Linux Network Servers

RAID

Objetivo: Entender os principais níveis de RAID, configurar RAID-1, verificar o estado do RAID, simular falhas no RAID.

Um servidor deve sofrer muitas manutenções durante seu período de uso? Não! Um servidor bem configurado e otimizado deve funcionar muito bem durante anos com poucas intervenções. Um servidor deve ser bem mais estável e confiável do que um desktop.

Um servidor pode fornecer vários serviços como DNS, HTTP, proxy, e-mail, banco de dados etc para um número de usuários grande ao mesmo tempo. Mas o que torna um servidor seguro?
O que torna um servidor seguro é o uso de componentes redundantes!

Quais são os problemas comuns que ocorre em um servidor? Problemas nos discos rígidos e fontes de alimentação. Os discos rígidos são confiáveis? Não são! Discos rígidos possuem partes mecânicas e estas são extremamente sujeitas a falhas.

Qual a vida média útil de um HD? Em torno de 5 anos. Você pode até dar sorte de funcionar por mais algum tempo bem. Depende da marca e dos seus cuidados com o disco.

Temos um servidor, e ele obviamente tem discos rígidos, porém o disco pifa. Um HD IDE é recomendável para um servidor? Não! Os HDs recomendados são os do tipo SCSI ou Sata. O Sata é uma opção mais barata!

O HD IDE não oferece suporte a NCQ. Mas o que é NCQ?
NCQ é um recurso que permite que a controladora do HD altere a sequência das leituras a fim de otimizar o processo.

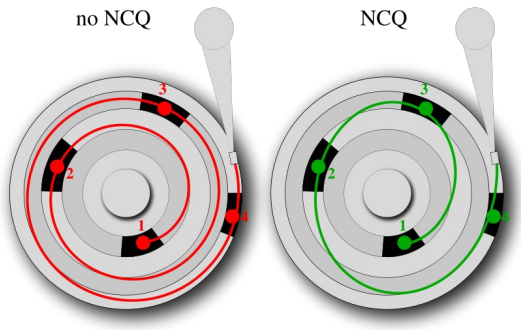

Ilustração 1: Modos de leitura de um HD sem NCQ e um com NCQ

O que poderia ser feito para garantir que esse servidor continue operando caso um disco pife?
Usar RAID!

Linux Network Servers

O RAID (Redundant Array of Inexpensive Disks) foi desenvolvido em 1988 como uma solução barata para garantir a disponibilidade da informação armazenada em discos, utilizando para isso uma configuração especial de discos rígidos, que podem oferecer redundância em caso de falhas e ganho de performance em escrita ou leitura, dependendo da configuração do conjunto RAID. O que significa redundância?

Redundância é ter componentes de reserva para substituir o componente principal mantendo disponibilidade de serviços. Existem fontes redundantes, matriz de discos redundantes, servidores redundantes. Manter redundância requer um custo!

Mas como funciona o RAID?
Os discos são agregados no que chamamos de níveis. Cada nível de agregação dos discos oferece vantagens e desvantagens. Quais são esses níveis? Os principais níveis de RAID utilizados hoje no mercado são os níveis 0, 1 e 5.

Como funciona um RAID 0?
No RAID 0 (stripping), vários discos são vistos como se fossem um só disco.

Os arquivos ficam fragmentados em vários discos, e com isso faço com que a leitura/gravação seja feita de forma simultânea, com isso, consegue-se uma taxa de leitura e gravação de dados.

No RAID 0, eu consigo usar todo o espaço do disco?
Sim. 100% do espaço em disco para dados. Então, se você tiver 2 discos de 80GB, você vai ter uma área útil de 160GB.

Ex. 4 Hds de 80GB = 1 de 320GB de área útil.

Qual a vantagem do RAID 0?
Ganho de desempenho, já que a leitura/gravação é feita de forma simultânea.

RAID 0 garante redundância?
Não. Este tipo de implementação vai utilizar o máximo disponível de discos no sistema, mas não vai te garantir redundância e de fato vai aumentar as chances do disco virtual falhar! Se um disco pifar, a controladora não consegue reconstruir os arquivos!

RAID0:

* Excelente gravação e leitura;
* Aproveitamento total de espaço;
* Nenhuma redundância;

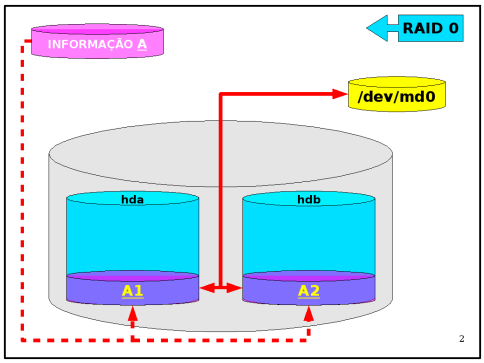

Ilustração 2: RAID 0

Linux Network Servers

Como funciona um RAID 1?

RAID 1 é conhecido como espelhamento, pois a ideia é espelhar as informações em um segundo disco. O sistema vai gravar os dados ao mesmo tempo nos dois discos. Implementar RAID 1 protege os dados, pois caso um dos discos falhe, o sistema continua funcionando normalmente.

O uso do RAID um necessita de dois discos ou qualquer número par, pois como já foi dito acima um para o sistema normal e outro para espelhar o primeiro. Vamos dizer um "HD clone". A desvantagem do RAID 1 é o custo, pois você vai ter dois Discos e a área útil de apenas 1.

Ex. Dois Hds de 80GB em RAID 1 = 80GB de área útil.

RAID1:

* Redundância, se um dos discos falhar o sistema continua funcionando.
* Você vai precisar de 2 Hds, mas só vai usar a área útil de um.
* Reduz um pouco o desempenho da escrita, pois o mesmo dado é gravado nos discos que estiverem em RAID 1.

RAID1 é backup?
Não!

DICA DE SEGURANÇA: RAID1 espelhado não é backup! Se você apagar um arquivo acidentalmente esse arquivo vai ser apagado em todos os discos! Sempre tenha uma CÓPIA dos dados.

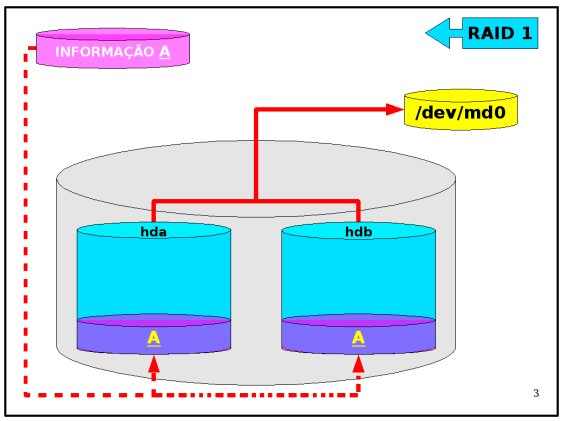

Ilustração 3: RAID 1

Linux Network Servers

E como funciona o RAID 5?

Este é o modo mais utilizado em servidores com um grande número de HDs. O RAID 5 usa um sistema de paridade para manter a integridade dos dados.
Os arquivos são divididos em fragmentos e, para cada grupo de fragmentos, é gerado um fragmento adicional, contendo códigos de paridade. Os códigos de correção são espalhados entre os discos.

Dessa forma, é possível gravar dados simultaneamente em todos os HDs, melhorando o desempenho.
O RAID 5 pode ser usado com a partir de 3 discos.

Independentemente da quantidade de discos usados, sempre temos sacrificado o espaço equivalente a um deles. Com 4 HDs de 1 TB, por exemplo, você ficaria com 3 TB de espaço disponível.

Os dados continuam seguros caso qualquer um dos HDs usados falhe, mas se um segundo HD falhar antes que o primeiro seja substituído (ou antes que a controladora tenha tempo de regravar os dados), todos os dados são perdidos. Você pode pensar no RAID 5 como um RAID 0 com uma camada de redundância.

Se você tiver seis discos em RAID 5 sua área útil será de 6-1

Ex. 6 discos de 80GB sua área útil será 80GB (6-1) = 80GB x 5 = 400GB

Numa composição de três discos os dados serão divididos em dois blocos, A1 e B1, sendo que os bits destes dois blocos serão comparado através de um XOR ("ou exclusivo").

O resultado será gravado no terceiro volume como P1. Além disso, os blocos de paridade são alternadamente gravados em cada disco, aumentando a tolerância. Exemplo:

A1 B1 P1
A2 P2 C2
P3 B3 C3

Qualquer um dos discos que falhar pode ser rapidamente reconstruído através de novas operações XOR entre os dados restantes. Tomemos por exemplo A1 = 01001100 e B1 = 10100101.

01001100 XOR
10100101

11101001

Agora suponha que o bloco B1 foi perdido. Para recuperá-lo basta aplicar um XOR entre A1 e P1.

01001100 XOR
11101001

10100101

A operação XOR significa que se são iguais "0 e 0" ou "1 e 1", então o resultado é falso (0).

Linux Network Servers

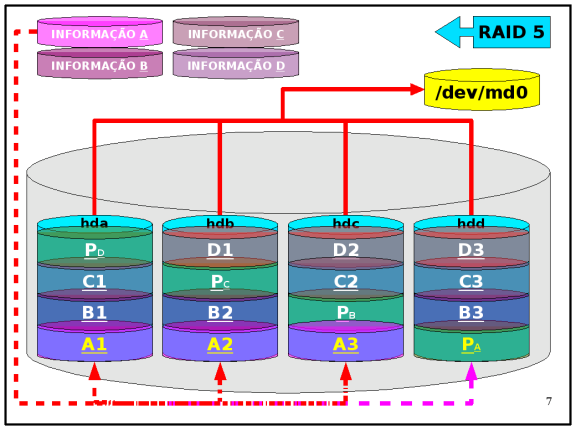

Ilustração 4: RAID 5

Prática!

Trabalhando com RAID - O objetivo deste exercício é criar RAID no mesmo HD, sendo assim possível fazer com HDs separados, uma vez que o Linux enxerga as partições como se fossem HDs diferentes, ou seja, hda1 é tratado como um dispositivo independente dos demais.

Se fôssemos construir um RAID sem planejamento você teria que ter espaço em disco ou um HD para espelhar os dados. No nosso caso, fizemos uma instalação personalizada e deixamos algumas partições para usar no RAID:

```
# fdisk /dev/hda
ou
# cfdisk /dev/hda
```

Verifique se estão instalados na sua distribuição o pacote mdadm:

```
a) via RPM (padrão RedHat)
# rpm -qa | grep mdadm

b) via dpkg (padrão)
# dpkg -l | grep mdadm

c) via aptitude
# aptitude install mdadm
```

Linux Network Servers

Na criação do RAID, o sistema vai sincronizar os discos. Vamos ativar a saída do arquivo de controle do kernel para um outro terminal que não está ativo para que possamos acompanhar a sincronização via arquivo /proc/mdstat. Para sair do comando a seguir basta digitar (CTRL+C):

```
# watch cat /proc/mdstat
```

Criando o RAID

Vamos criar o nosso RAID utilizando o nível 1:

```
# mdadm --create --verbose /dev/md0 --level=1 --raid-devices=2 /dev/hda9
/dev/hda10
```

Ele irá criar o nosso RAID e sincronizar os HDs. Vamos visualizar a sincronização no terminal F12 onde nós jogamos a saída do /proc/mdstat.
Depois de criarmos o nosso RAID, devemos editar o arquivo /etc/mdadm/mdadm.conf, que é usado para facilitar o manuseio e administração do nosso RAID:

```
# vi /etc/mdadm/mdadm.conf
DEVICE          /dev/hda9 /dev/hda10
ARRAY           /dev/md0 devices=/dev/hda9,/dev/hda10
```

Crie o sistema de arquivos (ext3):

```
# mke2fs -j /dev/md0
```

OBS: Seria o mesmo que fazer mkfs.ext3 /dev/md0

Iremos criar um ponto de montagem chamado /dados:

```
# mkdir /dados
```

OBS: Esse arquivo será utilizado para facilitar na hora de inicializar o nosso RAID, afim de não precisarmos criar parâmetros dos dispositivos.

Montando o RAID:

```
# mount -t ext3 /dev/md0 /dados
```

Linux Network Servers

Caso esteja utilizando RedHat (ou uma distro que siga o mesmo padrão), verifique a necessidade de criação de um label para o device. Se for necessário, faça-o:

```
# e2label /dev/md0 /dados
```

Adicione estas linhas no /etc/fstab:
Com label (Red Hat):

```
LABEL=/dev/md0  /dados   ext3   defaults   0  2
```

Sem label (Debian):

```
/dev/md0  /dados   ext3   defaults   0  2
```

Linux Network Servers

Troubleshooting

Verificando os dispositivos individualmente:

```
# mdadm -E /dev/hda9
```

/dev/hda9:
Magic : a92b4efc
Version : 00.90.00
UUID : 9bfe16a7:92778849:67d8fcec:2e7fc84b (local to host workaholic)
Creation Time : Tue Aug 18 23:05:14 2009
Raid Level : raid1 # Nível do RAID
Used Dev Size : 14643136 (13.96 GiB 14.99 GB)
Array Size : 14643136 (13.96 GiB 14.99 GB) # Tamanho da matriz
Raid Devices : 2 # Quantidade de dispositivos
Total Devices : 2 # Quantidade de dispositivos
Preferred Minor : 0

Update Time : Tue Aug 18 23:12:58 2009
State : active # Estado ativo
Active Devices : 2 # Quantidade de dispositivos ativos
Working Devices : 2 # Quantidade de dispositivos trabalhando
Failed Devices : 0 # Quantidade de dispositivos falhos
Spare Devices : 0
Checksum : 3efe10c - correct
Events : 3

Number	Major	Minor	RaidDevice	State this	
0	8	5	0	active sync	/dev/hda9
0	0	8	5	0 active sync	/dev/hda9
1	1	8	6	1 active sync	/dev/hda10

```
# mdadm -E /dev/hda10
```

Linux Network Servers

Verificando o relatório completo do status:

```
# mdadm --detail --scan
```

ARRAY **/dev/md0** level=**raid1** num-devices=**2** metadata=00.90 UUID=9bfe16a7:92778849:67d8fcec:2e7fc84b

```
# mdadm --detail /dev/md0
```

```
/dev/md0:
        Version : 00.90
  Creation Time : Tue Aug 18 23:05:14 2009
     Raid Level : raid1
     Array Size : 14643136 (13.96 GiB 14.99 GB)
  Used Dev Size : 14643136 (13.96 GiB 14.99 GB)
   Raid Devices : 2
  Total Devices : 2
Preferred Minor : 0
    Persistence : Superblock is persistent

    Update Time : Tue Aug 18 23:12:58 2009
          State : active, resyncing
Active Devices : 2
Working Devices : 2
Failed Devices : 0
  Spare Devices : 0

Rebuild Status : 21% complete

           UUID : 9bfe16a7:92778849:67d8fcec:2e7fc84b (local to host workaholic)
         Events : 0.3

    Number   Major   Minor   RaidDevice State
       0       8        5        0      active sync   /dev/hda9 # Disco ativo e síncrono
       1       8        6        1      active sync   /dev/hda10 # Disco ativo e síncrono
```

Linux Network Servers

Fazer alguns testes e, caso seja necessário, adicionar um disco ou removê-los simulando uma falha:

a) Simulando falha no HD:

Provavelmente, seu sistema continuará funcionando, sendo assim, é possível fazer a troca.

```
# mdadm /dev/md0 --remove /dev/hda9
```

mdadm: set /dev/hda9 faulty in /dev/md0 **(mensagem dizendo que o /dev/hda9 falhou)**

```
# mdadm --detail /dev/md0
```

```
/dev/md0:
        Version : 00.90
  Creation Time : Tue Aug 18 23:05:14 2009
     Raid Level : raid1
     Array Size : 14643136 (13.96 GiB 14.99 GB)
  Used Dev Size : 14643136 (13.96 GiB 14.99 GB)
   Raid Devices : 2
  Total Devices : 2
Preferred Minor : 0
    Persistence : Superblock is persistent

    Update Time : Tue Aug 18 23:17:11 2009
          State : active, degraded # estado que merece atenção, trocar o disco defeituoso
Active Devices : 1
Working Devices : 1
Failed Devices : 1
  Spare Devices : 0

           UUID : 9bfe16a7:92778849:67d8fcec:2e7fc84b (local to host workaholic)
         Events : 0.13

    Number   Major   Minor   RaidDevice State
       0       0        0        0      removed
       1       8        6        1      active sync   /dev/hda10 # Disco ativo

       2       8        5        -      faulty spare   /dev/hda9 # Disco falho
```

Linux Network Servers

b) Voltando o HD:

```
# mdadm /dev/md0 --add /dev/hda9
```

mdadm: re-added /dev/sda5 **(disco adicionado novamente)**

```
# mdadm --detail /dev/md0
```

```
/dev/md0:
        Version : 00.90
  Creation Time : Tue Aug 18 23:05:14 2009
     Raid Level : raid1
     Array Size : 14643136 (13.96 GiB 14.99 GB)
  Used Dev Size : 14643136 (13.96 GiB 14.99 GB)
   Raid Devices : 2
  Total Devices : 2
Preferred Minor : 0
    Persistence : Superblock is persistent

    Update Time : Tue Aug 18 23:19:15 2009
          State : active, degraded, recovering
Active Devices : 1
Working Devices : 2
Failed Devices : 0
  Spare Devices : 1

Rebuild Status : 1% complete # Status da reconstrução da matriz em porcentagem

           UUID : 9bfe16a7:92778849:67d8fcec:2e7fc84b (local to host workaholic)
         Events : 0.17

    Number   Major   Minor   RaidDevice State
       2       8        5        0      spare rebuilding   /dev/hda9 # Reconstruindo
       1       8        6        1      active sync   /dev/hda10 # Ativo
```

Linux Network Servers

Parando o RAID:

OBS: Para parar o RAID é necessário desmontar o ponto de montagem que está sendo utilizado.

```
# umount /dados
```

```
# mdadm -S /dev/md0
```

mdadm: stopped /dev/md0 **(RAID desativado)**

Voltando o RAID:

```
# mdadm -As /dev/md0
```

mdadm: /dev/md0 has been started with 1 drive (out of 2) and 1 spare. **(RAID ativado novamente)**

DICA LPI: Esteja familiarizado com o comando mdadm e o arquivo de configuração mdadm.conf, eles caem na prova.

Como principais vantagens, o RAID oferece:

* Ganho de desempenho no acesso para leitura ou gravação.
* Redundância em caso de falha em um dos discos.
* Uso múltiplo de varias unidades de discos.
* Facilidade em recuperação de conteúdo perdido.

Foi feito um RAID via software ou hardware?
Software!

E quais as diferenças entre eles?

Via Software: Feito por aplicativos e módulos do sistema operacional, o RAID via software só entra em funcionamento depois que o Kernel é carregado na memória do computador.

A principal vantagem é a facilidade de configuração e a flexibilidade, já que podemos trabalhar com vários discos diferentes. A principal desvantagem é a dependência da correta configuração do sistema operacional.

Via Hardware: Feito por uma placa controladora que conecta um disco ao outro. A principal vantagem é o desempenho, já que um RAID via hardware é mais rápido e independe do sistema operacional.

A principal desvantagem é que a placa controladora se torna um SPOF (SinglePoint of Failure), ou seja, é necessário ter uma controladora de discos igual ou compatível com a que você possui para o caso de falhas neste hardware. Outra desvantagem é o custo!