Procedimentos Eficazes de Instalação Automatizada em Nodos de um Agregado*

Rodrigo S. Alves, Diego F. Contessa, Clarissa C. Marquezan, Rafael B. Ávila, Tiarajú A. Diverio, Philippe O. A. Navaux

Instituto de Informática - Universidade Federal do Rio Grande do Sul - UFRGS
Av. Bento Gonçalves, 9500 - Porto Alegre RS - Telefone (51)3316 6846 - Fax: (51)3316 1576
{sanger, contessa, clarissa, avila, diverio, navaux}@inf.ufrgs.br

Introdução

Os agregados de computadores são cada vez mais utilizados por empresas e universidades quando o assunto é alto desempenho. Esses agregados estão sendo formados por dezenas, ou até mesmo centenas de nodos, o que acaba dificultando a sua implantação e manutenção. Uma grande desvantagem quando se têm diversos nodos é a instalação e a atualização do sistema operacional e das configurações, que podem fazer com que nem todas as máquinas possuam a mesma imagem. O processo de instalação de software em um agregado pode ser cansativo e lento. Ao alterar a imagem de algum nodo, seja por instalação de software ou atualização de sistema, por exemplo, há a necessidade de fazer a mesma alteração em todos os nodos do agregado. Sem a utilização de ferramentas esse processo pode consumir um tempo muito grande.

O objetivo do trabalho é apresentar um método desenvolvido para instalação automatizada do sistema operacional nos nodos de um agregado Linux, e realizar a comparação com outras ferramentas similares existentes. Com ele, em alguns passos simples, se pode ter um nodo completamente operacional, não sendo necessária a realização de toda a configuração que uma instalação manual exige. Esse método pretende ser uma solução personalizável e adaptável às necessidades de cada usuário.

Cabe ressaltar que o presente estudo está inserido no projeto existente entre a UFRGS e a Dell, do qual um dos frutos é o agregado de alto desempenho instalado no LabTeC - UFRGS/DELL (Laboratório de Tecnologia em Clusters). Esse agregado é composto por 20 nodos biprocessados (*dual* Pentium III 1,13 GHz), com 1 GB de memória RAM e 18 GB de capacidade de disco rígido SCSI.

Ferramentas de Instalação Automatizada Disponíveis

Existem diversas ferramentas para instalação automatizada dos nodos de um agregado. Algumas são específicas para uma determinada distribuição, como o FAI [FAI 02], exclusivo para Debian. Esta seção apresenta alguns desses sistemas, principalmente o SystemImager [SYS 02], que também foi utilizado na instalação do agregado do LabTeC e serviu como base de comparação para o método implementado.

O SystemImager é uma coleção de ferramentas escritas como *shell scripts* e *scripts* Perl. Foi projetado para possuir o mínimo possível de requisitos e ter uma

*Trabalho apoiado pelo CNPq e PIBIC/CNPq e pelo Convênio LabTeC - UFRGS/DELL

interface que esconda ao máximo seu funcionamento interno do administrador do sistema. Procedimentos como a preparação de um cliente cuja imagem será replicada e a obtenção e armazenamento pelo servidor da imagem a ser descarregada no cliente são facilitadas e têm sua complexidade escondida pelos utilitários do SystemImager. Além disso, permite o armazenamento e referência a várias imagens, de modo que cada nodo possa ter uma determinada imagem associada a si, pelo uso de links criados por um utilitário próprio. Disponibiliza ainda duas ferramentas que auxiliam na configuração de um servidor DHCP para prover informações como número IP, endereço de rede e máscara de sub-rede, indispensáveis quando um nodo a ser instalado é inicializado.

O Replicator [REP 00] é um sistema de duplicação de instalação de sistemas Debian Linux. Essa ferramenta procura levar em conta a diferença de hardware entre uma máquina e outra, configurando o software de acordo com o hardware. Outra ferramenta para a instalação de sistemas Debian é o FAI (*Fully Automatic Installation*), formada por *shell scripts* e *scripts* Perl. O BpBatch [BPB 02] é um inicializador remoto que utiliza o protocolo PXE, podendo clonar o esquema de partição de uma máquina, e instalar uma imagem nos clientes. O Partition Image [PAR 02] é uma ferramenta cuja maior diferença é o fato de suportar vários sistemas de arquivos.

Existem outros métodos onde é possível especificar quais pacotes deseja-se instalar automaticamente. Porém o método utilizado pelo SystemImager e pelo sistema implementado se mostra mais genérico, pois leva em consideração softwares que não estão na forma de pacotes, além de tratar de arquivos de configuração e mudanças no *kernel*. Isto se explica pela filosofia de funcionamento destas ferramentas, que se baseia na cópia e armazenamento da partição raiz inteira de um nodo devidamente instalado.

Apesar de existirem diversas ferramentas disponíveis, há a necessidade de uma maior adaptabilidade e personalização de uma solução de acordo com as necessidades de cada agregado, sem esquecer a simplicidade. Basta para isso que cada administrador implemente a sua solução de acordo com o conhecimento que tem do seu sistema.

Método de Instalação Automatizada Implementado

O procedimento implementado se baseia no que está descrito em [STE 01]. Nesse método, o nodo que está sendo instalado realiza a sua própria instalação, ou seja, o primeiro problema é disponibilizar um sistema temporário no nodo para que este possa executar um script de instalação, o qual efetua a instalação propriamente dita. O sistema operacional temporário a ser carregado no nodo é formado, informalmente, por duas partes principais: o *kernel* e um sistema de arquivos (árvore de diretórios). As diferentes possibilidades de realizar a instalação no método implementado consistem justamente das opções de localização de armazenamento dessas duas partes.

Basicamente a escolha está entre usar dispositivos como disquetes ou CD-ROM, ou utilizar a rede para acessar um servidor, isto tanto para a carga do *kernel* como do sistema provisório. Esta decisão deve ser tomada levando em conta os recursos disponíveis no agregado e o custo/benefício para sua implantação. O gerente do agregado deve procurar alternativas que não penalizem a manutenção dos nodos.

A utilização de dispositivos de armazenamento pode apresentar problemas, como o sistema de inicialização da máquina não poder ser carregado em apenas um disquete. Outro problema é a máquina a ser instalada não possuir um drive de CD-ROM. Também existem problemas quando se opta pela utilização da rede para

instalação de uma máquina. Um dos mais comuns é a placa de rede não suportar inicialização pela rede, inviabilizando esse procedimento.

Após alguns testes verificou-se que para a melhor utilização dos recursos disponíveis no *cluster* LabTeC - UFRGS/DELL, o qual foi utilizado para estas experiências, assim como para uma maior facilidade de manutenção, a opção mais adequada seria ter duas formas de carga do *kernel*, por disquete e pela rede, e carregar a árvore de diretórios por NFS [STH 91], isto é, montá-la como partição raiz a partir de uma imagem armazenada no servidor e trabalhar com esta árvore de diretórios apenas para disponibilizar um sistema onde o script de instalação possa ser executado.

Depois de tomada essa decisão, passa-se à implementação da solução propriamente dita. O primeiro passo de uma instalação automatizada onde o *kernel* e o sistema de arquivos temporário são obtidos pela rede é a aquisição da imagem de um nodo previamente instalado e configurado, a ser armazenada no servidor. Isso ocorre compactando a imagem de um cliente completamente instalado, e armazenando no servidor essa imagem que será replicada para todos os outros nodos do agregado.

Em seguida é preciso compilar um *kernel* que tenha suporte a exportar uma partição raiz por NFS. Esse recurso é chamado de NFS_ROOT. Esse *kernel* será usado para montar o sistema de arquivos temporário anteriormente citado, e será disponibilizado pelo servidor juntamente com as opções de inicialização através do protocolo TFTP. Cabe ressaltar que o servidor precisa ainda ser configurado para prover as informações de rede para um nodo a ser instalado através de, por exemplo, DHCP.

Neste momento, o nodo a ser instalado deve ser inicializado com o uso da opção de inicialização pela rede. Assim, sua placa de rede se comunicará com o servidor através do protocolo PXE, visando obter o *kernel* e o sistema de arquivos temporário. Feito este passo, o *kernel* é carregado e os *scripts* de inicialização são executados.

Uma das diferenças entre esse método e o que é descrito em [STE 01] é o fato do primeiro não substituir o *script* de inicialização padrão de sistemas UNIX (*init*). O sistema temporário que é carregado possui entre seus *scripts* de inicialização o *script* de instalação do nodo, que é executado após todos os outros *scripts* do sistema operacional temporário (montado por NFS_ROOT). Nesse *script* estão os comandos de particionamento e formatação do disco rígido, diretivas do local onde o sistema deve ser instalado e comandos para a descompactação e instalação efetiva do sistema. Ao final da execução desse *script*, o nodo é reiniciado e está praticamente pronto para ser utilizado, faltando apenas instalar um *boot loader*, permitindo assim sua inicialização.

Avaliação da Solução Apresentada

O método implementado é mais simples que o descrito em [STE 01]. Como o *init* não é substituído, são carregados na inicialização do sistema temporário alguns programas que não seriam necessários, já que sua finalidade é apenas executar o *script* de instalação. Por isso, foram feitas otimizações no sistema temporário, como a remoção de pacotes como *portmapper*, *nis*, *sysklogd*, *atd*, *cron*, entre outros. Essa alteração trouxe benefícios no que diz respeito à simplicidade de implementação, visto que a otimização efetuada é mais trivial que alterar o sistema de inicialização do Linux.

Um ponto positivo que tanto o SystemImager quanto o método implementado oferecem é o fato de permitirem que alterações sejam feitas diretamente na imagem que está guardada, simplesmente trabalhando com os arquivos no diretório onde eles estão

armazenados no servidor. Uma opção mais elegante e que traz facilidades como utilizar aplicativos como *apt-get* e *rpm* para gerenciar pacotes é o uso de *chroot* para trabalhar com a imagem armazenada no servidor como se fosse um sistema em execução.

Ferramentas que armazenam uma imagem no servidor para replicação nos nodos apresentam vantagens ao usar o utilitário *rsync* para a atualização dos clientes. O *rsync* torna a partição raiz do nodo idêntica à imagem armazenada no servidor. Após qualquer alteração nessa imagem, chama-se o *rsync* para atualizar todo o agregado sem tirá-lo do ar. O *rsync* transfere apenas as diferenças entre os arquivos, otimizando a operação.

Conclusões

O trabalho desenvolvido apresenta-se como uma solução viável para a manutenção de um agregado de alto desempenho. O principal ganho que se obtém usando um procedimento de instalação automatizada é no tempo necessário à instalação de um agregado, e também no momento da atualização dos nodos.

Cabe ressaltar que o sistema implementado não é o único disponível, mas é bastante flexível. Existem diversas outras ferramentas, amplamente difundidas entre os administradores. Porém, elas apresentam problemas de adaptabilidade. Os usuários mais avançados tendem a preferir ferramentas mais adaptáveis. No SystemImager, para obter uma personalização é preciso alterar *scripts* complexos e grandes, frutos da intenção da ferramenta de ser genérica para o uso com qualquer máquina. Com o sistema desenvolvido, é possível personalizar o processo de instalação de forma simples e rápida. Além disso, basta efetuar alterações em um único *script* e o procedimento é alterado, de acordo com as necessidades do administrador de sistema.

Outro ponto importante é o casamento do ganho no tempo de instalação com a diminuição da chance de ocorrência de falhas, principalmente humanas, que poderiam fazer com que se tivessem diferenças entre as imagens dos nodos. Com o procedimento desenvolvido, associado ao uso de *rsync*, garante-se que as imagens a serem instaladas em todos os nodos são iguais e que possuem as mesmas configurações e otimizações.

Referências

[BPB 02] BPBATCH. **BpBatch.** Disponível por WWW em http://www.bpbatch.org/ (14/11/2002).

[FAI 02] FAI. **FAI (Fully Automatic Installation).** Disponível por WWW em http://www.informatik.uni-koeln.de/fai/ (14/11/2002).

[PAR 02] PARTITION IMAGE. **Partition Image for Linux.** Disponível por WWW em http://www.partimage.org/ (14/11/2002).

[REP 00] REPLICATOR. **Replicator.** Disponível por WWW em http://replicator.sourceforge.net/ (14/11/2002).

[STE 01] STERLING, T. **Beowulf Cluster Computing with Linux**. Cambridge: MIT, 2001.

[STH 91] STERN, H. **Managing NFS and NIS.** O'Reilly & Associates Inc., 1991.

[SYS 02] SYSTEMIMAGER. **SystemImager.** Disponível por WWW em http://systemimager.org (14/11/2002).