Modelovanje i poređenje RAID performansi na hardverskom motherboard-RAID kontroleru

Borislav Đorđević,Valentina Timčenko, Slavica Boštjančič Rakas
Univerzitet u Beogradu, Institut Mihajlo Pupin, Beograd, Srbija
bora@impcomputers.com, valentina.timcenko@pupin.rs, slavica.bostjancic@pupin.rs

Nemanja Maček, Milan Kuprešanin
VISER
Beograd, Srbija
stonerhate@gmail.com, milan.kupresanin@gmail.com

***Sadržaj*—U radu su predstavljeni modelovanje i rezultati merenja performansi nekoliko različitih RAID konfiguracija pod MS Windows operativnim sistemom, na hardverskom motherboard-RAID kontroleru. Rad uključuje matematičko modelovanje RAID performansi, a model se validira kroz tumačenje dobijenih vrednosti za skup mera performansi. Performanse se mere korišćenjem HD Tune Pro benchmarka, za single disk i tri konfiguracije RAID arhitekture, RAID-0 RAID-1 i RAID-5, pri čemu se varira broj diskova od kojih se konfiguriše striping RAID-0.**

Ključne reči – RAID, NTFS, MS Windows, sistemi datoteka, performanse, HD Tune Pro

I. UVOD

RAID (*Redundant Array of Independent Disks*) je kombinacija više manjih hard diskova uređenih u niz, odnosno u polje diskova. Osnovne RAID prednosti su poboljšanje performansi, povećanje kapaciteta i povećanje pouzdanosti u odnosu na pojedinačni disk (*single disk*). Osnova poboljšanja performansi kod RAID sistema je u primeni *Striping* tehnike, kojom se delovi datoteke raspoređuju na više različitih diskova, a omogućeno je paralelno čitanje i upisivanje na više diskova. RAID se odlikuje visokom pouzdanošću koja se postiže redundantnim skladištenjem podataka primenom tehnike ogledala (*Mirroring*) ili tehnike parnosti. RAID se definiše kroz 7 osnovnih nivoa, počev od RAID-0 do RAID-6, pri čemu svaki od ovih nivoa obezbeđuje poboljšanje performansi na specifičan način, a u nekim slučajevima se kroz tehnike zaštite obezbeđuje i tolerancija na greške i otkaze diskova (*fault-tolerance*).

Cilj rada je da se postavi sopstveni matematički model osnovnih RAID nivoa od interesa (0, 1 i 5), a potom da se na osnovu njega ispitaju dobijene razlike njihovih performansi u slučaju NTFS (*New Technology File System*) [1, 2] fajl sistema na MS Windows operativnom sistemu, čime se obavlja validacija modela.

II. UVOD U RAID TEHNOLOGIJU

A. RAID-0 Striping level

Na RAID-0 nivou [3-5], deljenjem podataka u trake (*stripes*), podaci su raspoređeni ravnomerno na N diskova. Svi upisi i čitanja podataka se paralelno obavljaju sa svih diskova, čime se dobija na povećanju brzine prenosa podataka. Ukoliko su magistrale diskova dovoljno brze ovakva procedura može dovesti do N-tostrukog poboljšanja performansi. S obzirom da se ne vrši redudantno skladištenje podataka, otkaz bilo kog diska dovodi do gubitka podataka. RAID-0 je najbrži i najefikasniji RAID nivo, ali ne obezbeđuje zaštitu od otkaza.

B. RAID-1 Mirroring level

RAID-1 obezbeĒuje redudansu time što sve podatke upisuje na dva ili više diskova [6]. RAID-1 obezbeđuje slične brzine čitanja podataka i sporije upisivanje podataka kao u slučaju pojedinačnog diska, ali ukoliko dođe do otkaza nekog od diskova podaci neće biti izgubljeni. RAID-1 može da izdrži otkaz do N/2 diskova. Ovo je dobar *entry-level* redundantni sistem, s obzirom na to da su dovoljna samo dva diska, ali cena po megabitu može biti značajno veća jer se jedan od diskova koristi samo za dupliranje podataka.

C. RAID-5 Striping level

RAID-5 [3-5] nivo obezbeđuje redudansu time što se informacija o parnosti upisuje na sve dostupne diskove. Obezbeđuje veće brzine čitanja podataka i sporije upisivanje podataka u odnosu na jedan disk, ali ukoliko dođe do otkaza nekog od diskova podaci neće biti izgubljeni. RAID-5 obavlja blok-striping tehniku, a parnost je distrubuirana kroz sve diskove. RAID-5 je kvalitetno rešenje za sve vrste transfera, osim za male upise, gde ima jako slabe performanse.

D. Matematički model RAID konfiguracija

Matematički model za pojedinačne disk read/write transfere podrazumeva da je vreme trajanja za pojedinačno čitanje (*single read*, SR) i pojedinačan upis (*single write*, SW) predstavljeno kroz promenljive x_{SR}, x_{SW}, a srednje vreme pojedinačnog upisa i čitanja se može predstaviti u zavisnosti od komponenti definisanih kroz srednje vreme pozicioniranja (*seek time,* $\bar{x}_{seek}$), *latency time* ($\bar{x}_{latency}$) i vreme prenosa (*transfer time*, $\bar{x}_{transfer}$).

$$\bar{x}_{SR} = \bar{x}_{SW} = \bar{x}_{seek} + \bar{x}_{latency} + \bar{x}_{transfer} \quad (1)$$

Poseban ciklus upisa je *mali upis* koji se označava kao RMW (*read-modify-write*) ciklus:

$$\bar{x}_{RMW} = \bar{x}_{SR} + \bar{T}_{rotation} \quad (2)$$

Svako radno opterećenje sistema datoteka (*workload*) [6-7] se sastoji od nekoliko tipova upisa/čitanja, podeljenih u dve kategorije, slučajni i sekvencijalni upisi/čitanja (*Random Read/Write*, *SequentialRead/Write*), pa ukupno vreme za obradu definisanog radnog opterećenja sistema podataka za RAID zavisi od komponenti sekvencijalnog upisa ($T^{RAID}_{sequential_write}$) i čitanja ($T^{RAID}_{sequential_read}$), kao i slučajnog upisa ($T^{RAID}_{random_write}$) i čitanja ($T^{RAID}_{random_read}$):

$$T^{RAID}_{workload} = T^{RAID}_{random_read} + T^{RAID}_{sequential_read} + \\ + T^{RAID}_{random_write} + T^{RAID}_{sequential_write} \quad (3)$$

RAID-0 solidno ubrzava sekvencijalne upise-čitanja, upravo zbog primene *blok-striping* tehnike, koja omogućava paralelne upise/čitanja (*paralel streaming*) sa/na maksimalno N diskova, koji time praktično opslužuju jedan veliki zahtev SU (*Striping Unit*):

$$T^{RAID0}_{sequential_rw} \approx T_{SU_rw} / N \quad (4)$$

Zbog striping tehnike, RAID-0 značajno ubrzava slučajne (*random*) upise/čitanja, kao posledica konkurentih malih disk zahteva koji se realizuju na svih N diskova.

$$T^{RAID0}_{random_rw} \approx T_{random_rw_disk} / N \quad (5)$$

RAID-1, zbog postojanja duplih podataka, ubrzava sekvencijalno čitanje, pa se veliki transferi opslužuju simultanim čitanjem sa dva diska:

$$T^{RAID1}_{sequential_read} \approx T_{sequential_read_disk} / 2 \quad (6)$$

RAID-1 značajno ubrzava slučajna (*random*) čitanja, kao posledica konkurentih malih disk zahteva koji se realizuju na svih N diskova.

$$T^{RAID1}_{random_read} \approx T_{random_read_disk} / N \quad (7)$$

Ali RAID-1 ne može ubrzavati sekvencijalne upise:

$$T^{RAID1}_{sequential_write} \approx T_{sequential_single_disk} \quad (8)$$

RAID-1 ubzava *random* upise, koje se rade paralelno u parovima diskova kojih ima M=N/2:

$$T^{RAID1}_{radnom_write} \approx T_{radnom_write_disk} / M \quad (9)$$

Zbog prisustva *block-striping* tehnike RAID-5 solidno ubrzava sekvencijalna i slučajna čitanja podataka slično kao RAID-0, formule (4) i (5), ali su performanse ciklusa upisa zbog parnosti veoma složene. Obuhvaćena su tri specifična ciklusa upisa [5]:

1. FSW (*Full-Stripe Writes*) je upis cele trake bez parnosti, a parnost se izračunava bez čitanja:

$$p_{new1:N} \leftarrow d_{new1} \oplus d_{new2} \oplus d_{new3} \oplus d_{newN-1} \quad (10)$$

2. The RW (*Reconstruct Write*) predstavlja upis delimične trake, koji dalje inicira čitanje od najviše $N-2$ bloka podataka za potrebe izračunavanja parnosti:

$$p_{new1:N} \leftarrow d_1 \oplus d_2 \oplus d_3 \oplus d_{newi} \oplus ....d_{N-1} \quad (11)$$

3. RMW (*Read–Modify-Write*) predstavlja izuzetno male disk upise, manje od jednog disk bloka, i generiše 4 disk pristupa: 2 čitanja i 2 upisa.

RAID-5 ubrzava sekvencijalne upise samo u slučaju upisa celog stripa, zbog striping tehnike koja omogućava paralelni upis (*streaming*), ali se ovog puta izračunava i upisuje parnost, pa je nešto sporiji nego sekvencijalno čitanje. U slučaju RW pojavljuje se usporenje izazvano čitanjem određenog broja traka. Prvo se obavlja čitanje, zatim proračunavanje parnosti a na kraju se izvodi upis novih podataka i parnosti. Performanse zavise od odnosa FSW i RW ciklusa za radno opterećenje sistema podataka:

$$T^{RAID5}_{sequential_write} = F(T_{reconstruct_write}, T_{full_stripe_write}) \quad (12)$$

Za male disk upise performanse značajno opadaju. Za svaki ciklus upisa na svakom disku se realizuju 2 RMW ciklusa, čime su maksimalne perfomanse za RAID-5:

$$T^{RAID5}_{radnom_write} \approx T_{radnom_write_disk} / N \approx 2T_{RMW} / N \quad (13)$$

III. DETALJNA ANALIZA

Prilikom testiranja različitih RAID arhitektura primenjen je Intel Matrix RAID integrisan u Gigabyte GA-P55A-UD3 matičnoj ploči [9-10]. Intel Matrix RAID podržava sledeće nizove: RAID-0, RAID-1, RAID-1+0 i RAID-5. Testovi su obavljeni za RAID-0, RAID-1 i RAID-5 kao i na pojedinačnom disku. Za testiranja su korišćena tri Seagate Baracuda diska kapaciteta 500GB [11]. Instalacijom operativnog sisitema na četvrti nezavisni disk, omogućeno je testiranje čitanja i upisa podataka na diskove. Testiranje je obavljeno na sledećoj računarskoj konfiguraciji:

- Processor: Intel(R) i3 CPU 540 @ 3.07GHz
- Installed memory (RAM): 8.00 GB
- Hard disk: Seagate Baracuda
- File system: NTFS
- OS:Windows 7 Ultimate

RAID-ovi su testirani *benchmark* softverom **HD Tune Pro 5.50 [**12**],** koji obezbeđuje testiranja brzine upisa/čitanja

podataka. Testiranje je obavljeno korišćenjem dve opcije programa, **Benchmark** i **File benchmark,** na sledećim RAID konfiguracijama, single disk, RAID-0 (dva diska), RAID-0 (tri diska), RAID-1 (dva diska) i RAID-5 (tri diska). Tehničke karakteristike korišćenih diskova:

TABELA I. KARAKTERISTIKE SEAGATE BARACUDA HARD DISKOVA

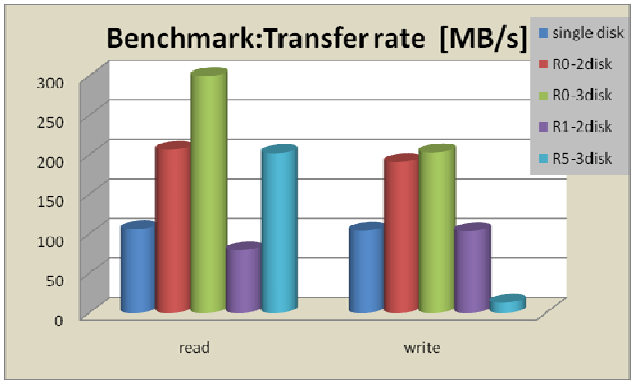
Specifikacije	500GB
Oznaka modela	ST500DM002
Interfejsi	SATA 6Gb/s NCQ
Performanse	
Broj obrtaja/min	7200
Keš	16MB
SATA brzine prenosa	6.0/3.0/1.5 Gb/s
Sr. brzina pretraživanja (čitanja)	11ms
Sr. brzina pretraživanja (upisa)	12ms
Sr. brzina čitanja/upisa	125Mb/s
Maks. brzina prenosa podataka	144Mb/s

A. Rezultati za HD Tune Benchmark opciju

Benchmark opcija softvera nudi četiri različita testa:

- ***Transfer rate:*** brzina prenosa podataka sa cele površine diska
- ***Acces time:*** vreme pristupa
- ***Burst rate:*** najveća brzina kojom se podaci mogu preneti
- ***CPU usage:*** izražen u procentima

Rezultati transfer rate za Benchmark opciju programa HD Tune Pro su prikazani na slici 1.



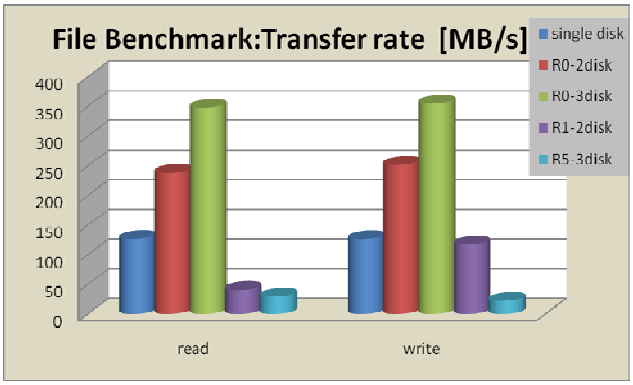
Slika 1. Rezultati transfer rate za Benchmark opciju.

U ovom testu dominiraju sekvencijalne komponente za čitanje/upis definisane u formuli 3. Za operaciju čitanja, RAID-0 i RAID-5 arhitekture ostvaruje ubedljivo najbolje performanse, a razlog je relacija predstavljena kroz formulu 4 za sekvencijalno čitanje. RAID-1 se nije pokazao kao dobro rešenje, odnosno formula (6) ne važi u ovom slučaju. Tada Matrix ne koristi činjenicu da ima duplikat podataka, pa se RAID-1 rešenje pokazalo kao slabije od pojedinačnog diska. Za upis, RAID-0 jako dobar, RAID-1 je sličan pojedinačnom disku (8), dok je RAID-5 pokazao loše performanse, uglavnom zbog FSW i RW ciklusa (10, 11, 12).

B. Rezultati za File Benchmark opciju

File benchmark test sastoji se iz dva dela: merenje brzine transfera i izvođenje testova na bazi veličine bloka. U ovom radu je primenjena samo prva opcija, brzina transfera, pri čemu se mere tri različita parametra:

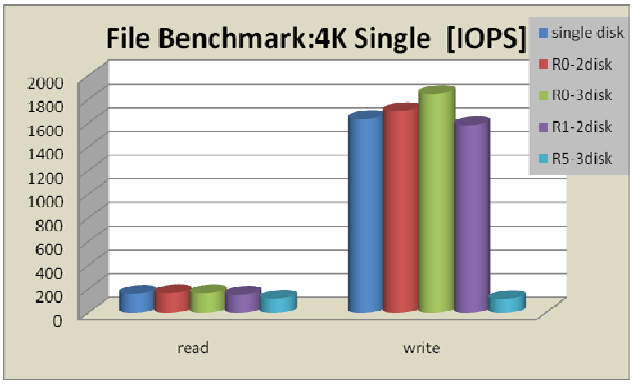
•**Sekvencijalni test**:
Sekvencijalna brzina se meri i prikazuje na grafikonu, a bazira se na merenju velikih fajl transfera.
Rezultati transfer rate za FileBenchmark opciju programa HDTunePro su prikazani na slici 2.



Slika 2. Rezultati transfer rate za FileBenchmark opciju.

Rezultati su veoma slični kao što su oni dobijeni u slučaju primene osnovnog Benchmark testa (slika 1), s tim što RAID-0 sada ima pravilnije ponašanje za upis po formuli 4, u odnosu na prethodni test.

• **4 KB random single test**: ovaj test meri performanse I/O operacija 4096 bajt blokova. Ovo je česta I/O operacija na tipičnom sistemu datoteka. Rezultati za **4 KB random single test** opciju programa HD Tune Pro su prikazani na slici 3.

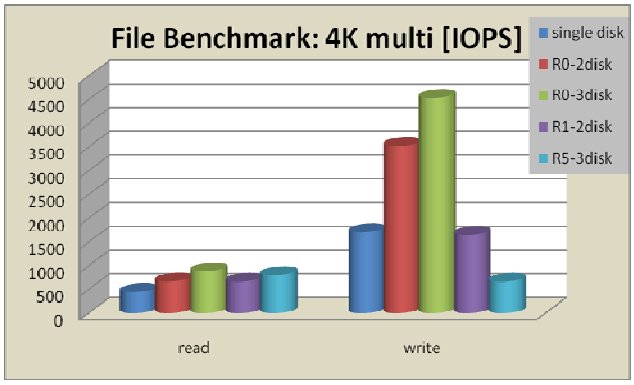


Slika 3. Rezultati za 4KB *random single* test.

U ovom testu dominiraju *random* komponente za čitanje/upis u skladu sa formulom 3, ali je zbog *single* prirode testa, broj konkurentnih zahteva manji. Za čitanje, sa grafika detektujemo da su svi RAID slabi po perfomansama i veoma slični, što znači da nijedan RAID nije uspeo da izvuče konkurenciju, tj da uposli više diskova istovremeno (formule

(5) i (7)). Za upis, RAID-0 je pokazao nešto bolje performanse u odnosu na ostale ispitivane konfiguracije (5), RAID-1 je sličan single disku (9), a RAID-5 jako loš, a to je posledica malih RMW upisa (13).

• **4 KB random multi test**: Ovaj test je sličan prethodnom testu izuzev što se više operacija istovremeno šalje na uređaj. Broj operacija se može unapred odrediti i imati vrednost od 2 do 64 (u našem slučaju 32). Rezultati za 4 KB *random-multi-test* opciju programa HD Tune Pro su prikazani na slici 4.



Slika 4. Rezultati za 4KB *random multi* test.

U testu dominiraju *random* komponente za čitanje/upis u formuli (3), ali je zbog više procesa u testu, broj konkurentnih zahteva osetno veći. Za čitanje, sa grafika detektovano je da su RAID-0 i RAID-5 dosta dobrih performansi, u skladu sa formulom (5), što znači da su uspeli da izvuku konkurenciju tj. da uposle veći broj diskova istovremeno, kao i RAID-1 po formuli (7). Za upis, RAID-0 je solidno bolji od ostalih (po formuli 5), RAID-1 je sličan *single* disku (9), a RAID-5 jako loš, a to je posledica malih RMW upisa (13).

ZAKLJUČAK

Osnovni zaključak ovog rada je da se RAID nivoi 0, 1 i 5 na relativno jednostavnom hardverskom kontroleru ponašaju očekivano, u skladu sa našim matematičkim modelom performansi. RAID-0 unosi solidno ubrzanje u odnosu na pojedinačni disk, kao i u odnosu na konkurentne RAID nivoe 1 i 5. To je izmereno i analizirano za radno opterećenje sa sekvencijanim i *random* osobinama. Zahvaljući *striping* tehnici, RAID-0 ubrzava sve sekvencijalne transfere, formula (4), a takođe ubrzava sve slučajne zahteve u kojima uspeva da realizuje konkurentnost, formula (5). RAID-1 ima slične osobine kao pojedinačni disk, zahvaljujući paralelnim SATA kanalima, koji ne dozvoljavaju osetan pad performansi upisa, a kombinacija od svega dva diska, omogučava samo relativno povećanje *random read* performansi, formula (7). RAID-5 se takođe ponaša očekivano. Za sekvencijalna i *random* čitanja, RAID-5 je dosta dobar, formule (4) i (5), međutim, za operaciju upisa postoji značajno usporenje koje unosi konfiguracija RAID-5. Razlog za takvo ponašanje konfiguracije RAID-5 je poznati "*small updates*" problem, tj. prisustvo RMW i RW ciklusa. Taj performansni nedostatak čini RAID-5 veoma osetljivim na cikluse u kojima dominiraju mali upisi, što se u našem slučaju očigledno dogodilo

Ovim rezultatima se ohrabruju svi korisnici hardverskih RAID da koriste RAID-0 ili RAID-1, dok sa RAID-5 se očekuje solidan pad performansi za opterećenja sa malim upisima. Ali treba naglasiti da kod RAID-0 ne postoji redudancija, a kod RAID-1 mora da se „plati" dupli disk prostor, tako da bolje performanse korisnici dobijaju na uštrb tih činjenica, tako da osnovu svojih potreba treba da izaberu konfiguraciju koju će koristiti.

ZAHVALNICA

Rad je finansiran od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije (TR32025, III43002, TR32037).

LITERATURA

[1] H. Custer,"Inside the Windows NT File System,"Microsoft Press, Redmond, WA, USA,1994.

[2] M. Shafiei, F. Iranmanesh, "Review NTFS Basics," Australian Journal of Basic & Applied Sciences, , Vol. 6 Issue 7, pp. 325-338, 2012.

[3] A. Lebrecht et al., "Analytical and Simulation Modelling of Zoned RAID Systems,"*Computer Journal*, 2010.

[4] A. Thomasian, "Clustered RAID Arrays and Their Access Costs,"*Computer Journal*, doi:10.1093/comjnl/bxh108, 2005.

[5] A. Thomasian et al., "Analysis of Rebuild Processing in RAID5 Disk Arrays,"*Computer Journal*, 2007.

[6] A. Thomasian, J. Xu, "Reliability and Performance of Mirrored Disk Organizations,"*Computer Journal*, 2008.

[7] E. Shriver et al., "An analytic behaviour model for disk drives with readahead caches and request reordering,"*in Proc. of 1998 ACM SIGMETRICS/Performance joint Int.Conf. on Measurement and modeling of computer systems,* pp. 182-191, 1998.

[8] M. Seltzer et al., "Journaling versus Soft Updates: Asynchronous Meta-data Protection in File Systems,"*USENIX Conf. Proc.*, San Diego, pp. 71-84,2000.

[9] "Intel Matrix Storage Technology," Inter Corporation, Available: http://www.intel.com/

[10] „Set Up a System with Matrix RAID", Inter Corporation, Available: http://www.intel.com/support/chipsets/imsm/

[11] „Seagate Barracuda hard discs", Seagate Technology LLC, Available: http://www.seagate.com

[12] „Hard disk/SSD utility HD Tune Pro 5.50", HD Tune, Available: http://www.hdtune.com

ABSTRACT

The paper presents the modeling and results of performance analysis of the RAID configurations for NTFS file system, under the MS Windows operating system. The paper includes mathematical modeling of RAID performances, a model is validated through the interpretation of performance. The observed performances were measured by the means of the HD TunePro benchmark application, taking into cosideration the single disk and three variations of RAID: striping RAID-0, striping RAID-5, and mirrored RAID-1.

MODELING AND COMPARISON OF RAID PERFORMANCE ON MOTHERBOARD HARDWARE-RAID CONTROLLER

Borislav Đorđević, Valentina Timčenko, Slavica Boštjančič Rakas, Nemanja Maček, Milan Kuprešanin