edu/`, 2011.

[3] AppScale: EC2 API Documentation. `http://code.google.com/p/appscale/wiki/EC2_API_Documentation`, 2011.

[4] AppScale: MapReduce API Documentation. `http://code.google.com/p/appscale/wiki/MapReduce_API_Documentation`, 2011.

[5] AppScale: Neptune API Documentation. `http://www.neptune-lang.org/`, 2011.

[6] Amazon: Elastic Compute Cloud. `http://aws.amazon.com/ec2/`, 2012.

[7] Google: App Engine. `https://developers.google.com/appengine/`, 2012.

[8] Chris Bunch, Navraj Chohan, Chandra Krintz, and Khawaja Shams. Neptune: a domain specific language for deploying hpc software on cloud platforms. In *Proceedings of the 2nd international workshop on Scientific cloud computing*, ScienceCloud '11, pages 59–68, New York, NY, USA, 2011. ACM.

[9] Navraj Chohan, Chris Bunch, Sydney Pang, Chandra Krintz, Nagy Mostafa, Sunil Soman, and Richard Wolski. Appscale: Scalable and open appengine application development and deployment. In *CloudComp*, pages 57–70, 2009.

[10] LaTeX project team. *LaTeX documentation*. `http://www.latex-project.org/guides/`.

[11] Puppet Labs. Puppet Faces API. `http://puppetlabs.com/faces/`, 2012.

[12] Garrick Staples. Torque resource manager. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, SC '06, New York, NY, USA, 2006. ACM.

[13] David Thomas, Chad Fowler, and Andrew Hunt. *Programming Ruby. The Pragmatic Programmer's Guide*. Pragmatic Programmers, 2004.

[14] J. Turnbull and J. McCune. *Pro Puppet*. Pro to Expert Series. Apress, 2011.