Tolerância a Falhas em Sistemas de Armazenamento de Dados

Márcio Joel Barth[1,2], Edvar Bergmann Araujo[2]

[1]Companhia de Processamento de Dados do RGS – PROCERGS
Praça dos Açorianos, s/n – 90010-340 – Porto Alegre – RS - Brasil

[2]Instituto de Ciências Exatas e Tecnológicas – Centro Universitário Feevale
RS 239, 2755 – 93352-000 – Novo Hamburgo – RS – Brasil

marcio-barth@procergs.rs.gov.br, edvar@feevale.br

Abstract. *The increasing use of computer networks as a way of sharing and accessing information resources have generated great pressure on the current data storage systems. Such systems became limited in their ability for manipulation and recovery of bigger data volumes, which became more and more heterogeneous and distributed in the network environment, arising the need for the implementation of management systems for data storage. These systems' new generations allow information to be gathered together in a single place of the company, thus allowing easier access to its localization, publication, safety, redundancy and administration. This way it is necessary that a storage system is supplied of appropriate techniques of fail tolerance to guarantee the wanted reliability, without the applications or users become aware of the employed techniques.*

Resumo. *A crescente utilização de redes de computadores como forma de compartilhamento e acesso a recursos de informação tem gerado grande pressão sobre os sistemas atuais de armazenamento de dados. Tais sistemas vêm apresentando limitações no que diz respeito à manipulação de grandes volumes de dados, que se tornam cada vez mais heterogêneos e distribuídos nos ambientes de rede, surgindo a necessidade da implantação de sistemas de gerenciamento de armazenamento de dados. As novas gerações desses sistemas permitem que informações sejam reunidas em um único lugar da empresa, facilitando assim a sua localização, publicação, segurança, redundância (backup) e gerenciamento. Desta forma, é necessário que um sistema de armazenamento seja suprido de técnicas de tolerância a falhas adequadas para garantir a confiabilidade desejada, sem que as aplicações ou os usuários tomem conhecimento das técnicas empregadas.*

1. Introdução

Recentemente, novas tecnologias como *Fibre Channel*[1], *Clustering*[2] e *Storage Networking* estão transformando o cenário de armazenamento que está caminhando em dois sentidos simultâneos: (i) recentralização, eliminando-se as ilhas existentes hoje; e (ii) externalização, separando o armazenamento da ligação física ('*bus attached*') com

[1] *Fibre Channel*: tecnologia de rede projetada para altas taxas de transferência entre dispositivos de armazenamento e computadores.
[2] *Clustering*: coleção de computadores que são interconectados (tipicamente em altas velocidades) para o propósito de promover maior disponibilidade de serviços e/ou desempenho (via balanceamento de carga). Geralmente os computadores em *cluster* possuem acesso a uma área de armazenamento comum, e utilizam softwares especiais para coordenar as atividades dos componentes dos computadores.

os servidores. Estas tendências são o que se chama de modelo de computação '*information centric*'.

No modelo "*information centric*", as informações são colocadas no centro do negócio e as plataformas de processamento são conectadas aos equipamentos de armazenamento. O modelo transcende plataformas e ambientes operacionais, tendo como objetivo a integração de todas as informações, fornecendo uma visão simples e única. É uma mudança significativa se comparado ao modelo '*server centric*' atual, onde o processador é a peça chave da computação e o limitador da capacidade de acesso às informações [HDS, 2003].

2. Tolerância a falhas em sistemas de armazenamento

Cada vez mais as empresas utilizam sistemas de armazenamento a fim de garantirem maior disponibilidade e confiabilidade das informações, em muitos casos configurando-se em modelos de armazenamento do tipo DAS, NAS ou SAN. Para garantir o perfeito funcionamento destes sistemas torna-se necessário que os subsistemas de armazenamento possuam a capacidade de continuar funcionando mesmo quando ocorrer falha em um disco ou outro componente (possivelmente com redução do nível de desempenho). A seguir são comentados vários aspectos referentes à tolerância a falhas em subsistemas de armazenamento que devem ser considerados.

2.1 RAID

O sistema RAID (*Redundant Array of Independent Disks*) – alguns fabricantes utilizam o termo "*Inexpensive*" no lugar de "*Independent*" - surgiu em 1987 pelos pesquisadores Patterson, Gibson e Katz, da Universidade da Califórnia, Berkeley. É um método que combina vários discos em uma única unidade lógica. Um *disk array* RAID oferece tolerância a falhas e melhores taxas de transferência do que um *drive* único ou um grupo de *drives* independente.

A configuração de um RAID pode ser realizada através do próprio sistema operacional (*software*), caso o mesmo ofereça o serviço, ou pela controladora (*hardware*), que neste caso, é o modo mais aconselhável por oferecer maior desempenho e liberar o sistema operacional desta tarefa [MOURA, 2001].

O RAID constitui-se a base para todas as funcionalidades esperadas num sistema de armazenamento em termos de proteção dos dados, tolerância a falhas, altos níveis de desempenho, grande capacidade de armazenamento e escalabilidade.

A implementação de um sistema RAID é possível utilizando-se as controladoras SCSI e IDE, que permitem a conexão e a configuração de vários discos a fim de obter-se as vantagens que o RAID proporciona.

Podem ser citadas três características fundamentais dos RAID's [SNIA,2003]:

- Vários discos acessados em paralelo fornecem uma taxa de I/O superior a de um único disco;
- O armazenamento de dados de modo redundante em vários discos oferece melhor tolerância à falha;

- O uso da tecnologia "*Hot-plug*" possibilita trocar um dispositivo com falha sem que o servidor tenha que ser desligado.

Há diferentes números que especificam o nível de segurança implementada em um produto RAID. Os níveis mais comuns são 0, 1 e 5 e estão entre os mais utilizados pelas corporações. A sua escolha deve ser determinada de acordo com as necessidades das aplicações que executam na empresa.

2.2 Falhas de interconexão

Técnicas de espelhamento e configuração RAID de discos protegem contra falhas em discos, mas os discos não são o único ponto que pode falhar. O barramento ou conexão de fibra SAN que conectam os dispositivos de armazenamento aos servidores, também pode falhar, tornando os dados inacessíveis.

O volume RAID na figura 1A possui uma interconexão entre cada disco e a controladora RAID do servidor. Se uma conexão falhar, somente um disco ficará inacessível e o volume poderá continuar funcionando através da reconstrução dos dados.

No volume da figura 1B, ao contrário, os discos compartilham a mesma conexão. Se a conexão falhar, mais de um disco poderá ficar inacessível e a reconstrução do volume poderá ficar comprometida. A conclusão é que o máximo de cuidado na configuração dos volumes é necessária para um nível máximo de tolerância a falhas [BARKER, 2002].

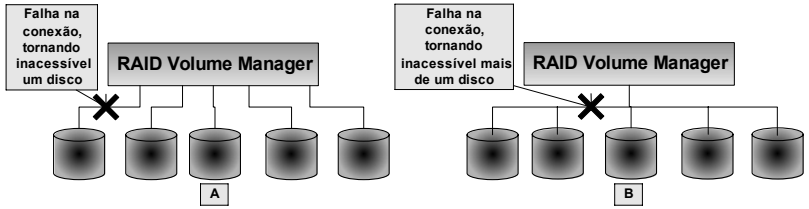

Figura 1- Interconexão individual de discos implementando tolerância a falha.

2.3 Falha nas controladoras RAID

Controladoras RAID são projetadas para garantir tolerância a falhas, embora a mesma possa falhar causando um impacto sobre a disponibilidade dos dados. Assim o controle para tolerância a falhas é tipicamente ativado com duas ou mais controladoras conectadas ao mesmo disco e servidor. Na figura 2, todos os discos estão conectados por duas controladoras RAID externas, que trocam mensagens entre si assegurando que ambas estão ativas. Caso uma controladora deixe de receber mensagens da outra, ela compreende que deve assumir a comunicação dos discos da controladora que falhou para não comprometer a disponibilidade dos dados [BARKER, 2002].

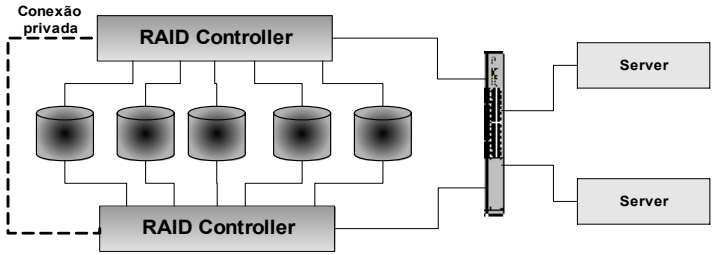

Figura 2 – Tolerância a falhas em controladoras RAID externas.

2.4 Tolerância a falhas transparentes e não transparentes

Mesmo que ambas as controladoras da figura 2, utilizem a mesma infra-estrutura para conectarem-se ao servidor, os volumes presentes devem necessariamente possuir diferentes endereçamentos (por exemplo, *Fibre Channel LUNs* (*Logical Unit Number*)). Quando uma controladora falha, a outra toma conta dos discos e apresenta seus volumes com o mesmo endereçamento utilizado pela controladora que falhou. Com a menor possibilidade de erro durante a transição, a troca de controladora é transparente para a aplicação. Os mesmos volumes são endereçados para o mesmo endereço, isso, é claro, se a outra controladora estiver gerenciando tudo.

Isto é uma solução elegante, mas ainda possui um ponto de falha, se a infra-estrutura da rede SAN falhar, então todo o acesso aos dados será perdido. Uma solução para este problema é conectar cada controladora RAID em um segmento SAN separado.

Com um subsistema de entrada/saída configurado desta forma, se um dos segmentos da SAN falhar, todos os dados permanecerão acessíveis, porque se a conexão até o servidor continuar funcionando o controle de todos os discos estará ativo pela outra controladora.

Alguns subsistemas equipam cada controladora RAID com duas ou mais conexões, como ilustrado na figura 3. Assim não somente servidores, mas também as controladoras RAID poderão estar conectadas aos dois segmentos da SAN. Esta configuração elimina a necessidade de substituir todos os discos para uma controladora se um segmento da SAN falhar [BARKER, 2002].

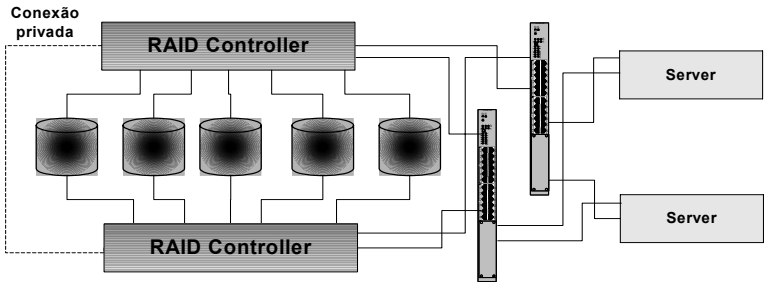

Figura 3 – Interconexão de controladoras RAID completamente redundantes.

2.5 Operações atômicas e integridade de dados

Muitas controladoras RAID possuem *cache* reverso para aumentar o desempenho de entrada/saída. Se uma controladora RAID ficar sem alimentação de energia e com uma atualização pela metade (por exemplo, uma cópia do espelhamento dos dados completa e a outra não), o volume pode retornar com dados corrompidos em algum momento no futuro. Por exemplo, se a energia falhar e somente uma cópia de um bloco de dados de um espelhamento estiver completa, uma leitura futura daquele bloco poderá retornar com resultados diferentes dependendo de qual disco espelhado for selecionado para satisfazer a solicitação de leitura.

Se a controladora RAID for realmente tolerante a falhas, deve proteger contra a perda de estado e conteúdo de *cache* quando faltar energia (ou quando a controladora falhar). A solução mais utilizada para este problema para as controladoras RAID interconectadas é a comunicação de mudança de estado e a respectiva mudança no conteúdo de sua *cache* reversa entre as controladoras. Independente da tecnologia implementada, o mais importante para os usuários de controladoras RAID é que os estados das operações e os dados de *cache* sejam preservados no caso da controladora falhar, não havendo assim, perda de dados [BARKER, 2002].

2.6 Replicação de sistema de armazenamento

Com a replicação dos dispositivos de armazenamento, todos os blocos gravados no sistema de armazenamento primário são replicados para dispositivos de igual capacidade para cada local secundário, sem considerar o significado dos blocos de dados replicados. Assim aumenta-se a tolerância a falhas através de um *site* para recuperação de desastres, que é um *datacenter* secundário localizado distante o suficiente para continuar as operações se o *site* primário sofrer um desastre irrecuperável [MOURA,2002].

Muitas tecnologias de replicação permitem ao administrador de sistema escolher entre duas opções:

Replicação síncrona: para cada atualização da aplicação gravada no *storage* principal é esperada uma validação do *storage* secundário antes que seja considerada completa a transação.

Replicação assíncrona: com o *storage* secundário podendo ficar defasado em relação ao *storage* principal, o mesmo grava as informações em tempos pré-determinados.

A replicação síncrona simplifica a conversão de dados entre um *storage* secundário em relação ao *storage* principal após um desastre, porque muito poucos dados estão no canal de transmissão. Mas a replicação síncrona de acordo com a forma de conexão utilizada pode afetar as aplicações adversamente, pois ficam esperando pelo *storage* secundário por longos tempos de resposta para completar uma transação.

A replicação assíncrona essencialmente elimina o tráfego da rede e o desempenho do *storage* secundário e dos tempos de resposta da aplicação. Embora sem atualização de transação executa a recuperação dos dados depois de um desastre de forma mais complexa do que a síncrona, pois as réplicas do *storage* secundário podem estar sensivelmente desatualizadas.

3 Conclusão

A grande dúvida das corporações sempre paira sobre a real necessidade e aplicabilidade de soluções tolerantes a falhas, pois além do alto custo sobre as soluções de armazenamento, a eficácia de implantação desses sistemas não pode ser mensurada. Sem dúvida, com a crescente demanda por recursos de armazenamento que devem estar disponíveis 24xForever, estes métodos de tolerância a falhas serão cada vez mais necessários para a disponibilidade e confiabilidade de qualquer sistema.

Desta forma, buscou-se a comprovação da aplicabilidade dos métodos citados no artigo, que se baseiam no método de *fail-stop* (são acionados somente quando existe a falha), através de um estudo de caso em duas empresas de grande porte que possuem equipamentos que já implementam parte das técnicas de tolerância a falhas. Com os testes realizados constatou-se que neste contexto a aplicabilidade dos métodos só é possível utilizando-se equipamentos de mesmo fornecedor, modelo e capacidade. Sendo assim, outros métodos de tolerância a falhas deverão surgir a fim de suprir outras necessidades como, por exemplo, garantir tolerância a falhas entre dispositivos de tamanhos e fornecedores diferentes.

4 Referências

BARKER, Richard & MASSIGLIA, Paul. Storage area networks essentials: a complete guide to understanding and implementing SANs. 2.ed. – New York: John Wiley & Sons, 2002. 498 p.

High Digital Storage. Pensando em armazenamento. Disponível em: http://www.hdsinfo.com.br/inpdfs.htm. Acessado em: agosto 2003.

MOURA, Giedre. Storage quer romper as barreiras. .**Network**, São Paulo: v. 3, n. 44, p. 28 – 34, out. 2002.

MOURA, Giedre. Todo o poder para o RAID. Network Computing. Disponível em: http://www.networkcomputing.com.br/noticias/artigo.asp?id=18674. Publicado em: 13 nov. 2001. Acessado em: agosto 2003.

Storage Network Industry Association. Disponível em: http://www.snia.org. Acessado em: agosto 2003.