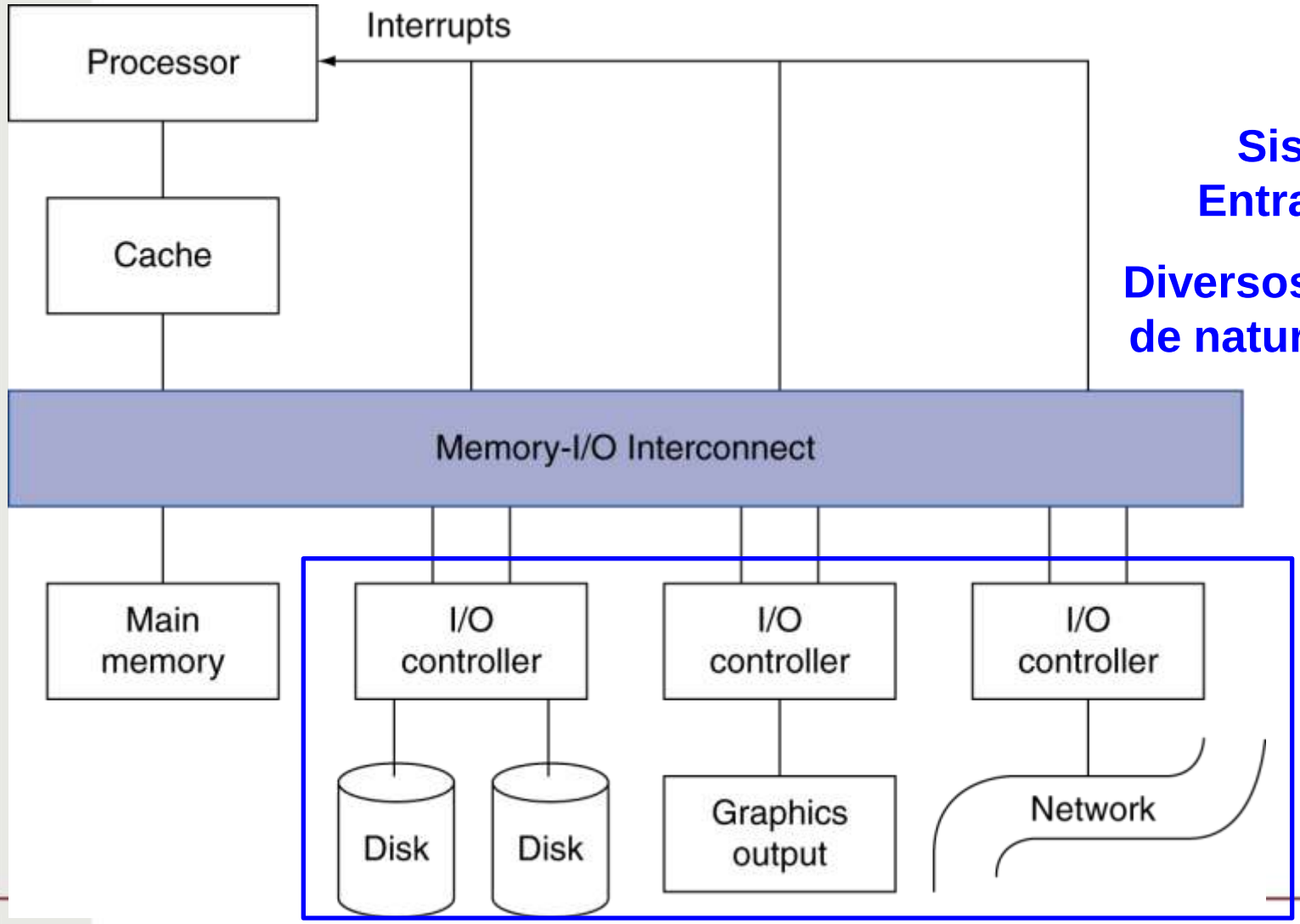
Infraestrutura de Hardware

Entrada/Saída: Armazenamento

Perguntas que Devem ser Respondidas ao Final do Curso

- Como um programa escrito em uma linguagem de alto nível é entendido e executado pelo HW?
- Qual é a interface entre SW e HW e como o SW instrui o HW a executar o que foi planejado?
- O que determina o desempenho de um programa e como ele pode ser melhorado?
- **Que técnicas um projetista de HW pode utilizar para melhorar o desempenho?**

Organização Típica de Sistemas Computacionais



Sistema de Entrada/Saída:

Diversos dispositivos de natureza diferente

Sistema de E/S e Desempenho do Sistema

- Melhora de desempenho de CPU: 60% por ano
- Desempenho de Sistemas de E/S: Limitado por Delays Mecânicos (disco E/S)
  - Melhora de 10% por ano
- Lei de Amdahl: Speed-up Limitado pelo Sub-Sistema mais lento!
- E/S : Gargalo
  - Reduz a fração do tempo na CPU
  - Reduz o valor de CPUs mais rápidas

Desempenho do Sistema com E/S

Problema:

Suponha que uma aplicação tenha $Tempo_{execução}$ = 100s, onde $Tempo_{CPU}$ = 90s e $Tempo_{E/S}$ = 10s

Considerando que o número de CPUs dobre a cada 2 anos, calcule o quanto o desempenho da aplicação vai melhorar em 6 anos (supondo que não há melhora no desempenho de E/S)

Desempenho com E/S

N anos	$Tempo_{CPU}$	$Tempo_{E/S}$	$Tempo_{total}$	*% tempo$_{I/O}$*
0	90	10	100	*10*
2	90/2=45	10	55	*18*
4	45/2 = 23	10	33	*31*
6	23/2 = 11	10	21	*47*

Melhora do $Desempenho_{CPU}$ = 90/11 = 8

Melhora do $Desempenho_{Total}$ =100/21 = 4.7

Dispositivos de E/S

- Podem diferir em:
  - Comportamento: entrada, saída, armazenamento
  - Interação: humano ou máquina
  - Taxa de transferência: dados/seg, operações/seg

Device	Behavior	Partner	Data rate (Mb/s)
Keyboard	input	human	0.0001
Mouse	input	human	0.0038
Laser printer	output	human	3.2
Magnetic disk	storage	machine	800-3000
Graphics display	output	human	800-8000
Network/LAN	input or output	machine	100-10000

Projeto de Sistema de E/S

- Considerações de projeto:
  - possibilidade de expandir o sistema
  - diversidade de dispositivos
  - comportamento no caso de falhas
  - custo
  - Desempenho
- **Maior ênfase em tolerância a falhas e custo**
- Desktops & sistemas embarcados
  - Importante a diversidade de dispositivos
- Servidores
  - Importante a expansibilidade de dispositivos e tolerância a falhas

Desempenho de Sistema de E/S

- Diferentes métricas, depende da aplicação
  - tempo de resposta(latência)
  - taxa de transferência (throughput)
    - Quantidade de dados transferidos
    - Quantidade de operações de E/S
- Desktops & sistemas embarcados
  - Importante a latência
- Servidores
  - Importante o throughput
- Difícil de medir, dependências:
  - características do dispositivo
  - conexão com o sistema
  - hierarquia de memória
  - sistema operacional (software de I/O)

Dependabilidade (Dependability)

- Dependabilidade é a propriedade que define a capacidade dos sistemas computacionais de prestar um serviço que se pode justificadamente confiar

- Atributos importantes:
  - Confiabilidade (reliability)
  - Segurança
  - Disponibilidade (availability)
  - Mantenabilidade

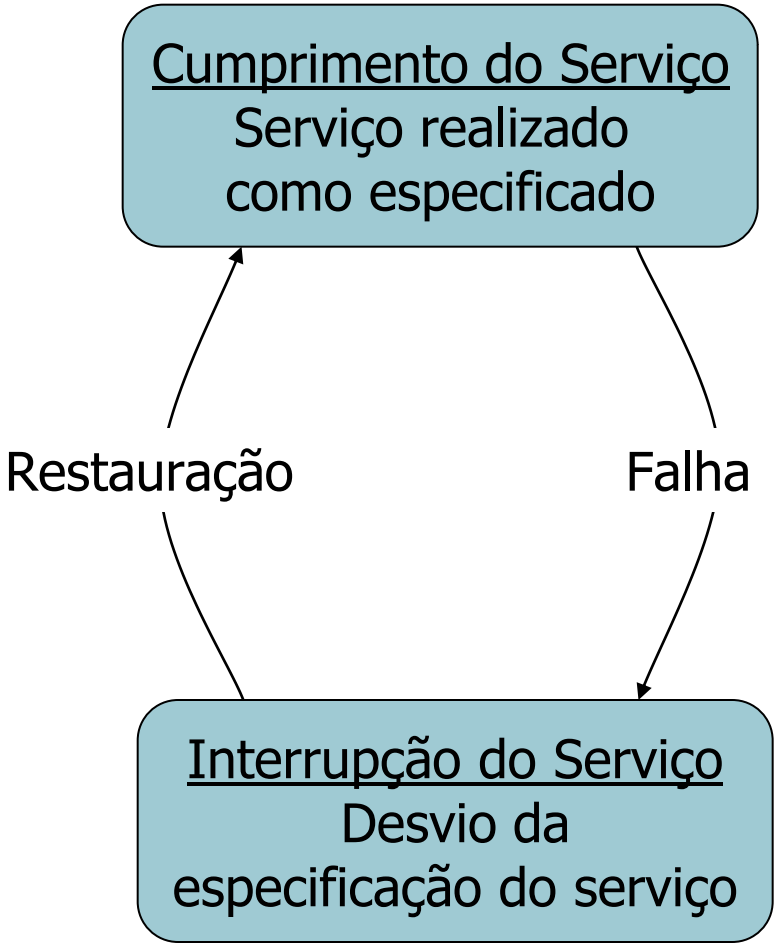
- Dependabilidade é importante em um sistema
  - **Particularmente para dispositivos de armazenamento**

Impacto da Dependabilidade em Um Sistema

Application	Cost of downtime per hour (thousands of $)	Annual losses (millions of $) with downtime of 1% (87.6 hrs/yr)	0.5% (43.8 hrs/yr)	0.1% (8.8 hrs/yr)
Brokerage operations	$6450	$565	$283	$56.5
Credit card authorization	$2600	$228	$114	$22.8
Package shipping services	$150	$13	$6.6	$1.3
Home shopping channel	$113	$9.9	$4.9	$1.0
Catalog sales center	$90	$7.9	$3.9	$0.8
Airline reservation center	$89	$7.9	$3.9	$0.8
Cellular service activation	$41	$3.6	$1.8	$0.4
Online network fees	$25	$2.2	$1.1	$0.2
ATM service fees	$14	$1.2	$0.6	$0.1

Figure 1.3 The cost of an unavailable system is shown by analyzing the cost of downtime (in terms of immediately lost revenue), assuming three different levels of availability, and that downtime is distributed uniformly. These data are from Kembel [2000] and were collected and analyzed by Contingency Planning Research.

Dependabilidade e Falhas



- **Falha (Fault) :** componente não executa como especificado
  - Pode ser permanente ou transiente
  - Pode ou não provocar a falha do sistema

Medindo a Dependabilidade

- Confiabilidade: Mean Time To Failure (MTTF).
- Interrupção de serviço : Mean Time To Repair (MTTR)
- Disponibilidade: MTTF/(MTTF + MTTR)

- Melhorando a disponibilidade

  Aumentando MTTF
  - Fault avoidance
  - Fault tolerance
  - Fault forecasting

  Diminuindo MTTR
  - Melhorando ferramentas e processos de diagnóstico e reparação

Armazenamento Secundário

- Importante para garantir dependabilidade de um sistema
  - Memória e cache → voláteis
  - Armazenamento secundário → dados persistidos
- Tem impacto no desempenho do sistema
  - Faltas na cache e memória → acesso a disco
  - Gargalo no tempo de execução
- Desafios:
  - Como aumentar dependabilidade?
  - Como aumentar desempenho?

Tipos de Armazenamento Secundário

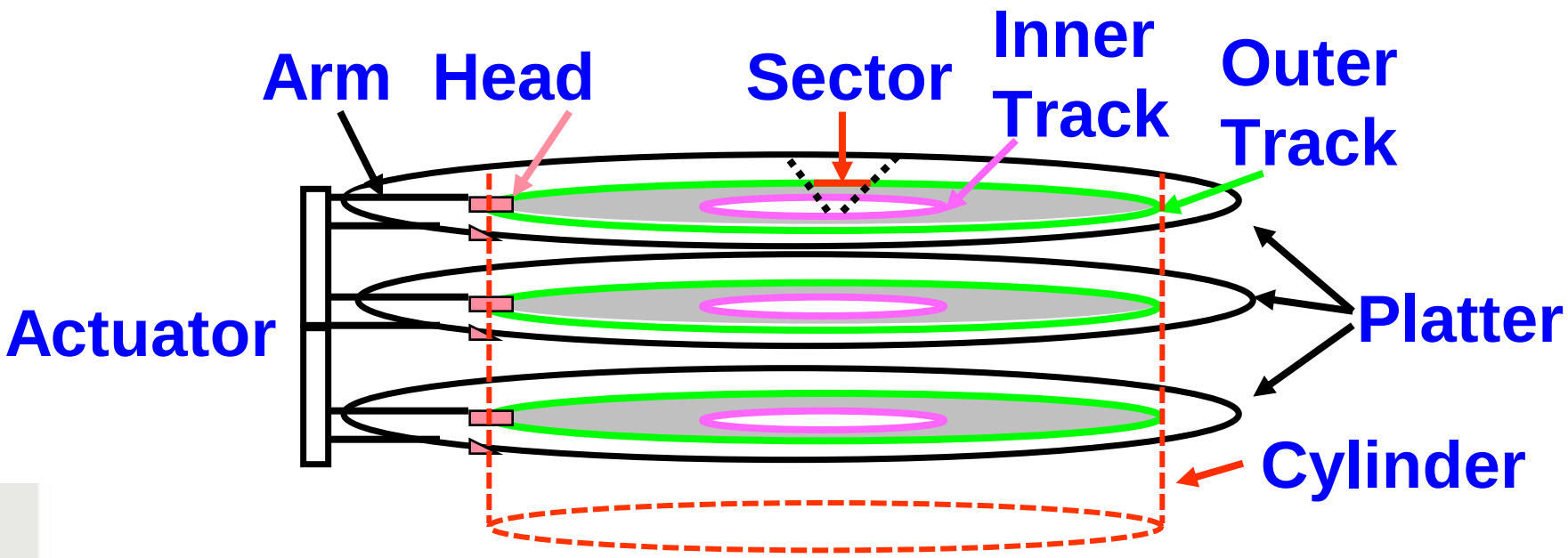
- Disco
  - Dispositivo magnético
    - Partes gravadas são magnetizadas
  - Possui componentes mecânicos

- Flash
  - Memória semicondutora (EEPROM)
  - Wearable
    - Após 100000-1000000 de escritas pode perder capacidade de armazenamento
    - Wear leveling : Redistribui dados em áreas menos usadas

Disco x Flash

- Velocidade de acesso
  - Flash 100-1000 mais rápida do que disco
- Preço
  - Disco até 40x mais barato (2008)
    - Mais popular em desktops e servidores
- Wearable
  - Disco sem limite para escrita
- Energia e Robustez
  - Flash mais eficiente e robusto
    - Mais popular em sistemas embarcados

Disco Rígido: Terminologia



- Vários pratos, com a informação armazenada magneticamente em ambas superfícies (usual)
- Bits armazenados em trilhas, que por sua vez são divididas em setores (e.g., 512 Bytes)
- Atuador move a cabeça (fim do braço,1/superfície) sobre a trilha ("seek"), seleciona a superfície, espera pelo setor passar sob a cabeça, então lê ou escreve
  - "Cilindro": todas as trilhas sob as cabeças

Setores do Disco e Acesso

- Cada setor armazena
  - ID do setor
  - Dados (512 bytes, 4096 bytes proposto)
  - Error correcting code (ECC)
    - Usado para esconder defeitos e erros de leitura
  - Campos de sincronização e espaços

- Acesso a setores envolve
  - Delays devidos a espera se outros acessos foram feitos antes
  - Procura (Seek): mover as cabeças
  - Latência rotacional
  - Transferência de dados
  - Overhead do controlador

Calculando o Desempenho de um Disco

Disk Latency = Seek Time + Rotation Time + Transfer Time + Controller Overhead

- **Seek Time**

  Depende do no. de trilhas e velocidade de seek do disco

  Fabricantes anunciam seek time "pessimistas"

  - Seek time real entre 25% - 33% do anunciado

- **Rotation Time(Latency)**

  Tempo para o setor girar embaixo da cabeça

  - depende da velocidade de rotação do disco
  - Tempo médio = 0,5 rotação/velocidade de rotação

- **Transfer Time**

  depende do tamanho do setor, densidade dos bits por trilha, tamanho da requisição, taxa de transferência

Exemplo : Cálculo de Acesso ao Disco

Problema:

Tamanho do Setor = 512 bytes
Velocidade de rotação = 15000 rpm
Tempo médio de seek = 4ms
Taxa de transferência = 100MB/s
Overhead do controlador = 0,2ms

Considerando que o disco está desocupado e tempo real de seek é 25% do anunciado, calcule o tempo de acesso ao disco

Exemplo : Cálculo de Acesso ao Disco

Problema:

Tamanho do Setor = 512 Bytes = 0,5 KB
Velocidade de rotação = 15000 rpm
Tempo médio de seek = 4ms
Taxa de transferência = 100MB/s
Overhead do controlador = 0,2ms

Calculando tempo, utilizando seek anunciado:

Disk Latency = Seek time + Rotation time + Transfer time + Controller Overhead
Disk Latency = 4ms + 0,5 rotacao/15000 rotacao/60 + 0,5KB/100MB/s + 0,2ms

Disk Latency = 4ms + 2ms + 0,005ms + 0,2ms = 6,205ms

Exemplo : Cálculo de Acesso ao Disco

Problema:

Tamanho do Setor = 512 Bytes = 0,5 KB
Velocidade de rotação = 15000 rpm
Tempo médio de seek = 4ms
Taxa de transferência = 100MB/s
Overhead do controlador = 0,2ms

Calculando tempo, utilizando seek real:

Disk Latency = Seek time + Rotation time + Transfer time + Controller Overhead
Disk Latency = 4ms x 0,25 + 0,5 rotacao/15000 rotacao/60 + 0,5KB/100MB/s + 0,2ms

Disk Latency = 1ms + 2ms + 0,005ms + 0,2ms = 3,205ms

Melhorando Desempenho: Localidade

- Fabricantes anunciam tempo médio de seek
  - Baseado em todos os seek possíveis
  - Localidade e escalonamento de acessos pelo S.O podem diminuir tempo de seek
    - Depende de aplicação e de algoritmo de escalonamento

Melhorando Desempenho: Controladores Inteligentes

- **Controladores com microprocessadores**

  Podem otimizar desempenho

  Apresentam interface alto nível permitindo que S.O enxergue blocos lógicos e escalone acessos

  Interfaces padrões (comandos, protocolos, interface elétrica) que permite conexão e transferências de dados entre computadores e periféricos

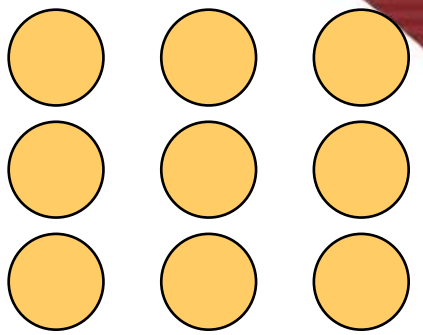
  - SCSI (Small Computer Systems Interface)
  - SATA (Serial Advanced Technology Attachment)

Melhorando Desempenho: Controladores Inteligentes

- **Controladores incluem caches**
  - Mantém dados acessados recentemente
  - Controlador implementa algoritmos que fazem a busca antecipada de setores que tem probabilidade maior de serem buscados
  - Evita seek e latência rotacional

Melhorando Desempenho : RAIDs:

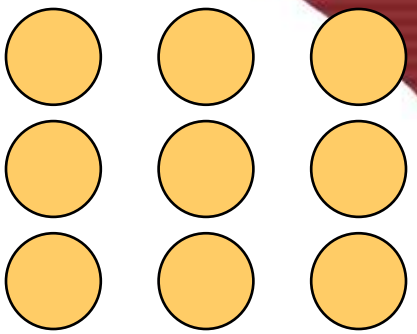
Redundant Array of
Inexpensive Disks



- Arrays de discos pequenos e baratos
  - Paralalelismo aumenta desempenho
    - Dado espalhado sobre múltiplos discos
    - Acessos múltiplos são feitos a vários discos simultaneamente

Melhorando Dependabilidade : RAIDs:

Redundant Array of
Inexpensive Disks



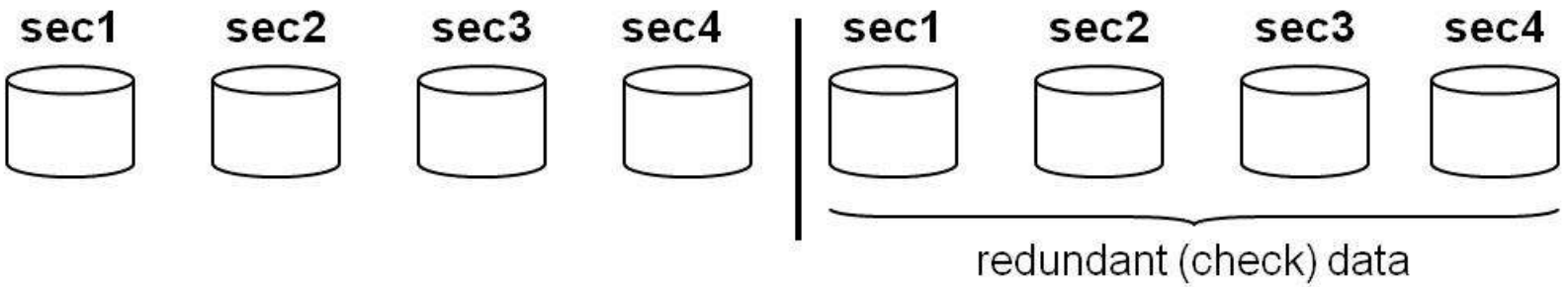
- Confiabilidade < disco UNICO

- MAS disponibilidade pode ser melhorada pela adição de discos redundantes(RAID)

  Informação perdida pode ser recuperada através da informação redundante $\rightarrow$ tolerância a falhas

RAID Nível 1

- RAID 1: Mirroring (Espelhamento)
  - Distribuição de blocos sobre discos múltiplos– striping
    - Múltiplos blocos podem ser acessados em paralelo aumentando o desempenho
  - N + N discos → dados replicados
  - Escrita deve ser feita nos discos de dados e nos redundantes
    - Em caso de falha, leitura do espelho



RAID Nível 3: Bit-Interleaved Parity

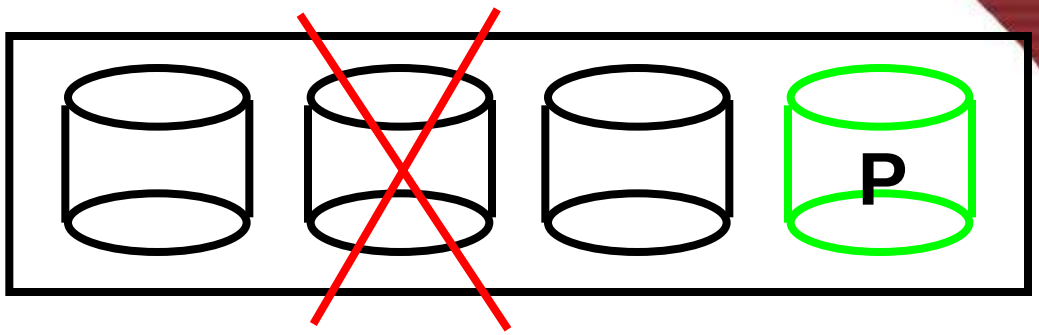
- **N + 1 discos**
  - Dados espalhados em N discos em nível de byte
  - Disco redundante armazena paridade
    - Não é necessário armazenar todo o dado
  - Leitura
    - Leitura em todos os discos
  - Escrita
    - Gera nova paridade e atualiza todos os discos
  - Em uma falha
    - Usa paridade para reconstruir dado

RAID Nível 3: Disco de Paridade

10010011
11001101
10010011
. . .

registro lógico

Striped registros físicos



1	1	1	1
0	1	0	1
0	0	0	0
1	0	1	0
0	1	0	1
0	1	0	1
1	0	1	0
1	1	1	1

P contem a soma dos discos por stripe mod 2 ("parity").
Se disco falha, basta subtrair P da soma dos outros discos para recuperar Informação

RAID Nível 4: Bit-Interleaved Parity

- **N + 1 discos**

  Dados espalhados em N discos em nível de bloco

  Disco redundante armazena paridade para um grupo de blocos

  Leitura

  - Leitura apenas no disco contendo o bloco

  Escrita

  - Leitura no disco contendo bloco modificado e disco de paridade
  - Gera nova paridade e atualiza disco de dado e de paridade

  Em uma falha

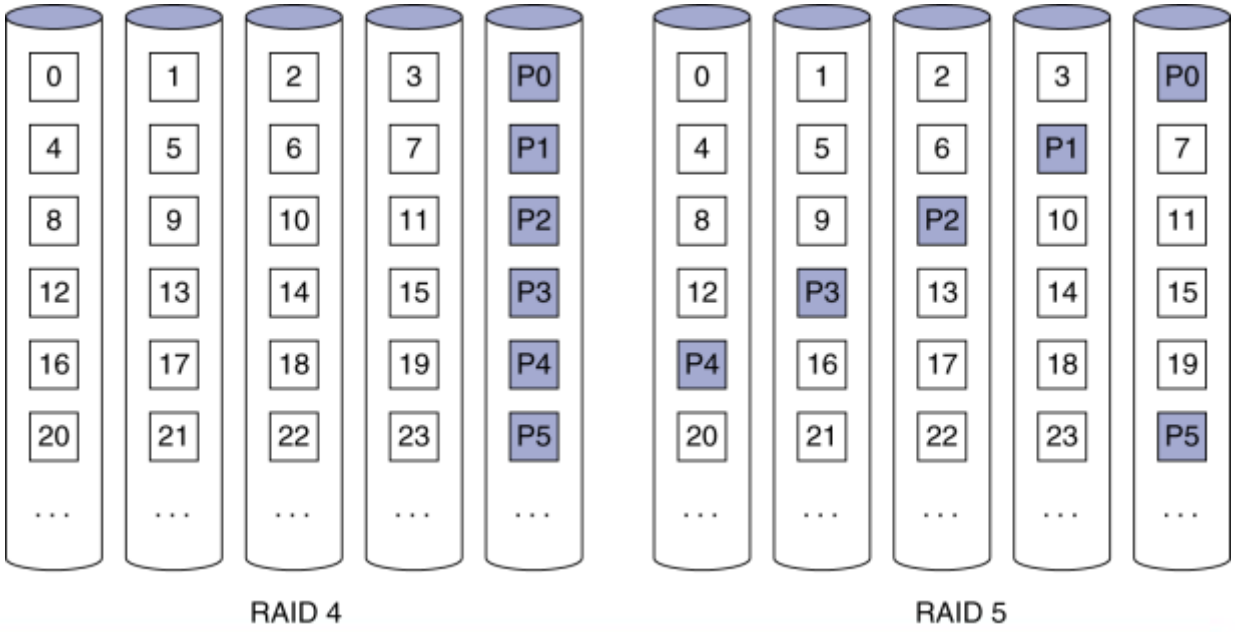
  - Usa paridade para reconstruir dado

RAID Nível 5: Paridade Distribuída

- **N + 1 discos**

  Parecido com RAID 4, mas blocos de paridade distribuídos pelos discos do RAID

  - Evita que disco de paridade seja gargalo
  - Permite escritas simultâneas a diferentes discos



Resumindo RAID…

- RAID pode melhorar desempenho e dependabilidade
  - Acessos paralelos
  - Tolerância a fallhas
- RAID 1 e RAID 5 muito utilizados
- RAID muito usado em servidores
  - Dependabilidade é essencial