Ext4 sistem datoteka na RAID-0 konfiguracijama u Linux okruženju

Valentina Timčenko, Borislav Đorđević
Institut Mihajlo Pupin, Univerzitet u Beogradu
Beograd, Srbija
valentina@kondor.imp.bg.ac.rs, bora@impcomputers.com

Slobodan Obradović
VIPOS
Valjevo, Srbija
sobradovic@viser.edu.rs

Sadržaj—**U radu su predstavljeni rezultati dobijeni ispitivanjem performansi ext4 sistema datoteka pod Linux operativnim sistemom na kernel verziji 2.6. Detaljno je izložena komparacija performansi ext4 sistema datoteka sa prethodnicima, ext3 i ext2. Osim toga, ispitivano je ponašanje sva tri sistema datoteka u slučaju jednog diska (*single drive*) i posebno za slučaj instalirane RAID-0 konfiguracije. Performanse su merene primenom Postmark benchmark aplikacije.**

Ključne reči - Linux, sistemi podataka, ext4, RAID-0

I. UVOD

Linux je operativni sistem razvijan na osnovu slobodno dostupnog softvera i putem otvorenog koda. Besplatan je i za razliku od Windows operativnog sistema ima integrisane specifične funkcionalnosti kojima su ostvarene stabilnost, pouzdanost i sigurnost u radu, čineći ga sofisticiranim i moćnim operativnim sistemom. Novije verzije Linux kernela uključuju podršku za rad sa visoko performansnim *journaling* sistemima datoteka, poput ext4/ext3, ReiserFS, XFS i JFS [1].

RAID (*Redundant Array of Inexpensive Disks*, RAID) rešenje je zasnovano na kombinovanju više manjih, jeftinih diskova na specifičan način uređenih u niz diskovih polja. Osnovna karakteristika RAID rešenja su poboljšanje performansi kao i veći kapacitet u odnosu na jedan disk. Osnovna poboljšanja performansi kod RAID su zasnovana na uvođenju tehnike deljenja podataka između različitih diskova kao i paralelno čitanje i upisivanje na više diskova. Osim visoke pouzdanosti, RAID obezbeđuju teleranciju na greške (*fault-tolerance*) koji se postiže redundantnim skladištenjem podataka primenom tehnika *mirroring* ili tehnike parnosti diskova. RAID se definiše kroz sedam osnovnih nivoa arhitekture, počev od RAID-0 do RAID-6, pri čemu svaka od ovih arhitektura obezbeđuje *fault-tolerance*, poboljšanje performansi i niz drugih mogućnosti.

II. CILJ I MOTIV RADA

Cilj rada je da se na sistematičan način predstave rezultati dobijeni ispitivanjem performansi ext4 sistema datoteka, poređenjem ostvarenih performansi sa performansama prethodnih verzija, ext2 i ext3 sistema datoteka, kao i ispitivanjem ponašanja sva tri sistema datoteka na RAID-0 konfiguraciji. Motivaciju za ovaj rad smo pronašli u činjenici da novi ext4 sistem datoteka uključuje brojna poboljšanja u odnosu na svoje prethodnike, prvenstveno u odnosu na ext3, ali i da je glomazniji jer je 64-bitni sistem datoteka, pa su moguća iznenađenja u testovima performansi.

III. SISTEM DATOTEKA EXT3 I NJEGOVE MANE

Svaka nova verzija ext sistema podataka ima za cilj da ostvari kompatibilnost sa prethodnom verzijom, pa tako i ext3 u odnosu na ext2 ima sličnu struktura podataka. S druge strane, ta činjenica je onemogućila implementaciju nekih savremenih karakteristika u ext3 sistem podataka, poput: *extents*, dinamička dodela indeksnih čvorova i *tail packing* (omogućava grupisanje krajnjih parcijalnih blokova više fajlova u jedan veći blok koji će moći da se iskoristi. Ovi blokovi se inače ne bi mogli koristiti zbog pojave interne fragmentacije) [2].

Mnogi Linux fajl sistemi, među njima i ext3, ne mogu biti provereni na greške sa softverskim alatom *fsck* (služi za proveru konzistencije fajl sistema), kada je sistem mauntovan već samo u trenucima podizanja sistema. Pokušaj da se izvrši provera sistema podataka koji je mauntovan može rezultirati detekcijom grešaka u izmenjenim podacima koji još uvek nisu zapisani na disk, čineći sistem podataka nekonzistentnim.

Osim toga, u ext3 sistemu podataka ne postoji *online* skup alata za defragmentaciju sistema podataka. Postoji *offline* ext2 defragmenter, *e2defrag*, ali da bi se on koristio potrebno je ext3 sistem podataka prvo konvertovati u ext2. Takođe, postoji opasnost da, kada se koristi programski alat *e2defrag* na ext3 sistemu podataka, potpuno uništi pojedine podatke usled nepoznavanja načina postupanja sa novim karakteristikama zastupljenim u ext3. Ipak, za defragmentaciju postoje specijalni programski alati, *shake* i *defrag*, pri čemu *shake* vrši alociranje slobodnog prostora za fajl tako što tu operaciju izvršava kao jednu operaciju, i time obezbeđuje da alokator pronađe kontinualni slobodni prostor. S druge strane, *defrag* kopira fajl preko istog fajla. Međutim, oba softverska alata mogu uspešno da funkcionišu samo kada je fajl sistem relativno prazan.

U literaturi se za moderni Linux sistem podataka navodi da pojavu fragmentacije drži na minimumu tako što sve blokove jednog fajla nastoji da smešta jedan blizu drugog, čak i u slučaju kada ne postoji mogućnost da ih smesti u uzastopne sektore. Iako je ext3 fajl sistem dosta otporniji na pojavu fragmentacije od FAT fajl sistema, i u ext3 se može vremenom pojaviti fragmentacija što se može primetiti ukoliko mu treba dosta vremena da upiše neke veće fajlove. Još jedna mana

dizajna ext3 je nemogućnost oporavka već izbrisanih fajlova, pri čemu ext3 *driver* trajno briše fajlove tako što briše njihove indeksne čvorove. Rešenje ovog problema potencijalno može da se obezbedi primenom novih tehnika i komercijalno dostupnih softverskih alata (npr.*UFS Explorer Standard Recovery version 4*) koje mogu da vrate izbrisane ili izgubljene fajlove koristeći se analizom dnevnika transakcija fajl sistema, ali, ne postoji nikakva garancija da će operacija biti uspešna.

IV. SISTEM DATOTEKA EXT4

Ext4 fajl sistem je po mnogim karakteristikama logički naslednik ext3 fajl sistema. Podržan je na većini danas korišćenih Linux distribucija. Za razliku od ext3, koji ima samo neke karakteristike dodate u odnosu na ext2 uz zadržavanje iste strukture podataka, ext4 uvodi dublje promene u odnosu na ext3, posebno sa aspekta strukture podataka, čineći ga pouzdanijim sistemom podataka. Osim toga, ext4 je u odnosu na ext3 je 64-bitni sistem podataka, čime veličina jednog fajla dostiže veličinu i do 16 TB [3], [4], [5], [6].

Jedna od najpozitivnijih karakteristika ext4 je kompatibilnost sa prethodnim ext2 i ext3 sistemima podataka. Moguće je instalirati ext2/ext3 sistem podataka, izmeniti nekoliko opcija i koristiti ga kao ext4. Postojeći podaci se neće izgubiti i ext4 fajl sistem će primenjivati nove strukture podataka i karakteristike na samo od tada novim podacima. Međutim, bez obzira na ova poboljšanja, preporuka je da se uvek obavi dodatno čuvanje pdoataka u vidu backup-a, posebno zbog ograničene kompatibilnosti ext4 i ext3 sistema podataka. Ova ograničenost se ogleda u tome da nije uvek moguće koristiti ext4 i mauntovati ga kao ext3 fajl sistem jer su im strukture podataka veoma različite.

Ext4 sistem podataka ima brojne nove karakteristike i tehnike koji nisu bile zastupljene u ext3: e*xtents*, *journaling checksumming*, zatim istovremena alokacija više blokova, odložena alokacija, brži *fsck*, *on-line* defragmentacija i veće veličine direktorijuma, koji mogu sadržati i do 64000 fajlova.

V. KONFIGURACIJA DISKOVA RAID-0

Na RAID-0 nivou podaci su raširenі ravnomerno preko N diskova deljenjem podataka u blokove, čime se dobija na povećanoj brzini prenosa podataka. Svi upisi i čitanja podataka se paralelno obavljaju sa svih N diskova. Ukoliko su magistrale diskova dovoljno brze ovakva procedura može dovesti do N-tostrukog poboljšanja performansi. S obzirom na to da se ne vrši redudantno skladištenje podataka, performanse su dobre međutim kvar na bilo kojem od diskova može dovesti do gubitaka podataka. U slučaju kvara diska izgubljeni podaci neće moći da se povrate. Ovaj nivo se u praksi često naziva i "*striping*" nivo. RAID-0 je najbrži i najefikasniji RAID niz ali ne nudi nikakvu zaštitu od grešaka ili kvarova [7], [8], [9].

VI. USLOVI I PRETPOSTAVKE ANALIZE

Pod uslovima i pretpostavkama podrazumeva se konfiguracija harvera, opis operativnog sistema i radnog okruženja kao i definisanje uslova testiranja i analize rezultata.

Hardverska konfiguracija se sastoji od nekoliko osnovnih komponeneti, predstavljenih u tabeli I. Za testiranje su odabrani diskovi iz serije HP SAS 10K.

TABELA I. HARDVERSKA KONFIGURACIJA

Server	HP Proliant ML350 G6
RAM	12 GB
Procesori	Intel(R)Xeon(R)
CPU Model	Quad-core E5506@2.13GHz
Broj jezgara CPU	4
CPU brzina	2133 MHz
L3 keš	12 MB
Kontroleri	
RAID	HP Smart Array P410i SAS
RAID keš memorija	256MB
Disk (*DualPort*)	HP Invent SAS 10K, 146GB,2.5"
Operativni sistem	Linux Fedora 13, kernel- 2.6.33.3-85.fc13

U pitanju su 3Gb SAS diskovi, veličine 2.5 inča, i sa kapacitetom od 146GB (Tabela II).

TABELA II. KARAKTERISTIKE DISKA

HP Invent SAS 10K, 146GB, 2.5", Hot-swap HD	
Kapacitet	146GB
Interfejs	SAS plug
Srednje vreme pozicioniranja	4 ms
Full stroke pretraživanje	8.1msec
Track-to-track pretraživanje	0.2msec
Rotaciona brzina	10,000 rpm
Maksimalna brzina bafera	6Gb/sec

Što se tiče operativnog sistema, izabrali smo jednu od najzastupljenijih Linux distribucija za PC arhitekture, Red Hat Linux Fedora 13 sa kernel verzijom 2.6.33.3-85.fc13.

RAID-0, formiran je u 3 različite konfiguracije, RAID-0 formiran od 2 diska R0-2, RAID-0 formiran od 3 diska R0-3 i RAID-0 formiran od 4 diska R0-4. Sistem podataka je organizovan u vidu logičkih particija (tabela III) [8, 10].

TABELA III. ORGANIZACIJA FAJL SISTEMA

Sistem podataka	***Veličina***	***Instaliran na***
LogVol00	50GB	/ root FS
LogVo10l	5GB	/ swap
……		
LogVol03	10GB	/ testing FS

Swap je definisan kao 5GB particija i realizovan je u vidu logičke grupe LogVol01, koja se na testiranom sistemu može naći prateći putanju /dev/mapper/VolGroup00-LogVol01. Prazan ext3 fajl sistem je kreiran u logičkoj grupi LogVol02, koja služi za potrebe testiranja i do nje se može doći putanjom /dev/mapper/VolGroup00-LogVol03. Fajl sistem korišćen za testiranje je iste veličine od 10GB za sve testove i testirane sisteme datoteka.

VII. DETALJNA ANALIZA

Za potrebe ovog rada korišćen je PostMark [11] softver koji simulira opterećenje Internet Mail servera. PostMark kreira veliki inicijalni skup (*pool*) slučajno generisanih datoteka na bilo kom mestu u sistemu podataka. 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.

A. Postmark test1(male datoteke)

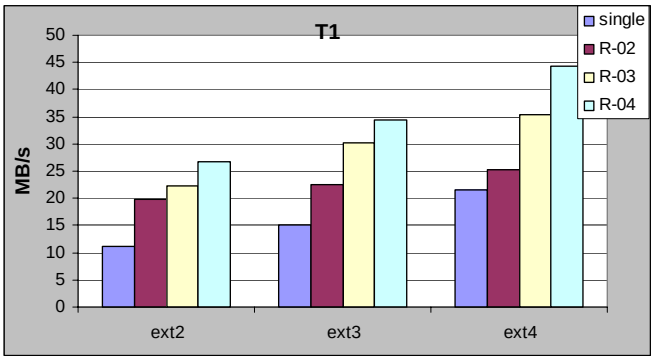
Datoteke zastupljene u ovom testiranju su relativno male, u opsegu od 1KB do 100KB. PostMark kofiguracija je:

- veličina pojedinačne datoteke 1000 100000
- broj kreiranih datoteka 4000
- broj izvršenih transakcija 50000

Ovakva konfiguracija generiše oko 1.6GB podataka za upis i čitanje, čime se ostvaruje se veliki broj I/O zahteva. Rezultati testa su predstavljeni na Sl. 1 i u tabeli IV:

TABELA IV. REZULTATI ZA POSTMARK TEST 1

MB/s	*ext2*	*ext3*	*ext4*
Single	11,02	15,09	21,58
R-02	19,92	22,56	25,24
R-03	22,20	30,20	35,32
R-04	26,79	34,48	44,40



Slika 1. Rezultati za Postmark test 1

U ovom testu malih datoteka, novi ext4 sistem datoteka pokazuje superiorne performanse u odnosu na svoja dva prethodnika. Sistem datoteka ext4 je 10-40% brži od ext3, dok je 30-100% brži od ext2, sistem datoteka ext3 je oko 10-35% brži od ext2. Najveće razlike između sistema podataka detektovane su na *single drive* konfiguraciji, dok su na RAID konfiguracijama razlike između testiranih sistema podataka manje. Uvođenje RAID konfiguracija i povećanje broja diskova u njima donosi različita ubrzavanja na različitim sistemima datoteka, od 10-80% .

B. Postmark test2 (ultra male datoteke)

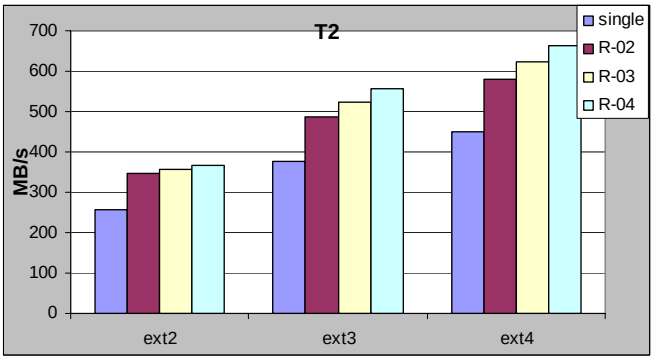
Ovo je takođe veoma intenzivan test jer uključuje veliki broj izuzetno malih datoteka, veličina u rasponu od 1bajta do 1KB. PostMark konfiguracija je:

- veličina pojedinačne datoteke 1 1000
- broj kreiranih datoteka 30000
- broj izvršenih transakcija 50000

Broj datoteka koji se kreira je povećan na 30000, čime se čita oko 14MB podataka i upisuje 32MB. Test generiše veliki broj metadata I/O zahteva. Rezultati testa su predstavljeni na Sl. 2 i u tabeli V:

TABELA V. REZULTATI ZA POSTMARK TEST 2

KB/s	*ext2*	*ext3*	*ext4*
single	258,02	377,53	449,45
R-02	348,33	487,65	580,54
R-03	357,26	522,65	622,21
R-04	366,66	557,32	663,48



Slika 2. Rezultati za Postmark test 2

U ovom testu ultra malih datoteka, novi ext4 sistem datoteka je takođe brži u odnosu na svoja dva prethodnika, pri čemu su ovog puta razlike manje nego u prethodnom testu. Sistem datoteka ext4 je 20% brži od ext3, dok je oko 70% brži od ext2, a sistem datoteka ext3 je oko 40% brži od ext2. Ovog puta razlike između FS detektovane na single drajv konfiguraciji i na RAID konfiguracijama su ujednačene. Uvođenje RAID konfiguracija i povećanje broja diskova u njima donosi različita ubrzavanja na različitim sistemima datoteka, od 3-30%.

C. Postmark test3 (veće datoteke)

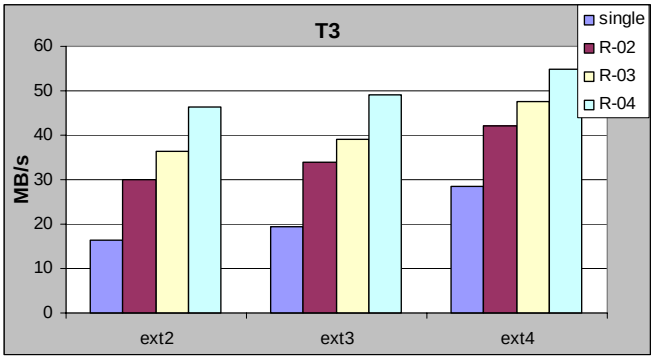
Ovo je veoma intenzivan test. PostMark konfiguracija:

- veličina pojedinačne datoteke 1000 300000
- broj kreiranih datoteka 4000
- broj izvršenih transakcija 50000

Ukupna količina podataka za čitanje, 4.7GB, i upis, 5.4GB, je znatno veća od količine sistemske memorije, u potpunosti eliminiše efekte svih mehanizama keširanja i proizvodi veliki broj I/O zahteva. Rezultati testa su prikazani na Sl.3 i tabeli VI:

TABELA VI. REZULTATI ZA POSTMARK TEST 3

MB/s	*ext2*	*ext3*	*ext4*
single	16,38	19,45	28,46
R-02	30,12	33,83	42,05
R-03	36,47	39,23	47,63
R-04	46,22	49,17	54,92



Slika 3. Rezultati za Postmark test 3

U ovom testu novi ext4 sistem datoteka nastavlja da pokazuje superiorne performanse u odnosu na svoja dva prethodnika. Sistem datoteka ext4 je 10-46% brži od ext3, dok je 20-70% brži od ext2, sistem datoteka ext3 je 6-18% brži od ext2. Kao i u prvom testu, najveće razlike između sistema podataka detektovane su na *single drive* konfiguraciji, dok su na RAID konfiguracijama razlike između sistema podataka manje. Uvođenje RAID konfiguracija i povećanje broja diskova u njima donosi ubrzanja u radu različitih sistema datoteka, od 15-80% , pri čemu je najveće ubrzanje postigla RAID-0 sa dva diska u odnosu na *single drive*.

VIII. ZAKLJUČAK

U ovom radu smo sumirali rezultate PostMark testova za tri bliska i kompatibilna sistema datoteka ext2, ext3 i ext4. Rezultati testova potvrđuju očekivanja. Novi 64-bitni sistem datoteka ext4 pokazuje superiorne osobine u odnosu na svoje prethodnike, ext2 i ext3, osetno ih pobeđujući u sva tri testa performansi. Ekstenti, brojne inovacione tehnike za alokaciju datoteka, poboljšana *journaling* tehnika i poboljšani bafeski keš mehanizam, doveli su do toga da ext4 pobeđuje svoje prethodnike u teškim uslovima, kao što je rad sa malim datotekama, što je maksimalno zastupljeno u našim testovima. Interesantno je da je ext3 bio bolji od ext2 u svim testovima, što znači *journaling* tehnika u kombinaciji sa keš mehanizmom, ne samo da ne usporava sistem, već poboljšava performanse. Ovim rezultatima ohrabrujemo sve korisnike Linux sistema da koriste novi ext4 sistem datoteka.

Drugi bitan zaključak je da postoji značajno ubrzanje operacija upisa i čitanja koje donosi RAID-0, od 30-80%, a više od dva puta u odnosu na *single drive* konfiguraciju. Konkretno ubrzanje za RAID-0 od dva, tri ili četiri diska dosta zavisi od vrste fajl sistema (ext2/ext3/ext4), veličine datoteka kao i ciklusa upisa koje nameće *journaling* tehnika. Najveći prinos ubrzanja je detektovan između RAID-0 sa dva diska u odnosu na *single drive* konfiguraciju.

ZAHVALNICA

Rad je finansiran od strane Ministarstva prosvete i nauke Republike Srbije (Projekti TR 032025 i TR 032037)

REFERENCE

[1] 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., San Diego, June 2000. pp. 71-84

[2] Tweedie S., "EXT3, Journaling Filesystem", July 2000.

[3] A. Mathur, M. Cao, S. Bhattacharya, A. Dilger, A. Tomas, L. Vivier, "The new ext4 filesystem: current status and future plans" in Proceedings of the Linux Symposium, Ottawa, Canada, June 2007.

[4] Roderick W. Smith, "Migrating to Ext4". DeveloperWorks. IBM, http://www.ibm.com/developerworks/linux/library/l-ext4/, April 2008.

[5] "Ext4 Howto", [Online]. https://ext4.wiki.kernel.org/index.php/Ext4_Howto#Bigger_File_System_and_File_Sizes, January 2011.

[6] „First benchmarks of ext4", [Online]. http://www.linuxinsight.com/first_benchmarks_of_the_ext4_file_system.html, October 2006.

[7] B. Baude, "RAID on Linux on POWER", IBM eServer Solutions Enablement, November 2005.

[8] A. Thomasian, J. Xu, „Reliability and Performance of Mirrored Disk Organizations" Computer Journal, January 2008.

[9] A. Lebrecht, N. Dingle, W. Knottenbel, „Analytical and Simulation Modelling of Zoned RAID Systems" in Computer Journal, June 2010.

[10] V. Danen, "Set up Logical Volume Manager in Linux", March 2007.

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

ABSTRACT

This paper represents a performance evaluation of new filesystem ext4 under the Linux Operating System on the kernel version 2.6. Paper includes a performance comparison of the ext4 file system with its predecessors, ext2 and ext3. This paper also examines the behavior of all three filesystems on the single drive and RAID-0 configurations. The performance is measured using Postmark benchmark software application.

EXT4 FILESYSTEM ON THE RAID-0 CONFIGURATION UNDER LINUX

Valentina Timčenko, Borislav Đorđević, Slobodan Obradović