NPFS: Um Sistema de Arquivos Paralelos em Rede

Hélio Crestana Guardia [1] Liria Matsumoto Sato [2]

1 Introdução

Utilizando um conjunto de discos locais [PAT 88] ou distribuídos entre os computadores de um sistema computacional paralelo [COR95, HAR93, LON94, MOY94], um **sistema de arquivos paralelos** visa combinar as taxas de transferência no acesso em paralelo a diversos discos para o aumento do *throughput* nas transmissões entre disco e memória, e a presença de diversos níveis de *buffers* para a diminuição da latência das requisições.

Este trabalho apresenta um sistema de arquivos paralelos para um ambiente computacional distribuído. Além de proporcionar mecanismos para o aumento da capacidade de armazenamento e da velocidade das operações de entrada e saída de dados, seu desenvolvimento visa a criação de uma ferramenta de testes para a avaliação de diferentes políticas que podem ser empregadas no armazenamento de dados em sistemas distribuídos.

2 NPFS: Um Sistema de Arquivos Paralelos em Rede

NPFS (*Network Parallel File System*) é um sistema de arquivos paralelos distribuídos destinado a um conjunto de estações de trabalho interligadas em rede. A distribuição dos dados é realizada com a técnica de *striping* [CRO 89], podendo-se selecionar o número de segmentos e o tamanho das unidades de distribuição. Além do aumento da capacidade de armazenamento e da velocidade das operações de E/S para aplicações seqüenciais, o sistema permite o acesso otimizado aos dados de aplicações paralelas que manipulam porções distintas de um arquivo compartilhado. Enquanto um conjunto básico de primitivas oferece uma semântica de acesso compatível com o conceito de arquivos no sistema operacional UNIX, a estrutura paralela dos arquivos pode ser mais bem aproveitada utilizando uma implementação das primitivas propostas para a interface de baixo nível de sistemas de arquivos paralelos (**SIO *Low Level* API**) [COR 96].

A arquitetura do sistema **NPFS** é baseada no modelo cliente / servidor. De maneira geral, a filosofia do sistema consiste em simplificar ao máximo a operação dos servidores, realizando todas as operações possíveis nos processos de aplicação que utilizam os arquivos paralelos. Assim, há uma dissociação entre as políticas de distribuição utilizadas e a operação dos servidores, que não mantêm informações sobre a distribuição dos dados. A localização dos dados necessários nas operações é realizada pelos próprios clientes, sem o auxílio de tabelas de armazenamento. O agrupamento de requisições contíguas provê uma otimização dos acessos aos discos nos servidores. Além da organização modular, a estruturação em processos tem o objetivo de isolar componentes e serviços e de reduzir as atribuições de elementos centralizados. As comunicações entre processos são realizadas utilizando os protocolos UDP/IP, sendo que a qualidade das comunicações é garantida pelo protocolo de

[1] DC/UFSCar – helio@dc.ufscar.br
[2] PCS/USP – liria@pcs.usp.br

comunicação do sistema **NPFS**. Mecanismos para tolerância a falhas de *hardware* ainda não estão implementados. Implementados como um conjunto de primitivas de software, os serviços do sistema **NPFS** que podem ser utilizados pelas aplicações incluem suporte para abrir, fechar, ler e escrever arquivos, além de ativar e desativar servidores.

3 Resultados

O paralelismo no acesso aos diversos discos faz com que o desempenho do sistema **NPFS** seja superior ao obtido com NFS. Como seria esperado, contudo, a comunicação através da rede impõe restrições ao desempenho do sistema desenvolvido. Para requisições que não envolvam fragmentações dos pacotes transmitidos, o paralelismo das operações supera as sobrecargas de comunicação, fazendo com que o desempenho do sistema **NPFS** nas operações de escrita de dados seja superior ao acesso local. Nas operações de leitura, contudo, o desempenho do sistema **NPFS** para acessos seqüenciais, foi inferior ao acesso local devido à leitura antecipada dos dados. Para grandes requisições, o elevado número de retransmissões ocorridas e o baixo desempenho das operações do sistema **NPFS** indicam que um protocolo de comunicação síncrono seria mais eficiente nesses casos. Nesses casos, o desempenho com o sistema **NPFS** foi inferior ao obtido com o sistema PIOUS [MOY 94].

4 Conclusões

Os resultados do sistema NPFS mostram a viabilidade de seus objetivos. Em um sistema de arquivos paralelos em rede, contudo, observa-se que o potencial para aumento de desempenho não está relacionado somente com o número de discos, mas depende também da rede de interconexão e dos tamanhos dos fragmentos utilizados. Além disso, o desempenho de uma aplicação paralela distribuída que apresenta um grande número de operações de E/S não está limitado apenas pelo desempenho dos sistemas de arquivos e de comunicação, mas também pelas interferências entre eles. Assim, mais do que simplesmente aumentar a taxa de transferência no acesso aos dados armazenados, a integração do sistema de arquivos com o mecanismo de suporte para os processos que os manipulam poderá reduzir o tráfego na rede e melhorar o uso de redes de estações de trabalho na execução de aplicações paralelas.

Referências Bibliográficas

[**COR 95**] CORBETT, P.; FEITELSON, D.; FINEBERG, S.; HSU, Y.; NITZBERG, B.; PROST, J.; SNIR, M.; TRAVERSAT, B. and WONG, P. “Overview of the MPI-IO parallel I/O interface”. In *IPPS’95 Workshop on Input/Output in Parallel and Distributed Systems*, p.1-15, April 1995.

[**COR 96**] CORBETT, P.F. et all. “Proposal for a Common Parallel File System Programming Interface”. 1996.

[**CRO 89**] CROCKETT, T.W. “File Concepts for Parallel I/O”. In *Proceedings of Supercomputing’89*, 1989.

[**HAR 93**] HARTMAN, J.H. and OUSTERHOUT, J.K. “The Zebra striped network file system”. In *Proceedings of the Fourteenth ACM Symposium on Operating Systems Principles*, p.29-43, 1993.

[**LON 94**] LONG, D.D.E.; CABRERA, L. “Swift/RAID: A Distributed RAID System”. Technical Report UCSC- CRL-94-06, 1994.

[**MOY 94**] MOYER, S.A and SUNDERAM, V.S. “*A Parallel I/O System for High-Performance Distributed Computing*”. Computer Science Technical Report CSTR-940101, 1994.

[**PAT 88**] PATTERSON, D.; KATZ, R. “A Case for Redundant Arrays of Inexpensive Disks (RAID)”. *Proceedings of the ACM SIGMod*. ACM p.109-16, New York, 1988.