Komparacija sistema datoteka na Linux kernelu 2.6

B. Đorđević, V. Timčenko, and D. Ilić

***Sadržaj* — Ovaj rad se koncentriše na komparaciju performansi između dva Linux sistema datoteka sa journaling opcijom, ext3 i ReiserFS. Glavni cilj rada je da obavi analizu uticaja na performanse zbog različitih journaling metoda, koje primenjuju na jednoj strani ext3, kao podrazumevani Linux sistem datoteka u odnosu na ReiserFS koji je optimizovan za rad sa malim datotekama. Performanse će biti merene korišćenjem Postmark benchmark programa, koji emulira Internet mail server sa parametrima koji su definisani od strane autora.**

***Ključne reči* — ext3, journaling tehnika, performanse, sistem datoteka, ReiserFS.**

I. Uvod

LINUX je moderan, sofisticiran i moćan operativni sistem. Novije verzije Linux kernela uključuju podršku za rad sa visoko performansnim journaling sistemima datoteka, poput ext3, ReiserFS, XFS i JFS sistema datoteka. Podrška za ext3 sistem datoteka, čiji je autor Dr. Stephen Tweedie, uključena je u većinu distribucija Linux-a, kao što su Red Hat, počev od verzije 7.2, i SuSE, počev od verzije 7.3.

Cilj ovog rada je da obavi komparaciju performansi dva Linux sistema datoteka koja primenjuju journaling tehniku. Journaling tehnika, pored povećanja pouzdanosti, može izazvati izvesno smanjenje performansi, zato što se pored upisa u sistem datoteka, obavlja upis u log datoteku (journal).

Cilj ovog rada je uporedna analiza performansi 32 bitnog ext3 sistema datoteka sa 32 bitnim Reiser. Sistemi datoteka su testirani u identičnom okruženju - svi testovi su obavljeni na identičnom hardveru i na identičnom rasporedu sistema datoteka na disku (sve je isto, samo se sistem datoteka za test formatira u ext3 i Reiser formatu).

Prilikom podizanja operativnog sistema proverava se integritet sistema datoteka. Gubitak integriteta se najčešće javlja kao posledica nasilnog obaranja sistema, odnosno promena u objektima sistema datoteka koje nisu blagovremeno ažurirane u tabeli indeksnih čvorova (inode tables), a može za posledicu imati gubitak podataka.

Opasnost od gubitka podataka umanjuje se uvođenjem journaling tehnike, odnosno dnevnika transakcija koji prati aktivnosti vezane za promenu meta-data oblasti, odnosno inode tabele, i objekata sistema datoteka [1], [2]. Dnevnik (journal, log) se ažurira pre promene sadržaja objekata i prati relativne promene u sistemu datoteka u odnosu na poslednje stabilno stanje. Transakcija se zatvara po obavljenom upisu i može biti ili u potpunosti prihvaćena ili odbijena. U slučaju oštećenja, izazvanog na primer nepravilnim gašenjem računara, sistem lako može da se rekonstruiše povratkom na stanje poslednje prihvaćene transakcije.

U ext3 sistemu datoteka prisutna su tri režima vođenja dnevnika transakcija: journal, ordered i writeback. [3], [4], [5]. Obavimo kratku diskusiju sva 3 režima journaling tehnike.

Journal je režim praćenja svih promena u sistemu datoteka, kako u meta-data oblasti, tako i u objektima, čime se pouzdanost sistema datoteka znatno uvećava na račun performansi. Redundansa koju ovaj režim rada unosi je velika.

Ordered je režim praćenja promena u meta-data oblasti, pri čemu se promene u objektima sistema datoteka upisuju pre ažuriranja inode tabele. Ovo je podrazumevani režim rada dnevnika, koji garantuje potpunu sinhronizaciju objekata sistema datoteka i meta-data oblasti. U odnosu na journal, ovaj režim karakteriše manja redundansa i veća brzina rada.

Writeback je režim praćenja promena u meta-data oblasti, pri čemu se inode tabela može ažurirati pre upisa promena u objekte sistema datoteka. Ovo je najbrži režim rada, ali ne garantuje konzistenciju meta-data oblasti, odnosno sinhronizaciju objekata sistema datoteka i meta-data oblasti.

Jedan od prvih sistema datoteka sa "journaling" opcijom je ReiserFS, verzija 3.6.x (prisutna u Linux kernelima počev od verzije 2.4). ReiserFS, koji je ime dobio po tvorcu, Hans Reiseru, značajno povećava performanse pri radu sa malim datotekama (small file performance), koje su kod ostalih journaling sistema datoteka veoma slabe. Brojni testovi pokazuju da je ReiserFS osetno brži od ext2/ext3, pri radu sa jako malim datotekama (manjim od 1KB). Dodatno, ReiserFS razrešava problem interne fragmentacije, čime se povećava efikasnost iskorišćenja diskova. Ovako visoke performanse pri radu sa malim datotekama ReiserFS postiže na osnovu optimizovanog B+ stabla (jedno po sistemu datoteka) i dinamičke alokacija i-node čvorova (za razliku od fiksne alokacije i-node koju koristi ext2). Dodatno, ReiserFS koristi promenljivu veličinu

Borislav Đorđević, Institut Mihajlo Pupin, Volgina 15, 11050 Beograd, Srbija; (e mail: bora@impcomputers.com)

Valentina Timčenko, Institut Mihajlo Pupin, Volgina 15, 11050 Beograd, Srbija; (e mail: valentina.timcenko@institutepupin.com)

Darko Ilić, Institut Mihajlo Pupin, Volgina 15, 11050 Beograd, Srbija; (e mail: darko@impcomputers.com)

sistemskog bloka, a male datoteke se upisuju u svoj direktorijum, zajedno sa svojom FCB (file control block) strukturom. U log dnevniku se ažuriraju samo promene u meta-data oblasti.

Loše osobine ReiserFS reflektuju se pri radu sa šupljim datotekama (sparse files), gde je ext2/ext3 daleko bolji. Takođe, ReiserFS radi sporije sa velikim datotekama u odnosu na ext2/ext3 [1] [2] [4] [5] [6].

II. METODOLOGIJA TESTIRANJA

Postoji nekoliko mogućih scenarija za određivanje performansi sistema datoteka. Testiranje se može obaviti pomoću svetski priznatog benchmark softvera, koji simulira različite vrste opterećenja, poput opterećenja Internet Service Provider-a ili NetNews servera. Drugi način uključuje korišćenje specijalnih testova, specijalno dizajniranih u te svrhe, poput testova sekvencijalnog i slučajnog čitanja i pisanja, kreiranja datoteka i simulacije rada u aplikaciji.

Za potrebe ovog rada korišćen je PostMark [7] softver koji simulira opterećenje Internet Mail servera. PostMark kreira veliki inicijalni skup (pool) slučajno generisanih datoteka na bilo kom mestu u fajl sistemu. Nad tim skupom se dalje vrše operacije kreiranja, čitanja, upisa i brisanja datoteka i određuje vreme potrebno za izvršavanje tih operacija. Redosled izvođenja operacija je slučajan čime se dobija na verodostojnosti simulacije. Broj datoteka, opseg njihove veličine i broj transakcija su u potpunosti konfigurabilni, a radi eliminisanja cache efekata preporučuje se kreiranje inicijalnog skupa sa što većim brojem datoteka (bar 10000) i izvršenje što većeg broja transakcija.

Konfiguraciju za testiranje performansi sistema datoteka odlikuju sledeći fundamentalni parametri: matična ploča, vrsta i radni takt procesora, količina i vrsta drugostepene keš memorije, količina i vrsta operativne (RAM) memorije, tip i model disk kontrolera, tip i model diska.

Peformanse ext2/ext3 i ReiserFS sistema datoteka su testirane na sledećoj konfiguraciji:

TABELA 1: KARAKTERISTIKE TESTING SISTEMA.

matična ploča	*Intel Server Board S845WD1-E*
procesor	Intel Pentuim IV 2.66GHz
L2 keš	L2 onboard cache 512KB
operativna memorija	512MB DIMM
disk kontroler	PATA
disk	DiamondMax Plus 8

Glavne karakteristike diska upotrebljenog u testu prikazane su u tabeli 2.

TABELA 2: KARAKTERISTIKE DIAMONDMAX® PLUS 8

Kapacitet	*40GB*
average seek time (prosečna brzina pristupa)	<10ms
brzina okretanja ploča (rpm)	7200
brzina interfejsa (MB/s)	133
veličina bafera (MB)	2

Testiranje je obavljeno na distribuciji Linux-a, Red Hat Fedora 5 sa stabilnom verzijom kernela 2.6.15-1.

Sistemi datoteka su kreirani u logičkim particijama na sledeći način:

Filesystem	**Size**	**Type**	**Description**
/dev/hda1	30G	ntfs	Non Linux
/dev/hda2	200M	ext3	boot FS
/dev/hda3	6.1G	ext3	root FS
/dev/hda5	2G	swap	swap
/dev/hda6	1.6G	ext3	testing FS
/dev/hda7	300M	ext3	auxiliary FS

Sistem datoteka /dev/hda6 je korišćen za testiranje performansi i najpre je kreiran kao generički ext2 sistem datoteka koji ne vodi dnevnik transakcija. Konverzija u ext3 izvršena je kreiranjem dnevnika transakcija pri aktiviranju sistema datoteka. Redom su kreirana tri tipa ext3 dnevnika čije su performanse određene, a nakon svakog testa je fajl sistem vraćen u generičko stanje. Dat je spisak komandi za aktiviranje sistema datoteka sa kreiranjem journal, ordered i writeback dnevnika transakcija i konvertovanje ext3 sistema datoteka u ext2, čime se dnevnik poništava:

-aktiviranje sistema datoteka sa kreiranjem journal dnevnika

```
#mount –o data=journal /dev/hda6 /testFS
```

-aktiviranje sistema datoteka sa kreiranjem ordered dnevnika

```
#mount –o data=ordered /dev/hda6 /testFS
```

-aktiviranje sistema datoteka sa kreiranjem writeback dnevnika

```
#mount –o data=writeback /dev/hda6 /testFS
```

-konvertovanje ext3 sistema datoteka u ext2 sa proverom integriteta sistema datoteka (fajl sistem mora biti neaktivan):

```
#tune2fs -O ^has_journal /dev/hda6
#fsck.ext2 -f /dev/hda6
```

III. REZULTATI TESTIRANJA

Izvršena su tri različita testa performansi nad različitim skupovima slučajno generisanih datoteka. Testovi su obavljeni, na ext2 sistemu datoteka, na sve tri journaling opcije ext3 sistema datoteka i na Reiser sistemu datoteka.

A. Test1

U prvom testu (testu malih i srednjih datoteka) je izvršeno 50.000 transakcija nad skupom od 2000 slučajno generisanih datoteka čije se veličine kreću u opsegu 1KB-

100KB, što rezultuje čitanjem i pisanjem približno 1.5GB podataka. Ova suma prevazilazi količinu sistemske memorije i generalno eliminiše efekte keširanja diskova.

PostMark konfiguracija:

- set size 1000 100000
- set number 2000
- set transactions 50000

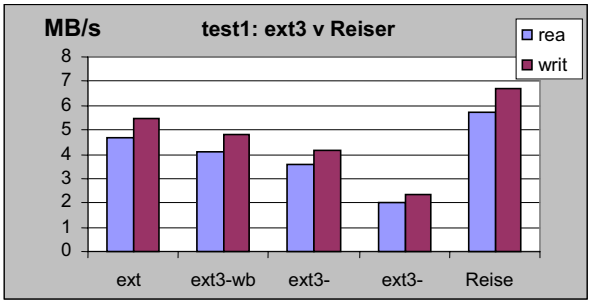
Rezultati testa dati su u tabelama 3 i 4, a grafički prikazani na slici 1.

TABELA 3: REZULTATI PRVOG TESTA ZA EXT3

MB/s	*ext2*	*ext3-wb*	*ext3-o*	*ext3-j*
read	4.67	4.12	3.58	2.02
write	5.46	4.82	4.19	2.36

TABELA 4: REZULTATI PRVOG TESTA ZA REISERFS

MB/s	*Reiser*
read	5.71
write	6.68



Sl. 1. Grafički prikaz performansi (prvi test).

U ovom testu malih datoteka, veliki broj I/O operacija uključujući i metadata operacije i filedata operacije. Zato se očekuje da na performanse imaju kombinovan uticaj i journaling tehnika i keširanje datoteka (file caching). U okviru testa malih datoteka, ReiserFS je superioran u odnosu na ext2 i sva tri ext3 journaling moda. ReiserFS je oko 30% brži od ext2 FS i najbrže ext3 opcije (writeback). ReiserFS je oko 1.5 puta brži od default ext3 opcije (ordered). Reiser je oko 3 puta brži od najsporije ext3 opcije (journal). Za ovakav test malih datoteka, ReiserFS je brži od ext2 i ext3, uglavnom zbog otimizacije za male datoteke, boljeg sopstvenog keširanja datoteka i raznih tehnika za optimizaciju.

B. *Test2*

U drugom testu (ultra male datoteke) je izvršeno 50.000 transakcija nad velikim skupom slučajno generisanih datoteka, 30000 datoteka, čije se veličine kreću u opsegu 1bajt-1KB, što rezultuje čitanjem i pisanjem približno oko 25MB podataka. Ovakva konfiguracija generiše veliki broj zahteva za ažuriranje meta-data oblasti, odnosno inode tabele.

PostMark konfiguracija:

- set size 1 1000
- set number 30000
- set transactions 50000

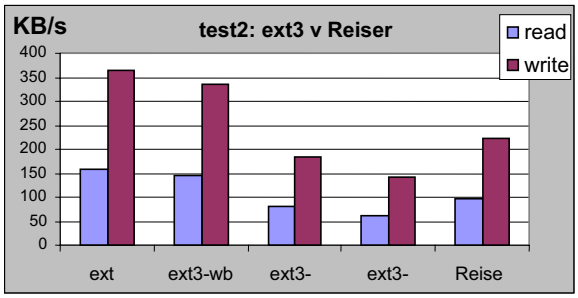
Rezultati testa dati su u tabeli 5 i 6, a grafički prikazani na slici 2.

TABELA 5: REZULTATI DRUGOG TESTA ZA EXT3

KB/s	*ext2*	*ext3-wb*	*ext3-o*	*ext3-j*
read	158.33	145.14	80.07	62.2
write	364.76	334.36	184.48	143.3

TABELA 6: REZULTATI DRUGOG TESTA ZA REISERFS

KB/s	*Reiser*
Read	96.76
Write	222.91



Sl. 2. Grafički prikaz performansi (drugi test).

Ovaj test uključuje ogroman broj veoma malih datoteka, pa samim tim i ogroman broj metadata operacija. Zato se očekuje da journaling i komponete file-keša, kao što su metadata keš i direktorijumski keš imaju dominantan uticaj na performanse. U ovom testu ultra malih datoteka, dogodilo se iznenađenje, ext3 je superioran u odnosu na ReiserFS, što nije bio slučaj na kernel verzijam 2.4. Sistemi datoteka ext2 i najbrži ext3 journaling mod su brži od ReiserFS, dok je ReiserFS malo brži od ext3 modova (ordered i journal mode). Za ovaj test, najbrži je ext2, a to je sistem datoteka bez journaling tehnike, dok svi ostali imaju pad performansi zbog journaling tehnike. Sistem datoteka ext2 je oko 70% brži od ReiserFS.

U ovakvom testu sa ultra malim datotekama, journaling tehnika, metadata i direktorijumski keš imaju dominantan uticaj na performanse. Očekivali smo da ReiserFS bude ubeljivo najbolji ali to se niije dogodilo.

C. *Test3*

U trećem testu (širok dijapazon malih i srednjih datoteka) je izvršeno 50.000 transakcija nad skupom od 2000 slučajno generisanih datoteka čija je maksimalna veličina povećana na 300KB, što rezultuje čitanjem i pisanjem približno 4,5GB podataka. Ovaj test je vrlo intenzivan - ukupna količina podataka za čitanje i upis je znatno veća od količine sistemske memorije i u potpunosti eliminiše efekte svih mehanizama keširanja.

PostMark konfiguracija:

- set size 1000 300000
- set number 2000
- set transactions 50000

Rezultati testa dati su u tabelama 7 i 8, a grafički

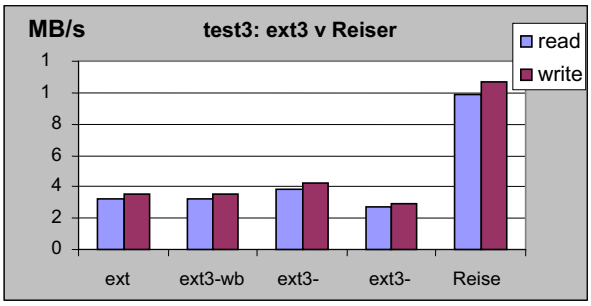
prikazani na slici 3.

TABELA 7: REZULTATI TREĆEG TESTA ZA EXT3

MB/s	*ext2*	*ext3-wb*	*ext3-o*	*ext3-j*
read	3.27	3.25	3.88	2.68
write	3.53	3.51	4.19	2.89

TABELA 8: REZULTATI TREĆEG TESTA ZA REISERFS

MB/s	*Reiser*
read	9.86
write	10.64



Sl. 3. Grafički prikaz performansi (treći test).

U okviru testa malih i srednjih datoteka, smanjuje se razlika u performansama između ext2 i ext3, dok ReiserFS ubedljivo pobeđuje i ext2 i sve opcije ext3. Sa povećanjem veličine test datoteka, filedata transferi dominiraju tako da filedata keširanje ima dominantan uticaj na performanse. ReiserFS je skoro 3 puta brži od ext2 i ext3 moda.

IV. ZAKLJUČAK

U ovom radu, napravili smo komparaciju između 32 bitnog ext2/ext3 FS i takođe 32bitnog ReiserFS, nadaleko čuvenog po svojoj optimizaciji u radu sa malim datotekama. Svi testovi su obavljeni na relativno malom sistemu datoteka (1GB). Iako smo očekivali da je Reiser bolji na ultra malim datotekama, tu se ipak dogodilo iznenađenje, ext3 je bio bolji. Međutim sa porastom veličine datoteka, sistem datoteka Reiser je dominantno bolji u odnosu na ext3, kao na primer u trećem testu, gde je skoro 3 puta bio bolji.

I ReiserFS i ext3 na kernelima 2.6 su doživeli brojna unapređenja u pogledu keširanja, tretmana metadata oblasti, kao journaling tehnike, a sve pomenute tehnike imaju kombinovani uticaj na performanse. U principu, mogli bi da konstatujemo da se sva 3 naša testa odnose na relativno male datoteke, u kojima Reiser ubedljivo pobeđuje u 2 od 3 testa.

LITERATURA

[1] G. Ganger, Y. Patt, "Metadata Update Performance in File Systems", OSDI Conf Proc., pp. 49-60, Monterey, CA, Nov. 1994.

[2] M. Seltzer, G. Ganger, M. McKusick, K. Smith, C. Soules, C. Stein, "Journaling versus Soft Updates: Asynchronous Meta-data Protection in File Systems", USENIX Conf. Proc., pp. 71-84, San Diego, CA, June 2000.

[3] K. M. Johnson, "Red Hat's New Journaling File System: ext3", www.redhat.com/support/wpapers/redhat/ext3/

[4] S. Tweedie S., "EXT3, Journaling Filesystem" July 20, 2000, http://olstrans.sourceforge.net/release/OLS2000-ext3/OLS2000-ext3.html

[5] D. Robbins, "Introducing ext3", Gentoo Technologies, Inc., http://www-106.ibm.com/developerworks/library/l-fs7/

[6] ReiserFS, http://www.namesys.com

[7] J. Katcher, "PostMark: A New File System Benchmark", Technical Report TR3022. Network Appliance Inc, Oct. 1997.

ABSTRACT

Abstract: — This paper concentrates on the Linux ext3 and Reiser filesystem performance comparison problem, under Linux kernel 2.6. Main goal this paper should achieve is analysis of performance impact due to a different journaling approach, implemented in 32 bit ext3 filesystem (default Linux filesystem) related to 32 bit Reiser filesystem, optimized for small file performances. The performance is measured using Postmark benchmark software, which emulates Internet mail server with environment defined by the authors.

LINUX FILE SYSTEMS COMPARISON ON THE KERNEL 2.6

Đorđević, B., Timčenko, V., Ilić, D.