SISTEMAS DE ENTRADA/SAÍDA

- ***Hardware* de E/S**
- **Mecanismos de E/S**
- **Interface de E/S das aplicações**
- **Estrutura de E/S do *kernel***
- **Gestão do disco**
- **Arquitecturas *RAID***

Hardware de E/S

- **Grande variedade de dispositivos de E/S**

- **Conceitos comuns**
  - **Porto (*port*) / porta - ponto de ligação**
    - » **registos: estado, controlo, dados de entrada, dados de saída**
  - **Barramento (*bus*) - conj. de "fios condutores" + protocolo (sinais, *timings*, …)**
  - **Controlador - electrónica que opera um porto / barramento / dispositivo**
    - » **registos de dados + sinais de controlo**
    - » **ex: controlador de disco (*built-in*)**
      - ***buffering, caching, bad-sector mapping, ...***

- **Comunicação**
  - » **instruções de E/S especiais**
  - » **E/S mapeada em memória**
  - » **híbrida -> instr. de E/S especiais + E/S map. mem. (ex: controlador gráfico)**

Dispositivos de E/S

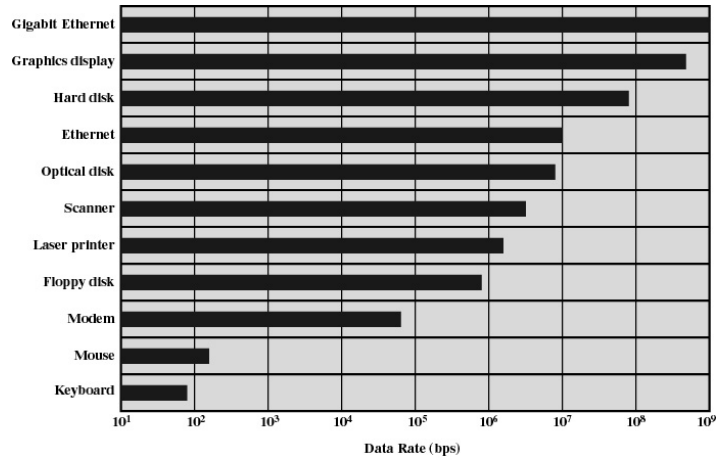
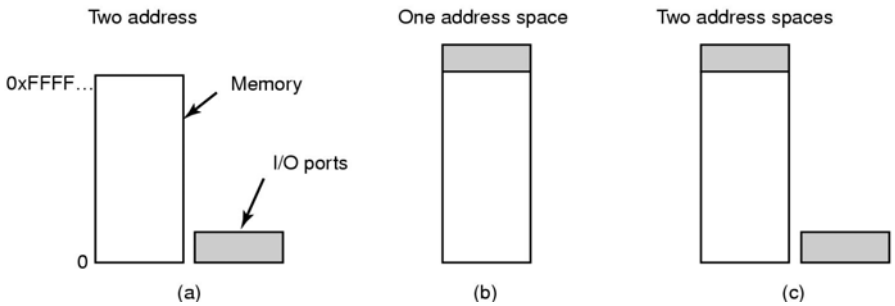


Figure 11.1 Typical I/O Device Data Rates

Comunicação com dispositivos de E/S

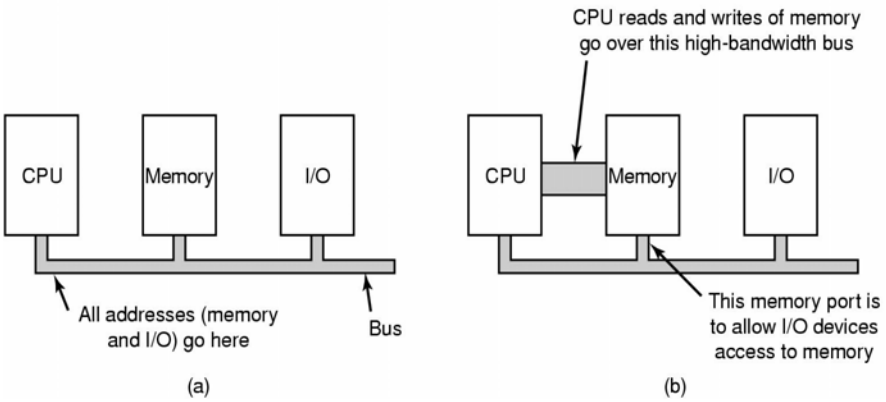


(a) Separate I/O and memory space
(b) Memory-mapped I/O
(c) Hybrid

Endereços de portas de E/S num *PC*

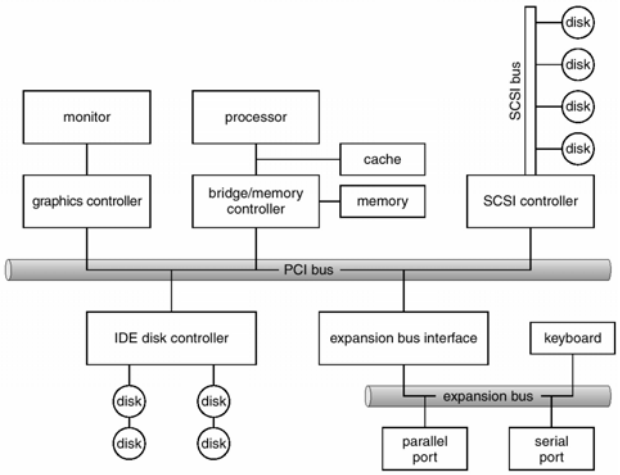
I/O address range (hexadecimal)	device
000-00F	DMA controller
020-021	interrupt controller
040-043	timer
200-20F	game controller
2F8-2FF	serial port (secondary)
320-32F	hard-disk controller
378-37F	parallel port
3D0-3DF	graphics controller
3F0-3F7	diskette-drive controller
3F8-3FF	serial port (primary)

Barramento

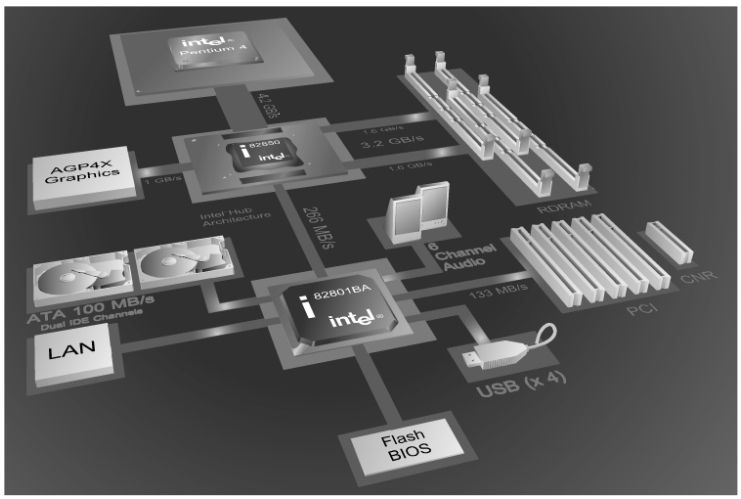


(a) Single-bus architecture
(b) Dual-bus memory architecture

Arquitectura típica de um barramento de um *PC*



Intel Pentium 4 + Intel 850 *chipset*

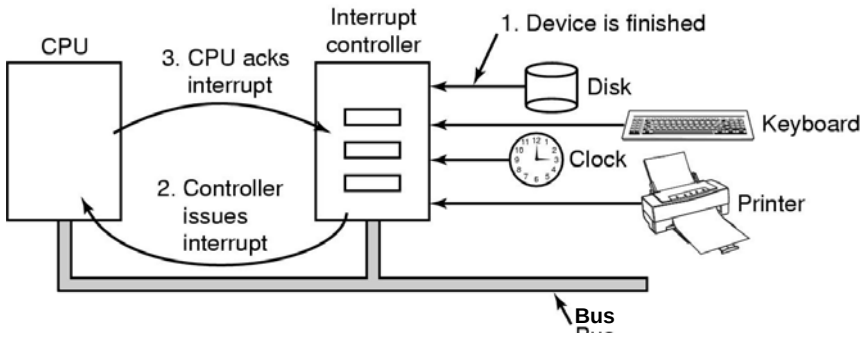


Mecanismo de E/S

- ***Polling / Busy-waiting***
  - **ciclo de leitura do *status register***
  - **eficiente (…mas se o dispositivo estiver frequentem. indisponível …)**
- **Interrupção**
  - **linhas de pedido de interrupção *(CPU)***
    - » **mascarável (interrup.s 0-31, no Pentium), não-mascarável (32-255)**
  - **controlador de interrupções**
    - » **vários níveis de interrupção**
  - **vector de interrupções**
    - » **contém endereços dos *handlers***
    - » **o mecanismo de interrupções aceita um índice deste vector**

  (chamadas ao sistema - implementadas via interrupção por *software*)
- ***DMA (Direct Memory Access)***
  - **evitar E/S programada para grandes transferências de dados**
  - **transferência directa entre dispositivo de E/S e memória**
  - **requer controlador de DMA**

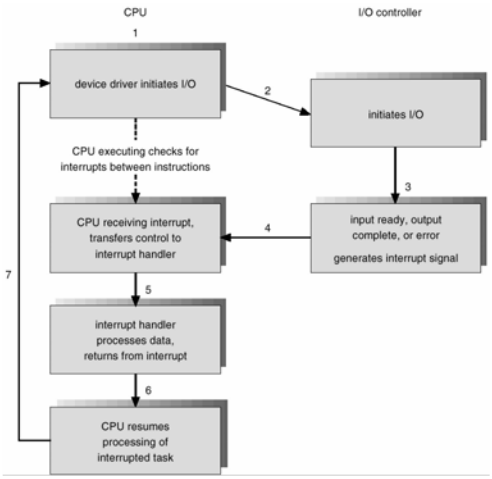
Interrupções



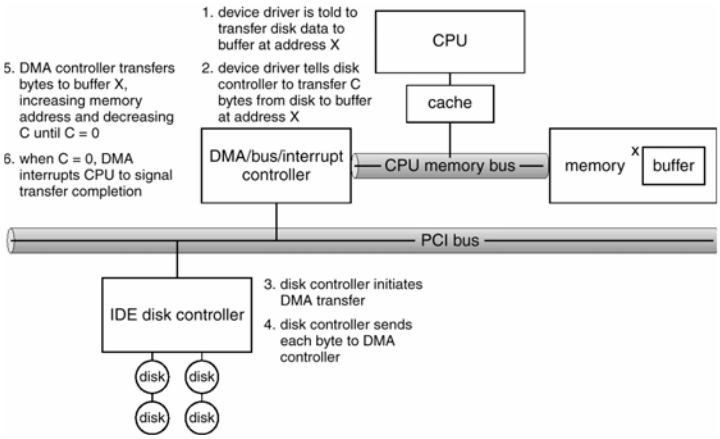
Interrupções do Intel Pentium

vector number	description
0	divide error
1	debug exception
2	null interrupt
3	breakpoint
4	INTO-detected overflow
5	bound range exception
6	invalid opcode
7	device not available
8	double fault
9	coprocessor segment overrun (reserved)
10	invalid task state segment
11	segment not present
12	stack fault
13	general protection
14	page fault
15	(Intel reserved, do not use)
16	floating-point error
17	alignment check
18	machine check
19Ð31	(Intel reserved, do not use)
32Ð255	maskable interrupts

E/S por interrupção



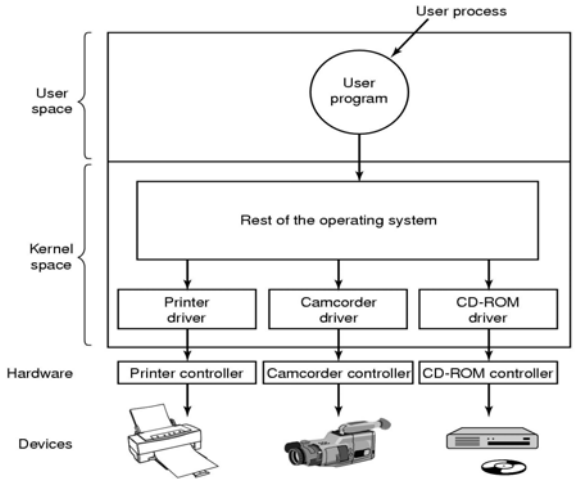
E/S por *DMA*



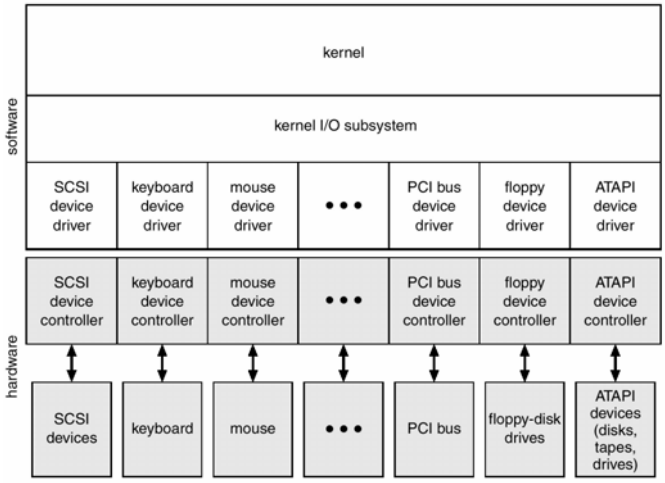
Interface de E/S das aplicações

- **Estruturação em camadas**
- **As diferenças entre dispositivos são encapsuladas em módulos do *kernel* que fornecem uma interface comum às aplicações**
- ***Device drivers* - "escondem" as diferenças entre os dispositivos do subsistema de E/S do *kernel***
- **As características dos dispositivos são muito variadas:**
  - **modo de transferência de dados (carácter / bloco)**
  - **método de acesso (sequencial / directo)**
  - **método de transferência (síncrono / assíncrono)**
  - **partilhado / dedicado**
  - **velocidade**
  - **leitura / escrita / leitura-escrita**

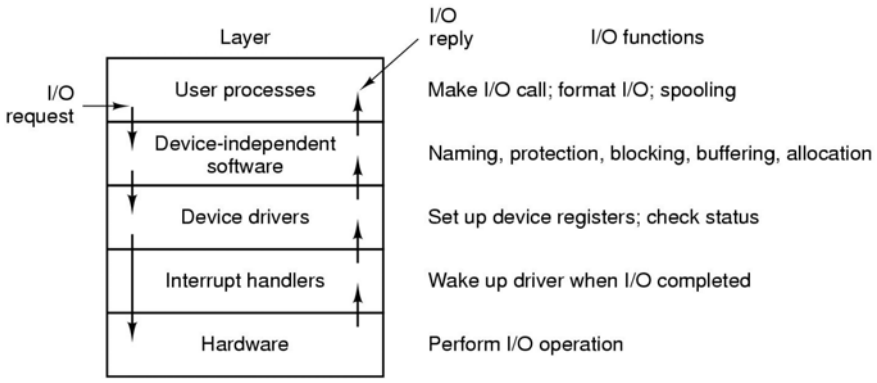
Camadas de *software* de E/S



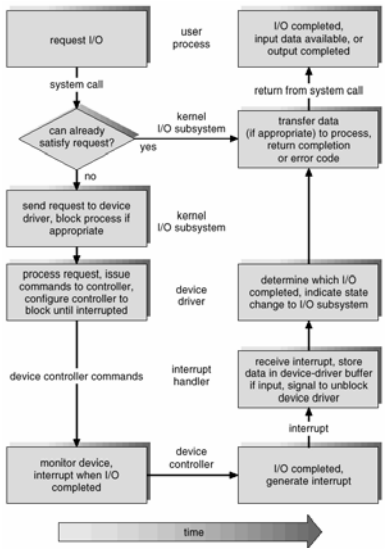
Estrutura de E/S do *kernel*



Camadas de *software* de E/S



"Ciclo de vida" de um pedido de E/S



Características dos dispositivos de E/S

aspect	variation	example
data-transfer mode	character block	terminal disk
access method	sequential random	modem CD-ROM
transfer schedule	synchronous asynchronous	tape keyboard
sharing	dedicated sharable	tape keyboard
device speed	latency seek time transfer rate delay between operations	
I/O direction	read only write only readÐwrite	CD-ROM graphics controller disk

Block devices & Character devices

- ***Block devices***
  - **ex: discos**
  - **comandos: read, write seek**
  - **acesso carácter a carácter / registos**
  - **possibilidade de acesso através de mapeamento de fich.s em memória**
- ***Caracter devices***
  - **ex: teclado, rato, porta série, …**
  - **comandos: get, put**
  - **acesso *raw* / *cooked***

Dispositivos de rede
Clocks & Timers

- **Dispositivos de rede**
  - **Interface própria diferente da dos discos**
    - » **chamadas: create, connect, listen, send, receive, select**
  - **Unix e Windows**
    - » **interface de *sockets***

- ***Clocks & Timers***
  - **funções básicas**
    - » **fornecer a hora actual**
    - » **fornecer o tempo decorrido**
    - » **estabelecer um temporizador p/ desencadear a operação X à hora T**

E/S com e sem bloqueio

- **Com bloqueio**
  - **se não houver dados a aplicação bloqueia**
  - **fácil de usar e de compreender**
  - **insuficiente para algumas necessidades**

- **Sem bloqueio**
  - **retorna os dados disponíveis**

- **Assíncrona**
  - **chamada retorna imediatamente antes da E/S se completar**
  - **a E/S completa-se posteriormente**
  - **a aplicação determina que a E/S terminou testando uma variável ou é informada através de um sinal**

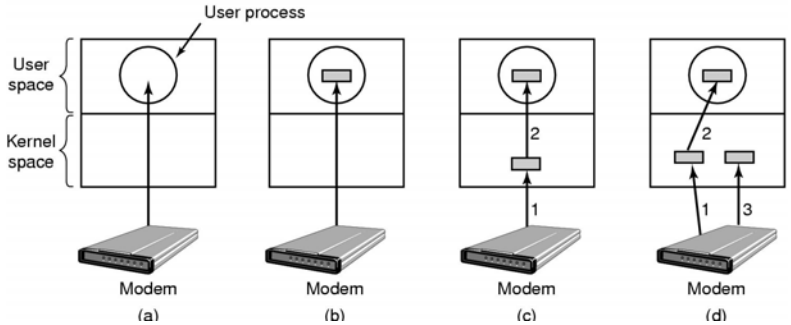
Subsistema de E/S do *kernel*

- **Escalonamento**
  - **melhorar a performance do sistema**
  - **partilhar de forma justa o acesso entre os processos**
  - **reduzir o tempo de espera pelo fim de uma operação de E/S**
- ***Buffering*** ***(buffer - mem. c/dados a transferir entre 2 dispos. ou aplic.-dispos.)***
  - **lidar c/diferenças de velocidade**
    - » **ex: transf. fich. recebido via *modem* p/ disco; *double buffering***
  - **adaptação entre dispositivos que têm diferentes tamanhos de transferência de dados**
    - » **ex: mensagem enviada em pacotes, tem de ser empacotada no receptor**
  - **suportar *copy semantics***
    - » **garantir que o que se manda escrever é aquilo que é escrito independentemente do os dados estarem a ser alterados na aplicação**

Subsistema de E/S do *kernel*

- ***Caching*** ***(cache- mem. rápida que contém cópias de dados)***
  - **apenas uma cópia**
  - **objectivo: melhorar a performance no acesso**
- ***Spooling* e *Device reservation***
  - ***spool* - *buffer* que contém saída para um dispositivo (ex: impressora) que não aceita dados "interlaçados"**
  - **o S.O. intercepta a saída p/ o dispositivo e envia-o para um ficheiro separado para cada aplicação, no disco**
  - **um *daemon* ou *thread* do *kernel* envia o ficheiro para a impressora quando o saída terminar**
  - ***Device reservation***
    - » **quando o S.O. suporta acesso exclusivo a um dispositivo**
    - » **=> chamadas adequadas p/ reservar e libertar o dispositivo**
- **Tratamento de erros**
  - **chamadas retornam código de erro**
    - » **em UNIX, variável *errno* permite obter mais informação**

Buffering



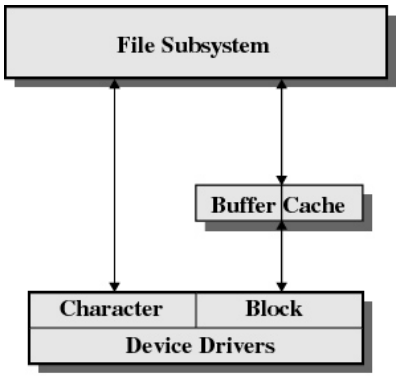
(a) Unbuffered input
(b) Buffering in user space
(c) Buffering in the kernel followed by copying to user space
(d) Double buffering in the kernel

Software de E/S independente do dispositivo

Functions of the device-independent I/O software :

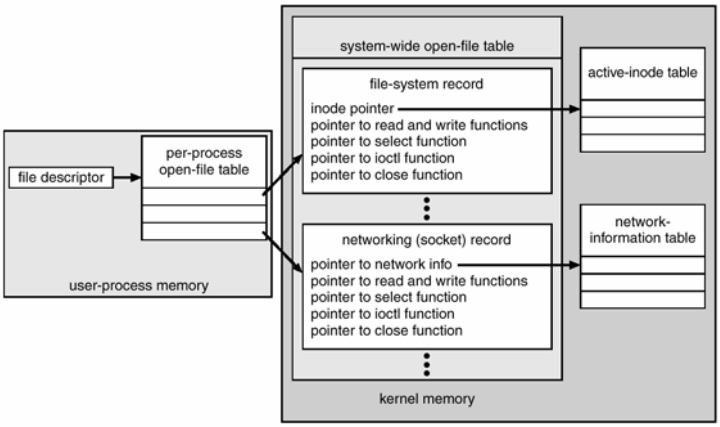
Uniform interfacing for device drivers
Buffering
Error reporting
Allocating and releasing dedicate devices
Providing a device-independent block size

Estrutura de E/S do *kernel* do UNIX



Sistemas Operativos
Sistemas de E/S & Disco

Estrutura de E/S do *kernel* do UNIX

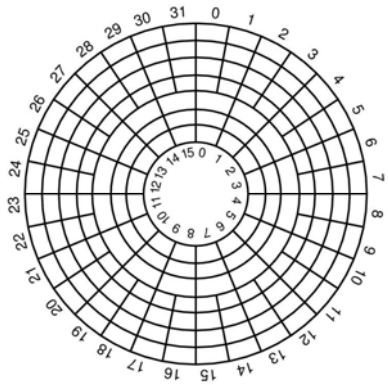


FEUP
MIEIC
Faculdade de Engenharia da Universidade do Porto

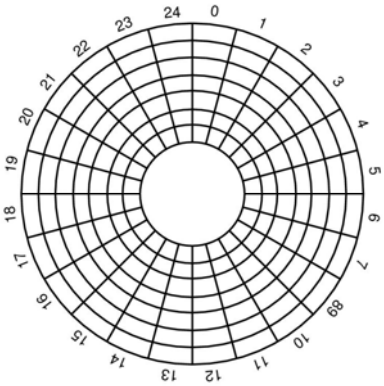
GESTÃO DO DISCO

- Estrutura do disco
- Escalonamento do disco
- Outros aspectos da gestão do disco
- Arquitecturas *RAID*

DISCOS



Physical geometry of a disk with two zones



A possible virtual geometry for this disk

Temporização de uma transferência do disco

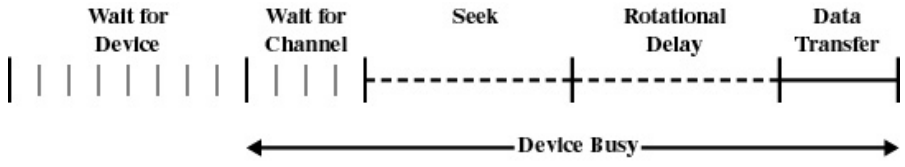


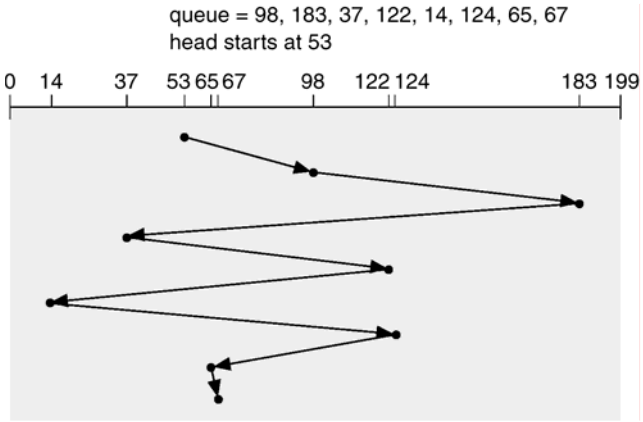
Figure 11.7 Timing of a Disk I/O Transfer

Algoritmos de escalonamento do disco

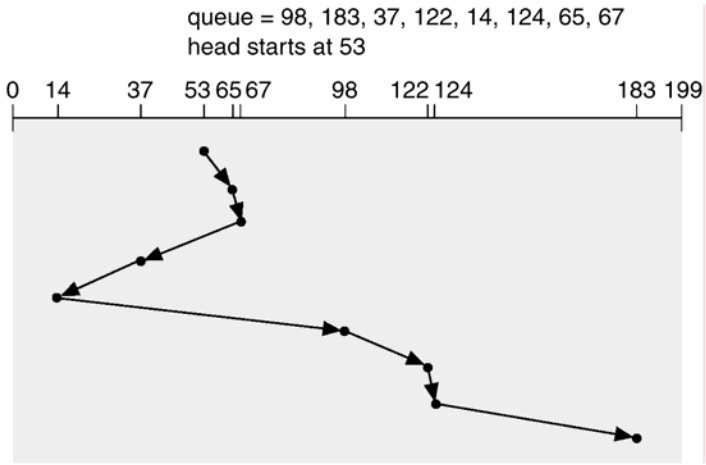
Table 11.3 Disk Scheduling Algorithms [WIED87]

Name	Description	Remarks
Selection according to requestor		
RSS	Random scheduling	For analysis and simulation
FIFO	First in first out	Fairest of them all
PRI	Priority by process	Control outside of disk queue management
LIFO	Last in first out	Maximize locality and resource utilization
Selection according to requested item:		
SSTF	Shortest service time first	High utilization, small queues
SCAN	Back and forth over disk	Better service distribution
C-SCAN	One way with fast return	Lower service variability

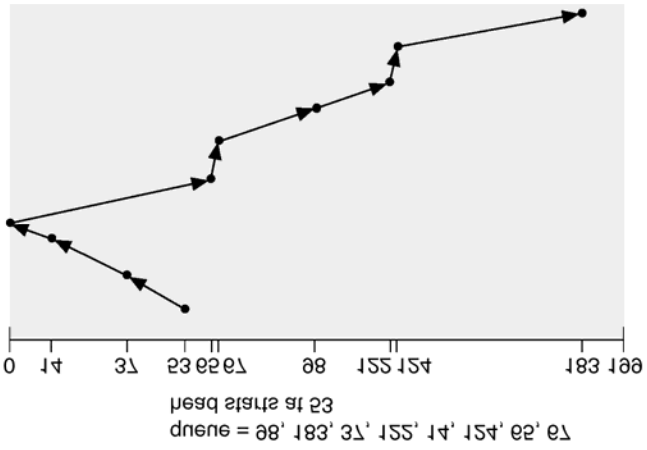
FCFS


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0 14 37 53 65 67 98 122 124 183 199

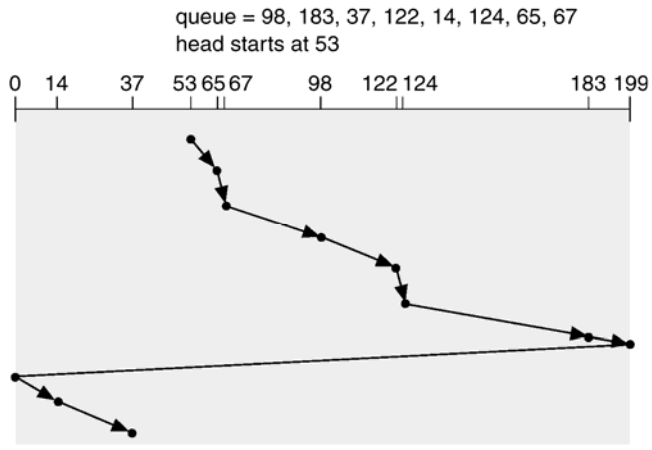
SSTF


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0 14 37 53 65 67 98 122 124 183 199

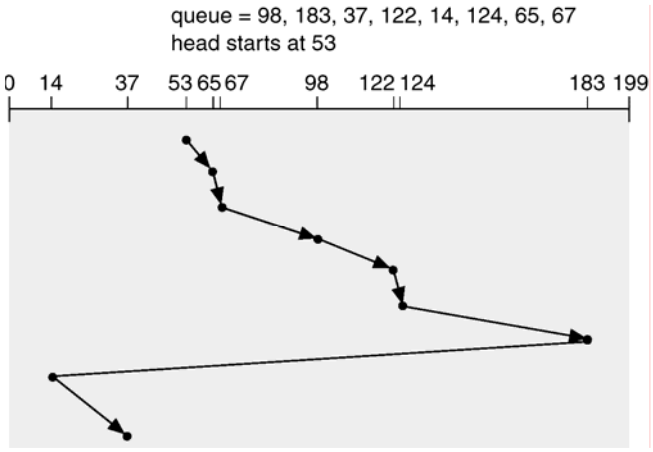
SCAN


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0 14 37 53 65 67 98 122 124 183 199

C-SCAN


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0 14 37 53 65 67 98 122 124 183 199

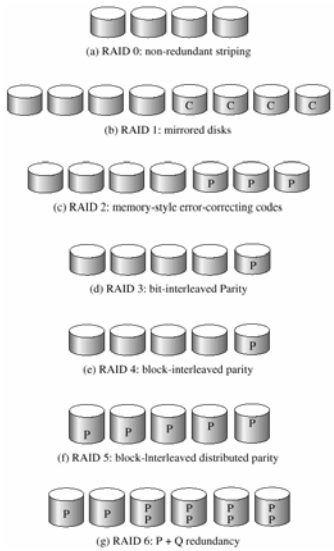
C-LOOK



Outros aspectos da gestão

- **Formatação de baixo nível ou formatação física**
  - **dividir o disco em sectores que o controlador possa ler / escrever**
- **Para usar o disco para guardar ficheiros o S.O. precisa de guardar as suas próprias estruturas no disco**
  - **dividir o disco num ou mais grupos de cilindros (partição)**
  - **criar um sistema de ficheiros (formatação lógica)**
- **Arranque do S.O.**
  - ***bootstrap* armazenado em ROM**
  - **programa de *boot* armazenado no disco**
- **Métodos para lidar com sectores defeituosos (*sector sparing*)**

Arquitecturas *RAID*



RAID **-** ***Redundant Array of Independent Disks***

I=Inexpensive **(no passado)**

- **alternativa a discos grandes (caros)**

I=Independent **(no presente)**

- **maior fiabiliadde**
  - ***mirroring/shadowing*** **(duplicação de dados)**
  - ***block interleaved parity*** **(menor redundância)**
- **maior taxa de transferência**

Arquitecturas *RAID*

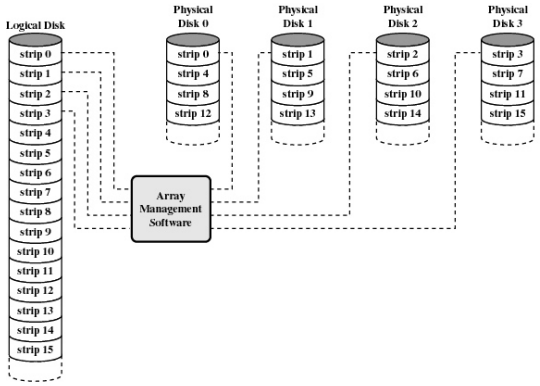
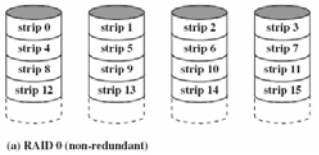


Figure 11.10 Data Mapping for a RAID Level 0 Array [MASS97]

- **não existe redundância, neste caso**
- **apenas *striping* - divisão de um bloco em sub-blocos que são distribuídos "circularmente" pelos vários discos**

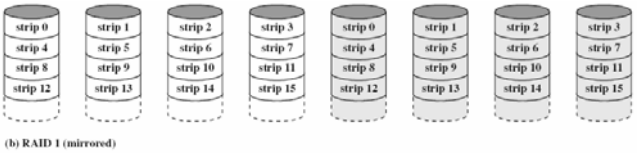
Arquitecturas *RAID*



(a) RAID 0 (non-redundant)

RAID 0

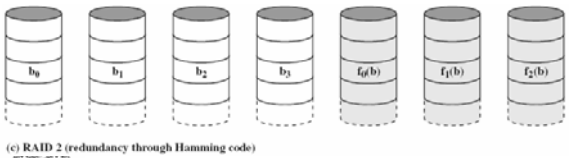
- sem redundância
- *striping*



(b) RAID 1 (mirrored)

RAID 1

- acrescenta redundância (*mirroring*)
- problema: 100% de redundância
- performance
  Read - boa
  Write - pouco melhor do que RAID 0



(c) RAID 2 (redundancy through Hamming code)

RAID 2

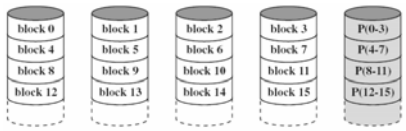
- cabeças sincronizadas
- pequenas *strips*
- *ECC - error correction code (Hamming code)*
- tipicam. não implementado

Arquitecturas *RAID*

(d) RAID 3 (bit-interleaved parity)

RAID 3

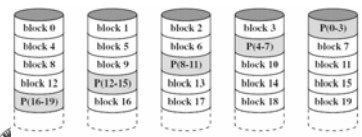
- cabeças sincronizadas
- pequenas *strips* (byte/word)
- simplesmente um *bit* de paridade em vez de *ECC;*
- disco extra p/guardar os *bits* de paridade
- exemplo de cálculo do *bit* de paridade para cada posição de *bit* de uma *strip*:
  - X4(i) = X3(i) ⊗ X2(i) ⊗ X1(i) ⊗ X0(i)
- em caso de avaria de um disco é possível recuperar a informação a partir dos outros exemplo: avaria do disco 1
  - X1(i) = X4(i) ⊗ X3(i) ⊗ X2(i) ⊗ X0(i)



(e) RAID 4 (block-level parity)

RAID 4

- cada *strip* do disco de paridade contém informação de paridade para todas as *strips* correspondentes
- escrita de um *strip* ⇒
  - acesso aos disco que contém o *strip* +
  - + acesso ao disco de paridade

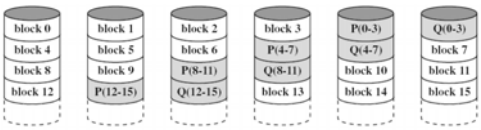


(f) RAID 5 (block-level distributed parity)

RAID 5

- RAID 4 c/ *strips* de paridade distribuídas
- evita a potencial sobreutilização do disco de paridades que pode ocorrer com RAID 4

Arquitecturas *RAID*



(g) RAID 6 (dual redundancy)

RAID 6

- duplo bloco de paridade (P e Q), geralmente calculadas usando códigos correctores de erros (ex: Reed-Solomon)
- maior fiabilidade: permite recuperar 2 discos avariados simultân.
- necessita de N+2 discos
- menor performance nas escritas do que RAID 5 (necessidade de escrever 2 blocos de paridade por cada bloco de dados)

a) RAID 0 + 1 with a single disk failure

RAID 0+1

- combinação de RAID 0 (⇒ performance) com RAID 1 (⇒ fiabilidade)
- os discos são *striped*
- as *strips* são *mirrored* noutros discos

b) RAID 1 + 0 with a single disk failure

RAID 1+0

- os discos são *mirrored* aos pares e os pares de discos resultantes são *striped*

Existem outras variantes de RAID