4.3.6 Sistemas de arquivos *journaling*

Embora os sistemas de arquivos estruturados com base em log sejam uma ideia interessante, eles não são largamente utilizados, em parte por conta da alta incompatibilidade com outros sistemas de arquivos existentes. Ainda assim, uma das ideias presentes nos LFSs, a de robustez diante da falha, pode ser facilmente aplicada em sistemas de arquivos mais convencionais. A premissa básica é a de manter um registro sobre o que o sistema de arquivos irá fazer antes que ele efetivamente o faça, de modo que, se o sistema falhar antes da execução do trabalho planejado, é possível, após a reinicialização do sistema, recorrer ao log para descobrir o que estava acontecendo no momento da parada e retomar o trabalho. Esse tipo de sistema de arquivos, denominado **sistemas de arquivos *journaling***, já está em uso: o sistema NTFS, da Microsoft, e os ext3 e ReiserFS, do Linux, são do tipo *journaling*. Faremos a seguir uma breve descrição do assunto.

Para compreender o problema, imagine uma operação corriqueira que acontece a todo instante: a remoção de um arquivo. No UNIX, essa operação é realizada em três etapas:

1. Remova o arquivo de seu diretório.
2. Libere o i-node para o conjunto de i-nodes livres.
3. Volte todos os blocos do disco para o conjunto de blocos livres no disco.

Desempenho, Redundância e Proteção de Dados

No final da década de 1980, pesquisadores da Universidade da Califórnia em Berkeley desenvolveram técnicas de gerenciamento de discos que otimizavam as operações de E/S e implementavam redundância e proteção de dados conhecidas como *RAID* (Redundant Arrays of Inexpensive Disk). As diferentes técnicas, utilizando múltiplos discos, foram publicadas em seis níveis (RAID 1-6). Estas técnicas tiveram grande aceitação no mercado e, posteriormente, um novo nível foi introduzido e denominado RAID 0.

As técnicas de RAID podem ser implementadas diretamente nos controladores de discos, conhecido como *subsistema RAID externo*, ou por software através do sistema operacional ou um produto gerenciador de discos, denominado *subsistema JBOD* (just a bunch of disks). A Fig. 12.10 ilustra a diferença entre as duas possíveis formas de implantação das técnicas de RAID.

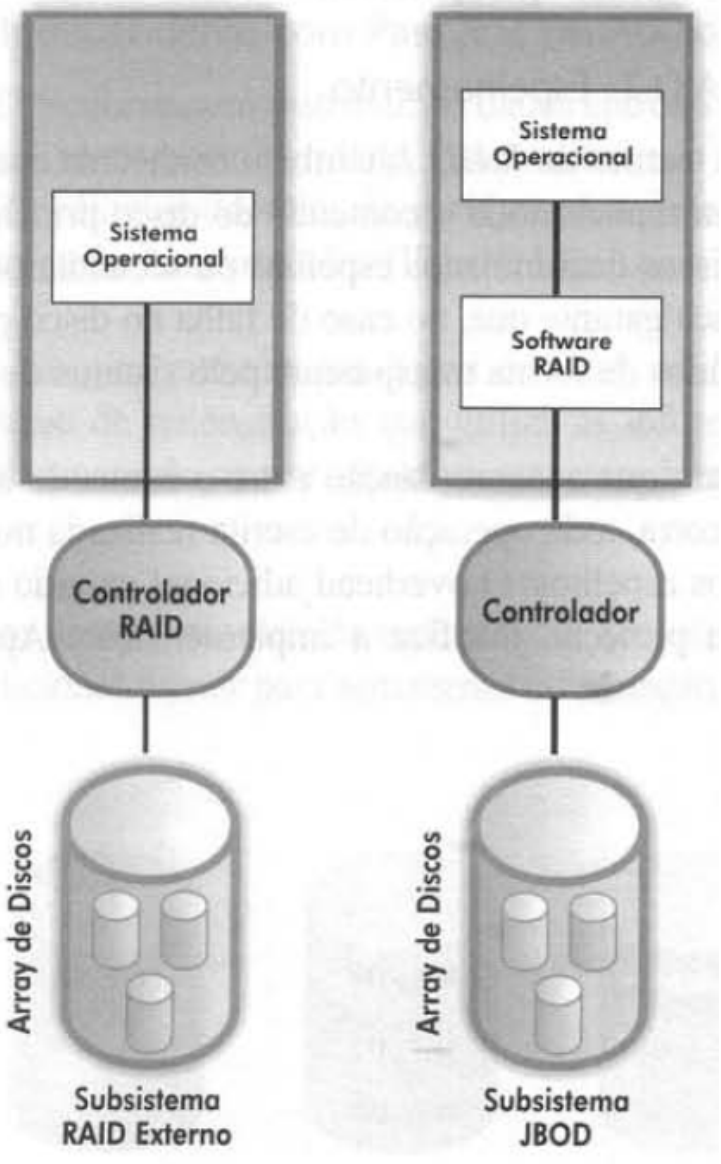

Fig. 12.10 Subsistema de discos.

- RAID 0: Striping

A técnica de *RAID 0*, também conhecida como *striping*, consiste na implementação do chamado disk striping, que é distribuir as operações de E/S entre os diversos discos físicos contidos no array com o intuito de otimizar o desempenho. Como os dados são divididos entre os diversos discos, as operações de E/S podem ser processadas paralelamente.

Para poder implementar o striping é preciso formar um conjunto de discos chamado de *stripe set*, onde cada disco é dividido em pedaços (stripes). Sempre que um arquivo é gravado, seus dados são divididos em pedaços iguais e espalhados simultaneamente pelos stripes dos diversos discos (Fig. 12.11).

Apesar da denominação RAID, esta técnica não implementa qualquer tipo de redundância, só sendo vantajosa no ganho de desempenho das operações de E/S. Caso um haja uma falha em qualquer disco do stripe set, os dados serão perdidos. Aplicações multimídia são beneficiadas com o uso desta técnica pois necessitam de alto desempenho nas operações com discos.

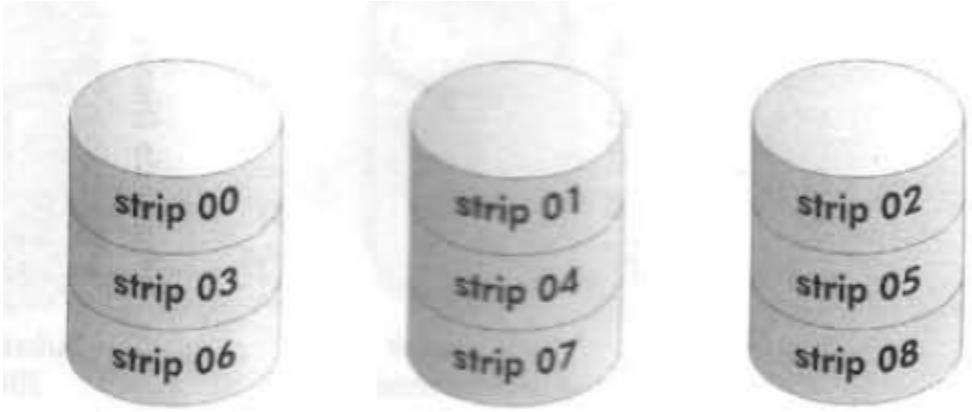

Fig. 12.11 RAID 0.

- RAID 1: Espelhamento

A técnica de *RAID 1*, também conhecida como *espelhamento* ou *mirroring*, consiste em replicar todo o conteúdo do disco principal, chamado primário, em um ou mais discos denominados espelhos ou secundários. A redundância oferecida por essa técnica garante que, no caso de falha no disco principal, os discos espelhos sejam utilizados de forma transparente pelo sistema de arquivos (Fig. 12.12).

Para que a sincronização entre o conteúdo do disco principal e dos discos espelhos ocorra, toda operação de escrita realizada no disco primário é replicada para os discos espelhos. O overhead adicional exigido nesta operação é pequeno e o benefício da proteção justifica a implementação. Apesar da vantagem proporcionada pela redundância oferecida por esta técnica, a capacidade útil do subsistema de discos com a implementação do RAID 1 é de apenas 50%.

A técnica de RAID 1 pode ser implementada por software em um subsistema JBOD ou por hardware diretamente pelo controlador de disco em um subsistema RAID externo. A implementação por software necessita que o sistema operacional ou algum produto gerenciador de discos ofereça esta facilidade.

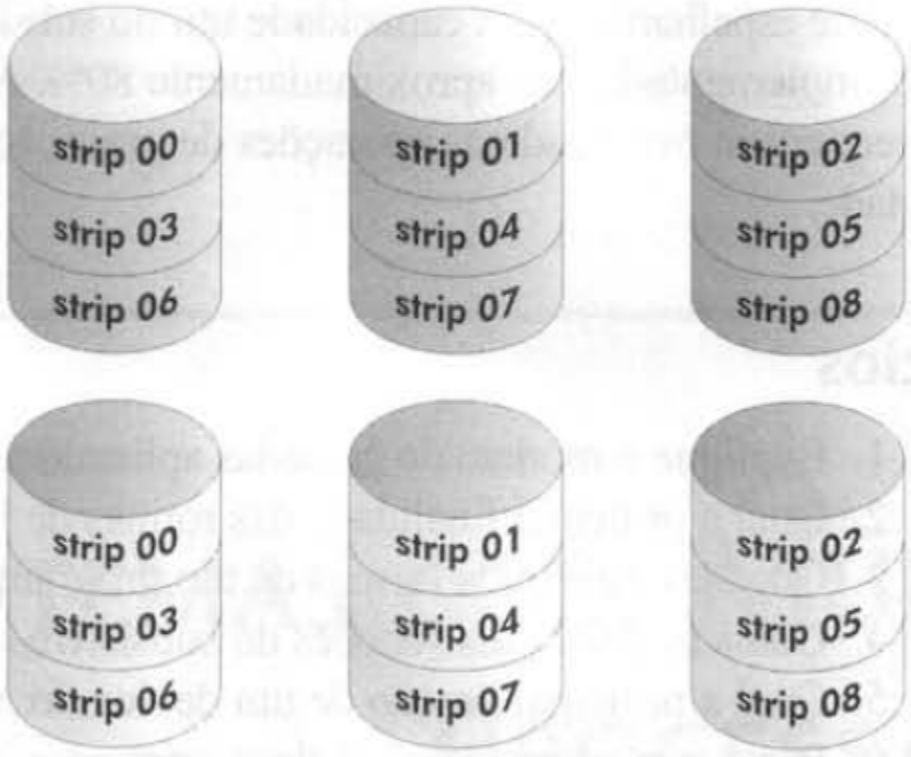

Fig. 12.12 RAID 1.

- RAID 5: Acesso Independente com Paridade Distribuída

A técnica de *RAID 5* consiste em distribuir os dados entre os discos do array e implementar redundância baseada em paridade. Este mecanismo de redundância é implementado através de cálculos do valor da paridade dos dados, que são armazenados nos discos do array junto com os dados (Fig. 12.13).

Fig. 12.13 RAID 5.

Caso haja uma falha em qualquer um dos discos do array, os dados podem ser recuperados por um algoritmo de reconstrução que utiliza as informações de paridade dos demais discos. Esta recuperação ocorre automaticamente e é transparente ao sistema de arquivos.

A principal vantagem de uma técnica de redundância que utiliza paridade é que esta requer um espaço adicional menor para armazenar informação de controle que a técnica de espelhamento. A capacidade útil do subsistema de discos com a técnica de RAID 5 implementada é de aproximadamente 80%. Apesar disso, esta técnica de redundância requer um overhead nas operações de gravação no disco em função do cálculo da paridade.