RAID in baza podatkov Oracle

Marija Kuhar[1], Borut Vovk[2], Miro Gradišar[3]

Izvleček

Varnost in zanesljivost informacijskih sistemov postajata vedno pomembnejša. V času, ko se poslovanje seli na spletne strani, je potreba po stalni dostopnosti podatkov vedno bolj pogosta. Hkrati se cena izpada sistema močno dviga, zaradi česar se pojavljajo nove strojne in programske rešitve, ki zagotavljajo boljšo zanesljivost, razpoložljivost in zmogljivost informacijskih sistemov.

Članek obravnava organizacijo diskov poimenovano RAID, ki poveča hitrost, razpoložljivost in zanesljivost računalniških sistemov. Predstavljeni so možni načini organizacije oziroma nivoji RAID, prednosti in slabosti le-teh, priporočljiva področja uporabe in analiza uporabe z vidika baze podatkov Oracle.

Ključne besede: RAID, zanesljivost, razpoložljivost, organizacija diskov, baza podatkov Oracle

Abstract

Raid and Oracle Database

Increased capacity requirements for network applications and better reliability in the time when significantly declined cost of storage per megabyte and high productivity losses give a fresh impetus to storage industry, which puts numerous solutions for improving overall better system performance on the market.

The article presents RAID storage systems and other supplements for improving overall availability and reliability of information systems. The article especially discusses the implementation of RAID systems in Oracle environment.

Key words: *RAID, reliability, availability, disk organization, Oracle database*

1. Uvod

Zadnjih nekaj deset let so se zmogljivosti procesorjev eksponentno večale. Približno vsakih 18 mesecev so se podvojile (1), kar pa ni moč reči za zmogljivosti diskov. V začetku sedemdesetih je bil povprečni čas dostopa do podatka na disku miniračunalnika nekje med 50 in 100 milisekundami. Dandanes se ti časi vrtijo okrog 10 milisekund. V mnogih vejah tehnične industrije je faktor sprememb 5 do 10 v 30 letih velika številka, v računalniški industriji, kjer dosega razvoj vrtoglave hitrosti, pa je tak napredek skromen. Razlika med zmogljivostjo procesorjev in diskov se tako iz leta v leto veča in seveda pomeni vedno večji problem, saj diski postajajo ozko grlo v procesu obdelave podatkov.

Na področju procesorjev se je veliko raziskovalo in tudi doseglo v smeri vzporednega procesiranja. Tako so znanstveniki v poznih osemdesetih začeli razmišljati tudi o paralelizmu na področju sistemov za shranjevanje podatkov. Leta 1988 so trije raziskovalci kalifornijske univerze David Patterson, Randy Katz in Garth Gibson objavili idejo o šestih različnih načinih paralelizma na področju organizacije diskov (2). Sistem je poimenovan RAID. RAID je kratica, ki danes pomeni *Redundant Array of Independent Disks* ali, če poskusimo to posloveniti – skupina neodvisnih diskov s preobiljem podatkov. Pri tem pomeni redundanca ali preobilje večkrat zapisane enake podatke ali osnovnim podatkom pridružene nadzorne podatke, ki so iz njih izračunani po določenem algoritmu, tako da je možno odkrivanje in odpravljanje napak. Glede tega bolj ali manj posrečenega prevoda velja omeniti, da je kratica RAID v začetku pomenila 'Redundant Array of Inexpensive Disks' – torej skupina poceni diskov, vendar je industrija idejo hitro spoznala za koristno in razvila praktične rešitve, ki so bile vse prej kot poceni. Zato so besedico 'Inexpensive' zamenjali za 'Independent'.

Glavna ideja je bila sestaviti črno škatlo, ki bo navzven izgledala kot en sam hiter in zanesljiv disk. Znotraj te črne škatle pa naj bi bilo več počasnejših in manj zanesljivih diskov ter krmilnik, ki naj bi skrbel za njihovo usklajeno delovanje.

Namen tega preglednega članka je predstaviti tehnologijo RAID in njeno praktično uporabo pri bazah podatkov Oracle, ki so v Sloveniji zelo razširjene. V nadaljevanju bomo opisali tehnologijo RAID in načine, kako lahko izboljšamo zmogljivost računalniškega sistema z uporabo različnih nivojev RAID, ki jih bomo med seboj primerjali. Opisali bomo različne možnosti uvedbe te tehnologije v prakso. Na koncu bomo podali pregled nad nivoji RAID, ki so zlasti zanimivi z vidika Oraclove baze podatkov in izvedli primerjavo tudi med njimi.

1 Grad d.d., Tržaška 118, 1000 Ljubljana, marija@grad.si
2 Gorenjska banka d.d., 4000 Kranj, borut.vovk@gbkr.si
3 Ekonomska fakulteta, Kardeljeva ploščad 17, 1000 Ljubljana, miro.gradisar@uni-lj.si

2. Tehnologija RAID

Razvoj tehnologije RAID koncem osemdesetih let so predvsem spodbujali naslednji takratni trendi (6):

- večanje potreb po velikih diskih zaradi novih omrežnih aplikacij
- hitrejši procesorji (stokratno povečanje zmogljivosti procesorjev napram samo štirikratnemu povečanju zmogljivosti diskov v istem obdobju) zahtevajo boljši vhodno izhodni (V/I) sistem
- zanesljivost sistema je v novih (omrežnih) aplikacijah vedno pomembnejša. Tehnologija RAID lahko prepreči izgubo podatkov, do katerih bi prišlo zaradi napak oziroma okvar diskov
- cena na enoto prostora na disku je močno padla in sistemi RAID so se razširili iz velikih računalniških centrov celo na področje delovnih postaj in namiznih računalnikov.

Zanesljivost delovanja kot velikokrat najpomembnejšo lastnost diskovnih sistemov lahko razčlenimo na (6):

- neobčutljivost na izpade (*fault tolerance*)
- majhno pogostost napak pri branju in pisanju
- visoko razpoložljivost.

V tabeli 1 so prikazani vzroki za odpovedi računalniških sistemov, kot jih je v svoji študiji leta 1995 predstavilo podjetje Intel Corporation.

Neobčutljivost na izpade posameznih komponent pomeni, naj bi bil sistem sestavljen tako, da posamezne komponente lahko odpovejo, vendar s tem ne povzročijo izpada sistema. Zrcaljenje diskov je najpogostejši način za doseganje neobčutljivosti na izpade, ki ga uporabljajo tudi sistemi RAID. Če odpove primarni disk, njegovo nalogo prevzame zrcalni disk in tako uporabnik opazi okvaro le kot počasnejše delovanje sistema.

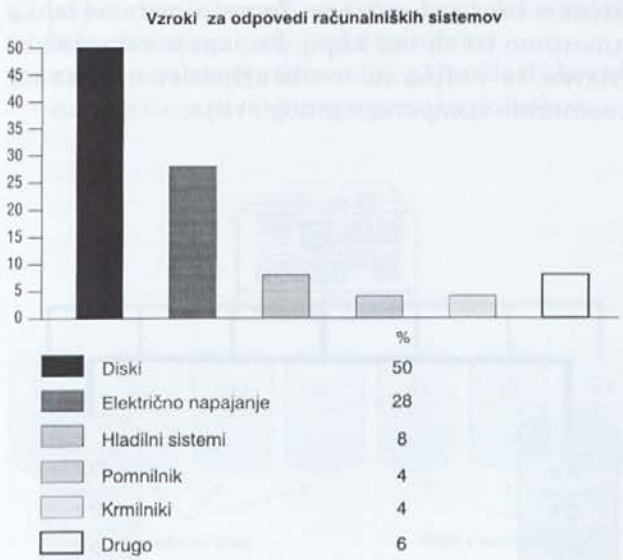


Tabela 1:
Razlogi za odpovedi računalniških sistemov (vir: Intel Corporation)

Majhna pogostost napak je naslednji pomemben dejavnik zanesljivosti. Napake sistema se lahko pojavljajo zaradi različnih vzrokov. Kadar le-te nastajajo zaradi vplivov okolja, na primer visoke temperature, so odpovedi posameznih komponent pogoste. Da se temu izognemo, je priporočljivo uporabljati komponente z visokim MTBF (kratica izhaja iz angleščine in pomeni povprečni čas med odpovedmi komponente - *mean time between failure*) in pomožne sisteme, kot so klimatske naprave, podvojeni sistemi napajanja in podobno. Nenazadnje lahko tudi programska oprema pripomore k majhni pogostosti napak. Programi, ki nadzorujejo delo sistema, lahko avtomatsko javljajo sumljive dogodke, ki nakazujejo na okvaro v obliki elektronske pošte ali sporočila na mobilni telefon dežurnega vzdrževalca.

Visoka razpoložljivost pomeni, da naj bi bil v nekem obdobju sistem čim dalj časa na razpolago. Visoka razpoložljivost je gotovo lastnost sistema, pri katerem niti okvara niti odprava te okvare ne povzročita izpada sistema. Visoko razpoložljivi diskovni sistemi morajo biti sestavljeni iz komponent, ki jih je mogoče zamenjati med delovanjem sistema. V ta namen se pogosto uporablja združevanje računalnikov v gruče (*clusters*), kjer je več strežnikov povezanih na isti diskovni sistem, da se omogoči uporabnikom dostop do programov in podatkov preko nadomestnega računalnika, če matični odpove. Seveda je v tem primeru za uporabnike ob delovanju vseh strežnikov pridobitev že to, da je zagotovljene več procesne moči za strežniško orientirane programe pri velikem številu uporabnikov.

2.1. Izboljšanje zmogljivosti sistema

V primerjavi z navadnimi diskovnimi sistemi omogočajo sistemi RAID izboljšanje zmogljivosti. Dejanske spremembe zmogljivosti diskovnega sistema RAID so odvisne od tega, kako je sistem zgrajen oziroma kakšen nivo RAID uporabljamo in od tega, kakšen je način dela s podatki. Pri uporabi zrcaljenja podatkov se zmogljivost v primerjavi z navadnimi enodiskovnimi sistemi pri branju podatkov povečajo, pri zapisovanju pa zmanjšajo. Pri sistemu zapisovanja paketov podatkov vzporedno na več diskov (*striping*), se poveča hitrost tako pri branju kot zapisovanju podatkov, vendar se zanesljivost sistema zmanjša. Nekateri sistemi RAID uporabljajo kombinacijo obeh principov in tako dosežejo hkrati pohitritev in povečanje zanesljivosti diskovnega sistema. Sistemi RAID so različnih tipov ali konfiguracij, ki jih literatura imenuje nivoji (*levels*).

2.2 Nivoji RAID

V uvodu omenjeni raziskovalci kalifornijske univerze so definirali 6 osnovnih nivojev RAID. Čeprav se govori o nivojih, tu ne gre za nikakršno hierarhijo, podrejenost

in nadrejenost posameznih nivojev. Gre le za 6 različnih konfiguracij, ki so pač tako poimenovane. Poleg teh šestih osnovnih konfiguracij se v praksi uporabljajo tudi kombinacije, izpeljane iz osnovnih. Opisi teh konfiguracij ter prednosti in slabosti so povzete po priporočilih proizvajalcev strojne opreme (8) (9) (11) in svetovalcih in združenjih (7) (10) (12).

2.2.1 RAID 0

Ta tehnika je poznana tudi kot *data striping*. Gre za razdeljevanje podatkov na več manjših blokov enake velikosti, ki omogočajo istočasno oziroma vzporedno branje ali pisanje na več diskov ali z njih in s tem občutno povečanje hitrosti. Skupina teh diskov je navzven vidna kot en sam velik in hiter disk.

Ker RAID 0 ne uporablja redundance podatkov, je tako od vseh konfiguracij glede prostora najvarčnejši. Prav odsotnost redundance pa je slaba stran. Ti sistemi niso neobčutljivi na izpade posameznih komponent, zato se v praksi največkrat uporabljajo le skupaj z drugimi nivoji RAID. Če odpove eden od diskov, izgubimo vse podatke in sistem stoji. Cena izpada je lahko bistveno višja kot prihranek pri nakupu diskovnega prostora. Učinek sistema je najboljši, kadar se uporablja čim večje število fizičnih diskov in kadar vsak krmilnik nadzira delo čim manj diskov, najbolje samo enega. Velikost posameznega bloka (*stripe*) podatkov mora biti skrbno pretehtana, sicer lahko dosežemo nasprotni učinek od zaželenega. S tako konfiguracijo dosežemo izjemno hitrost delovanja, ki je za nekatera področja zelo pomembna. Tako področje je na primer video, kjer je bistvena hitrost branja podatkov, saj je za kvaliteten zvok in sliko nujen neprekinjen dotok podatkov do procesorja.[4]. Značilnosti RAID 0 lahko strnemo v naslednje ugotovitve.

Prednosti:
- enostavnost sistema,
- preprosta izgradnja sistema,
- dobra zmogljivost.

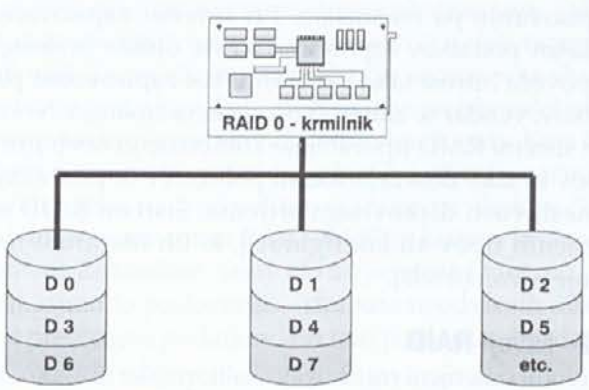
Slabosti:
- občutljivost na odpovedi,
- izguba samo enega diska povzroči izgubo vseh podatkov.

Priporočena področja uporabe:
- urejanje in predvajanje videa,
- urejanje slik, fotografij,
- aplikacije, ki potrebujejo velik pretok podatkov.

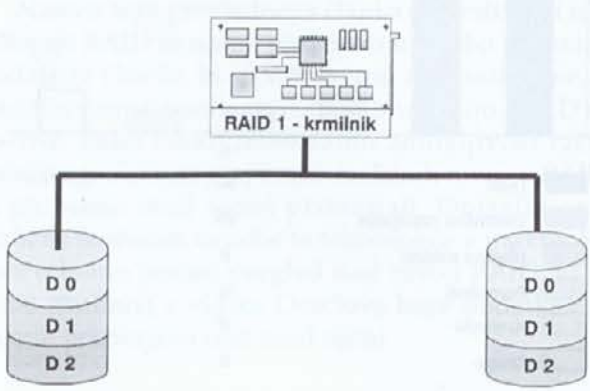
2.2.2 RAID 1

RAID 1 je najpreprostejši sistem, ki zagotavlja neobčutljivost na okvare posameznih komponent. Poznamo ga tudi pod imenom zrcaljenje. RAID 1 krmilnik razdeli diske v dve skupini. Vsak podatek vzporedno zapiše na obe skupini (slika 2). Za postavitev sistema sta potrebna vsaj dva diska. Če eden odpove, njegovo delo prevzame drugi in do izpada sistema ne pride. Pade le zmogljivost, dokler ne vstavimo nov disk. Ob izpadu še tega edinega diska odpove celoten sistem. Cena uvedbe sistema je zelo visoka, saj moramo celotno konfiguracijo podvojiti. Dodatno varnost pred odpovedmi sistema si zagotovimo s podvajanjem vseh komponent v sistemu (električno napajanje, pretvorniki, V/I vodila, itd.) in nakupom bolj kakovostnih komponent. Zaradi visoke ravni zaščite podatkov in enostavnosti uvajanja in vzdrževanja ta sistem priporočajo mnogi ponudniki sistemov za upravljanje z bazami podatkov, če že ne za celotno zbirko podatkov, pa vsaj za njene najpomembnejše dele. Zrcaljenje diskov je tudi edina konfiguracija RAID, kjer hitrost ni večja od hitrosti navadnih diskovnih sistemov brez redundance. Za večjo varnost lahko namestimo tri ali več kopij. Pri tem je zanesljivost sistema še boljša in neobčutljivost na okvare posameznih komponent mnogo višja.

4 *V tem primeru je priporočljivo uporabljati diske, kjer se ne izvaja tako imenovana termična rekalibracija – ponovno nastavljanje bralno pisalnega mehanizma diska zaradi spreminjanja dimenzij kot posledice pregrevanja.*



Slika 1: RAID 0 – »striping«



Slika 2: RAID 1 - zrcaljenje

Prednosti:
- teoretično dvakratna hitrost branja podatkov,
- visoka neobčutljivost na odpovedi komponent,
- pri odpovedi posamezne komponente ne izgubimo podatkov,
- enostavnost izgradnje sistema.

Slabosti:
- največja stopnja redundance (100%) od vseh nivojev RAID,
- najdražji sistem s stališča cene na enoto diskovnega prostora.

Priporočena področja uporabe:
- računovodski sistemi,
- finančni sistemi,
- aplikacije, ki potrebujejo zelo visoko razpoložljivost.

2.2.3. RAID 2

Nivo RAID 2 (slika 3) uporablja za odkrivanje in odpravo napak sistem, imenovan Hammingova koda. Za ta sistem potrebujemo vsaj 7 diskov. Za vsake 4 bite podatkov, kjer se vsak bit zapiše na svoj disk, se sproti izračunavajo in zapisujejo še 3 nadzorni biti. Ti nadzorni biti se izračunajo tako, da je v primeru odpovedi kateregakoli diska možno na podlagi informacij, zapisanih na ostalih šestih diskih izračunati vrednost okvarjenega bita in sistem nemoteno deluje. Slabost takih sistemov je veliko število diskov in krmilnikov, posebej če želimo zmanjšati odstotek redundantnih diskov. Šele pri uporabi 38 diskov (32 podatkovnih in 6 nadzornih) pade odstotek redundantnih bitov na 19%. Druga slabost je potrebna sinhronizacija diskov v sistemu, ki jo dosežemo tako, da diske namestimo na isto gred, kar je za proizvodnjo zelo zahtevno.

Prednosti:
- popravljanje napak »v živo«, med delovanjem, brez občutnega padca odzivnih časov,
- zaradi paralelnega dela velikega števila diskov obstaja možnost zelo hitrega pretoka podatkov,
- relativno preprost krmilnik v primerjavi s tistimi za RAID nivoje 3, 4 in 5.

Slabosti:
- na trgu se zaradi zapletenosti taki sistemi niso uveljavili,
- za implementacijo potrebujemo preveč diskov,
- delo vseh diskov mora biti sinhronizirano, kar ni lahko doseči.

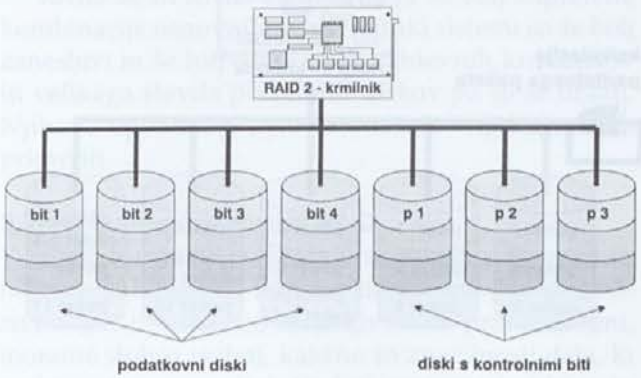
2.2.4 RAID 3

RAID 3 je poenostavljena verzija sistema RAID 2. Namesto Hammingove kode se za odkrivanje in odpravljanje napak uporablja le paritetni bit. V lihi paritetni shemi mora biti vsota vseh bitov liha, zato dobi paritetni bit vrednost 0 ali 1 glede na vsoto podatkovnih bitov. Podobno je v sodi paritetni shemi vsota vseh bitov soda. Postopek poteka tako, da se izračuna vsota podatkovnih bitov, nato pa se določi vrednost paritetnega bita, ki se zapiše na poseben vzporeden paritetni disk (slika 4). Če pride do izpada enega od diskov, ga moramo locirati, zatem pa lahko na podlagi informacij iz drugih diskov in paritetnega diska določimo vrednost podatka na disku, ki je odpovedal.

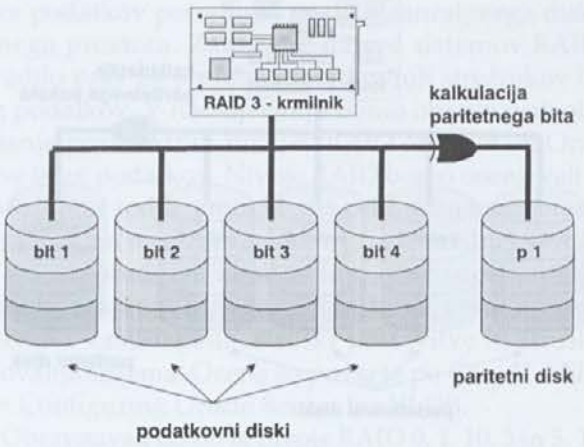
Prednosti:
- hitro branje in pisanje podatkov,
- odpoved enega diska ne vpliva močno na delovanje sistema,
- malo redundance podatkov.

Slabosti:
- kompleksni in dragi krmilniki,
- potrebno je zagotoviti sinhronizirano delovanje diskov,
- paritetni disk postane ozko grlo sistema pri pisanju velike količine podatkov.

Priporočena področja uporabe:
- urejanje in predvajanje videa,
- urejanje slik, fotografij,
- aplikacije, ki potrebujejo velik pretok podatkov.



Slika 3: RAID 2 – Hammingova koda



Slika 4: RAID 3 – paralelni prenos s pariteto

2.2.5 RAID 4

RAID 4 se od RAID 3 razlikuje po tem, da na posameznem disku ni zapisan le en podatkovni bit ampak en blok ali paket (slika 5). Tako se pri branju manjših količin podatkov lahko uporablja le enega ali samo potrebne diske, ne pa vse, kot je to treba pri RAID 3. Pri branju zato dosežemo zelo dobre rezultate, pri zapisovanju posebej manjših količin podatkov pa sistem deluje počasi, kajti tudi za najmanjšo spremembo podatkov znotraj bloka mora prej prebrati vse pripadajoče bloke na vseh podatkovnih diskih, da lahko obnovi pariteto. Prednost tega sistema je nizka cena dodatnega diskovnega prostora za kontrolne podatke v primerjavi z RAID 1 in fizično manj zapletena izdelava v primerjavi z RAID 2 in RAID 3. Zaradi slabih odzivnih časov pri zapisovanju podatkov se v strežniških bazah podatkov (npr. Oracle) ne uporablja, razen če aplikacija ne zahteva veliko sprotnega dodajanja, popravljanja in brisanja podatkov ali če nižja cena odtehta slabosti.

Prednosti:

- hitro in učinkovito branje podatkov,
- zaradi malo paritetnih diskov malo redundance podatkov.

Slabosti:

- zapleteni in dragi krmilniki,
- slaba zmogljivost pri zapisovanju in predvsem prepisovanju podatkov,
- počasno vzpostavljanje normalnega stanja ob izpadu ene komponente,
- pri pisanju velike količine podatkov paritetni disk postane ozko grlo sistema.

2.2.6 RAID 5

Sistem deluje podobno kot RAID 3 ali 4, le da je tu odpravljeno ozko grlo pri prepisovanju in brisanju podatkov. Paritetne informacije so namreč porazdeljene po vseh diskih v sistemu (slika 6). RAID 5 je danes na trgu najpogosteje uporabljan sistem take vrste. Zaradi majhne redundance podatkov je cenovno ugoden ob zelo dobrem odzivnem času pri branju podatkov in zmernem času zapisovanja. Ob izpadu enega diska le-tega lociramo in zamenjamo, potem pa sistem samodejno obnovi vsebino okvarjenega diska. En blok podatkov je lahko poljubne velikosti. Velja pa omeniti, da se z določanjem njegove velikosti da optimirati odzivne čase in konfiguracijo prilagoditi delovanju različnih sistemov. Pri obnavljanju podatkov po odpovedi in zamenjavi komponente je zmogljivost sistema močno zmanjšana.

V sistemih RAID 5 je najbolj problematično zapisovanje podatkov. Vsaka zahteva po zapisu sproži proces, sestavljen iz šestih korakov:

1. branje bloka, kamor naj bi zapisali nove podatke
2. branje pripadajočega primerjalnega bloka
3. odstranjevanje primerjalnih podatkov, ki naj bi bili prepisani
4. dodajanje novih primerjalnih podatkov
5. zapisovanje novih primerjalnih podatkov
6. zapisovanje novih podatkov.

Prednosti:

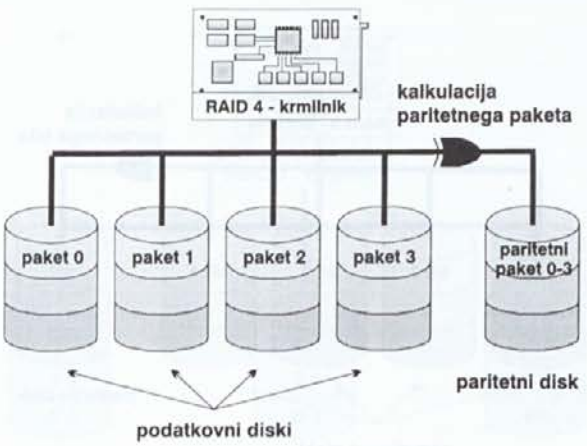
- hitro branje podatkov
- zadovoljiva hitrost zapisovanja podatkov
- malo redundantnih podatkov
- velika razširjenost.

Slabosti:

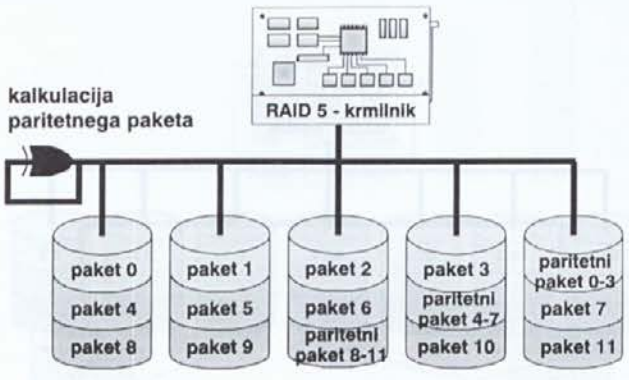
- zapleteni krmilniki
- zapleteno vzpostavljanje normalnega stanja ob izpadu ene komponente
- pri odpovedi ene komponente sistem sicer deluje, vendar se zmogljivost poslabša.

Priporočena področja uporabe:

- datotečni in aplikacijski strežniki
- strežniki za baze podatkov
- spletni strežniki in strežniki za elektronsko pošto
- intranetni strežniki.



Slika 5:
RAID 4 – neodvisni podatkovni diski s skupnim paritetnim diskom



Slika 6:
RAID 5 – neodvisni podatkovni diski s porazdeljenimi paritetnimi paketi

2.2.7 RAID 6

RAID 6 ima tako kot RAID 5 po več diskih porazdeljene paritetne informacije. Od RAID 5 se razlikuje v tem, da so paritetne informacije izračunane in zapisane dvakrat. Tako je ta sistem od nehibridnih (brez sestavljanja več osnovnih RAID konfiguracij skupaj) najzanesljivejši, vendar tudi zelo drag.

Prednosti:

- majhna občutljivost na odpovedi posameznih komponent
- idealna rešitev za sisteme, ki zahtevajo zelo veliko razpoložljivost in kjer si ne moremo privoščiti izgube podatkov.

Slabosti:

- zelo zapleteni in dragi krmilniki
- veliko redundantnih podatkov
- slabi odzivni časi pri zapisovanju podatkov
- zahteva po dodatnem disku zaradi dvonivojske paritete.

2.2.8 Drugi nivoji RAID

RAID 7 se kot izredno drag in zapleten sistem pojavlja zelo redko. Gre za asinhrono krmiljenje branja in pisanja podatkov, kar pomeni, da vsaki diskovni plošči pripada neodvisna gred, z internim operacijskim sistemom za krmiljenje in predpomnilnikom z vgrajenim paritetnim sistemom. RAID 7 je izredno zanesljiv, z odličnimi odzivnimi časi, vendar zaradi zahtevne izdelave zelo drag.

Skoraj vsi ponudniki diskovnih sistemov so razvili tako imenovane hibridne sisteme RAID, kjer se kombinirajo osnovni principi RAID. Najpogostejši so RAID 10, RAID 30 in RAID 50.

RAID 10 je kombinacija RAID 1 in RAID 0. Sistem je potemtakem tako kot RAID 1 dokaj neobčutljiv na odpovedi posameznih komponent in ima tako kot RAID 0 zelo kratek odzivni čas. Takšni sistemi so dragi, vendar vseeno precej razširjeni. Priporočljivi so za strežnike z bazami podatkov, kjer se zahtevata visoka zmogljivost in velika razpoložljivost.

RAID 30 in RAID 50 vsebujeta še bolj zapletene kombinacije osnovnih nivojev. Taki sistemi so še bolj zanesljivi in še hitrejši. Zaradi zahtevnih krmilnikov in velikega števila potrebnih diskov pa so še dražji. Njihova uporaba je zato smotrna le v dokaj redkih primerih.

2.2.9 Primerjava nivojev RAID

V tabeli 2 je predstavljena primerjava posameznih nivojev RAID (2). Če hočemo doseči sprejemljive odzivne čase in ustrezno zanesljivost ob primerni ceni, moramo dobro vedeti, kakšne so značilnosti dela, ki ga bo sistem opravljal. Vedeti moramo kolikšna je količina podatkov, kakšni so sprejemljivi odzivni časi pri branju, pisanju in obnovi sistema po okvari posamezne komponente ter kakšna je zahtevana varnost podatkov. Le tako se lahko odločimo za najprimernejši nivo RAID.

2.3. Izvedba RAID

Ponudniki sistemov za shranjevanje podatkov ponujajo nekaj načinov za namestitev RAID. Vsak od teh ima svoje prednosti in slabosti, ki jih je potrebno preučiti pred nakupom. Ti načini so:

- strežnik z vgrajenim krmilnikom za RAID, programsko krmiljenje RAID, v matično ploščo vgrajen RAID krmilnik ali kartica PCI s krmilnikom,
- zunanji RAID sistemi s poljem diskov v samostojnem ohišju, kjer je eden ali več takih sistemov povezanih s strežnikom ali delovno postajo. Posamičen samostojni RAID sistem lahko uporablja več strežnikov v večstrežniškem okolju,
- najboljša razpoložljivost podatkov je dosežena s kombiniranjem RAID sistemov in drugih komponent, ki povečujejo razpoložljivost računalniškega sistema, kot so brezprekinitveni napajalniki, gruče strežnikov, ko ob odpovedi enega prevzame delo drug strežnik itd.

Odpoved sistema povzroči stroške, med katere velikokrat ne moremo šteti le stroškov nakupa in zamenjave nove komponente, ampak tudi stroške izgube delovnega časa, izgubljenih podatkov, obnove podatkov, izgube dobrega imena podjetja in težav, ker stranke ne morejo dostopati do podatkov. Če so ti stroški veliki in če želimo povečati tudi hitrost delovanja, je pametno razmisliti o uvedbi sistema RAID. Nekateri proizvajalci sistemov RAID in uporabniki so v ta namen ustanovili organizacijo, imenovano RAID *Advisory Board* (RAD) (10), ki je definirala merila za zanesljivost sistema.

3. Sistemi RAID z vidika baze podatkov Oracle

Baze podatkov potrebujejo veliko zanesljivega diskovnega prostora. Zato se je največ sistemov RAID zgradilo prav za potrebe podatkovnih strežnikov in baz podatkov. V nadaljevanju bomo obravnavali obnašanje posameznih nivojev RAID ob uporabi Oraclove baze podatkov. Nivoje RAID bomo ocenjevali z naslednjimi sodili: zmogljivost pri naključnem branju, pri naključnem zapisovanju, pri zaporednem branju, pri zaporednem zapisovanju, pogostost izpada sistema, trajanje izpada, zmanjšanje zmogljivosti sistema v času izpada, stroški postavitve in stroški delovanja sistema. Ocene so povzete po Cary V. Millsap: Configuring Oracle Server for VLDB.

Obravnavali bomo le nivoje RAID 0, 1, 10, 3 in 5. Ti nivoji so v praksi najbolj razširjeni. Seveda je možno

Nivo RAID	Varnost podatkov	Odzivni čas pri branju	Odzivni čas pri pisanju	Odzivni čas pri obnovi sistema	Najmanjše potrebno število diskov	Primerna okolja za uporabo
RAID 0	slaba	zelo dober	zelo dober	N/A	N	nekritični podatki
RAID 1	odlična	zelo dober	dober	dober	2N x X (X = število nizov RAID)	majhne baze podatkov, transakcijski dnevnik, kritični podatki
RAID 2	dobra	zelo dober	dober	dober	N + 1	N/A
RAID 3	dobra	zaporedno branje: zelo dober transakcijski sistemi: slab	zaporedno pisanje: dober transakcijski sistemi: slab	ugoden	N + 1 (N = najmanj 2)	enouporabniška podatkovno intenzivna okolja (na primer obdelava videa)
RAID 4	dobra	zaporedno branje: dober transakcijski sistemi: dober	zaporedno pisanje: zelo dober transakcijski sistemi: slab	ugoden	N + 1 (N = najmanj 2)	baze podatkov in druge transakcijske obdelave z intenzivnim branjem podatkov
RAID 5	dobra	zaporedno branje: dober transakcijski sistemi: zelo dober	ugoden, razen če uporabljamo predpomnilnik za ponovno zapisovanje	slab	N + 1 (N = najmanj 2)	baze podatkov in druge transakcijske obdelave z intenzivnim branjem podatkov
RAID 6	odlična	zelo dober	slab	slab	N + 2	majhne in srednje velike baze podatkov s potrebo po veliki razpoložljivosti
RAID 10	odlična	zelo dober	ugoden	dober	2N x X (X = število RAID nizov)	podatkovno intenzivna okolja (dolgi zapisi)
RAID 30, RAID 50	odlična	zelo dober	ugoden	ugoden	N + 2 (N = najmanj 4, X = število RAID nizov)	srednje velike transakcijske baze podatkov in baze z velikim številom transakcij

Tabela 2: Primerjave nivojev RAID (Vir: http://www.del.com)

uporabiti tudi drugačne konfiguracije, predvsem hibridne ali sestavljene in pa konfiguracije z več kot enojnim senčenjem (3). O njihovih lastnostih se da sklepati na podlagi lastnosti osnovnih nivojev RAID.

3.1 RAID 0

Pravilno konfiguriran sistem RAID 0 lahko nudi izredno dobre rezultate pri vseh vrstah vhodno-izhodnih operacij, zaporednem (*sequential*) in naključnem (*random*) branju in zapisovanju podatkov. Tak sistem pride vpoštev le, kadar potrebujemo najcenejšo rešitev in zanesljivost ter razpoložljivost sistema nista tako pomembni.

Ocena sistema:

- Naključno branje: odlično, posebej če velikost zahtevanih podatkov, ki naj bi bili prebrani, sovpada z velikostjo bloka. Kadar so bloki premajhni, se zmogljivost pri naključnem branju lahko drastično poslabša.
- Naključno zapisovanje: enako kot branje.
- Zaporedno branje: odlično. Podobno kot pri naključnem branju je tudi tukaj zelo važna velikost posameznega bloka, ki mora biti usklajena z velikostjo bloka v Oraclovi bazi podatkov.
- Zaporedno zapisovanje - enako kot branje.
- Pogostost izpada sistema: velika. Izpad vsakega diska onemogoči delo baze podatkov in zahteva ponovno vzpostavitev sistema s pomočjo rezervnih kopij.
- Trajanje izpada: slabo. Pri vsakem izpadu je treba napako odkriti, pokvarjeno komponento zamenjati in obnoviti podatke. To je dolgotrajen proces.
- Zmanjšanje zmogljivosti v času izpada: slabo. Sistem v času izpada sploh ne deluje.
- Stroški postavitve sistema: odlično. RAID 0 je najcenejši od vseh RAID sistemov.
- Stroški delovanja: zelo slabo. Vsakokratno obnavljanje podatkov ob odpovedi posamezne komponente

zelo poveča skupne stroške sistema. Prav tako so postopki ob širjenju sistema z dokupom novih diskov precej zapleteni, kajti ves sistem je treba na novo postaviti.

3.2 RAID 1

Zrcaljenje diskov je najboljša tehnika za zmanjšanje pogostosti odpovedi. Rešitev je zelo uporabna, kadar želimo administratorju baze podatkov omogočiti razne posege v bazo, ki si jih pri drugih, cenejših sistemih ne bi mogli privoščiti.

Ocena sistema:

- Naključno branje: dobro. Obstajajo krmilniki, ki se za vsako branje posebej odločajo, katero od obeh kopij se splača uporabiti. Drugi diski lahko medtem delajo kaj drugega. Če krmilnik ne zna delati z vsakim diskom posebej, potem je zmogljivost enaka kot pri disku brez redundantnih podatkov.
- Naključno zapisovanje: dobro. V primeru, ko krmilnik zna delati z vsakim diskom posebej, je zmogljivost pri naključnem zapisovanju slabša kot pri samostojnem disku, če pa krmilnik omenjene lastnosti nima, potem je hitrost zapisovanja enaka samostojnemu disku.
- Zaporedno branje: zadovoljivo, enako kot pri navadnem, samostojnem disku.
- Zaporedno zapisovanje: zadovoljivo, enako kot pri navadnem, samostojnem disku.
- Pogostost izpada sistema: odlično. To je sistem, ki je najmanj občutljiv na odpovedi posameznih komponent, še posebej pri zrcaljenju na več diskov.
- Trajanje izpada: odlično. Če odpove samo ena komponenta, izpada sistema pravzaprav ni, ampak govorimo o delnem izpadu. Če odpovesta obe kopiji ali vse, če jih je več, seveda pride do izpada sistema.
- Zmanjšanje zmogljivosti v času izpada: odlično. Če odpove en disk, se hitrost delovanja ne spremeni. Po zamenjavi z novim se med obnovo podatkov hitrost začasno zmanjša.
- Stroški postavitve sistema: slabo. Potrebujemo dvakratno (ali celo večkratno) število diskov in RAID 1 krmilnike, ki sicer v primerjavi z RAID 3 in 5 niso najdražji, vendar vseeno dražji od navadnih.
- Stroški delovanja: zadovoljivo. Sistem ni najpreprostejši in najcenejši za vzdrževanje, vendar vseeno cenejši od tistih bolj zapletenih.

3.3. RAID 10

RAID 10 je kombiniran sistem, ki hkrati uporablja tehniko razporejanja blokov (RAID 0) in zrcaljenja (RAID 1). Tako dobimo zahvaljujoč zrcaljenju odlične rezultate z vidika neobčutljivosti na odpoved posameznih komponent in največje hitrosti pri vhodno izhodnih operacijah.

Ocena sistema:

- Naključno branje: odlično, če je nastavljena ustrezna velikost blokov. S krmilniki, ki znajo uporabljati optimizacijo RAID 1, kar pomeni branje samo enega diska, je hitrost branja celo večja kot pri RAID 0 sistemih.
- Naključno zapisovanje: odlično. Zaradi zahtev po dvojnem zapisovanju je sicer nekoliko slabše kot pri RAID 0, toda mnogo bolje kot pri RAID 5.
- Zaporedno branje: odlično. Enako kot pri naključnem branju je zelo pomembna usklajenost velikosti bloka z blokom v Oraclovi bazi.
- Zaporedno zapisovanje: odlično. Enako kot naključno zapisovanje.
- Pogostost izpada sistema: odlično. Enako kot RAID 1.
- Trajanje izpada: odlično. Enako kot RAID 1.
- Zmanjšanje zmogljivosti v času izpada: odlično. Enako kot pri RAID 1.
- Stroški postavitve sistema: slabo. Enako kot pri RAID 1, možni so celo dodatni stroški za zagotovitev principa RAID 0.
- Stroški delovanja: zadovoljivo. Tudi tukaj so seštete vse dobre in slabe lastnosti RAID 0 in RAID 1 sistemov. Vzdrževanje zahteva tehnično usposobljene kadre. Pri dograditvi sistema je potrebno letega na novo postaviti.

3.4. RAID 3

RAID 3 je odgovor na visoke stroške stoodstotnega podvajanja podatkov pri RAID 1. Ob odpovedi ene komponente sistem deluje dalje, medtem ko je komponento možno nadomestiti brez popolnega odklopa sistema, vendar je postopek zamuden in močno zmanjša čase izvajanja vhodno izhodnih operacij. Za večino baz podatkov ta rešitev ni primerna. Vpoštev pride le tam, kjer so stroški zelo pomembni in kjer je način dela tak, da ni veliko vpisovanja novih ter prepisovanja in brisanja obstoječih podatkov.

Ocena sistema:

- Naključno branje: slabo. Zaradi obveznega sinhroniziranega delovanja diskov je nemogoče vzporedno izvajati različne operacije.
- Naključno zapisovanje: slabo. Enako kot pri branju.
- Zaporedno branje: zelo dobro za sisteme z malo uporabniki, slabše za večuporabniške sisteme.
- Zaporedno zapisovanje: dobro za sisteme z malo uporabniki, slabše za večuporabniške sisteme.
- Pogostost izpada sistema: dobro. V primeru odpovedi ene komponente sistem ne odpove, pri odpovedi dveh pa je že potrebna obnovitev podatkov iz rezervnih kopij.
- Trajanje izpada: dobro. Ko izpad zaznamo, lociramo in zamenjamo pokvarjen disk, sistem sam vzpostavi podatke, ki manjkajo.

- Zmanjšanje zmogljivosti v času izpada: zadovoljivo. Dokler ne zamenjamo pokvarjene komponente, zmogljivost ni bistveno slabša. Ko pa namestimo nov disk in se sistem loti obnavljanja podatkov, zmogljivost začasno močno pade.
- Stroški postavitve sistema: zadovoljivo. Stroški odvečnih podatkov so relativno nizki, ker potrebujemo le en dodaten disk za nadzorne podatke. Potrebni so RAID 3 krmilniki, ki so v primerjavi s tistimi za RAID 0 in 1 dražji.
- Stroški izpada: zadovoljivo. Za vzdrževanje je potreben kader s posebnimi znanji, pri razširitvi je potrebno na novo postaviti sistem.

3.5. RAID 5

V RAID 5 sistemih je najbolj problematično zapisovanje podatkov, zaradi česar za baze podatkov ni priljubljen, je pa na trgu zelo razširjen na drugih področjih.

Ocena sistema:

- Naključno branje: odlično, če je sistem pravilno postavljen.
- Naključno zapisovanje: slabo. Delno lahko pomaga diskovni predpomnilnik, vendar se pri večjih količinah zapisanih podatkov tudi ta zapolni in sistem postane počasen.
- Zaporedno branje: odlično. Enako kot pri naključnem branju je pomembna pravilna postavitev sistema oziroma pravilna nastavitev parametrov.
- Zaporedno zapisovanje: dobro za sisteme z manj uporabniki. Pri večuporabniških sistemih in intenzivnem zapisovanju se diskovni predpomnilnik zapolni in zmogljivost močno pade.
- Pogostost izpada sistema: dobro. Sistem brez zaustavitve prenese odpoved enega diska, pri dveh istočasno pokvarjenih komponentah pa je potrebno obnoviti podatke iz rezervne kopije in medtem seveda sistem ne deluje.
- Trajanje izpada: dobro. Delni izpad traja toliko časa, da lociramo in zamenjamo odpovedano komponento.
- Zmanjšanje zmogljivosti v času izpada: zadovoljivo. V času delnega izpada, preden zamenjamo pokvarjeni disk, sistem deluje dobro, odzivnost se bistveno ne poslabša. Ko namestimo novo, prazno komponento in ko se prične samodejna obnova podatkov na novem disku, se delovanje sistema močno upočasni.
- Stroški postavitve sistema: zadovoljivo. Potrebujemo le en disk za odvečne podatke. Čim več diskov je v nizu, tem manjši je vpliv dodatnega diska na ceno celotnega sistema. Seveda se z večanjem niza zanesljivost sistema manjša. Krmilniki v RAID 5 so zapleteni in dragi, vendar precej razširjeni, kar jim kljub dragi proizvodnji znižuje ceno.
- Stroški delovanja: zadovoljivo. Inženirji, ki sistem vzdržujejo, morajo biti za to usposobljeni in imeti kar nekaj izkušenj. Ob razširitvi sistema je potrebno dokupiti cel nov niz ali če želimo dodati nove diske k že obstoječim nizom, je treba na novo določiti parametre.

3.6. Povzetek uporabe RAIDa v okolju Oracle

Tehnologija in cena diskov, vodil in krmilnikov se spreminjata. Zato je težko reči: »takšna konfiguracija je v takšnem primeru najboljša«. Vsekakor je najboljši pristop redno preverjanje razmer na trgu, to je razpoložljive ponudbe in vzdrževanje ter nadgrajevanje sistema, tako kot to narekujejo okoliščine in dovoljuje proračun (4).

V tabeli 3 (4) so predstavljene ocene delovanja posameznih RAID nivojev za specifične Oraclove datotečne podsisteme. Tabela je zasnovana tako, da je

Nivo RAID	Brez RAIDa	0	1	10	3	5
Zmogljivost z vidika kontrolnih datotek	2	1	2	1	5	3
Zmogljivost z vidika datotek *redo-log*	4	1	5	1	2	3
Zmogljivost z vidika sistemskih tabel	2	1	2	1	5	3
Zmogljivost z vidika sortiranja	4	1	5	1	2	3
Zmogljivost z vidika sistema *rollback*	2	1	2	1	5	5
Branje indeksiranih tabel	2	1	2	1	5	1
Zaporedno branje tabel	4	1	5	1	2	3
Intenzivno zapisovanje v bazo podatkov	1	1	2	1	5	5
Zaščita podatkov	4	5	1	1	2	2
Pristopni stroški in stroški vzdrževanja	1	1	5	5	3	3

Tabela 3: Primerjava nivojev RAID za baze podatkov Oracle

sistem RAID 10 vedno ocenjen z oceno 1 (najboljše), razen seveda pri stroških, kjer ima oceno 5 (najslabše). Drugi sistemi so ocenjeni glede na RAID 10.

4. Zaključek

Povečane zahteve po zmogljivosti strežniških sistemov in po večji zanesljivosti ob stalnem zniževanju cen sistemov za shrambo podatkov spodbujajo industrijo k razvoju vedno novih sistemov RAID in drugih dodatkov za splošno povečanje zmogljivosti prenosov V/I. Tehnike RAID, ki so bile doslej pretežno v domeni velikih sistemov, postajajo vedno bolj dostopne tudi manjšim podjetjem.

Sistemi RAID živijo v praksi že toliko časa in so že tako poznani, da so vsi proizvajalci podatkovnih baz, tudi Oracle, prilagodili svoje izdelke takim sistemom in izdali navodila in priporočila za to področje. Vendar pa se tudi programje baz podatkov hitro spreminja. Nove, izboljšane verzije, včasih z novimi funkcionalnostmi, včasih zanesljivejše, mnogokrat tudi ne, prihajajo kot po tekočem traku. Te novosti na področju programske opreme skupaj z novostmi in spremenjenimi karakteristikami strojne opreme strokovnjakom, ki morajo skrbeti za baze podatkov in načrtovati njihov razvoj, ne dovolijo počitka. Zahteve po razpoložljivosti delujoče baze podatkov se iz dneva v dan večajo. Varnost podatkov in zanesljivost, da ne pride do izgube podatkov, sta prav tako vedno pomembnejši. V času, ko se poslovanje seli na spletne strani, ko postaja zahteva po stalni dostopnosti podatkov vsakdanja, ko si ne moremo privoščiti zaustavitve sistema niti za izdelavo rezervnih kopij in sistemske posege in se cena izpada sistema drastično dviga, bodo sistemi RAID pridobivali na pomembnosti. V prihodnosti se pričakuje stopnjevanje takih zahtev, zato bo na teh področjih (varne, zanesljive baze podatkov, hitri, vedno večji diskovni sistemi) za strokovnjake in raziskovalce zagotovo dovolj dela in izzivov.

Pred uvedbo tehnike RAID je priporočljivo analizirati potrebe podjetja, predvsem dobro spoznati naravo baz podatkov, kajti pri posameznih nivojih RAID lahko z napačno izbiro dosežemo minimalne ali celo slabše rezultate kot pri navadnih diskih. Zaradi vse večjega pomena je razvoj na področju sistemov RAID in opreme za arhiviranje podatkov dokaj hiter in ga je potrebno spremljati, če želimo izbrati med vedno cenejšimi in boljšimi sistemi tistega, ki je najprimernejši za reševanje nalog v določenem praktičnem okolju, saj je dandanes zanesljiv in učinkovit informacijski sistem temelj dobrega poslovanja.

5. Literatura

1. Andrew S. Tanenbaum: Structured Computer Organization, Prentice Hall, 1999, četrta izdaja
2. Raid Technology White Paper, http: //www.dell.com/us/en/biz/topics/vectors_1999-raid.htm, marec 1999, 04.02.2001
3. E. Aronoff, K. Loney, N.Sonawalla: Oracle8 Advanced Tuning & Administration, Osborne/McGraw-Hill,
4. Cary V. Millsap: Configuring Oracle Server for VLDB, 01.03.1996,
5. Raid, http: //www6.tomshardware.com/storage/00q1/000329/, 04.02.2001
6. Raid Technology, The storage solution, http: // www.nstor.com, 04.02.2001
7. http: //www.fibrechannel.com, 04.02.2001
8. http: //www.raid-advisory.com, 12.05.2001
9. Abit Computer Corporation, BX133-RAID Users Manual, Rev. 1.00, Maj 2000
10. http://www.raid-advisory.com/rabguide.html#abouttherab, 14.03.2001
11. http://www.storagesearch.com/nstorart.html, 20.07.2001

◆

Marija Kuhar je diplomirala na višješolskem študiju pedagoške matematike in fizike na Fakulteti za naravoslovje in tehnologijo, smer matematika. Nato je pridobila univerzitetno izobrazbo na Fakulteti za organizacijske vede, smer Organizacijska informatika. Trenutno je vpisana na magistrski študij na Fakulteti za organizacijske vede, smer Management informacijskih sistemov. Zaposlena je v podjetju Grad d.d., kjer se ukvarja z razvojem integriranih informacijskih sistemov za srednja in mala podjetja.

◆

Borut Vovk je diplomiral na Fakulteti za organizacijske vede Univerze v Mariboru, smer informatika. Od takrat dalje je zaposlen v Gorenjski banki d.d. Kranj v Sektorju za informacijske sisteme. Sprva je opravljal dela programerja in analitika, od 1998 dalje pa je administrator baze podatkov. Njegovo delo je upravljanje in nadzor delovanja Oraclove baze podatkov. Prav tako se ukvarja z načrtovanjem in razvojem baze podatkov, podatkovnega skladišča in pripadajoče programske opreme v banki.

◆

Miro Gradišar je izredni profesor poslovne informatike. Diplomiral in magistriral je na Fakulteti za elektrotehniko v Ljubljani, doktoriral pa na Fakulteti za organizacijske vede v Kranju. Zaposlen je na Ekonomski fakulteti v Ljubljani, kjer predava predmete s področja poslovnih informacijskih sistemov, sodeluje pa tudi s Fakulteto za organizacijske vede. Osnovno raziskovalno področje je gradnja računalniških modelov za simulacijo in optimizacijo poslovnih procesov. Kot avtor ali soavtor je objavil pet knjig, 32 znanstvenih člankov in 46 referatov na domačih in tujih znanstvenih in strokovnih konferencah.

◆