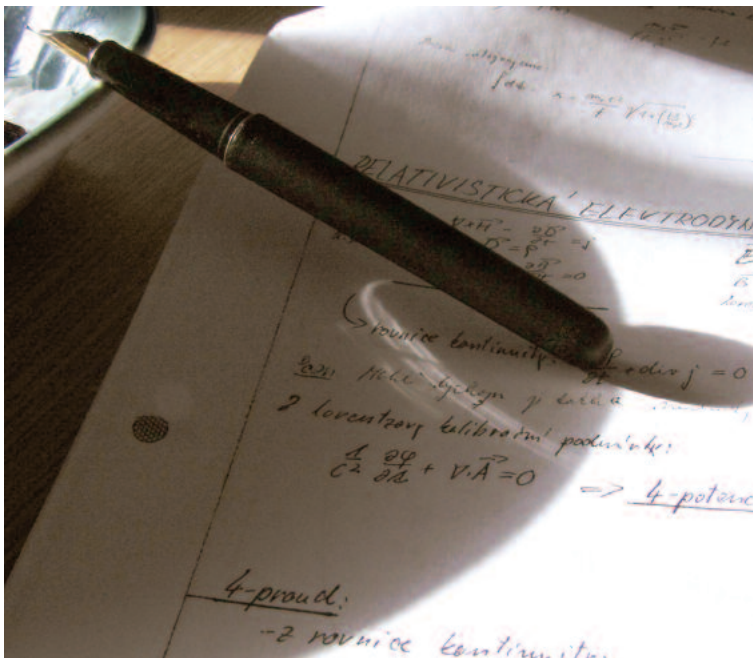
Ext 2/3, JFS, Reiser e XFS, quem ganha essa?

Provão!



Testes de desempenho de sistemas de arquivos são provas de fogo: servem para medir a velocidade de leitura e escrita de informações. A Linux Magazine colocou a performance do Ext 2/3, JFS, Reiserfs e XFS à prova.

POR MIRKO DÖLLE E JÖRG REITTER

Testes de desempenho são importantes indicadores de performance. Quando aplicados a sistemas de arquivos, eles indicam como o sistema reage a certas operações como leitura e escrita e o quanto é exigido do processador. Quem, por exemplo, armazena muitos arquivos pequenos em um diretório, necessita de um sistema de arquivos que vasculhe rapidamente o diretório e que se fragmente muito lentamente. Por outro lado, administradores de bancos de dados estão mais interessados em suporte a grandes arquivos e esperam uma excelente performance de leitura e escrita.

A Linux Magazine examinou o comportamento dos sistemas de arquivos Ext 2, Ext 3, JFS, ReiserFS e XFS com o conjunto de testes de desempenho Fsbench [1], que se compõe essencialmente de um script escrito em Python que chama os verdadeiros programas de comparação de desempenho: o Bonnie++ [2] e o Iozone [3], fornecendo-lhes os parâmetros apropriados (veja o Quadro "Ambiente de Testes").

Briga pela "pole"

Após a realização dos testes de desempenho, não nos causou espanto ver que o JFS foi, de longe, o vencedor. Entretanto, a versão para Linux do sistema de arquivos de código aberto da IBM traz claramente menos opções de ajustes que os outros sistemas (ver artigo à página 26). Contudo, ele se saiu bem em todos os testes, estando quase sempre entre os três primeiros. O segundo lugar foi um páreo duro, visto que os três concorrentes: o ReiserFS, o XFS e o Ext 2 travaram uma bela batalha pela colocação.

Finalmente, o consumo de CPU guiou a decisão, que deu o segundo lugar ao econômico XFS. O ReiserFS mostrou seu poderio com muitos arquivos pequenos em um mesmo diretório. Os bons valores de velocidade de escrita e leitura foram obtidos ao custo de uma carga de CPU razoavelmente grande. O uso do ReiserFS deveria, por isso, ter como pré-requisito um hardware mais poderoso. Em vez disso, os administradores de sistemas devem optar pelo JFS ou XFS, que se saem bem melhor com hardware menos poderoso.

O Ext 2 continua a convencer, com uma performance apenas ligeiramente pior que a dos três sistemas citados até agora. Ele realiza todas as tarefas com desempenho entre médio e bom e não sobrecarrega muito o processador. O Ext 3, entretanto, apresentou valores muito ruins – ao menos em nossos testes. Motivo: configuramos o Ext 3 para operar propositalmente não em seu modo mais rápido (*ordered*), mas no que oferece mais segurança (*journal*). Assim, o Ext 3 atualiza primeiro os metadados e depois os escreve imediatamente no sistema de arquivos.

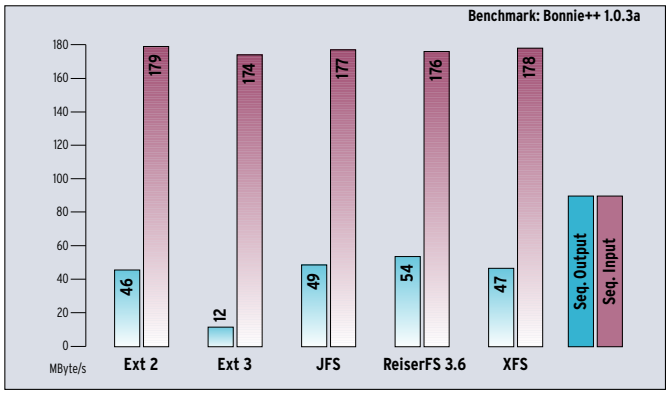


Figura 1: Taxas de transferência com arquivos de 1 GByte.

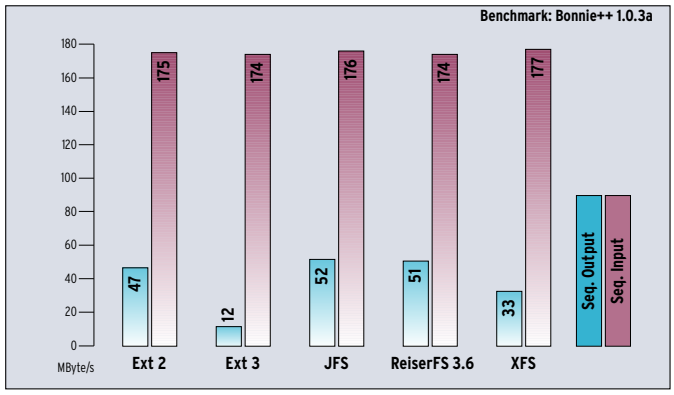


Figura 2: Taxas de transferência com arquivos de 4 GBytes.

Nos mínimos detalhes...

O Bonnie++ realiza diversos testes e os agrupa em saída seqüencial, entrada seqüencial, buscas aleatórias, criação seqüencial e criação aleatória de arquivos. A saída seqüencial mede a taxa de leitura por caractere e por bloco. Da mesma forma, o Bonnie mede o comportamento com a entrada seqüencial. Os gráficos mostram a performance de transferência de entrada e saída seqüenciais por bloco. Percebe-se imediatamente que o desempenho de leitura de todos os sistemas de arquivos são relativamente iguais, não importando se eles lidam com um arquivo de 1 GByte (Figura 1) ou de 4 GBytes (Figura 2). Totalmente diferente é a performance de escrita (Figura 3), na qual arquivos de 1, 2 e 4 GByte foram utilizadas nos testes.

O desempenho do Ext 2 ainda é bem semelhante ao do JFS, XFS e ReiserFS. O mesmo não se pode dizer do sistema de arquivos Ext 3. É uma surpresa que o modo journal emperre tanto o sistema de arquivos. Salta aos olhos a queda de performance com arquivos de 2 GB – problema que o ReiserFS também demonstra.

O sistema de arquivos JFS, da IBM, apresenta desempenho constante e, mesmo com tamanhos crescentes de arquivos, consegue um resultado melhor. Com o XFS acontece precisamente o inverso. O sistema de arquivos da SGI é muito eficiente com arquivos de 1GByte, contudo perde consideravelmente em performance com arquivos de 2 e 4 GByte. Até mesmo o Ext 2 é nitidamente mais rápido com arquivos de 4GByte.

Os resultados alcançados pelo Bonnie com as buscas aleatórias não são mostrados nos gráficos. Nesta categoria do teste, o programa acessa blocos escolhidos aleatoriamente. Aqui o Ext 2 e o Ext 3 ficam claramente na frente e mostram, com todos os tamanhos de arquivos, performances de um a dez por cento melhores que os outros sistemas.

No teste de criação seqüencial de arquivos o Ext 2 é realmente bom: cria em média aproximadamente quatro vezes mais arquivos por segundo que os outros. No teste de leitura o Ext3 e o XFS brigam pela "pole position". Somente no teste de exclusão de arquivos o ReiserFS demonstra que está bem acima da média e mostra em casos isolados um desempenho mais de 100 por cento melhor que o dos outros sistemas.

Ascensão e queda

O teste de criação aleatória mostra o mesmo quadro: o Ext 2 e o Ext 3 criam arquivos a uma taxa extremamente alta. Do outro lado, na leitura e na exclusão de arquivos, o ReiserFS tem desempenho nitidamente melhor. O XFS fica bem atrás da concorrência nos testes de criação. Por outro lado, ele não chega a usar nem 0,1% do tempo de CPU, sobretudo durante o teste de exclusão de arquivos (o que ocorre também com o JFS) e com isso alcança resultados de 1% a 400% melhores que o Ext2, que consome entre 2 e 7 por cento do tempo do processador. E por falar em utilização de processador, neste quesito o ReiserFS faz feio: apesar de excluir arquivos muito rapidamente, ele utiliza no mínimo 7 por cento da CPU, chegando a usar até 99 por cento, no teste de criação aleatória: exclusão de 10.000 arquivos de 2 GByte, o que não é exatamente o desempenho ideal para um servidor em ambiente de produção.

Iozone: Desempenho de leitura e escrita

O benchmark Iozone cria arquivos com 1, 2 e 4 GByte e utiliza blocos de valores entre 64 bytes e 16 kBytes para os testes. Como na análise anterior, tanto a medida da performance de escrita quanto da de leitura eram de nosso interesse. A Figura 4 mostra a performance de escrita com arquivos de 2 GByte. Foram empregados

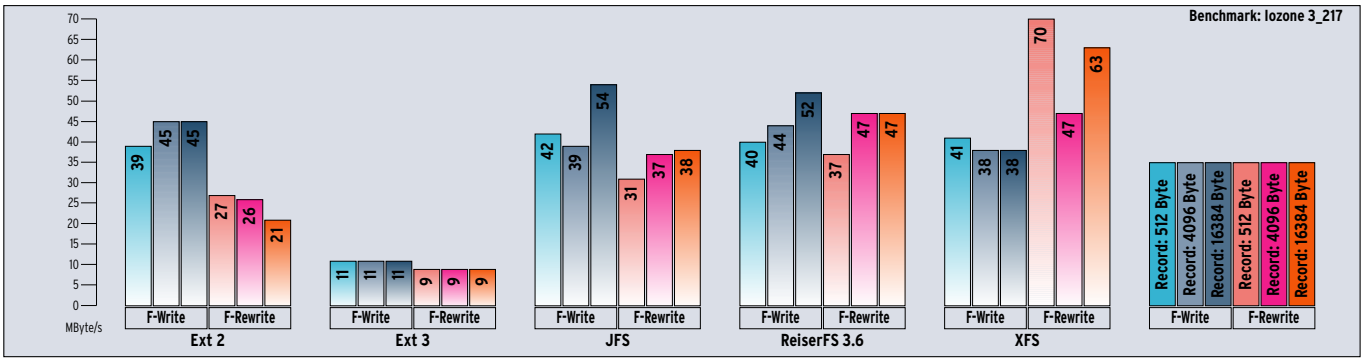


Figura 4: Desempenho de escrita com arquivos de 2 GBytes.

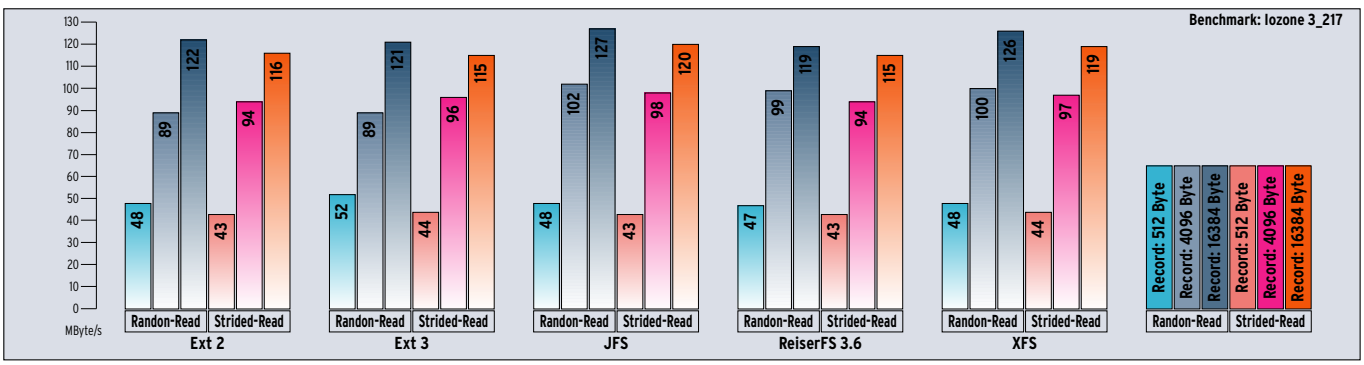


Figura 5: Desempenho de leitura com arquivos de 2 GBytes.

os testes F-Write e F-Rewrite nas medições. Estes utilizam a função fwrite() – da biblioteca *stdio* –, que escreve no sistema de arquivos nos modos "buffered" e "blocked".

Com o teste F-Write, o Iozone cria um novo arquivo, e isso gera um trabalho extra de gravação dos metadados, o que prejudica o desempenho. Já o teste F-Rewrite escreve num arquivo já existente. O gráfico da Figura 4 ilustra muito bem este comportamento, ao menos para o XFS e o ReiserFS. O XFS bate de longe a concorrência em ambos os testes. O desempenho do JFS é relativamente equilibrado. Entretanto, ele obviamente trabalha melhor quanto maior for o arquivo. Contudo, a relação entre os desempenhos sob o F-Write e o F-Rewrite é desigual, o que ocorre também com o Ext 2, cuja performance sob o F-Rewrite despenca. Aqui, o Ext 3 cai novamente na categoria dos "lanternas".

Os gráficos na Figura 5 mostram um desempenho homogêneo. Nela colocamos lado a lado os testes de leitura aleatória e leitura seqüencial intervalada ("strided"). O ouro olímpico nas duas modalidades vai para o JFS. As distâncias são, contudo, relativamente pequenas. Vê-se bem no gráfico as taxas de transferência crescentes quanto maior for o registro a ser lido. A princípio os resultados do Iozone no desempenho de leitura coincidem com os do Bonnie. Os dois indicam uma performance de leitura constantemente boa, indiferentemente se seqüencial ou aleatória e aplicada a pequenos ou grandes registros.

Outros comparativos de desempenho.

A Linux Magazine não é a única a realizar testes de desempenho em sistemas de arquivos. Usuários e os próprios desenvolvedores dos benchmarks publicam resultados na Web.

O site do Fsbench já aplicou a combinação Bonnie-Iozone ao kernel 2.6. Contudo, os resultados não diferem dos da Linux Magazine, que foram realizados com o Kernel 2.4.19, rodando no SuSE Linux Enterprise Server 8 (veja quadro "Ambiente de Testes"). O JFS e o XFS oferecem os melhores resultados de ponta a ponta. Para eles não faz diferença se os arquivos são grandes ou pequenos, nem se há muitos ou poucos deles em um diretório. Para a versão beta do ReiserFS 4 o resultado do Fsbench determina um aumento de desempenho de 65% sobre a versão 3.6, usada em nossos testes. Com isso o ReiserFS se projeta no quesito eficiência, mas ainda assim, não importa a versão, é um consumidor voraz de tempo de processador.

O Ext 2 novamente se sai bem, tanto no que diz respeito à performance nos testes de escrita e leitura quanto na demanda de CPU, esta última pelo menos duas vezes melhor que a do ReiserFS. Os resultados coincidem igualmente no que concerne ao Ext 3. No lento, porém seguro modo "journal", ele fica bem atrás dos outros sistemas de arquivos. Assim que ele é configurado nos modos "ordered" ou "writeback", quase não há mais diferenças entre os seus resultados e os do Ext 2.

O leitor encontrará outros testes comparativos nos sites da Guru-Labs [4], que comparam o Ext 3 com o ReiserFS 3.6, utilizando para isso o benchmark de código aberto *Postmark* da Netapps [5]. Este benchmark visa especificamente aplicativos como servidores de rede e email. Lá, o Ext 3 (em modo "ordered") é o vencedor indiscutível. ■

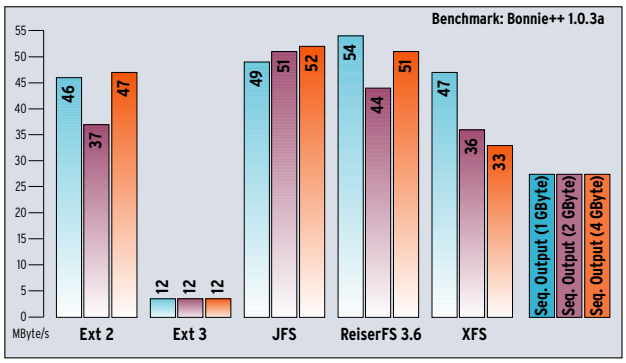


Figura 3: Taxas de transferência para arquivos de diversos tamanhos.

Ambiente de testes

O hardware que serviu de base para nossos testes foi o sistema de armazenamento conectado à rede Zero-One RM-250, da empresa alemã CTT (www.ctt.de, para usuários finais www.mailag.de), um sistema Pentium 4, rodando à 2,4 GHz, com 512 MByte de RAM, uma controladora RAID SATA modelo 8506-8 da empresa 3Ware e oito discos rígidos Maxtor, modelo 7Y250M0 (SATA). Cinco discos rígidos foram agrupados em uma configuração RAID5 com *hot-spare*, sendo que o sexto foi usado para a instalação do SuSE Linux Enterprise System 8 na sua configuração padrão (kernel 2.4.19-120).

O benchmark para sistemas de arquivos que foi utilizado vem de [1] e precisou apenas de ajustes mínimos para rodar no hardware de testes: primeiro os caminhos para os programas do benchmark (Bonnie++ e Iozone) não estavam corretos, depois o resultado do comando *mkfs.reiserfs -V*, que serve para indicar a versão do sistema de arquivos, retornou misteriosamente *1* ao invés de *0*, o que foi corrigido utilizando-se */bin/true*.

O programa Fsbench chama o Bonnie++ e o Iozone com os parâmetros correspondentes. Em nossos testes foi usadas a versão 1.03a do Bonnie++ e a versão 3_217 do Iozone, ambos compilados com o a versão 3.2 do compilador GCC, presente no SuSE Linux Enterprise Server 8.

Foram utilizados 16 GBytes de espaço no sistema RAID, pois a controladora 3Ware usa fatias de dados de 64 KBytes, mas estes dados são distribuídos igualmente por todo o sistema RAID.

A máquina utilizada para os testes comparativos foi um Zero-One RM-250 da empresa CTT, com oito discos rígidos SATA e uma controladora RAID 3Ware.

INFORMAÇÕES

[1] Fsbench: http://fsbench.netnation.com/

[2] Bonnie++: http://www.coker.com.au/bonnie++/

[3] Iozone: http://www.iozone.org/

[4] Guru Labs: http://www.gurulabs.com/ext3-reiserfs-2.html

[5] NetApps: http://www.netapp.com/tech_library/postmark.html