Struktura pamięci masowej (pomocniczej)

Struktura dysku

- Napędy dyskowe są adresowane jako duże jednowymiarowe tablice bloków logicznych, gdzie blok logiczny jest najmniejszą jednostką transmisji (zwykle 512B).
- Tablica ta jest sekwencyjnie odwzorowywana na sektory dysku:
  - sektor 0 jest pierwszym sektorem pierwszej ścieżki najbardziej zewnętrznego cylindra.
  - dalsze odwzorowanie następuje wzdłuż ścieżki, wzdłuż cylindra i dalej — od najbardziej zewnętrznego cylindra wgłąb.

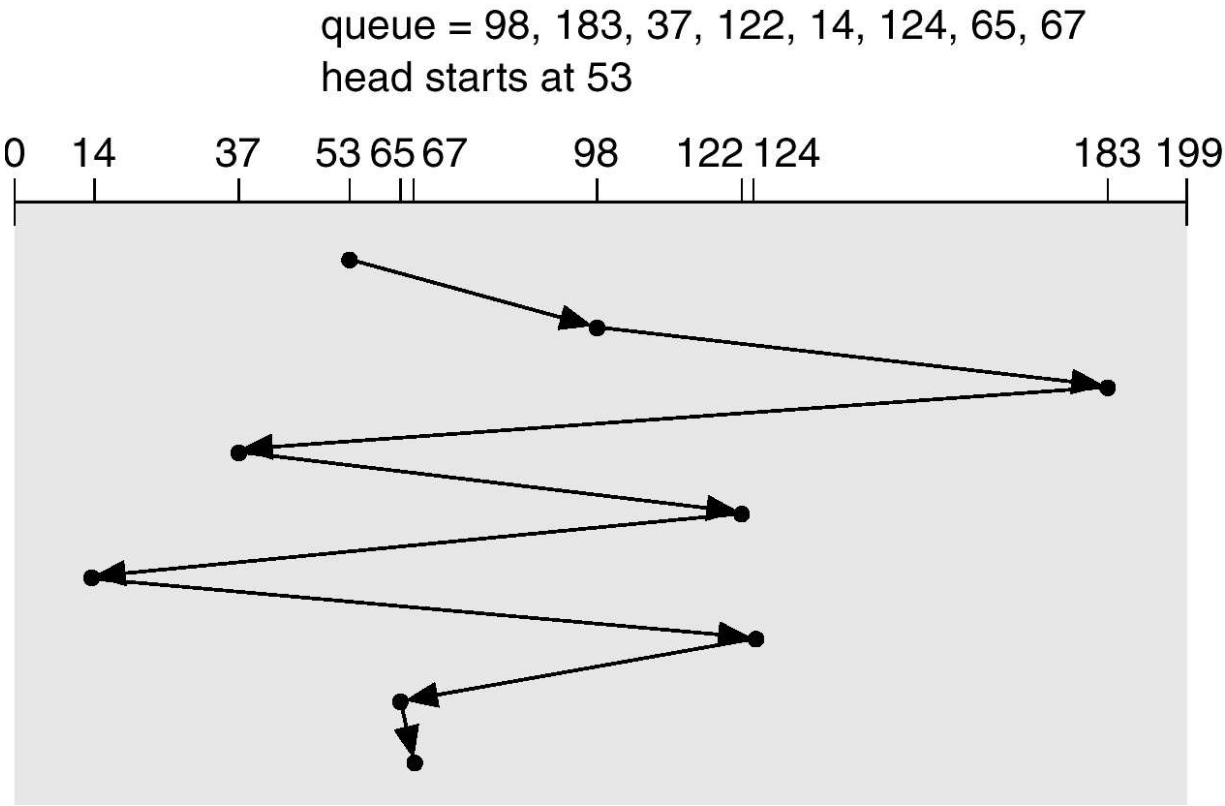
Planowanie dostępu do dysku

- System operacyjny odpowiada za ekonomiczne użytkowanie sprzętu, co oznacza szybki dostęp i szybkie przesyłanie danych.
- Czas dostępu składa się z dwóch składników:
  - *seek time* — czas przesunięcia głowicy do właściwego cylindra,
  - *rotation latency* — czas obrotu dysku do pozycji właściwego sektora.
- Dążymy do minimalizacji *seek time*.
- Czas szukania jest proporcjonalny do odległości między cylindrami.
- *Disk bandwidth* — szerokość pasma: łączna liczba przesłanych bajtów podzielona przez odcinek czasu pomiędzy pierwszym żądaniem usługi a zakończeniem przesyłania.

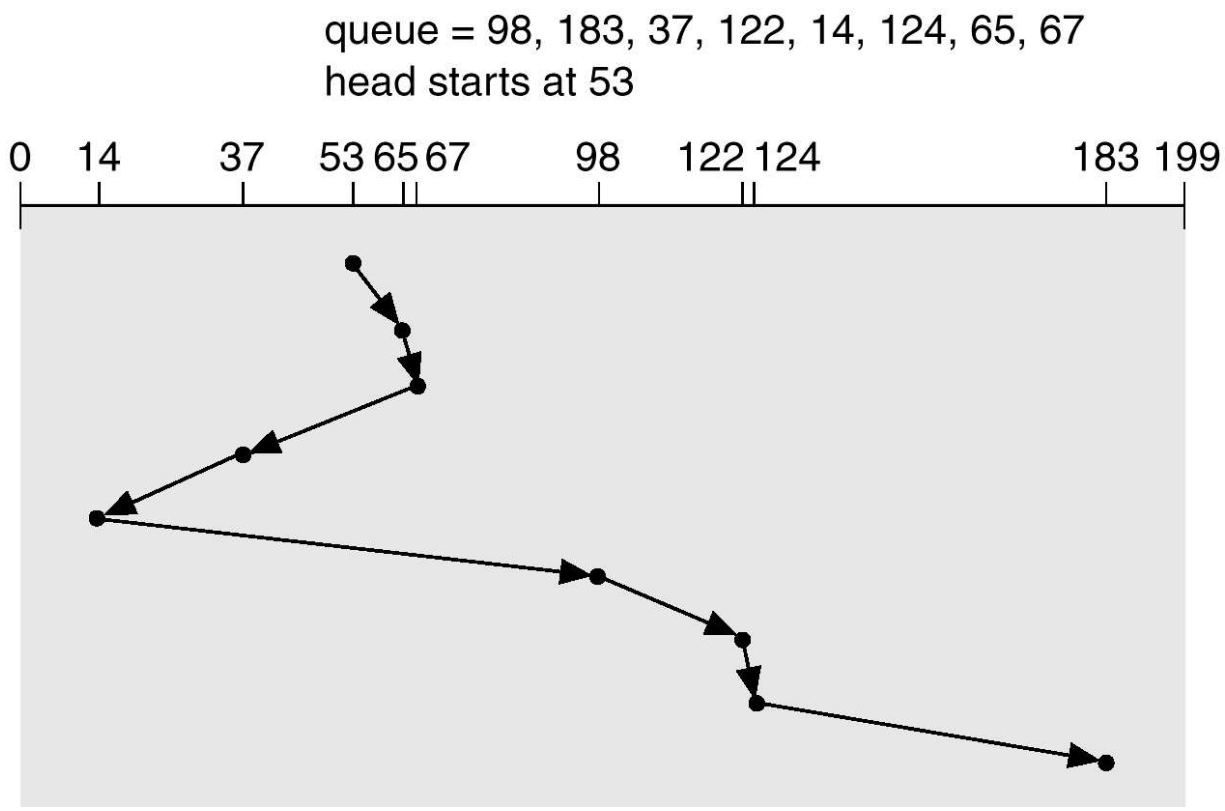
- Istnieją różne algorytmy szeregowania żądań dyskowych operacji we/wy.
- Będziemy je ilustrować kolejką żądań (z zakresu 0–199):

  98, 183, 37, 122, 14, 124, 65, 67

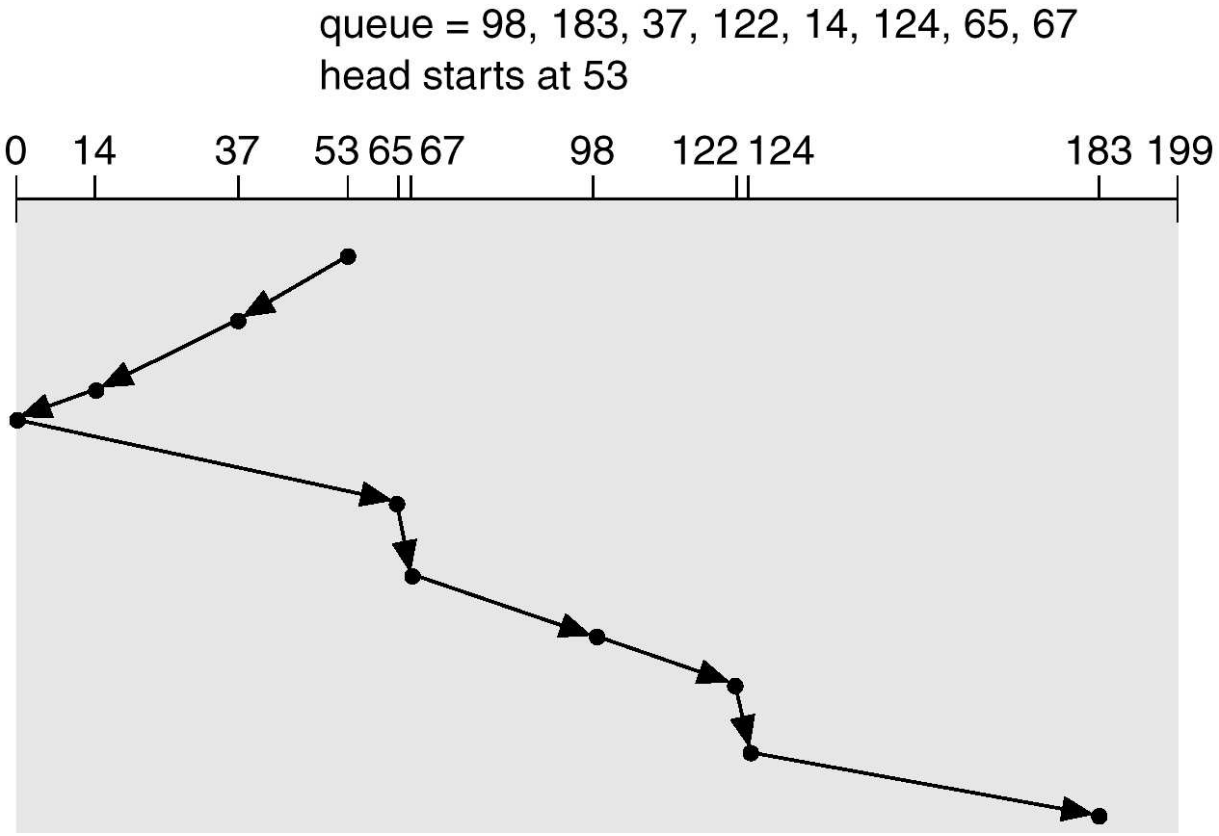
  Głowica wskazuje na cylinder 53.
- **FCFS** — łączna droga, którą pokonuje głowica to 640 cylindrów.

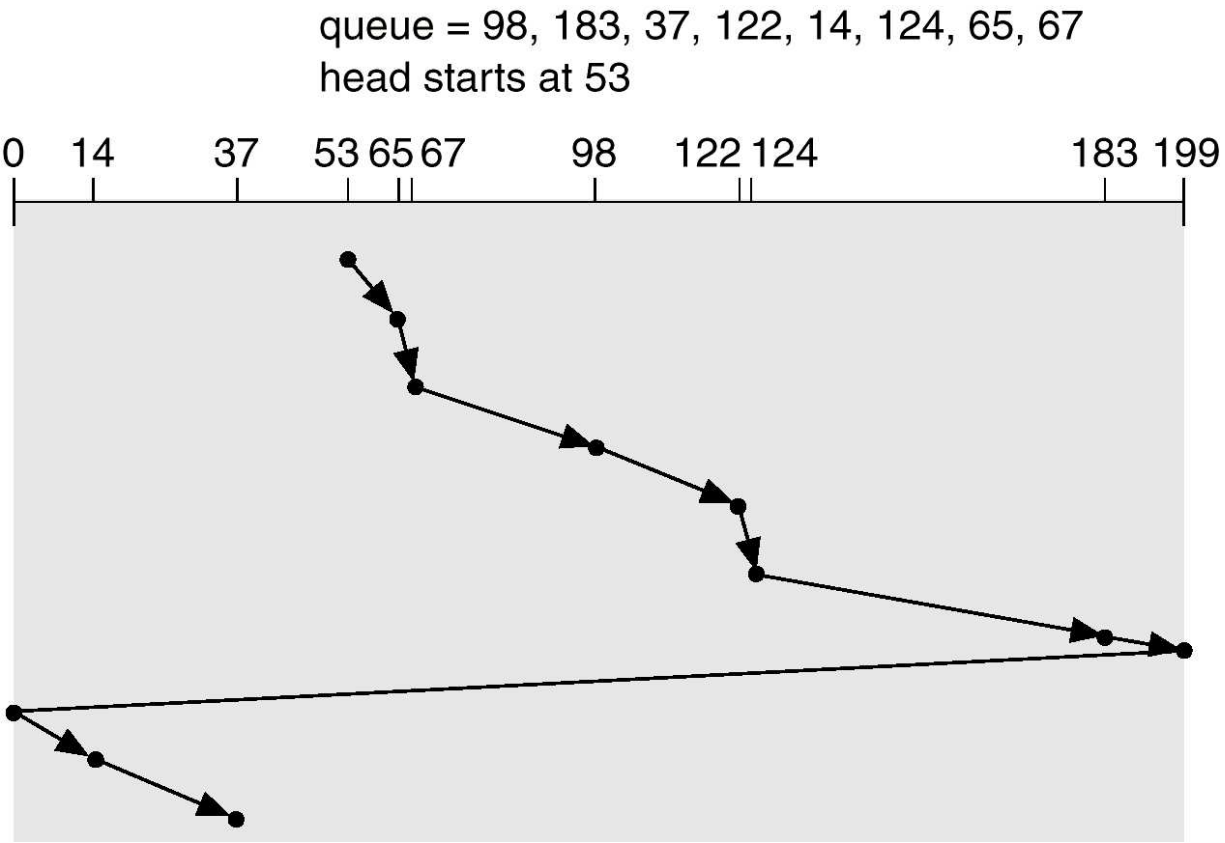


- **SSTF** — jako kolejne obsługuje żądanie z minimalnym czasem przeszukiwania w stosunku do bieżącej pozycji.
- SSTFjest jakąś wersją algorytmu SJF — może powodować zagłodzenie.
- Łączna droga, którąpokonuje głowica to 236 cylindrów.
- Algorytm **nie jest** optymalny — dla podanego przykładu istnieje krótsza droga.



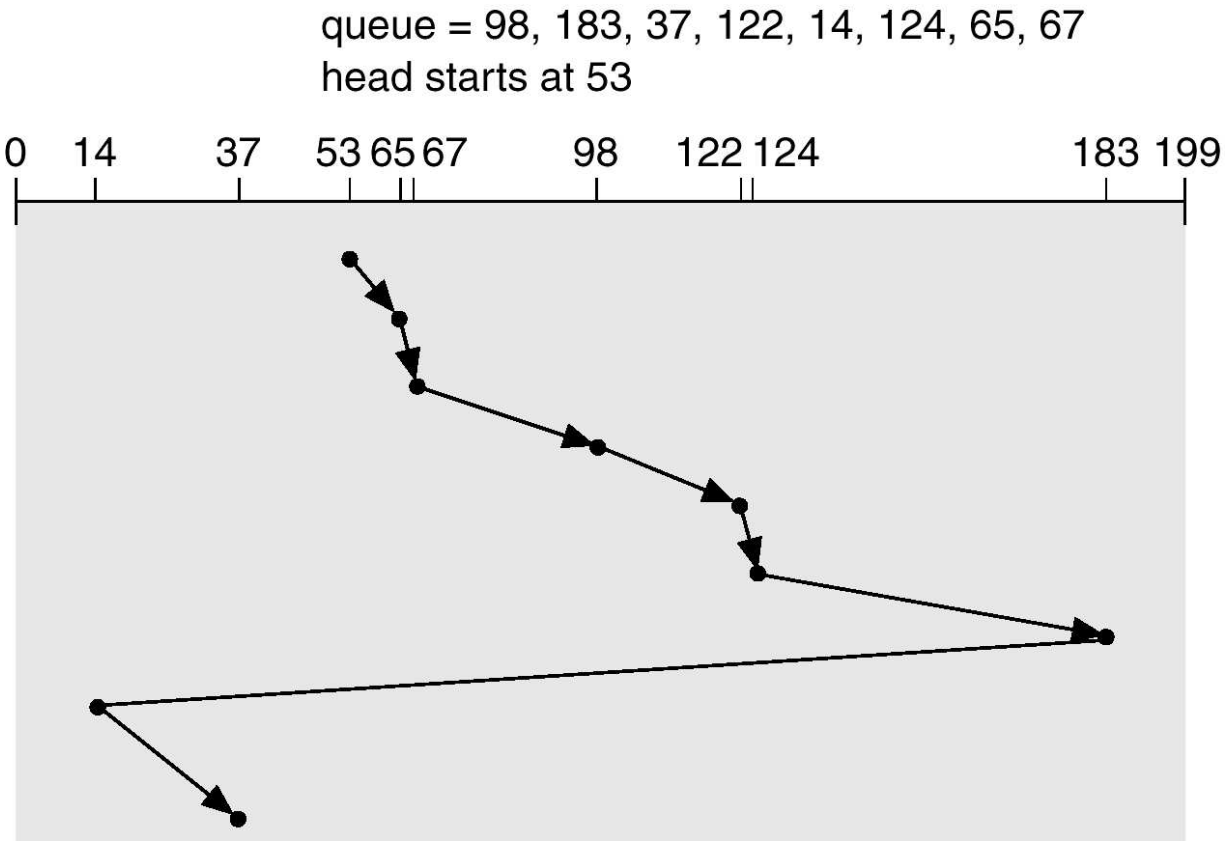
- **SCAN** — głowica przesuwa się od jednej krawędzi dysku do drugiej, obsługując żądania do mijanych cylindrów. Po dotarciu do skrajnego cylindra kierunek ruchu zmienia się na przeciwny.
- Niekiedy algorytm nazywany jest algorytmem windy.
- Łączna droga, którą pokonuje głowica wynosi 208 cylindrów.
- **C-SCAN** — głowica przemieszcza się od krawędzi dysku do krawędzi, po czym natychmiast wraca do początkowego położenia.
- Traktuje numery cylindrów jak listę cykliczną, na której ostatni cylinder poprzedza pierwszy.
- Zapewnia bardziej równomierny czas oczekiwania na obsłużenie zamówienia.





- **C-LOOK** — podobnie jak C-SCAN, ale głowica przemieszcza się

tylko od jednego skrajnego zamówienia do drugiego.

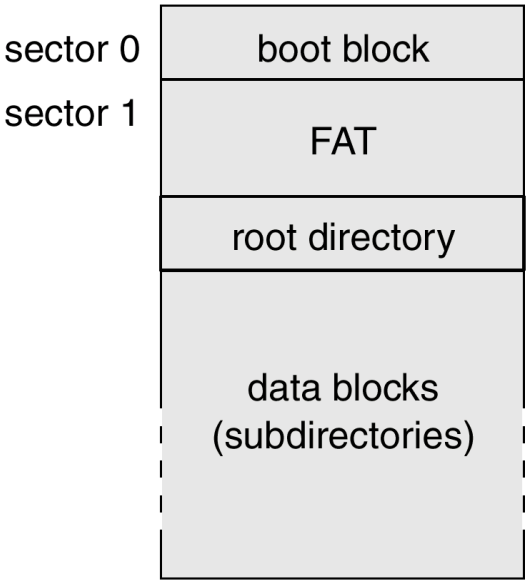


- Wybór algorytmu planowania dostępu do dysku.
  - SSTF jest dość powszechny i wygląda naturalnie
  - SCAN i C-SCAN mają lepszą wydajność w sytuacji znacznego obciążenia dysku
  - żądania i obsługa dysku zależy od metody przydziału bloków dla pliku.
  - obliczanie optymalnego sposobu obsługi dla danego ciągu zamówień może okazać się nieopłacalne
  - algorytm szeregowania powinien być oddzielnym modułem, który w razie potrzeby można zastąpić innym
  - jako algorytm domyślny nadają się SSFT albo LOOK
- System nie ma możliwości planowania dostępu pod kątem opóźnienia rotacyjnego — brak informacji o położeniu sektorów.

- Możliwa implementacja algorytmów dostępu w sterowniku wbudowanym w sprzęt napędu dysku.
- Planowanie na poziomie systemowym umożliwia realizację również innych niż wydajność celów, takich jak zagwarantowanie wykonania pewnych operacji przed innymi (np. pisanie przed czytaniem, stronicowanie przed we/wy aplikacji, itp.).

Zarządzanie dyskiem

- *Formatowanie niskopoziomowe (fizyczne)* — podział dysku na sektory czytelne dla kontrolera (nagłówki, zakończenia z *error-correcting code*).
- Aby przechowywać na dysku pliki system operacyjny musi jeszcze zapisać na nim swoje struktury danych:
  - podział na partycje (grupy cylindrów)
  - formatowanie logiczne każdej z partycji — zapisanie inicjalnych struktur systemu plików, np. mapa przydzielonych i zajętych obszarów (tablica FAT, i-węzły) pusty katalog początkowy, itp.
- *Boot block* — blok inicjujący system.
  - Program rozruchowy (bootstrap) ładuje i uruchamia jądro systemu; przechowywanie go w pamięci ROM może być kłopotliwe.
  - Boostrap loader (przechowywany w ROM) ładuje pełny program rozruchowy z ustalonego miejsca na dysku (zwykle tak zwany boot sector)

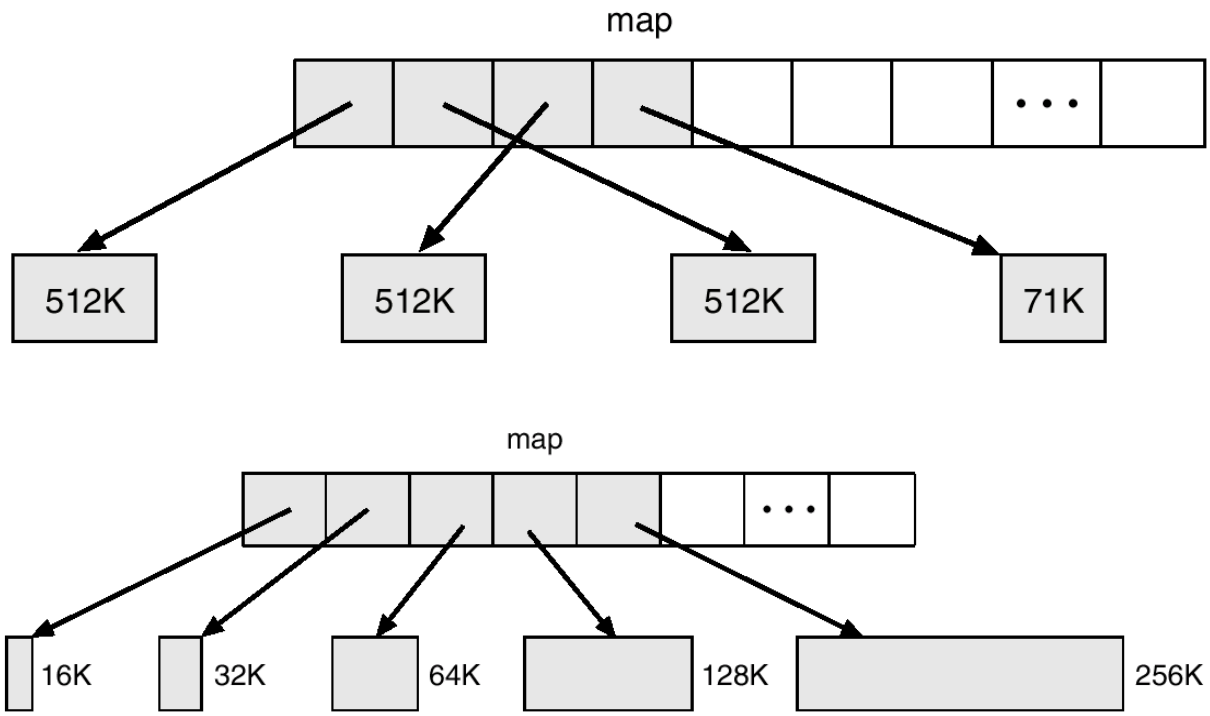


- Bloki uszkodzone
  - ręczna eliminacja np. w FAT specjalny znacznik
  - sektory zapasowe — sektory uszkodzone są zastępowane przez sektory z puli zapasowej
  - przeciąganie sektorów — przesuwanie o jeden blok grupy sektorów od miejsca uszkodzenia do najbliższego sektora zapasowego

Zarządzanie obszarem wymiany (swap space)

- System wykorzystuje go jako rozszerzenie pamięci głównej — sposób wykorzystania zależy od organizacji pamięci.
- Może być umieszczony w obrębie systemu plików ale najczęściej stanowi oddzielny obszar (partycję) z oddzielnym zarządcą optymalizowanym pod kątem szybkości.
- Zarządzanie bywa rozmaite w różnych systemach:
  - system 4.3BSD przydziela obszar wymiany mieszczący segment kodu i segment danych w momencie startu procesu.

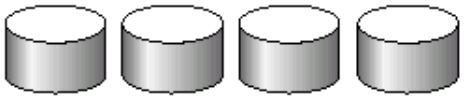
– Jądro używa dwu map wymiany: dla kodu o stałej wielkości bloku i dla danych o blokach różnej wielkości. W miarę wzrostu segmentu blok mniejszy zastępowany jest dwukrotnie większym.



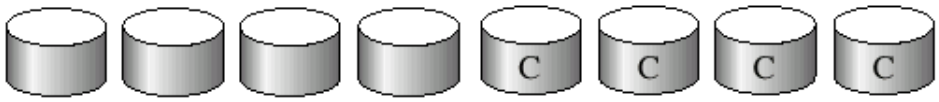
RAID (*redundant array of independent disks*)

- RAID — napędy wielodyskowe (dysków współpracujących) zapewniające niezawodność kosztem redundancji.
- Paskowanie dysku (disk striping) zwiększa szybkość — kilka dysków traktowane jest jak jedno urządzenie.
- Sześć różnych typów (level) organizacji.
  – mirroring — najprostsze ale kosztowne
  – przeplatanie bloków parzystości (typ 4), jeden blok parzystości dla kilku bloków z danymi — każdy bit w bloku parzystości odpowiada analogicznym bitom w blokach danych; pozwala na odzyskanie informacji przy awarii jednego z dysków

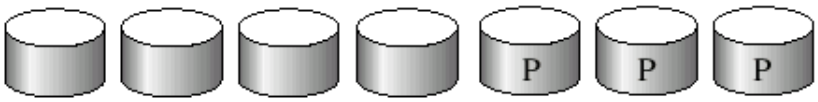
– dla typu 4: modyfikacja małej porcji danych wymaga aż 4 operacji dyskowych (jakich i dlaczego?)



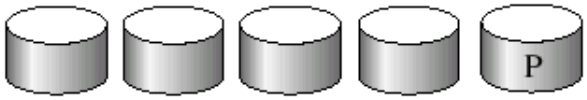
(a) RAID 0: non-redundant striping



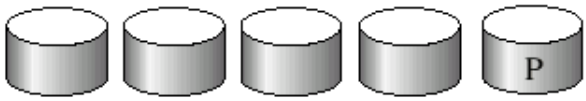
(b) RAID 1: mirrored disks



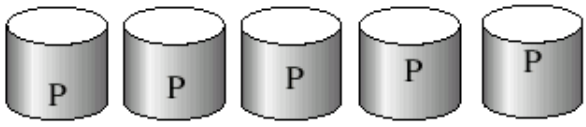
(c) RAID 2: memory-style error-correcting codes



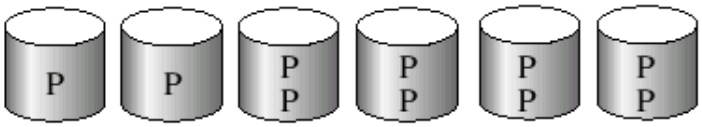
(d) RAID 3: bit-interleaved Parity



(e) RAID 4: block-interleaved parity



(f) RAID 5: block-lnterleaved distributed parity

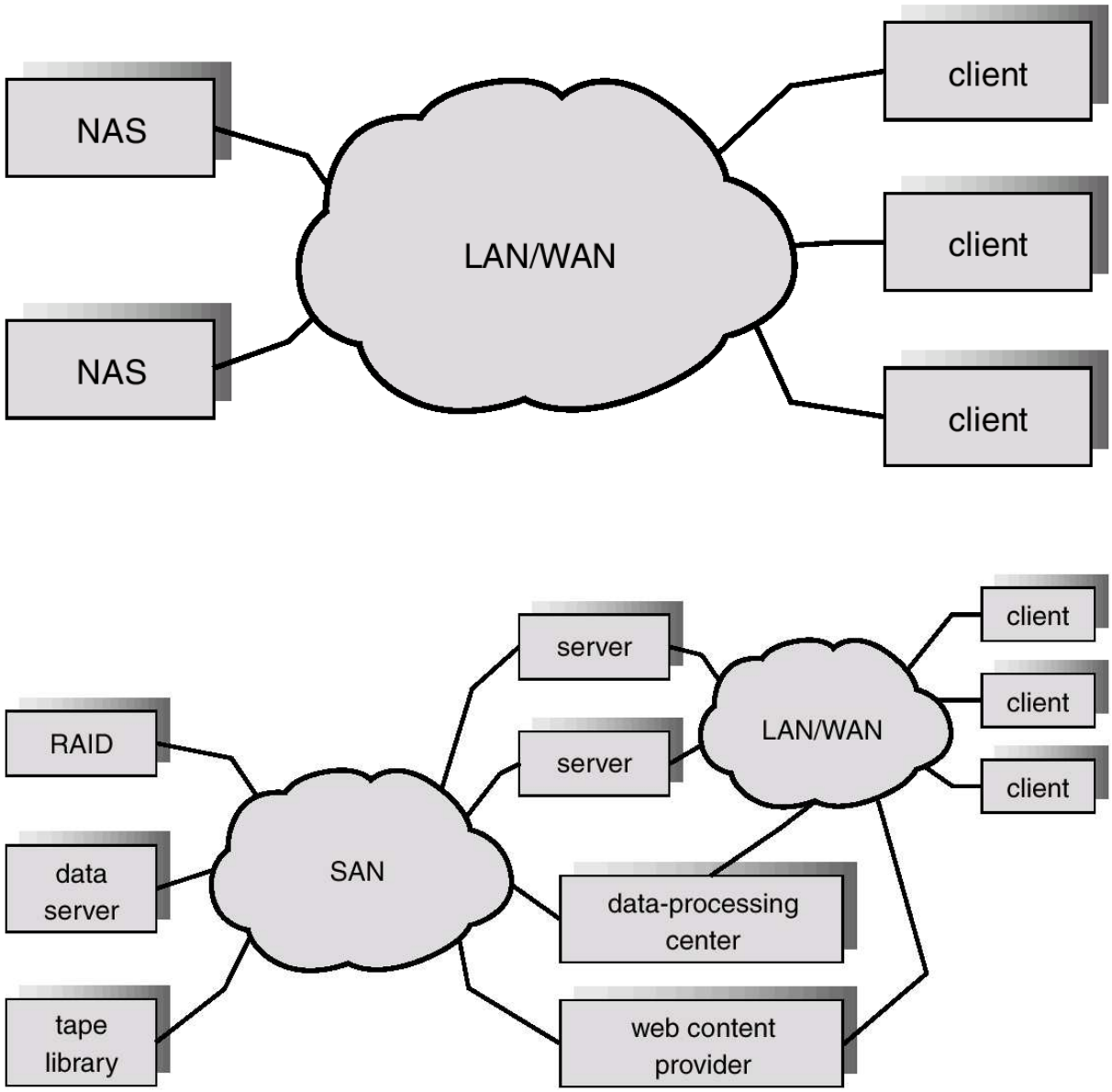


(g) RAID 6: P + Q redundancy

Przyłączanie dysków

Dwa podstawowe sposoby:

- Bezpośrednio — prze port I/O.
- Przez sieć (NAS - network attached storage).



Implementacja pamięci trwałej

- Niektóre techniki synchronizacyjne i bazodanowe wymagają dostępu do tak zwanej pamięci trwałej.
- Pamięć operacyjna – ulotna, pamięć dyskowa – nieulotna, pamięć trwała – informacja nigdy nie ulega stracie.
- Implementacja:
  - utrzymywanie kopii na wielu nieulotnych urządzeniach niezależnych od siebie pod względem awaryjności
  - modyfikowanie danych w sposób kontrolowany, tak aby mieć pewność odzyskania stabilnych danych w przypadku każdej awarii
  - utrzymywanie (co najmniej) dwóch bloków fizycznych dla każdego logicznego i realizacja zapisu danych najpierw do pierwszego, a po pomyślnym zakończeniu do drugiego bloku
  - w razie awarii sprawdza się każdą parę pod kątem błędów oraz identyczności zawartości

Pamięć trzeciorzędna

- Niskie koszty to element podstawowy dla pamięci trzeciorzędnej.
- Podstawą są nośniki wymienne, np. płyty CD
- Dyski wymienne:
  - dyski elastyczne (dyskietki) — słabo odporne mechanicznie
  - dyski magnetooptyczne (np. zip, technologia magnetyczna połączona z laserową) — głowica porusza się dalej, powierzchnia jest chroniona warstwą plastiku albo szkła

  - dyski optyczne – różne technologie zmieniające materiał pod wpływem światła laserowego
  - dyski WORM (jednorazowego zapisu CD i DVD) — trwałe i niezawodne
- Taśmy są znacznie tańsze niż dyski, mają większą pojemność ale są znacznie wolniejsze — nadają się do przechowywania danych, do których nie jest konieczny szybki dostęp.
- Istnieją wielkie instalacje taśmowe z automatycznymi zmieniaczami taśm. Taśmy są też automatycznie umieszczane w taśmotece (*near-line library*):
  - stacker — biblioteka z wieloma taśmami
  - silo — biblioteka z tysiącami taśm
- Możliwa jest automatyczna procedura składowania nieużywanych plików.

Zadania systemu operacyjnego

- Zarządzanie urządzeniami.
- Stworzenie abstrakcji maszyny wirtualnej na użytek aplikacji
- Dla dysków stałych realizuje dwie abstrakcje:
  - surowe urządzenie — tablica bloków
  - system plików — kolejkowanie i szeregowanie zamówień od różnych aplikacji

Jak system realizuje te zadania dla nośników wymiennych?

- Większość systemów traktuje dyski wymienne niemal identycznie jak stałe — nowy wolumin jest formatowany i tworzony jest na nim system

plików.

- Taśmy są traktowane jako surowy nośnik — aplikacja nie otwiera pliku na taśmie, natomiast otwiera przewijak jako urządzenie.
- Zwykle taśma jest rezerwowana do wyłącznego użytku aplikacji.
- Aplikacja decyduje jak wykorzystuje bloki na taśmie.
- Wykorzystanie danych na taśmie wymaga dokładnej znajomości struktury stworzonej przez aplikację.
- Zbiór operacji dostępnych dla taśmy jest inny niż dla dysku:
  - locate — podobnie jak seek, ale znajduje blok a nie ścieżkę,
  - read position — sprawdza pozycję głowicy
  - space — przesuwa głowicę o określoną liczbę bloków
  - pisanie powoduje utratę dostępu do informacji położonej za miejscem pisania — zapis powoduje automatyczne wpisanie znacznika EOT (end-of-tape)
- Problem nazw plików:
  - Generalnie nie jest proste zapewnienie jednoznacznej identyfikacji plików przez nazwy a zwłaszcza dostęp do plików na różnych komputerach czy pod innym systemem.
  - Dzisiejsze systemy problem nazewnictwa pozostawiają nierozwiązany.
  - Pewne typy nośników są dobrze ustandaryzowane – dyski CD mają tylko kilka formatów i typowy moduł obsługi napędu zna je wszystkie.
- Rozszerzenie hierarchii pamięci.
  - zwykle implementowane przez zastosowanie robota kasetowego

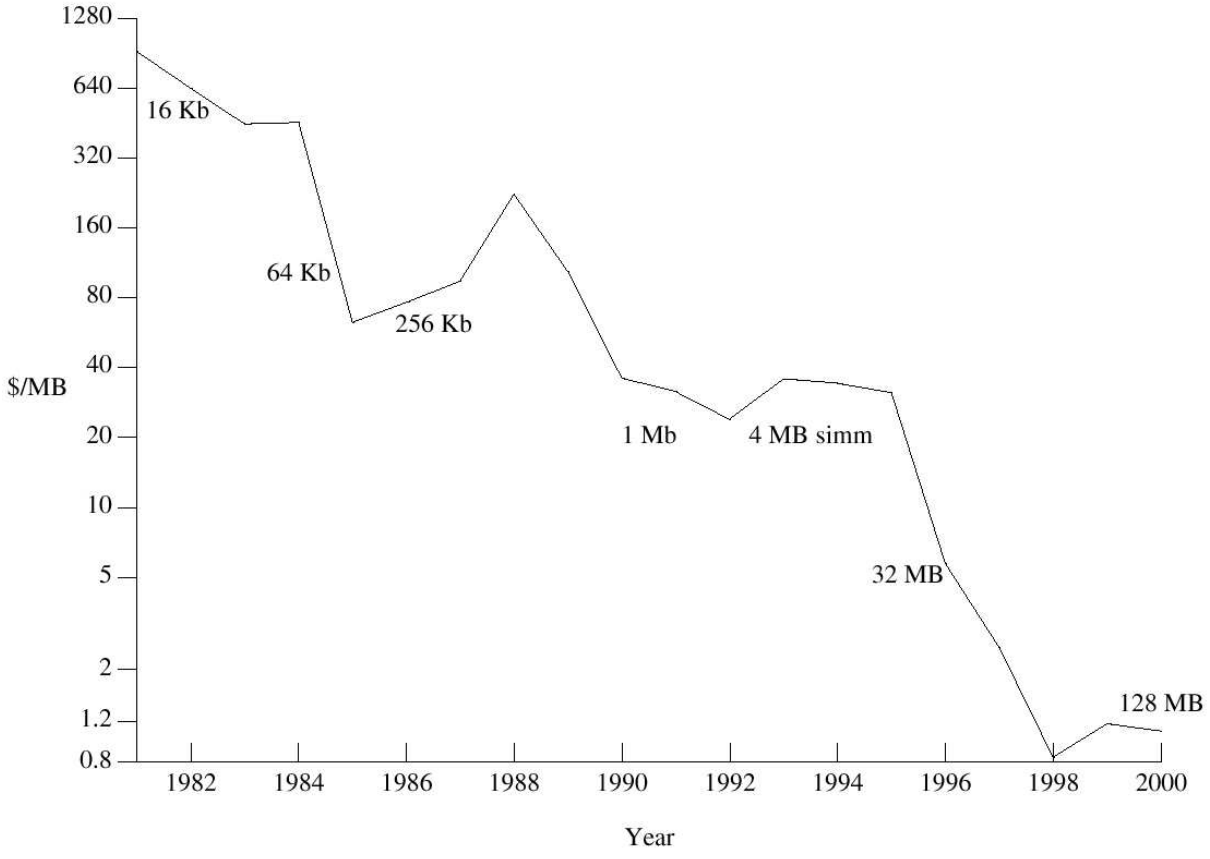
robotic jukebox — nie stosuje się do rozszerzania pamięci wirtualnej (zbyt wolne)

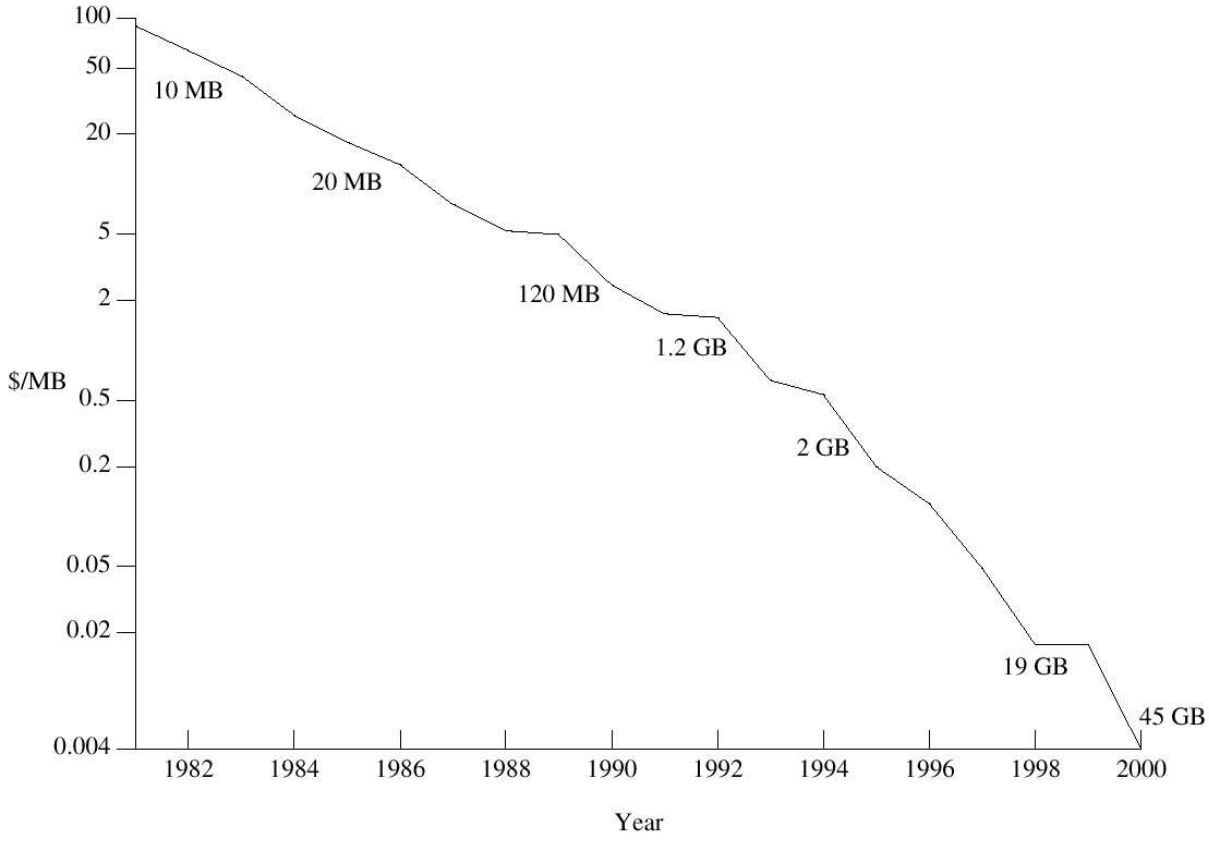
  - rozszerzenie systemu plików:
    * małe i często używane pliki są na dysku
    * duże, stare, nieaktywne pliki są archiwowane
  - HSM spotyka się w centrach superkomputerowych i innych instalacjach z wielką ilością danych

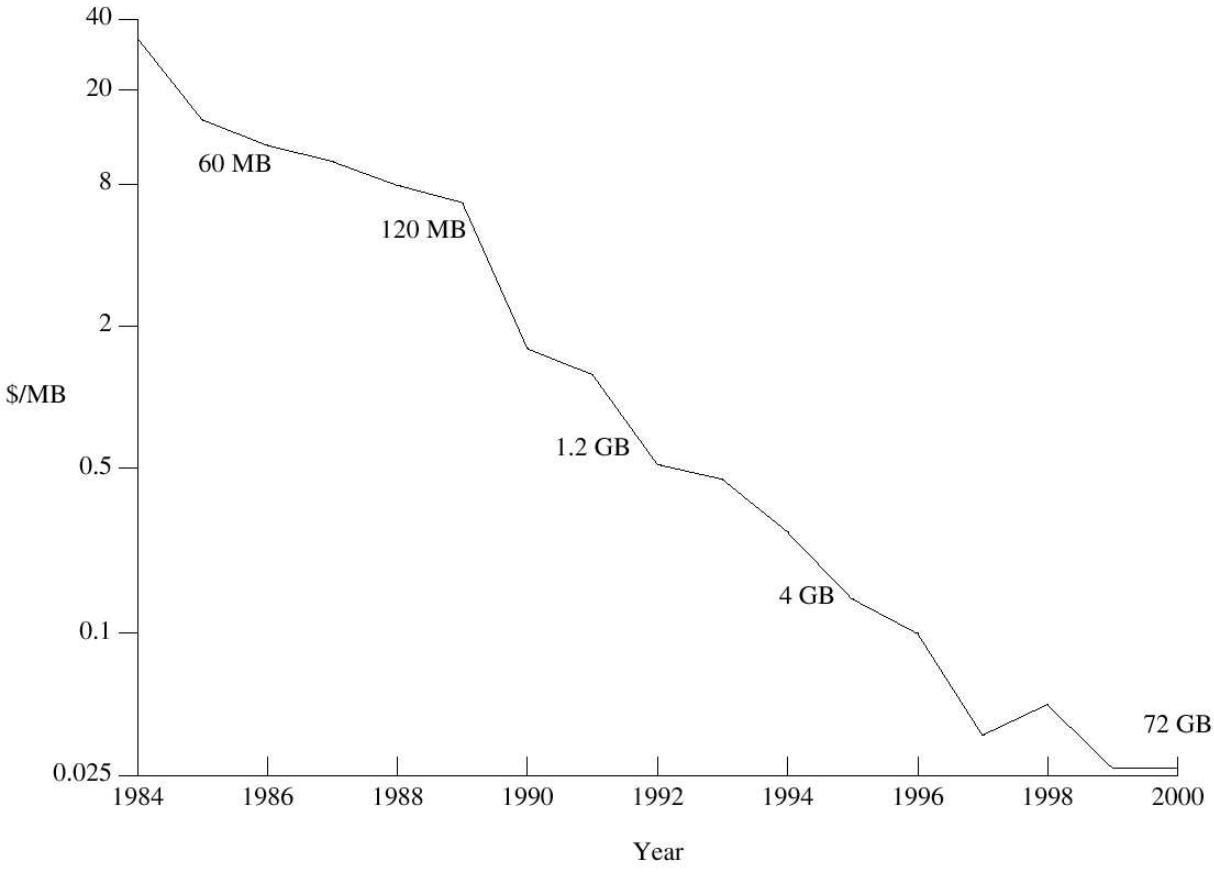
- Szybkość — dwa aspekty: przepustowość (szerokość pasma) i opóźnienie.
- Przepustowość stała — średnia szybkość przesyłania w trakcie długich transmisji.
- Przepustowość efektywna — średnia z całego czasu trwania operacji (łącznie z czasem szukania i czasem przełączania kaset).
- Najszybsze przewijaki mają większą przepustowość niż dyski wymienne i stałe(!)
- Opóźnienie dostępu — czas potrzebny do zlokalizowania danych.
  - dla dysków — rzędu milisekund
  - dla taśm — dziesiątki, setki sekund
  - dla taśm średnio tysiąc razy wolniej niż dla dysku
- Niski koszt pamięci trzeciorzędnej wynika z małego kosztu wielu woluminów na jeden kosztowny napęd.
- Niezawodność:
  - dyski stałe są nieco bardziej niezawodne niż kasety dyskowe
  - kasety z dyskami optycznymi uważa się za bardzo niezawodne
  - niezawodność taśm zależy od rodzaju napędu

  - awaria głowicy dysku stałego niszczy dysk całkowicie

- Koszty (ceny minimalne ogłoszeń z miesięcznika BYTE — rynek małych komputerów):
- Koszt pamięci DRAM (*dynamic random access memory*)
  - koszt pamięci waha się znacznie
  - pamięćoperacyjna jest znacznie droższa niż inne
- Koszt twardych dysków magnetycznych.
  - cena pamięci dyskowej jest konkurencyjna w stosunku do pamięci taśmowej (przy jednej taśmie)
  - porównywalne pojemności dysków i taśm
- Koszt pamięci taśmowej.




100
50
20
5
2
0.5
0.2
0.05
0.02
0.004
$/MB
10 MB
20 MB
120 MB
1.2 GB
2 GB
19 GB
45 GB
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
Year


40
20
8
2
0.5
0.1
0.025
$/MB
60 MB
120 MB
1.2 GB
4 GB
72 GB
1984
1986
1988
1990
1992
1994
1996
1998
2000
Year