# O Impacto do Suporte à Heterogeneidade no Mecanismo de Balanceamento de Servidores de Dados no dNFSp

Francieli Zanon Boito
fzboito@inf.ufrgs.br

Rodrigo Virote Kassick
rvkassick@inf.ufrgs.br

Philippe O. A. Navaux
navaux@inf.ufrgs.br

## Abstract

*dNFSp is a file system for cluster architectures where data is distributed among several data servers, called IODs. These IODs are organized in groups, called virtual data servers, or VIODs. In each VIOD, there is replication of all the data between the IODs. An IOD can be tied to more than one VIOD, and then it will have to answer more solicitations than the others, affecting the performance. Moreover, these shared servers will store data of more than one VIOD, and it can affect the functioning of the system when it will not have more space for data.*

*This paper presents a study on the changes that can be applied to the mechanism responsable of the balancing of IODs in VIODs to allow the suport to the heterogeneity of storage space on the data servers. This suport will allow the file system to completely use the capacity of storage of each data server. Proposals of modification in the mechanism will be presented for to deal with the situations when the IODs are full.*

## 1. Introdução

O dNFSp é um sistema de arquivos distribuído para *clusters*, onde o papel do servidor é realizado pelos metasservidores e pelos servidores de dados, ou *IODs* (de *Input/Output Daemons*). Análises sobre o seu funcionamento e desempenho podem ser encontradas em [1] e [3]. Existem ainda servidores de dados virtuais, ou VIODs, que são estruturas que, abstratamente, correspondem a um IOD, mas são compostas por 1 ou mais deles. Todos os dados de um VIOD estão replicados em todos os seus IODs, e o número mínimo de IODs por VIOD pode ser determinado pelo usuário. Quando não existem IODs suficientes para preencher esse número, alguns podem aparecer em mais do que um VIOD, e então dizemos que esse é um IOD compartilhado. A distribuição e a manutenção desta estrutura são responsabilidade do gerenciador de servidores de dados. [2]

Uma questão importante em sistemas de arquivos desse tipo é a heterogeneidade. Nem sempre é possível assumir que todos os nós envolvidos terão as mesmas capacidades de armazenamento, processamento, etc. É necessário que o sistema esteja preparado para tratar e explorar ao máximo capacidades diferentes de armazenamento entre seus servidores de dados, por exemplo. Sem este tratamento, a capacidade total de armazenamento acaba limitada pela capacidade do *menor* IOD.

## 2. O Mecanismo de Balanceamento

Quando um IOD é compartilhado, ele passa a responder requisições de mais de um vIOD, ou seja, as requisições feitas pelos clientes vão passar por ele mais vezes do que passariam se ele fosse de um único VIOD. Conforme aumentamos o número de VIODs aos quais ele está vinculado, isso pode transformá-lo em um gargalo. Além disso, ele conterá todos os dados de cada um dos VIODs nos quais ele aparece, diminuindo a parcela de sua capacidade de armazenamento disponível para cada um deles. No dNFSp atual, onde não há suporte à heterogeneidade implementado, o espaço de armazenamento acaba sendo, então, limitado principalmente por esses IODs compartilhados. Uma terceira desvantagem dos IODs compartilhados pode ainda aparecer em uma falha. Quando um dos servidores deixar o sistema por algum motivo, não será apenas um VIOD afetado, mas vários.

O balanceamento de IODs por VIOD é, portanto, uma tarefa importante executada pelo gerenciador dos servidores de dados. Em primeiro lugar, é necessário obedecer, sempre que possível, o número mínimo de IODs por VIODs determinado pelo usuário. Assim, quando um servidor de dados é inserido no sistema, o mecanismo deve procurar saber se existe algum VIOD que não obedece este critério. O grau de replicação nos servidores virtuais é uma configuração do usuário e, portanto, deve ser obedecida, mesmo que isso signifique a criação de IODs compartilhados.

Quando o sistema atingir um estado em que o número mínimo foi alcançado por todos os VIODs existentes, deve haver outro critério para o balanceamento. Sempre que um servidor de dados for inserido no sistema, este passa a tentar substituir IODs compartilhados, garantindo que estes serão

pouco numerosos e, preferencialmente, pouco compartilhados. Isso significa que, mesmo que aceite-se a existência de IODs compartilhados, deseja-se que, quando isto acontecer, que não aconteça muitas vezes para o mesmo IOD, pois quanto mais VIODs vinculados, piores o gargalo e o problema da capacidade de armazenamento.

## 3. O Tratamento da Heterogeneidade

Considerando que um IOD pode possuir capacidade de armazenamento diferente dos demais, é necessário saber o que fazer quando um deles se encontra cheio. Não podemos suspender as escritas ao VIOD inteiro, pois assim estaríamos desperdiçando a capacidade dos outros IODs integrantes. Quando isso ocorrer, o servidor cheio será passado para um estado de *full*. Não podendo mais escrever dados nesse IOD por falta de espaço, o VIOD ao qual ele está vinculado precisará de um novo IOD, que será acrescentado a ele em um estado de *append*, para manter o grau de replicação especificado pelo usuário do sistema e garantir o funcionamento correto do sistema.

Suspendendo as escritas ao IOD *full*, com o tempo ele pode acabar contendo informações desatualizadas. É importante que ele receba todas as requisições de escrita, e só as efetue quando se referirem a blocos já armazenados por ele. Quando não for o caso, ele deve repassar as requisições de escrita ao IOD *append*, adicionado ao IOD para suprir a sua falta. Assim, o IOD *append* não conterá todos os dados referentes ao VIOD, como os IODs comuns, mas apenas os que o *full* não puder gravar. Considerando que o conjunto dos dados referentes ao VIOD foi grande o suficiente para esgotar a capacidade de armazenamento do *full*, ele provavelmente esgotaria a do servidor *append* também, principalmente considerando que este normalmente será um IOD compartilhado com outro VIOD. A escolha do IOD a se tornar o *append* pode ser feita seguindo os critérios normais de escolha de compartilhados.

Para fins de contabilização de número de IODs por VIOD, IODs *append* e *full* serão considerados como um só, criando uma abstração de um único IOD, com capacidade maior de armazenamento. Em hipótese alguma um IOD *full* deve ser considerado candidato a se tornar um IOD compartilhado, uma vez que ele não possui espaço em disco disponível para armazenar informações de um novo VIOD. No entanto, um IOD *append* pode ser preferido para isso.

Um IOD *append* é um bom candidato a se tornar IOD compartilhado porque não possui todos os dados do VIOD ao qual ele está relacionado e, portanto, terá mais espaço disponível. Além disso, por não ter todos os dados, ele não precisa responder a todas as requisições de seu VIOD, como os IODs comuns. Isso diminuirá o gargalo formado e o problema do armazenamento.

## 4. Conclusões

Servidores de dados que estão vinculados a mais do que um servidor de dados virtual não são bons para o desempenho do sistema, pois precisam responder a mais requisições do que os outros, e, portanto, podem formar um gargalo. Além disso, por terem dados a guardar de cada VIOD, eles acabam limitando a capacidade de armazenamento para cada um deles.

O tratamento da heterogeneidade da capacidade de armazenamento dos nós que trabalham como IODs no sistema é importante para aproveitar ao máximo a capacidade de cada um. Para possibilitar esse tratamento, serão criados novos estados, chamados *full* e *append*. Eles implicarão mudanças no mecanismo de balanceamento. Além das mudanças necessárias, podem ser implementadas outras alterações que tirem proveito da nova condição do sistema para obter um balanceamento ainda melhor.

O suporte à heterogeneidade, da forma planejada e descrita neste trabalho, vai implicar certa perda de desempenho. Isso ocorrerá porque um bloco nem sempre será obtido de um único IOD, sendo preciso talvez procurar em IODs de *append*. No entanto, o objetivo dessa implementação é, por enquanto, aumentar a disponibilidade e o aproveitamento do espaço de armazenamento nos servidores. Após alcançado esse objetivo, testes serão executados com o objetivo de medir o quanto o desempenho será afetado.

## Referências


[1] R. B. Ávila. *Uma Proposta de Distribuição do Servidor de Arquivos em Clusters*. PhD thesis, PPGC-UFRGS/ID-IMAG, Porto Alegre/Grenoble, 2005.

[2] D. Conrad, R. Kassick, E. Hermann, and P. Navaux. Gerenciamento distribuido de servidores de dados. In *Anais da 7a. Escola Regional de Alto Desempenho, ERAD*, Porto Alegre, RS, 2007.

[3] R. Kassick, C. Machado, E. Hermann, R. Ávila, P. Navaux, and Y. Denneulin. Evaluating the performance of the dNFSP file system. In *Proc. of the 5th IEEE International Symposium on Cluster Computing and the Grid, CCGrid*, Cardiff, UK, 2005. Los Alamitos, IEEE Computer Society Press. CD-ROM Proceedings, ISBN 0-7803-9075-X.

[4] E. Luque, D. I. Rexachs, and J. Souza. Análise da distribuição de carga em um cluster heterogêneo. In *Proc. of Congreso Argentino de Ciencias de la Computación*, 2000.

[5] J. M. Perez, F. Garcia, J. Carretero, A. Calderon, and L. M. Sanchez. Data allocation and load balancing for heterogeneous cluster storage systems. In *Proc. of the 3th IEEE International Symposium on Cluster Computing and the Grid, CCGrid*, 2003.