Komparacija 64-bitnih sistema datoteka na Linux kernelu 2.6

B. Đorđević, and V. Timčenko

***Sadržaj* — Ovaj rad se koncentriše na komparaciju performansi JFS i XFS, dva glomazna 64-bitna Linux sistema datoteka sa journaling opcijom, implementiranih na Linux kernel verziji 2.6. Rad predstavlja analizu uticaja različitih journaling i keš metoda JFS sistema datoteka, kao IBM 64-bitnog sistem datoteka, u odnosu na XFS, kao SGI 64-bitnog sistema datoteka. Performanse su merene korišćenjem Postmark benchmark programa koji emulira Internet mail server sa parametrima koji su definisani od strane autora.**

***Ključne reči* — JFS, journaling tehnika, performanse, sistem datoteka, XFS.**

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*).

Težište rada je uporedna analiza performansi 64 bitnog JFS sistema datoteka, sa 64 bitnim XFS sistemom datoteka. 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 JFS odnosno XFS 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 (*i-node 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 i-node tabele, i objekata sistema datoteka [1], [2], [3], [4]. 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.

Borislav Đorđević, Institut Mihajlo Pupin, Volgina 15, 11060 Beograd, Srbija; (e mail: bora@impcomputers.com)

Valentina Timčenko, Institut Mihajlo Pupin, Volgina 15, 11060 Beograd, Srbija; (e mail: valentina.timcenko@institutepupin.com)

II. XFS I JFS

Sistem datoteka XFS je originalni proizvod kompanije Silicon Graphics, Inc (SGI), razvijen početkom devedesetih godina. Predstavlja robusni, puni 64-bitni sistem datoteka, sa brojnim kvalitetnim osobinama, prvenstveno namenjen za SGI IRIX, ali je napravljena i verzija za Linux.

Prva novina u dizajnu XFS [5] sistema datoteka je uvođenje alokacionih grupa, odnosno linearnih regiona jednake veličine, koji se definišu za svaki disk. Svaka alokaciona grupa ima svoju i-node tabelu i listu slobodnog prostora. Alokacione grupe su nezavisne i mogu učestvovati u paralelnim I/O operacijama i na taj način se omogućavaju istovremene paralelne I/O operacije na istom sistemu datoteka.

Interno, svaka alokaciona grupa koristi efikasna B+ stabla koja čuvaju informaciju o zonama slobodnog prostora i o slobodnim i-node čvorovima. XFS optimizuje alokaciju slobodnog prostora (koja je kritična po pitanju performansi upisa) putem odložene alokacije (*delayed allocation*).

Na bazi testova iz otvorene literature, mogu se postaviti zanimljivi zaključci. XFS uglavnom pobeđuje u radu sa velikim datotekama, ali u testovima sa intenzivnim brisanjem datoteka, pobeđuju ga i ReiserFS i ext3.

JFS je puni 64 bitni journaling sistem datoteka koji je namenjen prvenstveno za IBM servere, ali je takođe portiran i za Linux operativni sistem [6]. Kao i svi journaling sistemi, obezbeđuje visoku pouzdanost, brz oporavak sistema datoteka i visoke performanse. Po pitanju journaling tehnike, JFS upisuje u log samo metadata podatke.

JFS je extent baziran sistem datoteka, što mu omogućava da fleksibilno manipuliše datotekama na disku. Ekstent je sekvenca kontinualnih blokova, koji se dodeljuju datoteci i koju specificiraju tri parametra

<logical offset, length, physical>. Stablo je bazirano na B+ strukturi. JFS koristi različite veličine za sistemske blokove (512, 1024, 2048 i 4096 bajtova), što omogućava prilagođenje sistema datoteka prema potrebama.

I-node čvorovi i liste slobodnih blokova se održavaju dinamički, što svakako ima velike prednosti. Što se tiče direktorijuma, postoje dve moguće organizacije. Prva šema se koristi za male direktorijume i kod nje se sadržaj direktorijuma upisuje u njegov i-node, što drastično poboljšava performanse malih direktorijuma. Druga šema se koristi za velike direktorijume, realizovane u formi optimizovanog B+ stabla, koje optimizuje pretraživanje, unos novih objekata i brisanje.

JFS se dobro snalazi i sa šupljim (*sparse files*) i sa gusto popunjenim datotekama (*dense files*). Kao puni 64 bitni sistem datoteka, odličan je u slučaju kada se realizuju ogromni sistemi datoteka.

III. 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 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 JFS i XFS 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 Red Hat Fedora 5 distribuciji Linux-a, 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	.....	testing FS
/dev/hda7	300M	ext3	auxiliary FS

Sistem datoteka /dev/hda6 je korišćen za testiranje performansi i najpre je kreiran kao JFS, a potom kao XFS.

IV. REZULTATI TESTIRANJA

Izvršena su tri različita testa performansi nad različitim skupovima slučajno generisanih datoteka. Testovi su obavljeni nad 64-bitnim XFS i 64-bitnim JFS sistemom datoteka.

A. Test1

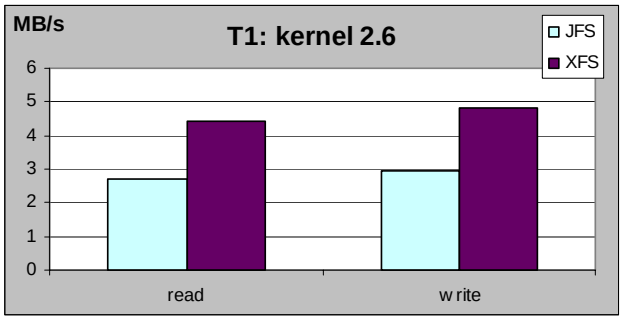
U prvom testu (testu malih i srednjih datoteka) je izvršeno 50.000 transakcija nad skupom od 2000 slučajno generisanih datoteka čije se veličine kreću u opsegu 1KB-100KB, što rezultuje čitanjem i pisanjem približno 1.5GB podataka. Ova suma prevazilazi količinu sistemske memorije i generalno eliminiše efekte keširanja diskova.

PostMark konfiguracija:
- set size 1000 100000
- set number 2000
- set transactions 50000

Rezultati testa dati su u tabeli 3, a grafički prikazani na slici 1.

TABELA 3: REZULTATI PRVOG TESTA

MB/s	***JFS***	***XFS***
read	2.7	4.44
write	2.93	4.83



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, SGI XFS je superioran u odnosu na IBM JFS. XFS je oko 1.65 puta brži od JFS. Za ovakav test malih datoteka, XFS je brži od JFS, zbog boljeg sopstvenog keširanja datoteka i raznih tehnika za optimizaciju.

B. Test2

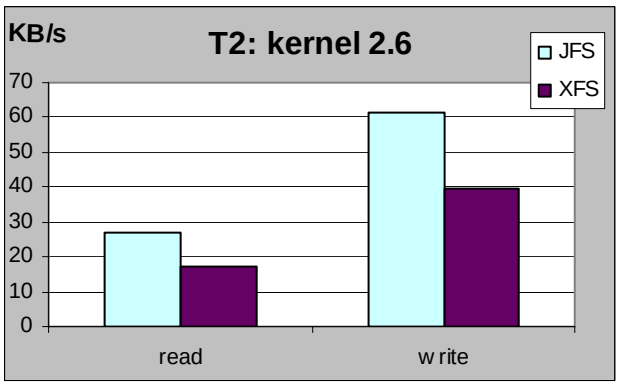
U drugom testu (ultra male datoteke) je izvršeno 50.000 transakcija nad velikim skupom slučajno generisanih datoteka, 30000 datoteka, čije se veličine kreću u opsegu 1bajt-1KB, što rezultuje čitanjem i pisanjem približno oko 25MB podataka. Ovakva konfiguracija generiše veliki broj zahteva za ažuriranje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

PostMark konfiguracija:
- set size 1 1000
- set number 30000
- set transactions 50000

Rezultati testa dati su u tabeli 4, a grafički prikazani na slici 2.

TABELA 4: REZULTATI DRUGOG TESTA

KB/s	*JFS*	*XFS*
read	26.74	17.2
write	61.61	39.63



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, ovog puta JFS je superioran u odnosu na XFS, što nije bio slučaj na kernel verzijama 2.4. Sistem datoteka JFS je brži oko 1.6 puta od XFS.

C. Test3

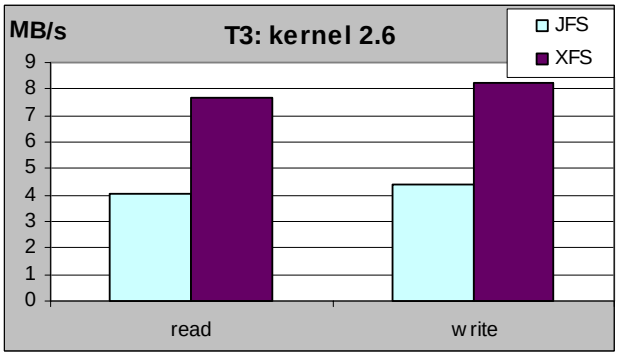
U trećem testu (širok dijapazon malih i srednjih datoteka) je izvršeno 50.000 transakcija nad skupom od 2000 slučajno generisanih datoteka čija je maksimalna veličina povećana na 300KB, što rezultuje čitanjem i pisanjem približno 4,5GB podataka. Ovaj test je vrlo intenzivan - ukupna količina podataka 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 tabeli 6, a grafički prikazani na slici 3.

TABELA 6: REZULTATI TREĆEG TESTA

MB/s	*JFS*	*XFS*
read	4.07	7.64
write	4.39	8.25



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

U okviru testa malih i srednjih datoteka, sa povećanjem veličine test datoteka, filedata transferi dominiraju tako da filedata keširanje ima dominantan uticaj na performanse.

U okviru testa malih i srednjih datoteka, SGI XFS je superioran u odnosu na IBM JFS. XFS je skoro dva puta brži od JFS. Za ovakav test malih datoteka, XFS je brži od JFS zbog boljeg sopstvenog keširanja datoteka i raznih tehnika za optimizaciju.

V. ZAKLJUČAK

U ovom radu, napravili smo komparaciju između robusnih 64 bitnih sistema datoteka, IBM JFS i SGI XFS. I IBM JFS i SGI XFS na kernelima 2.6 su doživeli brojna unapređenja u pogledu, keširanja, tretmana metadata oblasti, kao i journaling tehnike, a sve pomenute tehnike imaju kombinovani uticaj na performanse. To su praktično giganti u svetu sistema datotaka. Svi testovi su obavljeni na relativno malom sistemu datoteka (1GB) i sa relativno malim test datotekama (1bajt -300KB). Očekivali smo da su takvi test uslovi jako nepovoljni da bi u njima 64 bitni giganti mogli da pokažu solidne perfomanse.

SGI XFS se ipak pokazao kao daleko bolji u tim uslovima. U principu, mogli bi da konstatujemo da se sva tri naša testa odnose na relativno male datoteke, u kojima SGI XFS ubedljivo pobeđuje IBM JFS u dva od tri testa. Jedino u testu sa ultra malim datotekama (1bajt do 1K), IBM JFS ubedljivo pobedjuje sa preko 50%, dok sa povećanjem veličine test datoteka, solidnu prednost preuzima SGI XFS, koji se u dve vrste testa skoro duplo bolji u odnosu na IBM JFS.

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 to XFS", Gentoo Technologies, Inc.

[6] JFS, http://www.ibm.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 robust Linux 64-bit JFS and XFS filesystem performance comparison, both tested under Linux kernel 2.6. The goal is to perform specific performance analysis taking into consideration different journaling and caching approaches implemented in 64 bit IBM's JFS filesystem and 64 bit SGI's XFS filesystem. The performance is measured applying Postmark benchmark software, which emulates Internet mail server with environment parameters defined by the authors.

64 BIT LINUX FILE SYSTEMS COMPARISON ON THE KERNEL 2.6

Đorđević, B., Timčenko, V.