Um Sistema de Gerência de Réplicas para Grades baseado em redes Par-a-par

Diego S. Gomes, Marko Petek, Cláudio F. R. Geyer,
Tiarajú A. Diverio

Universidade Federal do Rio Grande do Sul
Porto Alegre, RS, Brasil
{dsgomes,petek,geyer,diverio}@inf.ufrgs.br

Introdução

Sistemas de Gerência de Dados para grades computacionais freqüentemente fornecem serviços para lidar com réplicas (cópias) de arquivos. A replicação de dados nesse ambiente visa aumentar a disponibilidade dos mesmos de forma a melhorar o desempenho e a garantir aspectos de tolerância a falhas. As operações típicas gerenciadas por um sistema de réplicas (*Replica Management System* – RMS) são: localização, transferência e armazenamento de arquivos. Em grades aonde as aplicações utilizam um grande volume de dados (grades de dados), o orquiestramento correto dessas operações é fundamental para se obter um bom desempenho.

O *toolkit* do Globus [FOS 97] fornece um componente para localização de réplicas na qual os índices usados para referenciar as réplicas de arquivos são distribuídos de maneira hierárquica. Para transferir arquivos, utiliza-se a ferramenta GridFTP em conjunto com um serviço de transferência confiável. Embora existam ainda serviços de mais alto nível, o Globus não gerencia o armazenamento dos dados. O usuário é quem deve, em princípio, especificar os locais (nodos) aonde colocá-los. Devido à dinamicidade da grade, é muito difícil um usuário saber, de antemão, aonde melhor colocar esses dados de forma a mantê-los próximos às aplicações.

Este trabalho apresenta um RMS para grades que considera o armazenamento dos dados. Utiliza-se a infra-estrutura de redes par-a-par (P2P) para fornecer adaptabilidade às variações da grade e para escalonar o sistema. Embora existam soluções específicas para localização de dados baseado em redes P2P (ex.: P-RLS), bem como para armazenamento (ex.: PAST), o uso de ambas em conjunto em um ambiente de grade é uma atividade ainda pouco explorada [SHO 03].

Descrição

Para que um arquivo seja gerenciado pelo sistema, o usuário precisa, antes de mais nada, registrá-lo. Um solicitação de registro é realizada em diferentes fases:

- **Fase 1 – transferência**: uma vez que o usuário pode requisitar o registro de um arquivo remoto, a primeira fase constitui-se de transferir o arquivo para o nodo aonde se está efetuando o registro;

- **Fase 2 – divisão**: a segunda etapa preocupa-se em dividir o arquivo em pedaços menores (no momento, os pedaços assumem um tamanho fixo de 10MB). Calcula-se então um *hashing* de cada parte do arquivo de forma a distribuí-los entre os diversos nodos da rede P2P (nodos da grade). O objetivo dessa divisão é balancear a carga de arquivos grandes (arquivos típicos encontrados em grades de dados). Por fim, os diferentes pedaços são transferidos para as máquinas que por eles são responsáveis na rede P2P.

- **Fase 3 – registro**: a última fase diz respeito ao registro, junto à rede P2P, de um bloco que descreve o arquivo. Este bloco contém informações como data de criação, identificação do usuário da grade que o armazenou (para posterior contabilidade do espaço ocupado) e uma lista dos *hashes* dos pedaços para que estes possam ser localizados independentemente. Outra informação importante nele contida é o fator de replicação k, que indica quantas cópias de cada parte do arquivo o sistema deve manter. O sistema criará novos pedaços do arquivo à medida que os nodos ingressam/saem da grade, de forma a manter a invariante de k cópias.

De posse de um nome identificador especificado na etapa de registro do arquivo, o usuário pode requisitar a localização do mesmo (ou de seus pedaços em separado), bem como pode solicitar uma operação de consulta. Consultar um arquivo significa transferir o arquivo para a máquina local. Nesta operação, o sistema faz a transferência em paralelo das diversas partes que compõem o arquivo. Como estas partes estão colocadas em diferentes nodos, diferentes fontes são contactadas simultaneamente (característica conhecida como *striping* de arquivos).

Por fim, uma operação de remoção de arquivo é fornecida. É através dela que o usuário indica que o sistema não deve mais gerenciar o arquivo, implicando na remoção das partes do mesmo dos diferentes nodos que compõem a grade. O protocolo para remoção utilizado pelo sistema garante que nodos temporariamente desconectados (quando solicitada a operação de remoção) não vão continuar a disponibilizar o arquivo removido.

Conclusões

Embora o sistema tenha sido projetado para o bom desempenho no armazenamento de arquivos grandes (maiores que 100MB), o uso do mesmo para arquivos menores ainda apresenta vantagens, tais como disponibilidade, adaptabilidade e segurança.

No momento, o sistema encontra-se em fase de prototipação, porém resultados preliminares de tempo de registro de réplicas já foram obtidos.

Referências

[FOS 97] FOSTER, I.; KESSELMAN, C. **Globus: A Metacomputing Infrastructure Toolkit**. 1997.

[SHO 03] SHOSHANI, A. **Data and Storage Management, Data Transport**. PPDG Review, Fermilab. 2003.