URZĄD WOJEWÓDZKI W SZCZECINIE

INSTYTUT BADAŃ SYSTEMOWYCH

Polskiej Akademii Nauk, Oddział w Szczecinie

MODELOWANIE ORGANIZACJI I SYSTEMY INFORMATYCZNE W GOSPODARCE REGIONU

Szczecin 1993

Modelowanie organizacji i systemy informatyczne w gospodarce ...

MODELOWANIE ORGANIZACJI I SYSTEMY INFORMATYCZNE W GOSPODARCE REGIONU

Praca pod redakcją
Prof. dr hab. Zygmunta DOWGIAŁŁO

Szczecin 1993

Publikacja zawiera referaty i doniesienia przygotowane na ogólnopolską konferencję zorganizowaną przez Urząd Wojewódzki w Szczecinie i Instytut Badań Systemowych PAN, Oddział w Szczecinie

Wykonano z oryginałów tekstowych dostarczonych przez autorów referatów

Publikacja finansowana ze środków Biura ds. Administracji Publicznej Urzędu Rady Ministrów

ISBN 83 - 85847 - 20 - 0

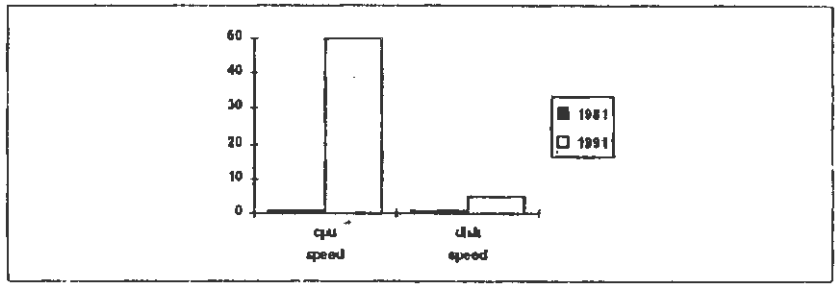
DRUK ZAKŁAD POLIGRAFICZNY
ul. Ku Słońcu 97, 71-046 SZCZECIN tel. 758-04

Marek JANAS
Krzysztof PANKOWSKI

MACIERZE DYSKOWE - CZY I DLACZEGO?

Potrzeba stosowania macierzy dyskowych

W ciągu ostatnich dziesięciu lat moc obliczeniowa komputerów klasy PC wzrosła ponad 50-krotnie. Pozwoliło to zbudować komputery o mocach dawniej zarezerwowanych dla większych i droższych maszyn. Jednak wzrost szybkości urządzeń pamięci masowych nie był aż tak duży. Na rysunku 1 pokazano skalę różnic tego postępu.



Pomimo tego, że pojemności dysków twardych znacznie wzrosły a ich ceny zmalały, to "osiągi" dysków wzrosły zaledwie trzy, czterokrotnie. Z tego względu najsilniejszymi maszynami obliczeniowymi lat 90-tych będą te, w których uda się zoptymalizować wydajność systemów dyskowych. Gdyby wydajność ta wzrosła w tej samej skali co moc obliczeniowa jednostek centralnych, mielibyśmy dziś prawdziwe superkomputery.

Telefoniczny System Techniczny to grupa wysokowyspecjalizowanych pracowników wyposażonych w odpowiednie narzędzia, którzy potrafią rozwiązać 90% problemów przez telefon.

Mamy przyjemność zaproponować państwu zakup sprzętu komputerowego renomowanej firmy Hyundai Electronics America, drukarek firmy OKI, Hewllet Packard oraz produktów innych renomowanych firm. Pragniemy zaznaczyć, że jesteśmy wyłącznym autoryzowanym dystrybutorem komputerów Hyundai na Pomorzu Zachodnim. Jesteśmy także autoryzowanym partnerem handlowym OKI Europe Ltd. Posiadamy status autoryzowanego dystrybutora lokalnego na obszar Pomorza Zachodniego produktów znanej firmy Data Technology Corporation (jest to producent renomowanych kart sieciowych i kontrolerów dysków twardych).

Zapraszamy do korzystania z naszej oferty. Dla nas każdy klient jest nowym członkiem wielkiej światowej rodziny miłośników PC-tów.

z wyrazami szacunku
Krzysztof Pankowski

Poziom RAID	Opis	Przyrost wydajności	Odporność na błędy
RAID 0	Striping	Równoległe operacje we/wy	Brak
RAID 1	Morroring	Brak	Tak (awaria jednego dysku)
RAID 2	Striping + detekcja błędów przy pomocy kodu Hamminga	Brak	Tak (awaria jednego dysku)
RAID 3	Striping z dodanym dyskiem parzystości	Równoległe operacje we/wy	Tak (awaria jednego dysku)
RAID 4	Striping z dodanym dyskiem parzystości: dyski nie nie synchronizowane	Równoległe operacje we/wy	Tak awaria jednego dysku)
RAID 5	Striping z rozproszoną parzystością	Równolegle operacje we/wy (wolniej niż RAID 0)	Tak (awaria jednego dysku

Spanning

Technika spanningu polega na traktowaniu kilku dysków w macierzy jako jednego dużego dysku. Pozwala to użytkownikowi na uniknięcie ograniczenia przestrzeni dyskowej poprzez łączenie istniejących już tanich zasobów i sukcesywne dodawanie nowych.

Na rysunku 2 pokazano przykładowo cztery 300 MB dyski połączone w jeden podsystem ze sterownikiem SCSI. Użytkownik widzi jeden dysk o pojemności 1200 MB zamiast czterech dysków 300 MB.
Dzięki temu administrator systemu nie musi się niepokoić kończącą się dostępną przestrzenią na danym dysku. Ponieważ całe 1200 MB jest teraz widziane jako jednolity zasób, można na nim tworzyć dowolną strukturę plików, nie martwiąc się już więcej o ograniczenia narzucane fizyczną wielkością poszczególnych dysków.

Nowe techniki pomagają zmniejszyć przestrzeń dzielącą wydajność jednostek centralnych i dysków. Techniki te opierają się na wykorzystaniu macierzy dyskowych (ang. *disk arrays*) do przyspieszenia operacji dyskowych. Techniki te, zwane w skrócie raid (ang. *Redundant Array of Inexpensive Disks*) można podzielić na kilka poziomów. Niektóre z nich pozwalają niemal czterokrotnie zwiększyć wydajność systemu dyskowego w stosunku do pojedynczego dysku. W macierzy dyskowej łączy się kilka dysków w określony sposób, aby osiągnąć jak największą wydajność systemu dyskowego i jednocześnie podnieść jego niezawodność do niemal pełnego bezpieczeństwa.

W tej pracy omówione zostaną poziomy RAID oraz plusy i minusy każdego z nich. Nie zamierzam wdawać się w szczegoły techniczne a jedynie wprowadzić w technikę macierzy dyskowych tych, którzy jeszcze nie mieli okazji się z nią zetknąć. Mam nadzieję, że te informacje okażą się pomocne w wyborze najlepszego dla Państwa systemu.

Definicje poziomów RAID

Poniższa tabela zawiera uproszczone definicje sześciu poziomów RAID. W dalszej części tego tekstu postaram się objaśnić je bardziej szczegółowo.

W swej najprostszej formie mirroring opiera się na dwóch dyskach podłączonych do jednego kontrolera. Zapis danych przebiega równolegle na oba dyski. W przypadku awarii jednego z nich system może kontynuować pracę używając sprawnego dysku. Nie ma przy tym znaczenia, który z dysków uległ uszkodzeniu. Oba zawierają tą samą informację i mogą spełniać rolę dysku systemowego.

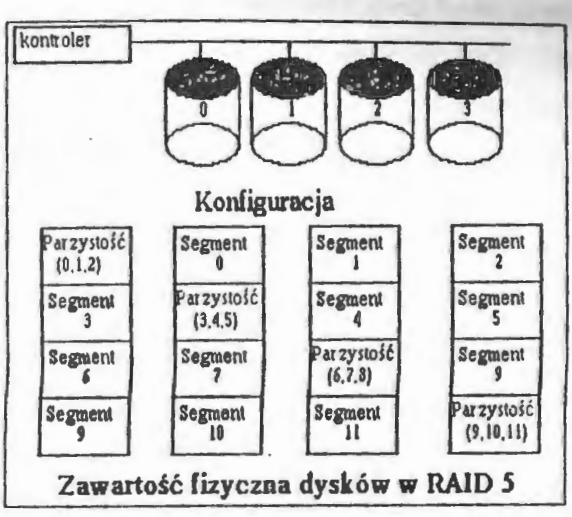
Nawet tak prosty system pozwala na zastosowanie pewnych metod optymalizacji operacji dyskowych. Jedną z nich jest rozkład obciążenia podczas operacji odczytu z dysku. Sytuacja taka zachodzi, gdy kilku użytkowników pragnie odczytać dane jednocześnie. Żądania odczytu mogą zostać rozdzielone pomiędzy oba dyski tak, aby były one równo obciążone. Metoda ta pozwala w istotny sposób usprawnić operacje odczytu, gdyż oba dyski mogą równocześnie czytać różne dane. Nie ma niestety sposobu na usprawnienie operacji zapisu przy stosowaniu mirroringu.

Jako dysk "zwierciadlany" możemy z powodzeniem stosować na przykład zapisywalne dyski optyczne. Jakkolwiek taki system nie będzie tak efektywny jak system dwóch dysków magnetycznych, to zawsze jest to alternatywa dla całkowitego braku zabezpieczenia.

Podsumowując: mirroring zapewnia bardzo dużą odporność na błędy oraz może przyspieszyć operacje odczytu z dysku. Jest on niestety bardzo drogi z powodu konieczności zastosowania dwóch dysków w systemie.

Striping z redundancją = RAID 2..5

Striping podnosi wydajność systemu dyskowego, gdyż umożliwia zapis i odczyt danych z prędkością powiększoną proporcjonalnie do ilości napędów dyskowych w systemie. Wadą tego rozwiązania jest jednak bardzo duża podatność na awarie dysków - uszkodzenie jednego uniemożliwia odczyt danych. Cały system dysków można zduplikować, czyli zastosować



klasyczny mirroring. Jeśli mielibyśmy cztery dyski połączone w system stripingu, to możemy dodać cztery identycznie połączone dyski dla zdublowania funkcji pierwszej czwórki. Jest to oczywiście drogie, choć może okazać się tańsze niż mirroring jednego dużego dysku.

Dla zapewnienia bezpieczeństwa zgromadzonych danych i umiarkowanej ceny należy zapewnić inny sposób dodania redundancji do informacji zapisanej na dysku. Można to osiągnąć przez wyliczenie sum kontrolnych i bądź wykorzystanie dodatkowego dysku do ich zgromadzenia (RAID 3), bądź rozproszenie ich wraz z danymi po wszystkich dyskach (RAID 5). Przykład rozproszenia sum kontrolnych w RAID 5 pokazano na rysunku.

Sumy kontrolne są tworzone przez operację XOR (Exclusive OR). Poniżej pokazano prosty przykład generowania bitu parzystości:

A	B	C	Dysk Parzystości
1	0	1	0

Należy pamiętać, że funkcja XOR daje w wyniku "1" wtedy i tylko

wtedy, gdy wśród jej argumentów jest nieparzysta ilość "1". Załóżmy, że uszkodzeniu uległ dysk B w naszym przykładzie. Możemy teraz odzyskać wartość z uszkodzonego dysku przez obliczenie funkcji XOR, której argumentami będą bity z dysków A i C oraz dysku parzystości. W jej wyniku otrzymamy brakujące "0". Podobnie można odzyskać informacje z dowolnego dysku. Gdyby rozszerzyć nasz przykład do siedmiu dysków, otrzymamy:

A	B	C	D	E	F	Dysk Parzystości
0	0	0	1	0	1	0

Gdyby teraz dysk B uległ uszkodzeniu, możemy jego zawartość odtworzyć poprzez obliczenie funkcji XOR z pozostałych sześciu dysków.
Jeżeli parzystość jest zapisywana na jednym dysku (RAID 3) i system musi obsłużyć kilka żądań zapisu jednocześnie, to dla każdej operacji zapisu musi być również zapisany dysk parzystości, co powoduje zator RAID 5 pozwala uniknąć takiej sytuacji dzięki rozproszeniu parzystości po wszystkich dyskach.
Ten typ zapisu daje równomierny rozkład wszystkich informacji na dysku, co umożliwia przykładowo zapis danych przez jednego użytkownika na dysk 1 i parzystości na dysk 2, podczas gdy drugi użytkownik może zapisywać dane na dysk 3 i parzystość na dysk 4. Pozwala to na znaczące przyspieszenie operacji dyskowych. Szczególnie jest to widoczne przy systemach wielodostępnych, gdzie wiele dyskowych operacji we/wy może zachodzić niemal równocześnie. System RAID 5 zapewnia większą wydajność niż system z pojedynczym dyskiem parzystości. Wydajność ta jest jednak mniejsza niż przy samym stripingu, gdyż system jest obciążony wyliczeniem i zapisem parzystości. Jeśli przykładowo część bloku danych jest modyfikowana jego niemodyfikowana część musi być również odczytana w celu wygenerowania parzystości dla całego bloku. Po wy-

generowaniu parzystości (co zajmuje pewien skończony czas procesora) należy ją jeszcze zapisać. Technika ta znana jest jako read-modify-write. Tak więc choć RAID 5 jest znacznie lepszy od RAID 0, który nie zapewnia żadnej redundancji, to jest on nieznacznie wolniejszy. Zwróćmy uwagę na lepszą wydajność, np. RAID i w systemie o przewadze operacji zapisu nad operacjami odczytu. Wynika ona z konieczności stosowania wyżej wymienionej techniki read-modify-write przez RAID 5. RAID 1 będzie oczywiście droższy, lecz problem wyboru między efektywnością w określonych warunkach a ceną należy do nabywcy systemu. Jeśli ilość odczytów rośnie w stosunku do ilości zapisów, RAID 5 staje się coraz bardziej interesujący.

Zapis danych w technice mirroringu (RAID 1) oraz rozpraszania danych i parzystości (RAID 5) powoduje powstanie informacji nadmiarowej. Różnica polega na jej ilości: dla RAIL 1 są to wszystkie dane, dla RAID 5 jedynie uzyskane za pomocą operacji XOR bity parzystości. W efekcie RAID 5 daje bardzo pewny mechanizm odzyskiwania utraconych danych bez nadmiernej utraty cennej powierzchni dysku.
Przy zastosowaniu tej techniki dla macierzy 5 dysków, około 20% powierzchni dysków będzie przeznaczone na bajty parzystości. Dla macierzy 10 dyskowej już tylko 10% powierzchni dysków będzie zajęte przez bajty parzystości. Im więcej dysków będzie w systemie, tym lepszy będzie stosunek cena/powierzchnia użyteczna dysku.

Podsumowując; RAID 5 łączy zalety stripingu i systemów z redundancją danych. Efektem tego połączenia jest system dyskowy dobrze nadający się do układów transakcyjnych np. systemów rezerwacji biletów iltp. Dla niektórych zastosowań należałoby rozpatrzyć zastosowanie systemu RAID 1 (omówiono to wcześniej).

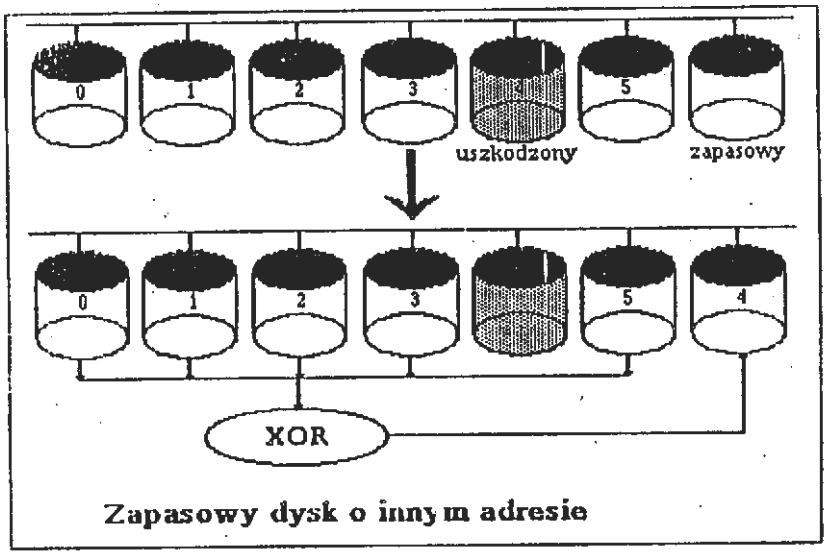
Generalnie jednak RAID 5 wydaje się być rozwiązaniem optymalnym,

zapewniając:
- dużą szybkość systemu dyskowego,
- niski koszt,
- bezpieczeństwo danych.

Odzyskiwanie danych po awarii dysku

Mirroring i RAID 5 oferują nowe sposoby odzyskiwania danych po awarii dysku. Ponieważ w obu systemach istnieje informacja nadmiarowa gromadzona w czasie rzeczywistym, można uzyskać potrzebne dane nawet po uszkodzeniu jednego z dysków. Ważną zaletą tego rozwiązania jest to, że wymiana uszkodzonego dysku nie musi być natychmiastowa, gdyż system może pracować z jednym dyskiem nie działającym. Oczywiście w takiej sytuacji macierz nie jest już odporna na uszkodzenia. Odbudowanie informacji na uszkodzonym dysku (po jego wymianie) musi nastąpić zanim dojdzie do uszkodzenia kolejnego dysku aby zapobiec utracie danych.
Po wymianie uszkodzonego dysku, jego poprzednia zawartość musi zostać odtworzona. w przypadku mirroringu cała potrzebna informacja znajduje się na dysku nieuszkodzonym. Wystarczy więc po prostu skopiować zawartość tego dysku na dysk docelowy. Kopiowanie takie jest znacznie szybsze od ściągania danych z pamięci taśmowej.

Jeśli dyski były połączone w system RAID 5, zawartość brakującego dysku jest odtwarzana przy wykorzystaniu danych zawartych na pozostałych, sprawnych dyskach. Dane z tych dysków są odczytywane (włącznie z sumami kontrolnymi) i zawartość brakującego dysku jest wyliczana i zapisywana na nowy dysk. Proces ten, pokazany na rysunku jest również znacząco szybszy od odzyskiwania danych z pamięci taśmowej.



Jeśli macierz posiada elastyczną strukturę można ją skonfigurować w ten sposób, by dysk zastępczy nie musiał znajdować się pod tym samym adresem co uszkodzony. Pokazuje to rysunek. Taka konfiguracja pozwala na jeszcze łatwiejszą wymianę dysku po awarii. Zapasowy dysk może być nawet podłączony zanim wystąpi awaria! Wtedy będzie on w stanie zastąpić którykolwiek z dysków w przypadku uszkodzenia. Dysk taki nazywa saię w macierzy Hot Spare.

Zastosowanie macierzy dyskowej w miejsce jednego dysku podnosi poziom bezpieczeństwa danych wielokrotnie. W tabeli podano podstawowe zależności dla różnych typów macierzy.

Typ Macierzy	Liczba Dysków	MTBF*	MTBDL*
Pojedynczy dysk	1	30.000 h	30.000 h
RAID 0	5	30.000 h	6.000 h
RAID 1	2	30.000 h	49.9 mln h
RAID 5	5	30.000 h	46.2 mln h

*) patrz słownik pojęć.

Dodatkowe zalety systemów macierzowych to:
- możliwość pracy systemu pomimo uszkodzenia,
- szybkość odbudowywania informacji po uszkodzeniu (znacznie większa niż dla pamięci taśmowych),
- 100% aktualność rezerwowych danych w momencie awarii (w systemach taśmowych odzyskujemy dane z ostatniego backupu).

Należy jednak pamiętać, że pełne bezpieczeństwo całego systemu można uzyskać jedynie zapewniając redundancje wszystkich jego składników: np. zasilaczy, sterowników dyskowych itp.

Optymalizacja odporności na błędy

Jak wiemy, RAID 0 (striping) może znacząco podnieść szybkość pracy systemu dyskowego w stosunku do pojedynczego dysku, gdyż dane na nim są rozproszone po kilku dyskach, przez co kilka operacji zapis/odczyt może zachodzić równocześnie.

Przez utworzenie zwierciadlanego odbicia tak połączonego zestawu dysków, możemy uzyskać bardzo szybki i bardzo niezawodny system dyskowy. Operacje odczytu będą nawet szybsze niż przy samym stripingu,

gdyż system może obsługiwać dwa żądania odczytu jednocześnie (jedno na każdy dysk). Zapis jest niemal równie szybki jak w systemie bez mirroringu, gdyż jedynie niewielki narzut czasu potrzebny jest na zapis informacji na oba podsystemy.

Bezpieczeństwo tak utworzonego systemu można jeszcze podnieść przez zastosowanie zdublowanego sterownika dyskowego. Dodatkowo otrzymujemy wtedy dalsze przyspieszenie operacji dyskowych.

Słownik pojęć

Disk Mirroring: dwa dyski lub podsystemy dyskowe podłączone do wspólnego kontrolera. Jeden z dysków pełni rolę zwierciadlanego odbicia drugiego. Dane są zapisywane jednocześnie na oba dyski. Ponieważ oba zawierają identyczną informację, każdy z nich może pełnić rolę dysku systemowego w przypadku awarii drugiego.

Disk Spanning: kilka dysków połączonych tak, by na zewnątrz widziane były jako pojedynczy dysk o odpowiedniej pojemności. Dane są zapisywane na różnych fizycznie dyskach w sposób przezroczysty dla użytkownika.

Disk Striping: dane są rozproszone po kilku dyskach zamiast być zgromadzone na jednym. Blok 1 zapisany będzie na dysku 0, blok 2 na dysku 1 itd. Gdy system osiągnie ostatni dostępny dysk zaczyna zapisywać kolejny wolny segment na dysku 0. Cały proces jest powtarzany do momentu, gdy wszystkie dane zostaną zapisane.

Duplexing: oznacza zastosowanie w systemie dwóch sterowników dyskowych zamiast jednego. Jeśli jeden z nich ulegnie uszkodzeniu jego zadania przejmuje drugi. Dodatkowo istnieje możliwość takiego zaprojektowania oprogramowania, aby oba kontrolery pracowały równolegle, obsługując różne dyski.

Host adapter: sterownik komunikujący się bezpośrednio z komputerem.

Hot Fix: oznacza zastąpienie uszkodzonego dysku innym dyskiem, znajdującym się już w macierzy. Uszkodzony dysk nie jest fizycznie usunięty. Zapasowy dysk jest po prostu zapisywany danymi, które uprzednio zawierał uszkodzony dysk i system dalej pracuje.

Hot Patch: system, który zawiera dyski zapasowe (Hot Spare).

Hot Spare: dysk podłączony do komputera, który jest w stanie przejąć rolę dysku uszkodzonego. Różnica pomiędzy hot spare i cold spare polega na tym, że cold spare znajduje się gdzieś na półce i musi dopiero zostać podłączony do systemu w razie awarii.

MTBDL: (Mean Time Between Data Loss) średni czas pomiędzy przypadkami całkowitej utraty danych.

MTBF: (Mean Time Between Failure) średni czas, w którym urządzenie pracuje bezawaryjnie.

RAID: (Redundant Array of Inexpensive Disks) metoda łączenia kilku tanich dysków w system zapewniający większą wydajność i bezpieczeństwo danych.

SEED: (Single Large expensive Disk) pojedynczy drogi dysk o dużej pojemności.