Técnicas para Armazenamento Distribuído com Alto Desempenho

Rodrigo Kassick, Everton Hermann, Philippe O. A. Navaux

Instituto de Informática UFRGS
Av. Bento Gonçalves 9000 Bloco 4
{rvkassick,ehermann,navaux} @ inf.ufrgs.br

Introdução

A utilização de *cluster* de computadores se tornou muito popular na área de computação paralela. Muitas aplicações científicas executadas neste tipo de plataforma efetuam grande quantidade de operações de entrada e saída em armazenamento secundário, local ou remoto. O desempenho de tais operações é, então, crucial e pode ser um fator determinante do tempo de execução de uma aplicação.

O aumento da diferença de velocidade entre processador e dispositivos de armazenamento agrava ainda mais esta situação. Uma alternativa para amenizar este problema com baixo custo é a utilização dos recursos de armazenamento dos nós de um cluster através de um sistema de arquivos paralelo. Isto permite agregar as capacidades de transferência de rede e disco de vários nós, oferecendo uma significativa melhoria de desempenho.

Sistemas de Arquivos Paralelos

A necessidade por desempenho em sistemas de armazenamento compartilhado fez surgir diversos trabalhos que objetivam sistemas escaláveis e que façam bom uso de padrões existentes de hardware e software. Tradicionalmente, estes sistemas dividem as tarefas de armazenamento entre um conjunto de nós, dedicando alguns ao gerenciamento de metadados – nome, tamanho ... – e outros ao de dados. Assim, é possível agregar a vazão de transferência de vários nós para permitir leituras e escritas mais rápidas. Dentre as ferramentas que implementam este modelo, pode-se destacar o PVFS, Lustre e o dNFSp2 [1].

O PVFS divide o sistema em servidores de dados (IOD's) metadados. Os dados são, por padrão, gravados por fatias nos IOD's. Os clientes que acessam um arquivo podem calcular a localização do bloco desejado a partir da posição a ser lida. Este esquema, além de melhorar o desempenho, provê uma área de armazenamento equivalente à soma do espaço de armazenamento dos IOD's.

O Lustre também diferencia as funções de armazenamento e gerência de metadados, mas os dados são armazenados de forma replicada em *Object Storage Servers*, os quais podem possuir diferentes dispositivos físicos de armazenamento, de discos comuns a sistemas proprietários. O servidor de metadados é centralizado, possuindo apenas réplicas para caso de falhas.

O dNFSp é um sistema de arquivos desenvolvido em parceria entre o laboratório ID da França e o GPPD da UFRGS. Desenvolvido a partir do NFSp, por sua vez inspirado

no PVFS1, ele também diferencia operações de dados e metadados. No dNFSp, no entanto, os clientes utilizam o protocolo NFS para comunicar com o servidor, dispensando alterações em suas configurações.

Outras Técnicas para Melhoria de Desempenho em IO

A divisão dos dados entre os IOD's tradicionalmente leva em conta apenas a posição do bloco dentro do arquivo, ignorando a maneira como os dados são acessados pelas aplicações. Através de uma distribuição de dados que leve em conta seu padrão de acesso, é possível diminuir a concorrência pelos recursos de armazenamento em uma mesma aplicação.

Para tanto, se fazem necessárias bibliotecas que permitam à aplicação definir o padrão de gravação de seus dados. A MPI-IO faz isto através de *views* e tipos de dados definidos pelo usuário. Tais informações são utilizadas pelo sistema de arquivos Cluster-File [2] para definir estratégias de armazenamento de forma diminuir a concorrência dos clientes nos mesmo servidores de dados e, conseqüentemente, aumentar o desempenho.

Outro fator importante é a concorrência entre aplicações distintas. Neste caso, a cada execução, o sistema deve avaliar a carga do sistema de arquivos para derivar um padrão de acesso com bom desempenho e *fairness* para todas as aplicações. Um exemplo de sistema com este objetivo é o aIOLi [3], desenvolvido também no laboratório ID-IMAG da França.

A distância em segmentos de rede entre os nós de dados e de processamento de uma aplicação também se tornam de grande importância em clusters de grande porte. Se a distância entre estes nós for grande ou se algum destes segmentos for muito concorrido, as aplicações podem ter sua performance degradada. Melhorias feitas no dNFSp permitem a criação dinâmica IOD's. Cluster's de grande porte podem fazer uso deste dinamismo para diminuir o custo de acesso aos dados nestes casos.

Conclusões

Este resumo apresentou um apanhado de características que devem ser levadas em conta para munir soluções de armazenamento em cluster com características de alto desempenho, visando permitir a aplicações paralelas com grandes necessidades de armazenamento executar sem grande perda de desempenho. Resta estudar como integrar estas soluções de forma a possuir um sistema de armazenamento escalável e com bom desempenho.

Referências

[1] R. B. Ávila. *Uma Proposta de Distribuição do Servidor de Arquivos em Clusters*. Tese de doutorado, Universidade Federal do Rio Grande do Sul, Brasil, 2005.

[2] F. Isaila, D. Singh, J. Carretero, F. Garcia, G. Szeder, and T. Moschny. Integrating logical and physical file models in the mpi-io implementation for "clusterfile". In *CCGRID '06*, 2006.

[3] A. Lebre, G. Huard, Y. Denneulin, and P. Sowa. I/o scheduling service for multi-application clusters, 2006. Accepted for publication in Cluster 2006.