Comparativo entre PVFS2 e GlusterFS em um Ambiente de Computação em Nuvem

Maurício A. Pillon, Mauro Hinz, Pedro C. B. de Azevedo Filho, Ricardo J. Pfitscher

[1]Departamento de Ciências da Computação
Universidade do Estado de Santa Catarina (UDESC) – Joinville, SC – Brazil

mauricio.pillon@udesc.br, {mmhinz,pedrocesar.ti,ricardo.pfitscher}@gmail.com

Abstract. *This paper presents a comparative study on the performance of two distributed file systems, the Parallel Virtual File System 2 (PVFS2) and the GlusterFS in a cloud computing environment. The comparative test is based on measuring throughput of reading and writing operations performed from a client to a server set virtualized data. The tabulated results indicate that PVFS2 maintains constant throughput for all file sizes. GlusterFS maintains high throughput for 1MB files with a significant decrease for larger files, and in these cases PVFS2 has underperformed.*

Resumo. *Este trabalho apresenta um estudo comparativo sobre o desempenho de dois sistemas de arquivos distribuídos, o Parallel Virtual File System 2 (PVFS2) e o GlusterFS, em um ambiente de computação em nuvem. O estudo comparativo baseia-se no teste de medição de vazão relacionada às operações de leitura e escrita realizadas a partir de um cliente para um conjunto de servidores de dados virtualizados. Os resultados tabulados indicam que PVFS2 mantém vazão constante, e próxima a largura de banda disponível, para todos os tamanhos de arquivos e que GlusterFS mantém alta vazão para arquivos de 1MB com queda significativa para arquivos maiores, e nestes casos tem desempenho inferior ao PVFS2.*

1. Introdução

O advento do uso de tecnologias de computação em nuvem apresenta uma utilização crescente nos últimos anos, e alguns trabalhos têm sido desenvolvido neste sentido [Beloglazov and Buyya 2010], [Rao et al. 2011], [Rimal et al. 2009], [Buyya et al. 2011], um dos principais benefícios é a entrega de recursos sobre demanda, ou seja, de acordo com as necessidades dos usuários.

Sistemas de arquivos distribuídos (SAD) objetivam prover formas de armazenamento de dados escaláveis e com altas taxas de vazão, onde, a adição de múltiplos nós, agregado a alta largura de banda, possibilita o aumento do desempenho de acesso a arquivos. Quando comparado ao sistema de arquivos centralizado, a tendência é de que o SAD tenha uma maior taxa de vazão. Além disto, SAD agregam outras características de sistemas distribuídos como disponibilidade e transparência.

Um SAD paralelo tende a apresentar uma taxa de vazão próximo ao limite do meio de comunicação, uma vez que os acessos aos arquivos são realizados a múltiplos discos

paralelamente, como exemplos típicos destes sistemas têm-se o PVFS2 [Haddad 2000] e o GlusterFS[1].

Tendo em vista que SAD podem oferecer disponibilidade e escalabilidade com o aumento de nós no sistema, e que ambientes de computação em nuvem oferecem a possibilidade de agregar recursos de acordo com a demanda, intui-se que é possível elaborar um ambiente de SAD sobre demanda, onde novas instâncias de SAD podem ser adicionadas de acordo com as necessidades das aplicações.

Sendo assim, este trabalho objetiva comparar dois SAD sobre a execução em um ambiente de nuvem computacional, a fim de identificar qual deles apresenta o melhor desempenho nestes ambientes. E traz como contribuição a característica do comportamento destes SAD sobre um cenário virtualizado, onde, PVFS2 mantém uma taxa de vazão constante e próxima a largura de banda disponível, e GlusterFS apresenta altas taxas de vazão para arquivos de tamanho próximo ao tamanho do bloco do sistema de arquivo local.

O restante deste trabalho está organizado da seguinte forma, na seção 2 são descritos os conceitos básicos sobre sistemas de arquivos distribuídos paralelos, bem como, faz se uma descrição das características de PVFS2 e GlusterFS. Na seção 3 é apresentado a avaliação experimental, com suas as características de implementação e os resultados obtidos. Na seção 4 são discutidos os principais trabalhos relacionados. Por fim, na seção 5 tem-se a conclusão.

2. Sistema de Arquivos Distribuídos

Estudos apontam que no decorrer dos últimos anos a taxa de acesso a disco não evoluiu na mesma proporção que as taxas de acesso a rede ou as taxas de processamento. Os servidores de dados podem assim tornar-se os gargalos em implementações que necessitam de acesso a dados remotamente. Neste sentido, os sistemas de arquivos distribuídos (SAD) buscam agregar recursos computacionais em prol de uma alta vazão no acesso aos dados [Goldman and de Carvalho 2005].

O principal objetivo de um sistema de arquivos distribuído é ser visto pelo usuário final, ou aplicação, como um sistema de arquivos íntegro e local [Levy and Silberschatz 1990]. Os mesmos conceitos aplicados a sistemas de arquivos convencionais podem ser utilizados aqui, ou seja, as terminologias utilizadas e abstrações permanecem as mesmas.

Os sistemas de arquivos baseados em *clusters* apresentam como principal finalidade, o armazenamento de dados em um ambiente de altíssima heterogeneidade e concorrência, devido à própria estrutura de um *cluster* computacional. Geralmente a adoção de um SAD em *clusters* está relacionada a sua ótima escalabilidade aliada a sua capacidade de abertura que permite aos SAD aumento de nós no sistema sem muita dificuldade. Outro fator importante a ser observado está relacionado a taxa de transferência da rede, já que em algumas implementações o SAD acaba sendo limitado pelo próprio *throughput* desta rede, sendo necessário do SAD uma certa ”inteligência” na utilização deste recurso.

A utilização de SAD pela comunidade científica é bastante grande. A busca por altas taxas de vazão e a necessidade de acesso a dados em diferentes locais, acabam por

[1]http://www.gluster.org/

fazer dos SAD uma alternativa viável. Sendo que a utilização do sistema pode variar devido a necessidade do usuário, como por exemplo, um SAD pode servir como base para uma aplicação distribuída ou paralela. Apresentando-se como estrutura para acesso a múltiplos dados distribuídos.

A grande parte das aplicações que executam neste ambiente trabalham de forma paralela e muitas delas implementam políticas e metodologias visando a tolerância a falhas. Entre as aplicações que utilizam esta arquitetura, encontra-se o PVFS2 e o GlusterFS, que possuem uma quantidade grande de seguidores [Barbosa et al. 2007].

2.1. PVFS2

O *Parallel Virtual File System 2* (PVFS2) apresenta algumas características marcantes em relação a desempenho, em linha gerais o sistema foi desenvolvido com a ideia de prover excelente desempenho no acesso aos dados em espaço de usuário [Carns et al. 2000], além de proporcionar um ótimo funcionamento em ambientes que manipulam grandes quantidades de dados.

O sistema permite a configuração de diversos servidores de *I/O* e servidores de metadados, inclusive permite a configuração de servidores de *I/O* e metadados executando sobre a mesma máquina (configuração recomendada). Os dados são divididos entre os servidores de *I/O* por escalonamento *Round-Robin* e as informações destes dados (metadados) são escalonadas aos servidores de metadados da mesma forma.

O PVFS2 é utilizado geralmente em *clusters*, onde as aplicações são desenvolvidas para o trabalho de forma paralela, o mesmo é desenvolvido sobre duas Interfaces, UNIX API e uma variante do MPI-IO (ROMIO)[2]. Porém, não apresenta uma semântica POSIX ”pura”. O que complicaria na sua integração com algumas aplicações não desenvolvidas especificamente para lidar com este sistema de arquivos.

Outras características gerais como consistência e cache não são muito exploradas pelo PVFS2, demonstrando mais uma vez o enfoque do projeto voltado a vazão. O PVFS2 permite acesso concorrente aos arquivos, mas não garante atomicidade operações, deixando a cargo da aplicação o tratamento relacionado à consistência. Quanto à cache é permitida a sua configuração, entretanto não é um padrão de configuração.

2.2. GlusterFS

O GlusterFS foi desenvolvido com foco principal em desempenho, entretanto, também traz características como disponibilidade e escalabilidade.

O sistema é desenvolvido sobre a API FUSE [3] que funciona como uma ponte para as chamadas de sistema (via VFS) acessarem os dados contidos no espaço de usuário. Apesar da dependência do FUSE (aplicado de forma nativa pelo kernel Linux), pode ser facilmente instalado e configurado, pois os pacotes estão disponíveis nos repositórios das principais distribuições de sistemas baseados em kernel Linux.

O GlusterFS não apresenta em sua arquitetura servidores de metadados e os dados são armazenados em volumes, conhecidos por *bricks*, estes, são espalhados por diver-

[2]http://www.mcs.anl.gov/research/projects/romio/

[3]http://fuse.sourceforge.net/

sos dispositivos de armazenamento disponibilizados pelos nós que compõem o sistema [ZResearch 2005].

Para realizar a distribuição dos dados o GlusterFS utiliza 4 tipos diferentes de escalonadores. São eles: *Random*, *Adaptive Least Usage* (ALU), *Non-uniform Filesystem Scheduler*(NUFA) e *Round-Robin*(RR). Outras características apresentadas pelo GlusterFS são: a adoção, por padrão, de cache nos clientes; a não garantia de consistência dos dados.

3. Avaliação Comparativa

A fim de avaliar o comportamento das ferramentas PVFS2 e GlusterFS, fez-se a implementação de um ambiente que simula uma nuvem computacional, onde, um cliente utiliza um SAD com servidores sobre máquinas virtuais na nuvem.

3.1. Cenário de testes

O cenário sobre o qual os SAD foram avaliados utiliza basicamente dois computadores, o primeiro atua como cliente enquanto o segundo atua como servidor de máquinas virtuais (MVs). As características de *hardware* dos dois computadores utilizados são idênticas, sendo, processador AMD Phenom II X4 B93 2,8 GHz (*quad-core*), 4 GB de memória RAM e disco 500 GB SATA. A conexão entre eles é provida por um *switch* a uma velocidade de 100Mbps.

No ambiente representado pela Figura 1, cada SAD dispunha de duas MVs, onde seu respectivo serviço estava configurado. Sobre o cliente os serviços-cliente do SAD foram configurados para conectar-se as respectivas MVs servidoras, e executou-se testes de desempenho com as operações de escrita, reescrita, leitura e releitura de arquivos.

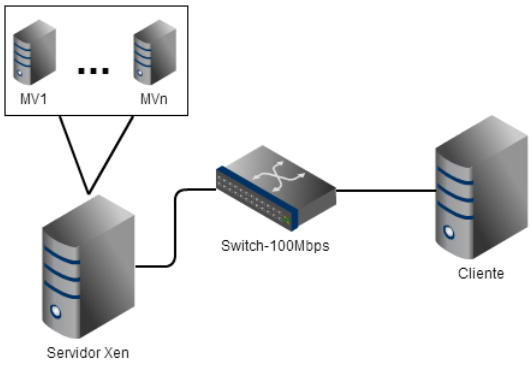

Figura 1. Ambiente de testes

Em termos de *software*, sobre o cliente executa-se o sistema operacional Debian Squeeze Kernel 2.6.32-5, no servidor o sistema operacional originalmente instalado é o Debian Squeeze 64 bits, Kernel 2.6.32-5, onde, sobre este instalou-se o *hypervisor* XEN[4] na versão 4.0.1. As máquinas virtuais executam através da técnica de paravirtualização e sistema operacional Debian Squeeze 64bits Kernel 2.6.32-5, a elas foi destinado uma VCPU, 256MB de memória e disco de 4GB. O sistemas de arquivos local do cliente, do servidor hospedeiro e das MVs é o ext3.

[4]http://www.xen.org/

O servidor e o cliente do PVFS2 foram instalados com sua configuração padrão, na versão 2.8.2. A ferramenta GlusterFS também foi executada de acordo com suas configurações padrão, na versão 3.2.7, contudo, a linha no arquivo configuração que mantinha a memória *cache* do cliente em 512MB foi apagada, a fim de que não afetasse os resultados, uma vez que PVFS2 por padrão não utiliza *cache* no cliente.

Para avaliar o desempenho dos SAD utilizou-se a ferramenta de *benchmark* de sistemas de arquivos *iozone* [5], com os testes de leitura, releitura, escrita e reescrita. A ferramenta executava no cliente e os testes foram feitos sobre arquivos com tamanhos entre 2^0 e 2^7 MB.

3.2. Resultados

A fim de obter dados estatisticamente comparáveis, executaram-se 25 repetições da execução da ferramenta *iozone* no cliente para cada SAD avaliado, além disto, cada execução foi independente, ou seja, enquanto avaliava-se um SAD o outro mantinha-se ocioso. Os resultados obtidos são expostos na Tabela 1 e ilustrados no gráfico da Figura 2 estes dados representam a média das 25 medições em cada SAD, e eliminou-se o maior e o menor valor obtido, desta forma o desvio padrão é inferior a 3% em todos os casos. Cada linha no gráfico representa a dupla (SAD,teste), onde o teste foram classificados como escrita (w), reescrita (rw), leitura(r) e releitura(rr).

Tabela 1. Comparativo entre PVFS2 e GlusterFS

	Escrita (MB/s)		Reescrita (MB/s)		Leitura (MB/s)		Releitura (MB/s)	
Tam. 2^nMB	PVFS2	GlusterFS	PVFS2	GlusterFS	PVFS2	GlusterFS	PVFS2	GlusterFS
1	10.40	253.31	10.78	248.91	10.85	1142.37	10.85	1155.52
2	10.49	12.00	10.78	11.97	10.84	20.75	10.83	20.74
3	10.62	7.57	10.77	7.59	10.76	13.78	10.82	13.81
4	10.68	6.39	10.64	6.42	10.81	11.80	10.84	11.80
5	10.65	5.95	10.72	5.98	10.83	11.01	10.84	11.01
6	10.68	5.77	10.77	5.78	10.84	10.64	10.84	10.61
7	10.28	5.68	10.75	5.47	10.84	10.48	10.84	10.33
8	10.70	5.63	10.74	5.45	10.81	10.40	10.84	10.07

A título de comparação, os mesmos experimentos foram executados sobre o sistema de arquivos local do cliente (ext3), e estes resultados são apresentados na Tabela 2.

3.3. Análise

Ao avaliar os resultados apresentados na seção 3.2 faz-se duas constatações:

1. Para arquivos de 1MB, o GlusterFS apresenta resultados muito superiores ao PVFS2, ultrapassando a vazão máxima da rede (12,5 MB/s) em aproximadamente 100 vezes, para os testes que envolvem leitura, e em aproximadamente 20 vezes, para testes que envolvem escrita. Esta característica pode ser atribuída à utilização da API FUSE pelo GlusterFS, mantendo os dados no espaço de usuário do cliente antes de realizar a transferência, assim, para os testes (escrita, leitura e releitura)

[5] http://www.iozone.org/

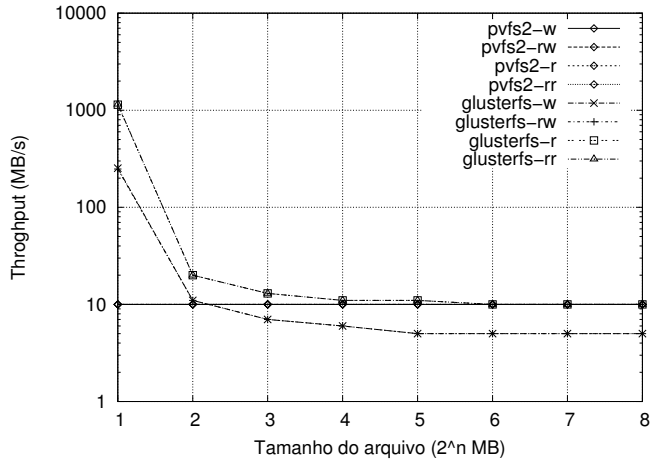

Figura 2. Comparativo entre PVFS2 e GlusterFS

Tabela 2. Medições sobre sistema de arquivos local

Tam 2^nMB	Escrita (MB/s)	Reescrita (MB/s)	Leitura(MB/s)	Releitura(MB/s)
1	254.77	1188.94	1188.94	1171.47
2	259.51	1203.98	1203.98	1547.25
3	284.99	1199.42	1199.42	2175.16
4	399.52	1357.48	1357.48	2241.90
5	536.88	1516.35	1516.35	2549.61
6	654.32	1976.23	1976.23	2709.13
7	713.51	2478.18	2478.18	2980.01
8	740.25	3129.14	3129.14	3151.61

com arquivo de 1MB a taxa de vazão obtida é comparável à mensurada sobre o sistema de arquivos local para o mesmo tamanho de arquivo (Tabela 2).

2. O PVFS2 mantém, para todos testes e tamanhos de arquivo, uma vazão constante e em aproximadamente 85% da vazão máxima da rede, enquanto, para os mesmos cenários, o GlusterFS apresenta diminuição de vazão conforme o crescimento do tamanho do arquivo.

Com base nestas duas conclusões pode-se afirmar que o sistema de arquivos distribuído PVFS2 tende a manter vazões de acesso a arquivos próximas a limitação do meio de transmissão de dados, enquanto, a ferramenta GlusterFS apresenta resultados favoráveis para arquivos de menor tamanho.

Contudo, ressalta-se que nos experimentos realizados as máquinas virtuais foram hospedadas pela mesma máquina física, e compartilhavam da mesma interface de rede. Em um ambiente de computação em Nuvem, espera-se que múltiplos nós do SAD sejam alocados em máquinas físicas distintas, e desta forma, com a menor concorrência sobre o meio de transmissão atingiria maiores taxas de vazão. [6]

[6]Em um teste realizado pela equipe do GlusterFS em conjunto com a Amazon Aws, verificou-se que conforme aumenta-se a largura de banda disponível e o número de nós, é possível aumentar de forma proporcional a taxa de vazão de acesso a arquivos [Gluster 2011]

4. Trabalhos Relacionados

Devido à forte aceitação da comunidade e sua base bem difundida, hoje tem-se um grande número de SADs, cada uma com sua proposta particular, além do PVFS2 e do GlusterFS, outros sistemas de mesmo prestígio podem ser citados como Lustre [System 2007], Ceph [WEIL et al. 2006], GPFS [Shmuck and Haskin 2002], entre outros.

Ao longo dos últimos anos, outros estudos buscaram avaliar o desempenho de SADs. Alguns deles avaliaram o impacto da estrutura de rede sob o funcionamento dos SADs [Apon et al. 2002]. Alguns estudos apresentam modificações na estrutura de rede buscando aumentar o *throughput*, como a adoção de Myrinet, Infiniband entre outras, como tecnologia de interligação [Kim et al. 2009].

Os estudos realizados por [Xu et al. 2010] [Macedo et al. 2009] fazem o uso de SAD's para melhorar o desempenho de aplicações que necessitam de um serviço para armazenamento de dados. Esta melhoria de desempenho foi verificada através da comparação com a solução não distribuída.

Neste trabalho realizou-se a análise de desempenho, voltado a vazão, em SAD sobre um ambiente que simula uma nuvem computacional. Os resultados obtidos, comparados a outros estudos realizados, apresentam resultados semelhantes. Em [Macedo and Azevedo 2011] os resultados do PVFS2 em uma rede local são comparáveis aos dados apresentados neste trabalho. Assim como, o comportamento linear na medição da vazão para tamanhos diferentes de arquivos no PVFS2, igualmente é observado em [Barbosa et al. 2007]. Outro resultado que chama a atenção é a queda brusca de rendimento do GlusterFS para arquivos grandes, também demonstrado em [Barbosa et al. 2007].

5. Conclusão

Ambientes de computação em nuvem têm se tornado foco de pesquisa na comunidade científica, a principal característica destes ambientes é a possibilidade de alocar recursos sobre demanda, o que facilita na adequação dos recursos à necessidade das aplicações dos clientes. Uma das linhas de pesquisa pouco abordadas nestes ambientes é a utilização de sistemas de arquivos distribuídos (SAD).

O propósito geral de um SAD é manter os dados disponíveis, escaláveis e com alta velocidade de acesso, neste sentido, este artigo propõe o uso de SAD sobre nuvens computacionais, onde, faz-se uma avaliação de dois típicos SAD utilizados na comunidade científica, o PVFS2 e o GlusterFS. A avaliação baseia-se no estudo sobre o funcionamento destas ferramentas e em testes de desempenho realizados sobre uma nuvem computacional simulada.

Os resultados obtidos neste trabalho permitem concluir que o sistema de arquivos GlusterFS mantém, para arquivos do tamanho do bloco de disco local, taxas de vazão comparáveis ao sistema de arquivos ext3, e apresenta queda de desempenho com o crescimento do tamanho do arquivo, no caso do maior tamanho de arquivo testado, a vazão obtida é próxima a 50% da disponível. O PVFS2, por sua vez, apresentou comportamento constante para todos os tamanhos de arquivos, entregando valores comparáveis a vazão da largura de banda de rede disponível.

Em relação a outros trabalhos desenvolvidos até hoje, não é notado um grande

impacto relacionado a adoção do ambiente virtualizado. Inclusive em alguns estudos são apresentadas taxas de vazão semelhantes as taxas apresentadas neste trabalho.

Por fim, destaca-se que para obtenção de altas taxas de vazão no acesso a arquivos em uma implementação de SAD sobre nuvem computacional, é necessário que os servidores de dados não utilizem redes compartilhadas e que se mantenha largura de banda de rede com valores altos e próximos ao desejado para vazão de acesso aos arquivos. Neste sentido, trabalhos futuros pretendem realizar a avaliação destes SAD sobre um ambiente com múltiplos nós em máquinas físicas distintas e com rede não compartilhada entre eles.

Referências

Apon, A. W., Wolinski, P. D., and Amerson, G. M. (2002). Sensitivity of cluster file system accesses to i/o server selection. *ACM/IEEE International Symposium on Cluster Computing and the Grid*, pages 183–192.

Barbosa, A. A., Greve, F., and Barreto, L. P. (2007). Avaliação de sistemas de arquivos distribuídos num ambiente de pequena escala. *IV Workshop de Sistemas Operacionais*.

Beloglazov, A. and Buyya, R. (2010). Energy efficient resource management in virtualized cloud data centers. In *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, CCGRID '10, pages 826–831, Washington, DC, USA. IEEE Computer Society.

Buyya, R., Garg, S., and Calheiros, R. (2011). SLA-oriented resource provisioning for cloud computing: Challenges, architecture, and solutions. In *Cloud and Service Computing (CSC), 2011 International Conference on*, pages 1–10.

Carns, P., III, W. L., Ross, R., and Thakur, R. (2000). PVFS: A parallel file system for linux clusters. *Proceedings of the 4th Annual Linux Showcase and Conference*, pages 317–327.

Gluster (2011). Performance in a gluster system.

Goldman, A. and de Carvalho, R. P. (2005). Sistemas de arquivos paralelos: Alternativas para a redução do gargalo no acesso ao sistema de arquivos.

Haddad, I. F. (2000). PVFS: A parallel virtual file system for linux clusters. *Linux J.*, 2000.

Kim, J., Kim, I., Kim, T., Eom, Y. I., Kim, H.-Y., and Kim, Y. (2009). Design and implementation of networking virtualization for cluster file systems. pages 79 –83.

Levy, E. and Silberschatz, A. (1990). Distributed file systems: concepts and examples. *ACM Comput. Surv.*, pages 321–374.

Macedo, D. and Azevedo, P. C. B. (2011). Análise comparativa entre soluções livres para construção de sistemas de arquivos ditribuídos. *IX ERRC - Escola Regional de Redes de Computadores*, pages 15–18.

Macedo, D. D. J., Perantunes, H. W. G., Dantas, M. A. R., and Wangenheim, A. V. (2009). An architecture for dicom medical images storage and retrieval adopting distributed file systems. *International Journal of High Performance Systems Architecture*.

Rao, J., Bu, X., Xu, C.-Z., and Wang, K. (2011). A distributed self-learning approach for elastic provisioning of virtualized cloud resources. In *Modeling, Analysis Simulation*

of Computer and Telecommunication Systems (MASCOTS), 2011 IEEE 19th International Symposium on, pages 45–54.

Rimal, B., Choi, E., and Lumb, I. (2009). A taxonomy and survey of cloud computing systems. In *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on*, pages 44–51.

Shmuck, F. and Haskin, R. (2002). GPFS: Shared disk file system for large computing clusters. *2nd USENIX Conference on File and Storage Technologies*.

System, C. F. (2007). Lustre: A scalable, high-performance file system. *Technical report*.

WEIL, S. A., Brandt, S. A., Miller, E. L., Long, D. D. E., and Maltzahn, C. (2006). Ceph: a scalable, high-performance distributed file system.

Xu, Y., Wang, L., Arteaga, D., Zhao, M., Liu, Y., and Figueiredo, R. (2010). Virtualization-based bandwidth management for parallel storage systems. pages 1 – 5.

ZResearch (2005). Glusterfs documentation.