Performance-Vergleich von Ext 2/3, JFS, ReiserFS und XFS

Pisa-Studie

Dateisysteme lassen sich ebenso wie Schüler an der Lese- und Schreibleistung messen. Das Linux-Magazin veranstaltet eine Abschlussprüfung für die Dateisysteme Ext 2/3, JFS, ReiserFS und XFS. Mirko Dölle, Jörg Reitter,

Benchmarks sind wichtige Indikatoren für die Leistungsstärke. Dateisystem-Benchmarks zeigen, wie das System auf bestimmte Operationen wie Lesen oder Schreiben reagiert und wie viel CPU-Last dies erzeugt. Wer beispielsweise viele kleine Dateien in einem Verzeichnis speichert, benötigt ein Dateisystem, das Verzeichnisse sehr schnell durchsucht. Dagegen suchen Datenbank-Administratoren eher nach Unterstützung für große Dateien und erwarten sehr gute Schreib- und Lese-Performance.

Das Linux-Magazin hat das Verhalten der Dateisysteme Ext 2, Ext 3, JFS, ReiserFS und XFS mit der neuen Benchmark-Suite Fsbench [1] untersucht. Sie besteht im Wesentlichen aus einem Python-Skript, das die eigentlichen Benchmark-Programme Bonnie++ [2] und Iozone [3] mit den entsprechenden Parametern versorgt und aufruft (siehe Kasten „So haben wir getestet“).

Kampf um die Plätze

Nach den Benchmark-Läufen staunten die Tester nicht schlecht, als JFS im Großen und Ganzen als Sieger hervorging. Immerhin bringt die Linux-Version von IBMs Open-Source-Dateisystem deutlich weniger Tuning-Optionen mit als die anderen Systeme (siehe Seite 41). Trotzdem machte es bei allen Tests eine gute Figur und war fast immer unter den ersten drei. Der zweite Platz war schwierig zu vergeben, da sich die Dateisysteme ReiserFS, XFS und Ext 2 einen harten Kampf um die Plätze lieferten.

Schließlich entschied der CPU-Hunger, der das genügsame XFS auf den zweiten Platz setzte. Reiser spielte seine Stärke bei vielen kleinen Dateien in einem Verzeichnis aus. Die guten Werte bei der Lese- und Schreibgeschwindigkeit werden aber durch eine teilweise enorme CPU-Belastung relativiert. ReiserFS sollte demnach auf einem leistungsfähigen Hardwarefundament beruhen, sonst ist der Server-Admin mit den Systemen JFS oder XFS besser dran.

Ext 2 wiederum überzeugt mit nur gering schwächerer Performance. Es erledigt alle Aufgaben mittel bis gut und setzt auch den Prozessor nicht allzu sehr unter Last. Ext 2 mit Journal (Ext 3) hingegen liefert sehr schlechte Werte – zumindest bei unserem Test. Grund: Wir haben Ext 3 nicht mit der Standardeinstellung »ordered«, sondern mit dem si-

Diagramm 1: Transferraten bei 1-GByte-Dateien

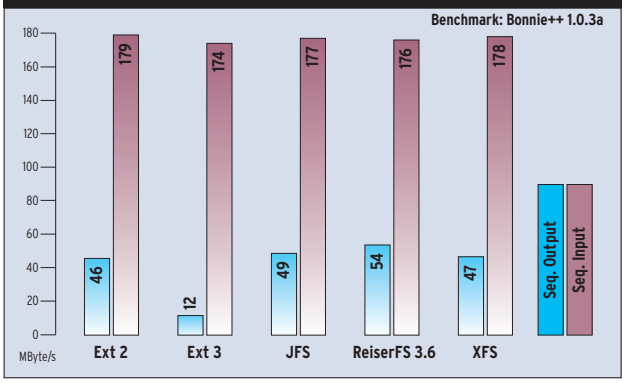


Diagramm 2: Transferraten bei 4-GByte-Dateien

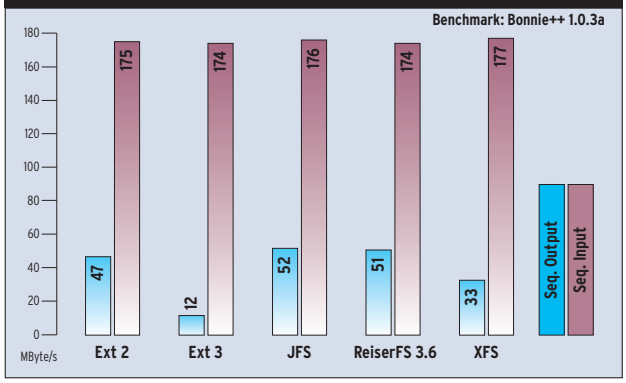
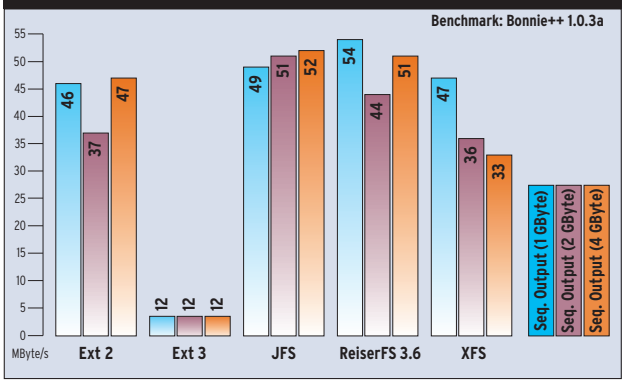


Diagramm 3: Transferrate bei verschiedener Größe



chersten Modus »journal« getestet. In ihm aktualisiert Ext 3 zuerst die Metadaten und schreibt sofort darauf die Daten ins Dateisystem.

Die Benchmark-Ergebnisse im Detail

Bonnie führt mehrere Tests durch und gruppiert sie in Sequential Output, Sequential Input, Random Seeks, Sequential Create und Random Create. Der Sequential Output misst die Lesegeschwindigkeit zeichen- und blockweise. Ebenfalls nach Zeichen und Blöcken bestimmt Bonnie den Durchsatz beim Sequential Input. Die Grafiken zeigen die Transferleistung bei blockweisem Sequential Out- und Input.

Sofort fällt auf, dass die Leseleistung bei allen Dateisystemen relativ gleich ist, egal ob es sich um eine 1-GByte-Datei (Diagramm 1) oder um eine mit 4 GByte handelt (Diagramm 2). Ganz anders sieht es bei der Schreibleistung aus (Diagramm 3). Hier geht es um den Transfer von Dateien mit 1, 2 und 4 GByte Größe.

Das Second-Extended-Filesystem hält mit den Großen locker mit, ganz im Gegensatz zum Journaling-System Ext 3. Dass der »journal«-Modus dieses Dateisystem dermaßen ausbremst, ist eine Überraschung! Auffallend ist auch der Leistungseinbruch bei 2-GByte-Dateien, den ReiserFS ebenfalls aufweist.

Stets performant gibt sich IBMs Journaling-Filesystem JFS, das mit zunehmender Dateigröße erstaunlicherweise einen besseren Durchsatz bringt. Genau umgekehrt sieht es bei XFS aus. SGIs Produkt ist zwar stark bei 1-GByte-Dateien, verliert bei 2 und 4 GByte jedoch erheblich. Sogar Ext 2 ist bei 4-GByte-Dateien deutlich schneller.

Nicht dargestellt sind die Ergebnisse, die Bonnie bei den Random-Seeks ermittelt hat. Dabei greift der Benchmark auf zufällig gewählte Blöcke zu. Hier haben

So haben wir getestet

Ausgangsbasis für den Test war das NAS-System Zero-One RM-250 von CTT – [www.ctt.de], für Endkunden [www.mailag.de] – mit Pentium-4 bei 2,4 GHz, 512 MByte RAM, einem SATA-Raid-Controller 8506-8 von 3Ware sowie acht Maxtor-Festplatten des Typs 7Y250M0 (SATA). Fünf Festplatten wurden als Raid-5 mit Hot-Spare konfiguriert, die sechste diente als Systemplatte für SuSE Linux Enterprise Server 8 in der Standardkonfiguration (Kernel 2.4.19-120).

Der Filesystem-Benchmark stammt von [1] und musste geringfügig angepasst werden, damit er auf dem Testsystem lief: Einmal stimmten die Pfade zu den Benchmark-Programmen Bonnie++ und Iozone nicht, zum anderen lieferte die Versionsabfrage von ReiserFS mittels »mkfs.reiserfs -V« unerklärlicherweise den Return Value »1« anstatt »0«. Der anschließende Aufruf von »/bin/true« korrigierte das Problem.

Der Fsbench ruft Bonnie++ und Iozone mit entsprechenden Parametern auf, für diesen Test kamen Bonnie++ in Version 1.03a sowie Iozone 3_217 zum Einsatz. Beide Benchmarks wurden mit dem GCC 3.2 von Suse Enterprise 8 übersetzt.

Von dem Raid-5-System wurden 16 GByte eingesetzt – der 3Ware-Controller verwendet ein Striping von 64 KByte, die Daten werden daher trotzdem über das gesamte Raid verteilt.

Abbildung 1: Der Testrechner Zero-One RM-250 von CTT ist mit acht SATA-Festplatten und einem entsprechenden Controller von 3Ware bestückt.

Ext 2 und Ext 3 die Nase vorn und zeigen bei allen Dateigrößen eine bis zehn Prozent bessere Leistung als die anderen Systeme. Im Sequential-Create-Test ist Ext 2 richtig gut. So legt es im Schnitt viermal so viele Dateien in der Sekunde an wie die anderen. Beim Read-Test streiten sich Ext 3 und XFS um die Krone. Erst im Delete-Test zeigt auch ReiserFS, dass es mehr als nur Mittelmaß ist, und stellt sich mit teilweise über 100 Prozent besserer Leistung als die anderen System dar.

Bonnie: Create und Delete

Das gleiche Bild zeigt sich beim Random-Create-Test. Ext 2 und auch Ext 3 legen die Testdateien rasend schnell an. Geht es ums Lesen und Löschen, liefert hingegen ReiserFS den deutlich besseren Durchsatz. XFS bleibt bei den Create-Tests durch die Bank weit hinter den Konkurrenten zurück. Allerdings schont es die Prozessorressourcen. Es nutzt wie JFS bei den Delete-Tests kaum mehr als 0,1 Prozent der CPU-Zeit.

Ganz übel sieht es bei ReiserFS aus, das zwar sehr schnell löscht, aber dafür mindestens sieben Prozent der CPU beansprucht. Maximal vereinnahmt es sogar 99 Prozent (Random-Create-Test: Delete mit 2 GByte Dateigröße und 10000 Dateien), was auf einem Produktivserver für Wartezeiten bei den Nutzern sorgt.

Iozone: Schreib- und Lese-Performance

Der Benchmark Iozone legt Dateien mit 1, 2 und 4 GByte an und benutzt für die Tests Blöcke von 64 Byte bis 16 KByte. Hier hat uns ebenfalls die Lese- und Schreibleistung interessiert. Das Diagramm 4 zeigt die Schreib-Performance bei 2-GByte-Dateien. Eingesetzt wurden die so genannten F-Write- und F-Rewrite-Tests, sie nutzen die Library-Funktion »fwrite()«, die Buffered- und Blocked-Schreibvorgänge ausführt.

Beim F-Write-Test legt Iozone eine neue Datei an, weshalb der Overhead der Metadaten-Sicherung die Leistung schmälert. Der F-Rewrite-Test schreibt in ein bereits vorhandenes File. Dies zeigt der Graph im Diagramm 4 recht gut, zumindest bei XFS und ReiserFS. XFS schlägt im Verhältnis von F-Write und F-Rewrite die Konkurrenz um Längen. JFS ist relativ ausgeglichen, kommt aber offenbar besser zurecht, je größer die Datei ist. Allerdings stimmt das Verhältnis F-Write und F-Rewrite nicht – genauso wenig wie bei Ext 2, das bei F-Rewrite stark einbricht.

Ein schönes Bild von gleichmäßig verteilter Leistungsfähigkeit zeigen die Graphen in Diagramm 5. Hier stehen sich die Tests Random-Read und Strided-Read gegenüber. Gold in beiden Disziplinen geht an JFS. Die Abstände sind jedoch gering. Gut im Graphen zu sehen sind die zunehmenden Transferleistungen, je größer der zu lesende Record ist. Im Prinzip decken sich bei der Lese-Performance die Ergebnisse von Iozone mit denen von Bonnie. Beide bescheinigen den Dateisystemen eine gleich bleibend gute Leistung bei sequenziellen und zufälligen Lesevorgängen und bei großen wie kleinen Record-Größen. ■

Infos

[1] Fsbench: [http://fsbench.netnation.com/]
[2] Bonnie++: [http://www.coker.com.au/bonnie++/]
[3] Iozone: [http://www.iozone.org]

Diagramm 4: Schreib-Performance bei 2-GByte-Datei

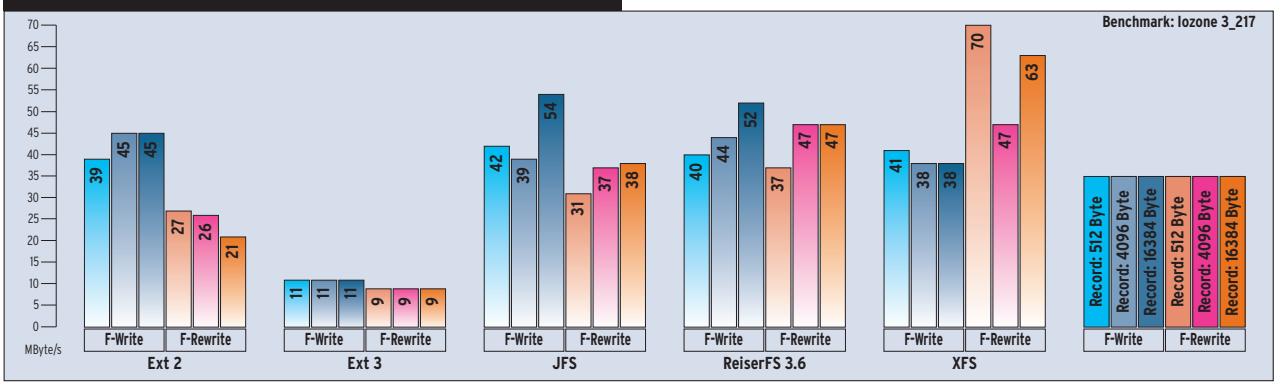


Diagramm 5: Lese-Performance bei 2-GByte-Datei