Avaliação de Sistemas de Arquivos Distribuídos num Ambiente de Pequena Escala

Amadeu Andrade Barbosa Jr., Fabíola Greve, Luciano Porto Barreto

[1]Universidade Federal da Bahia
Instituto de Matemática
Departamento de Ciência da Computação
Av. Adhemar de Barros, S/N, Campus de Ondina
CEP: 40170-110 - Salvador/BA

{amadeu,fabiola,lportoba}@dcc.ufba.br

Resumo. *Este trabalho apresenta um estudo experimental sobre a avaliação de sistemas de arquivos distribuídos num ambiente de pequena escala. A avaliação se portou sobre aspectos quantitativos (e.g., métricas de desempenho) e aspectos qualitativos (e.g., facilidade de administração e recuperação de falhas). Foram avaliados os sistemas de arquivos Lustre, PVFS2, OCFS2 com iSCSI e GlusterFS. Na avaliação efetuada, o Lustre apresentou melhor desempenho em operações de leitura enquanto o PVFS2 foi melhor em operações de escrita.*
Palavras-chave: *Sistemas de Arquivos Distribuídos, Armazenamento de Dados, Clusters*

Abstract. *This paper presents the evaluation of distributed filesystems in a small network environment. We present the results from a quantitative (e.g., performance metrics) and qualitative (e.g., easy of management, failure recovery) perspective. We analyzed four distributed filesystems: Lustre, PVFS2, OCFS2 with iSCSI and GlusterFS. Overall, Lustre presented best performance for read operations and PVFS2 performed better for write operations.*
Keywords: *Distributed Filesystems, Data Storage, Clusters*

1. Introdução

Com o avanço das técnicas de computação as aplicações se tornaram mais complexas, solicitando maior capacidade de processamento e armazenamento de informação. Aplicações nos campos da genética e meteorologia, por exemplo, requerem técnicas sofisticadas e eficientes no que concerne à organização e distribuição do processamento e da informação. Observamos, atualmente, considerável popularização de aglomerados de máquinas (*clusters*) e sua interligação em escala mundial através da internet, normalmente realizada por soluções de computação em grade (*grid computing*) [dos Santos and Cerqueira 2006]. Tais aplicações e ambientes computacionais requerem mecanismos adequados e eficientes quanto à recuperação e armazenamento de dados.

Diante deste cenário, observamos que as tecnologias de distribuição de processamento e dados, não somente precisam ser frequentemente aperfeiçoadas, mas também devem passar por um processo constante de validação e análise de desempenho. Em particular, é interessante o estabelecimento de um quadro comparativo entre as soluções de

sistemas de arquivos distribuídos, já que o acesso eficiente aos dados depende diretamente de uma boa estratégia de acesso aos discos e à rede.

O NFS (*Network File System*) é provavelmente o sistema de arquivos distribuído mais utilizado. Contudo, a abordagem do NFS para distribuição das soluções de armazenamento é precária, uma vez que só é possível distribuir o acesso dos clientes. Assim, o seu projeto não atende a certas aplicações de maior escala, pois a centralização do armazenamento da informação prejudica o desempenho geral. Dentre outros problemas, o NFS ainda apresenta deficiências históricas quanto à manutenção da consistência do cache dos clientes, à sobrecarga imposta pelo protocolo e à segurança de acesso aos dados.

As deficiências do NFS e outros sistemas de arquivos populares favoreceram o aparecimento de sistemas de arquivos mais sofisticados. Alguns destes permitem, por exemplo, o espalhamento de dados em diversos servidores, preservação de desempenho mesmo em ambientes com grande número de clientes, e funcionamento em ambientes compostos por computadores e arquiteturas de redes distintas. Tais soluções incluem sistemas de arquivos proprietários, a exemplo do TerraFS e GPFS [Oberg et al. 2005], ou de distribuição livre, tais como Lustre [Lustre 2002], PVFS2 [Latham et al. 2004], OCFS2 [Oracle 2007], GlusterFS [Gluster 2007], Ceph [Wang 2006] e pNFS [Hildebrand et al. 2007].

Esse trabalho visa realizar um estudo qualitativo e quantitativo do uso de alguns sistemas de arquivos atuais em um ambiente de pequena escala com máquinas e arquitetura de rede de baixo custo. Os sistemas de arquivos avaliados foram: Lustre, PVFS2, GlusterFS e OCFS2 com iSCSI. Alguns estudos sobre o desempenho dos sistemas de arquivos foram realizados [Oberg et al. 2005], mas estes consideram geralmente ambientes de larga escala, a exemplo de parques computacionais compostos por diversas máquinas interconectadas por canais de fibra ótica, o que diverge do cenário atual da maioria das instituições de pesquisa e empresas nacionais. Portanto, acreditamos que o presente trabalho auxilie os administradores de sistemas fornecendo subsídios iniciais para uma escolha mais adequada do sistema de arquivos a ser utilizado no seu ambiente de rede de acordo com suas limitações.

O restante deste artigo está estruturado da seguinte maneira. A seção 2 resume o funcionamento dos sistemas de arquivos testados. A seção 3 apresenta o ambiente de testes, a metodologia utilizada e os resultados da avaliação experimental. A seção 4 destaca alguns aspectos dos sistemas segundo critérios qualitativos. Finalmente, a seção 5 apresenta nossas considerações finais.

2. Sistemas de Arquivos Distribuídos Avaliados

O objetivo fundamental da distribuição na camada do sistema de arquivos para redes é permitir que os clientes, servidores e dispositivos de armazenamento troquem dados através da rede sem a dependência de um repositório único e centralizado de dados. Neste contexto, tais sistemas devem apresentar desempenho que justifique sua utilização e, na medida do possível, atender a outros aspectos importantes: transparência e consistência de dados, distribuição de carga, replicação automática e a capacidade de atendimento em escala global [Kon 1996]. Esse estudo se concentrou em sistemas de arquivos que se destinam a prover parte destas funcionalidades. São estes: Lustre, PVFS2, OCFS2 e GlusterFS.

2.1. Lustre

A arquitetura do Lustre [Lustre 2002] é composta por servidores de metadados (*MDS - Metadata Servers*), servidores de I/O (*OST - Object Storage Targets*) e *OSD (Object-Based Disks)*. Os *MDS*s contém a árvore de diretórios, os atributos de permissões e de estado de cada objeto de armazenamento, já os *OST*s são responsáveis pelo gerenciamento dos *OSD*s, que por sua vez, são os representantes dos discos físicos. É possível agrupar um conjunto destas entidades num *LOV (Logical Object Volume)*, que permite a existência de subgrupos de clientes diferentes tendo acesso à partes distintas do sistema de arquivos.

O Lustre foi projetado para funcionar num ambiente heterogêneo de máquinas e redes. Através do seu conceito de *portal* é possível fazer o roteamento entre redes de arquiteturas diferentes, utilizando uma camada de abstração de rede em sua API interna (*Network Abstraction Layer*). Também destina-se a ser escalável em grandes proporções [Lustre 2005]. Contudo, nosso interesse está em analisar a arquitetura de software no Lustre (aparentemente mais complexa) quanto ao desempenho de acesso aos arquivos em ambientes mais simples e com escala menor.

Quanto ao tratamento de falhas, atualmente, o Lustre permite somente a configuração simultânea de dois servidores de metadados: um primário e outro secundário. Quando algum servidor de I/O se desconecta no Lustre, os clientes recebem notificação dessa falha pelo servidor de metadados que, por sua vez, redireciona as operações para outro servidor de I/O disponível. Visto que o servidor de metadados não se responsabiliza pela criação de réplicas, os dados armazenados no servidor falho permanecem inacessíveis. Para aumentar a disponibilidade dos servidores de metadados, é possível integrar soluções externas como o DRBD [Reisner 2005]. Quanto à gerência, o Lustre é compatível com bases LDAP, o que facilita o processo de configuração e viabiliza a distribuição dessas informações. Além disso, o suporte ao protocolo SNMP permite o envio de informações relativas ao monitoramento dos servidores.

2.2. PVFS2

Assim como o Lustre, o PVFS2 (*Parallel Virtual File System*) é um sistema de arquivos que permite a configuração de múltiplos servidores de metadados e I/O. Este sistema permite a existência de vários servidores de metadados. Na prática, é permitido que todos os servidores sejam ao mesmo tempo servidores de metadados e servidores de I/O. Contudo não há estratégia nativa para efetuar redundância de servidores de I/O, o que requer uso de ferramentas adicionais. Um diferencial de projeto é permitir acesso concorrente aos arquivos, sem garantia quanto à consistência dos dados, devido a flexibilidade no uso de (*lock* para acesso aos arquivos [Latham et al. 2004]. O PVFS2 não implementa exatamente a semântica POSIX, o que inviabiliza a execução de certas aplicações, a exemplo daquelas que precisam travar o arquivo para leitura ou escrita.

O PVFS2 é muito usado em ambientes de *cluster*, onde as aplicações são geralmente escritas em linguagens com recursos de paralelização e necessitam de acesso rápido e compartilhado de I/O. Nestes ambientes específicos, é plausível considerar-se que as aplicações são projetadas de modo que dispensar o uso de (*locks* sem danos colaterais. O PVFS2 funciona em arquiteturas de rede diferentes, podendo ser usado em TCP/IP ou Infiniband, por exemplo. Contudo, o PVFS2 é incapaz de executar o roteamento entre as redes, como o Lustre.

2.3. OCFS2 com iSCSI

O OCFS2 [Oracle 2007] é um sistema de arquivos de propósito geral projetado para viabilizar o armazenamento das bases de dados do Oracle RAC (*Real Application Cluster*). Assim, o projeto do OCFS2 priorizou o tratamento de problemas relativos ao gerenciamento distribuído de *locks* a fim de melhorar o acesso concorrente. O OCFS2 provê garantias quanto à consistência dos metadados e a manutenção de logs transacionais para os clientes.

Uma solução mista foi adotada com o intuito de permitir o compartilhamento dos discos disponíveis nas diferentes máquinas, uma vez que o OCFS2 não se destina a prover armazenamento em diferentes servidores de I/O. Para tanto, escolhemos o iSCSI (*SCSI over IP*) [Mug 2003], o qual disponibiliza, através de um par ip/porta, um ou mais dispositivos de blocos locais. O iSCSI define duas entidades: o *iSCSI Target* que representa a máquina com seus discos locais e o *iSCSI Initiator* que mapeia os *targets* como discos SCSI emulados. Utilizando um *iSCSI Initiator* em cada cliente do cenário de testes, construímos um RAID0 com todos os *iSCSI Targets* e aplicamos o sistema de arquivos OCFS2.

2.4. GlusterFS

Uma alternativa à inserção de um sistema de arquivos diretamente no kernel do sistema operacional consiste na utilização de uma API que permita a execução de tais sistemas em modo usuário. O módulo *FUSE (Filesystem USErspace)* define um dispositivo virtual que funciona como uma ponte entre as chamadas de sistema ao VFS *Virtual Filesystem* do Linux e as bibliotecas do sistema de arquivos implementado em espaço de usuário.

O GlusterFS [Gluster 2007] é um sistema de arquivos implementado sobre a API do FUSE e cujos objetivos principais são: escalabilidade, desempenho e disponibilidade. Cada volume ou nó de armazenamento no GlusterFS é conhecido como *brick*. O GlusterFS define ainda dois componentes básicos: o *scheduler* e o *translator*. O *scheduler* especifica a estratégia de distribuição dos arquivos enquanto o *translator* define outras características, a exemplo da política de *caching*.

O GlusterFS disponibiliza quatro opções de (*schedulers*) para distribuição dos dados: *random*, *Adaptive Least Usage (ALU)*, *Non-uniform Filesystem Scheduler(NUFA)* e *Round-Robin(RR)*. O *random* efetua uma escolha aleatória. O *ALU* utiliza informações de carga do servidor. O *NUFA* ajusta a alocação priorizando acessos locais. Enfim, na estratégia *Round-Robin* cada cliente possui uma fila circular de servidores para uso. Os componentes dos *translators* fornecem funcionalidades adicionais, a exemplo da replicação de arquivos em qualquer quantidade entre os servidores, e a fusão de vários volumes de armazenamento (*bricks*) fornecendo a abstração de um grande volume único [Gluster 2007].

3. Avaliação Experimental

Os testes foram realizados num cenário de pequena escala composto por cinco (5) servidores e quatorze (14) clientes, conforme apresentado na Figura 3. Os servidores eram computadores AMD Opteron 246 bi-processados de 64bits à 2 GHz e foram divididos em: quatro servidores de I/O (cada qual com 2 GB de memória RAM) e um servidor para

metadados (com 4GB de memória RAM)[1]. Os clientes consistiram em quatorze máquinas Intel Pentium 4 de 32bits à 3 GHz com *Hyperthreading* habilitado e 512 MB de memória RAM.

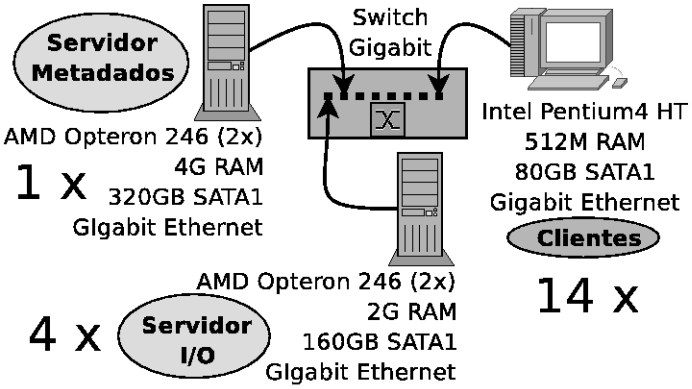


Figura 1. Cenário de testes: 5 servidores AMD Opteron e 14 clientes Intel Pentium 4

A rede de dados era composta por um *switch* Gigabit Ethernet de 24 portas com vazão total de 24 Gb/s, sendo todas as máquinas compatíveis com Gigabit Ethernet. O sistema operacional escolhido foi o Debian GNU/Linux cujas versões do kernel são apresentadas na Tabela 1.

Sistema de Arquivos	Intel Pentium4 HT 32bits	AMD Opteron 64bits
Lustre 1.4.7	kernel 2.6.12.6	kernel 2.6.12.6
PVFS2 2.6.3	kernel 2.6.18-4-686	kernel 2.6.18-4-amd64
OCFS2 1.2.1 + iSCSI	kernel 2.6.18-4-686	kernel 2.6.18-4-amd64
FUSE 2.6.3 + GlusterFS 1.3pre2.3	kernel 2.6.18-4-686	kernel 2.6.18-4-amd64

Tabela 1. Versões do kernel do linux para cada sistema de arquivos

3.1. Limites empíricos para a taxa de transferência da rede e dos discos

A fim de determinar o limite empírico das métricas de desempenho no ambiente de testes e poder avaliar o custo geral imposto pelo hardware/sistema operacional sobre a experimentação, obtivemos alguns resultados experimentais referentes à taxa de transferência dos discos rígidos e da rede de comunicação. Os discos rígidos utilizados atendem ao padrão SATA [SATA 2002] (especificação 1.0), cujo máximo teórico de transferência é 150 MB/s. Contudo, nos servidores AMD Opteron, observamos o desempenho médio de 69 MB/s, enquanto as estações Intel Pentium 4 obtiveram média de 62 MB/s. Estes valores foram obtidos através do programa sg_dd. Quanto à taxa de transferência da rede utilizada, temos um máximo teórico de 1 Gb/s, ou seja 125 MB/s.Utilizamos o pacote netpipe-tcp [NetPIPE 2005] para verificar o mesmo valor empiricamente. As máquinas Intel Pentium 4 obtiveram vazão máxima em torno de 600 Mb/s (75 MB/s) ao passo que os servidores AMD Opteron alcançaram 900 Mb/s (112,5 MB/s).

3.2. Metodologia de Avaliação

O desempenho dos sistemas de arquivos foi medido através da suíte de testes IOzone [IOzone 2006]. O uso dessa ferramenta permitiu: i) a medição da vazão dos acessos

[1]No caso dos sistemas OCFS2 e GlusterFS, não utilizamos o servidor de metadados, pois tal conceito não existe em suas arquiteturas.

de leitura e escrita no sistema de arquivos; ii) a compatibilidade entre as distintas arquiteturas; iii) a geração de arquivos tabulados, facilitando a plotagem de gráficos. O IOzone permite a análise do tempo decorrido e a vazão dos acessos no sistema de arquivos de acordo com a variação do tamanho do arquivo e do bloco de dados. Foram utilizados valores entre 16 MB e 4 GB para o tamanho dos arquivos e 64 KB e 16 MB para o bloco de dados. Coletamos cinco amostras de dados para os testes realizados.

No ambiente de testes, cada cliente executa uma instância do IOzone. Cada uma dessas instâncias executa em uma pasta diferente, dentro do ponto de montagem compartilhado, evitando, assim, que duas instâncias diferentes usem o mesmo diretório ou o mesmo arquivo. Um programa que monitora a atividade do IOzone verifica, a cada trinta segundos, o final da execução do teste em andamento para todas as máquinas clientes, a fim de evitar sobreposição nas execuções dos testes. As máquinas utilizadas eram dedicadas ao procedimento de testes, não havendo qualquer uso extra por outros programas ou usuários.

Conforme a estratégia adotada por [Oberg et al. 2005], utilizamos arquivos ao menos 50% maiores do que a quantidade memória RAM do cliente. Assim, procuramos evitar que os dados analisados sejam influenciados por questões de cache. Uma discussão mais ampla sobre a problemática do cache pode ser obtida em [de Carvalho 2005]. Nos experimentos realizados, não foram utilizados mecanismos de redundância disponíveis em alguns sistemas de arquivos, devido ao interesse em observar o desempenho na ausência de configurações que pudessem diminuir o desempenho geral. Por fim, optamos por avaliar o comportamento dos sistemas de arquivos com base na configuração padrão originalmente fornecida, sem efetuar otimizações específicas. Em particular, a configuração do PVFS2 limitou-se a avaliar sua capacidade de dividir os arquivos (*file stripping*) armazenados em múltiplos servidores de I/O. Já no caso do GlusterFS, utilizamos seu escalonador randômico em conjunto com o *unify translator* para prover a visão conjunta de todos os volumes de armazenamento.

3.3. Análise dos Resultados

A métrica fundamental da avaliação experimental consistiu na observação da taxa de transferência obtida por cada sistema de arquivos mediante operações de leitura e escrita com o IOzone. Esta avaliação foi realizada em duas etapas. Na primeira etapa, efetuamos a transmissão de arquivos de tamanhos distintos, entre 16 MB e 4 GB (com blocos de 16 MB). Na segunda etapa, observamos o comportamento do sistema na transferência de um arquivo de 1 GB para tamanhos de blocos distintos, entre 64 KB e 16 MB. Os resultados dos experimentos são apresentados nos gráficos das Figuras 2 a 4. Estes mostram a taxa de transferência média (em MB/s) obtida pelo sistema de arquivos frente à variação do tamanho do arquivo ou do bloco. Cabe ressaltar que a escala vertical dos gráficos varia em base 2 para melhor visualização dos resultados e que o desvio padrão é indicado por barras verticais.

Variação no tamanho dos arquivos

As Figuras 2 e 3 referem-se às operações de de leitura e escrita, respectivamente, frente à variação do tamanho do arquivos transferido, sendo o tamanho do registro (bloco) fixado em 16 MB.

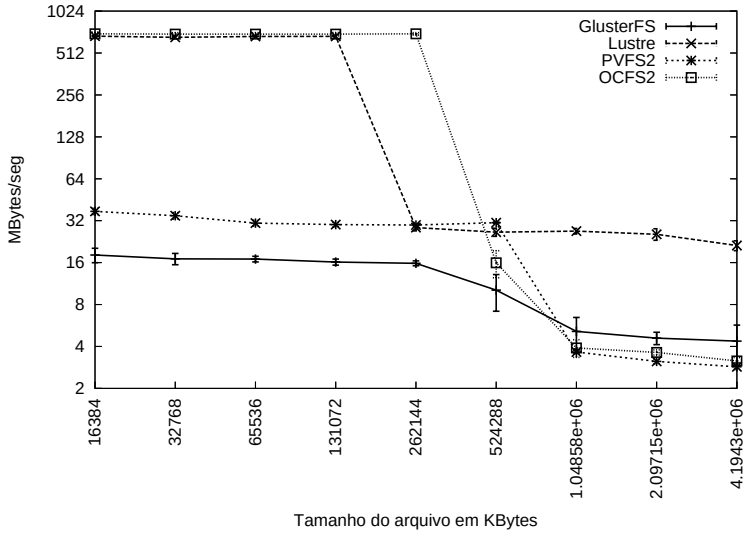


Figura 2. Taxa de transferência dos clientes em operações de leitura para diferentes tamanhos de arquivos

Operações de leitura. À luz da Figura 2, observa-se elevadas taxas de transferência para arquivos menores que 256 MB, exceto para o PVFS2 e GlusterFS. Creditamos esse comportamento ao fato do PVFS2 do GlusterFS não implementarem estratégias de cache para operações de leitura ou escrita em sua configuração básica. É possível observar que o Lustre e a solução combinada OCFS2 e iSCSI obtiveram desempenho superior para a operação de leitura quando comparados à solução do PVFS2 e GlusterFS. Para tamanhos de arquivos superiores a 1 GB, os sistemas apresentaram comportamento bastante similar. Dentre estes, o Lustre foi melhor com taxa de transferência em torno de 27 MB/s (desvio de 4%) para arquivos de 1 GB, mantendo tal comportamento para arquivos de 2 GB e 4 GB.

Operações de escrita. A Figura 3 apresenta os resultados obtidos para operações de escrita. Para arquivos de 1GB, o PVFS2 se destacou com média por cliente de 13 MB/s (1% de desvio). Em seguida, aparecem o Lustre com 12 MB/s (desvio de 7%), o GlusterFS com 10 MB/s (desvio de 7%) e o OCFS2 com 3 MB/s (desvio de 7%). Esse comportamento se repete para arquivos de 2 GB e 4 GB.

Variação no tamanho de bloco

Com o objetivo de avaliar o impacto do tamanho de bloco (registro) lidos e escritos através da rede, observamos o desempenho para arquivos de 1GB variando o tamanho do bloco entre 64 KB e 16 MB, apresentados nos gráficos da Figura 4.

Operações de leitura. A Figura 4(b) apresenta os resultados referentes às operações de leitura. Neste cenário, o Lustre apresentou taxas de transferências superiores aos outros sistemas de arquivos para operações de leitura (pico de 26 MB/s com desvio de 4,5%) considerando todos os tamanhos de registros.

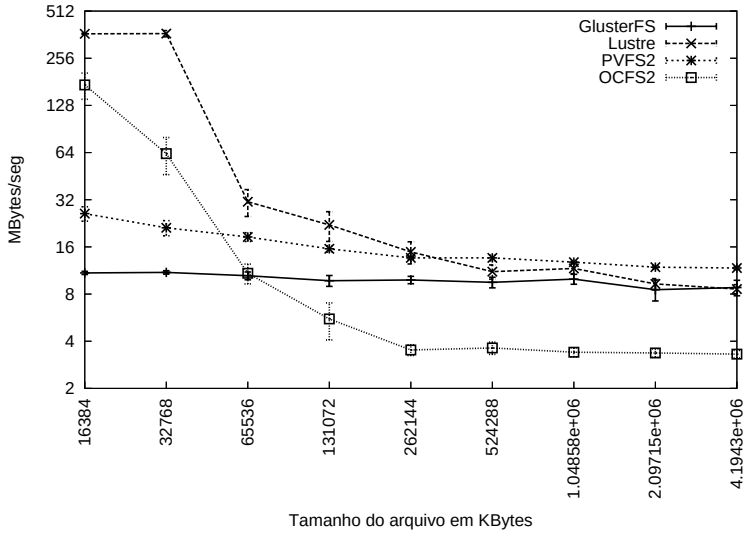


Figura 3. Taxa de transferência dos clientes em operações de escrita para diferentes tamanhos de arquivos

Operações de escrita. Em relação às operações de escrita, observamos pela Figura 4(a) que o PVFS2 apresenta os melhores resultados atingindo até 13 MB/s (desvio de 0,7%) para o registros de 4 MB.

A tabela 2 resume o desempenho dos sistemas de arquivos apresentando as melhores taxas de transferência de cada sistema de arquivos. Com base em nossos experimentos, aferimos que o Lustre se comporta melhor com registros de 16 MB e o PVFS2 com registros de 4 MB. Os melhores resultados do OCFS2 ocorreram para registros de 256 KB para operações de leitura e de 4 MB para escrita. O GlusterFS, por sua vez, apresentou bom desempenho para tamanho de registros entre 8 MB e 16 MB.

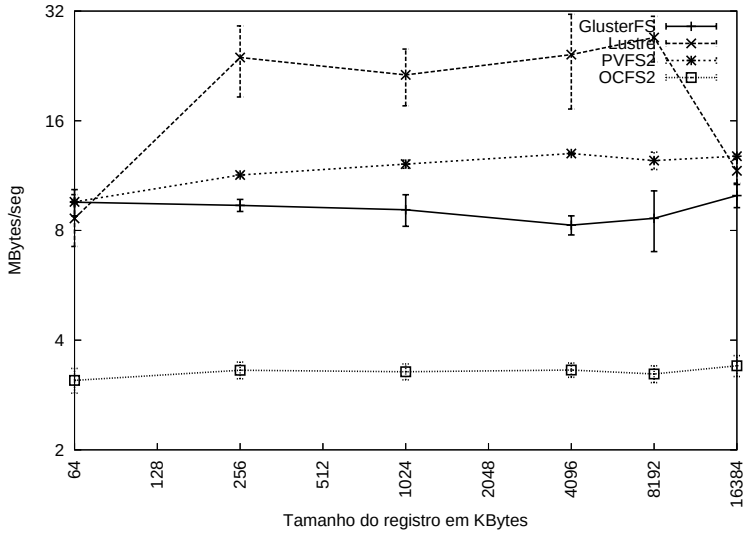
Sistema	leitura	tam. registro	escrita	tam. registro
Lustre	26 MB/s	16 MB	12 MB/s	16 MB
PVFS2	5 MB/s	4 MB	13 MB/s	4 MB
OCFS2+iSCSI	7 MB/s	256 KB	3 MB/s	4 MB
GlusterFS	5 MB/s	8 MB	10 MB/s	16 MB

Tabela 2. Resumo das melhores taxas de transferência obtidas para operações de leitura e escrita para um arquivo de 1 GB com variação do tamanho de bloco entre 64 KB e 16 MB

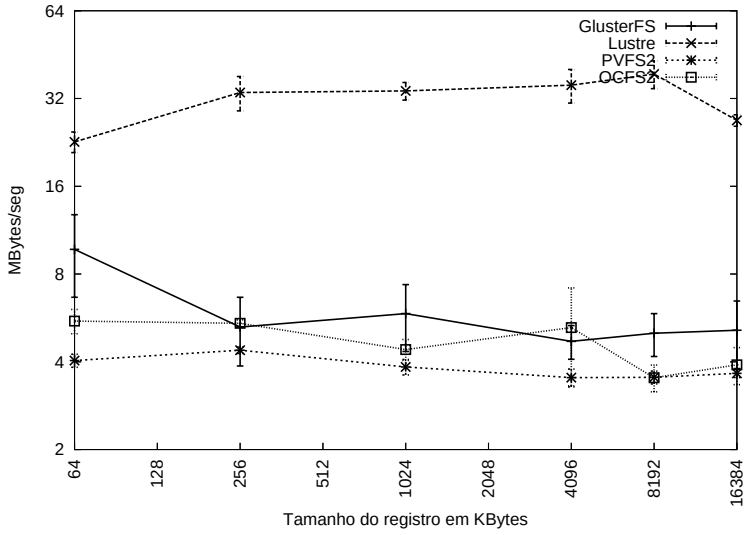
4. Avaliação Qualitativa

A tarefa cotidiana de administração de um sistema de arquivos distribuído, por vezes, pode requerer esforço maior do que adoção de uma solução centralizada. Por exemplo, a simples distribuição dos dados e processamento implica geralmente no aumento dos pontos de falha, o que requer cuidados específicos. Portanto, faz-se necessária a avaliação de outros aspectos importantes no momento de implantação, como a facilidade de configuração e administração do sistema distribuído. Essa seção visa fornecer uma idéia geral sobre os aspectos práticos de instalação e configuração dos sistemas distribuídos analisados.

O Lustre apresenta certas dificuldades quanto à instalação devido à dependência específica da versão 2.6.12.6 do kernel do Linux. Isso inviabilizou a aplicação de



(a) Operações de escrita



(b) Operações de leitura

Figura 4. Taxa de transferência dos clientes por tamanho do registro para arquivos de 1GB

atualizações (*patches*) das versões mais recentes do kernel. A dificuldade inicial de configuração é compensada pelo conjunto de ferramentas de manutenção que são capazes, por exemplo, de atualizar facilmente o estado da configuração em casos de adição de novos servidores.

O OCFS2 com o iSCSI, por sua vez, requer cuidados particulares quanto à administração. Por exemplo, é preciso ajustar as etapas de inicialização do servidor de arquivos de tal modo que os dispositivos de armazenamento SCSI via IP estejam disponíveis antes da inicialização do dispositivo de RAID. Além disso, a adição ou remoção de um servidor requer a reconstrução do arranjo RAID e nova formatação do dispositivo de blocos. Em contrapartida, o OCFS2 possui boas ferramentas para o redimensionamento de uma partição e adição de novos clientes em tempo de execução.

O PVFS2 possui ferramentas que facilitam bastante a tarefa de configuração e

administração, mas demanda reinício dos clientes em caso de adição de servidores. Um aspecto importante é que a falha de um servidor de I/O pode ser contornada caso o cliente escolha outro servidor em tempo de execução. Da mesma forma, a desconexão do cliente não traz impacto ao sistema devido à inexistência de estratégias de caching.

A configuração do GlusterFS se destaca pela dependência do FUSE, o que permite, por exemplo, utilizar um servidor através de uma conta comum de usuário (*i.e.*, sem privilégios de super-usuário). No GlusterFS, a adição de novos servidores também implica em reiniciar o serviço do cliente. Uma boa opção ao administrador é ativar o *translator AFR* para a replicação de arquivos a partir do seu tipo ou nome em qualquer quantidade desejada.

5. Conclusão

O estudo comparativo de sistemas de arquivos é importante para, dentre outros, guiar os administradores na escolha da opção mais adequada para o seu ambiente computacional. Este trabalho apresenta um tal quadro comparativo referente a aspectos tanto qualitativos quanto experimentais de quatro sistema de arquivos emergentes, a saber Lustre, PVFS2, OCFS2 com iSCSI e GlusterFS. Até onde sabemos, existem poucos trabalhos referentes ao estudo do desempenho de tais sistemas em ambientes de pequena escala.

No estudo realizado, identificou-se o Lustre como uma solução mais robusta e eficaz, sendo recomendado mesmo para ambientes com poucas máquinas e de baixo custo. O Lustre apresentou diversas vantagens, relativas à manutenção, capacidade de monitoramento (integrado com SNMP) e suporte à arquiteturas de redes diferentes. Além disso, em seu projeto, busca-se uma compatibilidade ao paralelismo do acesso aos *OST*s, o que trará benefícios às aplicações inerentemente paralelas.

Por outro lado, caso se busque soluções de armazenamento para uso conjunto com outras plataformas, tais como grades, sugere-se a adoção de soluções baseadas no FUSE, devido à facilidade de personalização e adição de novos recursos em espaço de usuário. Por exemplo, é possível integrar o GlusterFS a um pacote de software que provenha acesso compartilhado a dados distribuídos, tal como proposto em [dos Santos and Cerqueira 2006].

Outra opção de uso das soluções em FUSE seria para o estudo e avaliação de novas funcionalidades em sistemas de arquivos. Por exemplo, o sistema Ceph [Wang 2006] visa o provimento de alto desempenho e confiabilidade através do gerenciamento dos dados por funções pseudo-aleatórias no lugar do uso das tabelas de alocação. Infelizmente, neste trabalho não foi possível avaliar o seu desempenho, devido à problemas de instalação, o que sugere objeto de investigação futura.

Referências

de Carvalho, R. P. (2005). Sistemas de arquivos paralelos: Alternativas para redução de gargalo no acesso ao sistema de arquivos. Master's thesis, Universidade de São Paulo.

dos Santos, M. N. and Cerqueira, R. (2006). Gridfs: Targeting data sharing in grid environments. In *CCGRID '06: Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID'06)*, page 17, Washington, DC, USA. IEEE Computer Society.

Gluster, C. T. (2007). Glusterfs user guide. Maio de 2007: http://www.gluster.org/docs/index.php/GlusterFS_User_Guide.

Hildebrand, D., Adamson, A., and Honeyman, P. (2007). pnfs and linux: Working towards a heterogeneous future. Technical report, CITI - University of Michigan.

IOzone (2006). File system benchmark. Maio de 2007: http://www.iozone.org.

Kon, F. (1996). Distributed file systems past, present and future: A distributed file system for 2006. Technical report, University of Ilinois at Urbana-Campaign.

Latham, R., Miller, N., Ross, R., and Carns, P. (2004). An introdution to the second parallel virtual file system. Maio de 2007: http://www.pvfs.org/files/linuxworld-JAN2004-PVFS2.ps.

Lustre, C. F. S. I. (2002). Lustre: A scalable, high-performance filesystem. Technical report. Maio de 2007: http://www.lustre.org/docs/whitepaper.pdf.

Lustre, C. F. S. I. (2005). Selectig a scalable cluster file system. Technical report. Maio de 2007: http://www.lustre.org/docs/selecting-a-cfs.pdf.

Mug, M. (2003). Performance comparison between iscsi and other hardware and. software solutions. In *CHEP03: Computing in High Energy and Nuclear Physics*.

NetPIPE (2005). A network protocol independent performance evaluator. Maio de 2007: http://www.scl.ameslab.gov/netpipe/.

Oberg, J. C. M., Tufo, H. M., and Woitaszek, M. (2005). Shared parallel filesystems in heterogeneous linux multi-cluster environments. In *6th LCI International Conference on Linux Clusters: The HPC Revolution*.

Oracle (2007). Ocfs2 user's guide. Maio de 2007: http://oss.oracle.com/projects/ocfs2/dist/documentation/ocfs2_users_guide.pdf.

Reisner, P. (2005). Drbd v8: Replicated storage with shared disks semantics. In *12th International Linux System Technology Conference*.

SATA, I. O. (2002). Serial ata: A comparison with ultra ata technology. Maio de 2007: http://www.sata-io.org.

Wang, F. (2006). *Storage Management in Large Distributed Object-Based Storage Systems*. PhD thesis, University of California.