ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2000
Seria: ELEKTRONIKA z.12 Nr kol. 1492

Jerzy IHNATOWICZ
Politechnika Śląska, Instytut Elektroniki

MACIERZE DYSKOWE – BUDOWA, ZASTOSOWANIE, OGRANICZENIA

Streszczenie. W pracy dokonano próby uporządkowania podstawowych pojęć związanych z macierzami dysków komputerowych. Omówiono podstawowe koncepcje systemów RAID, ich właściwości oraz możliwości zastosowań. Omówiono rezultaty uzyskane w przykładowym systemie RAID, którego budowę oparto na pasywnej płycie głównej EISA.

THE IDEA, CONSTRUCTION AND APPLICATION OF COMPUTER DISK ARRAYS

Summary. In the paper the idea of the computer disk computer disk array is presented. The basic concepts of RAID subsystems and their properties are discussed. The practical results were verified with the model of RAID subsystem in the EISA passive backplane computer.

1. System RAID

1.1. Informacje ogólne

W roku 1987 Patterson, Gibson i Katz z Uniwersytetu Kalifornijskiego w Berkeley opublikowali artykuł pod tytułem „Macierze nadmiarowe złożone z niedrogich dysków (RAID)" („A Case for Redundant Arrays of Inexpensive Disks (RAID)"). W artykule tym opisano różne typ[y macierzy dyskowych nawiązujące do skrótu „RAID". Podstawową koncepcją RAID było złożenie wielu niedrogich dysków o stosunkowo małej pojemności w zespół – macierz dyskową – o pojemności i wydajności przekraczającej pojedynczy, kosztowny dysk o dużej pojemności (SLED – Single Large Expensive Drive). Taka macierz dyskowa jest obsługiwana przez system operacyjny komputera jak pojedynczy – fizyczny lub logiczny – napęd dyskowy.

Średni czas międzyawaryjny (MTBF) macierzy dyskowej jest mniejszy niż dla pojedynczego dysku i może być oszacowany jako MTBF pojedynczego dysku macierzy podzielony przez liczbę dysków tworzących tę macierz. Jest to powodem, dla którego MTBF macierzy dyskowej jest zazwyczaj zbyt mały dla praktycznych zastosowań. Odpowiednia struktura macierzy dyskowej, wykorzystująca zapis informacji nadmiarowej, może być jednak odporna na błędy i uszkodzenia dysków.

W artykule Pattersona i współpracowników omówiono architekturę pięciu typów macierzy dyskowych zapewniających odporność na błędy; różniących się wydajnością i możliwościami. Wprowadzenie na rynek przez czołowych producentów sterowników dysków komputerowych (DPT, Adaptec Inc.) obsługi systemów RAID bez korekcji błędów spopularyzowało dodatkową strukturę RAID określaną potocznie jako RAID-0.

1.2. Przelot dyskowy

Podstawowym dla zrozumienia zasady tworzenia macierzy dyskowych pojęciem jest "przeplot dyskowy", dający efekt połączenia (konkatenacji) wielu dysków w jedną jednostkę - dysk logiczny. (Zauważmy, że większość użytkowników komputerów jest przyzwyczajona do przeciwieństwa tej techniki, to znaczy do podziału jednego dysku fizycznego na szereg jednostek logicznych).

Przeplot dyskowy polega na podziale każdego dysku fizycznego na bloki, których pojemność może wynosić od jednego sektora (512 bajtów) do wielu megabajtów. Każdy kolejny blok jest związany z innym dyskiem, tak że zapis lub odczyt pliku z macierzy dyskowej jest związany z operacjami wejścia/wyjścia więcej niż jednego dysku fizycznego. Operacja taka może być porównana do rozdawania talii kart między poszczególnych graczy. Dobór rozmiarów bloków biorących udział w przeplocie jest zależny od przewidywanych zastosowań macierzy dyskowej.

Większość współczesnych systemów operacyjnych obsługujących jednocześnie wielu użytkowników, jak na przykład Unix czy oprogramowanie sieciowe Novell, zapewnia obsługę pojawiających się jednocześnie operacji zapisu/odczytu przez kilka dysków fizycznych. Łatwo jednak sprawdzić, że uzyskanie dobrej wydajności całego systemu wymaga takiego zrównoważenia operacji zapisu/odczytu między dyskami, aby każdy z nich był obciążony równomiernie. W typowych systemach wielodyskowych bez przeplotu pożądana równowaga obciążenia nie jest możliwa do uzyskania. Niektóre dyski zawierają bowiem pliki wykorzystywane częściej; żądanie dostępu do innych dysków jest zaś dosyć rzadkie. (Zwykle administratorzy systemów stosując metodę prób i błędów starają się odpowiednio skonfigurować zawartości poszczególnych dysków - uzyskany rezultat jest jednak zazwyczaj niewspółmiernie znikomy w stosunku do nakładu pracy). Stosując jednak przeplot dyskowy z wykorzystaniem bloków o rozmiarach odpowiadających długościom rekordów plików można uzyskać efekt równomiernego obciążenia wszystkich dysków tworzących macierz. Wszystkie dyski fizyczne będą więc aktywne podczas zwiększonego

obciążenia systemu. To zaś pozwoli na zwiększenie równolegle przebiegających operacji zapisu/odczytu w całej macierzy dyskowej.

1.3. Poziomy RAID

Struktura **RAID-0** jest określona jako beznadmiarowa, autonomiczna grupa przeplotu dyskowego. Macierz RAID-0 wykorzystuje zwykle duże bloki, jakkolwiek można również zastosować bloki o rozmiarach pojedynczych sektorów pod warunkiem użycia dysków z wzajemną synchronizacją wirowania. (Dyski takie - jak na przykład zastosowane przez autora niniejszej pracy dyski Micropolis 2112 - są jednak trudno dostępne w Polsce i przez to kosztowne). Synchronizacja wirowania dysków w macierzy dyskowej daje znakomite rezultaty - szczególnie w przypadku wykorzystywania macierzy dyskowej RAID-0 do obsługi dużych plików obrazów cyfrowych.

Awaria pojedynczego dysku powoduje awarię całej macierzy RAID-0 i utratę wszystkich zapisanych informacji. Struktura RAID-0 nie może być więc polecana jako wyłączny system przechowywania informacji, lecz raczej jako najszybszy i najbardziej wydajny system RAID przeznaczony do bieżącej obsługi transakcji między regularnym składowaniem (back-up) obsługiwanych zbiorów.

Struktura **RAID-1**, określana potocznie mianem "duplikowania dysków", jest złożona z par dysków fizycznych. Każdy dysk danej pary przechowuje identyczną informację, jakkolwiek para jest traktowana przez system operacyjny jak jeden dysk fizyczny. W systemie RAID-1 złożonym z par pojedynczych dysków nie stosuje się przeplotu dyskowego; każdy logiczny dysk pary w strukturze RAID-1 może być jednak podsystemem typu RAID-0. Aktualizacja zapisu informacji w systemie RAID-1 wymaga zapisu na każdy dysk pary; odczyt może odbywać się niezależnie. Wynika stąd, że duplikowanie dysków podwaja wydajność systemu w trybie czytania, pozostawiając wydajność w trybie zapisu bez zmian. Struktura RAID-1 jest najbardziej wydajna ze wszystkich struktur nadmiarowych (odpornych na błędy) w wielodostępnych systemach operacyjnych.

Struktura **RAID-2** wykorzystuje ideę przeplotu dyskowego, z tą różnicą, że niektóre z dysków fizycznych są przeznaczone do przechowywania kodów korekcyjnych ECC (Error Correction Codes). W większości współcześnie produkowanych dysków komputerowych kod korekcyjny jest zapisywany wewnątrz każdego sektora, stąd struktura RAID-2 nie wnosi nic ponad możliwości RAID-1 i nie jest obsługiwana w większości sterowników RAID. RAID-2 jest stosowana raczej wyjątkowo, tam gdzie użytkownik ma możliwość wyboru sposobu niskopoziomowego formatowania dysku i maksymalnego wykorzystania nośnika z pominięciem sektorów kontrolnych.

Macierz dyskowa typu **RAID-3** wykorzystuje (podobnie jak RAID-2) przeplot dyskowy w grupie dysków, z tym że jeden dysk każdej grupy jest przeznaczony do przechowywania dodatkowego kodu korekcyjnego oraz że bloki odpowiadają zwykle rozmiarom sektorów dysków. RAID-3 korzysta z ECC każdego z sektorów do stwierdzenia, czy informacja nie została przekłamana. W takim przypadku odtworzenie prawidłowej

informacji odbywa się przez wykonanie operacji OR (XOR) informacji z dysku "korekcyjnego" z informacją odczytywaną z pozostałych dysków. Optymalizacja wydajności systemu RAID-3 jest zapewniona przez zapis sektorów kolejno we wszystkich dyskach. Każda operacja zapisu/odczytu wymaga jednak dostępu do wszystkich dysków w macierzy, stąd współbieżna realizacja tych operacji nie jest możliwa. Macierze typu RAID-3 sprawdzają się najlepiej w systemach jednozadaniowych do obsługi dużych plików. Dodatkowym wymaganiem dla RAID-3 jest konieczność synchronizacji wirowania dysków; spełnienie tego warunku zapobiega degradacji wydajności systemu RAID-3 w przypadku transferu krótkich plików. Systemy RAID-3 w wersji ulepszonej stanowią strukturę typu RAID-5, stąd w chwili obecnej w większości przypadków RAID-3 nie jest obsługiwana przez współczesne sterowniki.

Wady podobne do występujących w RAID-3 ma również struktura **RAID-4**. RAID-4 jest w zasadzie identyczna z RAID-3, z tym jednak wyjątkiem, że w przeplocie dyskowym stosowane są dłuższe bloki. Można zatem zapewnić odczyt całego rekordu z pojedynczego dysku macierzy (z wyjątkiem dysku korekcyjnego), co umożliwia współbieżną realizację operacji czytania. Operacje zapisu muszą się jednak odbywać pojedynczo, gdyż każdy zapis wymaga aktualizacji zapisu na dysku korekcyjnym. Podobnie jak w przypadku RAID-3, struktura RAID-4 została ulepszona do postaci RAID-5 i nie jest obecnie wykorzystywana.

Macierz dyskowa typu **RAID-5** jest popularnie określona jako macierz dyskowa z wędrującym dyskiem korekcyjnym. Struktura RAID-5 pozwala na uniknięcie ograniczeń spowodowanych koniecznością oczekiwania na zakończenie operacji aktualizacji zapisu na dysku korekcyjnym - jak w macierzach RAID-3 oraz RAID-4. Podobnie jak w przypadku RAID-4, w macierzy dyskowej typu RAID-5 stosuje się przeplot dyskowy z dłuższymi blokami, co pozwala na współbieżną realizację operacji zapisu/odczytu. W odróżnieniu od RAID-4, każdy z dysków grupy przechowuje informację korekcyjną dla kolejnej serii bloków. Zauważmy, że w odróżnieniu od RAID-3 i RAID-4, macierz RAID-5 nie posiada wydzielonego dysku do zapisu informacji korekcyjnej; każdy z dysków zawiera zarówno informację podstawową,jak i korekcyjną (dla pozostałych dysków). Takie rozwiązanie daje możliwość równoległej obsługi odczytu dla wszystkich dysków w macierzy. Operacje zapisu wymagają dostępu do dwóch różnych dysków: dysku dla zapisu informacji podstawowej oraz dysku dla zapisu informacji korekcyjnej. Zatem w odróżnieniu od struktury RAID-4, zapis informacji korekcyjnej odbywa się na różnych dyskach fizycznych, co umożliwia współbieżną realizację również operacji zapisu.

1.4. Podsumowanie właściwości systemów RAID

Omówione podstawowe własności macierzy dyskowych RAID można zestawić następująco:

- **RAID-0** jest najszybszą i najbardziej wydajną macierzą dyskową, ale jest całkowicie nieodporna na błędy;
- **RAID-1** można polecić w zastosowaniach, gdzie powinna być zapewniona

odporność na błędy przy jednoczesnej dużej wydajności systemu dyskowego. RAID-1 jest jedyną konfiguracją macierzy składającej się tylko z dwóch dysków i odporną na błędy;

- **RAID-2** jest bardzo rzadko stosowane, jako że systemy korekcji błędów dla pojedynczych sektorów dysków są wbudowane w większości współcześnie produkowanych dysków komputerowych;
- **RAID-3** może być polecony w zastosowaniach jednostanowiskowych do obsługi długich plików. Struktura RAID-3 zapewnia wzrost wydajności systemu i odporność na błędy. Równoległa obsługa operacji we/wy nie jest możliwa; wymagana jest również synchronizacja wirowania dysków;
- **RAID-4** nie oferuje nic ponad możliwości RAID-5 i to przy braku możliwości stosowania równoległych operacji zapisu;
- **RAID-5** jest najlepszym wyborem w sytemach wielodostępnych, w których przeważają operacje odczytu informacji. Do stworzenia macierzy typu RAID-5 wymagane są co najmniej 3 dyski.

1.5. Informacja nadmiarowa w systemach RAID

Systemy RAID-5 zapewniają lepsze wykorzystanie pojemności dysków w porównaniu z systemami RAID-1, jako że zamiast zapisywania pełnej kopii informacji (jak w RAID-1), w systemach RAID-5 oprócz informacji podstawowej zapamiętywane są jedynie dodatkowe informacje korekcyjne. W rezultacie z sumarycznej pojemności każdych trzech dysków skonfigurowanych w macierz RAID-5, pojemność dwóch jest wykorzystywana do zapisu informacji podstawowej. To uzasadnia stwierdzenie, dlaczego systemy RAID-5 są bardziej "oszczędne" niż systemy RAID-1. Zysk na pojemności użytkowej systemu RAID-5 wiąże się, niestety, z utratą jego wydajności.

Podczas zapisu danych do macierzy RAID-5 informacja korekcyjna musi zostać zaktualizowana. Istnieją dwa sposoby realizacji tego zadania. Pierwszym - określanym jako bezpośredni - jest pełny zapis nowej informacji korekcyjnej. Jest to, niestety, sposób bardzo powolny: nowa informacja korekcyjna jest różnicą symetryczną (XOR) danych zapisywanych na każdym z dysków. Stąd zmiana informacji zapisanej na jednym z dysków wymaga uaktualnienia informacji korekcyjnej przez wykonanie operacji XOR z informacjami dotyczącymi aktualizowanego rekordu i zapisanymi na dyskach pozostałych. W rezultacie wykonanie operacji zapisu wymaga dostępu do wszystkich dysków macierzy RAID-5.

Drugim ze sposobów, zazwyczaj znacznie bardziej efektywnym i określanym jako warunkowa aktualizacja kodu korekcyjnego jest wyszukiwanie zmienionych bitów w zapisywanych danych i zmiana odpowiadających im bitów korekcyjnych (parzystości). Jest to realizowane następująco: W pierwszym kroku czytane są dane poprzednie. Następnie jest wyznaczana różnica symetryczna danych poprzednich z danymi nowymi. Otrzymany wynik stanowi maskę bitową zawierającą jedynki na pozycjach, które mają być zmienione. Maska bitowa jest poddawana operacji XOR ze starą informacją korekcyjną, przeczytaną z dysku

korekcyjnego dla danego bloku. Rezultat określa zmiany, jakich należy dokonać w kodzie korekcyjnym, który ma być zaktualizowany. Zauważmy, że tak rozumiana aktualizacja warunkowa wymaga jedynie dwóch operacji odczytu, dwóch operacji zapisu i dwóch operacji XOR zamiast trójki operacji odczytu, XOR i zapisu dla każdego dysku macierzy RAID-5.

Ceną płaconą w systemie RAID-5 za uniknięcie konieczności przechowywania pełnej kopii informacji podstawowej (jak w RAID-1) jest powiększenie czasu obsługi transakcji zapisu o czas zapisu informacji korekcyjnej. Z wykonanych w ramach niniejszej pracy pomiarów wydajności systemów RAID wynika, że wydajność zapisu w systemie RAID-5 wynosi od 30 do 60 procent wydajności zapisu w systemie RAID-1 zbudowanym z identycznych elementów i pracującym w tym samym środowisku. Stąd trudno zalecać stosowanie macierzy dyskowej RAID-5 tam, gdzie wymagania co do ogólnej dużej wydajności systemu dyskowego są ostre. Wyjątek stanowią systemy RAID-5 przeznaczone do obsługi baz danych, w których nie przeprowadza się operacji zapisu, bądź też operacje zapisu są przeprowadzone stosunkowo rzadko (na przykład w bazach danych zawierających teksty źródłowe).

1.6. Odtwarzanie informacji utraconej w systemach RAID

Podstawową wadą macierzy dyskowej RAID-0 jest utrata wszystkich informacji w przypadku awarii jednego dysku macierzy. Systemy RAID-0 wymagają więc dysków o szczególnie wysokim MTBF, a i to nie daje dobrej gwarancji zachowania danych. W przypadku stosowania systemu RAID-1 błędy pojedynczego dysku mają niewielki wpływ na wydajność całego systemu gdyż potrzebne dane mogą być odczytane z przechowywanego duplikatu. W systemie RAID-5 sytuacja jest znacznie bardziej skomplikowana: potrzebne dane muszą zostać odtworzone jako różnice symetryczne XOR między informacją korekcyjną i danymi odczytanymi z pozostałych dysków dotyczącymi odtwarzanego bloku informacji. Proces ten przebiega dosyć powoli i jest nazywany trybem pracy zdegradowanej. Łatwo wykazać, że tryb pracy zdegradowanej będzie tym wolniejszy, im więcej dysków będzie tworzyło macierz RAID-5.

1.7. Możliwości wymiany uszkodzonych dysków w systemach RAID

Jedną z najbardziej spektakularnych zalet stosowania w instalacjach komputerowych macierzy dyskowych typu RAID-1 i RAID-5 jest możliwość wymiany uszkodzonych dysków bez zatrzymywania pracy całego systemu. W przypadku uszkodzenia lub wyłączenia dysku dane mogą być odtworzone na tak zwanym aktywnym dysku zapasowym (Hot Spare) lub na dysku wymienionym w miejsce uszkodzonego. Aktywny dysk zapasowy podczas normalnej pracy macierzy dyskowej znajduje się w stanie uśpienia (stand-by); żadna z realizowanych transakcji odczytu/zapisu z tego dysku nie korzysta. Dysk ten zostaje włączony do macierzy dopiero w sytuacji awaryjnej lub na zlecenie administratora systemu RAID. (Podobnie, każdy z dysków macierzy RAID-1 lub RAID-5 może stać się na żądanie dyskiem zapasowym).

Zauważmy, że odtworzenie stanu dysku w systemie RAID-1 jest zadaniem prostym: realizowana w tle operacja kopiowania całej zawartości dysku na dysk duplikujący przebiega stosunkowo szybko. Zadanie odtwarzania zawartości dysku może być traktowane jako obsługa dodatkowego klienta macierzy dyskowej. W przypadku systemu RAID-5 odtwarzanie danych wiąże się z koniecznością syntezy poszczególnych rekordów na podstawie wyników operacji XOR z danymi przechowywanymi na dyskach pozostałych. Powoduje to dodatkową degradację wydajności trybu pracy zdegradowanej, tym większą, im więcej dysków tworzy macierz RAID-5.

Większość nowoczesnych sterowników RAID podejmuje automatyczne odtwarzanie stanu dysku w momencie jego trwałego uszkodzenia, wzrostu stopy błędów transmisji ponad ustaloną wartość lub wyłączenia dysku, pod warunkiem że macierz dyskowa zawiera co najmniej jeden aktywny dysk zapasowy. W przypadku braku dysku typu Hot Spare, system RAID-1 pracuje bez kopii danych podstawowych, zaś system RAID-5 przełącza się w tryb pracy zdegradowanej. W każdym jednak z tych przypadków system RAID funkcjonuje poprawnie. Zauważyć należy, że wymiana dysku jest szczególnym przypadkiem włączenia do systemu RAID dysku typu Hot Spare. Niektóre, bardziej zaawansowane technologicznie sterowniki RAID umożliwiają wyłączanie z obsługi wskazanego dysku i zarezerwowania go jako Hot Spare bez przerywania pracy systemu. Oczywiście, w takim przypadku musi nastąpić całkowita przebudowa zapisów na wszystkich dyskach systemu RAID, co może na dość długi czas (w praktyce nawet na kilka godzin) spowolnić pracę systemu.

2. Pamięć buforowa sterownika systemu RAID

2.1. Informacje ogólne

Większość nowoczesnych sterowników dysków komputerowych SCSI-2 zawiera wbudowaną niewielkiej pojemności (rzędu 1 MB) pamięć buforową o szybkim dostępie. W połączeniu z pamięcią buforową o podobnej pojemności, stosowaną w lepszych dyskach, tworzy to dość wydajny system buforowania transakcji odczytu/zapisu. Niektóre sterowniki umożliwiają rozbudowę pamięci buforowej do 64MB, przy czym pamięć ta zawiera zwykle mechanizmy automatycznej korekcji błędów z wykorzystaniem ECC, co pozwala na zwiększenie szybkości pracy magistrali SCSI. Niestety, pamięć taka jest bardzo kosztowna (16MB ECC kosztuje około 1500 $ USA).

Zysk z buforowania operacji dyskowych metodami sprzętowymi jest szczególnie duży w przypadku programów typu CAD, które zazwyczaj bardzo intensywnie korzystają z dysków oraz w przypadku programów obsługi serwerów systemów wielodostępnych Unix i Novell. Znaczącą poprawę wydajności systemu można zaobserwować nawet w przypadku obsługi baz danych o dużej liczbie niewielkich rekordów.

W szczególności, duża pamięć buforowa pozwala na zdecydowaną poprawę wydajności macierzy dyskowej typu RAID-5. Warto jednak pamiętać, że istnieje wiele przypadków, w których stosowanie pamięci buforowej nie daje znaczącej poprawy obsługi

dysków. Do nich należą programy, które korzystają z dysku rzadko (na przykład programy arkuszy kalkulacyjnych), niewielkie (do 6 użytkowników) systemy wielodostępne oraz programy operujące na długich sekwencjach plików (przykładem może być edycja filmu cyfrowego).

2.2. Buforowanie programowe i sprzętowe

Większość współczesnych systemów operacyjnych wykorzystuje intensywnie buforowanie programowe w celu uzyskania poprawy wydajności dysków. Niekiedy uzyskana poprawa wydajności może być dodatkowo powiększona w rezultacie zastosowania dodatkowego buforowania sprzętowego. Zwykle jednak algorytmy obsługi sprzętowej pamięci buforowej sterownika dysku (lub samego dysku) dublują buforowanie programowe realizowane przez system operacyjny. Taki rozwiązanie jest skuteczne w przypadku prostych systemów operacyjnych o nieskomplikowanym systemie buforowania programowego, lecz w przypadku systemów złożonych typu Unix czy Novell, w których poprawa wydajności obsługi jest zrealizowana poprzez buforowanie wszystkich transakcji systemowych, sprzętowe zdublowanie buforowania nic w praktyce nie daje, jako że buforowane są te same dane. Co więcej, wydajność buforowania programowego podczas czytania będzie zawsze lepsza, jako że bufor programowy jest zawsze szybciej dostępny dla programu rezydującego w tej samej pamięci operacyjnej.

Efektywne wykorzystywanie buforowania sprzętowego wymaga stosowania skomplikowanej obsługi pamięci buforowej z wykorzystaniem sprzętowej realizacji algorytmów kolejkowania i przetwarzania strumieniowego. Dopiero zastosowanie takich układów pozwala na faktyczną poprawę działania systemów operacyjnych z rozbudowanym buforowaniem programowym.

Podstawowym celem buforowania programowego w systemach operacyjnych jest zmniejszenie liczby operacji czytania z dysku. Jest bowiem oczywiste, że jeżeli dane żądane przez wykonywany program mogą być pobrane z pamięci buforowej, to operacja czytania z dysku może być wyeliminowana.

Buforowanie programowe może również skrócić czas zapisu na dysk w tym sensie, iż potwierdzenie dokonania zapisu może zostać wygenerowane już po umieszczeniu danych w pamięci buforowej i wyprzedzać rzeczywisty zapis. Sam zapis może być przeprowadzony dopiero w chwili, gdy dysk będzie faktycznie dostępny. Takie rozwiązanie jest jednak skuteczne tylko w przypadku słabo obciążonych systemów, obsługujących niewielką liczbę użytkowników. Łatwo się również przekonać, że programowe buforowanie zapisu może prowadzić do załamania obsługi urządzeń wymagających nieprzerwanego strumienia danych (na przykład urządzeń nagrywających płyty CD). Podstawowym problemem staje się bowiem niewystarczająca częstotliwość pojawiania się chwil "jałowej pracy" systemu operacyjnego, w których zachodzi możliwość przepisania informacji z bufora programowego na dysk.

Słabo obciążone systemy wielodostępne, obsługujące niewielką liczbę użytkowników, charakteryzują się stosunkowo długimi i częstymi chwilami wyczekiwania (pracy jałowej),

lecz wraz ze wzrostem liczby użytkowników chwile te występują coraz rzadziej i są coraz krótsze. W rezultacie bufor programowy zostaje zapełniany tak zwanymi "sektorami do wyczyszczenia", to znaczy zapisami sektorów dyskowych oczekującymi na przesłanie do dysku. Oczywiście, takie sektory zajmują część bufora, który mógłby być lepiej wykorzystany do znacznie wydajniejszego buforowania operacji odczytu.

Warunkiem uzyskania wydajnego buforowania operacji zapisu jest przyspieszenie opróżniania bufora. Nowoczesne sterowniki umożliwiają ponad 20. krotny wzrost szybkości opróżniania bufora: w pewnej chwili sterownik przejmuje kontrolę nad magistralą systemu i przesyła blok informacji do swojej pamięci, pozostawiając pusty bufor. Z punktu widzenia systemu operacyjnego jest to równoważne ze skróceniem czasu dostępu do dysku ze współczynnikiem 20, co pozwala na współbieżną realizację operacji zapisu na dysk z innymi działaniami systemu operacyjnego.

2.3. Sortowanie transakcji

Dalsza poprawa wydajności obsługi dysku, a co za tym idzie - poprawa wydajności systemu RAID, wymaga stosowania układów sortowania transakcji, tak aby dane były przesyłane na dysk w optymalnej kolejności. Kolejność ta musi zapewniać narastające numery cylindrów, głowic i numerów sektorów dysku. Realizacja tego algorytmu eliminuje niepotrzebne ruchy ramienia głowic dyskowych: jednokrotne przemieszczenie głowic od cylindra o najniższym numerze do cylindra o numerze najwyższym zapewnia możliwość zapisu dla całej oczekującej na zapis kolejki sektorów z pamięci buforowej sterownika dysków macierzy RAID.

Przedstawiony algorytm jest znany jako "obsługa windy". Zauważmy, jak bardzo można spowolnić dostęp do windy, jeśli winda będzie realizowała przejazdy według kolejności zgłoszeń, a nie na odcinku między najwyższym i najniższym zgłoszonym piętrem, zabierając po drodze pasażerów. Bez sortowania zgłoszeń winda może jeździć między dwoma sąsiednimi piętrami cały dzień.

Dodatkową korzyścią związaną z sortowaniem transakcji dyskowych jest możliwość grupowania pojedynczych operacji zapisu w bloki, co wydatnie zmniejsza zajętość magistrali SCSI.

2.4. Zapis wymuszony

Każda operacja buforowania zapisu wiąże się z niebezpieczeństwem utraty danych niezapisanych na dysku, jeśli wcześniej nastąpi załamanie lub zawieszenie się systemu. Sprawa jest tym bardziej poważna, że użytkownik systemu dyskowego otrzymuje informację o zakończeniu operacji zapisu na dysk już w chwili zapisu do bufora. Jeśli nawet chciałby sprawdzić poprawność zapisu przez wydanie polecenia "czytaj z dysku", to i tak w większości przypadków nastąpi odczyt zawartości z bufora, w którym znajdują się jeszcze wysłane dane do zapisu. Zjawisko to jest dobrze znane użytkownikom systemów

wielozadaniowych, którzy dość często przekonują się, że uzyskanie pewności co do losów przesłanego pliku wymaga ponownego sprawdzenia po upływie co najmniej kilkunastu minut.

Z uwagi na znacznie większy niż w innych systemach obsługi dysków stopień defragmentacji plików użytkownika w macierzach RAID, dobre sterowniki RAID realizują tak zwany wymuszony tryb zapisu. Polega on na bezwzględnej realizacji operacji zapisu na dysk całej zawartości pamięci buforowej, o ile przez czas dłuższy niż ustalony (zwykle 250 milisekund) system operacyjny polecenia zapisu nie wyda. Operacja taka jest realizowana nawet jeśli czas gromadzenia w pamięci buforowej danych jest wielokrotnie dłuższy od samego zapisu. Zmniejszenie ryzyka utraty danych dzięki zapisowi wymuszonemu pogarsza, rzecz jasna, całkowitą wydajność systemu, zwłaszcza jeśli dane przeznaczone do zapisu są często aktualizowane, przez co aktualizacja danych mogłaby się odbywać w pamięci buforowej.

Drugim, niezależnym kryterium inicjalizacji zapisu wymuszonego, jest procent zajętości pamięci buforowej danymi do zapisu. Zapis wymuszony w tym przypadku nie dotyczy całej pamięci buforowej, lecz takiej liczby sektorów, aby utrzymać ustaloną minimalną wielkość stale dostępnej pamięci buforowej. Badania wykazały, że istnieją dwie wartości zapewniające w praktyce utrzymanie wystarczającej wydajności i pewności działania systemów RAID. Pierwsza dotyczy zapisu wymuszonego w chwili zajętości połowy pamięci; druga w sytuacji gdy pamięć jest zajęta w 80 procentach. W każdym z tych przypadków zapis wymuszony trwa aż do zwolnienia 80 procent pamięci buforowej.

2.5. Czytanie wyprzedzające

Ostatnim z najważniejszych mechanizmów zapewniających efektywną pracę macierzy dyskowych RAID jest tak zwane czytanie wyprzedzające. Tryb ten umożliwia obniżenie wartości średniego czasu dostępu do dysku związanego z operacjami czytania. W większości przypadków dane zapisane na dysku dotyczące pojedynczego pliku są zapisane w kolejnych sektorach (przypadek ten może dotyczyć również sektorowego przeplotu dyskowego!). Zatem dość prawdopodobne jest, że następna operacja czytania będzie dotyczyła kolejnych sektorów. Niektóre sterowniki dyskowe prowadzą na bieżąco analizę częstości odwołań do poszczególnych sektorów dysku i sektory sąsiadujące bezpośrednio z najczęściej czytanymi kopiują do pamięci buforowej. W przypadku odwołania się do tych sektorów są one dostępne bez konieczności oczekiwania.

Dość dobre rezultaty stosowania zasady czytania wyprzedzającego ulegają pogorszeniu w przypadku wielozadaniowych systemów obsługujących większą liczbę użytkowników. Wzrasta wtedy liczba sektorów które - ze względu na wysokie prawdopodobieństwo odwołania się do nich - należałoby przechowywać stale w pamięci buforowej. Dodatkowo ten sam użytkownik będzie mógł być obsłużony przez system dopiero po zakończeniu obsługi pozostałych. Jeśli tymczasem sektory przeczytane w trybie wyprzedzającym zostaną usunięte (na przykład w rezultacie zapisu wymuszonego), to wydajność całego systemu ulegnie pogorszeniu. Warto również zauważyć, że zliczanie częstości odwołań do poszczególnych

sektorów jest w przypadku większej liczby użytkowników dosyć kłopotliwe. Jedynym rozwiązaniem staje się powiększanie pojemności pamięci buforowej sterownika dyskowego.

Jak stwierdzono, w praktyce najlepsze rezultaty działania systemów RAID uzyskuje się przy rezerwacji około 30 procent pamięci buforowej na sektory czytane z wyprzedzeniem, przy jednoczesnym zapewnieniu co najmniej 0,5 MB pamięci buforowej dla każdego aktywnego użytkownika systemu RAID.

2.6. Dobór pojemności pamięci buforowej

Przy projektowaniu wydajnego systemu RAID należy uwzględnić wpływ operacji buforowania programowego stosowanego przez system operacyjny na macierz dyskową. Zauważmy, że dobrze dobrane parametry czytania wyprzedzającego mogą wyeliminować wiele pozostałych operacji odczytu. W każdym jednak razie każda operacja zapisu na dysk musi być - wcześniej czy później - zrealizowana. W rezultacie pamięć buforowa systemu operacyjnego działa jak filtr odczytów z macierzy dyskowej. W miarę powiększania pojemności pamięci buforowej wykorzystywanej przez system operacyjny wzrasta liczba transkacji nie wymagających wykonania operacji czytania z dysku (granicznym przypadkiem jest tutaj pamięć buforowa o pojemności równej pojemności dysku). Zatem liczba operacji czytania maleje, podczas gdy liczba operacji zapisu pozostaje stała. Zauważmy przeto, że powiększanie pamięci operacyjnej komputera (wykorzystywanej między innymi przez system operacyjny jako pamięć buforowa) nie poprawia wydajności w operacjach zapisu. Znacznie lepiej jest powiększyć pamięć buforową sterownika współpracującego bezpośrednio z macierzą dyskową typu RAID.

Praktyka dowodzi, że przydział zbyt dużego fragmentu pamięci operacyjnej dla buforowania operacji odczytu/zapisu przez system operacyjny może spowodować wzrost liczby operacji zapisu/odczytu związanych z koniecznością stosowania plików wymiany w sytuacjach gdy program użytkowy korzysta z dużych plików danych. W efekcie, wraz z powiększaniem się rozmiarów pamięci buforowej, wydajność całego systemu maleje. Doświadczenia wykazały, że bezpieczną granicą wielkości pamięci buforowej systemu operacyjnego jest nie więcej niż 10 procent całej dostępnej pamięci operacyjnej komputera. Pamięć buforowa sterownika dysków w macierzy dyskowej typu RAID powinna być natomiast jak największa.

3. Przykładowa konstrukcja systemu RAID

3.1. Komputer, elementy, środowisko programowe

W ramach niniejszej pracy zbudowano macierz dyskową RAID obsługującą 4 dyski SCSI-2 oraz zbadano zachowanie się i wydajność systemu dla zastosowań związanych z automatyczną analizą mikroskopowych obrazów cyfrowych.

Do budowy systemów RAID użyto dysków ST 31051N (1.05GB) oraz typu ST

32155N (2.15GB). Dodatkowo, do prób z systemem wykorzystującym dyski synchronizowane, użyto dyski Micropolis 2112 (1.05 GB). Zastosowano sterownik RAID firmy DPT z modułami DM 4000 oraz CM 4000/16. Moduł DM 4000, zrealizowany z wykorzystaniem jednoukładowego procesora PIC, służy do obsługi struktur macierzy dyskowej RAID typu 0, 1, 5 oraz do prognozowania numerów sektorów przeznaczonych do czytania wyprzedzającego i do porządkowania (sortowania) transakcji. Moduł CM 4000/16 to pamięć buforowa sterownika. Jest to pamięć typu ECC o pojemności 16 MB. Badania przeprowadzono dla pamięci buforowej o łącznej pojemności 64MB ECC. Komputerem użytym do badań był komputer wykorzystujący pasywną płytę główną EISA oraz kartę procesora PIC486 LP /16 współpracującą bezpośrednio ze sterownikiem DPT w trybie C - Bus Mastering (Burst Mode). Badania wydajności systemów RAID-0, RAID-1 oraz RAID-5 przeprowadzono w środowisku MS Windows. Praktyczną wydajność macierzy dyskowych testowano podczas prac związanych z automatyczną analizą ilościową mikroskopowych obrazów cyfrowych. Typowe przetwarzane pliki miały objętość 0,25 MB.

3.2. Przebadane struktury RAID

W ramach niniejszej pracy zbadano następujące układy macierzy dyskowych RAID:

- RAID-0 złożoną z dwóch niesynchronizowanych dysków ST 31051N z przeplotem dyskowym z podziałem od 512 bajtów do 128 kB;
- RAID-0 złożoną z dwóch synchronicznie wirujących dysków Micropolis 2112 z przeplotem dyskowym z podziałem 512 bajtów;
- RAID-1 złożoną z dwóch niesynchronizowanych dysków ST 31051N;
- RAID-5 złożoną z podsystemu RAID-0 (2 x ST 31051N) oraz dwóch dysków ST 32155N.

Błędy działania dysków wywoływano wymuszając zakłócenia zasilania poszczególnych dysków. Dodatkowo sprawdzano praktyczne możliwości pracy w trakcie rekonstrukcji struktury RAID z wykorzystaniem aktywnego dysku zapasowego (Hot Spare).

3.3. Uwagi końcowe

Przeprowadzone badania i testy wykazały, że najwyższy wzrost wydajności całego systemu komputerowego w pracach związanych z automatyczną analizą obrazów cyfrowych uzyskano dla macierzy RAID-0 zbudowanej z dysków synchronizowanych i z przeplotem jednosektorowym (512B). Czytanie wyprzedzające było trafione, oczywiście, w ponad 99% przypadków. Jest to zrozumiałe, bo przy dwóch dyskach pracujących z krótkim przeplotem (512B) prawdopodobieństwo znalezienia dalszej części pliku w następnym sektorze jest bardzo bliskie jedności. Brak synchronizacji wirowania w strukturze RAID-0 nie spowodował widocznego dla użytkownika spowolnienia systemu, utrata wydajności o około 60% była natomiast dotkliwa podczas kopiowania długich plików. Po zwiększeniu przeplotu do 1 kB

nastąpił radykalny wzrost wydajności; dalsze powiększanie przeplotu nie wniosło już znacznych zmian.

Średni czas dostępu do dysku pracującego w macierzy spadł z około 8 ms do 0,1 ms i w zasadzie nie zależał od wielkości przeplotu. Ogólna szybkość transmisji zwiększyła się natomiast około dwukrotnie w porównianiu z szybkością dla tego samego dysku pracującego niezależnie. W przypadku długich bloków przeplotu spadała liczba trafionych czytań wyprzedzających, typowa wartość to od około 40 do 60%. Jest to jednak bardzo dużo w porównaniu z liczbą trafień dla dysku pracującego niezależnie, dla którego maksymalna liczba trafień po około 3 godzinach pracy nie przekroczyła nigdy 10-12%.

Badania macierzy RAID-1 i RAID-5 dowiodły rzeczywistej odporności systemu dyskowego na błędy. Wyłączanie z obsługi dysku duplikującego przebiegało całkowicie niezauważenie dla użytkownika. Wydajność zapisu w systemie RAID-1 była w przybliżeniu taka sama, jak dla dysku pracującego niezależnie, zauważalna była natomiast wydajność odczytu. Odtwarzanie kopii informacji podstawowej przebiegało praktycznie niezauważenie podczas normalnej pracy; widoczna degradacja wydajności występowała natomiast podczas przeszukiwania dysku.

Podobne wyniki uzyskano dla macierzy RAID-5, z tym że odtwarzanie całości zapisu z uszkodzonego dysku trwało - przy normalnym obciążeniu systemu - ponad godzinę. Pierwszy dostęp do odtwarzanej informacji był w niektórych przypadkach długi i wynosił do kilkudziesięciu sekund; użytkownik miał wrażenie, że system się zawieszał. W miarę upływu czasu, po odtworzeniu często czytanych rekordów, system działał już tylko nieco wolniej.

Literatura

1. CAE Specification: Data Storage Management: A Systems Approach Prentice Hall. New York 1997.
2. Gibson G.A.: Redundant Disk Arrays: Reliable, Parallel Secondary Storage MIT Press, Los Angeles 1992.
3. Wicker S.B.: Error Control systems for Digital Communication and Storage. Prentice Hall, New York 1995.

Recenzent: Prof.dr hab.inż. Janusz Walczak

Abstract

This paper describes various types of disk arrays, referred to by the acronym "RAID". The basic idea of RAID is to combine multiple small disk drives into an array of disk drives which yields performance exceeding that of single drive. This array of drives appears to the computer as a single logical storage unit or drive. Fundamental to "RAID" is "stripping", a method of concatenating multiple drives into one logical storage unit. These stripes are interleaved round-robin, so that combined space is composed alternately of stripes from each drive. Six types of array architectures, RAID-0 through RAID-5 were described, each offering different trade-offs in features and performance. Some of them are supported by modern SCSI-2 disk controllers. The fault-tolerant arrays RAID-1 and RAID-5 can be effectively used even for some special, disk space consuming applications as digital image processing. The best solution for high-efficient RAID-0 system is using of synchronized spindle drives.

The crucial point for disk array is the read/write caching system. The recommended solution is to increase the very fast cache memory of array disk controller, leaving the standard software caching unchanged. This is very important for multitaskk, multiuser operating systems.