CLICLUX: Cloner for LInux CLUsters. Una herramienta para instalación de clusters usando Linux

Andrés Barbieri*

Instituto de Investigación en Informática LIDI †
Facultad de Informática - Universidad Nacional de La Plata.

Resumen

Los clusters forman una plataforma fundamental para el cómputo paralelo. La proliferación de estos sistemas desde el gran impulso dado por Beowulf a aumentado día a día. Hoy llegan a estar en los primeros diez del *top500*. De cualquier forma, no todo es "enchufar y correr", es necesario administrar la gran potencia de cálculo que ofrecen para poder aprovecharlos al máximo. Un instancia de esta tarea es la instalación del software (sistema operativo, middlewares y aplicaciones) para el cómputo paralelo sobre estos equipos, cuestión que no resulta trivial cuando se tiene una cantidad de máquinas relativamente importante. En este trabajo se presenta una herramienta para clonación/instalación de sistemas para clusters.

Palabras Clave: Computación Paralela, Computación sobre Clusters, Instalación Distribuida, Sistemas de Imagen Unica.

1 Introducción

Los cluster forman la arquitectura que ofrece la mejor relación rendimiento/costo dentro de las arquitecturas paralelas. Esto ha sido demostrado por el proyecto Bewoulf [BWLUF] de la NASA, que aún continúa dando frutos. Una red de PCs homogéneas o heterogénea es posible en cualquier Universidad o Empresa Argentina, además, la existencia de una gran cantidad de herramientas para desarrollo ([PVM], implementaciones de [MPI]) de aplicaciones para esta plataforma la convierte en la mejor alternativa para el área del cómputo paralelo. Sin embargo, estas características no son fáciles de aprovechar si se tiene en cuenta el tiempo necesario para la instalación y administración del equipamiento. Las personas que desean tener un cluster para aprovecharlo como máquina paralela, primero deben aprender las herramientas que existen para su explotación y mantenimiento. Luego se debe proceder a su instalación máquina por máquina hasta tener todo el conjunto instalado. En este trabajo se presenta una solución a la tarea de instalación del software a un cluster llamada: **CLICLUX** . El principal objetivo de **CLICLUX** es disminuir el tiempo requerido y facilitar al usuario una serie de herramientas

*Becario de la UNLP. III-LIDI. Facultad de Informática. UNLP. barbieri@lidi.info.unlp.edu.ar
†III-LIDI miembro del Instituto de Investigación en Ciencia y Tecnología Informática (IICyTI). Calle 50 y 115 - (1900) La Plata, Argentina

para dejar "listo para correr" aplicaciones paralelas sobre este hardware. Si bien existen otras herramientas [OSCAR] con características similares **CLICLUX** busca ser lo más sencillo posible.

2 Implementación de CLICLUX

CLICLUX está destinado a trabajar sobre clusters con [GNU/Linux]. El instalador está basado en un mini sistema con las herramientas necesarias para la instalación. La implementación necesita básicamente un cliente `ftp` para mover los archivos hacia la máquina a clonar, el shell `bash` y un kernel con soporte de módulos para cargar la placa de red, TCP/IP dentro del kernel y con soporte de partición root en ramdisk.
Los archivos de instalación son scripts creados todos en `bash`. Si bien este no es el lenguaje de scripting más flexible comparado con `perl` o `tcl-tk` es el que menos recursos necesita (bibliotecas, memoria, disco) , y menos problemas de versiones tiene. Los scripts utilizan varios comandos de edición de textos como el `awk`, `egrep`, `ed` y `sed`. La configuración de las máquinas a clonar es generada editando un archivo común y un archivo por máquina.
Además de los scripts de instalación **CLICLUX** usa un kernel booteable desde diskette, un mini-GNU/Linux y un pool de drivers para placas ethernet.

3 Instalación con CLICLUX

3.1 Sistema Inicial

La primera etapa en la instalación de un cluster usando **CLICLUX** es la instalación del equipo denominado *master*. A partir de este se configuran los programas para ser clonados luego en el resto de los equipos. En esta etapa se debe seleccionar una distribución de GNU/Linux e instalarla en una máquina que formará parte del cluster. Durante esta fase se debe instalar el kernel del GNU/Linux, las herramientas básicas como shells, editores, lenguajes de desarrollo, manejo de red con TCP/IP y servicios de TCP/IP.
Una vez instalado GNU/Linux se deben agregar al equipo *master* las componentes a ser usadas en el cómputo paralelo. En este caso se instaló solamente PVM y LAM/MPI, pero se puede extender a otras herramientas como alguna implementación de BSP, herramientas para el manejo de colas como [PBS] o herramientas para la administración global como [C3].
Para la administración en lugar de C3, se desarrollo un conjunto mínimo de herramientas para la administración (scripts de línea de comando) que permiten el testeo global del cluster, apagado remoto, distribución y recolección de archivos de configuración. Estas herramientas están basadas en en los *r-commands* por lo cual éste es un requisito que se debe cubrir cuando se instala GNU/Linux.
Una vez que se tiene instalado y configurado el sistema en la máquina *master* se deben agregar en un archivo la lista del resto de las máquinas a ser clonadas. A partir de este archivo se configuran las herramientas a ser usadas en el cluster. Lo último en hacerse en esta etapa es la ejecución de un programa archivador que genera para cada directorio a partir del raíz una imagen a ser enviada a cada uno de los *clones*.

3.2 Configuración de los parámetros

La segunda etapa es la configuración de los parámetros de las máquinas a clonar. Primero se configura un archivo con los parámetros comunes para todas las máquinas y luego los valores particulares para cada equipo. Esto se hace editando archivos de texto los cuales contienen toda la información necesaria acerca de la misma. Los valores comunes que se necesitan son:

- Dirección IP de red y dirección de broadcast.
- Máscara de red.
- Direcciones de los DNS y nombre de dominio.
- IP del default router.

La información particular que se requiere es:

- Dirección de hardware de la máquina (MAC address).
- Dirección IP de la máquina.
- Nombre de la máquina.
- Archivo del dispositivo de disco de la máquina (`/dev`).
- Tabla de particiones de la máquina.
- Cantidad de memoria.
- Memoria swap.

3.3 Discos de Instalación

La tercera etapa para la instalación es la creación de los discos que se usarán para instalar el resto de las máquinas. En esta etapa se crean 4 discos:

1. Boot Disk.
2. Root Disk.
3. Network Drivers Disk.
4. Install Disk.

El primero tiene un kernel GNU/Linux para bootear. El segundo contiene un mini-GNU/Linux file system que se copia RAM con las bibliotecas y programas básicos que utilizan los scripts de instalación. El tercero contiene un pool de drivers para los diferentes chipsets de placas de red usadas en las máquinas a clonar. El último del los tiene los scripts de instalación.
Las imágenes de los diskettes son creadas y almacenadas en la máquina denominada *master* y se pueden copiar/extraer de la misma con un script. La cantidad de diskettes necesarios se puede reducir a uno si se utiliza NFS-server en la máquina *master*. EL kernel que se requiere debe tener soporte para NFS-root y estar compilado con el driver de la placa de red. Con esta herramienta el método por default es vía `ftp` debido a que genera menos complicaciones en cuanto a la configuración del equipo *master*. Una vez que se crearon los diskettes se debe configurar el `ftp` server en el equipo *master* a partir de donde se obtienen los archivos. En el caso de **CLICLUX** opta por usar una cuenta para copiar archivos y no usar el usuario *anonymous*.

3.4 Instalación del resto de las máquinas

Una vez que se tienen creado los 4 diskettes de instalación se debe proceder a la instalación del resto de las máquinas. Lo primero de esta etapa es arrancar una de las candidatas a clonar con el diskette número uno, luego se pedirá el dos y luego el tres. Al insertar el tercer diskette se procede a la detección del driver de la placa de red. Este proceso puede fallar al acceder al dispositivo de forma incorrecta por lo que es conveniente que el usuario seleccione el driver de la placa de red si sabe cuáles. Por último, con el cuarto diskette, se procede a la clonación efectiva de la máquina. El proceso de clonación lleva varios pasos:

1. Configuración de la dirección IP y sus parámetros.
2. Selección de la máquina *master*. a partir de la cual se hará la copia.
3. Particionado del disco rígido.
4. Creación de los file systems sobre el disco rígido.
5. Copiar las imágenes desde el equipo *master*.
6. Configurar la máquina de forma localmente
7. Cargar el boot loader.
8. rebootear.

Los pasos pueden ser ejecutados de forma interactiva o en lotes. Una vez que se clonó una máquina ésta puede ser utilizada como *master* para clonar otras y así acelerar el proceso de clonación.

4 Comparación con una herramienta existente

Para obtener una evaluación de la **CLICLUX** se buscó otra herramienta de características similares y de acceso libre, [OSCAR]. OSCAR es una "suite" de herramientas de acceso libre dedicadas al procesamiento paralelo integradas en un paquete listo para instalar sobre un cluster.
La comparación entre las dos herramientas se detalla a continuación separado por ítem:

Creación de máquina *master*: ambas herramientas requieren la instalación y configuración de una máquina considerada como *master*. Siempre se requiere instalar GNU/Linux. Luego una vez configurada la máquina *master* OSCAR instala los paquetes a ser usados en el procesamiento paralelo de forma automática sobre la misma. Por el contrario, con **CLICLUX** esta tarea se debe realizar de forma manual luego de la instalación de GNU/Linux. El inconveniente que se encuentra con OSCAR es que los paquetes vienen en formato RPM los cuales tienen muchas restricciones de acuerdo a la distribución de GNU/Linux que se está utilizando. Se experimentó el problema al tener instalada en la máquina *master* la versión de GNU/Linux RedHat 7.1 para i386, las versiones de OSCAR 1.2 y 2.2 fallaron. En el caso de **CLICLUX** como los paquetes se instalan de forma manual se pueden instalar en cualquier formato y seleccionar los paquetes que se deseen sin tener problemas.

Método de booteo remoto: OSCAR utiliza el método de booteo mediante las PROM que contienen las placas de red. Si no se posee PROM en la placa de red usa un programa que desde un diskette obtiene el kernel del sistema operativo usando el protocolo `tftp`. En el caso de **CLICLUX** el booteo es mediante un diskette directamente con el kernel. No se utiliza el booteo remoto. De acuerdo a la flexibilidad es mejor OSCAR pero requiere tener configurado el servicio `tftp` en la máquina *master*, cuestión que hace más sencillo a **CLICLUX** .

Cantidad de diskettes necesarios: para el booteo de las máquinas a clonar con OSCAR se requiere un solo diskette, con **CLICLUX** cuatro. En este sentido OSCAR es maás práctico.

Partición Root: La partición en la cual se monta el GNU/Linux, con OSCAR es una partición remota NFS root, lo cual requiere soporte de NFS root server en la máquina *master* y soporte de NFS root client y un file system independiente de dispositivos (`/dev`) en el kernel cargado. Para **CLICLUX** la partición que se levanta como root es una ramdisk local lo cual lo convierte en una opción mucho más sencilla.

Particionado de discos: La forma de particionar los discos en OSCAR es a "disco completo", al menos en la versión que funcionó con Red Hat 7.1, OSCAR-1.1. No hay forma de conservar particiones existentes. Con **CLICLUX** el particionado es aditivo, es decir si existen particiones pero aún hay una porción de disco sin usar se puede utilizar ésta sin "romper" el resto. Esto es conveniente si se quiere mantener booteo dual con otro sistema operativo.

Método de copia remota: La forma que se copian los archivos desde la máquina *master* a los clones con OSCAR es mediante NFS, en cambio con **CLICLUX** se usa `ftp` lo cual genera menos overheaden la red.

Partición de usuarios: la partición de usuarios (`/home`) montada con OSCAR es remota vía NFS, en cambio la de **CLICLUX** es local por default. Al ser remota facilita el acceso a los archivos desde cualquier máquina pero genera más carga sobre la red.

***Masters* disponibles:** con OSCAR las máquinas que se clonan deben ser reconfiguradas para ser usadas como *master*, en cambio con **CLICLUX** cada máquina clonada puede ser *master* para nuevos clones.

Interfaz de usuario: la interfaz de usuario de OSCAR está realizada utilizando tcl-tk y ejecuta perfectamente en X-Window. La interfaz de **CLICLUX** es meramente textual y gran parte de la configuración se debe hacer editando manualmente los archivos. En este sentido es mucho mejor la versión de OSCAR, pero existe el inconveniente que no se sabe qué archivos modifica y en ocasiones es necesario desinstalar y comenzar nuevamente, problema que con **CLICLUX** no existe debido a que sólo se deben modificar los archivos de configuración.

5 Conclusiones y Trabajos a Futuro

La herramienta **CLICLUX** se utilizó para instalar un cluster de 8 PC homogéneas conectadas mediante una red ethernet switcheada de 100MB/s. El resultado fue satisfactorio.
Con OSCAR la instalación tuvo varios inconvenientes y se debió terminar "a mano". Surgieron

problemas con los archivos de dispositivos (`/dev`) y se debió cambiar el kernel para que soporte [DEVFS].
Desde el punto de vista del usuario final OSCAR es más **amigable** que **CLICLUX** , pero es conocido por ejemplos reales que el ser amigable no necesariamente significa más **simple** o **robusto** y cuando algo no funciona es difícil encontrar el problema. En este sentido **CLICLUX** es mucho más simple y robusto logrando resultados equivalentes. Para usar **CLICLUX** se requiere editar una serie de archivos, cuestión que es más tediosa pero ayuda al momento de resolver problemas cuando las cosas no funcionan. Con respecto a la **flexibilidad** existen cuestiones en que uno es más flexible que otro. Con respecto al particionado de disco y a las distribuciones que se soportan **CLICLUX** es más flexible, con respecto al booteo y a la forma de instalar las herramientas, OSCAR es más flexible.
Antes de concluir se puede destacar que OSCAR es una herramienta que tiene más maduración que **CLICLUX** y que este último necesita ser mejorado para que esté disponible para usuarios finales, además de agregarle la documentación. Otro aspecto importante es la cuestión de la GUI que necesita **CLICLUX** para ser más amigable.
Como resultado final obtenemos una herramienta útil muy simple y práctica para el clonado de clusters que necesita ser mejorada.

Referencias

[GNU/Linux] GNU/Linux. URL: http://www.linux.org/.

[BWLUF] Beowulf HOWTO, v1.1.1. Jacek Radajewski and Douglas Eadline. Homepage http://www.beowulf.org. 22 November 1998.

[PVM] PVM: Parallel Virtual Machine. A User's Guide and Tutorial for Networked Parallel Computing. Al Geist, Adam Beguelin, Jack Dongarra, Weicheng Jiang, Robert Manchek and Vaidy Sunderam. The MIT Press, Cambridge, Massachusetts. 1994.

[MPI] MPI: The Complete Reference. Marc Snier, Steve Otto, Steven Huss-Lederman, David Walker and Jack Dongarra. The MIT Press, Cambridge, Massachusetts. 1996.

[OSCAR] OSCAR: Open Source Cluster Application Resources. Michael J. Brim, Timothy G. Mattson, Stephen L. Scott. URL: http://www.csm.ornl.gov/oscar/papers.html.

[PBS] PBS: Portable Batch System. Altair Grid Technologies, LLC, 2003. URL: http://www.openpbs.org/.

[C3] Cluster Command & Control (C3) Tool Suite. Michael Brim, Ray Flanery, Al Geist, Brian Luethke, and Stephen Scott D Computer Science & Mathematics Division Oak Ridge National Laboratory Oak Ridge, TN 37830-6367 USA.

[DEVFS] Linux Devfs (Device File System) FAQ. Richard Gooch. 20-AUG-2002. URL: http://www.atnf.csiro.au/people/rgooch/linux/docs/devfs.html.