Rocks and Rolls: Sistemas de Computación de Alto Rendimiento de Fácil Despliegue

Rocks and Rolls: Easily Deployable HPC Systems for Everyone

Jorge I. Zuluaga
Grupo de Física y Astrofísica Computacional, FACom, Instituto de Física, Universidad de Antioquia, Colombia
jzuluaga@fisica.udea.edu.co

Resumen

Rocks and Rolls es una distribución de Linux y un paquete de herramientas especialmente diseñadas para el despliegue y la administración de sistemas de computación, almacenamiento y visualización distribuida con un mínimo costo administrativo y un alto nivel de flexibilidad y personalización. Rocks es hoy por hoy una de las soluciones más populares en el mundo de la computación en clusters y ha sido utilizada para el montaje de sistemas para HPC a pequeña y gran escala. Las problemáticas básicas del montaje, instalación y administración de un sistema para HPC, las soluciones ofrecidas por Rocks para esas mismas problemáticas, algunos detalles del funcionamiento del original sistema de instalación y configuración de la herramienta y la enumeración de algunas experiencias significativas con el uso de Rocks a nivel mundial en general y localmente en Colombia son presentadas en este trabajo.

Abstract

Rocks and Rolls is a Linux distribution and a toolkit specially designed for the deploy and management of distributed computing, storage or visualization systems at a very low administrative cost and a high level of flexibility and customization capabilities. Rocks is today one of the most populars solutions in the Cluster Computing community and it has been used for the deploy and use of HPC systems at large and small scales. The basic problems faced in the installation and management of HPC systems, the solutions offered by Rocks to these problems, several details of its original installation system and the enumeration of several significative experience in the world in general and locally in Colombia are presented in this work.

1. Introducción

El montaje y la instalación de un sistema de computo intensivo es un proceso que para quien nunca lo ha hecho o visto hacer puede parecer de entrada verdaderamente intimidante. Y no es para menos. Suponga por un momento que su jefe inmediato o sus propias necesidades (casí igualmente demandantes) le pide montar un sistema que sea capaz de incrementar en un factor indeterminado el poder de computo que tiene a su disposición, contando para ello con el más reciente conjunto de herramientas para computo paralelo, compiladores y librerias, un sistema de ejecución de procesos por colas y un sistema de almacenamiento distribuído que le permita guardar esas grandes cantidades de información que producen sus más exigentes aplicaciones y que superan con creces el tamaño de los más grandes discos duros que puede conseguir. Todo, eso sí, con la condición de utilizar software completamente gratuito y de poner aquellas máquinas que por distintas razones usted dejo de utilizar hace un rato a trabajar como una sóla o esas costosas estaciones de trabajo que adquirió sin saber todavía lo complicado que era todo esto.

Naturalmente además esta obligado a hacerlo en el menor tiempo posible: no puede tomarse muchos días para montar un sistema que se supone le va a ahorrar tiempo! Pero las condiciones no acaban ahí: el sistema resultante debe estar formado en la medida de sus posibilidades por nodos con configuración y versiones de los paquetes practicamente idénticos y debe ser también posible hacer una actualización del sistema con el mínimo costo administrativo. Esto sin contar con que si las cosas salen bien eventualmente su jefé o sus necesidades lo podrían obligar a hacer crecer el sistema y usted debe estar preparado para que lo que instaló escale apropiadamente.

Este intimidante panorama es el que enfrantemos normalmente todos los administradores de sistemas de

computo distribuído alrededor del mundo. Naturalmente desde que existe la profesión muchas soluciones a estas problemáticas han sido creadas por ingeniosos equipos de desarrolladores en la misma situación que nosotros o en situaciones peores (Sclyd, SCE, Warewulf, OSCAR, etc.). Esto no es nada nuevo. Sin embargo encontrar una herramienta que sea capaz de ofrecer solución a casí cualquiera de los anteriores condicionantes y problemas, manteniendo al mismo tiempo un nivel de exigencia relativamente bajo sobre lo que debe saber o manejar el administrador del sistema, parecería hasta hace algunos años una tarea imposible.

Hasta hace muy poco los administradores de clusters se veían obligados a dominar el uso de múltiples herramientas para realizar por separado todas esas tareas. En otros casos si bien una misma herramienta podía ser capaz de ofrecer un ambiente unificado para el montaje, instalación y configuración de un sistema para HPC su uso estaba limitado a ambientes altamente homogéneos o era necesario un conocimiento a profundidad de la arquitectura del sistema. Esto fue así hasta que nació Rocks [10]. Este sistema de despliegue rápido, configuración y administración eficiente de clusters, ha sido creado para resolver si no todas, la mayor parte de las más importantes necesidades que el administrador de sistemas HPC tiene a la hora de montar un cluster. Además Rocks se presenta como una herramienta altamente configurable y flexible que permite al desarrollador crear y configurar sus propias componentes que encajan naturalmente en una arquitectura muy bien diseñada para estos propósitos [4]. Los conocimientos requeridos para la instalación y puesta a punto de un sistema para HPC corriendo Rocks son mínimos y el tiempo para su puesta en producción es uno de los más pequeños de todo el mundo de las plataformas para HPC.

Este artículo presenta a Rocks (si es que necesita realmente una presentación.) En la sección 2 se exponen las principales dificultades que se enfrentan al montar y configurar un sistema para HPC y algunas de las necesidades específicas para su eficiente administración y actualización; allí mismo se describen algunas de las soluciones creadas con Rocks para superar estos inconvenientes. La sección 3 describe brevemente el novedoso y muy potente sistema de instalación no supervisada basado en el sistema Kickstart de Red Hat. La sección 4 presenta algunos detalles de la configuración y administración de los servicios, actualizaciones e instalación de paquetes en un sistema HPC que corre Rocks. Finalmente la sección 6 enumera algunos de casos importantes de utilización de Rocks a nivel mundial y algunas de nuestras experiencias localmente en Colombia con el uso de la herramienta.

El propósito principal de este trabajo es el de introducir la herramienta a aquellos que no la conozcan o que tengan apenas información de ella pero no la hayan utilizado realmente. Quienes conocen y han usado Rocks es posible que encuentren aquí razones adicionales para quererlo aún más.

2. Los retos administrativos del HPC

Cada vez es más común el uso de Clusters para resolver problemas en muy diversas disciplinas científicas y técnicas. Su uso en la solución a problemas que los modelos tradicionales de computación no pueden resolver o resuelven en tiempos prohibitivamente grandes es cada vez más extendido. Pero el montaje y la administración de un sistema de computación de alto rendimiento requiere normalmente un alto nivel de experticia o la disposición de un grupo de personas con destrezas específicas y complementarias. Estos últimos condicionantes son díficiles de satisfacer en general pero muy especialmente en nuestra región (Latinoamérica) y más en ciertos paises donde la inversión en este tipo de tecnologías esta apenas comenzando. Es por ello que contar con herramientas que simplifiquen y faciliten el montaje de este tipo de sistemas con un mínimo de conocimientos y que permitan un acceso practicamente inmediato a plataformas para computo distribuído es una imperiosa necesidad. Esto es especialmente cierto en la Ciencia donde los usuarios requieren soluciones casí inmediatas a su necesidad de disponer cada vez un mayor poder de computo.

¿Cuáles son los retos que enfrenta el administrador de un sistema de computo distribuido cuando asume la tarea de hacer operativo y dar mantenimiento a un sistema como estos? Como una estrategia para describir apropiadamente las características de Rocks, en los siguientes paragrafos enumeraremos las que consideramos son las más importantes dificultades y/u obstáculos que enfrentamos al instalar y poner en funcionamiento una plataforma para HPC y la estrategia de Rocks para superarlas.

2.1 Problema 1: el sistema de base

Administrador. Desde el primer momento en el que comenzamos con la instalación de un sistema como estos enfrentamos el primer problema: ¿cuál es la distribución que debemos utilizar para instalar el sistema operativo base en la plataforma? ¿cuáles son los paquetes básicos que se requieren para empezar a

hacer HPC y cómo puedo agregar otros en el camino? Estas aunque parecen preguntas sencillas encierran decisiones importantes que con el tiempo aprendemos a valorar mejor. Normalmente un administrador de HPC se decanta por una distribución dada, o bien porque es la que mejor conoce o porque sabe de antemano que cuenta con software preinstalado para HPC.

Rocks. Rocks es una distribución basada en Red Hat Linux una ampliamente reconocida y muy popular distribución que cuenta con una extensa comunidad de desarrolladores y usuarios que garantizan un continuo ritmo de actualización en los paquetes y la disposición de abundante material de ayuda. Adicionalmente el sistema de instalación de Red Hat, el Red Hat Package Manager (RPM) se ha convertido en lo que podríamos llamar un *standard de facto* en la distribución de paquetes para el sistema operativo Linux. Adicionalmente Red Hat tiene un ingenioso y robusto sistema para realizar instalaciones y configuraciones no supervisadas de una máquina, *Kickstart*, un mecanismo que se encuentra en la base del funcionamiento de los procedimientos de automatización de instalación en Rocks.
El sistema de base para HPC que es instalado con Rocks (HPC Roll) cuenta con herramientas estándar para utilizar la plataforma para cálculo distribuído. Todo esto sin mencionar el hecho de que la configuración del sistema se prepara automáticamente para que las componentes compartan la información de configuración (incluyendo información de autenticación de los usuarios) sin necesidad alguna de intervención administrativa o de configuración después de la instalación.

2.2 Problema 2: la "golden image"

Administrador: el proceso de instalación de las componentes de computo del sistema (nodos de computo) en muchos sistemas procede preparando la que se llama una "golden image" de la distribución [12]. La instalación normalmente implica que esta imágen de base se distribuye a través de la red en los distintos nodos del sistema. Sin embargo el uso de imágenes exhibe una serie de inconvenientes que la ponen en desventaja respecto a otras estrategias. El primero es que pone en aprietos al administrador cuando se enfrenta ante un sistema de recursos heterogeneos. En ese caso es menester por ejemplo contar con tantas imágenes del sistema como arquitecturas distintas encontremos en la plataforma. Adicionalmente la creación de imágenes y su tranferencia puede ser un proceso tedioso y muy prolongado (una imágen bit a bit comprimida pesa algo menos que su tamaño total que puede llegar a ser de varios gigabytes). Después de la instalación de la imágen a menudo se requieren tareas de reconfiguración para poner a punto el nodo en el que es instalada. Existen alternativas para la creación de imágenes. Entre ellas se encuentra el uso de nodos diskless en los que realmente no hay sistema operativo instalado en los discos locales (no es necesaria la imágen) y el sistema que ejecutan las máquinas es cargado en memoria desde el servidor principal del sistema (en lo sucesivo *frontend*). Este mecanismo tiene la desventaja de utilizar masivamente la red cada que una máquina debe arrancar nuevamente. Otros sistemas utilizan mecanismos más ingeniosos como la creación de una imágen usando la lista de archivos en lugar de los archivos mismos. Aquí sin embargo nos enfrentamos nuevamente al problema de la transferencia por la red de los archivos durante la instalación.
En otras ocasiones el administrador toma la decisión de hacer la instalación de cada nodo independientemente. Esta que puede ser una decisión práctica para clusters con un número pequeño de nodos es a todas luces impráctica cuando se tiene un sistema formado por muchísimos nodos. La configuración manual esta también afectada por el hecho de que durante el proceso de configuración algunos errores pueden cometerse con el respectivo problema de sincronización entre el friontend y el nodo o entre los nodos mismos.

Rocks. La solución que da Rocks al problema de la instalación sistemática y masiva del sistema operativo en los nodos esta entre sus más originales características y será descrita con algún detalle en la sección 3. Basta con decir que Rocks no utiliza un sistema de instalación por imágenes sino que se vale del sistema de Kickstart diseñado por Red Hat. Con este sistema el instalador de un nodo primero construye usando las especificaciones publicadas en el frontend un archivo de Kickstart especialmente diseñado para ese nodo. A partir de ese archivo de Kickstart y usando el servidor web del frontend descarga desde este último los paquetes que se van a instalar, los instala y además los configura después de la instalación. En tiempos recientes los desarrolladores de Rocks diseñaron el que se denomina "Avalanche Installer" [11] que permite la instalación eficiente de muchos nodos simultaneamente sin la esperada sobrecarga del frontend, mediante un esquema de comunicaciones peer-to-peer usando para ello los estándares y protocolos propios de BitTorrent (ver sección 3).
Kickstart tiene la ventaja adicional de quitar a los desarrolladores de Rocks el peso del desarrollo de los ahora sofisticados métodos de detección de hardware y

carga de los módulos necesarios para su operación, algo que hace parte intrínseca del propio Kickstart.

2.3 Problema 3: la instalación a ciegas

Administrador. La instalación de los nodos en un sistema de computo distribuido es normalmente un proceso que se lleva a cabo a ciegas. Los nodos por razones naturales no cuentan con interfaces de entrada y sálida por lo que es díficil reconocer que es lo que esta pasando en ellos mientras se realiza la instalación. Este problema normalmente es subsanado con estrategias como la de contar con switches del tipo KVM (Keyboard-Video-Mouse) que permiten alternar entre los distintos nodos un único o pocos dispositivos sálida y la entrada. Esta estrategía tiene tres inconvenientes importantes. El primero es el costo del equipo de switcheo (que puede superar los USD$500 para un equipo con una decena de puertos), el segundo el exceso de conexiones y cableado y el tercero la incapacidad para escalar apropiadamente.

Rocks. Los desarrolladores de Rocks diseñaron un sistema para monitorear el avance del proceso de instalación capturando y transportando a través de la red la sálida estándar del gestor de instalación Anaconda. El sistema, llamado eKV (*Ethernet Keyboard Video*) permite obtener información del proceso de instalación desde la consola del frontend una vez el sistema ha activado la interfaz de red (que ocurre poco después de iniciado el proceso de instalación) [10]. Con este mecanismo se descarta practicamente la necesidad de un dispositivo como el KVM para realizar el control y la instalación de un cluster Rocks.

2.4 Problema 4: la sincronización de los nodos y la actualización del sistema

Administrador. Este es uno de los más delicados e incómodos problemas que se enfrentan durante la administración de un sistema de computo distribuido. Normalmente para que el sistema funcione de forma estable los paquetes instalados del software en el frontend y los nodos deben coincidir tanto en versión como en configuración. Cuando se utiliza un sistema de imágenes esto esta naturalmente garantizado al menos durante la instalación del sistema pero es mucho más dificil de garantizar una vez el sistema esta ya instalado. Cualquier actualización de los paquetes en el frontend o en uno de los nodos debería propagarse a toda la plataforma lo que normalmente es una tarea engorrosa y díficil de conseguir sin errores. Una solución normalmente utilizada por muchos es la de mantener los paquetes en la versión original en la que fueron instalados para evitar desestabilizar la plataforma, pero esto se hace al costo de no actualizar e instalar por ejemplo importantes parches de seguridad. En el mismo orden de dificultades se encuentra la instalación de nuevo software. En instalaciones tradicionales la instalación de ese software requiere un proceso de propagación del mismo a través del sistema y la ejecución de tareas de instalación automaticas en cada nodo. Esto sin mencionar la configuración en cada nodo que nuevamente se debe garantizar igual a la configuración en los nodos restantes.

Rocks. La solución a Rocks para el problema de la sincronización entre los nodos es realmente simple: reinstalar completamente los nodos cuando se produce una actualización en la plataforma. Esta filosofía llamada informalmente "fire and forget" ha sido ampliamente debatida en la comunidad de HPC [10]. Sin embargo todas las experiencias demuestran que si bien es una decisión extrema, es comparativamente, muy sencilla y sobre todo muy efectiva para mantener la sincronización al más alto nivel. Para quienes piensan que la reinstalación implica la perdida de algo hay que comentar que durante este proceso Rocks esta configurado para respetar los datos contenidos en particiones distintas a la partición raíz que es la única cuyo contenido debería cambiar con la reinstalación.
Con esta filosofía de "fire and forget" la instalación de nuevas aplicaciones es trivial. La nueva aplicación se agrega a la distribución que es instalada en los nodos y los nodos se reinstalan nuevamente desde cero.

2.5 Problema 5: el bombardeo a los sistemas de autenticación

Administrador. Este es uno de los problemas más sutiles de la administración y uso de un cluster. La inmensa mayoría de los sistemas de computo distribuido se valen del que ha terminado por convertirse en el *standard de facto* de los sistemas para administración de la información de configuración del sistema: NIS (Network Information System). Los que no utilizan este sistema se valen de otro distinto pero que sufre del mismo tipo de problema: LDAP. La principal limitante de estos sistemas operando en plataformas de computo paralelo es el hecho de que normalmente la autenticación de una instancia de un proceso paralelo en uno de los nodos implica un requerimiento al servidor de información que se encuentra en el frontend. En sistemas donde esos requerimientos se producen de forma permanente pero

no simultanea, ellos funcionan muy bien. Sin embargo al ejecutar un proceso paralelo las distintas instancias del proceso bombardean inmisericordemente el sistema de información y pueden encontrar en el proceso fallas que les impiden utilizar el recurso. En síntesis estos sistemas de autenticación no escalan de forma apropiada y aunque funcionan muy bien en plataformas pequeñas no lo hacen cuando el número de procesadores y de procesos que requieren autenticación crece considerablemente.

Rocks. Los desarrolladores de Rocks han resuelto este problema diseñando un sistema de información llamado 411 [2], que se basa en la replicación agresiva de la información de autenticación (y otra información de configuración) a través de todo el cluster. Todas las operaciones de acceso a la información de los usuarios se realizan en cada nodo y por tanto se distribuyen efectivamente. El sistema 411 mantiene automáticamente actualizado los archivos de configuración del sistema y permite también al administrador ejecutar una actualización forzada a través de todo el cluster. 411 además garantiza la seguridad de la información que circula a través de la red de la plataforma usando para ello canales seguros con encriptación del tipo SSL/TLS. El sistema ha sido probado exhaustivamente, tanto en sus aspectos de seguridad como en los de rendimiento y escalabilidad demostrando su capacidad para satisfacer los requerimientos de hasta las más grandes y congestionadas plataformas.

2.6 Problema 6: la personalización de la plataforma

Administrador. La última dificultad que importante que se reconoce en este trabajo es la poca flexibilidad que ofrecen las herramientas para la personalización del sistema de acuerdo a las necesidades y destinos que se le vaya a dar a la plataforma. Es un hecho bien reconocido que la mayor parte de las herramientas y distribuciones específicas desarrolladas para el montaje de plataformas distribuidas están orientadas a la utilización de esas plataformas especialmente para HPC [8]. ¿Qué pasa si se desea o requiere la instalación de una plataforma que de respuesta a otras necesidades (almacenamiento distribuido, visualización, entre otras)? En la mayoría de los casos el administrador de un sistema como estos debe proceder él mismo a modificar la distribución que instaló en el sistema para adaptarla a las necesidades específicas con las que la esta montando. Este procedimiento implica la desinstalación o deshabilitación de paquetes y servicios y la instalación y habilitación de otros nuevos. Todo esto a un costo administrativo muy alto.

Rocks. Aquí es donde aparece el concepto de "Rolls" [4]. Rocks and Rolls. Con el obejto de ofrecer una adecuada flexibilidad en la configuración inicial del sistema, Rocks permite configurar desde la instalación del frontend el perfil que tendrá la plataforma. Esto se hace mediante unos paquetes especialmente preparados y configurados y conocidos como "Rolls" que contienen las herramientas específicas necesarias para realizar unas u otras tareas (sobre los Rolls volveremos en la siguiente sección). La instalación de un sistema usando Rocks implica la elección del conjunto de Rolls que deben instalarse deacuerdo a las necesidades. Esto permite desde el principio contar con un frontend que tiene solamente los paquetes que necesita y que no requiere la desinstalación de paquetes que para el propósito con el que es creada la plataforma no son necesarios.

Como se menciono al principio hay que tener presentes que existen en el mercado diversas alternativas que permiten resolver algunos o todos los problemas descritos en las subsecciones anteriores. Sistemas como Sclyd, OSCAR, Warewulf, LCFG, LinuxBIOS, OpenMosix entre muchos otros se encuentran entre las más importantes y conocidas alternativas. Sin embargo la mayoría de ellos o bien no poseen todas las características que permiten a Rocks and Rolls resolver de forma natural y sencilla los problemas arriba descritos o bien sus soluciones exigen a veces conocimientos avanzados del manejo y la arquitectura del sistema de computo distribuido.

3. El ciclo de vida de Rocks: detrás de camaras

Si la anterior argumentación todavía no lo ha convencido del poder especial que tiene Rocks and Rolls para poner al alcance de todos sistemas para HPC de fácil despliegue, tal vez necesita ponerse manos a la obra y probarlo. La instalación, configuración y uso de Rocks es una tarea, como era de esperarse muy sencilla y ampliamente documentada (ver [7] y [13] para documentación en castellano inclusive). Más complejo pero igualmente interesante es lo que pasa en el sistema internamente cuando se realizan algunas operaciones fundamentales. Para los más curiosos entender precisamente que ocurre cuando se realiza la instalación de un sistema para HPC usando Rocks ayuda a entender mejor la manera como funciona la

herramienta y ofrece una visión de la maquinaria interna que puede ayudar considerablemente en las tareas de personalización y administración del sistema.

3.1. Instalación del Frontend

La instalación de Rocks se realiza en 3 etapas básicas.

La primera de esas etapas implica la instalación del frontend en un proceso bastante convencional y común a todas las instalaciones de Red Hat. La única particularidad para resaltar en esta etapa es la configuración del sistema que se instalará en el frontend y que será posteriormente heredado a los nodos. En esta fase el instalador permite escoger uno a uno los Rolls que van configurando el carácter de la plataforma (si será una plataforma de computo intensivo, una orientada al cálculo en grid, una para el almacenamiento, etc. o una combinación personalizada de todas ellas.)

Como se menciono en la sección precedente la posibilidad de configurar la distribución que se instala en el frontend desde el momento mismo de la instalación es una ventaja muy grande para el administrador que se ve descargado de la tarea de modificar una instalación prediseñada para adaptarla a sus condiciones partículares.

3.2. Preparación de la distribución para los nodos

La segunda etapa de la instalación del cluster es la personalización y preparación de la(s) distribución(es) que será(n) utilizada(s) para la instalación de los recursos (appliances) que se conectan a la plataforma. Esta es la fase equivalente a la preparación de la "golden image" y en su diseño ha trabajado el equipo de desarrollo de Rocks desde el surgimiento de la herramienta en el 2000.

Como se había adelantado en la sección 2 Rocks utiliza un original mecanismo para la descripción y construcción de la distribución que será utilizada para instalar el cluster [4], [8].

En primer lugar el frontend viene dotado de una colección completa de diversas versiones de los paquetes binarios (en RPM) a partir de los cuales se instalará el sistema. En el momento de la instalación de los nodos esos paquetes son transferidos e instalados de acuerdo a lo dictado por un script de instalación de Kickstart. Este último script representa en sí mismo la descripción de la instalación que será puesta en cada nodo. Sin embargo el contenido de ese script dependerá de forma a veces importante del tipo del recurso que se este instalando. Rocks clasifica los recursos en 2 tipos básicamente: frontend y compute node. Otros tipos pueden ser también definidos (login node, web server, file server, visualization engine, etc.) dependiendo del tipo de plataforma que se este construyendo. Además la arquitectura de los recursos podría ser diversa conteniendo el cluster por ejemplo máquinas de arquitecturas i386 y x86 o con interfaces de red diversas (ethernet y myrinet). La posibilidad de Rocks para adaptarse a estas diversas condiciones de instalación es una de sus más reconocidas fortalezas. Pero ¿como lo logra? Es bien reconocido que los scripts de instalación del sistema Kickstart tienen muy poca capacidad para adaptarse a condiciones variables y no permiten, excepto por la disposición de un script por cada configuración, una aproximación programática para la solución del problema de la configuración a instalar en un determinado recurso.

Rocks utiliza un novedoso y original sistema para construir durante la instalación el archivo Kickstart que será utilizado en particular por el recurso que esta instalándose. El sistema, que utiliza un software construído con un lenguaje descriptivo (XML) usa como base un grafo directo con el que se describen las distintas componentes de la instalación (nodo) y el orden o la relación de instalación entre ellas (conectores). Esta aproximación utiliza además estrategias de modularización que dan una increible flexibilidad al sistema. Cada nodo del grafo es un módulo aparte y puede contener a su vez un pequeño grafo que describe la manera como debe ser instalado y configurado el servicio respectivo [8].

Los denominados grafos de Kickstart son las entidades fundamentales del sistema de instalación de Rocks al contener la información completa sobre la construcción de la distribución de cualquier recurso que se agremie en el cluster.

Los grafos de Kickstart permiten construir la distribución de Rocks que será instalada en los recursos del cluster. La construcción de esa distribución dista sin embargo de ser una "golden image". Todos los archivos que pueden ser utilizados por los recursos en el momento de la instalación se reúnen en un directorio de la distribución que puede ser accedido a través de servidores tftp o http en el frontend. Pero los binarios que constituyen la distribución están allí en la forma de enlaces simbólicos a los archivos reales que se encuentran en otro gran repositorio de binarios que reposa en el frontend. Esto hace que la distribución sea muy liviana (tan sólo unas decenas de megabytes para la distribución de un nodo típico) en contraste con el tamaño real de los archivos binarios que pueden sumar unos centenares de megabytes. La distribución

contiene también todos los módulos de los grafos de Kickstart que permitirán en su momento construir el script maestro con el que Red Hat procederá a la instalación.

3.3. Instalación de los nodos

La última etapa de instalación de Rocks and Rolls en la plataforma es la instalación misma de los nodos o los recursos que se agremian en el cluster.

La instalación (de un nodo de computo por ejemplo) normalmente procede iniciando el nodo usando PXE (*Preboot Execution Environment*). El nodo se conecta al servidor DHCP instalado en el frontend que le asigna una IP y le entrega un kernel para comenzar el proceso de instalación. Entre tanto en el frontend un aplicativo especial (insert-ethers) captura la información del nodo y alimenta una base de datos SQL con el par (Mac Address, IP) que sirve al sistema para la generación automática de distintos archivos de configuración (otra característica original de Rocks). Una vez iniciado el nodo se ejecuta en el frontend un script CGI que comienza el proceso de construcción del script de Kickstart que será entregado al instalador Anaconda para la instalación y configuración de los paquetes. En la versión actual la construcción del archivo de Kickstart que puede ser un proceso que consume recursos de computo y es distribuída (una parte se realiza en el servidor y otra en el nodo mismo). Una vez concluida la construcción del script de Kickstart la instalación procede de la manera convencional.

A partir de 2006 (versión 4.1) Rocks implemento un nuevo y muy eficiente sistema para la instalación simultánea de muchos nodos en un gran cluster. El sistema conocido como "Avalanche Installer" [11] se basa en mecanismos y herramientas provenientes del mundo de las comunicaciones peer-to-peer. El Avalance Installer de rocks utiliza un sistema trazador de BitTorrent para conectar en una red virtual de usuarios que comparten los binarios de instalación a los nodos que están en un momento dado realizando el proceso. Con este sistema se reduce considerablemente las exigencias sobre el servidor http del frontend y se delega a una red de clientes peer-to-peer, incrementando efectivamente el ancho de banda del proceso y reduciendo considerablemente el tiempo de instalación, tal y como lo demuestran los experimentos hechos con el sistema [11].

4. Rocks and Rolls en acción

Desde el surgimiento mismo de la herramienta en el año 2000 el número de usuarios de Rocks, es decir de clusters instalados y configurados usando la herramienta, crece a un ritmo constante tal y como lo constanta el propio sistema de registro de Rocks que se puede acceder a través de su sitio web http://ww.rocksclusters.org. Actualmente Rocks es utilizado en más de 1000 clusters alrededor del mundo (1191 registrados para la fecha en la que se preparo este artículo) que suman un total de alrededor de 54000 CPUs con un poder conjunto conjunto de más de 220 Tflops. Dentro de la lista estan registrado clusters que tienen desde 2 máquinas hasta grandes plataformas con más de 1200 nodos. Algunos de ellos están en las listas del top500 entre los que resalta el cluster Jaws del Multi High-Performance Computing Center (MPHCC) que ocupa la posición 16 de la última lista publicada en Junio de 2007 y el Tungsten 2 de la NCSA que se encontraba en la posición 108 la última vez que fué reportado.

El poder de Rocks para instalar y configurar un cluster de altas prestaciones en tiempo record fue probado en Noviembre de 2002 (apenas 2 años después de ser liberada la primera versión de la herramienta) cuando un "metacluster" formado por la fusión de otros dos cada uno con 128 procesadores fue montado, desplegado y probado en el lapso de 40 horas en el Scripps Institution of Oceanograhy (SIO) [9]. El éxito de la operación fué confirmado por el hecho de que la plataforma paso a ocupar inmediatamente la posición 233 del top500 después de las pruebas realizadas con la herramienta.

Otro de los escenarios en los que Rocks esta empezando a resaltar por sus especiales características (fácil despliegue y administración y flexibilidad en la personalización del sistema) es en el montaje de sistemas Grid. En los casos existosos que se reportan Rocks ha sido utilizado para la instalación y configuración de los clusters que se asocian en el Grid. Rolls particulares han sido especialmente desarrollados para crear instalaciones homogeneas de herramientas y servicios en los clusters que se conectan a la ciberinfraestructura.

Entre las experiencias bien conocidas de Rocks como sistema de base en Grids se encuentran la del Thai Grid del National Grid Center en Thailandia, el GEON Grid (GEO Science Network) en los Estados Unidos, el CL Grid la iniciativa de Grid Nacional en Chile, entre otros.

Localmente y en especial en la ciudad de Medellín Rocks ha sido utilizado para el montaje de practicamente todos los sistemas de computo intensivo con los que cuenta por ejemplo la Universidad de Antioquia incluyendo su recien adquirido Centro de Simulación y Cálculo Avanzado. Algunos de esos clusters (especialmente los operados por grupos de

investigación particulares) son administrados por personal con una experiencia mínima en la adminsitración de servidores Linux y mucho menos en la administración de sistemas de computo distribuido. Aún así esas plataformas pudieron ser instaladas y configuradas sin inconvenientes y son activamente utilizadas en la investigación en Ciencias e Ingeniería. Un proyecto de documentación completa en castellano de la instalación, configuración, administración y uso de Rocks fue iniciado en el Instituto de Física de la misma Universidad en Noviembre de 2006 y es actualmente reconocido como una contribución a la comunidad de Rocks por su equipo de desarrollo [13]. Una interesante iniciativa ha surgido también en el seno de la comunidad académica de usuarios de Rocks en Colombia. Se trata del desarrollo de un *know how* para el montaje de clusters parcialmente dedicados usando equipos de salas de computo públicas que puedan cumplir la misión doble de prestar el servicio a investigadores que requieran de una plataforma de computo distribuido en instituciones donde adquirir un recurso dedicado sea muy costoso y al mismo tiempo ofrecer los servucios propios de una sala incluyendo el uso del sistema operativo Windows [5], [6].

También el Frente de Operación y Configuración de Grid Colombia, la iniciativa de Grid Nacional en nuestro país, viene estudiando la posibilidad de utilizar Rocks como sistema base para la instalación de los recursos del Grid. Las primeras pruebas de interconectividad y configuración ya comenzaron y usan esta filosofía [1].

7. Agradecimientos

El autor agradece al Instituto de Física de la Universidad de Antioquia por la oportunidad que me dió al ponerme hace algunos años frente a una sala de máquinas inutilizadas, con conocimientos mínimos de montaje y administración de clusters y con un jefé muy exigente. Todo ello me obligó a descubrir las bondades de Rocks and Rolls y me permitió satisfacer las elevadas demandas y los estrechos calendarios que se me impusieron. Los recursos necesarios para elaborar este trabajo y participar en la Conferencia de Computación de Alto Rendimiento (CLCAR-2007) fueron proveidos por la vicerrectoría de Docencia y la Decanatura de la Facultad de Ciencias Exactas y Naturales de la Universidad de Antioquia.

6. Referencias

[1] A. Ospina, J. Zuluaga, "Pruebas de conectividad y procesamiento distribuido con Clusters Beowulf usando la Red Regional de Antioquia de Tecnología Avanzada (RUANA)," Proyecto de investigación, 2007, CIDI, UPB, Código 092A-06/07-20.

[2] Federico D Sacerdoti, Mason J. Katz, and Philip M. Papadopoulos, "411 on Scalable Password Service," July 2005, IEEE High Performance Distributed Computing Conference, North Carolina.

[3] Federico D. Sacerdoti, Sandeep Chandra, and Karan Bhatia "Grid Systems Deployment & Management Using Rocks," September 2004, IEEE International Conference on Cluster Computing, San Diego.

[4] Greg Bruno, Mason J. Katz, Federico D. Sacerdoti, and Philip M. Papadopoulos, "Rolls: Modifying a Standard System Installer to Support User-Customizable Cluster Frontend Appliances," IEEE International Conference on Cluster Computing, San Diego, September 2004.

[5] J. Zuluaga, A. Ospina "Consideraciones metodológicas para Instalación y Configuracion de una Sala Cluster," Revista OMEGA, No. 19, ISSN 1692-0872, 2007.

[6] J. Zuluaga, A. Ospina "Installation and Configuration of a Cluster-Room as a Low Cost Solution for the Access to Distributed Computing Technologies in Latin America," Conferencia Latinoamericana de Computación de Alto Rendimiento. En este mismo volumen, 2007.

[7] J. Zuluaga, D. Mejía. "Instalación y Uso de una Sala Cluster usando NPACI Rocks (partes 1, 2 y 3)." Taller presentado durante el Encuentro de Investigación sobre Tecnologías de la Información Aplicadas a la Solución de Problemas, EITI2005: "El uso del paralelismo en las soluciones informáticas," ISBN 958655895-0, (2005).

[8] Mason J. Katz, Philip M. Papadopoulos, and Greg Bruno "Leveraging Standard Core Technologies to Programmatically," April 2002, Cluster 2002: IEEE International Conference on Cluster Computing.

[9] Philip M. Papadopoulos, Caroline A. Papadopoulos, Mason J. Katz, William J. Link, and Greg Bruno, "Configuring Large High-Performance Clusters at Lightspeed: A Case Study," Clusters and Computational Grids for Scientific Computing, December 2002.

[10] Philip M. Papadopoulos., Mason J. Katz, and Greg Bruno, "NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters," Concurrency and Computation: Practice and Experience Special Issue: Cluster 2001.

[11] Rocks Group, "The Rocks Avalanche Installer. A two-page white paper that provides an overview of the Rocks Avalanche Installer." (introduced in Rocks v4.1, code-named "Fuji").

[12] T. L. Sterling, J. Salmon, D. J. Becker, Savarese, and D. F. Savarese. "How to Build a Beowulf: A Guide to the Implementation and Application of PC Clusters," MIT Press, 1999.

[13] J. Zuluaga, "Linux Clustering con Rocks: una guía práctica," disponible en línea en el sitio de Rocks Clusters. Disponible en línea en http://www.rocksclusters.org.