PERFORMANSE EXT3 FAJL SISTEMA PRI RAZLIČITIM REŽIMIMA VOĐENJA DNEVNIKA TRANSAKCIJA

Borislav Đorđević, Nemanja Maček, Dragan Pleskonjić, Stanislav Mišković, Borislav Krneta, Predrag Gavrilović
Viša elektrotehnička škola u Beogradu
e-mail: bora@impcomputers.com

Sadržaj – *Uvod, režimi vođenja dnevnika transakcija u ext3 fajl sistemu pod Linux-om, metodologija testiranja, test konfiguracija, parametri operativnog sistema, rezultati tesiranja, zaključak*

1. UVOD

Linux je moderan, sofisticiran i moćan operativni sistem. Novije verzije Linux kernela uključuju podršku za rad sa visokoperformansnim journaling fajl sistemima, poput ext3, ReiserFS, XFS i JFS fajl sistema. Podrška za ext3 fajl sistem, 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 analiza uticaja dnevnika transakcija na performanse fajl sistema, odnosno uporedna analiza performansi ext3 fajl sistema pri različitim režimima vođenja dnevnika. Fajl sistemi su testirani u identičnom okruženju - svi testovi su obavljeni na identičnom hardveru sa identičnim parametrima fajl sistema (verzija kernela i osnova fajl sistema - ext2).

2. REŽIMI VOĐENJA DNEVNIKA TRANSAKCIJA U EXT3 FAJL SISTEMU

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

Opasnost od gubitka podataka umanjuje se uvođenjem dnevnika transakcija koji prati aktivnosti vezane za promenu meta-data oblasti, odnosno i-node tabele, i objekata fajl sistema. Dnevnik (journal, log) se ažurira pre promene sadržaja objekata i prati relativne promene u fajl sistemu u odnosu na poslednje stabilno stanje. Transakcija se zatvara po obavljenom upisu i može biti ili u potpunosti prihvaćena ili odbijena. U slučaju oštetećenja, izazvanog npr. nepravilnim gašenjem računara, fajl sistem može lako rekonstruisati povratkom na stanje poslednje prihvaćene transakcije.

U ext3 fajl sistemu prisutna su tri režima vođenja dnevnika transakcija: **journal**, **ordered** i **writeback**.

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

Ordered je režim praćenja promena u meta-data oblasti, pri čemu se promene u objektima fajl sistema upisuju pre ažuriranja i-node tabele. Ovo je podrazumevati režim rada dnevnika, koji garantuje potpunu sinhronizaciju objekata fajl sistema 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 i-node tabela može ažurirati pre upisa promena u objekte fajl sistema. Ovo je najbrži režim rada, ali ne garantuje konzistenciju meta-data oblasti, odnosno sinhronizaciju objekata fajl sistema i meta-data oblasti.

3. METODOLOGIJA TESTIRANJA

Postoji nekoliko mogućih scenarija za određivanje performansi fajl sistema. 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 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.

4. TEST KONFIGURACIJA

Konfiguraciju za testiranje performansi fajl sistema 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 ext3 fajl sistema su testirane na sledećoj konfiguraciji:

- Matična ploča: D815EE2U
- Procesor: Intel Celeron, 1200MHz
- Drugostepena keš memorija: L2 onboard cache 256KB ECC
- Operativna memorija: 128MB DIMM
- Disk kontroler: Adaptec 29160 (AIC-7899) U2W SCSI
- Disk: Quantum Atlas V

Adaptec 29160 je izabran kao reprezentativni kontroler u klasi SCSI kontrolera (non-RAID), dizajniran da opsužuje servere pri nižim i srednjim opterećenjem. Ovaj jednokanalni SCSI kontroler sa 64-bitnom magistalom je idealan za

povezivanje Ultra160 SCSI (LVD) diskova, poput diskova iz serije Quantum Atlas-V, kao i drugih internih i eksternih uređaja.

Karakteristike Adaptec 29160 kontrolera i Quantum Atlas-V diska date su u tabelama 1 i 2 respektivno.

Tabela 1. *Karakteristike Adaptec 29160 kontrolera*

broj SCSI kanala	jedan (single channel)
radno okruženje	serveri pri nižim i srednjim opterećenjem
brzina interfejsa	160 MB/sec
magistrala	64 bit PCI
konektori za interne uređaje	68 pin LVD SCSI
	68 pin Ultra Wide SCSI
	50 pin Eltra SCSI
konektori za eksterne uređaje	68 pin LVD SCSI

Tabela 2. *Karakteristike Quantum Atlas-V diskova*

average seek time (prosečna brzina pristupa)	6.3ms
full stroke seek (brzina pristupa s kraja na kraj)	15ms
track-to-track seek (brzina pristupa sledećoj stazi)	0.8ms
brzina okretanja ploča	7200 obrtaja u minuti
brzina interfejsa	160 MB/sec
veličina bafera	4 MB

5. PARAMETRI OPERATIVNOG SISTEMA

Testiranje je vršeno na jednoj od najboljih i često prisutnih distribucija Linux-a, Red Hat 8.0 sa stabilnom verzijom kernela 2.4.18-14.

Fajl sistemi su kreirani u logičkim particijama na sledeći način:

- boot fajl sistem /dev/sda5, veličine 99MB
- swap particija /dev/sda6, veličine 256MB
- root fajl sistem /dev/sda7, veličine 2.3GB
- test fajl sistem /dev/sda8, veličine 1.3GB

Fajl sistem /dev/sda8 je korišćen za testiranje performansi i najpre je kreiran kao generički ext2 fajl sistem koji ne vodi dnevnik transakcija. Konverzija u ext3 izvršena je kreiranjem dnevnika transakcija pri aktiviranju fajl sistema. 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 fajl sistema sa kreiranjem **journal**, **ordered** i **writeback** dnevnika transakcija i konvertovanje ext3 fajl sistema u ext2, čime se dnevnik poništava:

- aktiviranje fajl sistema sa kreiranjem **journal** dnevnika
  #mount –o data "journal" /dev/sda8 /testFS
- aktiviranje fajl sistema sa kreiranjem **ordered** dnevnika
  #mount –o data "ordered" /dev/sda8 /testFS
- aktiviranje fajl sistema sa kreiranjem writeback dnevnika
  #mount –o data "writeback" /dev/sda8 /testFS
- konvertovanje ext3 fajl sistema u ext2 sa proverom integriteta fajl sistema (fajl sistem mora biti neaktivan):
  #tune2fs -O ^has_journal /dev/sda8
  #fsck.ext2 -f /dev/sda8

6. REZULTATI TESTIRANJA

Izvršena su tri procene performansi nad različitim skupovima slučajno generisanih datoteka.

1. test

U prvom testu je izvršeno 50000 transakcija nad skupom od 2000 slučajno generisanih datoteka čije se veličine kreću u opsegu 1KB-90KB, što rezultuje čitanjem i pisanjem približno 140MB podataka. Ova suma prevazilazi količinu sistemske memorije i generalno eliminiše efekte keširanja diskova.

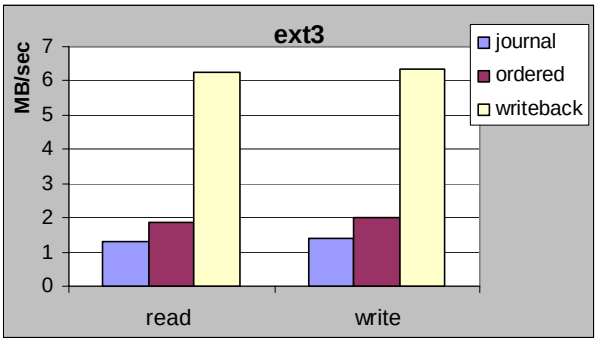
PostMark konfiguracija:

- set size 1000 90000
- set number 2000
- set transactions 50000

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

Tabela 3. *Rezultati prvog testa*

datoteka/sec	**journal**	**ordered**	**writeback**
kreiranje dat.	2000	2000	500
kreiranje sa trans.	220	314	1038
čitanje	219	312	1035
upis u datoteku	216	308	1035
brisanje	2340	2340	336
brisanje sa trans.	217	310	1045
protok pri čitanju i upisu			
čitanje (MB/sec)	1.3	1.85	6.25
upis (MB/sec)	1.41	2	6.37



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

U ovom testu, performanse writeback režima su znatno veće u odnosu na ostale režime rada dnevnika.

2. test

U prvom testu je izvršeno 50000 transakcija nad skupom od 4000 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.

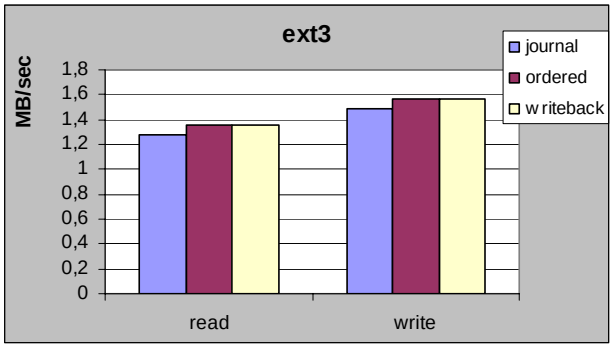
PostMark konfiguracija:

- set size 1000 300000
- set number 4000
- set transactions 50000

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

Tabela 4. *Rezultati drugog testa*

datoteka/sec	**journal**	**ordered**	**writeback**
kreiranje dat.	63	85	88
kreiranje sa trans.	7	7	7
čitanje	7	7	7
upis u datoteku	6	7	7
brisanje	154	166	160
brisanje sa trans.	7	7	7
protok pri čitanju i upisu			
čitanje (MB/sec)	1.28	1.36	1.36
upis (MB/sec)	1.49	1.57	1.57



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

Razlike u performanse režima vođenja dnevnika transakcija su male u slučaju intenzivnih testova, a **journal** se opet pokazuje kao najsporiji ražim rada.

3. test

U prvom testu je izvršeno 50000 transakcija nad velikim skupom slučajno generisanih datoteka čije se veličine kreću u opsegu 1bajt-1KB, što rezultuje čitanjem približno 9GB podataka i pisanjem približno 25MB podataka. Ovakva konfiguracija generiše veliki broj zahteva za ažuriranje meta-data oblasti, odnosno i-node tabele.

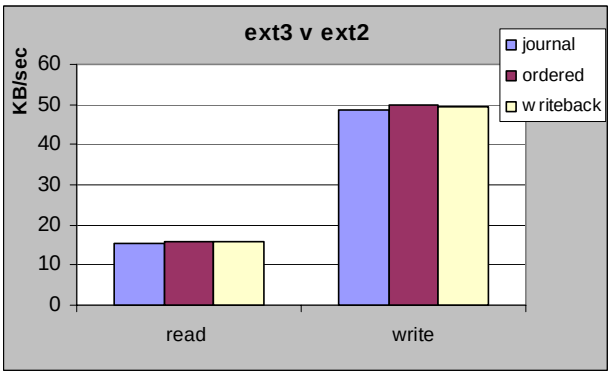
PostMark konfiguracija:

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

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

Tabela 5. *Rezultati trećeg testa*

datoteka/sec	**journal**	**ordered**	**writeback**
kreiranje dat.	277	256	256
kreiranje sa trans.	36	38	38
čitanje	36	38	37
upis u datoteku	36	38	38
brisanje	4992	4992	4992
brisanje sa trans.	36	38	38
protok pri čitanju i upisu			
čitanje (MB/sec)	15.44	15.92	15.71
upis (MB/sec)	48.51	50.04	49.36



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

7. ZAKLJUČAK

Fajl sistemi sa dnevnikom transakcija znatno povećavaju pouzdanost fajl sistema (umanjuju šanse gubitka integriteta fajl sistema) na račun performansi. Na osnovu izvršene procene performansi nad različitim skupovima slučajno generisanih datoteka u identičnom okruženju izveden je sledeći zaključak:

Najpouzadniji ext3 režim rada dnevnika transakcija pokazao je najgore performanse u svim testovima. Izuzetak predstavljaju interaktivni sekvencijalni testovi, detaljno opisani u tekstovima Daniel Robbinsa, koji nisu predmet ovog rada. Performanse ostala dva režima su solidne - **writeback** je pokazao najbolje rezultate u svim testovima, naročito pri radu sa manjom količinom podataka, dok je ordered pokazao idealan odnos povećanja pouzdanosti na račun performansi.

LITERATURA

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

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

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

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

[5] 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.

Abstract - This paper concentrates on the Linux ext3 filesystem performance comparison problem. Main goal this paper should achieve is analysis of performance impact due to a different journaling approaches, implemented in ext3 filesystem. The performance is measured using Postmark benchmark software, which emulates Internet mail server environment defined by the authors.

EXT3 FILESYSTEM JOURNALING MODES PERFORMANSE COMPARISION

Borislav Đorđević, Nemanja Maček, Dragan Pleskonjić, Stanislav Mišković, Borislav Krneta, Predrag Gavrilović