POREĐENJE PERFORMANSI RAID-0 I RAID-5 U LINUX OKRUŽENJU POMOĆU TIOBENH BENHMARKA
EXAMINATION OF HARDWARE RAID SOLUTIONS UNDER LINUX: RAID-0 V RAID-5 USING BY TIOBENH

Đorđević Borislav, Institut Mihajlo Pupin, Beograd

Valentina Timčenko, Institut Mihajlo Pupin, Beograd

Slobodan Obradović, VIPOS, Valjevo

Sadržaj – *Rad predstavlja ispitivanje hardverskih RAID rešenja pod Linux operativnim sistemom. Rad uključuje komparaciju performansi dva popularna RAID nivoa, RAID-0 i RAID-5 pod Linux kernel verzijom 2.6. Performanse se mere korišćenjem Tiobenh benchmark programa..*

Abstract - *This paper represents an evaluation of hardware RAID solutions under the Linux Operating System. Paper includes a performance comparison of two very popular RAID levels such as RAID-0 and RAID-5 under Linux kernel version 2.6. The performance is measured using Tiobench benchmark software application*

1. UVOD

RAID je kombinacija više manjih, jeftinih diskova uređenih u niz diskovih polja. Osnovna karakteristika RAID rešenja su poboljšanje performansi kao i veći kapacitet u odnosu na jedan disk. Osnova poboljšanja performansi kod RAID je tehnika deljenja podataka između različitih diskova kao i paralelno čitanje i upisivanje na više diskova [1].

Osim toga, RAID se odlikuje visokom pouzdanošću a *fault-tolerance* se postiže preko redundantnog skladištenja podataka primenom tehnika *mirroring* ili tehnike parnosti diskova. RAID se definiše kroz šest osnovnih nivoa arhitekture, počev od RAID-1 do RAID-6, pri čemu svaka od ovih arhitektura obezbeđuje *fault-tolerance*, poboljšanje performansi kao i niz drugih mogućnosti. Ovi nivoi su osnova za mnoge druge ugnežđene RAIDX+Y nivoe koji se mogu realizovati. Ugnežđeni RAID1+0 nivo je trenutno najpopularniji među RAID konfiguracijama. Postoje dva pristupa realizaciji RAID arhitektura: hardverska (Hw-RAID) i softverska (Sw-RAID). Svaka od ovih realizacija nudi specifične mogućnosti i performanse.

U poslednjih deset godina došlo je velikih poboljšanja u arhitekturi računara a s aspekta RAID tehnologije to se ogleda u povećanoj brzini okretanja diskova, od 10K-15K okretaja, skraćenim vremenom pozicioniranja, poboljšanim mehaničkim kašnjenima i boljim tehnikama keš baferisanja. Osim toga, na tržištu se mogu naći visoko kvalitetni disk interfejsi, naročito su cenjeni SAS i FC, kojima se postiže prenos podataka od 3 Gb/sec ili 4 Gb/sec. Takođe su poboljšane brze sistemske magistrale kao što su paralelna PCI i serijska PCI-express. Paralelna PCI magistrala je 64-bit sistemska magistrala zasnovana na taktu od 64MHz sa teorijski mogućim protokom od 533 MB/sec. PCI express je serijska magistrala komutacionog tipa sa visoko kvalitetnim gigabitnim linkovima.

Linux je moderan, sofisticiran i moćan operativni sistem. Novije verzije Linux kernela uključuju podršku za rad sa visoko performansnim journaling sistemima datoteka, poput 32bitnih ext3 i ReiserFS, i 64bitnih XFS i JFS sistema datoteka razvijenih u Silicon Graphics i IBM laboratorijama respektivno[2], [3], [4], [5].

2. CILJ I MOTIV RADA

Cilj ovog rada je da se opiše rad i performanse dva različita RAID nivoa, kao i da se uporede njihove brzine rada u odnosu na rešenja koja nisu zasnovana na RAID tehnologiji. U odnosu na rešenja koja ne podrazumevaju upotrebu RAID nivoa, RAID nudi visoku pouzdanost i *hot-swap* mogućnosti, tj. "vruću zamenu" diskova. Osim toga, mogućnost paralelnog upisa i čitanja sa više diskova dodatno poboljšava kvalitet performansi RAID rešenja. RAID tehnologija se zasniva na implementaciji dve tehnike obrade podataka "Parity" i "Mirror" (check/generate) koje se primenjuju na više diskova istovremeno i ne mogu se naći u slučaju ostalih, Non-RAID tehnologija. Naše rešenje se fokusira na evaluaciji performansi RAID i Non-RAID konfiguracija pod Linux operativnim sistemom. Najpre smo izabrali server na kojem smo instalirali RAID-SAS kontroler zatim smo odabrali RAID konfiguraciju i podesili parametre na željene vrednosti. Posebna pažna je posvećena testiranju različitih RAID nivoa u *fer-play* uslovima a koji podrazumevaju testiranje pod istim ili sličnim uslovima, u istom hardverskom okruženju i pod istim operativnim sistemom. Iste veličina fajl sistema za sve testirane nivoe omogućava jednake uslove pri korišćenju slobodnog prostora na diskovima (oko 50GB). Procedura testiranja podrazumeva da se posebna pažnja usmeri na sledeće elemente: (a) operativni sistem (b) interfejsi diskova (c) Hardverski RAID disk kontroler (d) RAID- nivo i (e) diskovi.

3. UVOD U RAID TEHNOLOGIJU

Deljenje podataka – Data Striping

"Striping" je jedna od fundamentalnih prednosti RAID tehnologije nad drugim. Predstavlja metod kojim se više diskova povezuje u jednu logičku disk jedinicu. A Striping

podrazumeva deljenje svakog diska u delove koji se nazivaju strajpovi (*stripes*), čija najmanja veličina može biti veličine jednog sektora, (512 bajtova) a može biti i do veličine nekoliko megabajtova. Strajpovi se zatim raspoređuju po različitim diskovima tako da je ukupan prostor za skladištenje organizovan u *round-robin* stilu, pri čemu se strajpovi podataka kružno redom čuvaju svaki na sledećem dostupnom disku. Prostor za skladištenje ovim poprilično podseća na špil izmešanih karata. Aplikativno okruženje definiše veličinu korišćenih strajpova u sistemu.

RAID-0
Na RAID0 nivou podaci su raširenи 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 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. RAID0 je najbrži i najefikasniji RAID niz ali ne nudi nikakvu zaštitu od grešaka ili kvarova.
RAID-5
RAID Nivo 5 obezbeđuje redudansu time što informaciju o parnosti upisuje na sve diskove Nivo 5 obezbeđuje nešto veće brzine čitanja podataka i nešto sporije upisivanje podataka nego što je to moguće u radu sa samo jednim diskom, ali ukoliko dođe do kvara nekog od diskova podaci neće biti izgubljeni. Uklanjanem ili kvarom jednog od N diskova informacije će i dalje biti sačuvane. RAID 5 obavlja striping tehniku na nivou blokova, a parnost je distrubuirana kroz sve diskove. RAID 5 je odličan za sve vrste transfera, osim za male upise, gde ispoljava jako loše performanse.

4. EXT3
Postoji veliki broj trenutno dostupnih Linux journaling fajl sistema[8]. Najpoznatiji je XFS, journaling fajl sistem razvijen u kompaniji Silicon Graphics a zatim je postao dostupan i kao open source fajl sistem. ReiserFS, je takođe jedan od popularnijih fajl sistema razvijen isključivo za Linux okruženje, JFS prvenstveno namenjen na IBM mašinama a zatim kao *open source* postao dostupan i za Linux. Možda najčešće primenjivani je ext3 fajl sistem implementiran na Red Hat i drugim Linux distribucijama. ext3 fajl sistem je journaling verzija Linuxovog ext2 fajl sistema. U potpunosti je kompatibilan sa svojim predhodnikom, ext2. Ext3 fajl sistem je podržan na 2.4.16 i novijim verzijama kernela ali da bi u potpunosti došao do izražaja neophodno ga je aktivirati kroz konfiguracioni terminal fajl sistema pri podizanju kernela. Linux distribucije Red Hat 7.2 i SUSE 7.3 uključuju podršku za ext3 fajl sistem. Ext3 se može koristiti jedino ukoliko je podrška za ext3 kompajlirana u okviru željenog kernela i ukoliko je primenjena najnovija verzija e2fsprogs Linux alatki.

5. USLOVI I PRETPOSTAVKE

Konfiguracija sistema
Pod uslovima i pretpostavkama ćemo podrazumevati konfiguraciju harvera, opis operativnog sistema i radnog okruženja kao i definisanje uslova kod kojim izvodimo testiranja [9]. Hardverska konfiguracija se sastoji od nekoliko osnovnih komponeneti: (a) CPU i njegova učestanost (b) matična ploča servera (c) sistemska RAM memorija (d) drugonivovska CPU keš memorija (e) RAID disk kontroler (f) RAID nivo i (h) diskovi. Svi testovi predstavljeni u ovom radu su obavljeni na konfiguraciji sistema tabelarno predstavljeno u tabeli 5.1.:

Server	HP Proliant ML350 G5
RAM	4 GB
Procesori	Intel(R)Xeon(R)
CPU Model	Quad-core 410@2.33GHz
Nmbr. of CPU cores	4
CPU brzina	2333MHz
L2 keš	2 x 6 MB
Kontroleri	
RAID	HP SmartArray E200i (SAS)
RAID keš memorija	128MB
Disk (jedan port)	HP SAS 10K, 146GB, 2.5" t
Operativni sistem	Linux Fedora 8, 2.6.23.1-42

Tabela 5.1 Konfiguracija sistema

Diskovi
Za potrebe testiranja smo izabrali diskove serije HP SAS 10K. U pitanju su 3Gb SAS diskovi, veličine 2.5 inča, i sa kapacitetom od 146GB (Tabela 5.2):

HP SAS 10K, 146GB, 2.5"Single Port, Hot-swap HD	
kapacitet	146GB
interfejs	SAS plug
srednje vreme pozicioniranja	4ms
full stroke pretraživanje	8.1msec
track-to-track pretraživanje	0.2msec
rotaciona brzina	10,000 rpm
Maks. brzina bafera	3Gb/sec

Tabela 5.2 Diskovi korišćeni pri izvođenju testiranja

Osobine operativnog sistema i fajl sistema
Izabrana je Red Hat Linux verzija i Fedora 8 sa kernel verzijom 2.6.23.1-42. Ovo je jedna od najpopularnijih Linux distribucija za PC arhitekture. Za potrebe testiranja, fajlsistem je organizovan u vidu logičkih particija [10] kao što je predstavljeno u tabeli 5.3. Swap je definisan kao 2GB swap particija i realizovan u vidu logičke grupe (logical volume group), koju smo imenovali LogVol01. Grupa LogVol01 se može naći na testiranom sistemu prateći putanju /dev/mapper/VolGroup00-LogVol01.

Filesystem	**Size**		**Mounted on**
LogVol00	90-530GB	/	root FS
LogVo10l	2GB		swap
LogVol02	50GB	/test	testing FS

Tabela 5.3 FS layout

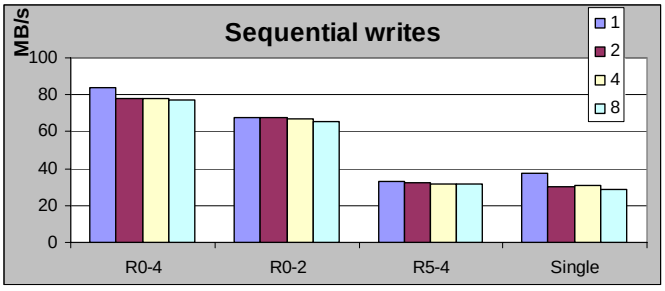
Prazan ext3 fajl sistem je kreiran u logičkoj grupi LogVol02, specijalno kreiranoj za potrebe testiranja i do nje se može doći putanjom /dev/mapper/VolGroup00-LogVol02. Fajl sistem koji koristimo za testiranje je u potpunosti iste veličine za sve testove i testirane RAID nivoe.

6. DETALJNA ANALIZA

Tiobench je višenitni I/O benchmark za Linux (ili bilo koji UNIX sistem sa podrškom za POSIX threads biblioteku). Autor benchmarka je Mika Kuoppala. Program tiobench može izršavati predefinisane ili konfigurabilne testove [11].
U našem slučaju, izvršavali smo test "tiobench.pl - t 1 -f 1792 -r 4000 - b 4096", pri čemu je f veličina datoteke, r je broj prolazaka i b je veličina bloka. Izdvojili smo rezultate za sequential i random write test. Prva kolona predstavlja veličinu disk bloka za testiranje. Rezultati for sequential write test su prikazani u tabeli 6.1 i slici 6.1:

MB/s	R0-4	R0-2	R5-4	Single
1	83.65	67.87	33.33	37.18
2	78.24	67.46	32.31	30.44
4	77.6	66.6	31.66	30.74
8	76.99	65.15	31.78	28.52

Tabela 6.1 Rezultati za Tiobench sequential writes test

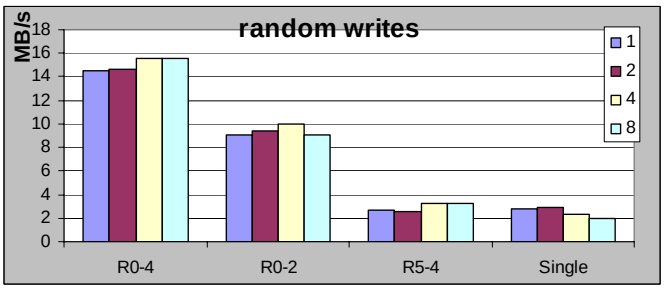


Slika 6.1 Rezultati za Tiobench sequential writes test

Konfiguracija RAID0-4 (RAID-0 kreiran od 4 diska) pokazuje superiorne performanse u ovom testu. RAID0-4 je od 17-21% brži od RAID0-2 (kreiran od 2 diska). RAID0-4 je oko150% brži od od RAID5 (RAID-5 kreiran od 4 diska) i oko 150% brži od jednog diska.
RAID0-2 je solidno brža od RAID5 (oko 100%) a i oko 150% brži od jednog diska. RAID5 i jedan disk imaju slične performanse za ovakvu vrstu testa.Rezultati for sequential write test su prikazani u tabeli 6.2 i slici 6.2:

MB/s	R0-4	R0-2	R5-4	Single
1	14.51	9.08	2.66	2.8
2	14.64	9.37	2.6	2.88
4	15.53	9.97	3.25	2.36
8	15.52	9.08	3.21	2.01

Tabela 6.2 Rezultati za Tiobench random writes test



Slika 6.2 Rezultati za Tiobench random writes test

Konfiguracija RAID0-4 (RAID-0 kreiran od 4 diska) pokazuje superiorne performanse u ovom testu. RAID0-4 je od 50-70% brži od RAID0-2 (kreiran od 2 diska). RAID0-4 je oko 5 puta brži od od RAID5 (RAID-5 kreiran od 4 diska) i oko 5-6 puta brži od jednog diska. RAID0-2 je solidno brža od RAID5 (oko 3 puta%) a i oko 3-4 brži od jednog diska. RAID5 i jedan disk imaju slične performanse za ovakvu vrstu testa

7. ZAKLJUČAK

Zapažamo da je razlika u performansama između različitih RAID nivoa, manja u sequential *write* testu, a veoma dominantna u random write testu, gde je RAID 5 osobito loš.
RAID 0 je, kao što se očekivalo, ubedljivo najbolja konfiguracija, zato što nema nikakve dodatne cikluse upisa. Primetili smo da u random write testu, performanse RAID-0, rastu skoro linearno sa povećanjem broja diskova (na primer, RAID 0-4 je uglavnom 2 puta brži od RAID 0-2).
U Tiobench testu, RAID 5 je drastično sporiji od RAID 0 konfiguracija, a dosta sličan jednom disku.
U većini testova, osobito u Random, male datoteke i prenos malih količina podataka sa čestim upisima (*small updates*) su dominantne i to je fatalno za performanse RAID-5.
Razlog za takvo ponašanje RAID 5 je čuveni "*small updates*" problem. Taj performansni nedostatak čini RAID-5 veoma osetljivim na cikluse u kojima dominiraju mali upisi, što se u našem slučaju očigledno dogodilo. Ima više tehnika, koje mogu da poboljšaju performanse za RAID-5 u slučaju malih upisa: keširanje parnosti i parity log. Naši testovi pokazuju da je RAID 5 do 5 puta sporiji u odnosu RAID 0, a to znači da količina RAID keša od 128MB na RAID-SAS kontroleru, nije bila dovoljna da kompenzuje small write problem.

7. LITERATURA

[1] B. Baude "RAID on Linux on POWER", IBM eServer Solutions Enablement, nov2005
[2] Tweedie S., "EXT3, Journaling Filesystem" 20 July, 2000
[3] Bill von Hagen, "Exploring the ext3 Filesystem", April 5, 2002
[4] Robbins D., "Introducing ext3", Gentoo Technologies, Inc., Updated October 9, 2005
[5] Robbins D., "Surprises in ext3", Gentoo Technologies, Inc. 1 December 2001
[6] The Software-RAID HOWTO, Jakob Østergaard and Emilio BuesoApril 2004
[7] Hardware RAID vs. Software RAID: Which Implementation is Best for my Application?, Company: Adaptec Published: January 2008
[8] 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.
[9] A. Silberschatz, P. Galvin, Operating System Concepts. Addison-Wesley, 2007
[10] lvm V. Danen "Set up Logical Volume Manager in Linux", Mar 09 2007
[11] [http://sourceforge.net/projects/tiobench/