Sebenta de SO

Bash Scripting, Filesystems, Sofs17, Programação concorrente, Process switching, Processor Scheduling and Memory Management

PEDRO MARTINS

February 1, 2018

Contents

1 Sistemas de Computação **13**
1.1 Vista simplificada de um sistema de computação . . . 13
1.2 Vista geral . . . 14
1.2.1 Extended Machine . . . 14
Tipos de funções da extended machine . . . 15
1.2.2 Resource Manager . . . 16
1.3 Evolução dos Sistemas Operativos . . . 17

2 Taxonomia de Sistemas Operativos **17**
2.1 Classificação com base no tipo de processamento . . . 17
2.1.1 Multiprogrammed batch . . . 17
2.1.2 Interactive System (Time-Sharing) . . . 18
2.1.3 Real Time System . . . 18
2.1.4 Network Operating System . . . 19
2.1.5 Distributed Operating System . . . 19
2.2 Classificação com base no propósito . . . 20

3 Multiprocessing vs Multiprogramming **20**
3.1 Paralelismo . . . 20
3.2 Concorrência . . . 21

4 Estrutura Interna de um Sistema Operativo **21**
4.1 Design de um sistema operativo . . . 21
4.1.1 Monolithic system . . . 22
4.1.2 Layered Approach: Divisão por camadas . . . 22
4.1.3 Microkernel . . . 23
4.1.4 Virtual machine (hypervisors) . . . 24
4.1.5 Client-Server . . . 25
4.1.6 Exokernels . . . 25
4.2 Estruturas Internas do Unix/Linux e Windows . . . 26
4.2.1 Estrutura Interna do Unix (tradicional) . . . 26
4.2.2 Estrutura Global do Unix . . . 27
4.2.3 Estrutra do Kernel Unix . . . 28
4.2.4 Estrutura do Kernel Windows . . . 29

5 Shell Scripting **30**
5.1 Exercise 1 - Command Overview . . . 30
5.2 Exercise 2 - Redirect input and output . . . 31
5.2.1 1 . . . 31
5.2.2 2 . . . 31
5.2.3 3 . . . 31
5.2.4 4 . . . 31
5.3 Exercise 3 - Using special characters . . . 31
5.3.1 1 . . . 31
5.3.2 2 . . . 31
5.3.3 3 . . . 31

5.3.4 4 . . . 32
5.4 Exercise 4 - Declaring and using variables . . . 32
5.4.1 1 . . . 32
5.4.2 2 . . . 32
5.4.3 3 . . . 32
5.5 Exercise 5 - Declaring and using functions . . . 32
5.5.1 1 . . . 32
5.5.2 2 . . . 33
5.6 Exercise 6 . . . 33
5.6.1 1, 2 e 3 . . . 33
5.7 Exercise 7 . . . 33
5.7.1 1 . . . 33
5.7.2 2 . . . 33
5.7.3 3 . . . 34
5.7.4 4 . . . 34
5.7.5 5 . . . 34
5.8 Exercise 8 - The multiple choice case construction . . . 34
5.8.1 1 Case stament . . . 34
5.8.2 1 . . . 34
5.8.3 2 . . . 35
5.9 10 - The repetitive while and until constructions . . . 35
5.9.1 1 . . . 35
5.10 Exercise 11 - Script Files . . . 35

6 1 **35**
6.1 Exercise 12 - Bash supports both indexed and associative arrays . . . 36
6.1.1 1 . . . 36
6.1.2 2 . . . 36

7 sofs2017 **37**

8 Organização das aulas durante o sofs17 **37**

9 Introduction **37**
9.1 File as an abstract data type . . . 39
9.1.1 Operações em ficheiros . . . 40
9.2 FUSE . . . 41
9.2.1 Infrastrutura . . . 41

10 SOFS17 Architecture **42**
10.1 List of free inodes . . . 43
10.2 List of free clusters . . . 44
10.2.1 Retrieval Chache . . . 44
10.2.2 Insertion cache . . . 44
10.2.3 Allocation . . . 45
10.3 List of clusters used by a file (inode) . . . 45
10.3.1 Considerações: . . . 46
10.3.2 Campo i_1 . . . 47
10.3.3 Campo i_2 . . . 47

10.3.4 NullReference . . . 47
10.4 Directories . . . 47

11 Formatting **48**

12 Code Structure **48**
12.1 Rawdisk . . . 49
12.2 Dealers . . . 49
12.2.1 sbdealer . . . 49
12.2.2 itdealer . . . 49
12.2.3 bmdealer . . . 50
12.2.4 czdealer . . . 50
12.2.5 ocdelaer . . . 50
12.3 ilayers . . . 50
12.3.1 inodeattr . . . 50
12.3.2 freelists . . . 50
12.3.3 filecluster . . . 50
12.3.4 direntries . . . 50
12.4 syscalls . . . 50
12.5 fusecallbacks . . . 50
12.6 probing . . . 50
12.7 exception . . . 51

13 createDisk **51**
13.1 Exemplo de utilização . . . 51
13.2 Implementação . . . 51

14 showblock **52**
14.1 Utilização . . . 52
14.1.1 Opções de Visualização . . . 52
14.2 Exemplos . . . 52

15 rawlevel **54**
15.1 Módulos . . . 54

16 rawdisk **54**
16.1 Macros . . . 54
16.2 Funções . . . 55
16.3 Utilização . . . 56
16.3.1 No Options . . . 56
16.3.2 Set name . . . 56
16.3.3 Set inodes . . . 57
16.3.4 Zero Mode . . . 57

17 computeStruture **57**
17.1 Algoritmo . . . 58
17.2 Utilização . . . 58
17.2.1 Parameters . . . 58
17.3 Testes . . . 58

17.3.1 1000 blocos, 125 inodes (nblocos/8) . . . 58

18 fillInSuperBlock **59**
18.1 Algoritmo . . . 59
18.2 Utilização . . . 60
18.2.1 Parameters . . . 60

19 fillInInodeTable **60**
19.1 Algoritmia . . . 60
19.2 Utilização . . . 61
19.2.1 Parameters . . . 61

20 fillInFreeClusterTable **61**
20.1 Algoritmo . . . 62
20.1.1 Considerações . . . 62
20.2 Utilização . . . 63
20.2.1 Parameters . . . 63
20.2.2 Data Structure . . . 63
20.3 Testes . . . 63

21 fillInRootDir **65**
21.1 Algoritmia . . . 65
21.2 Utilização . . . 66
21.2.1 Parameters . . . 66

22 resetClusters **66**
22.1 Algoritmia . . . 66
22.2 Utilização . . . 66
22.2.1 Parameters . . . 66

23 freelists **66**
23.1 soAllocInode . . . 67
23.2 soFreeCluster . . . 67
23.3 soFreeInode . . . 67
23.4 soReplenish . . . 68
23.5 soDeplete . . . 68

24 Cenário Inicial **68**
24.1 freeinode . . . 69
24.2 inserir inode 200 . . . 69
24.3 soAllocateInode . . . 69
24.4 iOpen . . . 70
24.5 iSave . . . 70
24.6 iClose . . . 70
24.7 Interface com os inodes é suposto usar uma estrutura de inodes . . . 70

25 mais difícis (5) **70**

26 intermédias (3) **70**

27 mais triviais (1) **70**
27.1 Utilização . . . 70
27.2 Uma posição para referência dupla indireta . . . 71
27.3 Doxygen . . . 71
27.3.1 uint32_t soGetFileCluster (int ih, . . . 72
O que é preciso fazer: . . . 73
Testes . . . 73
27.4 Your command: . . . 74
27.4.1 void soReadFileCluster (int ih, . . . 74
27.5 soFreeFileCLusters . . . 75

28 soGetDirEntry **75**

29 soRenameDirEntry **75**

30 soTraversePath **76**

31 soAddDirEntry **76**

32 soDeleteDirEntry **76**

33 Extra **77**

34 soRenameDirEntry **77**

35 soDeleteDirEntry **77**

36 soGetDirEntry **77**
36.1 iOpen . . . 77
36.2 Main syscalls . . . 78
36.3 Other syscalls . . . 78
36.4 soLink . . . 78
36.5 unLink . . . 78
36.6 soRename . . . 79
36.7 soMKnod . . . 79
36.8 soRead . . . 79
36.9 soTrucnate . . . 80
36.10 soMkdir . . . 80
36.11 soReadDIr . . . 80
36.12 soSymlink . . . 80
36.13 so ReadLink . . . 80

37 Make **81**

38 soFreeFileCLusters **81**
41.1 Notes . . . 82

42 3 Nov 2017 **82**

43 Unlink **82**

44 Remove **82**

45 mtime vs ctime **82**

46 Conceitos Introdutórios **84**
46.1 Exclusão Mútua . . . 84

47 Acesso a um Recurso **84**

48 Acesso a Memória Partilhada **85**
48.1 Relação Produtor-Consumidor . . . 86
48.1.1 Produtor . . . 86
48.1.2 Consumidor . . . 86

49 Acesso a uma Zona Crítica **87**
49.1 Tipos de Soluções . . . 87
49.2 Alternância Estrita *(Strict Alternation)* . . . 88
49.3 Eliminar a Alternância Estrita . . . 88
49.4 Garantir a exclusão mútua . . . 89
49.5 Garantir que não ocorre deadlock . . . 89
49.6 Mediar os acessos de forma determinística: *Dekker agorithm* . . . 90
49.7 Dijkstra algorithm (1966) . . . 91
49.8 Peterson Algorithm (1981) . . . 92
49.9 Generalized Peterson Algorithm (1981) . . . 93

50 Soluções de Hardware **94**
50.1 Desativar as interrupções . . . 94
50.2 Instruções Especiais em Hardware . . . 94
50.2.1 Test and Set (TAS primitive) . . . 94
50.2.2 Compare and Swap . . . 95
50.3 Busy Waiting . . . 95
50.4 Block and wake-up . . . 96

51 Semáforos **97**
51.1 Implementação . . . 97
51.1.1 Operações . . . 98
51.1.2 Solução típica de sistemas *uniprocessor* . . . 98
51.2 Bounded Buffer Problem . . . 99
51.2.1 Como Implementar usando semáforos? . . . 100
51.3 Análise de Semáforos . . . 102
51.3.1 Vantagens . . . 102
51.3.2 Desvantagens . . . 102
51.3.3 Problemas do uso de semáforos . . . 102
51.4 Semáforos em Unix/Linux . . . 102

52 Monitores **103**
52.1 Implementação . . . 104
52.2 Tipos de Monitores . . . 105
52.2.1 Hoare Monitor . . . 105
52.2.2 Brinch Hansen Monitor . . . 106
52.2.3 Lampson/Redell Monitors . . . 107

52.3 Bounded-Buffer Problem usando Monitores . . . 107
52.4 POSIX support for monitors . . . 109

53 Message-passing **109**
53.1 Direct vs Indirect . . . 110
53.1.1 Symmetric direct communication . . . 110
53.2 Asymmetric direct communications . . . 110
53.3 Comunicação Indireta . . . 110
53.4 Implementação . . . 110
53.5 Buffering . . . 111
53.6 Bound-Buffer Problem usando mensagens . . . 111
53.7 Message Passing in Unix/Linux . . . 112

54 Shared Memory in Unix/Linux **112**
54.1 POSIX Shared Memory . . . 113
54.2 System V Shared Memory . . . 113

55 Deadlock **113**
55.1 Condições necessárias para a ocorrência de deadlock . . . 114
55.1.1 O Problema da Exclusão Mútua . . . 115
55.2 Jantar dos Filósofos . . . 115
55.3 Prevenção de Deadlock . . . 116
55.3.1 Negar a exclusão mútua . . . 117
55.3.2 Negar *hold and wait* . . . 117
55.3.3 Negar *no preemption* . . . 117
55.3.4 Negar a espera circular . . . 118
55.4 Deadlock Avoidance . . . 119
55.4.1 Condições para lançar um novo processo . . . 119
55.4.2 Algoritmo dos Banqueiros . . . 120
Algoritmo dos banqueiros aplicado ao Jantar dos filósofos . . . 120
55.5 Deadlock Detection . . . 121

56 Processes and Threads **123**
56.1 Arquitectura típica de um computador . . . 123
56.2 Programa vs Processo . . . 123
56.3 Execução num ambiente multiprogramado . . . 123
56.4 Modelo de Processos . . . 124
56.5 Diagrama de Estados de um Processo . . . 125
56.5.1 Swap Area . . . 126
56.5.2 Temporalidade na vida dos processos . . . 127
56.6 State Diagram of a Unix Process . . . 129
56.7 Supervisor preempting . . . 130
56.8 Unix – traditional login . . . 130
56.9 Criação de Processos . . . 131
56.10 Execução de um programa em C/C++ . . . 134
56.11 Argumentos passados pela linha de comandos e variáveis de ambiente . . . 135
56.12 Espaço de Endereçamento de um Processo em Linux . . . 135
56.12.1 Process Control Table . . . 136

57 Threads **137**
57.1 Diagrama de Estados de uma thread . . . 139
57.2 Vantagens de Multithreading . . . 139
57.3 Estrutura de um programa multithreaded . . . 140
57.4 Implementação de Multithreading . . . 140
57.4.1 Libraria pthread . . . 141
57.5 Threads em Linux . . . 142

58 Process Switching **143**
58.1 Exception Handling . . . 145
58.2 Processing a process switching . . . 146

59 Processor Scheduling **146**
59.1 Scheduler . . . 147
59.1.1 Long-Term Scheduling . . . 147
59.1.2 Medium Term Scheduling . . . 147
59.1.3 Short-Term Scheduling . . . 148
59.2 Critérios de Scheduling . . . 148
59.2.1 User oriented . . . 148
59.2.2 System oriented . . . 149
59.3 Preemption & Non-Preemption . . . 149
59.4 Scheduling . . . 150
59.4.1 Favouring Fearness . . . 150
59.4.2 Priorities . . . 151
Prioridades Estáticas . . . 151
Prioridades Dinâmicas . . . 152
Shortest job first (SJF) / Shortest process next (SPN) . . . 152
59.5 Scheduling Policies . . . 153
59.5.1 First Come, First Serve (FCFS) . . . 153
59.5.2 Round-Robin . . . 154
59.5.3 Shortest Process Next (SPN) ou Shortest Job First (SJF) . . . 154
59.5.4 Linux . . . 155
Algoritmo Tradicional . . . 155
59.6 Novo Algoritmo . . . 156

60 Introdução à Gestão de Memória **157**
60.1 Porquê a gestão de memória . . . 157
60.2 Hierarquia da memória . . . 158
60.2.1 Memória Cache . . . 159
60.2.2 Memória Secundária . . . 160
60.2.3 Princípio da Localidade da Referência . . . 160
60.3 Gestão da memória num ambiente multiprogramado . . . 160
60.4 Espaço de Endereçamento . . . 161
60.4.1 Exemplo . . . 164
60.4.2 Espaço de endereçamento lógico vs físico . . . 166

61 Arquitecturas de Memória Particionadas **166**
61.1 Arquitectura de partições fixas . . . 166
61.1.1 Vantagens e Desvantagens . . . 168

61.2 Arquitectura de posições variáveis . . . 169
61.2.1 Gestão do espaço . . . 169
61.2.2 Exemplo . . . 170
61.2.3 Políticas de Escalonamento . . . 172
61.2.4 Vantagens vs Desvantagens . . . 173

62 Organização da memória real 174
62.1 Tradução de um endereço lógico num endereço físico . . . 175
62.2 Memória real e o ciclo de vida de um processo . . . 175
62.2.1 Criação de um processo . . . 176
62.2.2 Ciclo de Vida do processo . . . 176
62.2.3 Fim de Vida do processo . . . 176

63 Organização da memória virtual 176
63.1 Tradução de um endereço lógico num endereço físico . . . 179
63.1.1 Acesso à memória . . . 179
63.2 Ciclo de vida de um processo . . . 181
63.2.1 Criação de um processo . . . 181
63.2.2 Ao longo da execução . . . 182
63.2.3 Término de um processo . . . 182
63.3 Exceção por falta de bloco . . . 183
63.3.1 Sequência de instruções . . . 183
63.4 Acesso à memória . . . 186
63.5 Vantagens e Desvantagens . . . 187
63.5.1 Vantagens . . . 187
63.5.2 Desvantagens . . . 187

64 Arquitectura Segmentada 188
64.1 Tipos de Segmentos: . . . 189
64.2 Tradução de um endereço lógico num endereço físico . . . 190
64.3 Conclusão . . . 190

65 Arquitectura Segmentada/Paginada 190
65.1 Tradução de um endereço lógico num endereço físico numa arquitectura segmento-paginada . . . 192
65.2 Vantagens vs Desvantagens . . . 193
65.2.1 Vantagens . . . 193
65.2.2 Desvantagens . . . 193

66 Políticas de Substituição de páginas em memória 195
66.1 Algoritmo LRU - Least Recently Used . . . 197
66.1.1 Algoritmo NRU - Not Recently used . . . 197
66.2 Algoritmo FIFO - First In, First Out . . . 198
66.3 Algoritmo da Segunda Oportunidade . . . 198
66.4 Algoritmo do relógio . . . 198

67 Working set 199

68 Demand paging vs prepaging 200
68.1 Substituição global vs substituição local . . . 200

List of Tables

3 Estrutura de um sistema operativo por camadas - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos* . . . 23
8 Banker's Algorithm Example . . . 120
10 Comparação entre os diferentes tipos de memórias de um sistema computacional . . . 159
11 Distribuição da ocupação da memória . . . 171

List of Figures

1 Esquema típico de um sistema de computação . . . 13
2 Diagrama em camadas de um sistema de operação . . . 14
3 Visão de um sistema operativo do tipo Extended Machine . . . 15
4 Visão de um sistema operativo do tipo Extended Machine . . . 16
5 Multiprogrammed batch . . . 18
6 Interactive system (Time-Sharing) . . . 18
7 Real Time System . . . 19
8 Networking Operating System . . . 19
9 Exemplo de multiplexing temporal: Os programas A e B estão a ser executados de forma concorrente num sistema single processor . . . 21
10 Diagrama de um `kernel` monolítico - imagem retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos* . . . 22
11 Estrutura de um sistema operativo que usa microkernel - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos* . . . 24
12 Estrutura de uma virtual machine - Imagem retirada da Wikipedia . . . 25
13 Estrutura Interna do Unix - Tradicional . . . 26
14 Estrutura Global do Sistema Linux - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos* . . . 27
15 Estrutura do Kernel do Linux - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos* 28
16 Estrutura Interna do Kernel do Windows - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos* . . . 29
17 FUSE diagram with sofs17 . . . 42
18 Organização de um disco formatado em sofs17 . . . 43
19 Caches and Reference bitmap blocks . . . 45
20 Code Strucuture to be developed . . . 49
21 Diagrama da estrutura interna de um Monitor de Hoare . . . 105
22 Diagrama da estrutura interna de um Monitor de Brinch Hansen . . . 106
23 Diagrama da estrutura interna de um Monitor de Lampson/Redell . . . 107
24 Ciclo de Vida de um filósofo . . . 115
25 Negar *hold and wait* . . . 117
26 Negar a condição de *no preemption* dos recursos . . . 118
27 Negar a condição de espera circular no acesso aos recursos . . . 119
28 Algoritmo dos banqueiros aplicado ao Jantar dos filósofos . . . 121
29 Arquitectura típica de um computador . . . 123
30 Exemplo de execução num ambiente multiprogramado . . . 124
31 Diagrama de Estados do Processador - Básico . . . 125
32 Diagrama de Estados do Processador - Com Memória de Swap . . . 127

33 Diagrama de Estados do Processador - Com Processos Temporalmente Finitos . . . 128
34 Diagrama de Estados do Processador - Com Memória de Swap . . . 129
35 Diagrama do Login em Linux . . . 130
36 Criação de Processos . . . 131
37 Execução de um programa em C/C++ . . . 134
38 Espaço de endereçamento de um processo em Linux . . . 135
39 Process Control Table . . . 137
40 Single threading vs Multithreading . . . 138
41 Diagrama de estados de uma thread . . . 139
42 Exemplo do uso da libraria pthread . . . 141
43 Diagrama de estados completo para um processador `multithreading` . . . 144
44 Algoritmo a seguir para tratar de exceções normais . . . 145
45 Algoritmo a seguir para efetuar uma process switching . . . 146
46 Identificação dos diferentes tipos de schedulers no diagrama de estados dos processos . . . 147
47 Espaço de endereçamento de um processo em Linux . . . 150
48 Espaço de endereçamento de um processo em Linux . . . 151
49 Problema de Scheduling . . . 153
50 Política FCFS . . . 153
51 Política Round-Robin . . . 154
52 Relembrando o diagrama de um sistema Computacional . . . 157
53 Grau de ocupação do processador em função do número de processos concorrentes residentes em memória principal em simultâneo . . . 158
54 Hierarquia da Memória num sistema de computação . . . 159
55 Diagrama da inclusão da gestão de memória com o scheduling de baixo nível do processador . . . 161
56 Construção do espaço de endereçamento de um programa após compilação e linkagem . . . 161
57 Diagrama da divisão do espaço de endereçamento de um programa . . . 163
58 Divisão em partições fixas mutuamente exclusivas com diferentes tamanhos . . . 167
59 Divisão da memória em partições de tamanho variável . . . 169
60 Diagrama da Memória Particionada . . . 171
61 Espaço de endereçamento real de um processo . . . 174
62 Tradução de um endereço lógico num endereço físico . . . 175
63 Espaço de endereçamento completamente em memória virtual . . . 177
64 Espaço de endereçamento apenas parcialmente em memória virtual . . . 178
65 Diagrama de blocos da decomposição de um endereço lógico num endereço físico . . . 179
66 Diagrama de eventos e estrutura de uma organização em memória virtual . . . 181
67 Estrutura de uma organização de memória em arquitectura paginada . . . 184
68 Exemplo da ocupação da memória principal e swap num arquitectura paginada . . . 185
69 Diagrama de blocos para efetuar o acesso . . . 186
70 Exemplo de memória segmentada . . . 189
71 Diagrama de blocos da operação de tradução de um endereço lógico num endereço físico . . . 190
72 Estrutura de uma arquitectura segmento-paginada . . . 191
73 Tradução de um endereço lógico num endereço físico numa arquitectura segmento-paginada . . . 192
74 Conteúdo de cada entrada da tabela de segmentação . . . 193
75 Conteúdo de cada entrada da tabela de paginação de cada segmento . . . 193
76 Divisão da memória em frames . . . 195
77 Exemplos do estado das listas biligadas . . . 196
78 Algoritmo do Relógio . . . 199

1 Sistemas de Computação

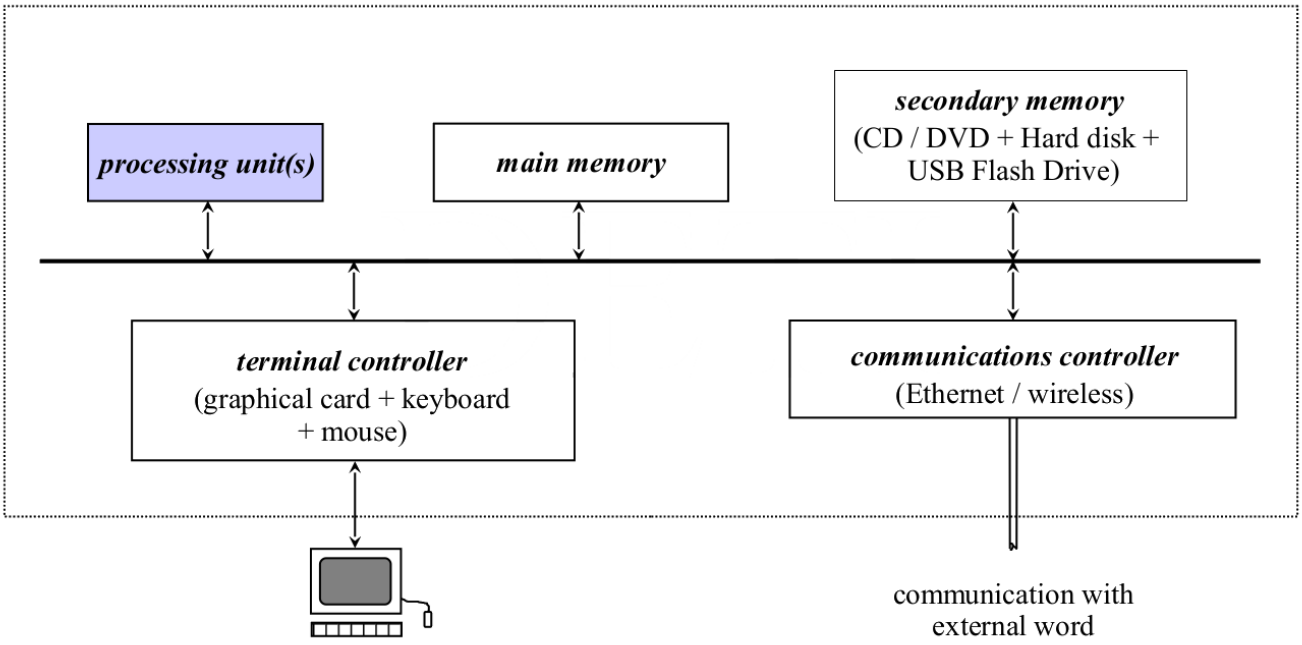


Figure 1: Esquema típico de um sistema de computação

1.1 Vista simplificada de um sistema de computação

Um sistema operativo é um sistema/programa base que é executado pelo sistema computacional.

- Controla diretamente o `hardware`
- Providencia uma **camada de abstração** para que os restantes **programas** possam **interagir com o `hardware`** de forma indireta

Podem ser classificados em dois tipos:

1. **gráficos:**
    - utilizam um contexto de **janelas** num ambiente gráfico
    - os elementos principais de interação são os **ícones** e os **menus**
    - a principal ferramente de `input` da interação humana é o rato
2. **textuais *(shell)*:**
    - baseado em comandos introduzidos através do teclado
    - uma linguagem de scripting/comandos[1]

Os dois tipos não são mutualmente exclusivas.

- Windows: sistema operativo gráfico que pode lançar uma aplicação para ambiente textual
- Linux: sistema operativo textual que pode lançar ambiente gráfico

[1] ficheiro em código fonte de compilação separada

1.2 Vista geral

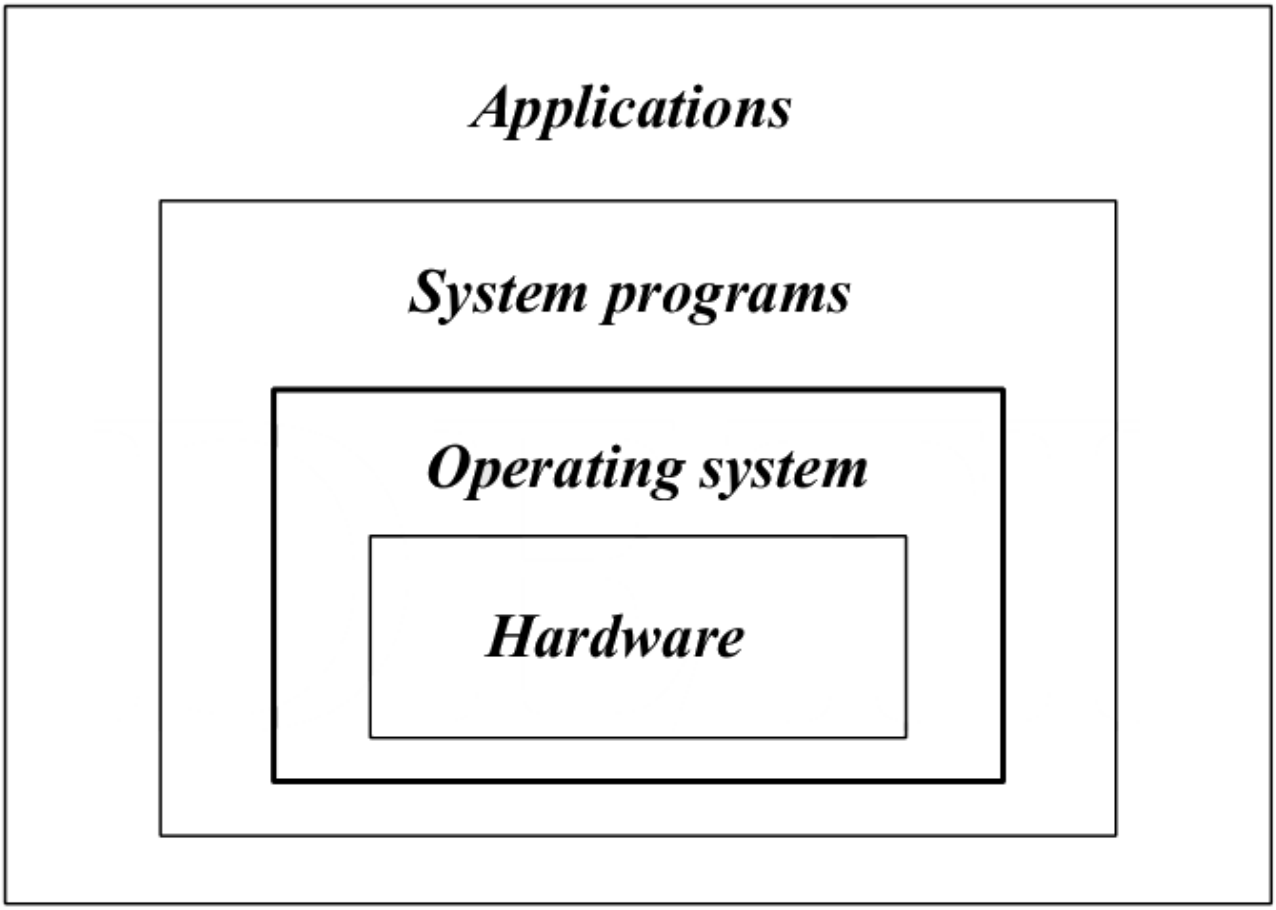


Figure 2: Diagrama em camadas de um sistema de operação

Os sistemas de operação podem ser vistos segundo duas perspectivas:

1. `Extended Machines`
2. `Resource Manager`

1.2.1 Extended Machine

O sistema operativo fornece **níveis de abstração** (APIs) para que os programas possam aceder a partes físicas do sistema, criando uma **"máquina virtual":**

- Os programas e programadores têm uma visão virtual do computador, um **modelo funcional**
  - Liberta os programadores de serem obrigados a saber os detalhes do hardware
- Acesso a componentes do sistema mediado através de `system calls`
  - Executa o core da sua função em root (com permissões de super user)
  - Existem funções que só podem correr em super user
  - Todas as chamadas ao sistema são interrupções

  - Interface uniforme com o `hardware`
  - Permite as aplicações serem **portáteis** entre sistemas de computação **estruturalmente diferentes**

- O sistema operativo controla o **espaço de endereçamento físico** criando uma camada de abstração **(memória virtual)**

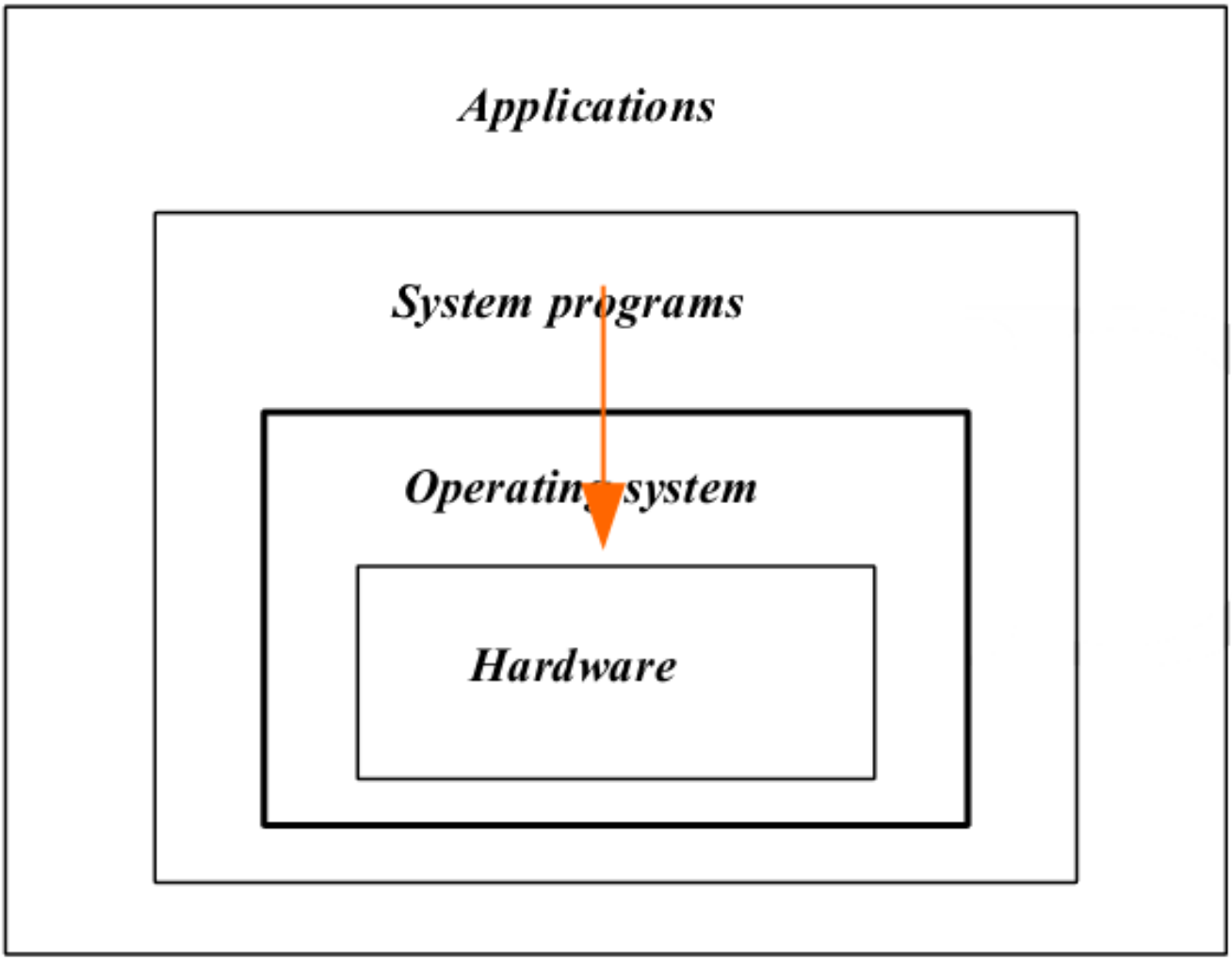


Figure 3: Visão de um sistema operativo do tipo Extended Machine

Tipos de funções da extended machine

- Criar um ambiente interativo que sirva de interface máquina-utilizador
- Disponibilizar mecanismos para desenvolver, testar e validar programas
- Disponibilizar mecanismos que controlem e monitorizem a execução de programas, incluindo a sua intercomunicação e e sincronização
- Isolar os espaços de endereçamento de cada programa e gerir o espaço de cada um deles tendo em conta as limitações físicas da memória principal do sistema
- Organizar a memória secundária [2] em sistema de ficheiros
- Definir um modelo geral de acesso aos dispositivos de I/O, independentemente das suas características individuais

[2] De um para um

- Detetar situações de erros e estabelecer protocolos para lidar com essas situações

1.2.2 Resource Manager

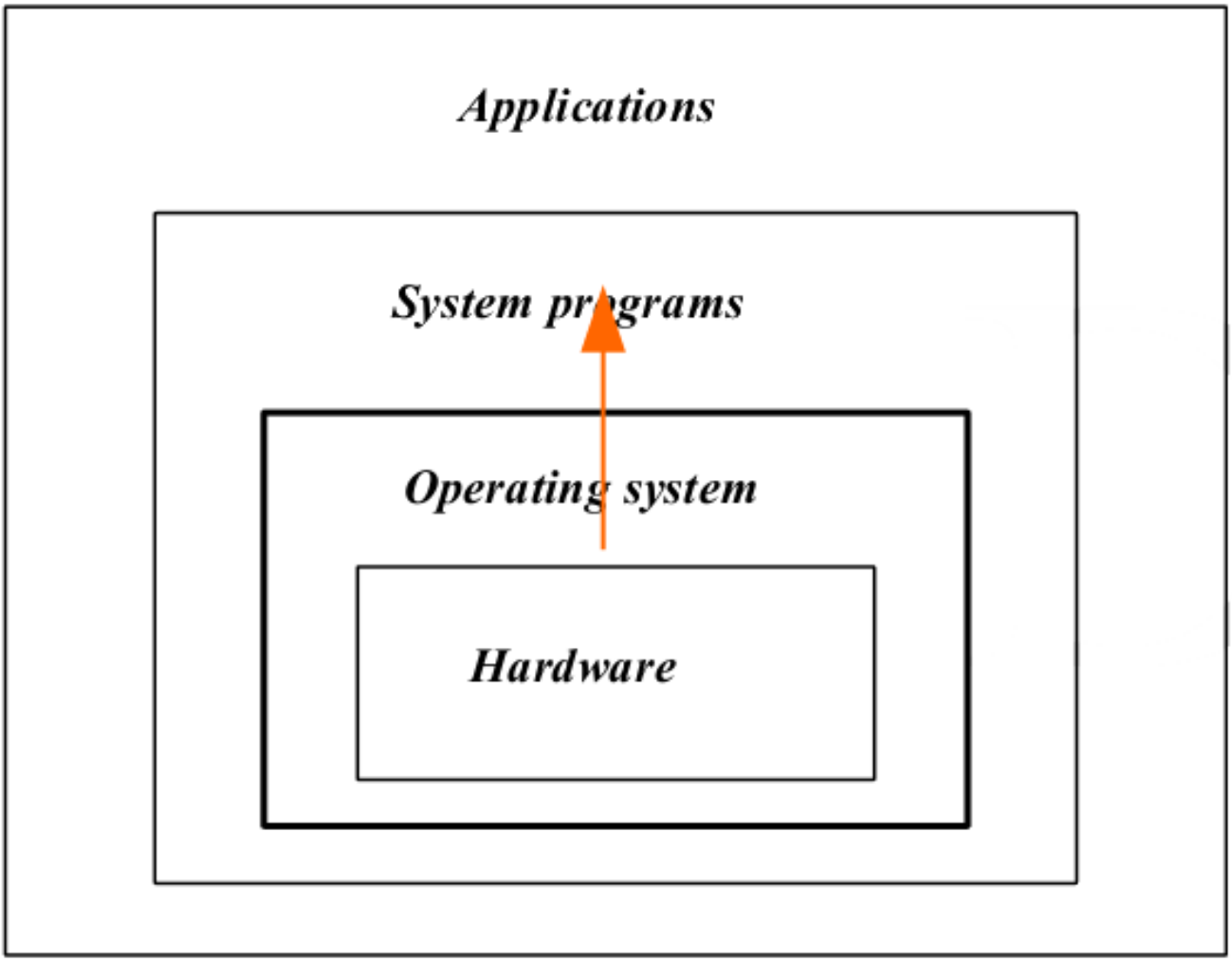


Figure 4: Visão de um sistema operativo do tipo Extended Machine

Sistema computacional composto por um conjunto de recursos:

- processador(es)
- memória
  - principal
  - secundária
- dispositivos de I/O e respetivos controladores

O sistema operativo é visto como um programa que gere todos estes recursos, efetuando uma gestão controlada e ordenada dos recursos pelos diferentes programas que tentam aceder a estes. O seu objetivo é **maximizar a performance** do sistema, tentando garantir a maior eficiência no uso dos recursos, que são **multiplexados no tempo e no espaço**.

1.3 Evolução dos Sistemas Operativos

Primórdios : Sistema Electromecânico

- 1ª Geração: 1945 - 1955
  - Vacuum tubes
  - electromechanical relays -No operating system -programed in system
  - Program has full control of the machine
  - Cartões perfurada (ENIAC)
- 2ª geração: Transistores individuais
- 4ª Geração (1980 - presente)

Technology	Notes
LSI/VLSI	Standard Operation systems (MS-DOS, Macintosh, Windows, Unix)
personal computers (microcomputers)	Network operation systems
network	

- 5ª Geração (1990 - presente)

Technology	Notes
Broadband, wireless	mobile operation systems (Symbian, iOS, Android)
system on chip	cloud computing
smartphone	ubiquitous computing

2 Taxonomia de Sistemas Operativos

2.1 Classificação com base no tipo de processamento

- Processamento em série
- Batch Processing
  - Single
  - Multiprogrammed batch
- Time-sharing System
- Real-time system
- Network system
- Distributed System

2.1.1 Multiprogrammed batch

- **Propósito:** Otimizar a utilização do processador

- **Método de Otimização:** Enquanto um programa está à espera pela conclusão de uma operação de I/O, outro programa usa o processador

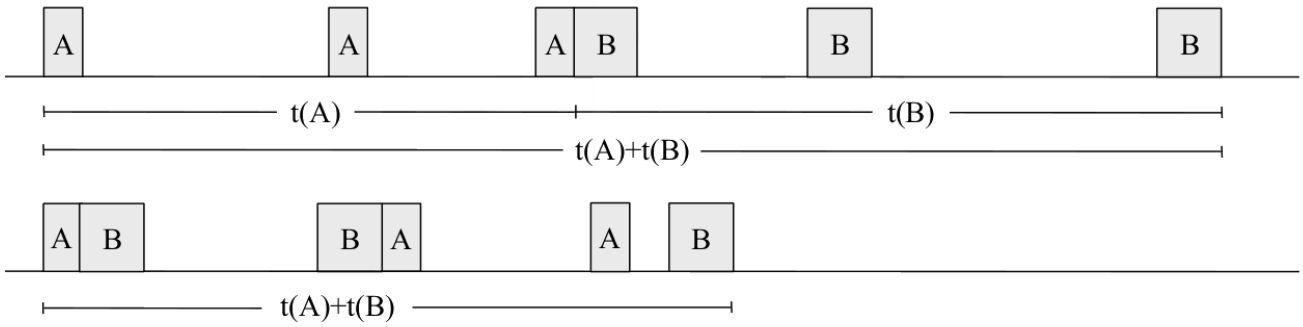


Figure 5: Multiprogrammed batch

2.1.2 Interactive System (Time-Sharing)

- **Propósito:**
  - Proporcionar uma interface `user-friendly`
  - Minimizar o tempo de resposta a pedidos externos
- **Método:**
  - Vários utilizadores mas cada um no seu terminal
  - Todos os terminais têm comunicação direta e em simultâneo com o sistema
  - Usando multiprogramação, o uso do processador é multiplexado no tempo, sendo atribuído um time-quantum a cada utilizador
  - No `macrotempo` é criada a ilusão ao utilizador que possui o sistema só para si

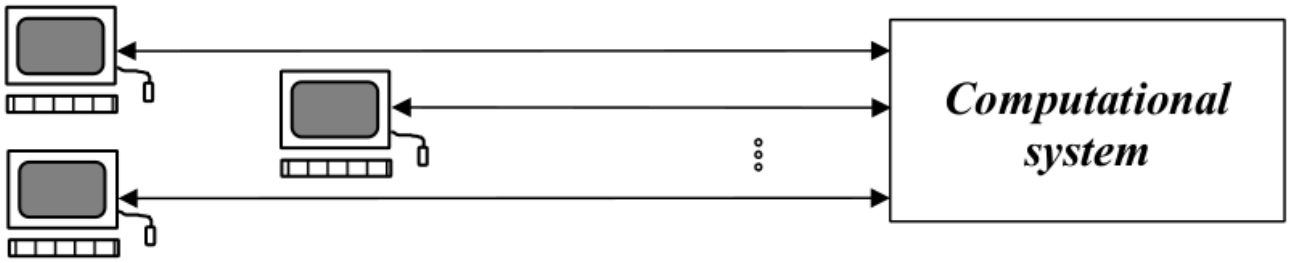


Figure 6: Interactive system (Time-Sharing)

2.1.3 Real Time System

- **Propósito:** Monitorizar e (re)agir processo físicos
- **Método:** Variante do Sistema Interativo que permite import limites máximos aos tempos de resposta para diferentes classes de eventos externos

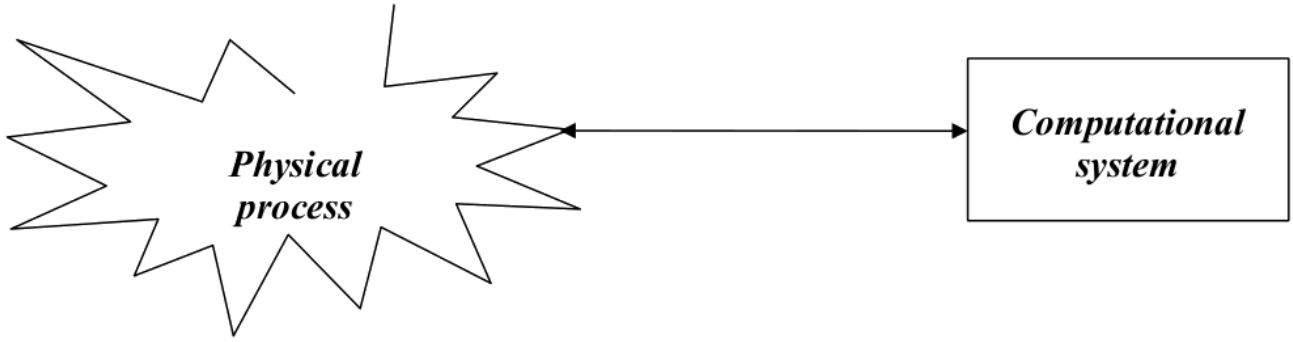


Figure 7: Real Time System

2.1.4 Network Operating System

- **Propósito:** Obter vantagem com as interconexões de `hardware` existentes de sistemas computacionais para estabelecer um conjunto de serviços comuns a uma comunidade.

A máquina é mantêm a sua individualidade mas está dotada de um conjunto de primitivas que permite a comunicação com outras máquinas da mesma rede:

- partilha de ficheiros (ftp)
- acesso a sistemas de ficheiros remotos (NFS)
- Partilha de recursos (e.g. impressoras)
- Acesso a sistemas computacionais remotos:
  - telnet
  - remote login
  - ssh
- servidores de email
- Acesso à internet e/ou Intranet

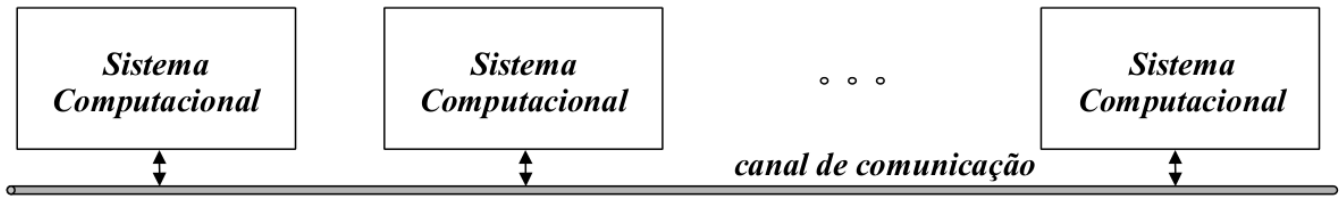


Figure 8: Networking Operating System

2.1.5 Distributed Operating System

- **Propósito:** Criar uma rede de computadores para explorar as vantagens de usar **sistemas multiprocessador**, estabelecendo uma cada de abstração onde o utilizador vê a computação paralela distribuída por todos os computadores da rede como uma única entidade
- **Metodologia:** Tem de garantir uma completa **transparência ao utilizador** no acesso ao processador e outros recursos partilhados (e.g. memória, dados) e permitir:

  - distribuição da carga de `jobs` (programas a executar) de forma dinâmica e estática
  - automaticamente aumentar a sua capacidade de processamento de forma dinâmica se
    * um novo computador se ligar à rede
    * forem incorporados novos processadores/computadores na rede
  - a paralelização de operações
  - implementação de mecanismos tolerantes a falhas

2.2 Classificação com base no propósito

- Mainframe
- Servidor
- Multiprocessador
- Computador Pessoal
- Real time
- Handheld
- Sistemas Embutidos
- Nós de sensores
- Smart Card

3 Multiprocessing vs Multiprogramming

3.1 Paralelismo

- Habilidade de um computador **executar simultaneamente** um ou mais programas
- Necessita de possuir uma estrutura multicore
  - Ou processadores com mais que um core
  - Ou múltiplos processadores por máquina
  - Ou uma estrutura distribuída
  - Ou uma combinação das anteriores

Se um sistema suporta este tipo de arquitectura, suporta **multiprocessamento**

O **multiprocessamento** pode ser feito com diferentes arquitecturas:

- **SMTP** - symmetric processing (SMP)
  - Computadores de uso pessoal
  - Vários processadores
  - A memória principal é partilhada por todos os processadores
  - Cada core possui cache própria
  - Tem de existir **mecanismos de exclusão mútua** para o hardware de suporte ao multiprocessamento
  - Cada processador vê toda a memória (como memória virtual) apesar de ter o acesso limitado
- **Planar Mesh**
  - Cada processador liga a 4 memória adjacentes

3.2 Concorrência

- Ilusão criada por um sistema computacional de "aparentemente" ser capaz de executar mais programas em simultâneo do que o seu número de processadores
- Os processador(es) devem ser atribuídos a diferentes programas de forma multiplexada no tempo

Se um sistema suporta este tipo de arquitectura suporta **multiprogramação**

Figure 9: Exemplo de multiplexing temporal: Os programas A e B estão a ser executados de forma concorrente num sistema single processor

4 Estrutura Interna de um Sistema Operativo

Um sistema operativo deve:

- Implementar um ambiente gráfico para interagir com o utilizador
- Permitir mais do que um utilizador
  - Tanto simultânea como separadamente
- Ter capacidade de ser `multitasking`, i.e., executar vários programas ao mesmo tempo
- Implementar memória virtual
- Permitir o acesso, de forma transparente ao utilizador, a:
  - sistemas de ficheiros locais e/ou remotos (i.e., na rede)
  - dispositivos de I/O, independentemente da sua funcionalidade
- Permitir a ligação da máquina por rede a outras máquinas
- Conter um bom conjunto de `device drivers`
- Permitir a ligação de dispositivos `plug and play` [3]

4.1 Design de um sistema operativo

Por estas razões, um sistema operativo é **complexo**, com milhões de linhas de código. O design e implementação do seu `kernel`pode seguir as seguintes filosofias:

- Monolithic
- Layered (por camada)
- Microkernels
- Client-Server Model
- Virtual Machines
- Exokernels

[3] ficheiro em código fonte de compilação separada

4.1.1 Monolithic system

- A perspectiva mais utilizada
- Só existe um **único programa** a ser executado em `kernel mode`
- Um **único entry point**
  - Todos os pedidos ao sistema são feitos usando este `entry-point`
- Comunicação com o sistema através de `syscall` [4]
  - Implementadas por um conjunto de rotinas
  - Existe ainda outro conjunto de funções auxiliares para a system call
- Qualquer parte do sistema (aka `kernel`) pode "ver" qualquer outra parte do sistema
  - **Vantagem:** eficiência no acesso a informação e dados
  - **Desvantagem:** Sistema difícil de testar e modificar

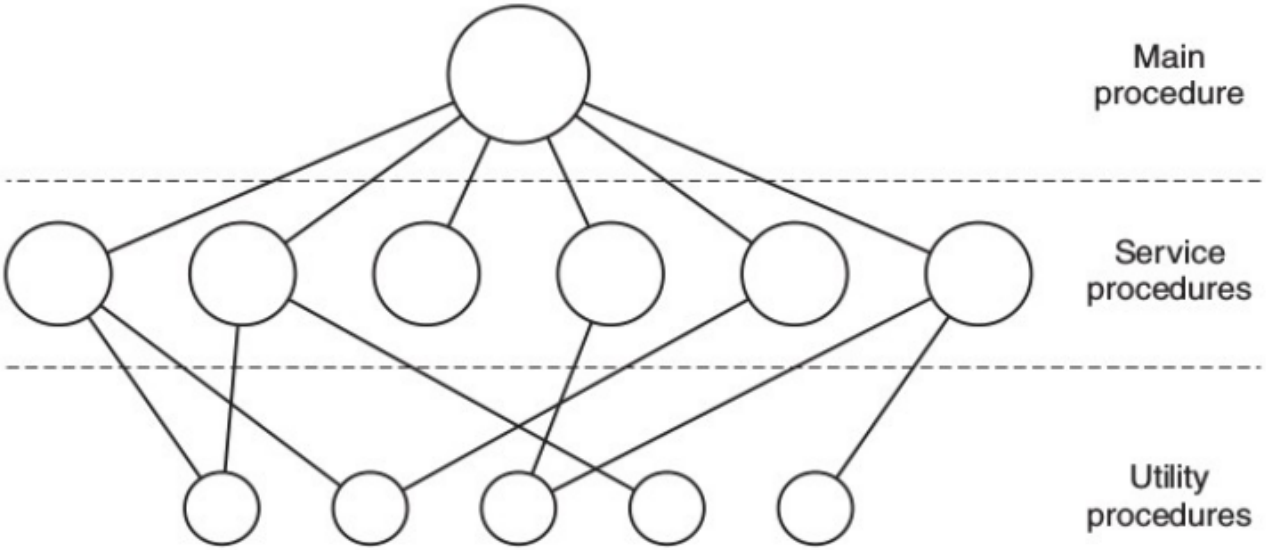


Figure 10: Diagrama de um `kernel` monolítico - imagem retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos*

4.1.2 Layered Approach: Divisão por camadas

- Perspetiva modular
  - O sistema operativo é constituído por um conjunto de camadas, com diferentes níveis hierárquicos
- A interação **só é possível entre camadas adjacentes**
  - Uma função da camada mais superior não pode chamar uma função da camada mais abaixo
  - Tem de chamar uma função da camada imediatamente abaixo que irá tratar de chamar funções das camadas mais abaixo (estilo `sofs`)
- Não é simples de projetar
  - É preciso definir claramente que funcionalidades em que camada, o que pode ser difícil de decidir

[4] De um para um

- **Fácil de testar e modificar**, mas uma **grande perda de eficiência**
  - A eficiência pode piorar se a divisão de funções não for bem feita
  - Existe um `overhead` adicional causado pelo chamada de funções entre as várias camadas
- Facilita a divisão de funções entre o modo de utilizador e o modo de `kernel`

Table 3: Estrutura de um sistema operativo por camadas - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos*

Layer	Function
5	Operador
4	Programas do Utilizador
3	Gestão de dispositivos de I/O
2	Comunicação Operator- Process
1	Memory and drum management
0	Alocação do processador e gestão do ambiente multiprogramado

4.1.3 Microkernel

- Posso ter **modularidade** sem ser obrigado a usar camadas em níveis hierárquicos diferentes
- Defino um conjunto de módulos de "pequena dimensão", com funcionalidades bem definidas
  - apenas o `microkernel` é executado em `kernel space`, com permissões de `root`
  - todos os outros módulos são executados em `user space` e comunicam entre si usando os mecanismos de comunicação providenciados pelo `microkernel`
  - Os módulos que são executados em `user space` podem ser lançados no startup ou dinamicamente à medida que são precisos (dispositivos `plug-and-play` [5])
- O `microkernel` é responsável por:
  - Gestão de Processos
  - Gestão da Memória
  - Implementar sistemas simples de comunicação interprocess
  - Escalonamento do Processador (Processor Scheduling)
  - Tratar das interrupções
- Sistema robusto
  - Manipulação de um filesystem é feita em `user space`. Se houver problemas a integridade do sistema físico não é afetada

[5] ficheiro em código fonte de compilação separada

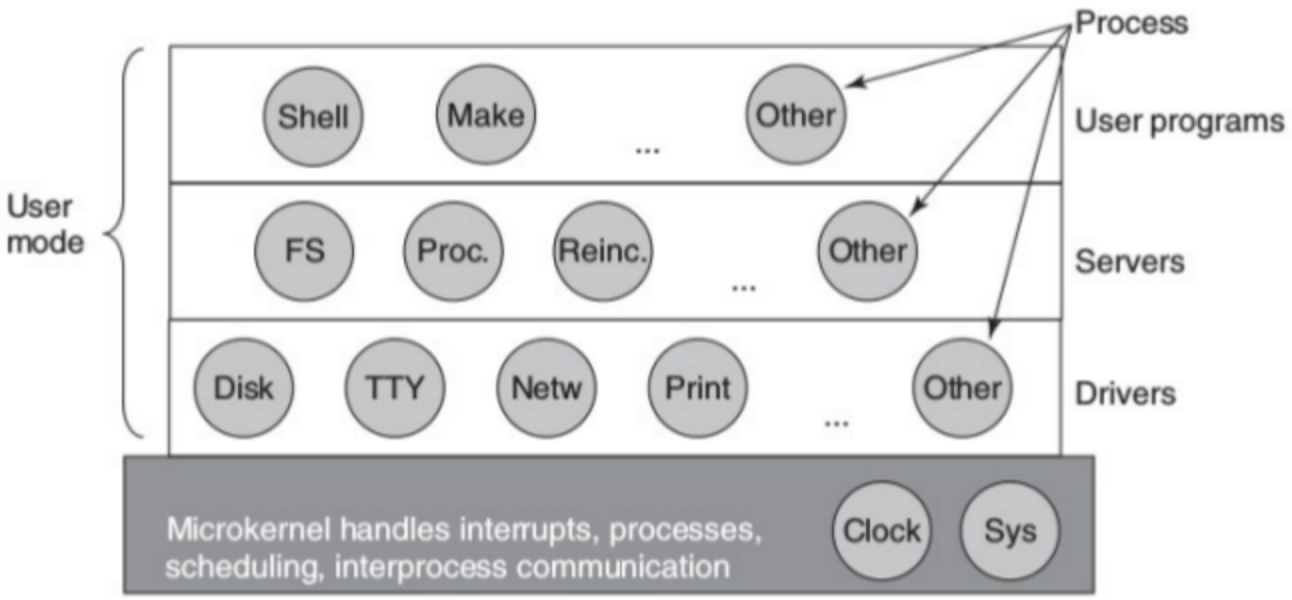


Figure 11: Estrutura de um sistema operativo que usa microkernel - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos*

4.1.4 Virtual machine (hypervisors)

- Criam plataformas virtuais onde podem ser instalados `guess OSs`
- Existem dois tipos de hypervisors
  - Type-1 (**`native`** `hypervisor`): executa o `guest OS` **diretamente** no `hardware` da máquina `host` (máquina física onde a máquina virtual vai ser executada). Exemplos:
    * z/VM
    * Xen
    * Hyper-V
    * VMware ESX
  - Type-2 (`hosted supervisor`): executa o `guest OS` **indiretamente** no `hardware` da máquina, sendo a máquina virtual executada "em cima" do sistema operativo do `host`. Exemplos:
    * VirtualBox
    * VMware Workstation
    * Parallels
- Existem exemplos de `hypervisors` híbridos, que tanto podem ser executar o `guest OS` indiretamente (por cima do sistema operativo) ou diretamente no `hardware` da máquina:
  - KVM
  - bhyve

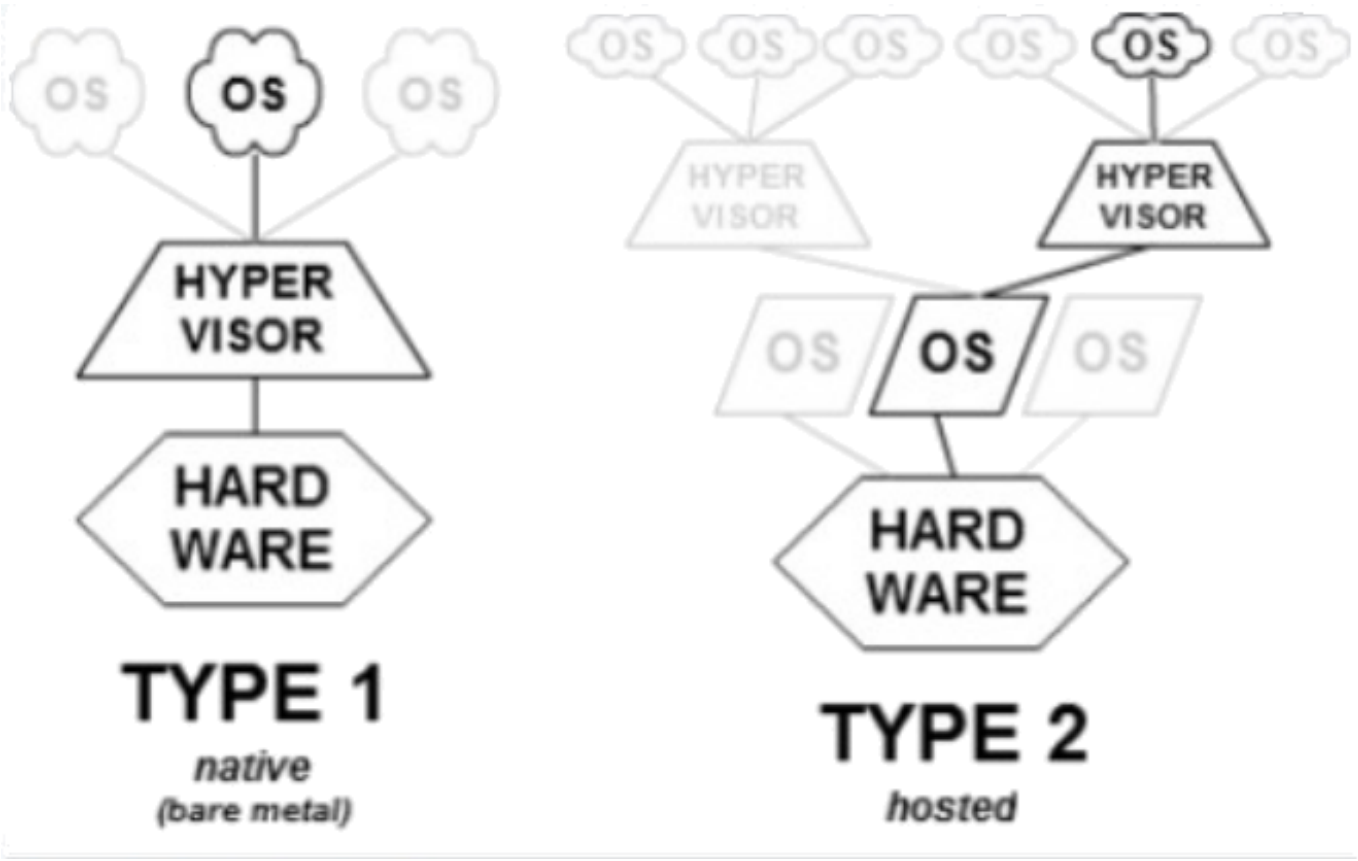


Figure 12: Estrutura de uma virtual machine - Imagem retirada da Wikipedia

4.1.5 Client-Server

- Implementação modular, baseada na relação cliente-servidor
  - A comunicação é feita através de **pedidos e respostas**
  - Para isso é usada `message-passing`
- Pode estar presente um `microkernel` que manipula operações de baixo nível
- Pode ser generalizado e usado em sistemas `multimachine`

4.1.6 Exokernels

- Usa um `kernel` com funcionalidades reduzidas
  - Apenas providencia algumas abstrações de `hardware`
- Segue a filosofia de *"Em vez de clonar a máquina virtual, divido-a"*
  - Os recursos são **divididos em partições**, em vez de clonados
  - Os recursos são alocados às `virtual machines` e a sua utilização é controlada pelo `microkernel`
- Permite a implementação de **camadas de abstração personalizadas** consoante as necessidades
- **Eficiente:** Poupa uma camada destinada a efetuar o mapeamento

4.2 Estruturas Internas do Unix/Linux e Windows

4.2.1 Estrutura Interna do Unix (tradicional)

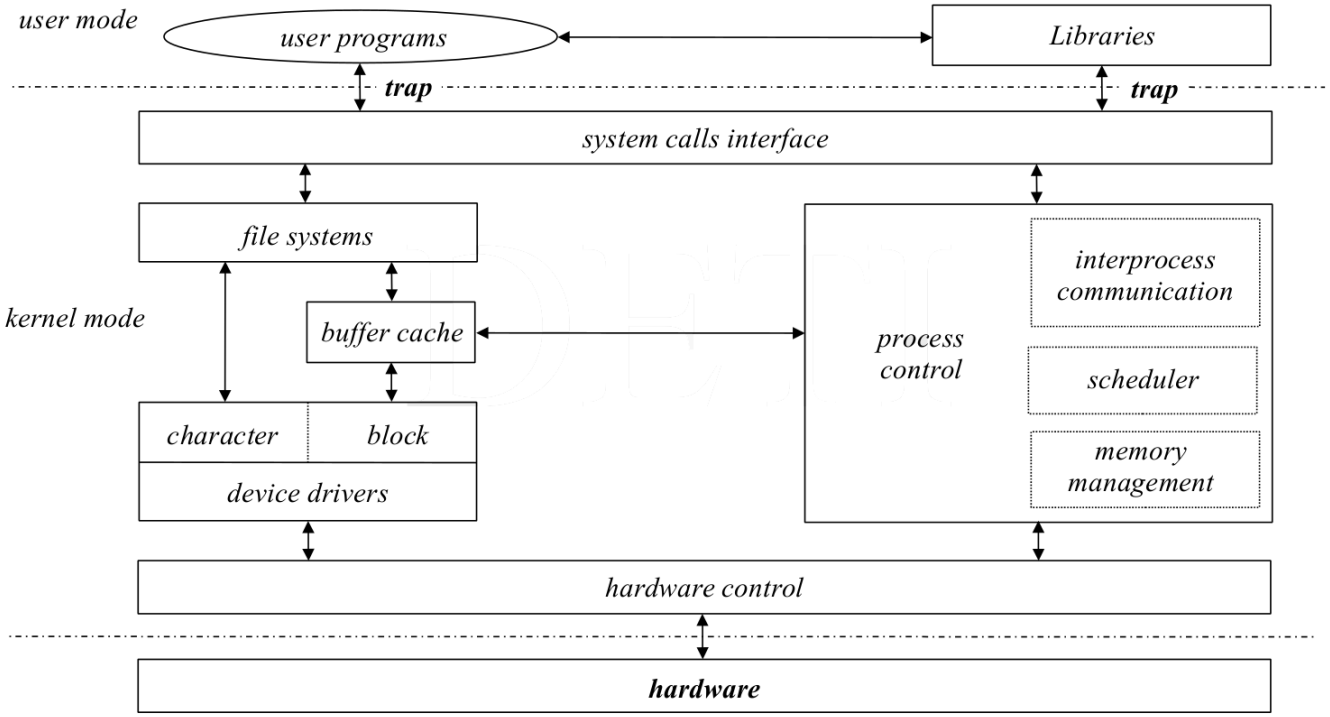


Figure 13: Estrutura Interna do Unix - Tradicional

Legenda:

- `trap`: interrupção por software (única instrução que muda o modo de execução)
- `buffercache`: espaço do disco onde são mantidos todos os ficheiros em cache (aka abertos)
  - **desmontar uma pen:** forçar a escrita da buffer cache para a pen

Unix considera tudo como sendo ficheiros: - ou blocos (buffer cache) - ou bytes

`open`, `close`, `fork` **não são system calls**. São funções de biblioteca que acedem às `system call` (implementadas no `kernel`). São um interface amigável para o utilizador ter acesso a estas funcionalidades.

4.2.2 Estrutura Global do Unix

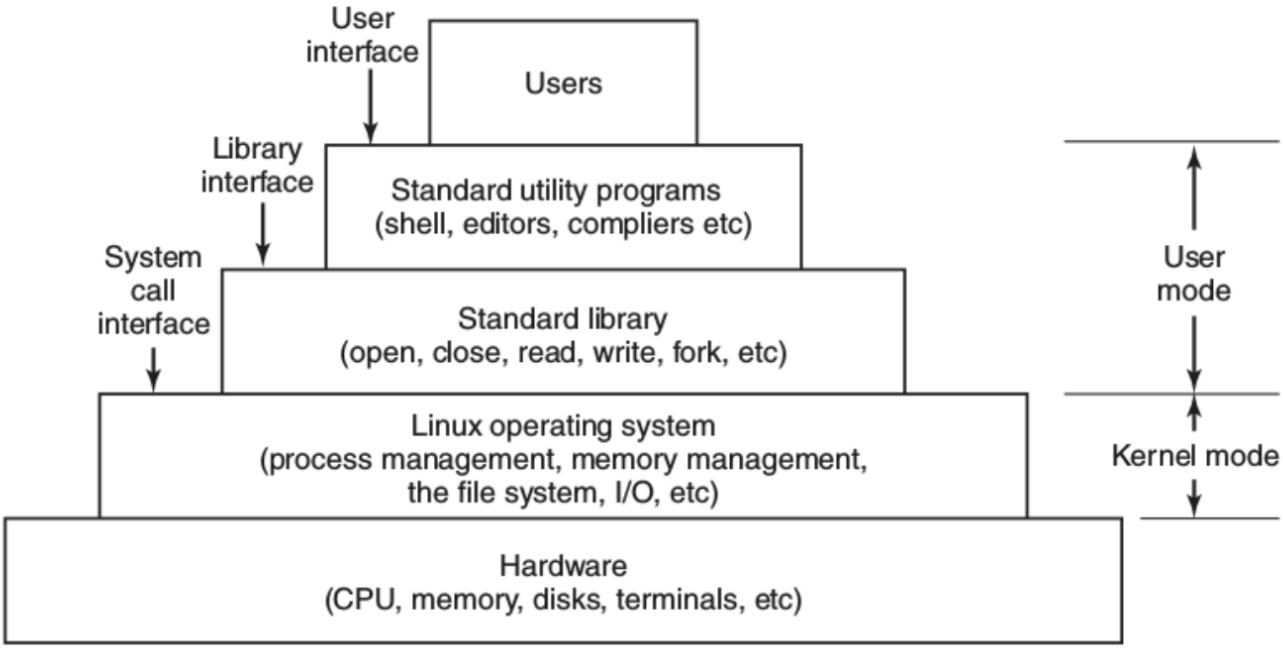


Figure 14: Estrutura Global do Sistema Linux - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos*

4.2.3 Estrutra do Kernel Unix

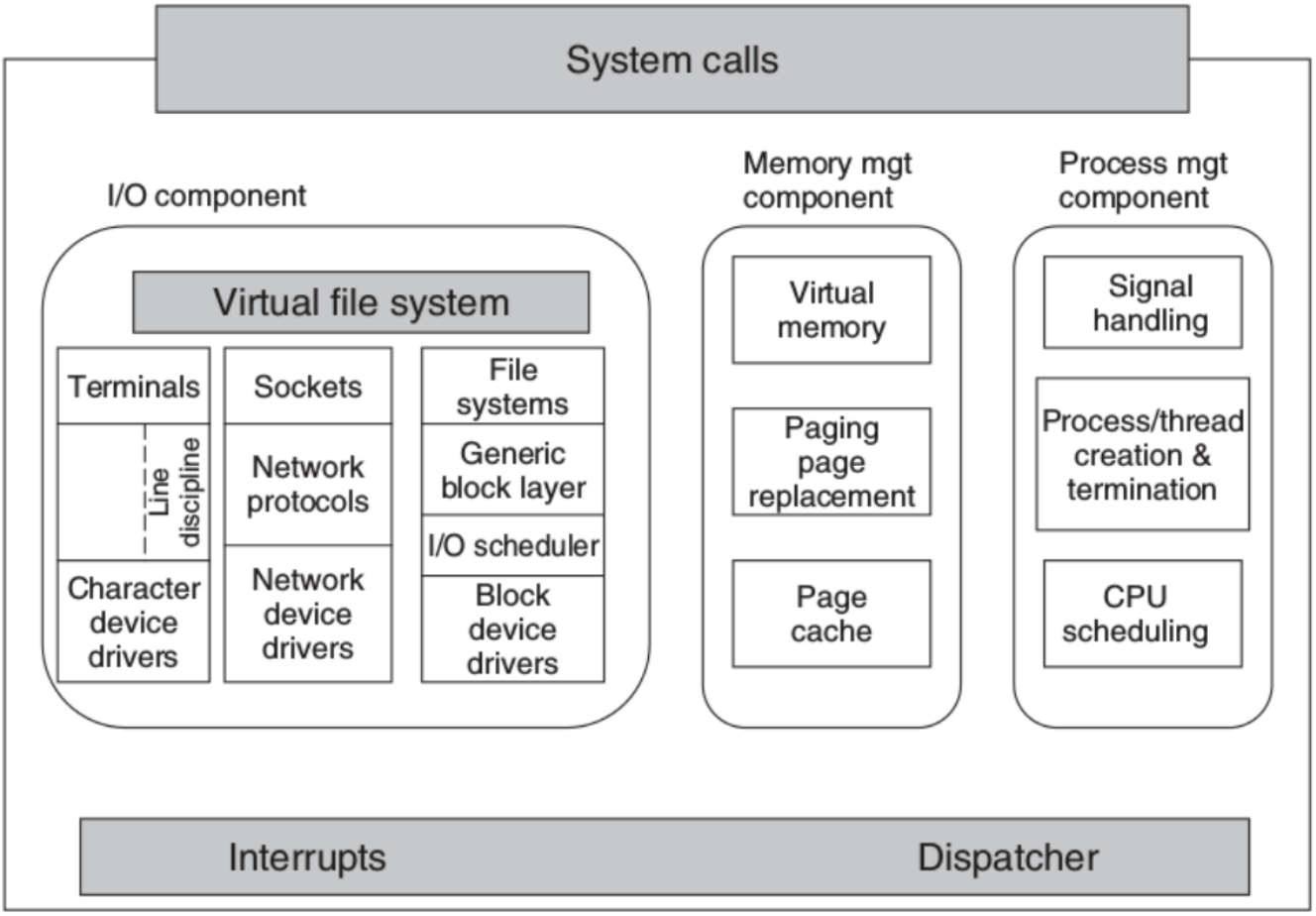


Figure 15: Estrutura do Kernel do Linux - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos*

4.2.4 Estrutura do Kernel Windows

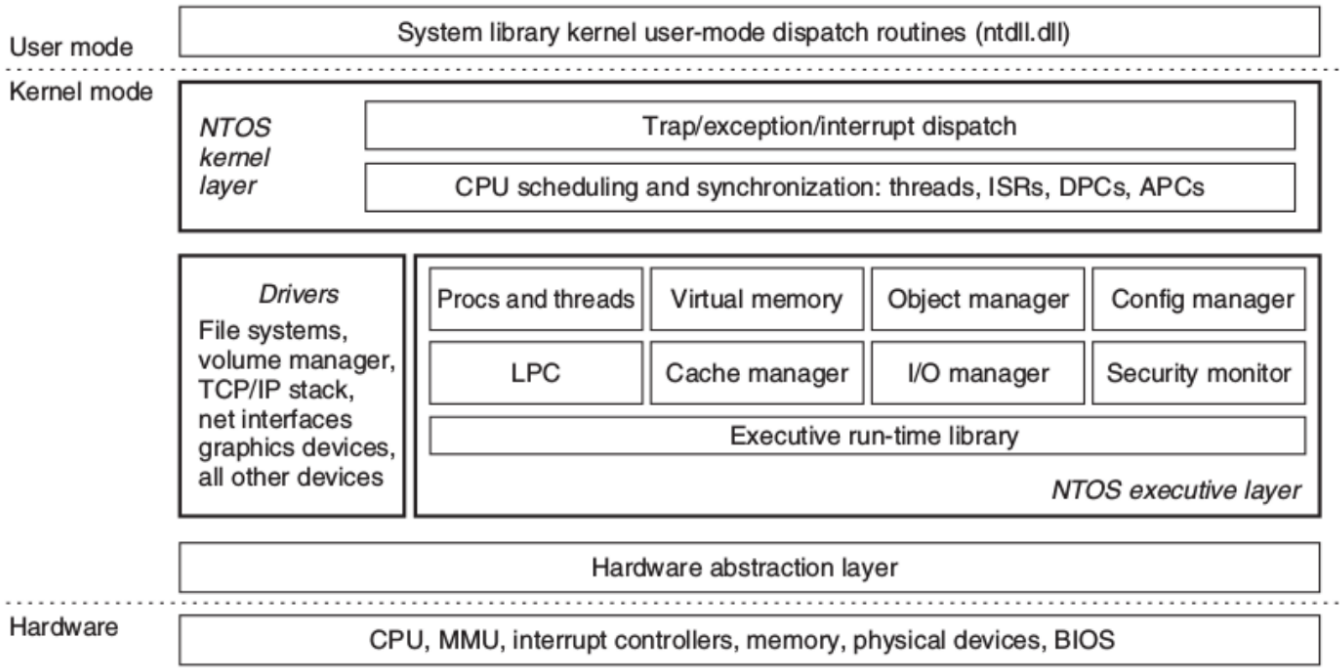


Figure 16: Estrutura Interna do Kernel do Windows - Retirada do livro *Modern Operating Systems, Andrew Tanenbaum & Herbert Bos*

5 Shell Scripting

5.1 Exercise 1 - Command Overview

- **man** : Documentação dos comandos
- **ls** : Listar ficheiros de uma pasta
- **mkdir** : Criar uma pasta
- **pwd** : Caminho absoluto do diretório corrente
- **rm** : Remover ficheiros
- **mv** : Renomear ficheiros ou mover ficheiros/pastas entre pastas
- **cat** : Imprimir um ficheiro para o stdout
- **echo** : Imprimir para o stdout uma mensagem
- **less** : paginar um ficheiro (não mostra o texto literal)
- **head** : mostrar as primeiras 10 linhas de um ficheiro
- **tail** : mostrar as ultimas 10 linhas de m ficheiro
- **cp** : copiar ficheiros
- **diff** : mostrar as diferenças linha a linha entre dois ficheiros
- **wc** : contar linhas, palavras e caracteres de um ficheiro
- **sort** : ordenar ficheiros
- **grep** : pesquisa de padroes em ficheiros
- **sed** : transformacoes de texto
- **tr** : substituir, modificar ou apagar caracteres do stdin e imprimir no stdout
- **cut** : imprimir partes de um ficheiro para o stdout
- **paste** : imprimir linhas de um ficheiro separadas por tabs para o stdout
- **tee** : Redireciona para o nome do ficheiro passado como argumento e para o stdout

5.2 Exercise 2 - Redirect input and output

5.2.1 1

> : redirecionar o output do comando anterior do stdout para um ficheiro
>> : append do output do comando anterior do stdout para um ficheiro

5.2.2 2

2> : redireciona o stderror para um ficheiro

5.2.3 3

| : redireciona o stdout de um comando para o stdin do comando seguinte

5.2.4 4

2>&1 : redireciona o stderror para o stdout
1>&2 : redireciona o stdout para o stderror

5.3 Exercise 3 - Using special characters

5.3.1 1

touch : criar ficheiros caso o ficheiro não exista. Alterar a data de modificação caso o ficheiro exista
a* : [REGEX] Lista todos os ficheiros que o primeiro caracter seja um a, independentemente do número de ficheiros
a? : [REGEX] Lista todos os ficheiros começados por a e com mais 1 caracter
* : [REGEX] Lista qualquer ficheiro independentemente do numero de caracteres

5.3.2 2

[ac] : [REGEX] Lista os ficheiros com os caracteres entre
[a-c] : [REGEX] Lista os ficheiros com os caracteres entre a e c
[ab]* : [REGEX] Lista os ficheiros com os caracteres {a, b} independentemente do número de caracteres

5.3.3 3

o \ antes de um caracter especial desativa as capacidades especiais do stdout

a* : [REGEX] Lista todos os ficheiros começados por a independentemente do número de caracteres
a* : Lista o ficheiro com o nome a*
a? : [REGEX] Lista todos os ficheiros começados por a e com mais um caracter
a\? : Lista o ficheiro com o nome a?
a\[: Lista o ficheiro com o nome a[
a\\ : Lista o ficheiro com o nome a

5.3.4 4

Usando '' ou "" podemos desativar o significado de caracteres especiais
a* : [REGEX] Lista todos os ficheiros começados por a independentemente do número de caracteres
'a*' : Seleciona o ficheiro a*
"a*" : Seleciona o ficheiro a*

5.4 Exercise 4 - Declaring and using variables

5.4.1 1

<variable name>=…. : Atribuição de variáveis em bash. Não deve ter espaço entre o nome da variável e a atribuição
$<variable name> : lê o valor da variável (em bash existe diferença entre atribuir um valor a uma variável e ler o valor da variável). Pode se atribuir nome de ficheiros e usar REGEX (p.e. z=a*)
${<variable name>} : lê o valor da variável (em bash existe diferença entre atribuir m valor a uma variável e ler o valor da variável)
${<variable name>}<etc> : Concatena o valor da variável com o que está à frente ()

5.4.2 2

- $<variable name> : Acede ao valor da variável
- "$<varibale name>" : Acede ao valor da variável (não aplica quaisquer caracteres especiais). P.e. se v=a, "$v" *será igual a a* em vez de todos os ficheiros começados por a com mais um caracter adicional
- '$<variable name>' : Ignora a leitura da variável e de um possível REGEX, devolvendo $<variable name>

5.4.3 3

- ${<variable name>:start:numero de caracteres} : trata a variável como string, criando uma substring começando no caracter start com o numero de caracteres especificado. Pode ter espaços entre os :
- ${<variable name\>/<search substring\>/<replace substring>} : Procura uma substring na variable name e substitui por outra substring indicada

5.5 Exercise 5 - Declaring and using functions

5.5.1 1

Para declarar uma função:

```
1 <nome_da_funcao>()
2 {
3     # corpo da função
4 }
```

5.5.2 2

`$#` : Número de argumentos de uma função
`$1` : Primeiro argumento
`$2` : Segundo argumento
`$*` : Todos os argumentos - Ignora sequencias de white space dentro das aspas na passagem de argumentos da bash
`$@` : Todos os argumentos - Ignora sequencias de white space dentro das aspas na passagem de argumentos da bash
`"$*"` : Todos os argumentos - Preserva a forma dos argumentos passados entre aspas (i.e., o white space)
`"$@"` : Todos os argumentos - Preserva a forma dos argumentos passados entre aspas (i.e., o white space)

5.6 Exercise 6

5.6.1 1, 2 e 3

- `{ .....\}` : Agrupar commandos (pode ser redirecionado o stdout usando | ou >). A lista de comandos é executada na mesma instância da bash em que é chamada (contexto global de execução, com variáveis globais)

- (…….) : Agrupar comandos (pode ser redirecionado o stdout usando | ou >). O grupo de comandos é executado noutra instância da bash (contexto próprio de execução, com variáveis locais)

5.7 Exercise 7

5.7.1 1

`$?` : Valor de retorno de um comando (semelhante a C/C++). Se for `'1'` existe um erro na execução do comando. Se for `'0'` está tudo bem

5.7.2 2

```
1  echo -e : Faz parse de códigos de cores
2
3  "e\33m ... \e[0m" : Código de cores que define a cor de sucesso
4  "e\31m ... \e[0m" : Código de cores que define a cor de erro
```

Estrutura de um if:

```
1  if <cond>
2  then
3      <statment>
4  else
5      <statment>
6  fi
```

5.7.3 3

Os parentesis retos na condição do if (p.e. `if [ -f $1]`) que chamam a função test devem estar com pelo menos um espaço entre os outros caracteres

5.7.4 4

Os operadores têm de estar com pelo menos um espaço de intervalo
! : Operador not

5.7.5 5

&& : Operador and
|| : Operador or

5.8 Exercise 8 - The multiple choice case construction

5.8.1 1 Case stament

```
case <variavel para selecionar> in
    <cond1>) <statment 1>;;
    <cond2>) <statment 2>;;
    <cond3>) <statment 3>;;
    ....
esac
```

Onde:
- A pode ser `$#`, `$*` ou `$1`
- O ;; no final da `<ação #>` equivale ao fim da branch (break em C)
- O | permite a definição de uma várias alternativas (condições) para o mesmo case (e consequentemente ação)
- O * segnifica qualquer valor. Ao ser colocado em último permite selecionar todas as outras opções que ainda não forma cobertas (equivalente ao default em C)
Exercise 9 - The repetitive for contruction

5.8.2 1

A syntax de um for é:

```
for <variavel de iteracao> in <lista de objectos para iterar>
do
    <statment>
done
```

Onde `<lista de objectos para iterar>` podem ser ficheiros e/ou pastas e podem ser usados caracteres especiais como `a*`

5.8.3 2

5.9 10 - The repetitive while and until constructions

5.9.1 1

Estrutura de um while:

```
while [ <condicao de paragem> ]
do
    <statment>
    shift
done
```

Estrutura de um until

```
until [ <condição de paragem> ]
do
    <statment>
    shift
done
```

Onde a condição de paragem pode ser escrita como: `<variable> <condição de teste> <fim>`

As condições de teste podem ser:

```
-gt : greater than
-eq : equal
-lt : less than
-le : less or equal than
-ge : greater or equal than
```

`shift` é uma palavra equivalente ao continue em C

5.10 Exercise 11 - Script Files

6 1

O cabeçalho do ficheiro de script é:

```
#!/bin/bash
# The previous line (comment) tells the operating system that
#  this script is to be executed in bash
#
```

Condições usadas:

```
[ $# -ne 1 ] : número de argumentos diferente de 1
! [ -f $1 ] : O primeiro argumento dado não é um ficheiro
1>&2 - Redirecionar o stderror para o stdout
```

6.1 Exercise 12 - Bash supports both indexed and associative arrays

6.1.1 1

Os indices de um array não são continuos e não podem ser negativos

A declaração explicita dos arrays pode ser feita fazendo: `declare -a <array>[<idx>]=<value>`
Outras operações:
- **Atribuição**: `<array>[<idx>]=<valor>`
- **Leitura**: `${<array>[<idx]}`
- **Leitura de todos os elementos do array**: `${a[*]}`
- **Número de elementos do array**: `${#a[*]}`
- **Lista dos indices do array** : `${!a[*]}`

Os indices podem ser obtidos com expressões aritméticas
A iteração pelos indices é feita da mesma forma que em python
- **Iterar na lista de elementos**: `for <variavel> in ${<array>[*]}`
- **Iterar na lista de indices**: `for <variavel> in ${!a[*]}`
Exemplo de código para imprimir os indices e os elementos

```
for v in ${!a[*]}
do
    echo "a[$i] = ${a[$i]}"
done
```

6.1.2 2

A declaração de arrays associativos tem de ser feita de forma explicita `declare -A <array>`

- A **atribuição de valores** para um **array associativo**: `<array>["<key>"]=<value>`
- **Listar os elementos no array** : `${<array>[*]}`
- **Listar o número de elementos no array** : `${#<array>[*]}`
- **Listar os indices usados no array** : `${!<array>[*]}`

Exemplo de código para percorrer as keys e imprimir as keys e os values

```
for i in ${!arr[*]}
do
    echo "Key = $i | Value = ${arr[$i]}\"
done
```

7 sofs2017

> *The sofs17 is a simple and limited file system, based on the ext2 file system, which was designed for purely educational purposes and is intended to be developed in the practical classes of the Operating Systems course in academic year of 2017/2018. The physical support is a regular file from any other file system.*

- Sistema simples e limitado
- Baseado no ext2
- Suporte físico: um ficheiro regular de outro sistema operativo
  - Este ficheiro será formatado para imitar uma unidade física formatada no formato sofs17

8 Organização das aulas durante o sofs17

2 horas:
- 1h30 : interagir relativamente ao trabalho pendente
- 0h30 : falar da próxima camada de software

9 Introduction

- Durante a execução de um programa, ele manipula informação (produz, acede e/ou modifica).
- Esta informação tem de ser guardada exteriormente (**mass storage**)
  - discos magnéticos
  - discos ópticos
  - SSD
  - …
- **mass storage** (armazenamento de massa): dispositivos organizados em arrays de blocos
  - 256 bytes até 8 Kbytes por bloco
  - os blocos são numerados sequencialmente (LBA model)
  - o acesso para R/W é efetuado através de um ID (identification number)

Block 0	Block 1	Block 2	Block 3	…	Block NTBK-1

Cada bloco tem BKZS bytes de informação - O acesso ao disco é feito bloco a bloco: - **Não é possível modificar um único byte**

> *Direct access to the contents of the device* ***should not be allowed to the application programmer.***

Porque:

- Um sistema de ficheiros é complexo
- A sua estrutura interna precisa de *enforce quality criteria* para garantir:
  - eficiência
  - integridade
  - partilha de acessos
- O utilizador não sabe o conteúdo de cada bloco de dados nem em que blocos a informação do ficheiro x está.

Daí a necessidade/exigência da existência de um *uniform interaction model* (Nível de abstração).

ficheiro:

- unidade lógica de armazenamento de massa
- *abstract data type*, sobre o ponto de vista do programador
  - composto por um conjunto de atributos e operações
- tipos:
  - NTFS
  - ext3
  - FAT*
  - UDF
  - APFS
  - …

> *Is the operating system's responsability to provide a set of from the file system point of view: system calls that implement such abstract data type. These system calls should be a simple and safe interface with the mass storage device. The component of the operating system dedicated size — the size in bytes of the file's data to this task is the file system*

Ou seja, operações de leitura e escrita **são sempre efetuadas no contexto de ficheiros**, através de syscall disponibilizadas pelo OS.

A interface de comunicação com o OS é a mesma, mas diferentes sistemas de ficheiros obrigam a diferentes técnicas e manipulação do filesystem, que são transparentes para o programador.

9.1 File as an abstract data type

Os atributos de um ficheiros dependem da implementação do sistema de ficheiros.

Os mais comuns:

- **name**:
- **internal identifier**: ID númerico (e interno - o user desconhece) que é usado pelo OS para aceder ao ficheiro
- **size**: tamanho do ficheiro em bytes
- **ownership**: Identificação de quêm o ficheiro pertence (usado para controlo de acessos)
- **permissions**: Atributos que em conjunto com a ownership (des)autorizam o acesso ao ficheiro
  - Possíveis permissões:
    * r: read
    * w: write
    * x: execute
    * d: directory
  - Nos diretórios, execução `x` significa que eu tenho permissões para atravessar o diretório (posso não ter permissões nem para ler nem para escrever, mas posso seguir no diretório para chegar a outro path)
- **acess monitoring**: data do último acesso e última modificação
- **localization of the data**: identificação dos clusters onde os dados do ficheiros estão guardados
- **type**: tipo dos ficheiros:
  1. ordinary or regular: qualquer ficheiro ”normal” para o utilizador [ID= -]
     - .txt
     - .doc
     - .png
     - .avi
     - .mp3
     - .pdf
     - .c
     - .exe
     - …
  2. directory: um tipo de **ficheiro** interno, com um formato pre-definido, usado para localizar outros ficheiros ou diretórios, permitindo visualizar o sistemas de ficheiros como uma árvore de diretórios e fichieros [ID= d]
  3. shortcut (symbolic link): ficheiro interno, com um formato predefinido, que contém uma referência para outro ficheiro/diretório [ID= s]
     - ref pode ser absoluta ou relativa
  4. character device(**special file**): *represents a device handles in bytes* [ID= c]
  5. block device(**special file**): *rep esents a device handles in block* [ID= b]
  6. socket(**special file**): *represents a file used for inter-process communication* [ID= s]
  7. named pipe: ***another special file*** *used for inter-process communication* [ID = p]

ID (`ls -ll`)	meaning
-	ordinary/regular file
d	directory
s	symbolic link

ID (`ls -ll`)	meaning
c	character device
b	block device
s	socket
p	named pipe

No sofs17 só serão considerados os três primeiros tipos de ficheiros.

9.1.1 Operações em ficheiros

- Dependem do OS
- Todas as operações estão disponíveis **apenas** através de syscalls (funções que funcionam como *entry-points* para o OS
- Syscalls em Linux para os tipos de ficheiros a usar no sofs17:

```
1 /*[TODO] Inserir descrição das operações para o teste*/
```

  - Comun aos três:
    * open
    * close
    * chmod
    * chown
    * utime
    * stat
    * rename
  - Comun para **ficheiros regulares** e **shortcuts**:
    * link
    * unlink
  - Só para **ficheiros regulares**:
    * mknod
    * read
    * write
    * truncate
    * lseek
  - Só para **diretórios**
    * mkdir
    * rmdir
    * getdents
  - Só para **shortcuts**
    * symlink
    * readlink

A descrição destas syscalls pode ser obtida executando num terminal o comando:

```
1 man 2 <syscall>
```

9.2 FUSE

Inserir um novo filesystem num OS requer: 1. Integração do software que implementa o novo filesystem no kernel 2. Instanciação de um ou mais dispositivos que usam o formato do novo filesystem

> *In monolitic kernels, the integration task involves the recompilation of the kernel, including the sofware that implements the new file system. In modular kernels, the new software should be compiled and linked separalely and attached to the kernel at run time.*

Tarefa morosa e difícil, que requer *deep knowledge of the hosting system* - **OUT OF THE SCOPE OF SO**

> FUSE (File system in User SpacE) is a canny solution that **allows for the implementation of file systems in user space** (memory where normal user programs run). Thus, **any effect of flaws of the suporting software are restricted to the user space, keeping the kernel imune to them.**

O novo filesystem é executado em cima do FUSE com permissões de user e não de root. Assim, certas operações que poderiam danificar fisicamente os dispositivos estão interditas e erros no código não geram kernel-panics.

Isola-se a execução deste novo filesystem do kernel.

9.2.1 Infrastrutura

- **Interface com o filesystem nativo**: funciona como mediador entre as syscalls do sistema nativo e as implementadas em user sapce
- **Implementation library**:
  - Estruturas de dados
  - Protótipos de funções (que devem ser desenvolvidas pelo user para criar o filesystem específico)
  - Métodos para instanciar e integrar o novo filesystem com o kernel

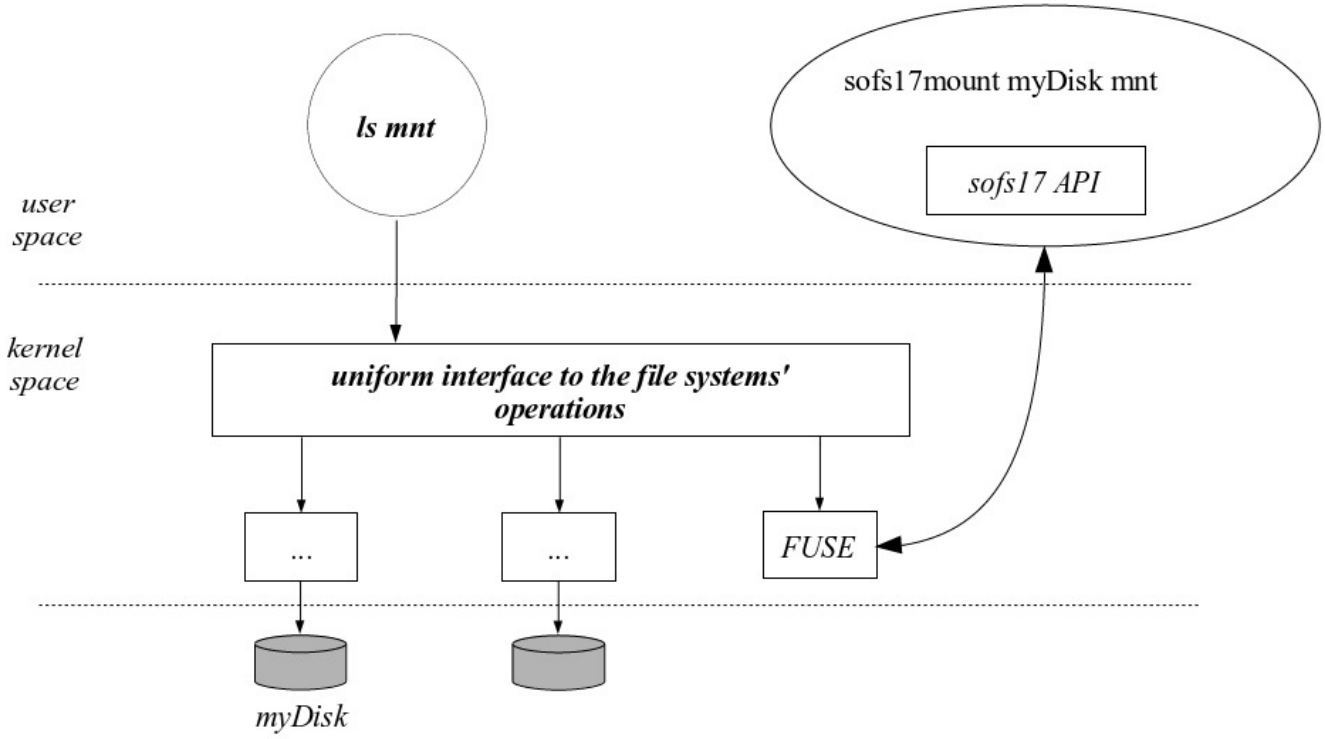


Figure 17: FUSE diagram with sofs17

10 SOFS17 Architecture

- Um disco é um conjunto de blocos numerados
  - No sofs17 cada bloco tem 512 bytes
- Os elementos principais na definição da arquitectura do sofs2017 são:
  - **superblock**: estrutura de dados guardada no bloco 0. Contém atributos globais para
    * o disco como um todo
    * outras estruturas de dados
  - **inode**: estrutura de dados que contém **todos os atributos de um ficheiro, excepto o nome**
    * Existe um região contínua no disco reservada para guardar todos os inodes (inode table)
    * A identificação de um inode é feita com um indíce que representa a sua posição relativa na inode table
  - **directory**: *special file* que permite a implementação de uma hierarquia (árvore) para acesso aos ficheiros
    * É composto por um conjunto de entradas (*directory entries*) em que cada uma associa um nome a um inode
    * Assume-se que o diretório de raiz (*root*) está associado ao inode 0
  - **disk blocks**: usados para guardar os dados
    * Estão organizados em grupos de 4 blocos contínuos -> **clusters**
    * A identificação de um cluster é dada através de um indíce que identifica a posição relativa do cluster na cluster zone
  - **cluster**: Para cada cluster existe um bit correspondente que representa o seu estado (vazio/preenchido)
    * Estes bits estão guardados no sistema de ficheiros numa área chamada *reference bitmpa table*

De forma geral, os N blocks de um disco formatado em sof17 organizam-se em 4 áreas:

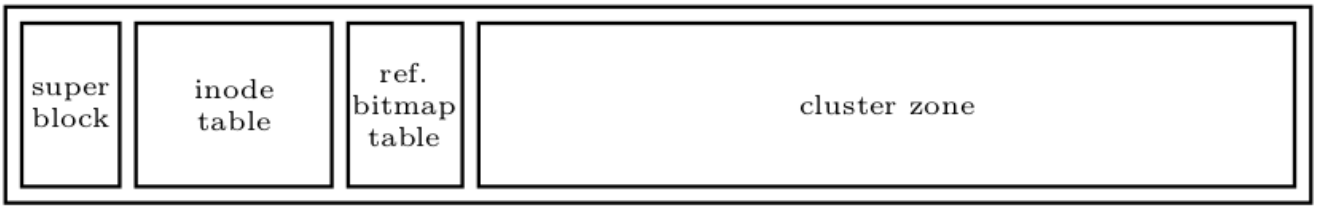


Figure 18: Organização de um disco formatado em sofs17

10.1 List of free inodes

- O número de inodes num disco sofs17 é **fixo após a formatação**.
- Quando um novo ficheiro é criado, deve-lhe ser atribuido um inode. Para isso é preciso:
  - Definir uma política (conjunto de regras) para decidir que free inode será usado
  - Definir e guardar no disco uma estrutura de dados adequada à implementação desta política

---

> *In sofs17 a FIFO policy is used, meaning that the first free inode to be used is the oldest one. The implementation is based in a double linked list of free inodes, built using the inodes themselves.*

- Na estrutura de inodes, existem dois campos que guardam os indíces do próximo inode e do inode anterior vazios (criam uma lista ligada)
- Estas listas ligadas de inodes são circulares, ou seja:
  - O *previous* inode livre do primeiro free inode é o último free inode
  - o *next* free inode do último inode é o primeiro free inode
  - Assim:
  - Cada numero da lista paonta sempre para o seguinte.
  - O previous aponta sempre para o elemento aterior.
  - Só preciso de saber a tail porque a previous do head é a tail
- No *superblock*, dois campos guardam o número total de free inodes e um indíce para o primeiro free inode
- O número de inodes por default é [NUM_BLOCKS]/8

Correspondência univoca entre o inode e o nome do ficheiro

- stat : mostra a estatísticas do ficheiro (filesize, blocks, ID Block, device, inode, links e datas de aceso, modificação e change)
  - Ficheiro `.` : diretório atual
  - Ficheiro `..` : diretório atual

10.2 List of free clusters

- Tal como os inodes, o número de clusters num disco é fixo após a formatação.
- Para manipular a estrutura de clusters é necessário:
  - Definir uma maneira de representar o estado (livre/usado) de todos os clusters no disco
  - Definir uma política para decidir que cluster (que estea livre) deve ser usado quanto é necessário um cluster
  - Definir e guardar no disco uma estrutura de dados adequada para representar o sistema de clusters e permitir a implementação dos pontos acima

Concretamente no sofs17:
- Existe uma estrutura de *bit map* unívoca que mapeia o estado de um cluster - Esta estrutura é formada por um blocos contínuos no disco (logo após a *inode table*) - Cada bit funciona como uma variável booleana que classifica o cluster que referencia como vazio ou ocupado - Existem duas caches guardadas no superblock que são usadas para guardar referências diretas para os clusters. São: - **retrieval cache**: - **insertion cache**: - Considera-se que um cluster está livre nas seguintes condições - A sua referência está em qualquer uma das caches - Ou o seu bit correspondente no bit map indica que está vazio

- As duas caches têm como função melhorar a eficiência de operações de **alocação**(atribiuir um cluster livre a um novo ficheiro a ser guardado em disco) e **libertação**(remover as referências para um dado cluster).
  - Na maioria das vezes as duas operações só precisam de aceder ao superblock e não fazem mais di que um acesso ao disco

10.2.1 Retrieval Chache

Serve para guardar as referências após eliminar um ficheiro. Se o disco tiver vazio, a referência deve ser max e não 0. O valor 0 significa que está cheio a retrieval cach está cheia.

10.2.2 Insertion cache

Serve para guardar as referências de ficheiros a inserir. Se o cache estiver vazia, a referência deve ser 0. O valor 0 significa que a insertion cache está cheia.

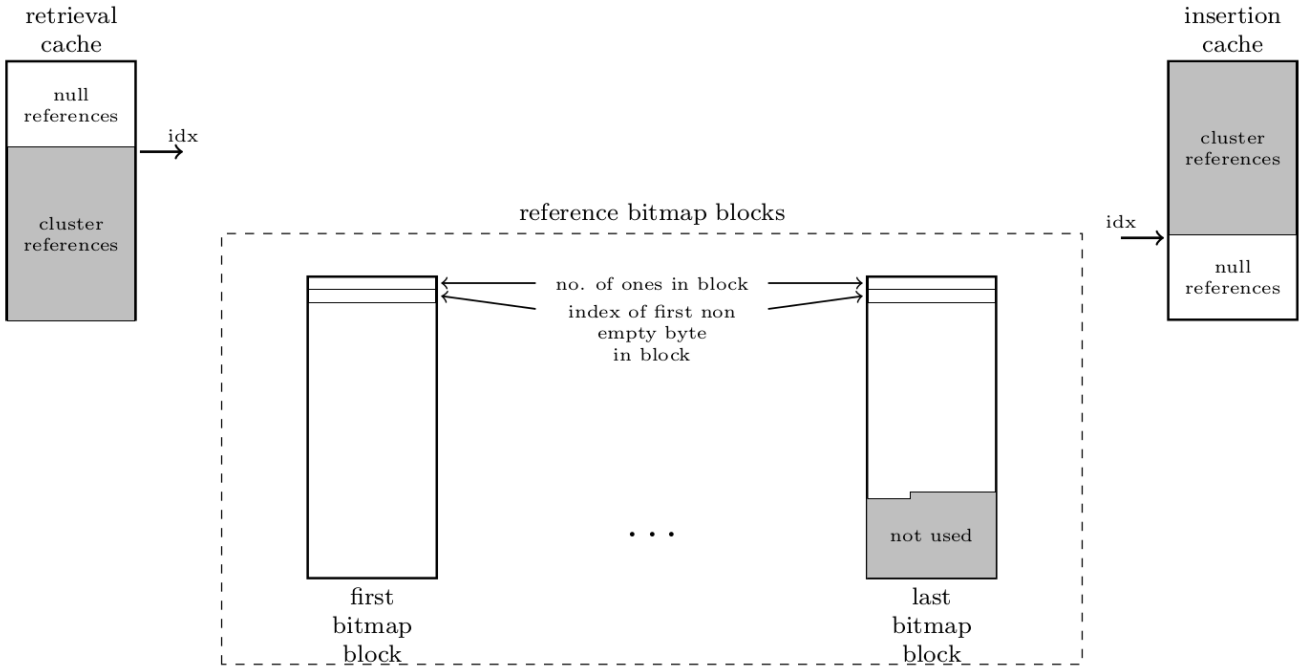


Figure 19: Caches and Reference bitmap blocks

10.2.3 Allocation

1. Uma referência para um cluster livre é obtida da retrieval cache
    1. Se a cache tiver vazia, são transferidas várias referências do bit map para a cache antes de se obter a referência para o cluster livre
    2. As referências transferidas para os clusters são transferidas de forma sequencial
2. Um byte iglobal (guardado no superblock) indica a localização no bit map de onde a transferência deve começar
    1. Assim cria-se rotatividade no uso dos clusters
3. Os clusters não funcionam estritamente como uma FIFO.
    1. Se a *retrieval cache* e o bit map estão vazios, as referências presentes na inserion cache são transferidas da insertion cache para o disco
    2. Se se efetua uma release operation a referência para o novo cluster livre é inserida na inserion cache
    3. Se esta cache está cheia, então as referências para a cache são transferidas para o bit map, antes de se proceder como anteriormente

10.3 List of clusters used by a file (inode)

> ***Clusters are not shared among files, thus, an in-use cluster belongs to a single file***

O número de clusters usado por um ficheiro é $N_c = roundup(\frac{size}{ClusterSize})$, onde:

- **size**: tamanho em bytes de um ficheiro
- **ClusterSize**: tamanho de um cluster em bytes

10.3.1 Considerações:

- N_c pode ser muito elevado.
  - Um disco com um block size de 512 bytes e com um *cluster size* de 4 blocos. Se o ficheiro a guardar tiver 2 GByte são necessários 1 milhão de clusters
- N_c pode ser nulo (0):
  - Se o ficheiro tiver 0 bytes, N_c = 0

> *Thus, it is impractical that all the clusters used by a file are contiguous in disk. The data structure used to represent the sequence of clusters used by a file must be flexible, growing as necessary.*

A **escrita** e a **leitura** no disco **não são sequenciais**, mas sim aleatórias.

Exemplo: pretendemos aceder ao indice *j* de um ficheiro. Para obter o cluster que contém esse ficheiro precisamos de saber o indice do cluster do ponto de vista de um ficheiro, $ClusterIndex = \frac{j}{ClusterSize}$

Para obter a localização do ficheiro no disco, temos de obter o número do cluster usando a estrutura do filesystem. e No sofs17:
- a *data strucuture* definida é dinâmica e permite uam identificação rápida de qualquer data cluster. - Cada inode permite o acesso a um array dinâmico, ***d***, que identifica a sequencia de clusters usados para guardar os dados associados com um ficheiro. - Sendo ***ClusterSize*** o tamanho em bytes de um cluster, temos:

Cluster	Descrição
d[0]	Número do cluster que contem os primeiros ClusterSize bytes
d1	Número do cluster que contem os segundos ClusterSize bytes
.	…
.	…
d[$N_c - 1$]	Número do cluster que contém os últimos ClusterSize bytes

O array ***d*** não é guardado num único local:
- Os **primeiros 6 elementos** são diretamente guardados no inode, no campo d (referência direta) - Os **próximos elementos, se existirem, são referenciados através dos campos: - i_1: referência indireta - i_2**: referªencia indireta dupla

10.3.2 Campo i_1

- É usado para estender indiretamente o array ***d***
- O primeiro elemento, $i_1[0]$ é usado para referenciar um cluster onde cada bloco é um endereço para uma posição no disco (cluster) onde estão guardados os dados do ficheiro
- Permite extender o array d de d[6] para d[RPC+6-1]=d[RPC+5]
  - **RPC** é o número de referências para clusters que podem ser guardadas num cluster

10.3.3 Campo i_2

- Se mesmo assim não for possível guardar os dados do ficheiro usando referência indireta simples, pode ser usada referência indireta dupla.
- O campo i_2 do inode é usado para referenciar um cluster em que cada bloco do cluster referenciad um cluster de dados
- É usado para extender o array de referências indiretas i_1 usando as referências indiretas do cluster. Assim temos um array de referências indiretas de $i_1[1]$ até $i_1[RPC]$.
- O primeiro cluster do array de referências indiretas duplas é $i_1[1]$ ($i_1[0]$ corresponde às referências diretas).
  - Traduzindo para o array de ***d*** corresponde aos segmentos do array entre d[RPC + 6] e $d[2 * RPC + 5]$

10.3.4 NullReference

- É usada para representar uma referência que não existe
- Exemplos:
  - se d1 for uma NullReference, o ficheiro não contém o index de cluster 1
  - se i_1, representando $i_1[0]$ é equal to NullReference, significa que entre d[6] até d[RPC+5] todos os indices são NullReference e o o ficheiro não contém estes indices
  - se i_2 for uma NullReference, significa que entre $i_1[1]$ to $i_1[RPC]$ são NullReferences e portanto d[RPC+6] até d[RPC^2 + RPC+ 5] são NullReferences e o ficheiro não contém esses indices

10.4 Directories

- Um diretório pode ser visto como um array de entrada para diretórios.
- No sofs17:
  - Um diretório é uma estrutura de dados composta por um array de bytes com tamanho fixo. Usados para guardar:
    - nome
    - referência que associa o diretório a um inode
  - A estrutura de dados foi definida para que um cluster suporte apenas um número inteiro de diretórios
    - As primeiras duas entradas `"."` e `".."` representam o diretório atual e o diretório pai
    - Um diretório pode tomar um de três estados:
      - **in-use**: contém o nome e o inode number de um ficheiro que existe (seja ele um regular file, diretório ou atalho)
      - **deleted**: o nome contém o 1 e último caracter trocado, passando a ser `\0 name[0:end-1]` (ou seja, uma null string, mas com o nome recuperável). Mantém o slot no diretório.
      - **clean**: Todos os caracteres do nome são `\0` e o reference field é uma NullReference

  - Quando um cluster é adicionado a um diretório, primeiro deve formatado como uma sequência de diretórios de entrada limpas.
    * O tamanho do diretório é sempre um múltiplo do tamanho de um cluster e nunca pode "encolher" (devido à forma como o delete está implementado

11 Formatting

A operação de formatação deve preencher todos os blocos do disco para criar um disco sofs17 vazio.

Um disco formatado contém:
- root directory
- duas entradas, `"."` e `".."`, que apontam para o inode 0 (root directory)

A operação de formatação deve:

- escolher o valor apropriado para o número de inodes, o número de clusters e o número de blocos usados pelo bit map
  - tem de ter em consideração o número de inodes especificado pelo utilizador e número total de blocos no disco.
  - todos os blocos no disco devem ser usados. Se não forem usados para outros propósitos devem ser adicionados à inode table
- Preencher a tabela de inodes:
  - inode number 0 é o root directory
  - todos os outros inodes estão livres
  - A lista de inodes livres começa no inode número 1 e termina no último inode
- Preencher o bit map, sabendo que:
  - O cluster 0 está a ser usado pelo diretório root
  - Todos os outros clusters estão livres
- Preencher o root directory, ocupando o cluster número 0
- Preencher com zeros todos os clusters livres, se especificado pela ferramenta de formatação

12 Code Structure

A estrutura do código é apresentada abaixo:

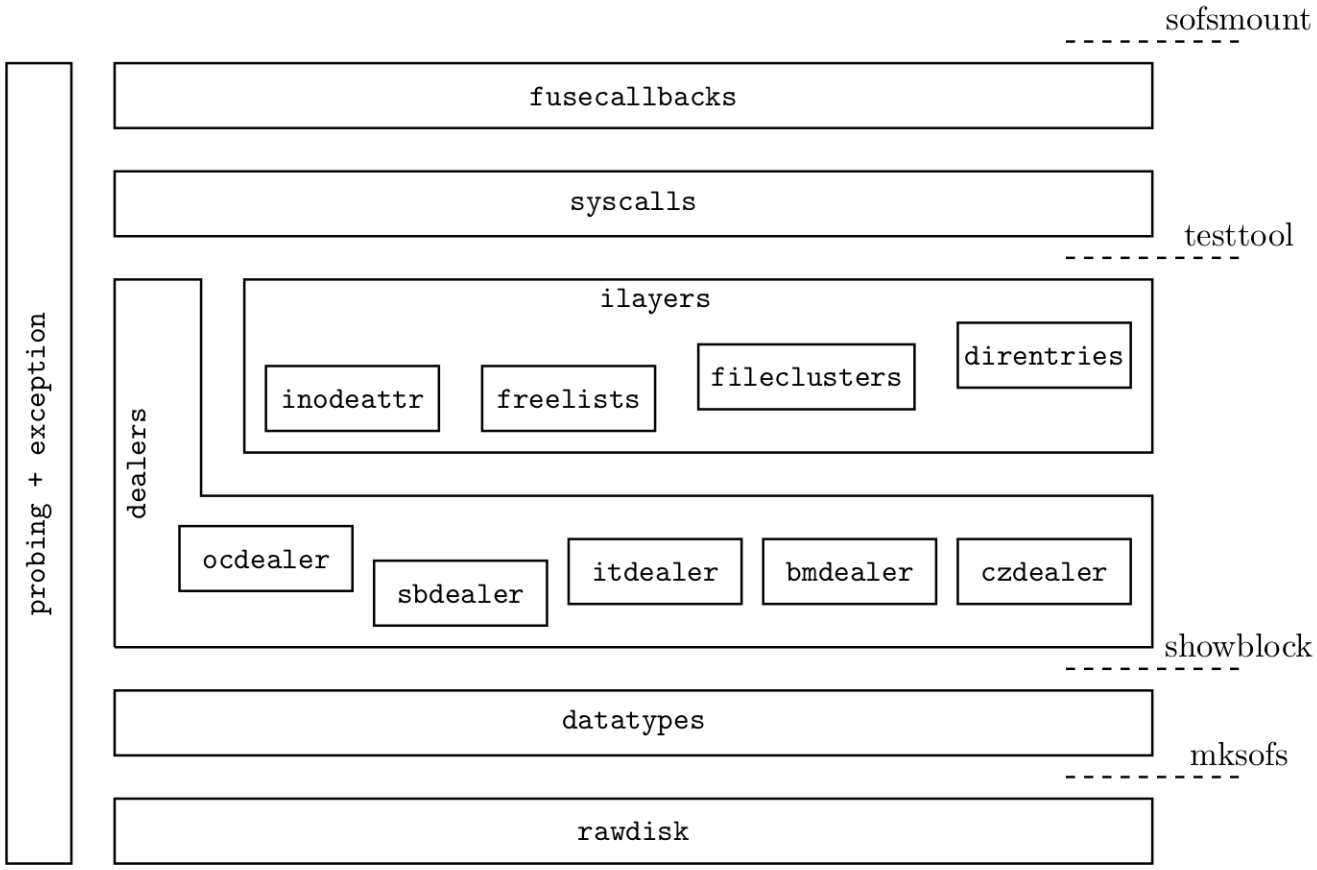


Figure 20: Code Strucuture to be developed

12.1 Rawdisk

Implementa o acesso físico ao disco

12.2 Dealers

- Implementam o acesso ao superblock, inodes, bit map e clusters
- São opcionais (só são feitas se os alunos desejarem ter notas mais altas)

12.2.1 sbdealer

Acesso ao superblock

12.2.2 itdealer

Acesso à inode table e aos inodes

12.2.3 bmdealer

Acesso às referências da bit map table

12.2.4 czdealer

Acesso à cluster zone, usando as cluster references

12.2.5 ocdelaer

Open/close the dealers

12.3 ilayers

Funções intermédias Obrigatórias

12.3.1 inodeattr

Lida com a manipulação dos campos especiais dos inodes

12.3.2 freelists

Manipular a lista dos inodes livres e a lista de clusters livres

12.3.3 filecluster

Lidar com os clusters de um inode (file clusters associados a um ficheiro)

12.3.4 direntries

Lidar com entradas de diretórios

12.4 syscalls

versão das syscalls de sistema adaptadas ao *sofs17* Cada grupo Só irá implementar 6 das 24 utilizadas.

12.5 fusecallbacks

Interface com FUSE

12.6 probing

Biblioteca para debug

12.7 exception

o tipo de exceções lançadas em caso de erro

- **datatypes**: um conjunto de constantes que podem ser usadas para aceder aos ficheiros
  - InodesPerBlock
  - ReferencesPerBlock
  - ReferencesPerCluster
  - ReferencesPerBitmapBlock
  - BlocksPerCluster
  - CLusterSize
  - DirentriesPerCluster
  - NullReference

13 createDisk

- Cria um disco **não formatado** que serve de suporte a um sistema de ficheiros.
- Na prática, um disco é um ficheiro que possui uma estrutura de blocos fixa.
- Apenas é garantida que a estrutura do disco possui:
  - o número desejado de clusters
  - o número desejado de bytes por cluster
- Para o disco ser um sistema de ficheiros válido é necessário formatá-lo com ferramentas adequadas para o tipo de sistemas de ficheiros pretendido

13.1 Exemplo de utilização

```
./createDisk <diskfile> <numblocks>
```

O *output* após a execução do script para um disco com 1000 blocos é:

```
./createDisk <diskfle> 1000
1000+0 records in
1000+0 records out
512000 bytes (512 kB) copied, 0.05734 s, 8.9 MB/s
```

13.2 Implementação

O createDisk usa o comando *dd* para escrever para o disco/ficheiro e preenche-o com valores aleatórios obtidos do */dev/u-random*.

```
#!/bin/bash

if [ $# != 2 ]; then
        echo "$0 diskfile numblocks"
```

```
        exit 1
fi

dd if=/dev/urandom of=$1 bs=512 count=$2
```

14 showblock

- Permite visualizar a informação contida numa sequência de blocos do disco:
  - Os dados dos blocos podem ser formatados para serem facilmente interpretáveis por humanos
  - A formatação dos dados dos blocos pode ser feita de acordo com a função de cada um dos blocos

14.1 Utilização

```
# showblock -h imprime a ajuda
./showblock [ OPTION ] <disk filename>
```

14.1.1 Opções de Visualização

Option	Description
-x	show block(s) as hexadecimal data
-a	show block(s) as ascii/hexadecimal data
-s	show block(s) as superblock data
-i	show block(s) as inode entries
-d	show block(s) as directory entries
-r	show block(s) as cluster references
-b	show block(s) as bitmap references

14.2 Exemplos

```
# Showblock para o bloco 1 em hexadecimal
./showblock -x 1 disk.sofs17
0000: fd 41 02 00 e8 03 00 00 e8 03 00 00 00 08 00 00 01 00 00 00 dc ee f9 59 dc ee f9 59
    dc ee f9 59
0020: 00 00 00 00 ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
    ff ff ff ff
0040: 00 02 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 02 00 00 00
    8f 00 00 00
0060: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
    ff ff ff ff
```

```
0080: 00 02 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 03 00 00 00
    01 00 00 00
00a0: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
    ff ff ff ff
00c0: 00 02 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 04 00 00 00
    02 00 00 00
00e0: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
    ff ff ff ff
0100: 00 02 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 05 00 00 00
    03 00 00 00
0120: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
    ff ff ff ff
0140: 00 02 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 06 00 00 00
    04 00 00 00
0160: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
    ff ff ff ff
0180: 00 02 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 07 00 00 00
    05 00 00 00
01a0: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
    ff ff ff ff
01c0: 00 02 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 08 00 00 00
    06 00 00 00
01e0: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
    ff ff ff ff

# A informação não é diretamente percetível por humanos.
# Sabendo que este bloco corresponde ao primeiro bloco da inode table,
# se executarmos o mesmo comando mas o output vier formatado para inodes, temos:
./showblock -i 1 disk.sofs17
Inode #0
type =  directory, permissions = rwxrwxr-x, lnkcnt = 2, owner = 1000, group = 1000
size in bytes = 2048, size in clusters = 1
atime = Wed Nov  1 15:57:16 2017, mtime = Wed Nov  1 15:57:16 2017, ctime = Wed Nov  1
    15:57:16 2017
d[] = {0 (nil) (nil) (nil) (nil) (nil)}, i1 = (nil), i2 = (nil)
----------------
Inode #1
type = free clean, permissions = ---------, lnkcnt = 0, owner = 0, group = 0
size in bytes = 0, size in clusters = 0
next = 2, , prev = 143
d[] = {(nil) (nil) (nil) (nil) (nil) (nil)}, i1 = (nil), i2 = (nil)
----------------
Inode #2
type = free clean, permissions = ---------, lnkcnt = 0, owner = 0, group = 0
size in bytes = 0, size in clusters = 0
next = 3, , prev = 1
d[] = {(nil) (nil) (nil) (nil) (nil) (nil)}, i1 = (nil), i2 = (nil)
----------------
Inode #3
type = free clean, permissions = ---------, lnkcnt = 0, owner = 0, group = 0
size in bytes = 0, size in clusters = 0
next = 4, , prev = 2
```

```
d[] = {(nil) (nil) (nil) (nil) (nil) (nil)}, i1 = (nil), i2 = (nil)
----------------
Inode #4
type = free clean, permissions = ---------, lnkcnt = 0, owner = 0, group = 0
size in bytes = 0, size in clusters = 0
next = 5, , prev = 3
d[] = {(nil) (nil) (nil) (nil) (nil) (nil)}, i1 = (nil), i2 = (nil)
----------------
Inode #5
type = free clean, permissions = ---------, lnkcnt = 0, owner = 0, group = 0
size in bytes = 0, size in clusters = 0
next = 6, , prev = 4
d[] = {(nil) (nil) (nil) (nil) (nil) (nil)}, i1 = (nil), i2 = (nil)
----------------
Inode #6
type = free clean, permissions = ---------, lnkcnt = 0, owner = 0, group = 0
size in bytes = 0, size in clusters = 0
next = 7, , prev = 5
d[] = {(nil) (nil) (nil) (nil) (nil) (nil)}, i1 = (nil), i2 = (nil)
----------------
Inode #7
type = free clean, permissions = ---------, lnkcnt = 0, owner = 0, group = 0
size in bytes = 0, size in clusters = 0
next = 8, , prev = 6
d[] = {(nil) (nil) (nil) (nil) (nil) (nil)}, i1 = (nil), i2 = (nil)
----------------
```

15 rawlevel

- Camada que permite a manipulação do disco ao nível do bloco

15.1 Módulos

- **mksofs** : Formatador
- **rawdisk** : Acesso aos blocos do disco

16 rawdisk

- Permite o acesso aos blocos do disco
  - Os blocos são a menor unidade lógica no *filesystem*
- Medeia o acesso direto ao disco, impedindo que ocorram erros que podem danificar a estrutura do sistema de ficheiros

16.1 Macros

```
// block size (in bytes)
#define BlockSize (512U)
```

16.2 Funções

```
void soOpenRawDisk (const char *devname, uint32_t *np=NULL)
```

- Abre o dispositivo de armazenamento, criando um canal de comunicação com esse dispositivo
  - Supoem que o dispositivo está fechado e mais nenhum canal de comunicação para esse dispositivo está aberto
  - O dispositivo de armazenamento tem de existir
  - O dispositivo de armazenamento tem de ter um tamanho múltiplo do block size

16.2.1

```
void soCloseRawDisk (void)
```

- Fecha o dispositivo de armazenamento e o canal de comunicação.

16.2.2

```
void soReadRawBlock(uint32_t n, void *buf)
```

- Lê um bloco de dados do dispositivo
- **Parametros:**
  - *n*: número físico do bloco de dados no disco de onde a informação vai ser lida
  - *buf*: ponteiro para o buffer para onde os dados vão ser lidos

16.2.3

```
void soWriteRawBlock ( uint32_t  n, void * buf)
```

- Escreve um bloco de dados do dispositivo
- **Parametros:**
  - *n*: número físico do bloco de dados no disco onde a informação vai ser escrita
  - *buf*: ponteiro para o buffer que contém os dados a ser escritos # msksofs
- Script responsável por formatar o disco
- Cria um disco utilizável
  - Manipula os blocos do disco para implementar o *sofs17 filesystem*

16.3 Utilização

```
USAGE:
Sinopsis: mksofs [OPTIONS] supp-file
  OPTIONS:
   -n name --- set volume name (default: "sofs17_disk")
   -i num  --- set number of inodes (default: N/8, where N = number of blocks)
   -z      --- set zero mode (default: not zero)
   -q      --- set quiet mode (default: not quiet)
   -h      --- print this help
```

- Posso usar mais do que uma das opções na mesma execução ## Exemplos

16.3.1 No Options

- Basta indicar o número do ficheiro
- Por default, o número de inodes é o número de clusters/8

```
./mksofs ../disk.sofs17

Trying to install a 125-inodes SOFS17 file system in ../disk.sofs17.
Computing disk structure...
Filling in the superblock fields...
Filling in the table of inodes...
Filling in the bitmap of free clusters...
Filling in the root directory...
144-inodes SOFS17 file system was successfully installed in ../disk.sofs17.
```

16.3.2 Set name

```
$ ./mksofs.bin64 disk.sofs17 -n "my disk"

Trying to install a 125-inodes SOFS17 file system in disk.sofs17.
  Computing disk structure... done.
  Filling in the superblock fields... done.
  Filling in the table of inodes... done.
  Filling in the bitmap of free clusters... done.
  Filling in the root directory... done.
A 144-inodes SOFS17 file system was successfully installed in disk.sofs17.

$ ./showblock disk.sofs17 -s 0
Header:
   Magic number: 0x50F5
   Version number: 0x2017
   Volume name: my disk
   Properly unmounted: yes
   Number of mounts: 0
   Total number of blocks in the device: 1000
Inode table metadata:
```

```
20     First block of the inode table: 1
21     (...)
```

16.3.3 Set inodes

- O formatador tenta formatar o disco para o número desejado de inodes

```
1  ./mksofs.bin64 disk.sofs17 -i 2000
2
3  Trying to install a 2000-inodes SOFS17 file system in disk.sofs17.
4    Computing disk structure... done.
5    Filling in the superblock fields... done.
6    Filling in the table of inodes... done.
7    Filling in the bitmap of free clusters... done.
8    Filling in the root directory... done.
9  A 2000-inodes SOFS17 file system was successfully installed in disk.sofs17.
```

- Pode não ser possível formatar o disco para o número desejado de inodes
- Nesse caso, o formatador usa o número de inodes possível imediatamente superior ao pretendido

```
1  ./mksofs.bin64 disk.sofs17 -i 100
2
3  Trying to install a 100-inodes SOFS17 file system in disk.sofs17.
4    Computing disk structure... done.
5    Filling in the superblock fields... done.
6    Filling in the table of inodes... done.
7    Filling in the bitmap of free clusters... done.
8    Filling in the root directory... done.
9  A 112-inodes SOFS17 file system was successfully installed in disk.sofs17.
```

16.3.4 Zero Mode

- Ao usar a opção `-z` todos os clusters livres são preenchidos com zeros

```
1  ./mksofs ../disk.sofs17 -z
2
3  Trying to install a 125-inodes SOFS17 file system in ../disk.sofs17.
4  Computing disk structure...
5  Filling in the superblock fields...
6  Filling in the table of inodes... .
7  Filling in the bitmap of free clusters...
8  Filling in the root directory...
9  Filling in free clusters with zeros... cstart: 24, ctotal: 244
10 A 144-inodes SOFS17 file system was successfully installed in ../disk.sofs17.
```

17 computeStruture

- Calcula a divisão da estruturas no disco

  - número de clusters
  - número de blocos para inodes
  - número de blocos para reference map
-

17.1 Algoritmo

- No mínimo têm de existir 6 blocos no disco
  - 1 superblock
  - 0 inodes
  - 1 reference map
  - 1 cluster de dados
- Por default o número de inodes é $N_{inodes} = \frac{N_{clusters}}{8}$
- Caso o número de inodes não seja divisível por 8, é preciso alocar mais um bloco para os inodes
- O número temporário de blocos livres (falta o reference map) é:

$$N_{blocosdisco} - N_{blocosinodes} - 1$$

  - o "1" corresponde ao superblock
- O número de clusters é o resultado da divisão do número de blocos livres pelo número de blocos por cluster
- Através do número de clusters pode ser estimado o número de blocos necessários para a reference map
- Depois dessa estimativa é possível calcular o número de blocos restantes e atribuí-los à inode table

17.2 Utilização

```
1 void computeStructure(  uint32_t    ntotal,
2                         uint32_t    itotal,
3                         uint32_t *  itsizep,
4                         uint32_t *  rmsizep,
5                         uint32_t *  ctotalp
6                       )
```

17.2.1 Parameters

- **ntotal:** total number of blocks of the device
- **itotal:** requested number of inodes
- **itsizep:** pointer to mem where to store the size of inode table in blocks
- **rmsizep:** pointer to mem where to store the size of cluster reference table in blocks
- **ctotalp:** pointer to mem where to store the number of clusters

17.3 Testes

17.3.1 1000 blocos, 125 inodes (nblocos/8)

- Começamos por calcular o número de blocos necessários para os inodes

  - Existem 8 inodes por bloco

```
125 / 8
120   15
  5
```

- Obtemos 15 blocos para inodes
- E 5 blocos que sobram
- Se permitimos que 4 sejam usados para um cluster, temos 16 inodes
- O número de clusters é $1000 blocos - 16 inodes - 1 superblock = 983 blocos$
- O número de clusters para dados é:

```
983 / 4
980   245
  3
```

- Passamos a ter um sistema de ficheiros $15 + 3 = 18 blocos para inodes$
  - Isto equivale a ter $18 \times 8 = 144 inodes$ e não os 125 como inicialmente se desejava

18 fillInSuperBlock

- Preenche os campos dos superblock
- O *magic number* deve ser 0xFFFF
- As caches estão no *superblock*

18.1 Algoritmo

- Atribuições a serem feitas:
  - o magic number (identifica se o sistema é Big-Endian ou Little-Endian)
  - *version number*
  - nome do disco
    * Tem de ser truncado caso ultrapasse o *PARTITION_NAME_SIZE*
  - Número total de blocos
- Indicar que o disco ainda está unmounted
- Reset ao número de mounts
- Inode table metadata
  - **itstart:** Bloco onde começa a tabela de inodes
  - **itsize:** Número de blocos da inode table
  - **itotal:** Número total de inodes
  - **ifree:** Número de inodes livres
  - **ihead:** Índice para a head do primeiro inode
- Free Cluster table metadata

  - **rmstart:** bloco onde começa a reference map
  - **rmsize:** número de blocos usados pela reference table
  - **rmidx:** Primeira referência (*root dir*)
- Clusters metadata
  - **czstart:** bloco onde começa a cluster zone
  - **ctotal:** número total de clusters
  - **cfree:** número de clusters livres
- Retrieval cache
  - Inicializar com NullReferences
  - idx -> última posição da cache
- Insertion cache
  - Inicializar com NullReferences
  - idx -> primeira posição da cache

18.2 Utilização

```
void fillInSuperBlock(  const char * name,
                        uint32_t     ntotal,
                        uint32_t     itsize,
                        uint32_t     rmsize
                      )
```

18.2.1 Parameters

- **name:** volume name
- **ntotal:** the total number of blocks in the device
- **itsize:** the number of blocks used by the inode table
- **rmsize:** the number of blocks used by the cluster reference table

19 fillInInodeTable

- Preenche os blocos da inode table
- O inode **0** deve ser preenchido considerando que está a ser usado pelo diretório raiz
- Todos os outros inodes estão livres

19.1 Algoritmia

- Para cada bloco da inode table
  - Criar a lista biligada
  - Referências para os clusters:
    * Preencher com NullReference as *direct references (d)*

    * Preencher com NullReference as *indirect references (i1)*
    * Preencher com NullReference as *double direct references (i2)*

- Inicializar o inode da root directory (*inode 0*)
  - **mode:** permissões
  - **lnkcnt:** link count - número de caminhos que chegam a este (2: `.`, `..`)
  - **owner:**
  - **group:**
  - **size:** Tamanho do inode (1 cluster)
  - **clucnt:** file size in bytes
  - Modificar access times
  - Apontar para o root dir usando as referências diretas (_d[0])

19.2 Utilização

```
1 void fillInInodeTable(  uint32_t itstart,
2                         uint32_t itsize
3                       )
```

19.2.1 Parameters

- **itstart:** physical number of the first block used by the inode table
- **itsize:** number of blocks of the inode table

20 fillInFreeClusterTable

- Preeche a Free Cluster Table:
  - Estrutura que indica se os clusters estão livres ou ocupados
- Existe uma correspondência unívoca entre bits na *reference cluster table* e clusters no disco
- Os bits na tabela de referências crescem da:
  - Lower blocks to upper blocks
  - Lower bytes to upper bytes
  - Most Significant Bytes (MSB) to Least Significant Bytes (LSB)
- Assim, o bit "0" corresponde ao bit mais significativo do primeiro byte do primeiro bloco da tabela de bitmap
- Valor do bit:
  - **"1":** Cluster Livre
  - **"0":** Cluster Ocupado
- Em geral, o número de bits na tabela é maior que o número de clutsers
  - Os bits não usados (ou seja, que não correspondem ao estado de nenhum cluster) devem ser inicializados como se fosse usados
- Pode existir mais do que um bloco para a reference map table

20.1 Algoritmo

- Começa-se por definir algumas constantes:

```
#define CLUSTER_IN_USE 0
#define CLUSTER_FREE   1

#define BYTE_FREE   0xFF
#define BYTE_IN_USE 0x00

#define ROOT_DIR_MAP_MASK 0x7F
```

- Calcula-se:
  - Número de Clusters que não ficam referenciados num bloco completo

$$nExtraClusters = c_{total}\%ReferencesPerBitmapBlock$$

  - Número de Clusters da reference zone que estão totalmente ocupados

$$nFullRefBlocks = \frac{c_{total}}{ReferencesPerBitmapBlock}$$

  - Número de Blocos para a Reference Bitmap zone

$$nRefBlocks = nFullRefBlocks + (nExtraClusters! = 0)$$

  - Byte no último bloco de referências onde começam as referências não válidas

$$byteStartFreeBitmapPos = \frac{nExtraClusters}{8}$$

  - Bit no byte acima onde começam as referências não válidas

$$bitStartFreeBitmapPos = nExtraClusters\%8$$

- Blocos de referências completos
  - Para cada bloco de referências completo o número de referências é ReferencesPerBitmapBlock
  - O indice é 0 (o primeiro cluster livre está no início do bloco)
- Bloco parcialmente completo
  - Preencher o cnt com o número de clusters que esse bloco tem
  - Colocar os clusters não válidos como usados
- Referência do cluster 0
  - Indicar que está a ser usada pela root dir
  - Decrementar o count, porque existe menos um cluster vazio

20.1.1 Considerações

- O número de clusters pode ser inferior ou superior ao tamanho de um bloco.
- O primeiro bit do primeiro bloco deve estar em use
- Todos os bits que não referenciem um cluster devem ser colocados como em uso

20.2 Utilização

```
void fillInFreeClusterTable(uint32_t rmstart,
                            uint32_t ctotal
                           )
```

20.2.1 Parameters

- **rmstart:** the number of the fisrt block used by the bit table
- **ctotal:** the total number of clusters

20.2.2 Data Structure

SORefBlock : estrutura dos Reference bitmap Block data type

```
struct SORefBlock
{
    /** \brief number of references in block */
    uint16_t cnt;
    /** \brief index of first non-empty byte */
    uint16_t idx;
    /** \brief bit map */
    uint8_t map[ReferenceBytesPerBitmapBlock];
};
```

20.3 Testes

```
# 1000 blocks, 144-inodes, mksofs.bin64
block range: 19
cnt = 244, idx = 0
0000: 7f ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
    ff ff f8 00
0020: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0040: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0060: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0080: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
00a0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
00c0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
00e0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0100: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
```

```
0120: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0140: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0160: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0180: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
01a0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
01c0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
01e0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00


# 200 blocks, 48-inodes, mksofs.bin64
block range: 7
cnt = 47, idx = 0
0000: 7f ff ff ff ff ff 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0020: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0040: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0060: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0080: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
00a0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
00c0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
00e0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0100: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0120: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0140: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0160: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
0180: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
01a0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
01c0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
    00 00 00 00
01e0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00


# 10000 blocks, 1264 inodes, mksofs17.bin64
```

```
block range: 159
cnt = 2459, idx = 0
0000: 7f ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
0020: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
0040: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
0060: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
0080: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
00a0: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
00c0: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
00e0: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
0100: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff
0120: ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff f0 00 00 00 00 00 00 00 00 00 00 00
0140: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
0160: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
0180: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
01a0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
01c0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
01e0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
```

21 fillInRootDir

- A root dir ocupa um cluster
- Os dois primeiros slots estão reservados para as entradas “.” e “..”
- Todos os os outros slots devem estar limpos:
  - o campo *name* preenchido com zeros
  - o campo *inode* preenchido com NullReference

21.1 Algoritmia

- Colocar todos os blocos com zeros
- Na entrada 0
  - nome: “.”

  - inode: 0 (raiz)

- Na entrada 1
  - nome: ".."
  - inode: 0 (raiz)

21.2 Utilização

```
1 void fillInRootDir(uint32_t rtstart)
```

21.2.1 Parameters

- **rtstart:** number of the block where the root cluster starts.

22 resetClusters

- Escrever com zeros um conjunto de clusters

22.1 Algoritmia

- Percorrer a sequência de clusters desejada e escrever zeros

22.2 Utilização

```
1 void resetClusters( uint32_t cstart,
2                     uint32_t ctotal
3                   )
```

22.2.1 Parameters

- **cstart:** number of the block of the first free cluster
- **ctotal:** number of clusters to be filled

23 freelists

- Funções para manipular a lista de inodes livres e a lista de clusters livres.
- A lista de inodes livres é mantida usando uma lista biligada de inodes
  - Política FIFO
- A lista de clusters é mantida com duas caches:
  - Retrieval Cache: Clusters livres para serem alocados

  - Insertion Cache: Clusters que foram libertados
- A lista de clusters segue uma estrutura parecida com FIFO
  - Não é bem FIFO porque existe rotatividade nos clusters
    * Impede que exista uma escrita desigual nos clusters
    * Antes de um cluster libertado puder voltar a ser escrito, todos os outros clusters no disco têm de ser escritos
    * Aumenta o tempo útil do disco

23.1 soAllocInode

uint32_t soAllocInode (uint32_t type)

Allocate a free inode.

An inode is retrieved from the list of free inodes, marked in use, associated to the legal file type passed as a parameter and is generally initialized.

Parameters

```
1       type the inode type (it must represent either a file, or a
2            directory, or a symbolic link)
```

Returns the number of the allocated inode

23.2 soFreeCluster

void soFreeCluster (uint32_t cn)

Free the referenced cluster.

Parameters

```
1        cn the number of the cluster to be freed
```

23.3 soFreeInode

void soFreeInode (uint32_t in)

Free the referenced inode.

The inode is inserted into the list of free inodes.

Parameters

```
1        in number of the inode to be freed
```

23.4 soReplenish

void soReplenish ( )

replenish the retrieval cache

References to free clusters should be transfered from the free cluster table (bit map) or insertion cache to the retrieval cache. Nothing should be done if the retrieval cache is not empty. The insertion cache should only be used if there are no bits at one in the map. Only a single block should be processed, even if it is not enough to fulfill the retrieval cache. The block to be processes is the one pointed to by the rmidx field of the superblock. This field should be updated if the processing of the current block reaches its end.

23.5 soDeplete

[MISING IN DOXYGEN]

Functions

uint32_t soAllocInode (uint32_t type) Allocate a free inode. More…

```
1    void  soFreeInode (uint32_t in)
2          Free the referenced inode. More...
```

uint32_t soAllocCluster () Allocate a free cluster. More…

```
1    void  soReplenish ()
2          replenish the retrieval cache More...
3
4    void  soFreeCluster (uint32_t cn)
5          Free the referenced cluster. More...
6
7    void  soDeplete ()
8          Deplete the insertion cache.
```

Detailed Description

Functions to manage the list of free inodes and the list of free clusters.

24 Cenário Inicial

campo ihead superblock: 101 campo ifree : 3

inode previous|next

101 7|5

5 10|7

7 5/101

24.1 freeinode

- Quero libertar o nó numero 200
- mante o tipo
- apenas altera flag para free
- o ficheiro passa a ser deleted file
- o nó que é libertado é obviamente o ultimo

```
# Estado inicial
----------------
Inode #3
type =  regular file, permissions = rw-rw-r--, lnkcnt = 1, owner = 1000, group = 1000
size in bytes = 42000, size in clusters = 7
atime = Thu Oct 26 23:02:47 2017, mtime = Thu Oct 26 23:02:47 2017, ctime = Thu Oct 26
    23:02:47 2017
d[] = {11 (nil) (nil) (nil) 12 13}, i1 = 14, i2 = (nil)
----------------
Após chamar o freeinode para o inode 3
---------------
Inode #3
type = free regular file, permissions = rw-rw-r--, lnkcnt = 1, owner = 1000, group = 1000
size in bytes = 42000, size in clusters = 7
next = 10, , prev = 1
d[] = {11 (nil) (nil) (nil) 12 13}, i1 = 14, i2 = (nil)
----------------
- Mantem o size in clusters
```

24.2 inserir inode 200

200 7|101

101 200|5

5 10/7

7 5|200

200 7|101

ihead:101 ifree: 4

24.3 soAllocateInode

- Retirar o nó da lista de inodes
- O primeiro no a retirar é o que apontado pelo ihead

ihead:5 ifree:2

5 7|7

7 5|5

Só pode existir uma cópia do suoerbloco

24.4 iOpen

- Abrir o inode
  - so o volta a ler do disco se já foi aberto
  - só posso pedir um pointer para esse inode
  - devolve me um inode handler
  - in: inode number
  - ih: inode handler
  - ip: inode pointer

24.5 iSave

- Guardar um inode

24.6 iClose

- Guardar o ficheiro

24.7 Interface com os inodes é suposto usar uma estrutura de inodes

função replentish tem como função transferir blocos para a cache - transforma bits em referências - os bits que forem transferidos vão passar de 1 a zero - rmidx: primeiro byte do map de bit que tem bits a 1 - comece por aqui. Antes não há bits a 1

25 mais difícis (5)

soReplenish (Patricia) soDeplete (Bernardo)

26 intermédias (3)

soAllocateInode (panda) soFreeInode (Gradim)

27 mais triviais (1)

soAllocClusters (Pedro) soFreeClusters (Mica) # AllocInode - Aloca um inode livre da estrutura de inodes -

27.1 Utilização

```
1 uint32_t soAllocInode ( uint32_t  type )
```

Allocate a free inode.

An inode is retrieved from the list of free inodes, marked in use, associated to the legal file type passed as a parameter and is generally initialized.

Parameters

```
1       type the inode type (it must represent either a file, or a directory, or a symbolic
            link)
```

Returns the number of the allocated inode

- Se for possível, aloca o inode que está na HEAD
- Incrementa a HEAD
  - tem de verificar se a inode table está no fim e se tem de voltar aos inodes que entretanto foram libertados
- Decrementa o número de inodes livres
- O inode tem de ser corretamente inicializado # Inodes
- Existem 6 posições para referência direta aos clusters do ficheiro
  - d[0 … 5]
- Uma posição para referência indireta
  - i1
  - Extende o array de d[6 … 517]
-

27.2 Uma posição para referência dupla indireta

- Cada ficheiro possui um inode
  - O número máximo de ficheiros num disco é o número máximo de inodes
- Um inode ocupa 64 bytes
  - Logo num disco com 512 bytes por bloco, existem 8 inodes em cada bloco # Fileclusters Functions to manage the clusters belonging by a file

27.3 Doxygen

uint32_t soGetFileCluster (int ih, uint32_t fcn) Get the cluster number of a given file cluster. More…

uint32_t soAllocFileCluster (int ih, uint32_t fcn) Associate a cluster to a given file cluster position. More…

```
1   void  soFreeFileClusters (int ih, uint32_t ffcn)
2         Free all file clusters from the given position on.
3         More...
4
5   void  soReadFileCluster (int ih, uint32_t fcn, void *buf)
6         Read a file cluster. More...
```

```
7
8   void  soWriteFileCluster (int ih, uint32_t fcn, void *buf)
9         Write a data cluster. More...
```

Detailed Description

Functions to manage the clusters belonging by a file.

Author Artur Pereira - 2008-2009, 2016-2017 Miguel Oliveira e Silva - 2009, 2017 António Rui Borges - 2010-2015

Remarks In case an error occurs, every function throws an SOException

Function Documentation

uint32_t soAllocFileCluster (int ih, uint32_t fcn)

Associate a cluster to a given file cluster position.

Parameters

```
1       ih  inode handler
2       fcn file cluster number
```

Returns the number of the allocated cluster

void soFreeFileClusters (int ih, uint32_t ffcn)

Free all file clusters from the given position on.

Parameters

```
1       ih   inode handler
2       ffcn first file cluster number
```

27.3.1 uint32_t soGetFileCluster (int ih,

```
1                           uint32_t  fcn
2                         )
```

Get the cluster number of a given file cluster.

Parameters

```
1       ih  inode handler
2       fcn file cluster number
```

Returns the number of the corresponding cluster

- Parece me que apenas tenho de retornar o endereço do cluster
- Não é preciso retornar tudo
- Mandar mail ao professor
- Perguntar panda
- Para que servem as fnções do prof??
- Aquilo que faz é usar o inode handler para saber onde está no disk e o file cluster number para obter a referência
- É preciso rever as duas estruturas

  - superblock
  - inode
  - cluster

O que é preciso fazer:

- Obter o inode
- Se o cluster index estiver nos 6 primeiros
  - Sai direto da estrutura de inodes
- Se o cluster index for referenciado diretamente (i_1)
  - está no cluster de referências
  - Ler esse cluster do disco
  - Calcular novo index (subtrair 6?)
  - Ler Retornar a referência em que este está
- Se o cluste index estiver no cluster de referências indiretas (i_2)
  - Calcular dois novos indexes:
    * index no cluster de referências indiretas
    * index no cluster de referências diretas
  - Ler a referência do disco
  - Retornar o valor
- A testtool já trata de fazer o iOpen
- A soGetFileCluster é chamada com o indice do inode
- É preciso usar a `iGetPointer` para obter o ponteiro para a estrutura

Testes

+===+ | testing functions | +=============================
| q - exit | sb - show block | | fd - format disk | spd - set probe depths | +—————————————–+———————————————-+ | ai - alloc inode | fi - free inode | | ac - alloc cluster | fc - free cluster | | r - replenish | d - deplete | +—————————————–+———————————————-+ | gfc - get file cluster | afc - alloc file cluster | | ffc - free file clusters | - NOT USED | | rfc - read file cluster | wfc - write file cluster | +—————————————–+———————————————-+ | gde - get dir entry | ade - add dir entry | | rde - rename dir entry | dde - delete dir entry | | tp - traverse path | - NOT USED | +—————————————–+———————————————-+ + cia - check inode access | sia - set inode access + + iil - increment inode lnkcnt | dil - decrement inode lnkcnt + +===+

Your command: gfc Inode number: 1 File cluster index: 7 (711)–> iOpen(1) (711)–> –iOpenBin(1) (403)–> soGetFileCluster(0, 7) (403)–> –soGetFileClusterBin(0, 7) (712)–> iGetPointer(0) (712)–> –iGetPointerBin(0) (714)–> iClose(0) (714)–> –iCloseBin(0) Cluster number (nil) retrieved +===+ | testing functions | +===+ | q - exit | sb - show block | | fd - format disk | spd - set probe depths | +—————————————–+———————————————-+ | ai - alloc inode | fi - free inode | | ac - alloc cluster | fc - free cluster | | r - replenish | d - deplete | +—————————————–+———————————————-+ | gfc - get file cluster | afc - alloc file cluster | | ffc - free file clusters | - NOT USED | | rfc - read file cluster | wfc - write file cluster | +—————————————–+———————————————-+ | gde - get dir entry | ade - add dir entry | | rde - rename dir entry | dde - delete dir entry | | tp - traverse path | - NOT USED | +—————————————–+———————————————-+ + cia - check inode access | sia - set inode access + + iil - increment inode lnkcnt | dil - decrement inode lnkcnt + +===+

Your command: gfc Inode number: 1 File cluster index: 8 (711)–> iOpen(1) (711)–> –iOpenBin(1) (403)–> soGetFileCluster(0, 8) (403)–> –soGetFileClusterBin(0, 8) (712)–> iGetPointer(0) (712)–> –iGetPointerBin(0) (714)–> iClose(0) (714)–> –iCloseBin(0) Cluster number (nil) retrieved +===+ | testing functions | +===+ | q - exit | sb - show block | | fd - format disk | spd - set probe depths | +———————————–+———————————-+ | ai - alloc inode | fi - free inode | | ac - alloc cluster | fc - free cluster | | r - replenish | d - deplete | +———————————–+———————————-+ | gfc - get file cluster | afc - alloc file cluster | | ffc - free file clusters | - NOT USED | | rfc - read file cluster | wfc - write file cluster | +———————————–+———————————-+ | gde - get dir entry | ade - add dir entry | | rde - rename dir entry | dde - delete dir entry | | tp - traverse path | - NOT USED | +———————————–+———————————-+ + cia - check inode access | sia - set inode access + + iil - increment inode lnkcnt | dil - decrement inode lnkcnt + +===+

Your command: gfc Inode number: 0 File cluster index: 1 (711)–> iOpen(0) (711)–> –iOpenBin(0) (851)–> sbGetPointer() (851)–> –sbGetPointerBin() (951)–> soReadRawBlock(1, 0x7fffd273b450) (403)–> soGetFileCluster(1, 1) (403)–> –soGetFileClusterBin(1, 1) (712)–> iGetPointer(1) (712)–> –iGetPointerBin(1) (714)–> iClose(1) (714)–> –iCloseBin(1) Cluster number (nil) retrieved +===+ | testing functions | +===+ | q - exit | sb - show block | | fd - format disk | spd - set probe depths | +———————————–+———————————-+ | ai - alloc inode | fi - free inode | | ac - alloc cluster | fc - free cluster | | r - replenish | d - deplete | +———————————–+———————————-+ | gfc - get file cluster | afc - alloc file cluster | | ffc - free file clusters | - NOT USED | | rfc - read file cluster | wfc - write file cluster | +———————————–+———————————-+ | gde - get dir entry | ade - add dir entry | | rde - rename dir entry | dde - delete dir entry | | tp - traverse path | - NOT USED | +———————————–+———————————-+ + cia - check inode access | sia - set inode access + + iil - increment inode lnkcnt | dil - decrement inode lnkcnt + +===+

27.4 Your command:

27.4.1 void soReadFileCluster (int ih,

```
                        uint32_t  fcn,
                        void *    buf
                      )
```

Read a file cluster.

Data is read from a specific data cluster which is supposed to belong to an inode associated to a file (a regular file, a directory or a symbolic link).

If the referred file cluster has not been allocated yet, the returned data will consist of a byte stream filled with the character null (ascii code 0).

Parameters

```
      ih  inode handler
      fcn file cluster number
      buf pointer to the buffer where data must be read into
```

void soWriteFileCluster (int ih, uint32_t fcn, void * buf)

Write a data cluster.

Data is written into a specific data cluster which is supposed to belong to an inode associated to a file (a regular file, a directory or a symbolic link).

If the referred cluster has not been allocated yet, it will be allocated now so that the data can be stored as its contents.

Parameters

```
ih  inode handler
fcn file cluster number
buf pointer to the buffer containing data to be written
```

27.5 soFreeFileCLusters

- Liberta todos os clusters do inode começando na posição atual
- Se o inode ficar sem clusters, é apagado # direntries

28 soGetDirEntry

- Obtem o inode associado ao nome da função
- É preciso fazer o parse do nome do diretório para chegar ao diretório pretendido
- Chama a traverse Path
- Tem de verificar se a entrada já existe

uint32_t soGetDirEntry (int pih, const char * name)

Get the inode associated to the given name.

The directory contents, seen as an array of directory entries, is parsed to find an entry whose name is name.

The name must also be a base name and not a path, that is, it can not contain the character "/".

Parameters

```
pih  inode handler of the parent directory
name the name entry to be searched for
```

Returns the corresponding inode number

29 soRenameDirEntry

- Renomeia a entrada de um diretório

void soRenameDirEntry (int pih, const char * name, const char * newName)

Rename an entry of a directory.

A direntry associated from the given directory is renamed.

Parameters

```
pih     inode handler of the parent inode
name    current name of the entry
newName new name for the entry
```

30 soTraversePath

- Obtem o inode associado com um dado caminho
- Atravessa a estrutura do sistema do sistema de ficheiros para obter o inode cujo nome do ficheiro é a componente mais à direita do caminho
- O caminho deve ser absoluto
- Todos elementos do caminho (com exceção do último) devem ser diretório ou symbolic links com permissão de travers (x)

uint32_t soTraversePath (char * path)

Get the inode associated to the given path.

The directory hierarchy of the file system is traversed to find an entry whose name is the rightmost component of path. The path is supposed to be absolute and each component of path, with the exception of the rightmost one, should be a directory name or symbolic link name to a path.

The process that calls the operation must have execution (x) permission on all the components of the path with exception of the rightmost one.

Parameters

```
1       path the path to be traversed
```

Returns the corresponding inode number

31 soAddDirEntry

- Adiciona uma nova entrada no diretório pai
- Uma direntry é adicionada ligando o parent inode ao child inode
  - O lnkcnt do inode filho **não** é incrementado nesta função

void soAddDirEntry (int pih, const char * name, uint32_t cin)

Add a new entry to the parent directory.

A direntry is added connecting the parent inode to the child inode. The refcount of the child inode is not incremented by this function.

Parameters

```
1       pih  inode handler of the parent inode
2       name name of the entry
3       cin  number of the child inode
```

32 soDeleteDirEntry

- Remove uma entrada do parent directory
- O lnkcnt do inode filho **não** é decrementado

uint32_t soDeleteDirEntry (int pih, const char * name, bool clean = false)

Remove an entry from a parent directory.

A direntry associated from the given directory is deleted. The refcount of the child inode is not decremented by this function.

Parameters

```
pih   inode handler of the parent inode
name  name of the entry
clean if true (different than zero) clean the corresponding dir entry, otherwise
    keep it dirty
```

Returns the inode number in the deleted entry

33 Extra

filesystem check - atribui ficheiros para a o disco sempre que uma função falahe gera uma exceção

34 soRenameDirEntry

- dá um novo nome ao diretório

35 soDeleteDirEntry

- remove o diretório

36 soGetDirEntry

- quando alguém a nível superior quer abrir/escrever, ver permissões precisa de sabe o inode # itdealer
- Conjunto de funções que manipulam diretamente a estrutura de inodes

36.1 iOpen

- Abre um inode
  - Transfere o seu conteudo para a memória
  - Set ao *usecount*
- Caso o inode já esteja aberto, incrementa a *usecount*
- Em qualquer dos casos devolve o handler (referência para a posição de memória) para o inode # Syscalls

36.2 Main syscalls

- **soLInk:** Cria um link para um ficheiro
- **soMkdir:** Cria um diretório
- **soMknod:** Cria um ficheiro regular com tamanho nulo
- **soRead:** Lê os dados de um ficheiro regular previamente aberto
- **soReaddir:** Lê uma entrada para um diretório de um dado diretório
- **soReadLink:** Lê um *symbolic link*
- **soRename:** Muda um nome de um ficheiro ou a sua localização na estrutura de diretórios
- **soRmdir:** Remover um diretório
  - O diretório de ve estar vazio
- **soSymlink:** Cria um symbolic link com o caminho desejado
- **soTruncate:** Trunca o tamanho de um regular file para o desejado
- **soUnlink:** Remove um link para um ficheiro através de um diretório
  - Remove também o ficheiro se o lnkcnt = 0
- **soWrite:** Escreve dados num regular file previamente aberto

36.3 Other syscalls

- **exceptions** : sofs17 exception definition module "'cpp struct SOException:public std::exception int en; ///< (system) error number const char *msg; ///< name of function that has thrown the exception

digraph x{ a -> b [label="b"] a -> c [label="a"] a -> d [label="d"] } # Existem duas camadas de syscalls - main syscalls - 12 funções (temos de saber para o mini teste) - other syscalss

36.4 soLink

- Usar a transverse path para saber qual o nó que está na ponta
- link("/b", "a/c")
  - saber qual é o nó que está na pomnta do /b
    * uso o traverse para saber
    * abro e pergunto para saber o tipo
    * base name
    * verifico se é o diretório e tenho permissoes de excrita
    * verifico se ja tem o ficheirp que quero criar
    * chamar idirentry para criar o diretorio
    * chamar increment link count

36.5 unLink

- Não apaga o ficheiro
- Quebra a ligaão
- dde - delete dir entry
- decrementa o link count

- se o tiver 0 links
  - chama a free inode para libertar o inode
  - chama a free cluster para libertar o cluster
- APgar ficheiros é derivado do unlink

Enquanto o ficheiro estive aberto não pode ser destruido É o close do sistema operativo que apaga um fucheiro

36.6 soRename

- função complicada
- soRename(“/b”, “/a/c”)
- OU é um rename se o novo pathe e o path antigo orem iguais

```
1     soRename(”/b”, ”/c”)
```

- Equivale no caso do move a fazer delete direentry e add direntry
- Não tem o link nem o dec
- O nó de destino passa a ser o mesmo

36.7 soMKnod

- Cria um nod do tipo ficheiro
- COrrepsonde a fazer:
  - Começar por criar um inode: alloc indode,
  - add dir entry
  - increment link count
- Tem de validar primeiro:
  - Verificar se o “/a” existe, é um diretório e tem permisssoes de escrita
  - Verificar se o “/c” não existe

36.8 soRead

- Posso quer ler dois bytes e ter de ler dois clusters
  - Ultimo byte do 1º cluster
  - Primeiro byte do 2º cluster
- A função read não pode ler para a lém do fim de ficheiro
- O size é que determina o fim do ficheiroo
- Indiretamento o write també, pode alocar clusters
- Tipicamente o wrtie tem de ler priemiro caos vá alterar parcialemtne um cluster
- O que interessa em termos de miniteste é o papel e efeito da função, não como o ocódigo é feito
- O que interessa é a consequemncia da execuação de u comando

36.9 soTrucnate

- Alçtera o tamanho de um gicheiro ou para cima ou para baixo
- È assim que se cria buracos num ficheiro
  - Trunco e vou acrescentando
- Gunção joga com o size e
- Trunco o tamanho para 10
- Depois volto a troncar para 20
  - Ou no truncar para cima ou no truncar para baixo tenho de garantir que nºao existe lixo entre as zonas dos meus dados
    * Ou escrevo zeros., ou escrevo NullReferences
      · Null References dentro do size são lidas como zeros
  - O ficheiro é o mesmo, estou só a alterar o tamanho que esse ficheiro tem no disco
  - O truncate não quer saber o que lá está, simplesmente trunca os dados inteirores
  - Ou eu ponho zeros quando encolhi, ou ponho zeros quando abro
- Se fize ro fopne de um ficheiro já abero, o SO chama a truncate e mete os dados desse ficheiro a zero

36.10 soMkdir

- Sempre ue há u novo direentry é preciso adicionar o linkcount
- Pressuposto: só se pode aoagar diretórios vazios
- Ler man2 para saber os erros que tem de emitir

36.11 soReadDIr

- É usado para fazer o ls
- Leê entradas de um diretorio
- Sempre que esta funlção é lida, ele vai ler a próxima direntry
- Em cadainvocaçáo eu tenho que lhe dzwr qenaotos bytes já passei
- Posso ter de processar duas entradas para lhe dar uma . QUando a fubnção rreaddir devolve zero, já não existem mais entradas naquele deiretorio

36.12 soSymlink

- Creates a symbolic link

36.13 so ReadLink

- Valor de retorno do symbolic link # HOW to use sofs17 (so1718 - Aula prática 29 Sep) ## Documentação

```
# Gerar documentação
cd ./doc
doxygen
```

A documentação fica na pasta `./doc/html/`

37 Make

```
# O make compila sempre tudo e não somente o conteudo da pasta
make
make -C <path_to_start>     % indica o caminho onde o make começar
```

Na linkagem necessita da biblioteca fuse.h. Está contida na biblioteca libfuse-dev que pode ser instalada com:

```
sudo apt-get install libfuse-dev
```

[TODO] mksofs - msksofs : formatador para o sistema de ficheiros sofs17

Para já as funções

- compute structure:
  - nao altera os dados no disco.
  - Apenas calcula os blocos de inodes, clusters, etc.
- cada função vai preencher a àrea do disco respetiva
  - **fillInSuperBlock** : computes the structural division of the disk
  - **fillInInodeTable** :

38 soFreeFileCLusters

- Apagamento da esquerda para a direita
- As posições são alteradas e escritas com Null Reference
- o size só é alterado por syscalls e não ao libertar um inode/cluster

39

- Um diretório tem um size múltiplo do cluster
- Os diretórios crescem cluster a cluster

40

uint32_t soGetDirEntry (int pih, const char * name)

Get the inode associated to the given name.

The directory contents, seen as an array of directory entries, is parsed to find an entry whose name is name.

The name must also be a base name and not a path, that is, it can not contain the character “/”.

Parameters

```
      pih  inode handler of the parent directory
      name the name entry to be searched for
```

Returns the corresponding inode number

41

uint32_t soDeleteDirEntry (int pih, const char * name, bool clean = false)

Remove an entry from a parent directory.

A direntry associated from the given directory is deleted. The refcount of the child inode is not decremented by this function.

Parameters

```
pih   inode handler of the parent inode
name  name of the entry
clean if true (different than zero) clean the corresponding dir entry, otherwise
    keep it dirty
```

Returns the inode number in the deleted entry

Comentários - soWriteRawBlock - char blk[blockSize] - SOSuperblock sb; - SOTnode it[inodesPerBlock] - soWriteRawBlock(uint32_t n, void *buf) - blk - fsb - it

41.1 Notes

função alloc cluster tem de verificar se ficou tudo bem no disco função replentish transfer da reference bitmao block para a retrieval cache

Capacidade: $(6 + 2^9 + (2^9)^2) \cdot 2^1 1$

42 3 Nov 2017

- As direntries não mexem no lnkcnt
- Add mexe no size

43 Unlink

1. dde
2. dec
3. if(dec ==0) 3.1 ffc 3.2 fi

O dec devolve o devolve

44 Remove

45 mtime vs ctime

- **mtime:** conteudo do ficheiro
- **ctime:** metadados do ficheiro

- Não podemos ter nenhum diretório apontado por dois diretórios

```
ln: 'ddd/': hard link not allowwd for directory
```

46 Conceitos Introdutórios

Num ambiente multiprogramado, os processos podem ser:

- Independentes:
  - Nunca interagem desde a sua criação à sua destruição
  - Só possuem uma interação implícita: **competir por recursos do sistema**
    - e.g.: jobs num sistema bacth, processos de diferentes utilizadores
  - É da responsabilidade do sistema operativo garantir que a atribuição de recursos é feita de forma controlada
    - É preciso garantir que não ocorre perda de informação
    - Só **um processo pode usar um recurso num intervalo de tempo** - ***Mutual Exclusive Access***
- Cooperativos:
  - **Partilham Informação** e/ou **Comunicam** entre si
  - Para **partilharem** informação precisam de ter acesso a um **espaço de endereçamento comum**
  - A comunicação entre processos pode ser feita através de:
    - Endereço de memória comum
    - Canal de comunicação que liga os processos
  - É da **responsabilidade do processo** garantir que o acesso à zona de memória partilhada ou ao canal de comunicação é feito de forma controlada para não ocorrerem perdas de informação
    - Só **um processo pode usar um recurso num intervalo de tempo** - ***Mutual Exclusive Access***
    - Tipicamente, o canal de comunicação é um recurso do sistema, pelo quais os **processos competem**

O acesso a um recurso/área partilhada é efetuada através de código. Para evitar a perda de informação, o código de acesso (também denominado zona crítica) deve evitar incorrer em **race conditions**.

46.1 Exclusão Mútua

Ao forçar a ocorrência de exclusão mútua no acesso a um recurso/área partilhada, podemos originar:

- **deadlock:**
  - Vários processos estão em espera **eternamente** pelas condições/eventos que lhe permitem aceder à sua respetiva **zona crítica**
    - Pode ser provado que estas condições/eventos **nunca se irão verificar**
  - Causa o bloqueio da execução das operações
- **starvation:**
  - Na competição por acesso a uma zona crítica por vários processos, verificam-se um conjunto de circunstâncias na qual novos processos, com maior prioridade no acesso às suas zonas críticas, continuam a aparecer e **tomar posse dos recursos partilhados**
  - O acesso dos processos mais antigos à sua zona crítica é sucessivamente adiado

47 Acesso a um Recurso

No acesso a um recurso é preciso garantir que não ocorrem **race conditions**. Para isso, **antes** do acesso ao recurso propriamente dito é preciso **desativar o acesso** a esse recurso pelos **outros processos** (reclamar *ownership*) e após o acesso é preciso restaurar as condições iniciais, ou seja, **libertar o acesso** ao recurso.

```c
/* processes competing for a resource - p = 0, 1, ..., N-1 */
void main (unsigned int p)
{
    forever
    {
        do_something();
        access_resource(p);
        do_something_else();
    }
}

void acess_resource(unsigned int p)
{
    enter_critical_section(p);
    use_resource();     //  critical section
    leave_critical_section(p);
}
```

48 Acesso a Memória Partilhada

O acesso à memória partilhada é muito semelhante ao aceso a um recurso (podemos ver a memória partilhada como um recurso partilhado entre vários processos).

Assim, à semelhança do acesso a um recurso, é preciso **bloquear o acesso de outros processos à memória partilhada** antes de aceder ao recurso e após aceder, **reativar o acesso a memória partilhada** pelos outros processos.

```c
/* shared data structure */
shared DATA d;

/* processes sharing data - p = 0, 1, ..., N-1 */
void main (unsigned int p)
{
    forever
    {
        do_something();
        access_shared_area(p);
        do_something_else();
    }
}

void access_shared_area(unsigned int p)
{
    enter_critical_section(p);
    manipulate_shared_area(); // critical section
    leave_critical_section(p);
}
```

48.1 Relação Produtor-Consumidor

O acesso a um recurso/memória partilhada pode ser visto como um problema Produtor-Consumidor:

- Um processo acede para **armazenar dados**, **escrevendo** na memória partilhada *(Produtor)*
- Outro processo acede para **obter dados**, **lendo** da memória partilhada *(Consumidor)*

48.1.1 Produtor

O produtor "produz informação" que quer guardar na FIFO e enquanto não puder efetuar a sua escrita, aguarda até puder **bloquear** e **tomar posse** do zona de memória partilhada

```
/* communicating data structure: FIFO of fixed size */
shared FIFO fifo;

/* producer processes - p = 0, 1, ..., N-1 */
void main (unsigned int p)
{
    DATA val;
    bool done;


    forever
    {
        produce_data(&val);
        done = false;
        do
        {
            // Beginning of Critical Section
            enter_critical_section(p);
            if (fifo.notFull())
            {
                fifo.insert(val);
                done = true;
            }
            leave_critical_section(p);
            // End of Critical Section
        } while (!done);
        do_something_else();
    }
}
```

48.1.2 Consumidor

O consumidor quer ler informação que precisa de obter da FIFO e enquanto não puder efetuar a sua leitura, aguarda até puder **bloquear** e **tomar posse** do zona de memória partilhada

```
/* communicating data structure: FIFO of fixed size */
shared FIFO fifo;

```

```
4  /* consumer processes - p = 0, 1, ..., M-1 */
5  void main (unsigned int p)
6  {
7  DATA val;
8  bool done;
9      forever
10     {
11         done = false;
12         do
13         {
14             // Beginning of Critical Section
15             enter_critical_section(p);
16             if (fifo.notEmpty())
17             {
18                 fifo.retrieve(&val);
19                 done = true;
20             }
21             leave_critical_section(p);
22             // End of Critical Section
23         } while (!done);
24         consume_data(val);
25         do_something_else();
26     }
27 }
```

49 Acesso a uma Zona Crítica

Ao aceder a uma zona crítica devem ser verificados as seguintes condições:

- **Efective Mutual Exclusion:** O **acesso** a uma **zona crítica** associada com o mesmo recurso/memória partilhada só pode ser **permitida a um processo de cada vez** entre **todos os processos** a competir pelo acesso a esse mesmo recurso/memória partilhada
- **Independência** do número de processos intervenientes e na sua velocidade relativa de execução
- Um processo fora da sua zona crítica não pode impedir outro processo de entrar na sua zona crítica
- Um processo **não deve ter de esperar indefinidamente** após pedir acesso ao recurso/memória partilhada para que possa aceder à sua zona crítica
- O período de tempo que um processo está na sua **zona crítica** deve ser **finito**

49.1 Tipos de Soluções

Para controlar o acesso às zonas críticas normalmente é usado um endereço de memória. A gestão pode ser efetuada por:

- **Software:**
  - A solução é baseada nas instruções típicas de acesso à memória
  - Leitura e Escrita são independentes e correspondem a instruções diferentes
- **Hardware:**
  - A solução é baseada num conjunto de instruções especiais de acesso à memória
  - Estas instruções permitem ler e de seguida escrever na memória, de forma **atómica**

49.2 Alternância Estrita *(Strict Alternation)*

Não é uma solução válida

- Depende da velocidade relativa de execução dos processos intervenientes
- O processo com menos acessos impõe o ritmo de acessos aos restantes processos
- Um processo fora da zona crítica não pode prevenir outro processo de entrar na sua zona crítica
- Se não for o seu turno, um processo é obrigado a esperar, mesmo que não exista mais nenhum processo a pedir acesso ao recurso/memória partilhada

```
/* control data structure */
#define R       /* process id = 0, 1, ..., R-1 */

shared unsigned int access_turn = 0;
void enter_critical_section(unsigned int own_pid)
{
    while (own_pid != access_turn);
}

void leave_critical_section(unsigned int own_pid)
{
    if (own_pid == access_turn)
        access_turn = (access_turn + 1) % R;
}
```

49.3 Eliminar a Alternância Estrita

```
/* control data structure */
#define R 2     /* process id = 0, 1 */

shared bool is_in[R] = {false, false};

void enter_critical_section(unsigned int own_pid)
{
    unsigned int other_pid_ = 1 - own_pid;

    while (is_in[other_pid]);
    is_in[own_pid] = true;
}

void leave_critical_section(unsigned int own_pid)
{
    is_in[own_pid] = false;
}
```

Esta solução não é válida porque não garante **exclusão mútua**.

Assume que:

- P_0 entra na função `enters_critical_section` e testa `is_in[1]`, que retorna Falso

- P_1 entra na função `enter_critical_section` e testa `is_in[0]`, que retorna Falso
- P_1 altera `is_in[0]` para *true* e entra na zona crítica
- P_0 altera `is_in[1]` para *true* e entra na zona crítica

Assim, ambos os processos entra na sua zona crítica **no mesmo intervalo de tempo**.

O principal problema desta implementação advém de **testar primeiro** a variável de controlo do **outro processo** e só **depois** alterar a **sua variável** de controlo.

49.4 Garantir a exclusão mútua

```
1  /* control data structure */
2  #define R 2     /* process id = 0, 1 */
3
4  shared bool want_enter[R] = {false, false};
5
6  void enter_critical_section(unsigned int own_pid)
7  {
8      unsigned int other_pid_ = 1 - own_pid;
9
10     want_enter[own_pid] = true;
11     while (want_enter[other_pid]);
12 }
13
14 void leave_critical_section(unsigned int own_pid)
15 {
16     want_enter[own_pid] = false;
17 }
```

Esta solução, apesar de **resolver a exclusão mútua**, **não é válida** porque podem ocorrer situações de **deadlock**.

Assume que:

- P_0 entra na função `enter_critical_section` e efetua o *set* de `want_enter[0]`
- P_1 entra na função `enter_critical_section` e efetua o *set* de `want_enter[1]`
- P_1 testa `want_enter[0]` e, como é *true*, **fica em espera** para entrar na zona crítica
- P_0 testa `want_enter[1]` e, como é *true*, **fica em espera** para entrar na zona crítica

Com **ambos os processos em espera** para entrar na zona crítica e **nenhum processo na zona crítica** entramos numa situação de **deadlock**.

Para resolver a situação de deadlock, **pelo menos um dos processos** tem recuar na intenção de aceder à zona crítica.

49.5 Garantir que não ocorre deadlock

```
1  /* control data structure */
2  #define R 2     /* process id = 0, 1 */
3
4  shared bool want_enter[R] = {false, false};
5
6  void enter_critical_section(unsigned int own_pid)
```

```
{
    unsigned int other_pid_ = 1 - own_pid;

    want_enter[own_pid] = true;
    while (want_enter[other_pid])
    {
        want_enter[own_pid] = false;    // go back
        random_dealy();
        want_enter[own_pid] = true;     // attempt a to go to the critical section
    }
}

void leave_critical_section(unsigned int own_pid)
{
    want_enter[own_pid] = false;
}
```

A solução é quase válida. Mesmo um dos processos a recuar ainda é possível ocorrerem situações de **deadlock** e **starvation**:

- Se ambos os processos **recuarem ao "mesmo tempo"** (devido ao `random_delay()` ser igual), entramos numa situação de **starvation**
- Se ambos os processos **avançarem ao "mesmo tempo"** (devido ao `random_delay()` ser igual), entramos numa situação de **deadlock**

A solução para **mediar os acessos** tem de ser **determinística** e não aleatória.

49.6 Mediar os acessos de forma determinística: *Dekker agorithm*

```
/* control data structure */
#define R 2     /* process id = 0, 1 */

shared bool want_enter[R] = {false, false};
shared uint p_w_priority = 0;

void enter_critical_section(unsigned int own_pid)
{
    unsigned int other_pid_ = 1 - own_pid;

    want_enter[own_pid] = true;
    while (want_enter[other_pid])
    {
        if (own_pid != p_w_priority)            // If the process is not the priority
            process
        {
            want_enter[own_pid] = false;        // go back
            while (own_pid != p_w_priority);    // waits to access to his critical section
                 while
                                                //  its is not the priority process
            want_enter[own_pid] = true;         // attempt to go to his critical section
        }
    }
```

```
}

void leave_critical_section(unsigned int own_pid)
{
    unsigned int other_pid_ = 1 - own_pid;
    p_w_priority = other_pid;                   // when leaving the its critical section,
        assign the
                                                // priority to the other process
    want_enter[own_pid] = false;
}
```

É uma **solução válida**:

- Garante exclusão mútua no acesso à zona crítica através de um mecanismo de alternância para resolver o conflito de acessos
- **deadlock** e **starvation não estão presentes**
- Não são feitas suposições relativas ao tempo de execução dos processos, i.e., o algoritmo é **independente** do tempo de execução dos processos

No entanto, **não pode ser generalizado** para mais do que 2 processos e garantir que continuam a ser satisfeitas as condições de **exclusão mútua** e a ausência de **deadlock** e **starvation**

49.7 Dijkstra algorithm (1966)

```
/* control data structure */
#define R 2     /* process id = 0, 1 */

shared uint want_enter[R] = {NO, NO, ..., NO};
shared uint p_w_priority = 0;

void enter_critical_section(uint own_pid)
{
    uint n;
    do
    {
        want_enter[own_pid] = WANT;                 // attempt to access to the critical
            section
        while (own_pid != p_w_priority)             // While the process is not the
            priority process
        {
            if (want_enter[p_w_priority] == NO)     // Wait for the priority process to
                leave its critical section
                p_w_priority = own_pid;
        }

        want_enter[own_pid] = DECIDED;              // Mark as the next process to access
            to its critical section

        for (n = 0; n < R; n++)             // Search if another process is already
            entering its critical section
        {
```

```
23              if (n != own_pid && want_enter[n] == DECIDED)   // If so, abort attempt to
                    ensure mutual exclusion
24                  break;
25          }
26      } while(n < R);
27  }
28
29  void leave_critical_section(unsigned int own_pid)
30  {
31      p_w_priority = (own_pid + 1) % R;           // when leaving the its critical section,
            assign the
32                                                  // priority to the next process
33      want_enter[own_pid] = false;
34  }
```

Pode sofrer de **starvation** se quando um processo iniciar a saída da zona crítica e alterar `p_w_priority`, atribuindo a prioridade a outro processo, outro processo tentar aceder à zona crítica, sendo a sua execução interrompida no for. Em situações "especiais", este fenómeno pode ocorrer sempre para o mesmo processo, o que faz com que ele nunca entre na sua zona crítica

49.8 Peterson Algorithm (1981)

```
1   /* control data structure */
2   #define R 2     /* process id = 0, 1 */
3
4   shared bool want_enter[R] = {false, false};
5   shared uint last;
6
7   void enter_critical_section(uint own_pid)
8   {
9       unsigned int other_pid_ = 1 - own_pid;
10
11      want_enter[own_pid] = true;
12      last = own_pid;
13      while ( (want_enter[other_pid]) && (last == own_pid) );     // Only enters the
            critical section when no other
14                                                                  // process wants to enter
                                                                        and the last request
15                                                                  // to enter is made by the
                                                                        current process
16  }
17
18  void leave_critical_section(unsigned int own_pid)
19  {
20      want_enter[own_pid] = false;
21  }
```

O algoritmo de *Peterson* usa a **ordem de chegada** de pedidos para resolver conflitos:

- Cada processo tem de **escrever o seu ID numa variável partilhada** (*last*), que indica qual foi o último processo a pedir para entrar na zona crítica

- A **leitura seguinte** é que vai determinar qual é o processo que foi o último a escrever e portanto qual o processo que deve entrar na zona crítica

	P_0 quer entrar		P_1 quer entrar	
	P_1 não quer entrar	P_1 quer entrar	P_0 não quer entrar	P_0 quer entrar
last = P_0	P_0 entra	P_1 entra	-	P_1 entra
last = P_1	-	P_0 entra	P_1 entra	P_0 entra

É uma solução válida que:

- Garante exclusão mútua
- Previne deadlock e starvation
- É independente da velocidade relativa dos processos
- Pode ser generalizada para mais do que dois processos (variável partilhada -> fila de espera)

49.9 Generalized Peterson Algorithm (1981)

```
/* control data structure */
#define R ...   /* process id = 0, 1, ..., R-1 */

shared bool want_enter[R] = {-1, -1, ..., -1};
shared uint last[R-1];

void enter_critical_section(uint own_pid)
{
    for (uint i = 0; i < R -1; i++)
    {
        want_enter[own_pid] = i;

        last[i] = own_pid;

        do
        {
            test = false;
            for (uint j = 0; j < R; j++)
            {
                if (j != own_pid)
                    test = test || (want_enter[j] >= i)
            }
        } while ( test && (last[i] == own_pid) );       // Only enters the critical
            section when no other
                                                                // process wants to enter
                                                                    and the last request
                                                                // to enter is made by the
                                                                    current process
    }
}

void leave_critical_section(unsigned int own_pid)
```

```
30 {
31     want_enter[own_pid] = -1;
32 }
```

> *needs clarification*

50 Soluções de Hardware

50.1 Desativar as interrupções

Num ambiente computacional com **um único processador:**

- A alternância entre processos, num ambiente **multiprogramado**, é sempre causada por um evento/dispositivo externo
  - **real time clock (RTC):** origina a transição de time-out em sistemas *preemptive*
  - **device controller:** pode causar transições *preemptive* no caso de um fenómeno de *wake up* de um **processo mais prioritário**
  - Em qualquer dos casos, o **processador é interrompido** e a execução do processo atual parada
- A garantia de acesso em **exclusão mútua** pode ser feita desativando as interrupções
- No entanto, só pode ser efetuada em **modo kernel**
  - Senão código malicioso ou com *bugs* poderia bloquear completamente o sistema

Num ambiente computacional **multiprocessador**, desativar as interrupções num único processador não tem qualquer efeito.

Todos os outro processadores (ou *cores*) continuam a responder às interrupções.

50.2 Instruções Especiais em Hardware

50.2.1 Test and Set (TAS primitive)

A função de hardware, `test_and_set` se for implementada atomicamente (i.e., sem interrupções) pode ser utilizada para construir a primitiva **lock**, que permite a entrada na zona crítica

Usando esta primitiva, é possível criar a função *lock*, que permite entrar na zona crítica

```
1  shared bool flag = false;
2
3  bool test_and_set(bool * flag)
4  {
5      bool prev = *flag;
6      *flag = true;
7      return prev;
8  }
9
10 void lock(bool * flag)
11 {
12     while (test_and_set(flag);  // Stays locked until and unlock operation is used
13 }
```

```
14
15 void unlock(bool * flag)
16 {
17     *flag = false;
18 }
```

50.2.2 Compare and Swap

Se implementada de forma atómica, a função `compare_and_set` pode ser usada para implementar a primitiva lock, que permite a entrada na zona crítica

O comportamento esperado é que coloque a variável a 1 sabendo que estava a 0 quando a função foi chamada e vice-versa.

```
1  shared int value = 0;
2
3  int compare_and_swap(int * value, int expected, int new_value)
4  {
5      int v = *value;
6      if (*value == expected)
7          *value = new_value;
8      return v;
9  }
10
11 void lock(int * flag)
12 {
13     while (compare_and_swap(&flag, 0, 1) != 0);
14 }
15
16 void unlock(bool * flag)
17 {
18     *flag = 0;
19 }
```

50.3 Busy Waiting

Ambas as funções anteriores são suportadas nos *Instruction Sets* de alguns processadores, implementadas de forma atómica

No entanto, ambas as soluções anteriores sofrem de **busy waiting**. A primitiva lock está no seu **estado ON** (usando o CPU) **enquanto espera** que se verifique a condição de acesso à zona crítica. Este tipo de soluções são conhecidas como **spinlocks**, porque o processo oscila em torno da variável enquanto espera pelo acesso

Em sistemas **uniprocessor**, o **busy_waiting** é **indesejado** porque causa:

- **Perda de eficiência:** O **time quantum** de um processo está a ser desperdiçado porque não está a ser usado para nada
- ** Risco de deadlock**: Se um** processo mais prioritário** tenciona efetuar um **lock** enquanto um processo menos prioritário está na sua zona crítica, **nenhum deles pode prosseguir**.
  - O processo menos prioritário tenta executar um unlock, mas não consegue ganhar acesso a um *time qantum* do CPU devido ao processo mais prioritário
  - O processo mais prioritário não consegue entrar na sua zona crítica porque o processo menos prioritário ainda não saiu da sua zona crítica

Em sistemas **multiprocessador** com **memória partilhada**, situações de busy waiting podem ser menos críticas, uma vez que a troca de processos *(preempt)* tem custos temporais associados. É preciso:

- guardar o estado do processo atual
  - variáveis
  - stack
  - $PC
- copiar para memória o código do novo processo

50.4 Block and wake-up

Em **sistemas uniprocessor** (e em geral nos restantes sistemas), existe a o requerimento de **bloquear um processo** enquanto este está à espera para entrar na sua zona crítica

A implementação das funções `enter_critical_section` e `leave_critical_section` continua a precisar de operações atómicas.

```
#define  R ... /* process id = 0, 1, ..., R-1 */

shared unsigned int access = 1;     // Note that access is an integer, not a boolean

void enter_critical_section(unsigned int own_pid)
{
    // Beginning of atomic operation
    if (access == 0)
        block(own_pid);

    else access -= 1;
    // Ending of atomic operation
}

void leave_critical_section(unsigned int own_pid)
{
    // Beginning of atomic operation
    if (there_are_blocked_processes)
        wake_up_one();
     else access += 1;
    // Ending of atomic operation
}
```

```
/* producers - p = 0, 1, ..., N-1 */
void producer(unsigned int p)
{
    DATA data;
    forever
    {
        produce_data(&data);
        bool done = false;
        do
        {
```

```
            lock(p);
            if (fifo.notFull())
            {
                fifo.insert(data);
                done = true;
            }
        unlock(p);
    } while (!done);
    do_something_else();
    }
}
```

```
/* consumers - c = 0, 1, ..., M-1 */
void consumer(unsigned int c)
{
    DATA data;
    forever
    {
        bool done = false;
        do
        {
            lock(c);
            if (fifo.notEmpty())
            {
                fifo.retrieve(&data);
                done = true;
            }
            unlock(c);
        } while (!done);
        consume_data(data);
        do_something_else();
    }
}
```

51 Semáforos

No ficheiro `IPC.md` são indicadas as condições e informação base para:

- Sincronizar a entrada na zona crítica
- Para serem usadas em programação concorrente
- Criar zonas que garantam a exclusão mútua

Semáforos são **mecanismos** que permitem por implementar estas condições e **sincronizar a atividade** de **entidades concorrentes em ambiente multiprogramado**

Não são nada mais do que **mecanismos de sincronização**.

51.1 Implementação

Um semáforo é implementado através de:

- Um tipo/estrutura de dados
- Duas operações **atómicas**:
  - down (ou wait)
  - up (ou signal/post)

```
typedef struct
{
    unsigned int val;   /* can not be negative */
    PROCESS *queue;     /* queue of waiting blocked processes */
} SEMAPHORE;
```

51.1.1 Operações

As únicas operações permitidas são o **incremento**, up, ou **decremento**, down, da variável de controlo. A variável de controlo, `val`, **só pode ser manipulada através destas operações!**

Não existe uma função de leitura nem de escrita para `val`.

- `down`
  - **bloqueia** o processo se `val == 0`
  - **decrementa** `val` se `val != 0`
- `up`
  - Se a `queue` não estiver vazia, **acorda** um dos processos
  - O processo a ser acordado depende da **política implementada**
  - **Incrementa** `val` se a `queue` estiver vazia

51.1.2 Solução típica de sistemas *uniprocessor*

```
/* array of semaphores defined in kernel */
#define R /* semid = 0, 1, ..., R-1 */

static SEMAPHORE sem[R];

void sem_down(unsigned int semid)
{
    disable_interruptions;
    if (sem[semid].val == 0)
        block_on_sem(getpid(), semid);
    else
        sem[semid].val -= 1;
    enable_interruptions;
}

void sem_up(unsigned int semid)
{
    disable_interruptions;
    if (sem[sem_id].queue != NULL)
        wake_up_one_on_sem(semid);
```

```
21      else
22          sem[semid].val += 1;
23      enable_interruptions;
24  }
```

A solução apresentada é típica de um sistema *uniprocessor* porque recorre à diretivas **disable_interruptions** e **enable_interruptions** para garantir a exclusão mútua no acesso à zona crítica.

Só é possível garantir a exclusão mútua nestas condições se o sistema só possuir um único processador, porque as diretivas irão impedir a interrupção do processo que está na posse do processador devido a eventos externos. Esta solução não funciona para um sistema multiprocessador porque ao executar a diretiva **disable_interruptions**, só estamos a **desativar as interrupções para um único processador**. Nada impede que noutro processador esteja a correr um processo que vá aceder à mesma zona de memória partilhada, não sendo garantida a exclusão mútua para sistemas multiprocessador.

Uma solução alternativa seria a extensão do **disable_interruptions** a todos os processadores. No entanto, iriamos estar a impedir a troca de processos noutros processadores do sistema que poderiam nem sequer tentar aceder às variáveis de memória partilhada.

51.2 Bounded Buffer Problem

```
1   shared FIFO fifo; /* fixed-size FIFO memory */
2
3   /* producers - p = 0, 1, ..., N-1 */
4   void producer(unsigned int p)
5   {
6       DATA data;
7       forever
8       {
9           produce_data(&data);
10          bool done = false;
11          do
12          {
13              lock(p);
14              if (fifo.notFull())
15              {
16                  fifo.insert(data);
17                  done = true;
18              }
19              unlock(p);
20          } while (!done);
21      do_something_else();
22      }
23  }
24
25  /* consumers - c = 0, 1, ..., M-1 */
26  void consumer(unsigned int c)
27  {
28      DATA data;
29      forever
30      {
31          bool done = false;
```

```
32          do
33          {
34              lock(c);
35              if (fifo.notEmpty())
36              {
37                  fifo.retrieve(&data);
38                  done = true;
39              }
40              unlock(c);
41          } while (!done);
42          consume_data(data);
43          do_something_else();
44      }
45  }
```

51.2.1 Como Implementar usando semáforos?

A solução para o *Bounded-buffer Problem* usando semáforos tem de:

- Garantir **exclusão mútua**
- Ausência de busy waiting

```
1   shared FIFO fifo;   /*fixed-size FIFO memory */
2   shared sem access;  /*semaphore to control mutual exclusion */
3   shared sem nslots;  /*semaphore to control number of available slots*/
4   shared sem nitems;  /*semaphore to control number of available items */
5
6
7   /* producers - p = 0, 1, ..., N-1 */
8   void producer(unsigned int p)
9   {
10      DATA val;
11
12      forever
13      {
14          produce_data(&val);
15          sem_down(nslots);
16          sem_down(access);
17          fifo.insert(val);
18          sem_up(access);
19          sem_up(nitems);
20          do_something_else();
21      }
22  }
23
24  /* consumers - c = 0, 1, ..., M-1 */
25  void consumer(unsigned int c)
26  {
27      DATA val;
28
29      forever
```

```
30      {
31          sem_down(nitems);
32          sem_down(access);
33          fifo.retrieve(&val);
34          sem_up(access);
35          sem_up(nslots);
36          consume_data(val);
37          do_something_else();
38      }
39  }
```

Não são necessárias as funções `fifo.empty()` e `fifo.full()` porque são implementadas indiretamente pelas variáveis:

- **nitens:** Número de "produtos" prontos a serem "consumidos"
  - Acaba por implementar, indiretamente, a funcionalidade de verificar se a FIFO está empty
- **nslots:** Número de slots livres no semáforo. Indica quantos mais "produtos" podem ser produzidos pelo "consumidor"
  - Acaba por implementar, indiretamente, a funcionalidade de verificar se a FIFO está full

Uma alternativa **ERRADA** a uma implementação com semáforos é apresentada abaixo:

```
1   shared FIFO fifo;   /*fixed-size FIFO memory */
2   shared sem access;  /*semaphore to control mutual exclusion */
3   shared sem nslots;  /*semaphore to control number of available slots*/
4   shared sem nitems;  /*semaphore to control number of available items */
5
6
7   /* producers - p = 0, 1, ..., N-1 */
8   void producer(unsigned int p)
9   {
10      DATA val;
11
12      forever
13      {
14          produce_data(&val);
15          sem_down(access);       // WRONG SOLUTION! The order of this
16          sem_down(nslots);       // two lines are changed
17          fifo.insert(val);
18          sem_up(access);
19          sem_up(nitems);
20          do_something_else();
21      }
22  }
23
24  /* consumers - c = 0, 1, ..., M-1 */
25  void consumer(unsigned int c)
26  {
27      DATA val;
28
29      forever
30      {
31          sem_down(nitems);
```

```
32          sem_down(access);
33          fifo.retrieve(&val);
34          sem_up(access);
35          sem_up(nslots);
36          consume_data(val);
37          do_something_else();
38      }
39  }
```

A diferença entre esta solução e a anterior está na troca de ordem de instruções `sem_down(access)` e `sem_down(nslots)`. A função `sem_down`, ao contrário das funções anteriores, **decrementa** a variável, não tenta decrementar.

Assim, o produtor tenta aceder à sua zona crítica sem primeiro decrementar o número de slots livres para ele guardar os resultados da sua produção *(needs_clarification)*

51.3 Análise de Semáforos

51.3.1 Vantagens

- **Operam ao nível do sistema operativo:**
  - As operações dos semáforos são implementadas no *kernel*
  - São disponibilizadas aos utilizadores através de *system_calls*
- São **genéricos** e **modulares**
  - por serem implementações de baixo nível, ganham **versatilidade**
  - Podem ser usados em qualquer tipo de situação de programação concorrente

51.3.2 Desvantagens

- Usam **primitivas de baixo nível**, o que implica que o programador necessita de conhecer os **princípios da programação concorrente**, uma vez que são aplicadas numa filosofia *bottom-up* - Facilmente ocorrem **race conditions** - Facilmente se geram situações de **deadlock**, uma vez que **a ordem das operações atómicas são relevantes**
- São tanto usados para implementar **exclusão mútua** como para **sincronizar processos**

51.3.3 Problemas do uso de semáforos

Como tanto usados para implementar **exclusão mútua** como para **sincronizar processos**, se as condições de acesso não forem satisfeitas, os processos são bloqueados **antes** de entrarem nas suas regiões críticas.

- Solução sujeita a erros, especialmente em situações complexas
  - pode existir **mais do que um ponto de sincronismos** ao longo do programa

51.4 Semáforos em Unix/Linux

POSIX:

- Suportam as operações de `down` e `up`

  - `sem_wait`
  - `sem_trywait`
  - `sem_timedwait`
  - `sem_post`

- Dois tipos de semáforos:
  - **named semaphores:**
    * São criados num sistema de ficheiros virtual (e.g. /dev/sem)
    * Suportam as operações:
      · `sem_open`
      · `sem_close`
      · `sem_unlink`
  - **unnamed semaphores:**
    * São *memory based*
    * Suportam as operações
      · `sem_init`
      · `sem_destroy`

System V:

- Suporta as operações:
  - `semget` : criação
  - `semop` : as diretivas `up` e `down`
  - `semctl` : outras operações

52 Monitores

Mecanismo de sincronização de alto nível para resolver os problemas de sincronização entre processos, numa perspetiva **top-down**. Propostos independentemente por Hoare e Brinch Hansen

Seguindo esta filosofia, a **exclusão mútua** e **sincronização** são tratadas **separadamente**, devendo os processos:

1. Entrar na sua zona crítica
2. Bloquear caso não possuam condições para continuar

Os monitores são uma solução que suporta nativamente a exclusão mútua, onde uma aplicação é vista como um conjunto de *threads* que competem para terem acesso a uma estrutura de dados partilhada, sendo que esta estrutura só pode ser acedida pelos métodos do monitor.

Um monitor assume que todos os seus métodos **têm de ser executados em exclusão mútua**:

- Se uma *thread* chama um **método de acesso** enquanto outra *thread* está a executar outro método de acesso, a sua **execução é bloqueada** até a outra terminar a execução do método

A sincronização entre threads é obtida usando **variáveis condicionais**:

- `wait`: A *thread* é bloqueada e colocada fora do monitor
- `signal`: Se existirem outras *threads* bloqueadas, uma é escolhida para ser "acordada"

52.1 Implementação

```
monitor example
{
/* internal shared data structure */
DATA data;

condition c; /* condition variable */

/* access methods */
method_1 (...)
{
    ...
}
method_2 (...)
{
    ...
}

...

/* initialization code */
...
```

52.2 Tipos de Monitores

52.2.1 Hoare Monitor

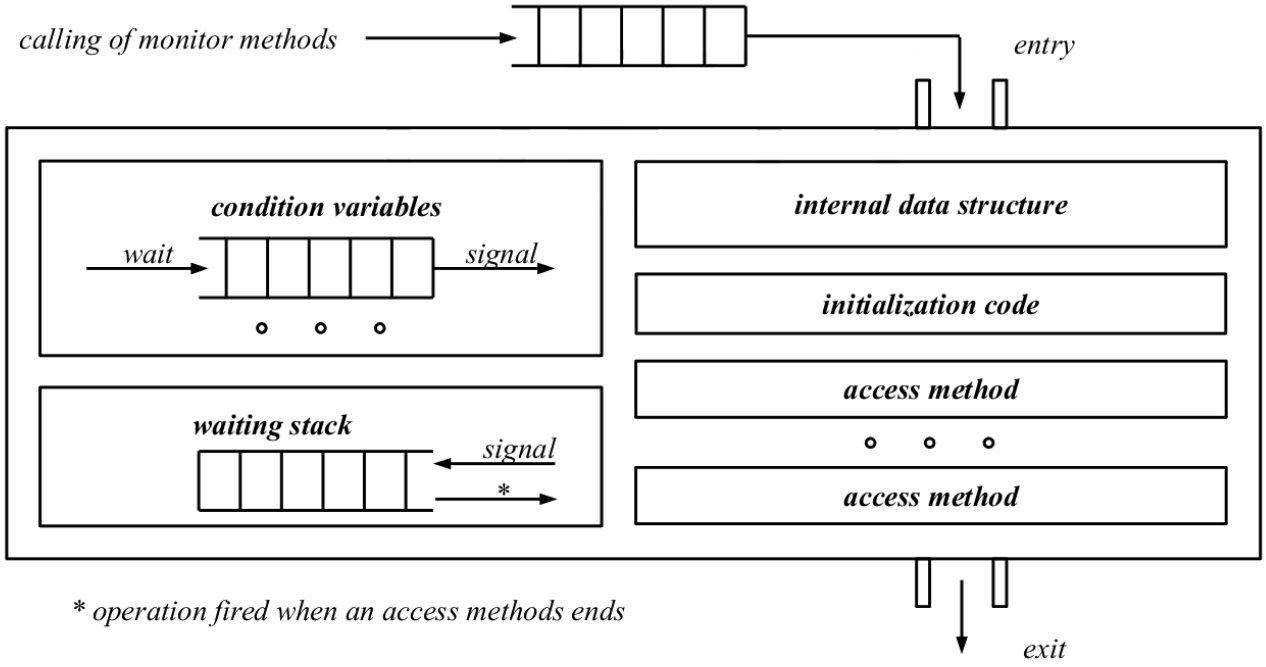


Figure 21: Diagrama da estrutura interna de um Monitor de Hoare

- Monitor de aplicação geral
- Precisa de uma stack para os processos que efetuaram um `wait` e são colocados em espera
- Dentro do monitor só se encontra a *thread* a ser executada por ele
- Quando existe um `signal`, uma *thread* é **acordada** e posta em execução

52.2.2 Brinch Hansen Monitor

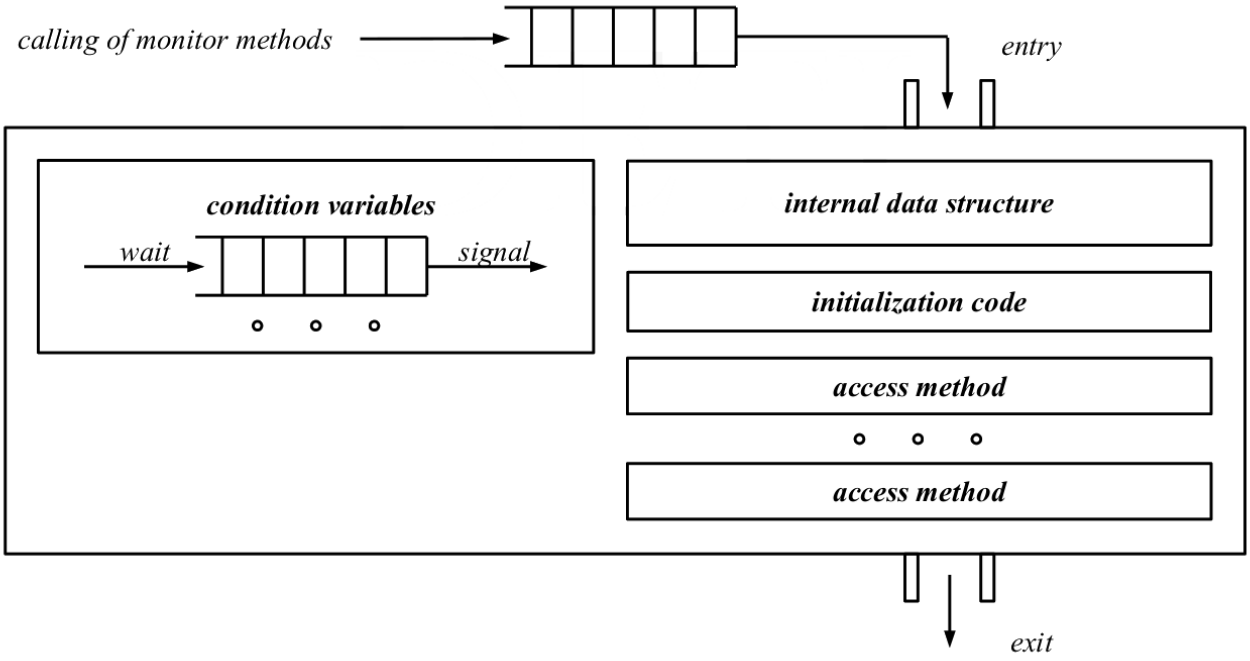


Figure 22: Diagrama da estrutura interna de um Monitor de Brinch Hansen

- A última instrução dos métodos do monitor é `signal`
  - Após o `signal` a *thread* sai do monitor
- **Fácil de implementar:** não requer nenhuma estrutura externa ao monitor
- **Restritiva: Obriga** a que cada método só possa possuir uma instrução de `signal`

52.2.3 Lampson/Redell Monitors

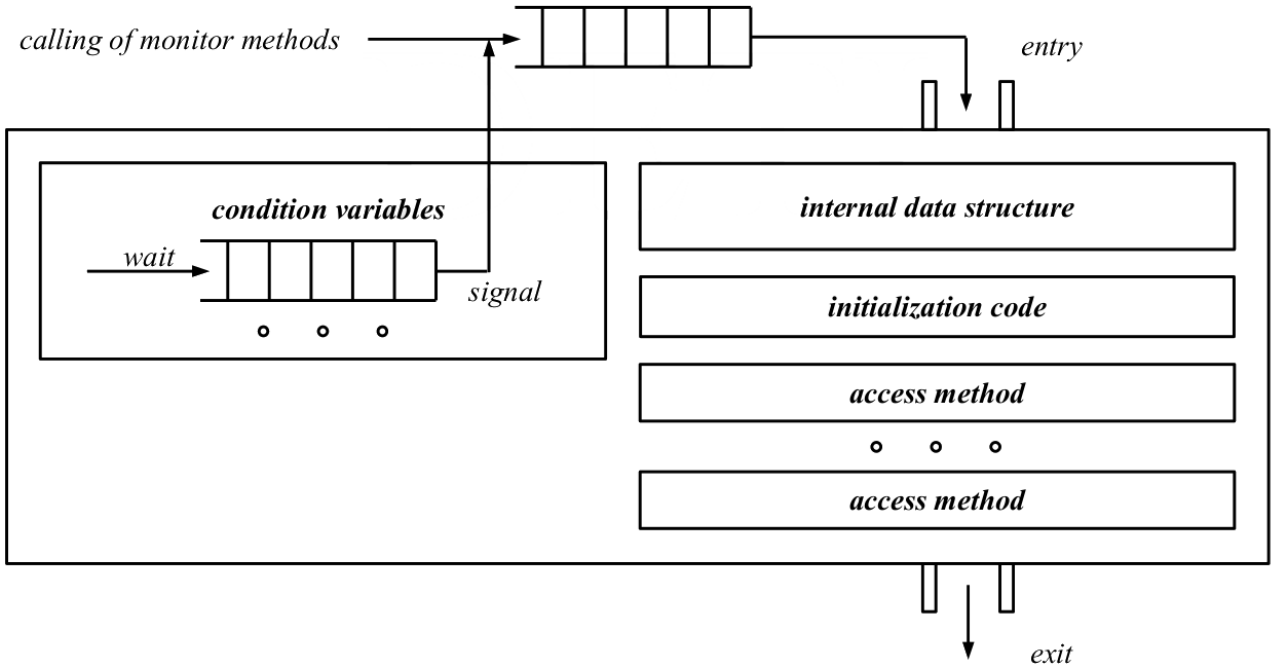


Figure 23: Diagrama da estrutura interna de um Monitor de Lampson/Redell

- A *thread* que faz o `signal` é a que continua a sua execução (entrando no monitor)
- A *thread* que é acordada devido ao `signal` fica fora do monitor, **competindo pelo acesso** ao monitor
- Pode causar **starvation**.
  - Não existem garantias que a **thread** que foi acordada e fica em competição por acesso vá ter acesso
  - Pode ser **acordada** e voltar a **bloquear**
  - Enquanto está em `ready` nada garante que outra *thread* não dê um `signal` e passe para o estado `ready`
  - A *thread* que tinha sido acordada volta a ser **bloqueada**

52.3 Bounded-Buffer Problem usando Monitores

```
1  shared FIFO fifo;            /* fixed-size FIFO memory */
2  shared mutex access;         /* mutex to control mutual exclusion */
3  shared cond nslots;      /* condition variable to control availability of slots*/
4  shared cond nitems;      /* condition variable to control availability of items */
5
6  /* producers - p = 0, 1, ..., N-1 */
7  void producer(unsigned int p)
8  {
9      DATA data;
10     forever
11     {
12         produce_data(&data);
13         lock(access);
```

```
14          if/while (fifo.isFull())
15          {
16              wait(nslots, access);
17          }
18          fifo.insert(data);
19          unlock(access);
20          signal(nitems);
21          do_something_else();
22      }
23  }
24
25  /* consumers - c = 0, 1, ..., M-1 */
26  void consumer(unsigned int c)
27  {
28      DATA data;
29      forever
30      {
31          lock(access);
32          if/while (fifo.isEmpty())
33          {
34              wait(nitems, access);
35          }
36          fifo.retrieve(&data);
37          unlock(access);
38          signal(nslots);
39          consume_data(data);
40          do_something_else();
41      }
42  }
```

O uso de `if`/`while` deve-se às diferentes implementações de monitores:

- `if`: **Brinch Hansen**
  - quando a *thread* efetua o `signal` sai imediatamente do monitor, podendo entrar logo outra *thread*
- `while`: **Lamson Redell**
  - A *thread* acordada fica à espera que a *thread* que deu o `signal` termine para que possa **disputar** o acesso
- O `wait` internamente vai **largar a exclusão mútua**
  - Se não larga a exclusão mútua, mais nenhum processo consegue entrar
  - Um wait na verdade é um `lock(..)` seguido de `unlock(...)`
- Depois de efetuar uma **inserção**, é preciso efetuar um `signal` do nitems
- Depois de efetuar um **retrieval** é preciso fazer um `signal` do nslots
  - Em comparação, num semáforo quando faço o up é sempre incrementado o seu valor
- Quando uma *thread* emite um `signal` relativo a uma variável de transmissão, ela só **emite** quando alguém está à escuta
  - O `wait` só pode ser feito se a FIFO estiver cheia
  - O `signal` pode ser sempre feito

É necessário existir a `fifo.empty()` e a `fifo.full()` porque as variáveis de controlo não são semáforos binários.

O valor inicial do **mutex** é 0.

52.4 POSIX support for monitors

A criação e sincronização de *threads* usa o *Standard POSIX, IEEE 1003.1c*.

O *standard* define uma API para a **criação** e **sincronização** de *threads*, implementada em Unix pela biblioteca *pthread*

O conceito de monitor **não existe**, mas a biblioteca permite ser usada para criar monitores *Lampsom/Redell* em C/C++, usando:

- `mutexes`
- `variáveis de condição`

As funções disponíveis são:

- `ptread_create`: **cria** uma nova *thread* (similar ao *fork*)
- `ptread_exit`: equivalente à `exit`
- `ptread_join`: equivalente à `waitpid`
- `ptread_self`: equivalente à `getpid`
- `pthread_mutex_*`: manipulação de **mutexes**
- `ptread_cond_*`: manipulação de **variáveis condicionais**
- `ptread_once`: inicialização

53 Message-passing

Os processos podem comunicar entre si usando **mensagens**.

- Não existe a necessidade de possuírem memória partilhada
- Mecanismos válidos quer para sistemas **uniprocessor** quer para sistemas **multiprocessador**

A **comunicação** é efetuada através de **duas operações**:

- `send`
- `receive`

Requer a existência de um **canal de comunicação**. Existem 3 implementações possíveis:

1. **Endereçamento direto/indireto**
2. Comunicação **síncrona/assíncrona**
    - Só o `sender` é que indica o **destinatário**
    - O destinatário **não indica** o `sender`
    - Quando existem **caixas partilhadas**, normalmente usam-se mecanismos com políticas de **round-robin**
        1. Lê o processo N
        2. Lê o processo $N + 1$
        3. etc…
    - No entanto, outros métodos podem ser usados
3. **Automatic or explicit buffering**

53.1 Direct vs Indirect

53.1.1 Symmetric direct communication

O processo que pretende comunicar deve **explicitar o nome do destinatário/remetente:**

- Quando o `sender` envia uma mensagem tem de indicar o **destinatário**
  - `send(P, message`
- O destinatário tem de indicar de quem **quer receber** (`sender`)
  - `receive(P, message)`

A comunicação entre os **dois processos** envolvidos é **peer-to-peer**, e é estabelecida automaticamente entre entre um conjunto de processos comunicantes, só existindo **um canal de comunicação**

53.2 Asymmetric direct communications

Só o `sender` tem de explicitar o destinatário:

- `send(P, message`:
- `receive(id, message)`: receive mensagens de qualquer processo

53.3 Comunicação Indireta

As mensagens são enviadas para uma **mailbox** (caixa de mensagens) ou **ports**, e o `receiver` vai buscar as mensagens a uma `poll`

- `send(M, message`
- `receive(M, message)`

O canal de comunicação possui as seguintes propriedades:

- Só é estabelecido se o **par de processos** comunicantes possui uma **`mailbox` partilhada**
- Pode estar associado a **mais do que dois processos**
- Entre um par de processos pode existir **mais do que um link** (uma mailbox por cada processo)

Questões que se levantam. Se **mais do que um processo** tentar **receber uma mensagem da mesma `mailbox`**…

- … é permitido?
  - Se sim. qual dos processos deve ser bem sucedido em ler a mensagem?

53.4 Implementação

Existem várias opções para implementar o **send** e **receive**, que podem ser combinadas entre si:

- **blocking send:** o `sender` **envia** a mensagem e fica **bloqueado** até a mensagem ser entregue ao processo ou mailbox destinatária
- **nonblocking send:** o `sender` após **enviar** a mensagem, **continua** a sua execução
- **blocking receive:** o `receiver` bloqueia-se até estar disponível uma mensagem para si
- **nonblocking receiver:** o `receiver` devolve a uma mensagem válida quando tiver ou uma indicação de que não existe uma mensagem válida quando não tiver

53.5 Buffering

O link pode usar várias políticas de implementação:

- **Zero Capacity:**
  - Não existe uma `queue`
  - O `sender` só pode enviar uma mensagem de cada vez. e o envio é **bloqueante**
  - O `receiver` lê uma mensagem de cada vez, podendo ser bloqueante ou não
- **Bounded Capacity:**
  - A `queue` possui uma capacidade finita
  - Quando está cheia, o `sender` bloqueia o envio até possuir espaço disponível
- **Unbounded Capacity:**
  - A `queue` possui uma capacidade (potencialmente) infinita
  - Tanto o `sender` como o `receiver` podem ser **não bloqueantes**

53.6 Bound-Buffer Problem usando mensagens

```
1  shared FIFO fifo;            /* fixed-size FIFO memory */
2  shared mutex access;         /* mutex to control mutual exclusion */
3  shared cond nslots;      /* condition variable to control availability of slots*/
4  shared cond nitems;      /* condition variable to control availability of items */
5
6  /* producers - p = 0, 1, ..., N-1 */
7  void producer(unsigned int p)
8  {
9      DATA data;
10     MESSAGE msg;
11
12     forever
13     {
14         produce_data(&val);
15         make_message(msg, data);
16         send(msg);
17         do_something_else();
18     }
19 }
20
21 /* consumers - c = 0, 1, ..., M-1 */
22 void consumer(unsigned int c)
23 {
24     DATA data;
25     MESSAGE msg;
26
27     forever
28     {
29         receive(msg);
30         extract_data(data, msg);
31         consume_data(data);
```

```
32          do_something_else();
33      }
34  }
```

53.7 Message Passing in Unix/Linux

System V:

- Existe uma fila de mensagens de **diferentes tipos**, representados por um inteiro
- `send` **bloqueante** se **não existir espaço disponível**
- A recepção possui um argumento para especificar o **tipo de mensagem a receber**:
  - Um tipo específico
  - Qualquer tipo
  - Um conjunto de tipos
- Qualquer que seja a política de recepção de mensagens:
  - É sempre **obtida** a mensagem **mais antiga** de uma dado tipo(s)
  - A implementação do `receive` pode ser **blocking** ou **nonblocking**
- System calls:
  - `msgget`
  - `msgsnd`
  - `msgrcv`
  - `msgctl`

POSIX

- Existe uma **priority** **`queue`**
- `send` **bloqueante** se **não existir espaço disponível**
- `receive` obtêm a mensagem **mais antiga** com a **maior prioridade**
  - Pode ser blocking ou nonblocking
- Funções:
  - `mq_open`
  - `mq_send`
  - `mq_receive`

54 Shared Memory in Unix/Linux

- É um recurso gerido pelo sistema operativo

Os espaços de endereçamento são **independentes** de processo para processo, mas o **espaço de endereçamento** é virtual, podendo a mesma **região de memória física**(memória real) estar mapeada em mais do que uma **memórias virtuais**

54.1 POSIX Shared Memory

- Criação:
  - `shm_open`
  - `ftruncate`
- Mapeamento:
  - `mmap`
  - `munmap`
- Outras operações:
  - `close`
  - `shm_unlink`
  - `fchmod`
  - …

54.2 System V Shared Memory

- Criação:
  - `shmget`
- Mapeamento:
  - `shmat`
  - `shmdt`
- Outras operações:
  - `shmctl`

55 Deadlock

- **recurso:** algo que um processo precisa para prosseguir com a sua execução. Podem ser:
  - **componentes físicos** do sistema computacional, como:
    * processador
    * memória
    * dispositivos de I/O
    * …
  - **estruturas de dados partilhadas**. Podem estar definidas
    * Ao nível do sistema operativo
      · PCT
      · Canais de Comunicação
    * Entre vários processos de uma aplicação

Os recursos podem ser:

- **preemptable:** podem ser retirados aos processos que estão na sua posse por entidades externas

  - processador
  - regiões de memória usadas no espaço de endereçamento de um processo

- **non-preemptable:** os recursos só podem ser libertados pelos processos que estão na sua posse
  - impressoras
  - regiões de memória partilhada que requerem acesso por exclusão mútua

O **deadlock** só é importante nos recursos **non-preemptable**.

O caso mais simples de deadlock ocorre quando:

1. O processo P_0 pede a posse do recurso A
   - É lhe dada a posse do recurso A, e o processo P_0 passa a possuir o recurso A em sua posse
2. O processo P_1 pede a posse do recurso B
   - É lhe dada a posse do recurso B, e o processo P_1 passa a possuir o recurso B em sua posse
3. O processo P_0 pede agora a posse do recurso B
   - Como o recurso B está na posse do processo P_1, é lhe negado
   - O processo P_0 fica em espera que o recurso B seja libertado para puder continuar a sua execução
   - No entanto, o processo P_0 não liberta o recurso A
4. O processo P_1 necessita do recurso A
   - Como o recurso A está na posse do processo P_0, é lhe negado
   - O processo P_1 fica em espera que o recurso A seja libertado para puder continuar a sua execução
   - No entanto, o processo P_1 não liberta o recurso B
5. Estamos numa situação de **deadlock**. Nenhum dos processos vai libertar o recurso que está na sua posse mas cada um deles precisa do recurso que está na posse do outro

55.1 Condições necessárias para a ocorrência de deadlock

Existem 4 condições necessárias para a ocorrência de **deadlock:**

1. **exclusão mútua:**
   - Pelo menos um dos recursos fica em posse de um processo de forma não partilhável
   - Obriga a que outro processo que precise do recurso espere que este seja libertado
2. **hold and wait:**
   - Um processo mantêm em posse pelo menos um recurso enquanto espera por outro recurso que está na posse de outro processo
3. **no preemption:**
   - Os recursos em causa são non preemptive, o que implica que só o processo na posse do recurso o pode libertar
4. **espera circular:**
   - é necessário um conjunto de processos em espera tais que cada um deles precise de um recurso que está na posse de outro processo nesse conjunto

Se **existir deadlock**, todas estas condições se verificam. *(A => B)*

Se **uma delas não se verifica**, não há deadlock. *(~B => ~A)*

55.1.1 O Problema da Exclusão Mútua

Dijkstra em 1965 enunciou um conjunto de regras para garantir o acesso **em exclusão mútua** por processo em competição por recursos de memória partilhados entre eles.[6]

1. **Exclusão Mútua:** Dois processos não podem entrar nas suas zonas críticas ao mesmo tempo
2. **Livre de Deadlock:** Se um process está a tentar entrar na sua zona crítica, eventualmente algum processo (não necessariamente o que está a tentar entrar), mas entra na sua zona crítica
3. **Livre de Starvation:** Se um processo está atentar entrar na sua zona crítica, então eventualmente esse processo entra na sua zona crítica
4. **First In First Out:** Nenhum processo a iniciar pode entrar na sua zona crítica antes de um processo que já está à espera do seu turno para entrar na sua zona crítica

55.2 Jantar dos Filósofos

- 5 filósofos sentados à volta de uma mesa, com comida à sua frente
  - Para comer, cada filósofo precisa de 2 garfos, um à sua esquerda e outro à sua direita
  - Cada filósofo alterna entre períodos de tempo em que medita ou come
- Cada **filósofo** é um **processo/thread** diferente
- Os **garfos** são os **recursos**

Uma possível solução para o problema é:

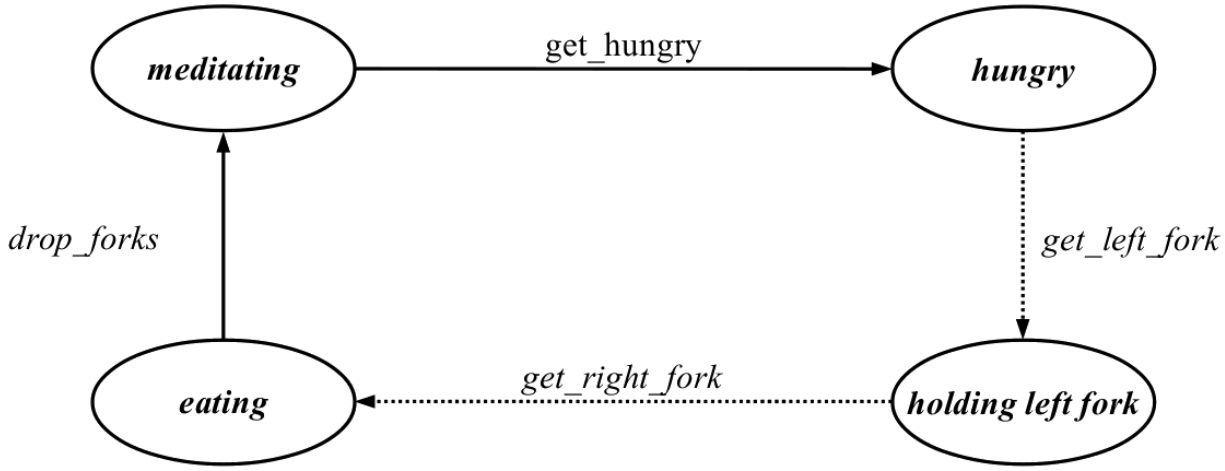


Figure 24: Ciclo de Vida de um filósofo

```
1  enum {MEDITATING, HUNGRY, HOLDING, EATING};
2
3  typedef struct TablePlace
4  {
5      int state;
```

[6] ficheiro em código fonte de compilação separada

```c
} TablePlace;

typedef struct Table
{
    Int semid;
    int nplaces;
    TablePlace place[0];
} Table;

int set_table(unsigned int n, FILE *logp);
int get_hungry(unsigned int f);
int get_left_fork(unsigned int f);
int get_right_fork(unsigned int f);
int drop_forks(unsigned int f);
```

Quando um filósofo fica *hungry*:

1. Obtém o garfo à sua esquerda
2. Obtém o garfo à sua direita

A solução **pode sofrer de deadlock:**

1. **exclusão mútua:**
   - Os garfos são partilháveis
2. **hold and wait:**
   - Se conseguir adquirir o `left_fork`, o filósofo fica no estado `holding_left_fork` até conseguir obter o `right_fork` e não liberta o `left_fork`
3. **no preemption:**
   - Os garfos são recursos non preemptive. Só o filósofo é que pode libertar os seus garfos após obter a sua posse e no fim de comer
4. **espera circular:**
   - Os garfos são partilhados por todos os filósofos de forma circular
     - O garfo à esquerda de um filósofo, `left_fork` é o garfo à direita do outro, `right_fork`

Se todos os filósofos estiverem a pensar e decidirem comer, pegando todos no garfo à sua esquerda ao mesmo tempo, entramos numa situação de **deadlock**.

55.3 Prevenção de Deadlock

Se uma das condições necessárias para a ocorrência de deadlock não se verificar, não ocorre deadlock.

As **políticas de prevenção de deadlock** são bastantes **restritas**, **pouco efetivas** e **difíceis de aplicar** em várias situações.

- **Negar a exclusão mútua** só pode ser aplicada a **recursos partilhados**
- **Negar *hold and wait*** requer **conhecimento *a priori* dos recursos necessários** e considera sempre o pior caso, no qual os recursos são todos necessários em simultâneo (o que pode não ser verdade)
- **Negar *no preemption***, impondo a libertação (e posterior reaquisição) de recursos adquiridos por processos que não têm condições (aka, todos os recursos que precisam) para continuar a execução pode originar grandes atrasos na execução da tarefa
- **Negar a *circular wait*** pode resultar numa má gestão de recursos

55.3.1 Negar a exclusão mútua

- Só é possível se os recursos puderem ser partilhados, senão podemos incorrer em **race conditions**
- Não é possível no jantar dos filósofos, porque os garfos não podem ser partilhados entre os filósofos
- Não é a condição mais vulgar a negar para prevenir *deadlock*

55.3.2 Negar *hold and wait*

- É possível fazê-lo se um processo é obrigado a pedir todos os recursos que vai precisar antes de iniciar, em vez de ir obtendo os recursos à medida que precisa deles
- Pode ocorrer **starvation**, porque um processo pode nunca ter condições para obter nenhum recurso
  - É comum usar *aging mechanisms* to para resolver este problema
- No jantar dos filósofos, quando um filósofo quer comer, passa a adquirir os dois garfos ao mesmo tempo
  - Se estes não tiverem disponíveis, o filósofo espera no `hungry state`, podendo ocorrer **starvation**

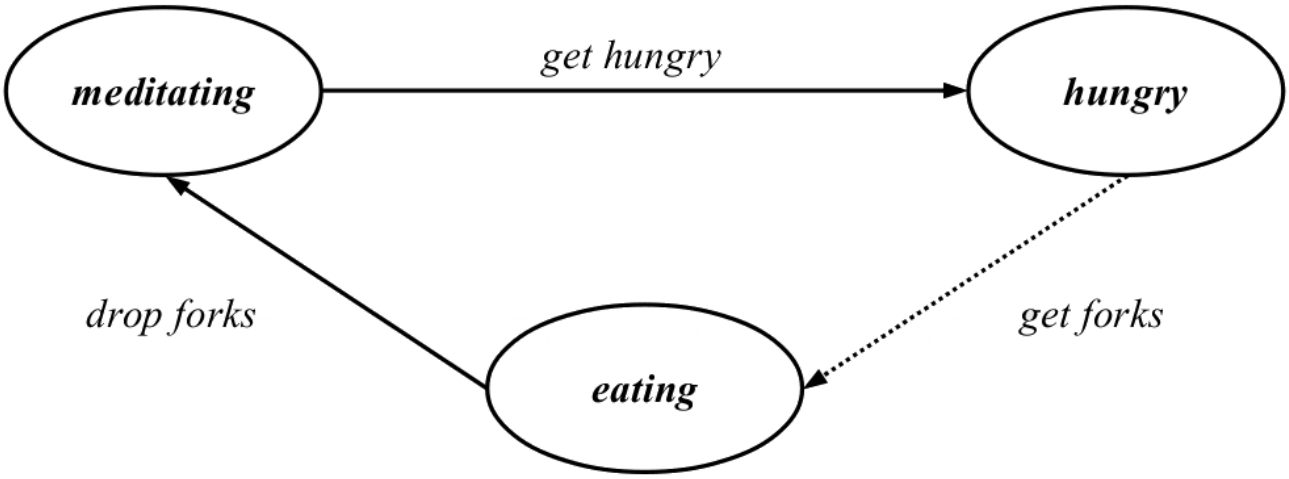


Figure 25: Negar *hold and wait*

Solução equivalente à proposta por Dijkstra.

55.3.3 Negar *no preemption*

- A condição de os recursos serem *non preemptive* pode ser implementada fazendo um processo libertar o(s) recurso(s) que possui se não conseguir adquirir o próximo recurso que precisa para continuar em execução
- Posteriormente o processo tenta novamente adquirir esses recursos
- Pode ocorrer **starvation** and **busy waiting**
  - podem ser usados *aging mechanisms* para resolver a starvation

  - para evitar busy waiting, o processo pode ser bloqueado e acordado quando o recurso for libertado

- No janta dos filósofos, o filósofo tenta adquirir o `left_fork`
  - Se conseguir, tenta adquirir o `right_fork`
    * Se conseguir, come
    * Se não conseguir, liberta o `left_fork` e volta ao estado `hungry`

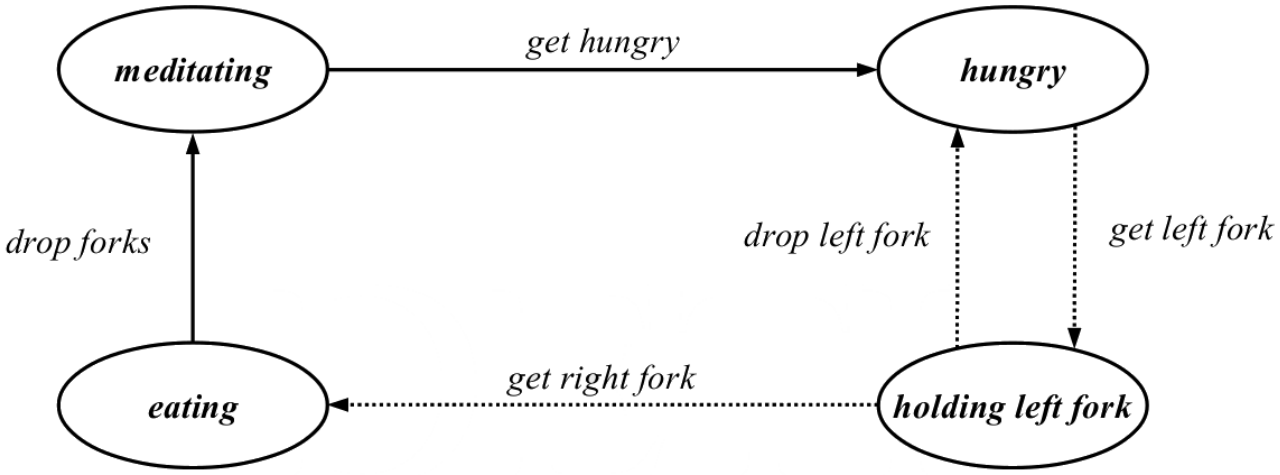


Figure 26: Negar a condição de *no preemption* dos recursos

55.3.4 Negar a espera circular

- Através do uso de IDs atribuídos a cada recurso e impondo uma ordem de acesso (ascendente ou descendente) é possível evitar sempre a espera em círculo
- Pode ocorrer **starvation**
- No jantar dos filósofos, isto implica que nalgumas situações, um dos filósofos vai precisar de adquirir primeiro o `right_fork` e de seguida o `left_fork`
  - A cada filósofo é atribuído um número entre 0 e N
  - A cada garfo é atribuído um ID (e.g., igual ao ID do filósofo à sua direita ou esquerda)
  - Cada filósofo adquire primeiro o garfo com o menor ID
  - obriga a que os filósofos 0 a N-2 adquiram primeiro o `left_fork` enquanto o filósofo N-1 adquirir primeiro o `right_fork`

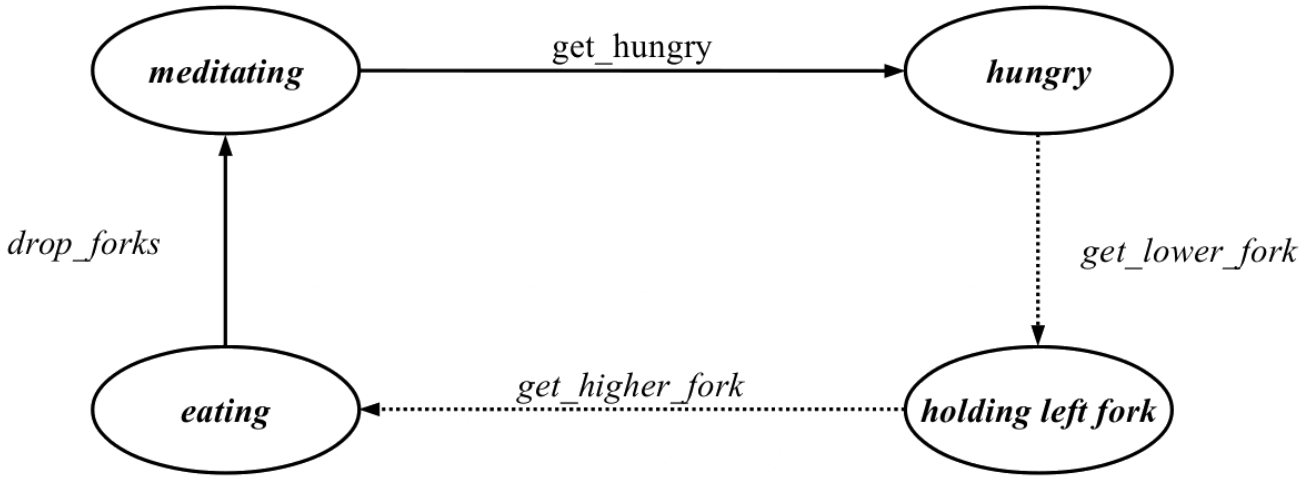


Figure 27: Negar a condição de espera circular no acesso aos recursos

55.4 Deadlock Avoidance

Forma menos restritiva para resolver situações de deadlock, em que **nenhuma das condições necessárias à ocorrência de deadlock é negada**. Em contrapartida, o sistema é **monitorizado continuamente** e um recurso **não é atribuído** se como consequência o sistema entrar num **estado inseguro/instável**

Um estado é considerado seguro se existe uma sequência de atribuição de recursos na qual todos os processos possa terminar a sua execução (não ocorrendo *deadlock*).

Caso contrário, poderá ocorrer deadlock (pode não ocorrer, mas estamos a considerar o pior caso) e o estado é considerado inseguro.

Implica que:

- exista uma lista de todos os recursos do sistema
- os processos intervenientes têm de declarar *a priori* todas as suas necessidades em termos de recursos

55.4.1 Condições para lançar um novo processo

Considerando:

- NTR_i - o número total de recursos do tipo `i` *(i = 0, 1, …, N-1)*
- $R_{i,j}$: o número de recursos do tipo `i` requeridos pelo processo `j`, *(i=0, 1, …, N-1 e j=0, 1, …, M-1)*

O sistema pode impedir um novo processo, `M`, de ser executado se a sua terminação não pode ser garantida. Para que existam certezas que um novo processo pode ser terminado após ser lançado, tem de se verificar:

$$NTR_i \geq R_{i,M} + \sum_{j=0}^{M-1} R_{i,j}$$

55.4.2 Algoritmo dos Banqueiros

Considerando:

- NTR_i: o número total de recursos do tipo `i` *(i = 0, 1, …, N-1)*
- $R_{i,j}$: o número de recursos do tipo `i` requeridos pelo processo `j`, *(i=0, 1, …, N-1 e j=0, 1, …, M-1)*
- $A_{i,j}$: o número de recursos do tipo `i` atribuídos/em posse do processo `j`, *(i=0, 1, …, N-1 e j=0, 1, …, M-1)*

Um novo recurso do tipo `i` só pode ser atribuído a um processo **se e só se** existe uma sequência $j' = f(i,j)$ tal que:

$$R_{i,j'} - A_{i,j'} < \sum_{k \geq ji'}^{M-1} A_{i,k}$$

Table 8: Banker's Algorithm Example

		A	B	C	D
	total	6	5	7	6
	free	3	1	1	2
	p1	3	3	2	2
maximum	p2	1	2	3	4
	p3	1	3	5	0
	p1	1	2	2	1
	p2	1	0	3	3
	p3	1	2	1	0
	p1	2	1	0	1
needed	p2	0	2	0	1
	p3	0	1	4	0
	p1	0	0	0	0
	p2	0	0	0	0
	p3	0	0	0	0

Para verificar se posso atribuir recursos a um processo, aos recursos `free` subtraio os recursos `needed`, ficando com os recursos que sobram. Em seguida simulo o que aconteceria se atribuisse o recurso ao processo, tendo em consideração que o processo pode usar o novo recurso que lhe foi atribuído sem libertar os que já possui em sua posse (estou a avaliar o pior caso, para garantir que não há deadlock)

Se o processo **p3** pedir 2 recursos do tipo C, o **pedido é negado**, porque **só existe 1 disponível**

Se o processo **p3** pedir 1 recurso do tipo B, o **pedido é negado**, porque apesar de existir 1 recurso desse tipo disponível, ao **longo da sua execução processo vai necessitar de 4** e só **existe 1 disponível**, podendo originar uma situação de **deadlock**, logo o **acesso ao recurso é negado**

Algoritmo dos banqueiros aplicado ao Jantar dos filósofos

- Cada filósofo primeiro obtém o `left_fork` e depois o `right_fork`

- No entanto, se um dos filósofos tentar obter um `left_fork` e o filósofo à sua esquerda já tem na sua posse um `left_fork`, o acesso do filósofo sem garfos ao `left_fork` é negado para não ocorrer **deadlock**

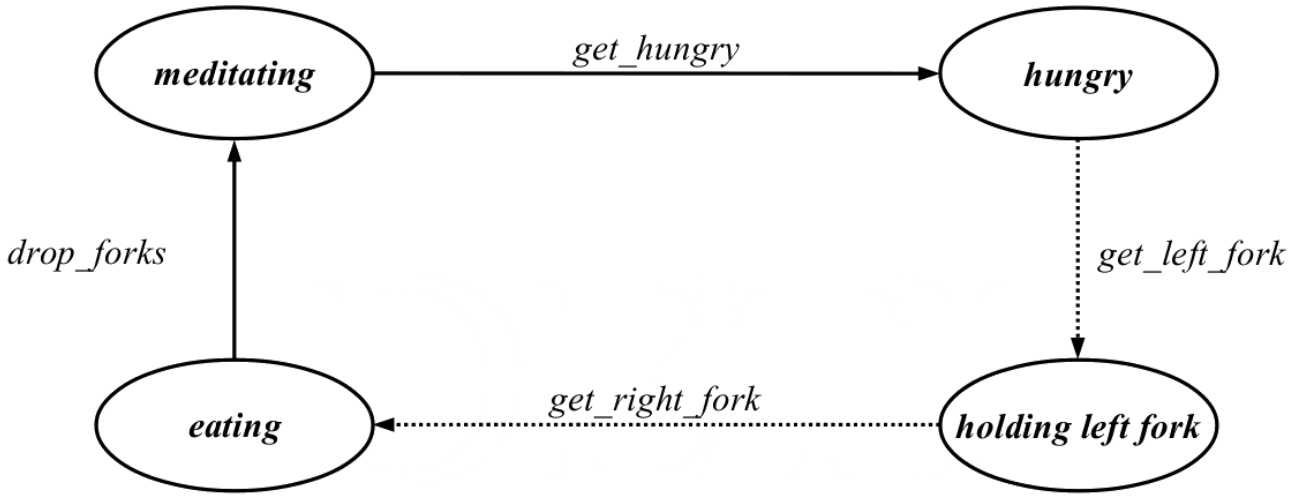


Figure 28: Algoritmo dos banqueiros aplicado ao Jantar dos filósofos

55.5 Deadlock Detection

Não são usados mecanismos nem para prevenir nem par evitar o deadlock, podendo ocorrer situações de deadlock:

- O estado dos sistema deve ser examinado para determinar se ocorreu uma situação de deadlock
  - É preciso verificar se existe uma **dependência circular de recursos** entre os processos
  - Periodicamente é executado um algoritmo que verifica o estado do registo de recursos:
    * recursos `free` vs recursos `granted` vs recursos `needed`
  - Se tiver ocorrido uma situação de deadlock, o SO deve possuir uma **rotina de recuperação** de situações de deadlock e executá-la
- Alternativamente, de um ponto de vista "arrogante", o problema pode ser ignorado

Se **ocorrer uma situação de deadlock**, a rotina de recuperação deve ser posta em prática com o objetivo de interromper a dependência circular de processos e recursos.

Existem três métodos para recuperar de deadlock:

- **Libertar recursos de um processo**, se possível
  - É atividade de um processo é suspensa até se puder devolver o recurso que lhe foi retirado
  - Requer que o estado do processo seja guardado e em seguida recarregado
  - Método eficiente
- **Rollback**
  - O estado de execução dos diferentes processos é guardado periodicamente
  - Um dos processos envolvidos na situação de deadlock é *rolled back* para o instante temporal em que o recurso lhe foi atribuído
  - A recurso é assim libertado do processo

- **Matar o processo**
  - Quando um processo entra em deadlock, é terminado
  - Método radical mas fácil de implementar

Alternativamente, existe sempre a opção de não fazer nada, entrando o processo em deadlock. Nestas situações, o utilizador é que é responsável por corrigir as situações de deadlock, por exemplo, terminando o programa com `CTRL + C`

56 Processes and Threads

56.1 Arquitectura típica de um computador

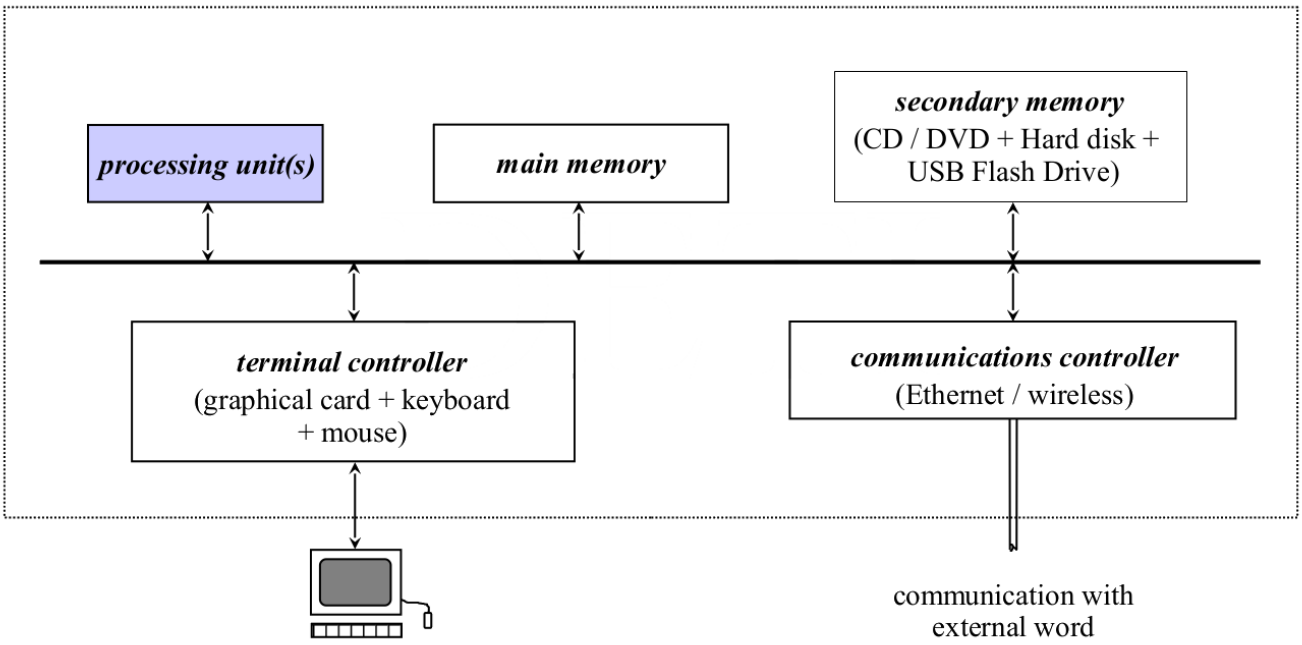


Figure 29: Arquitectura típica de um computador

56.2 Programa vs Processo

- **programa**: conjunto de instruções que definem como uma tarefa é executada num computador
  - É apenas um **conjunto de instruções** (código máquina), nada mais
  - Para realizar essas **funções/instruções/tarefas** o código (ou a versão compilada dele) tem de ser executado(a)
- **processo:** Entidade que representa a **execução de um programa**
  - Representa a sua atividade
  - Tem associado a si:
    * código que ao contrário do programa está armazenado num endereço de memória (addressing space)
    * **dados** (valores das diferentes variáveis) da execução corrente
    * valores atuais dos registos internos do processador
    * dados dos I/Os, ou seja, dados que estão a ser transferidos entre dispositivos de input e output
    * Estado da execução do programa, ou seja, qual a próxima execução a ser executada (registo PC)
  - Podem existir diferentes processos do mesmo programa
    * Ambiente **multiprogramado** - mais processos que processadores

56.3 Execução num ambiente multiprogramado

O sistema assume que o processo que está na posse do processador irá **ser interrompido**, pudendo assim executar outro processo e dar a "sensação" em **macro tempo** de **simultaneidade**. Nestas situações, o OS é responsável por:

- tratar da **mudança do contexto de execução**, guardando
  - o valor dos registos internos
  - o valor das variáveis
  - o endereço da próxima instrução a ser executada
- chamar o novo processo que vai ocupar agora o CPU e:
  - Esperar que o novo processo termine a realização das suas operações **ou**
  - Interromper o processo, **parando a sua execução no** processador quando este esgotar o seu `time quantum`

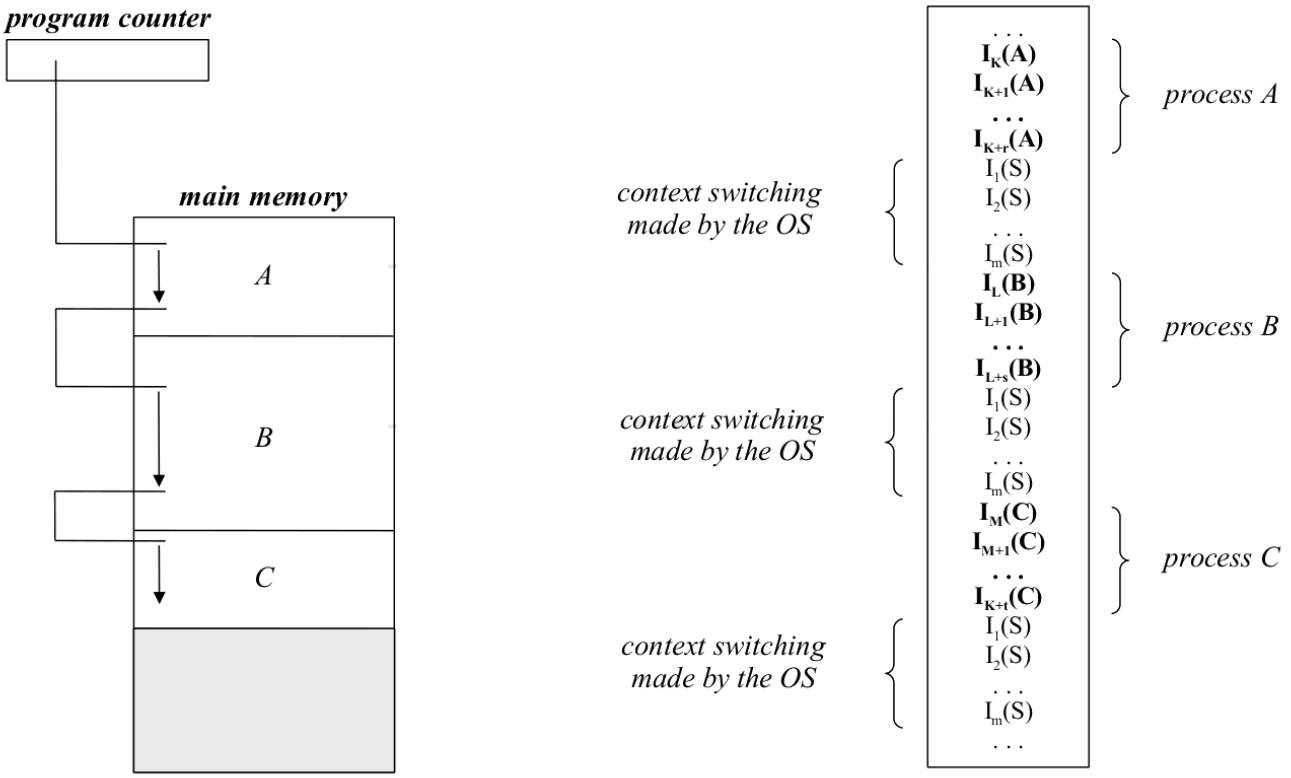


Figure 30: Exemplo de execução num ambiente multiprogramado

56.4 Modelo de Processos

Num ambiente **multiprogramado**, devido à constante **troca de processos**, é difícil expressar uma modelo para o processador. Devido ao elevado numero de processo e ao multiprogramming, torna-se difícil de saber qual o processo que está a ser executado e qual a fila de processos as ser executada.

É mais fácil assumir que o ambiente multiprogramado pode ser representado por um **conjunto de processadores virtuais**, estando um processo atribuído a cada um.

O processador virtual está: - **ON:** se o processo que lhe está atribuído está a ser executado - **OFF:** se o processo que lhe está atribuído não está a ser executado

Para este modelo temos ainda de assumir que: - Só um dos processadores é que pode estar ativo num dado período de tempo - O número de **processadores virtuais ativos é menor** (ou igual, se for um ambiente `single processor`) ao número de **processadores reais** - A execução de um processo **não é afetada** pelo instante temporal nem a localização no código em

que o processo é interrompido e é efetuado o switching - Não existem restrições do número de vezes que qualquer processo pode ser interrompido, quer seja executado total ou parcialmente

A operação de **switching entre processos** e consequentemente entre processadores virtuais ocorre de forma não **controlada** pelo programa a correr no CPU

Uma **operação de switching** é equivalente a efetuar o `Turning Off` de um processo virtual e o `Turning On` de outro processo virtual.

- `Turning Off` implica **guardar** todo o **contexto de execução**
- `Turning On` implica carregar todo o contexto de execução, **restaurando o estado do programa** quando foi interrompido

56.5 Diagrama de Estados de um Processo

Durante a sua existência, um processo pode assumir diferentes estados, dependendo das condições em que se encontra:

- **run:** O processo está em execução, tendo a posse do processador
- **blocked:** O processo está bloqueado à **espera de um evento externo** para estar em condições retomar a sua execução. Esse evento externo pode ser:
  - Acesso a um recurso da máquina
  - Fim de uma operação de I/O
  - …
- **ready:** O processo está pronto a ser executado, mas está à espera que o processador lhe dê a ordem de `start`/`resume` para puder retomar a sua execução.

As **transições entre estados** normalmente resultam de **intervenções externas ao processo**, mas podem depender de situações em que o processo força uma transição: - termina a sua execução antes de terminar o seu `time quantum` - Leitura/Escrita em I/O (scanf/printf)

Mesmo que um processo **não abandone o processador por sua iniciativa**, o `scheduler` é responsável por **planear o uso do processador pelos diferentes processos**.

O `(Process)Scheduler` é um módulo do kernel que **monitoriza e gere as transições entre processos**. Assim, um `while (1)` não é executado *ad eternum*. Um processador `multiprocess` só permite que o ciclo infinito seja executado quando é atribuído `CPU time` ao processo.

Existem diferentes políticas que permitem controlar a execução destas transições

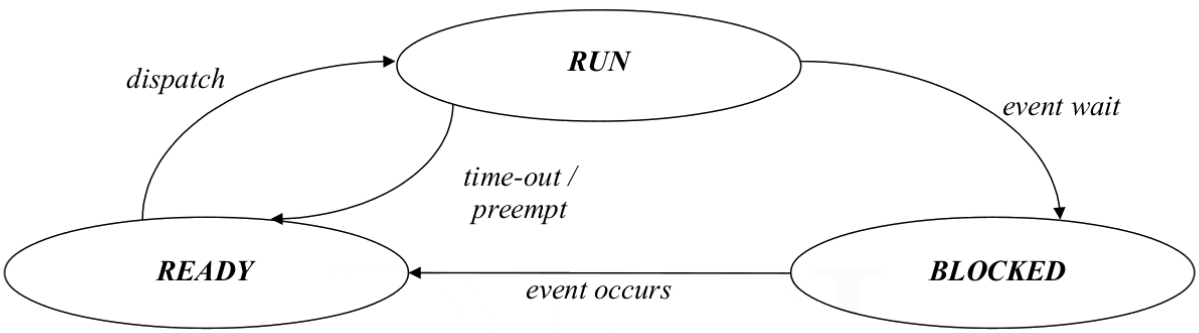


Figure 31: Diagrama de Estados do Processador - Básico

`Triggers` das transições entre estados:

- **dispatch:**
  - O processo que estava em modo `run` perdeu o acesso ao processador.
  - Do conjunto de processos prontos a serem executados, tem de ser escolhido **um** para ser executado, sendo lhe atribuído o processador.
  - A escolha feita pelo `dispatcher` pode basear-se em:
    * um sistema de prioridades
    * requisitos temporais
    * aleatoriedade
    * divisão igual do CPU
- **event wait:**
  - O processo que estava a ser executado sai do estado `run`, não estando em execução no processador.
    * Ou porque é impedido de continuar pelo scheduler
    * Ou por iniciativa do próprio processo.
      · *scanf*
      · *printf*
  - O CPU guarda o estado de execução do processo
  - O processo fica em estado `blocked` à **espera da ocorrência de um evento externo**, `event occurs`
- **event occurs:**
  - Ocorreu o evento que o processo estava à espera
  - O processo transita do estado `blocked` para o estado `ready`, ficando em fila de espera para que lhe seja atribuído o processador
- **time_out:**
  - O processo esgotou a sua janela temporal, `time quantum`
  - Através de uma interrupção em *hardware*, o sistema operativo vai forçar a saída do processo do processador
  - Transita para o estado *ready* até lhe ser atribuído um novo `time-quantum` do CPU
  - A transição por time out ocorre em qualquer momento do código.
  - Os sistemas podem ter `time quantum` diferentes e os `time slots` alocados não têm de ser necessariamente iguais entre dois sistemas.
- **preempt**:
  - O processo que possui a posse do processador tem uma prioridade mais baixa do que um processo que acordou e está pronto a correr (estado `ready`)
  - O processo que está a correr no processador é **removido** e transita para o estado `ready`
  - Passa a ser **executado** o processo de **maior prioridade**

56.5.1 Swap Area

O diagram de estados apresentado não leva em consideração que a **memória principal** (RAM) é **finita**. Isto implica que o número de **processos coexistentes em memória é limitado**.

É necessário usar a **memória secundária** (Disco Rígido) para **extender a memória principal** e aumentar a capacidade de armazenamento dos estados dos processos.

A **memória** **swap** pode ser:

- uma partição de um disco
- um ficheiro

Qualquer processo que **não esteja a correr** por ser `swapped out`, libertando memória principal para outros processos

Qualquer processo `swapped out` pode ser `swapped in`, **quando existir memória principal disponível**

Ao diagrama de estados tem de ser adicionados: - dois novos estados: - **suspended-ready:** Um processo no estado `ready` foi `swapped-out` - **suspended-blocked:** O processo no estado `blocked` foi `swapped-out` - dois novos tipos de transições: - **suspend:** O processo é `swapped out` - **activate:** O processo é `swapped in`

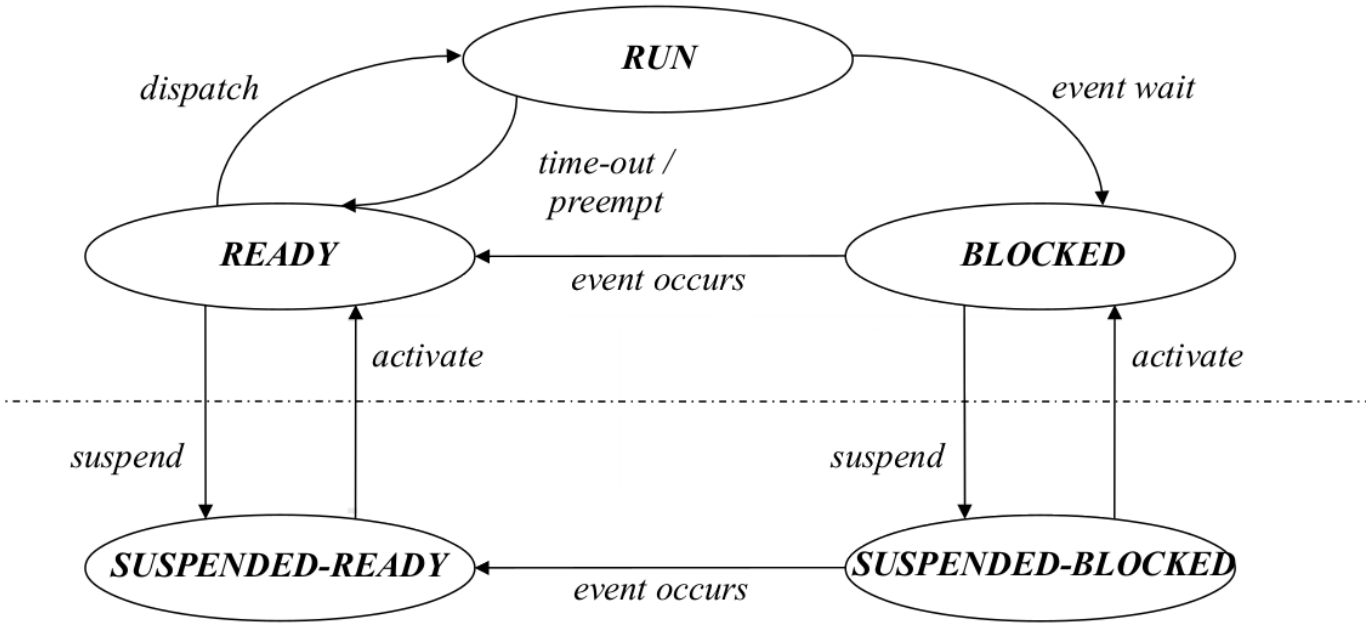


Figure 32: Diagrama de Estados do Processador - Com Memória de Swap

56.5.2 Temporalidade na vida dos processos

O diagrama assume que os processos são **intemporais**. Excluindo alguns processos de sistema, todos os processos são **temporais**, i.e.:

1. Nascem/São criados
2. Existem (por algum tempo)
3. Morrem/Terminam

Para introduzi a temporalidade no diagrama de estados, são necessários dois novos estados: - **new:** - O processo foi criado - Ainda não foi atribuído à `pool` de processos a serem executados - A estrutura de dados associado ao processo é inicializada - **terminated:** - O processo foi descartado da fila de processos executáveis - Antes de ser descartado, existem ações que tem de tomar *(needs clarification)*

Em consequência dos novos estados, passam a existir três novas transições: - **admit:** O processo é admitido pelo OS para a `pool` de processos executáveis - **exit:** O processo informa o SO que terminou a sua execução - **abort:** Um processo é forçado a terminar. - Ocorreu um `fatal error` - Um processo autorizado abortou a sua execução

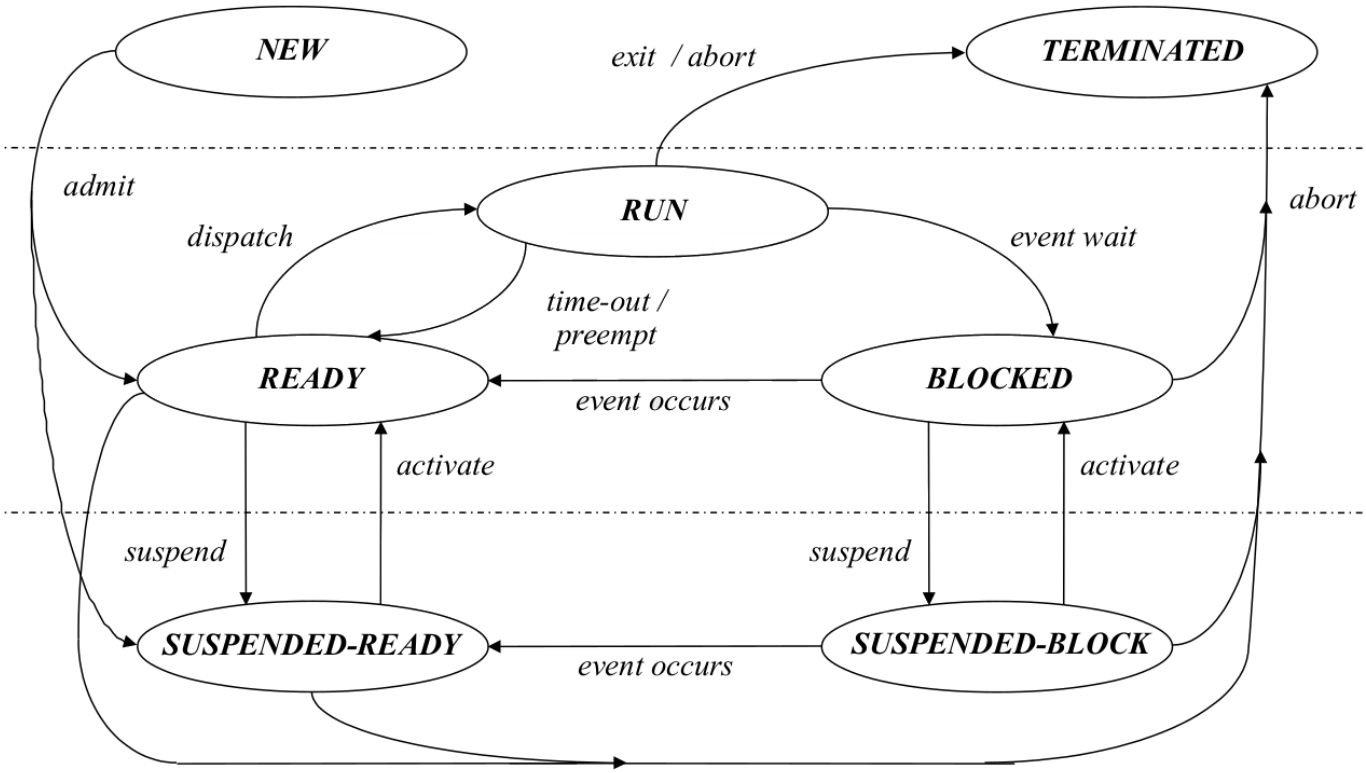


Figure 33: Diagrama de Estados do Processador - Com Processos Temporalmente Finitos

56.6 State Diagram of a Unix Process

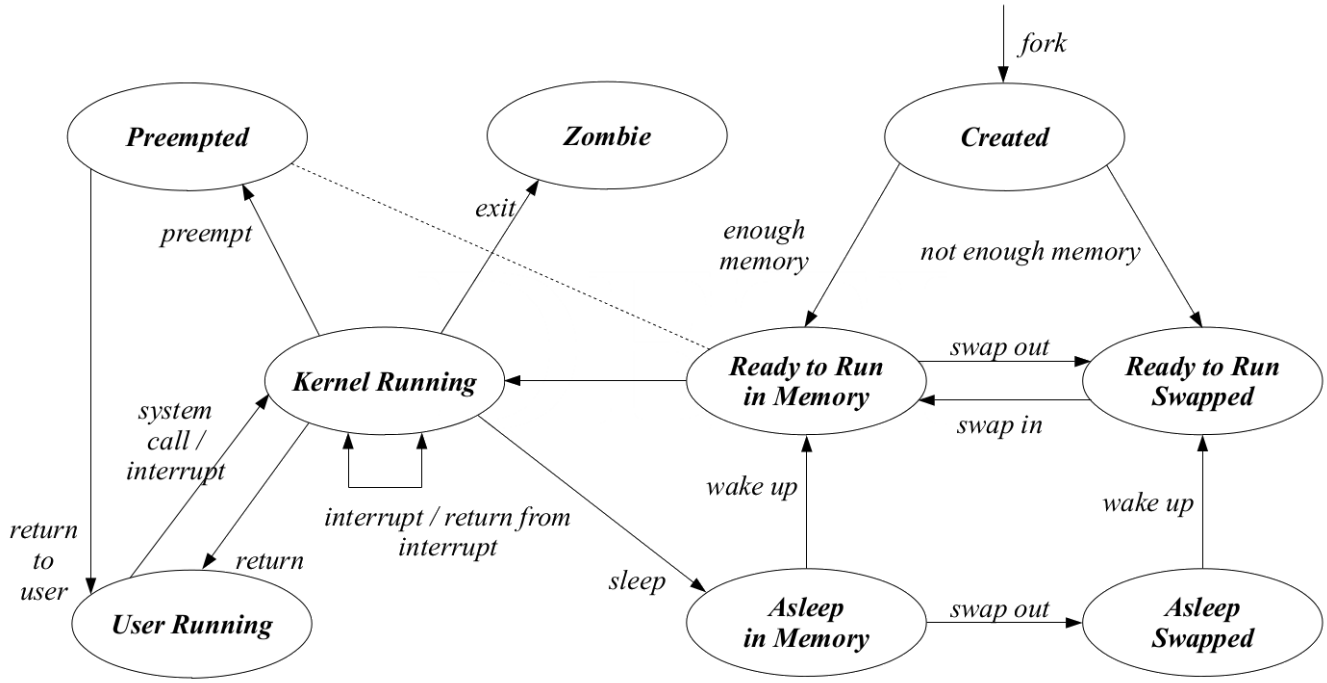


Figure 34: Diagrama de Estados do Processador - Com Memória de Swap

As três diferenças entre o diagrama de estados de um processo e o diagrama de estados do sistema Unix são

1. Existem **dois estados run**
    1. **kernel running**
    2. **user running**
    - Diferem **no modo** como o processador **executa o código máquina**, existindo **mais instruções e diretivas disponíveis** no modo supervisor (`root`)
2. O estado **ready** é dividido em dois estados:
    1. **ready to run in memory:** O processo está pronto para ser executado/continuar a execução, estando guardado o seu estado em memória
    2. **preempted:** O processo que estava a ser executado foi **forçado a sair do processador** porque **existe um processo mais prioritário para ser executado**
    - Estes **estados são equivalentes** porque:
        - estão ambos **armazenado na memória principal**
        - quando um processo é `preempted` continua pronto a ser executado (não precisando de nenhuma informação de I/O)
        - Partilham a mesma fila (`queue`) de processos, logo são tratados de forma idêntica pelo OS
    - Quando um **processo do utilizador abandona o modo de supervisor** (corre com permissões `root`), **pode ser preempted**
3. A transição de `time-out` que existe no diagrama dos estados de um processo em UNIX é coberta pela transição `preempted`

56.7 Supervisor preempting

Tradicionalmente, a **execução** de um processo **em modo supervisor** (`root`) implicava que a execução do processo **não pudesse ser** interrompida, ou seja, o processo não pode ser **`preempted`**. Ou seja, o UNIX não permitia **real-time processing**

Nas novas versões o código está dividido em **regiões atómicas**, onde a **execução não pode ser interrompida** para garantir a **preservação de informação das estruturas de dados a manipular**. Fora das regiões atómicas é seguro interromper a execução do código

Cria-se assim uma nova transição, **`return`** `to kernel` entre os estados `preempted` e `kernel running`.

56.8 Unix – traditional login

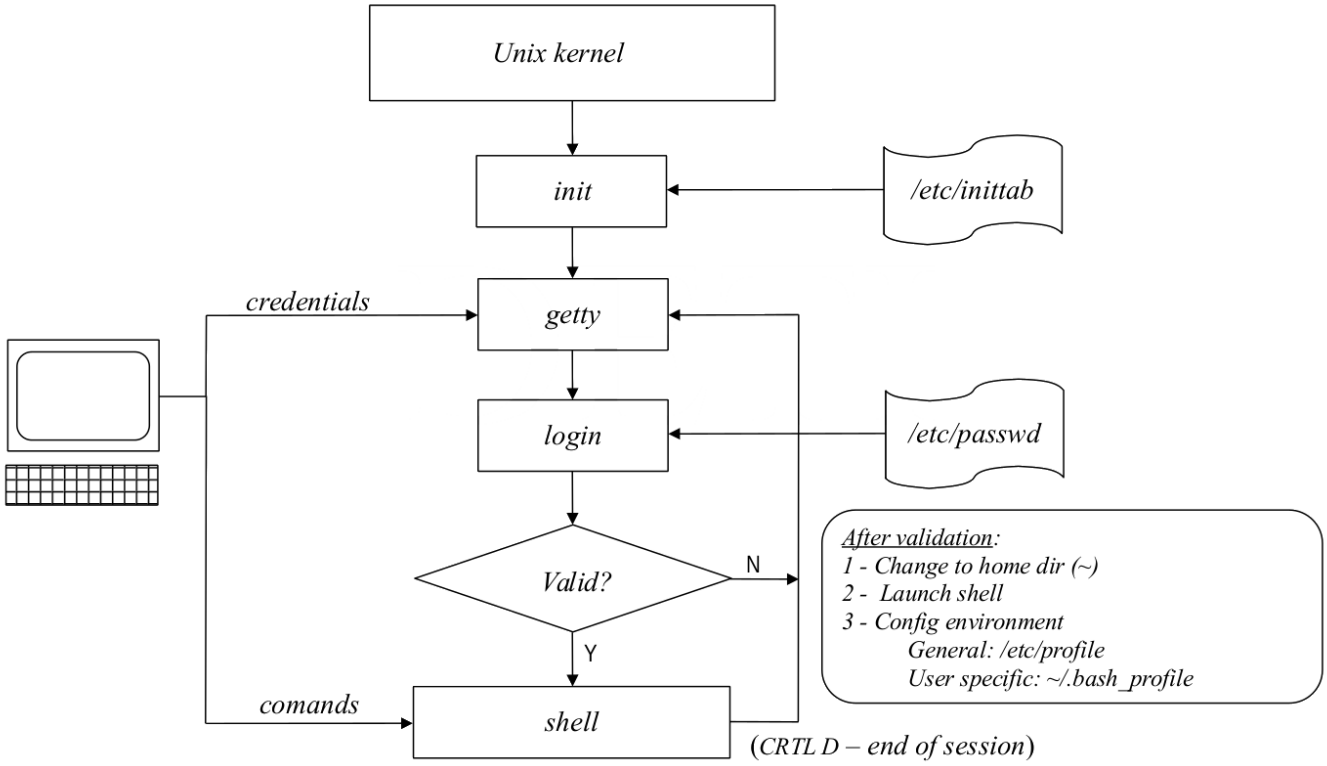


Figure 35: Diagrama do Login em Linux

56.9 Criação de Processos

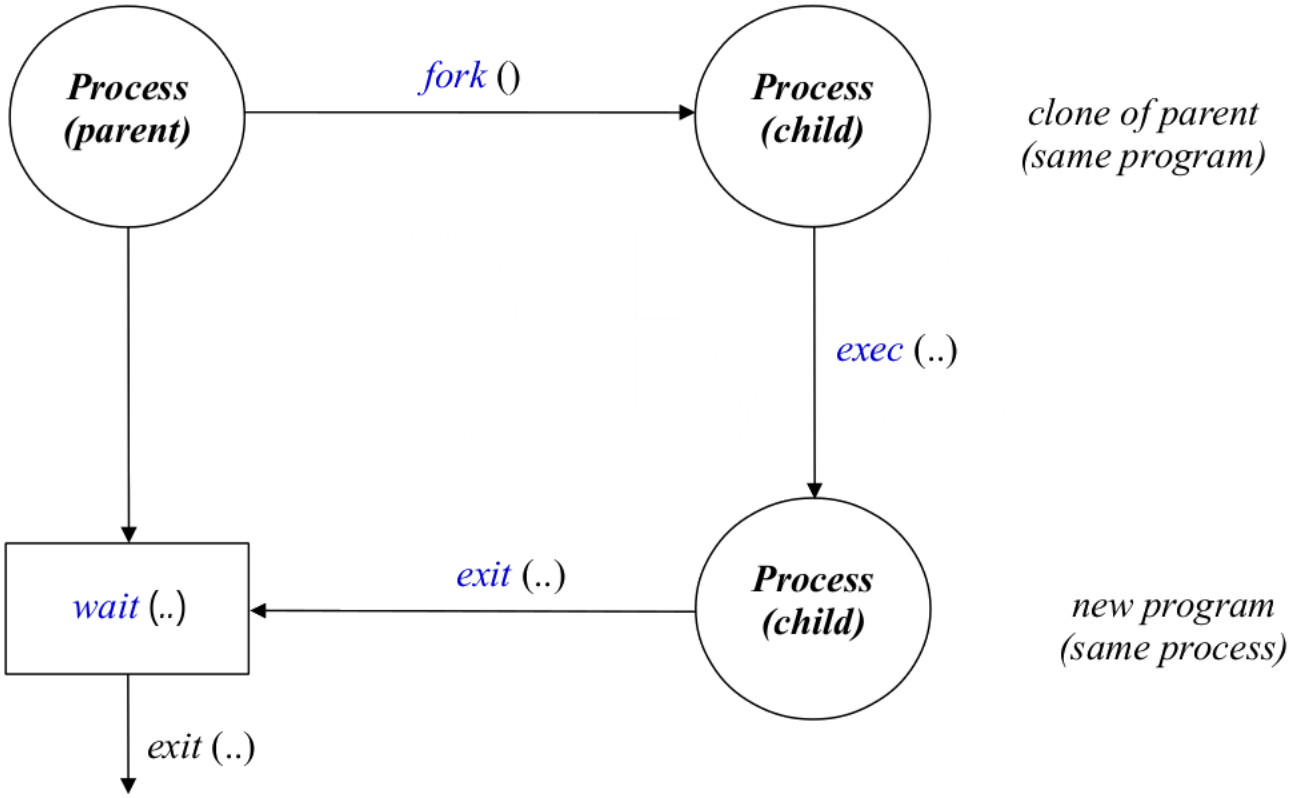


Figure 36: Criação de Processos

```c
#include <stdio.h>
#include <stdlib.h>
#include <sys/types.h>
#include <unistd.h>

int main(void)
{
    printf("Before the fork:\n");
    printf(" PID = %d, PPID = %d.\n",
    getpid(), getppid());

    fork();

    printf("After the fork:\n");
    printf(" PID = %d, PPID = %d. Who am I?\n", getpid(), getppid());

    return EXIT_SUCCESS;
}
```

- **fork**: **clona** o processo existente, criando uma **réplica**

  - O estado de execução é igual, incluindo o PC *(Program Counter)*
  - O **mesmo programa** é executado em **processos diferentes**
  - Não existem garantias que o pai execute primeiro que o filho
    * Tudo depende do `time quantum` que o processo pai ocupa antes do `fork`
    * É um ambiente multiprogramado: os dois programas correm em **simultâneo no micro tempo**

- O **espaço de endereçamento** dos dois processos é **igual**
  - É seguida uma filosofia **copy-on-write**. Só é efetuada a cópia da página de memória se o **processo filho modificar** os conteúdos das variáveis

Existem variáveis diferentes:

- **PPID:** Parent Process ID
- **PID:** Process ID

```c
#include <stdio.h>
#include <stdlib.h>
#include <sys/types.h>
#include <unistd.h>

int main(void)
{
    printf("Before the fork:\n");
    printf(" PID = %d, PPID = %d.\n",
    getpid(), getppid());

    int ret = fork();

    printf("After the fork:\n");
    printf(" PID = %d, PPID = %d, ret = %d\n", getpid(), getppid(), ret);

    return EXIT_SUCCESS;
}
```

O retorno da instrução `fork` é diferente entre o processo pai e filho:

- pai: `PID` do filho
- filho: `0`

O retorno do `fork` pode ser usado como variável booleana, de modo a **distinguir o código a correr no filho e no pai**

```c
#include <stdio.h>
#include <stdlib.h>
#include <sys/types.h>
#include <unistd.h>

int main(void)
{
    printf("Before the fork:\n");
    printf(" PID = %d, PPID = %d.\n", getpid(), getppid());

    int ret = fork();
```

```
12
13      if (ret == 0)
14      {
15          printf("I'm the child:\n");
16          printf(" PID = %d, PPID = %d\n", getpid(), getppid());
17      }
18      else
19      {
20          printf("I'm the parent:\n");
21          printf(" PID = %d, PPID = %d\n", getpid(), getppid());
22      }
23
24      printf("After the fork:\n");
25      printf(" PID = %d, PPID = %d, ret = %d\n", getpid(), getppid(), ret);
26
27      return EXIT_SUCCESS;
28  }
```

O `fork` por si só não possui grande interesse. O interesse é puder executar um programa diferente no filho.

- **`exec` system call:** Executar um programa diferente no processo filho
- **`wait` system call:** O pai esperar pela conclusão do programa a correr nos filhos

```
1   #include <stdio.h>
2   #include <stdlib.h>
3   #include <sys/types.h>
4   #include <unistd.h>
5
6   int main(void)
7   {
8       /* check arguments */
9       if (argc != 2)
10      {
11          fprintf(stderr, "spawn <path to file>\n");
12          exit(EXIT_FAILURE);
13      }
14      char *aplic = argv[1];
15
16      printf("==============================\n");
17
18      /* clone phase */
19      int pid;
20      if ((pid = fork()) < 0)
21      {
22          perror("Fail cloning process");
23          exit(EXIT_FAILURE);
24      }
25
26      /* exec and wait phases */
27      if (pid != 0) // only runs in parent process
28      {
29          int status;
```

```
30          while (wait(&status) == -1);
31          printf("================================\n");
32          printf("Process %d (child of %d)""
33           ends with status %d\n",
34          pid, getpid(), WEXITSTATUS(status));
35      }
36      else // this only runs in the child process
37      {
38          execl(aplic, aplic, NULL);
39          perror("Fail launching program");
40          exit(EXIT_FAILURE);
41      }
42  }
```

O `fork` pode **não ser bem sucedido**, ocorrendo um `fork failure`.

56.10 Execução de um programa em C/C++

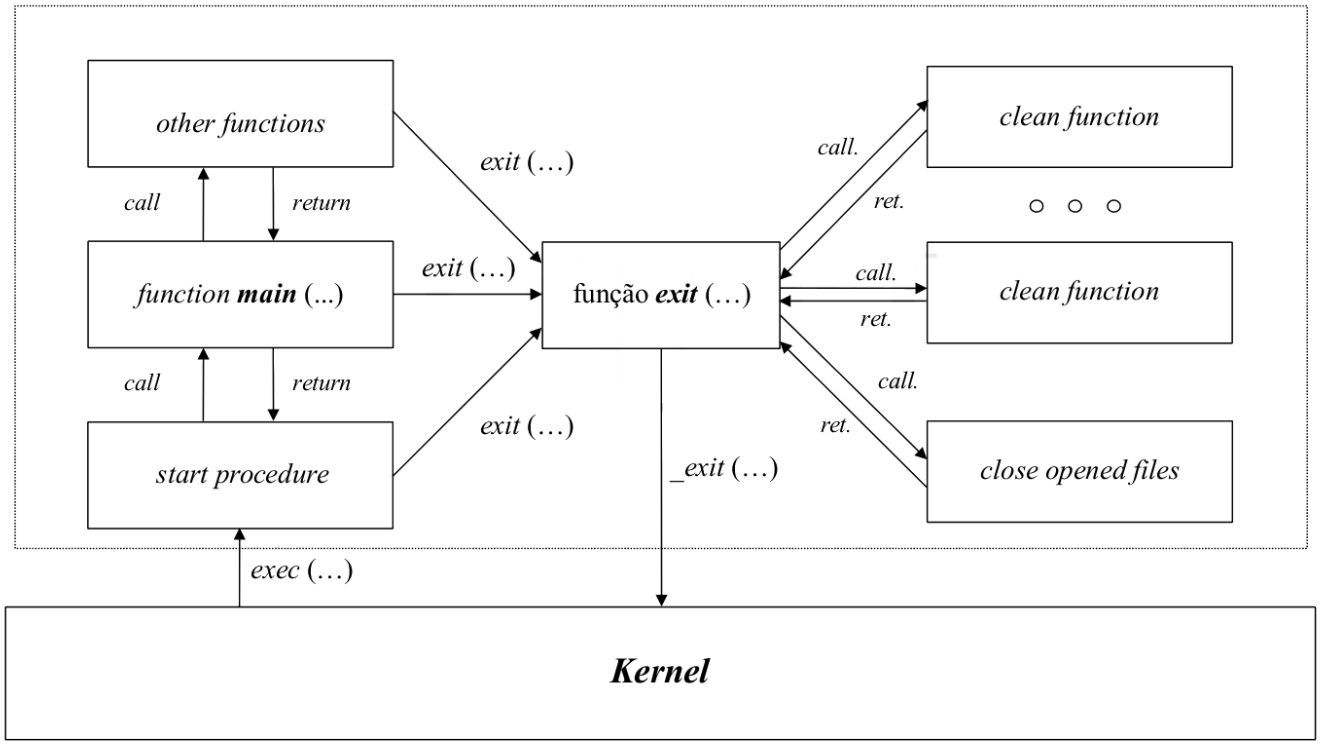


Figure 37: Execução de um programa em C/C++

- Em C/C++ o nome de uma função é um ponteiro para a sua função.
- Em C/C++ um include não inclui a biblioteca
  - Indica ao programa que vou usar uma função que tem esta assinatura
- **atexit:** Regista uma função para ser chamada no fim da execução normal do programa
  - São chamadas em ordem inversa ao seu registo

56.11 Argumentos passados pela linha de comandos e variáveis de ambiente

- **argv:** array de strings que representa um conjunto de parâmetros passados ao programa
  - **argv[0]** é a referência do programa
- **env** é um array de strings onde cada string representa uma variável do tipo: `name-value`
- **getenv** devolve o valor de uma variável definida no array **env**

56.12 Espaço de Endereçamento de um Processo em Linux

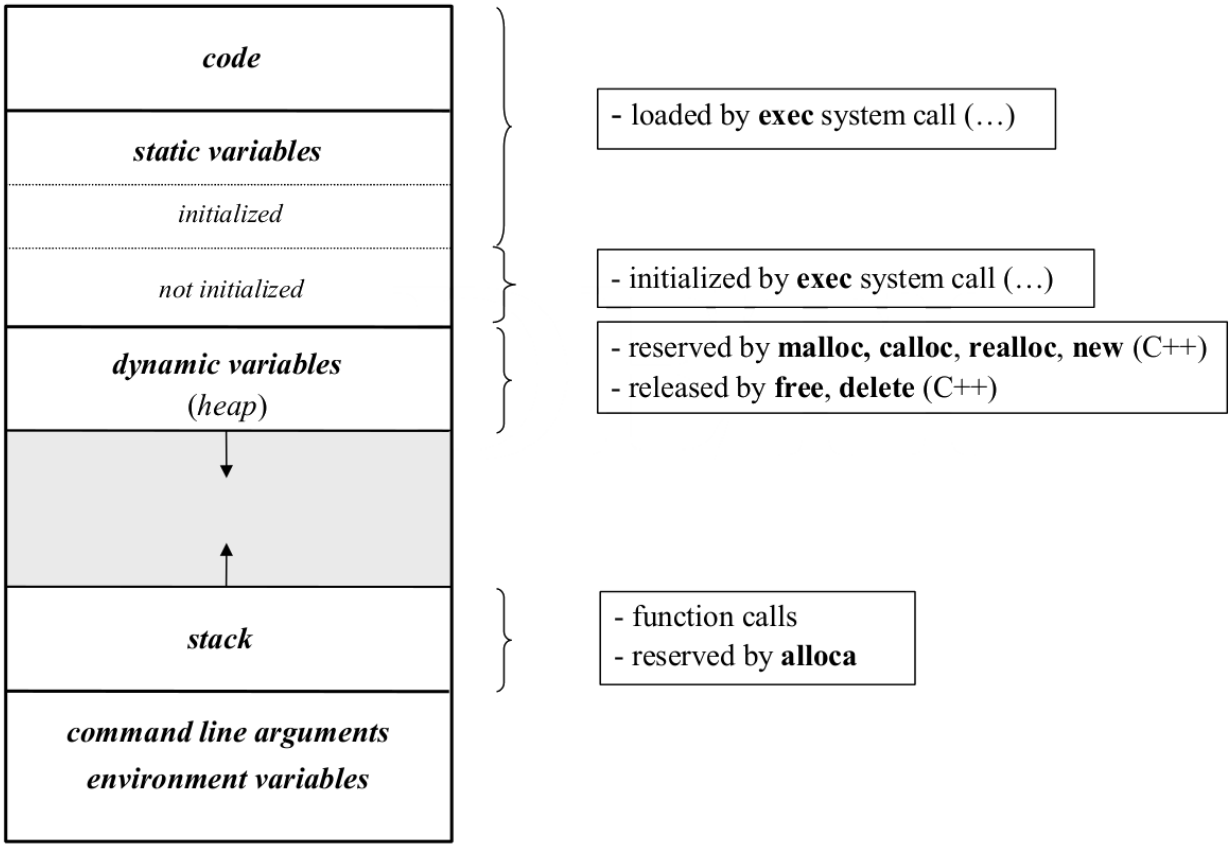


Figure 38: Espaço de endereçamento de um processo em Linux

- As variáveis que existem no processo pai também existem no processo filho (clonadas)
- As variáveis globais são comuns aos dois processos
- Os endereços das variáveis são todos iguais porque o espaço de endereçamento é igual (memória virtual)
- Cada processo tem as suas variáveis, residindo numa página de memória física diferente
- Quando o processo é clonado, o espaço de dados só é clonado quando um processo escreve numa variável, ou seja, após a modificação é que são efetuadas as cópias dos dados
- O programa acede a um endereço de memória virtual e depois existe hardware que trata de alocar esse endereço de memória de virtual num endereço de memória física

- Posso ter dois processos com memórias virtuais distintas mas fisicamente estarem ligados *ao mesmo endereço de memória*
- Quando faço um `fork` não posso assumir que existem variáveis partilhadas entre os processos

56.12.1 Process Control Table

É um array de `process control block`, uma estrutura de dados mantida pelo sistema operativo para guardar a informação relativa todos os processos.

O `process controlo block` é usado para guardar a informação relativa a apenas um processo, possuindo os campos:

- `identification`: id do processo, processo-pai, utilizador e grupo de utilizadores a que pertence
- `address space`: endereço de memória/swap onde se encontra:
  - código
  - dados
  - stack
- `processo state`: valor dos registos internos (incluindo o PC e o `stack pointer`) quando ocorre o switching entre processos
- `I/O context`: canais de comunicação e respetivos buffers que o processo tem associados a si
- `state`: estado de execução, prioridade e eventos
- `statistic`: *start time*, *CPU time*

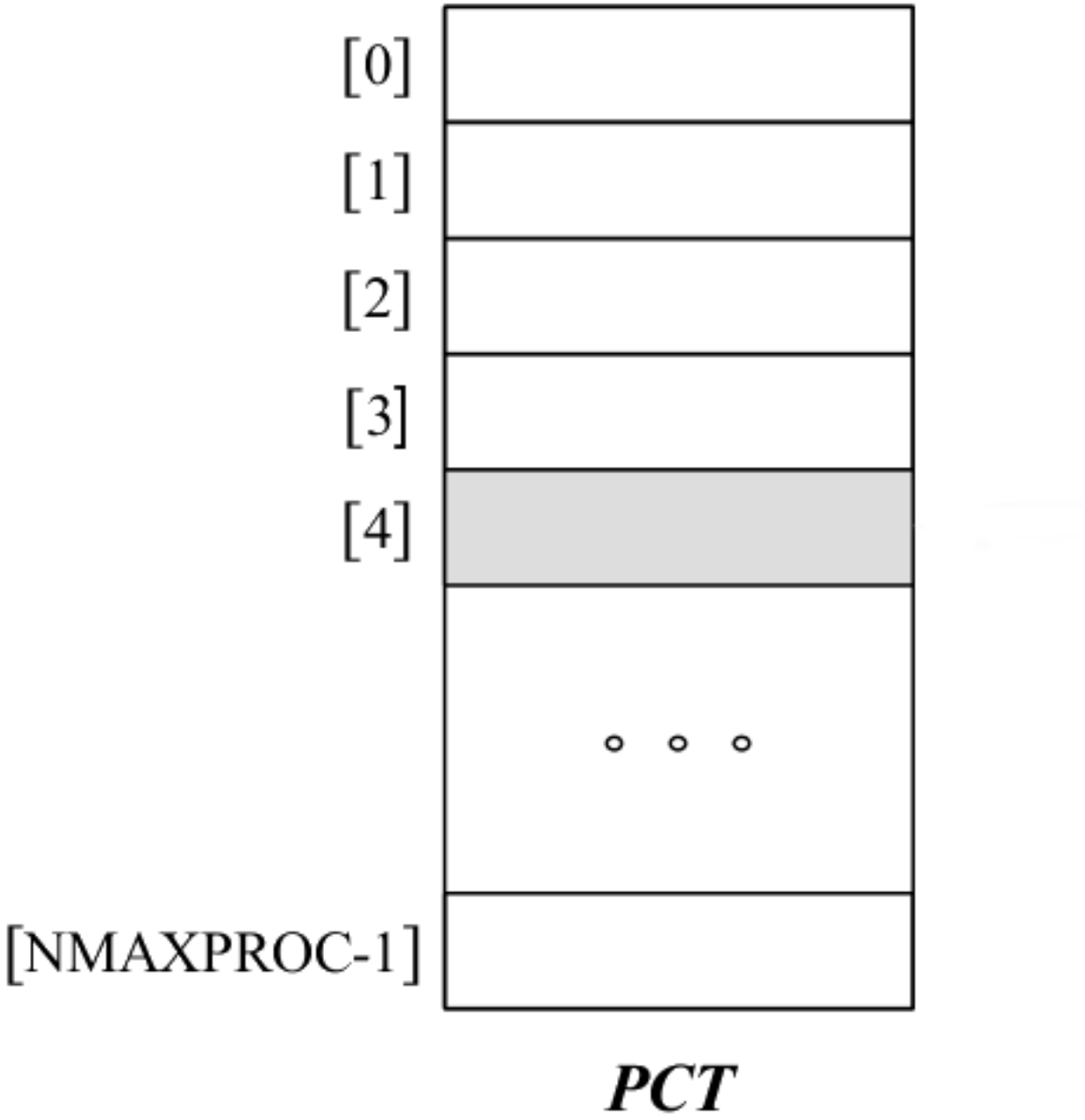


Figure 39: Process Control Table

57 Threads

Num sistema operativo tradicional, um **processo** inclui:

- um **espaço de endereçamento**
- um **conjunto de canais de comunicação** com dispositivos de I/O

- uma única **`thread` de controlo** que:
  - incorpora os **registos do processador** (incluindo o PC)
  - **stack**

Existem duas stacks no sistema operativo:

- `user stack`: cada `processo`/`thread` possui a sua (em memória virtual e corre com privilégios do `user`)
- `system stack`: global a todo o sistema operativo (no `kernel`)

Podendo estes dois componentes serem **geridos de forma independente**.

Visto que uma `thread` é apenas um **componente de execução** dentro de um processo, várias `threads` **independentes** podem coexistir no mesmo processo, **partilhando** o mesmo **espaço de endereçamento** e o mesmo contexto de **acesso aos dispositivos de I/O**. Isto é **`multithreading`**.

Na prática, as `threads` podem ser vistas como *light weight processes*

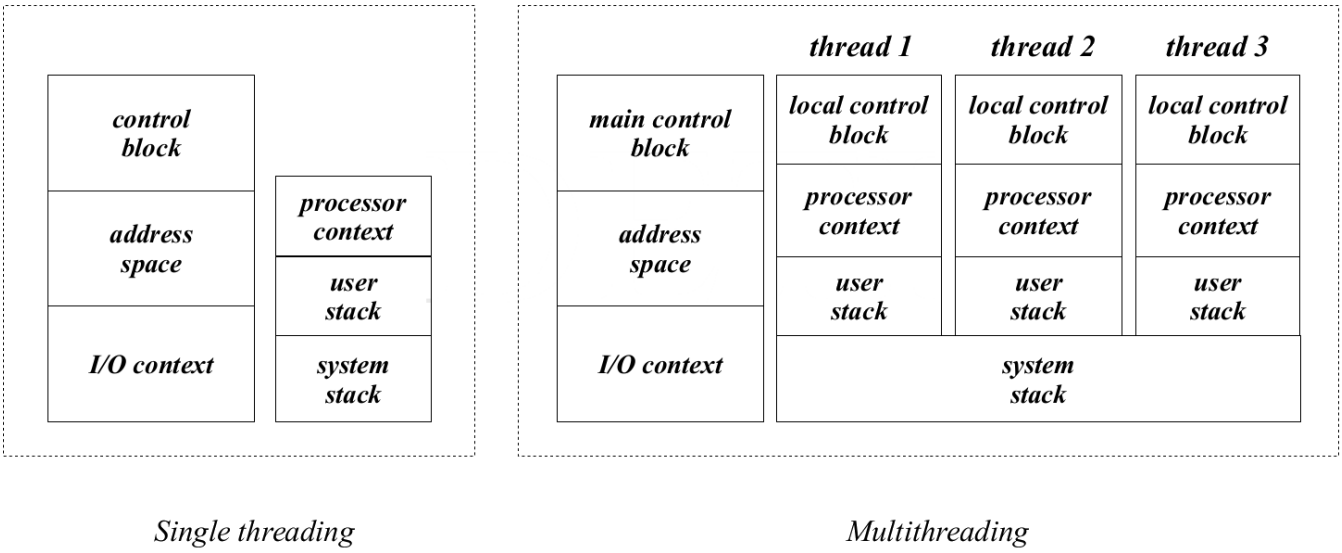


Figure 40: Single threading vs Multithreading

O controlo passa a ser centralizado na `thread` principal que gere o processo. A `user stack`, o **contexto de execução do processador** passa a ser dividido por todas as `threads`.

57.1 Diagrama de Estados de uma thread

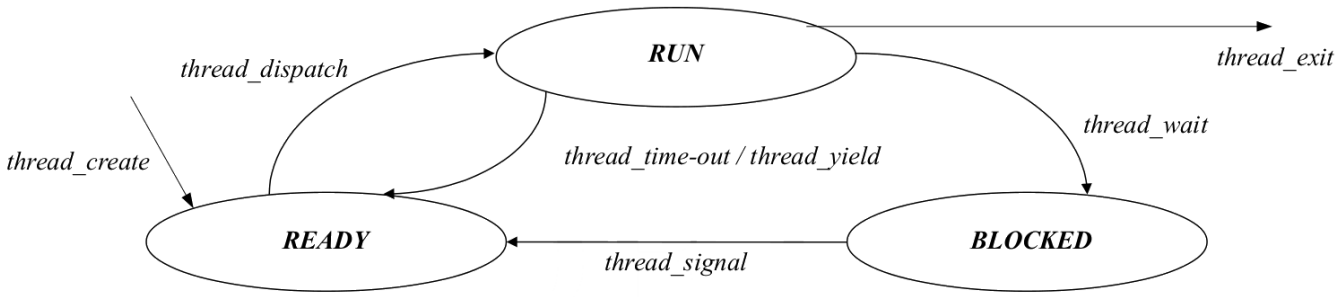


Figure 41: Diagrama de estados de uma thread

O diagrama de estados de um `thread` é mais simplificado do que o de um processo, porque só são "necessários" os estados que interagem **diretamente com o processador**:

```
- 'run'
- 'ready'
- 'blocked'
```

Os estados `suspend-ready` e `suspended-blocked` estão relacionados com o **espaço de endereçamento** do **processo** e com a zona onde estes dados estão guardados, dizendo respeito ao **processo e não à thread**

Os estado **new** e `terminated`não estão presentes, porque a gestão do ambiente multiprogramado prende-se com a restrição do número de `threads` que um processo pode ter, logo dizem respeito ao processo

57.2 Vantagens de Multithreading

- **facilita a implementação** (em certas aplicações):
  - Existem aplicações em que **decompor a solução** num conjunto de `threads` que correm paralelamente facilita a implementação
  - Como o `address space` e o `I/O context` são partilhados por todas as `threads`, `multithreading` favorece esta decomposição
- **melhor utilização dos recursos**
  - A criação, destruição e `switching` é mais eficiente usando `threads` em vez de processos
- **melhor performance**
  - Em aplicações `I/O driven`, `multithreading` permite que **várias atividades se sobreponham**, **aumentando a rapidez** da sua execução
- **`multiprocessing`**
  - É possível **paralelismo em tempo real** se o processador possuir **múltiplos CPUs**

57.3 Estrutura de um programa multithreaded

- Cada `thread` está tipicamente associada com a execução de uma função que implementa alguma atividade em específico
- A **comunicação entre `threads`** é efetuada através da estrutura de dados do **processo**, que é vista pelas `threads` como uma estrutura de dados global
- o **programa principal** também é uma `thread`
  - A 1ª a ser criada
  - Por norma a última a ser destruída

57.4 Implementação de Multithreading

`user level threads`:

- Implementadas por uma biblioteca
  - Suporta a criação e gestão das `threads` sem intervenção do `kernel`
- Correm com permissões do utilizador
- Solução versátil e portável
- Quando uma `thread` executa uma `system call` bloqueante, **todo o processo bloqueia** (o *kernel* só "vê" o processo)
- Quando passo variáveis a `threads`, elas têm de ser estáticas ou dinâmicas

`kernel level threads`

- As `threads` são implementadas diretamente ao nível do kernel
- Menos versáteis e portáteis
- Quando uma `thread` executa uma `system call` bloqueante, **outra *thread* pode entrar em execução**

57.4.1 Libraria pthread

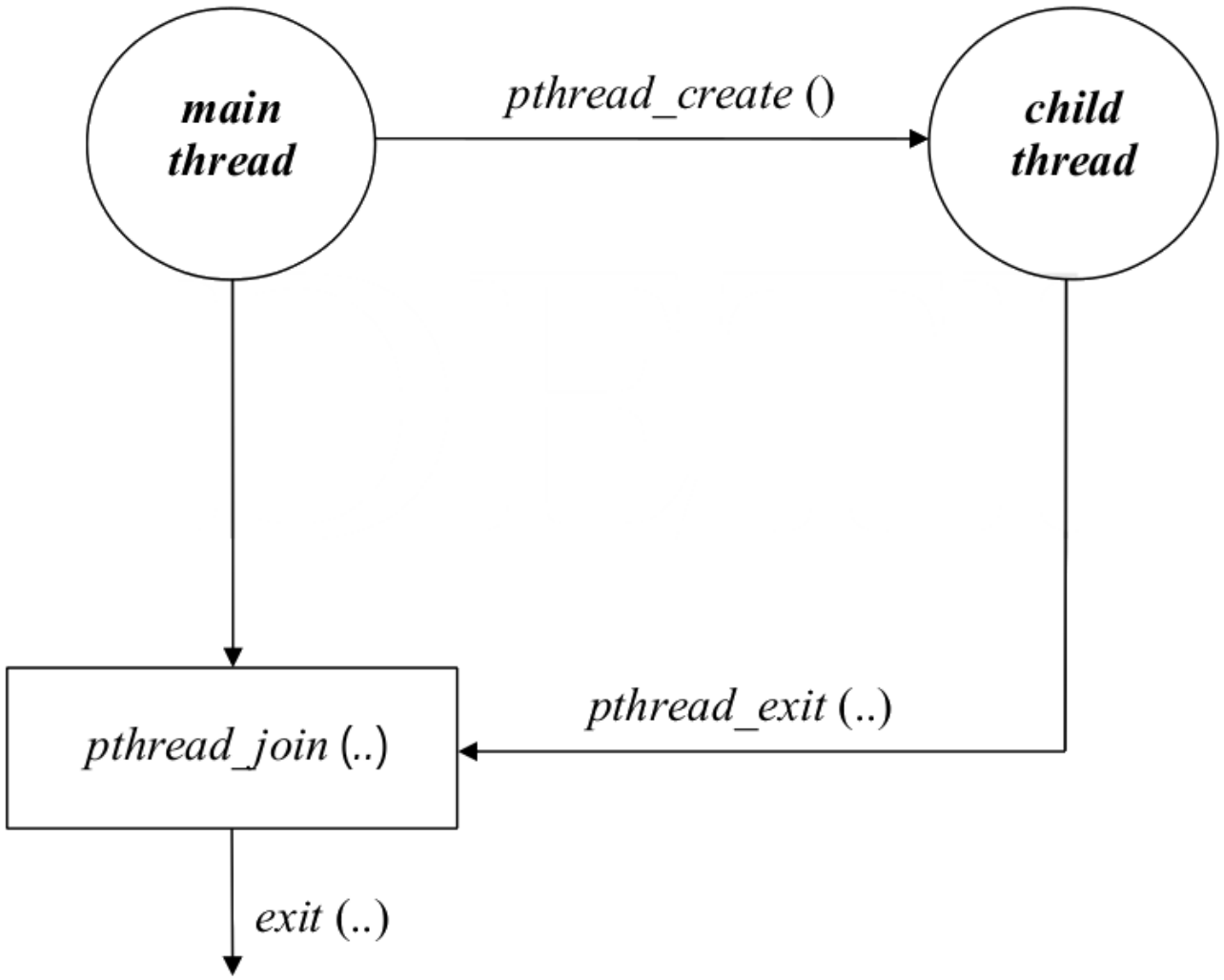


Figure 42: Exemplo do uso da libraria pthread

```c
#include <stdio.h>
#include <stdlib.h>
#include <pthread.h>

/* return status */
int status;

/* child thread */
void *threadChild (void *par)
{
    printf (”I’m the child thread!\n”);
    status = EXIT_SUCCESS;
    pthread_exit (&status);
}

```

```c
/* main thread */
int main (int argc, char *argv[])
{
    /* launching the child thread */
    pthread_t thr;
    if (pthread_create (&thr, NULL, threadChild, NULL) != 0)
    {
        perror ("Fail launching thread");
        return EXIT_FAILURE;
    }

    /* waits for child termination */
    if (pthread_join (thr, NULL) != 0)
    {
        perror ("Fail joining child");
        return EXIT_FAILURE;
    }

    printf ("Child ends; status %d.\n", status);
    return EXIT_SUCCESS;
}
```

57.5 Threads em Linux

2 system calls para criar processos filhos:

- fork:
  - cria uma novo processo que é uma **cópia integral** do processo atual
  - o address space é I/O context é duplicado
- clone:
  - cria um novo processo que pode partilhar elementos com o pai
  - Podem ser partilhados
    * espaço de endereçamento
    * tabela de file descriptors
    * tabela de signal handlers
  - O processo filho executa uma dada função

Do ponto de vista do kernel, processos e threads são **tratados de forma semelhante**

Threads do **mesmo processo** foma um thread group e possuem o **mesmo** thread group identifier (TGID). Este é o valor retornado pela função getpid() quando aplicada a um grupo de threads

As várias threads podem ser distinguidas dentro de um **grupo de threads** pelo seu unique thread identifier (TID). É o valor retornado pela função gettid()

```c
#include <stdio.h>
#include <stdlib.h>
#include <pthread.h>
#include <unistd.h>
```

```c
#include <unistd.h>
#include <sys/types.h>

pid_t gettid()
{
    return syscall(SYS_gettid);
}

/* child thread */
int status;
void *threadChild (void *par)
{
    /* There is no glibc wrapper, so it was to be called
     * indirectly through a system call
     */

    printf ("Child: PPID: %d, PID: %d, TID: %d\n", getppid(), getpid(), gettid());
    status = EXIT_SUCCESS;
    pthread_exit (&status);
}
```

O `TID` da *main thread* é a mesma que o `PID` do processo, **porque são a mesma entidade**.

Para efetuar a compilação, tenho de indicar que a biblioteca pthread tem de ser usada na linkagem:

```
g++ -o x thread.cpp -pthread
```

58 Process Switching

Revisitando a o diagrama de estados de um processador `multithreading`

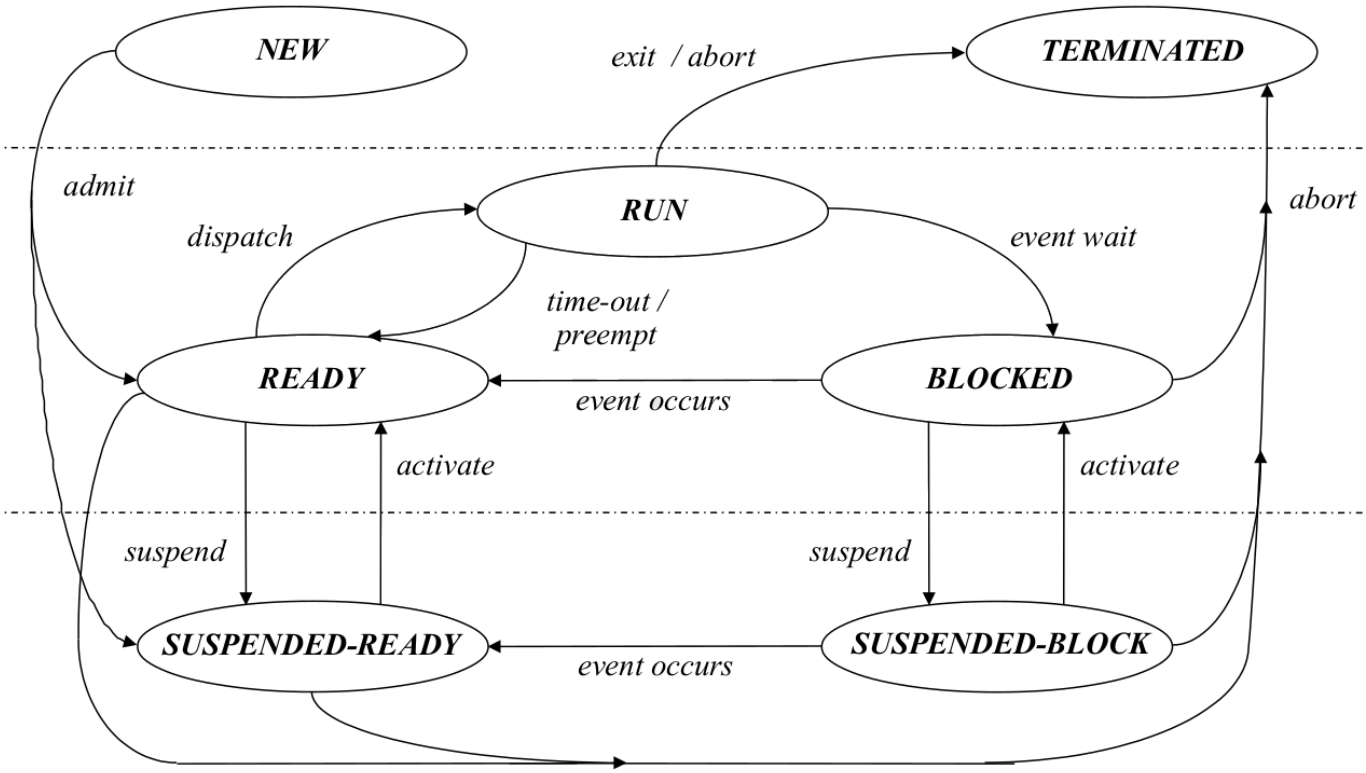


Figure 43: Diagrama de estados completo para um processador `multithreading`

Os processadores atuais possuem **dois modos de funcionamento:**

1. `supervisor mode`
   - Todas as instruções podem ser executadas
   - É um modo **privilegiado**, **reservado para o sistema operativo**
   - O modo em que o **sistema operativo devia funcionar**, para garantir que pode aceder a todas as funcionalidades do processador
2. `user mode`
   - Só uma **secção reduzida do instruction set** é que pode ser executada
   - Instruções de I/O e outras que modifiquem os registos de controlo não são executadas em `user mode`
   - É o **modo normal de operação**

A **troca entre os dois modos de operação**, `switching`, só é possível através de `exceções`. Uma `exceção` pode ser causada por:

- Interrupção de um dispositivo de I/O
- Instrução ilegal
  - divisão por zero
  - bus error
  - …
- trap instruction (aka interrupção por *software*)

As **funções do `kernel`**, incluindo as `system calls` só podem ser lançadas por:

- **hardware** $\Longrightarrow$ `interrupção`
- **traps** $\Longrightarrow$ `intreeupção por software`

O ambiente de operação nestas condições é denominado de `exception handling`

58.1 Exception Handling

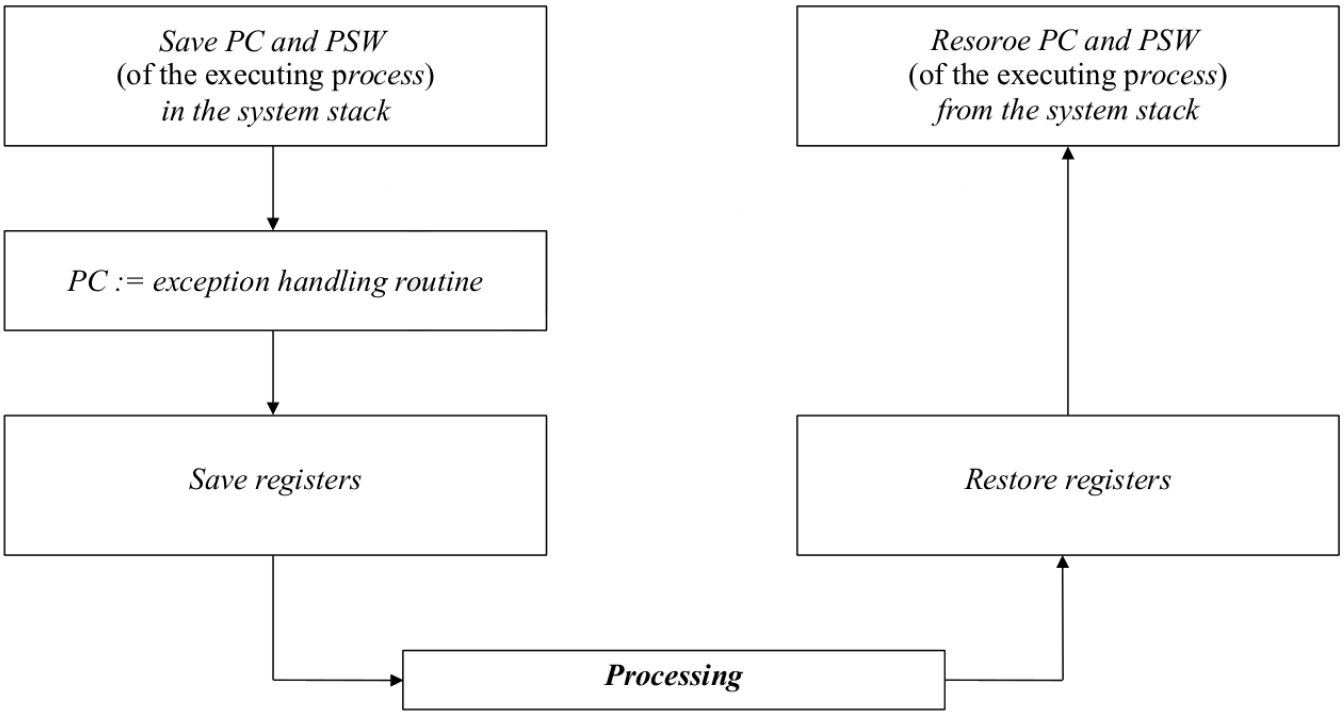


Figure 44: Algoritmo a seguir para tratar de exceções normais

A **troca do contexto de execução** é feita guardando o estado dos registos PC e PSW na stack do sistema, saltando para a rotina de interrupção e em seguida salvaguardando os registos que a rotina de tratamento da exceção vai precisar de modificar. No fim, os valores dos registos são restaurados e o programa resume a sua execução

58.2 Processing a process switching

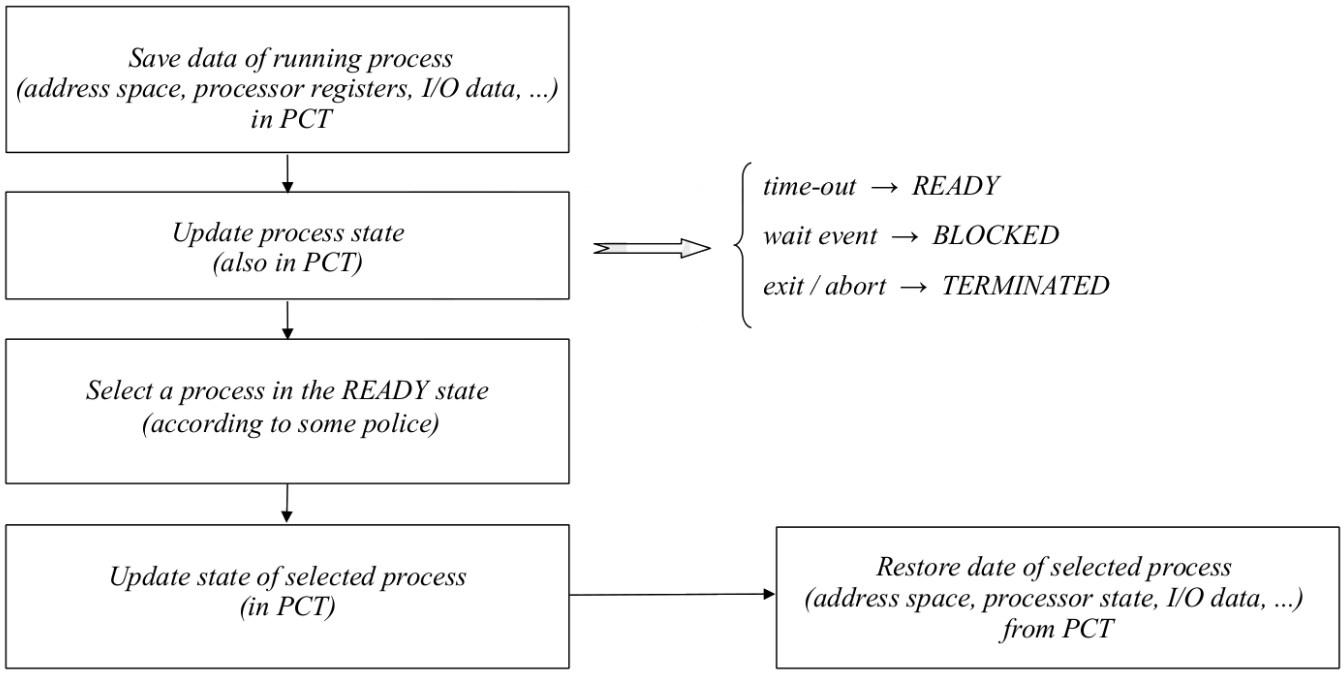


Figure 45: Algoritmo a seguir para efetuar uma process switching

O algoritmo é bastante parecido com o tratamento de exceções:

1. Salvaguardar todos os dados relacionados com o processo atual
2. Efetuar a troca para um novo processo
3. Correr esse novo processo
4. Restaurar os dados e a execução do processo anterior

59 Processor Scheduling

A execução de um processo é uma sequência alternada de períodos de:

- `CPU burst`, causado pela execução de instruções do CPU
- `I/O burst`, causados pela espera do resultado de pedidos a dispositivos de I/O

O processo pode então ser classificado como:

- `I/O bound` se possuir muitos e curtos `CPU bursts`
- `CPU bound` se possuir poucos e longos `CPU bursts`

O objetivo da **multiprogramação** é obter vantagem dos períodos de `I/O burst` para permitir outros processos terem acesso ao processador. A componente do sistema responsável por esta gestão é o `scheduler`.

A funcionalidade principal do `scheduler` é decidir da `poll` de processos prontos para serem executados que coexistem no sistema:

- quais é que devem ser executados?

- quando?
- por quanto tempo?
- porque ordem?

59.1 Scheduler

Revisitando o diagrama de estados do processador, identificamos três `schedulers`

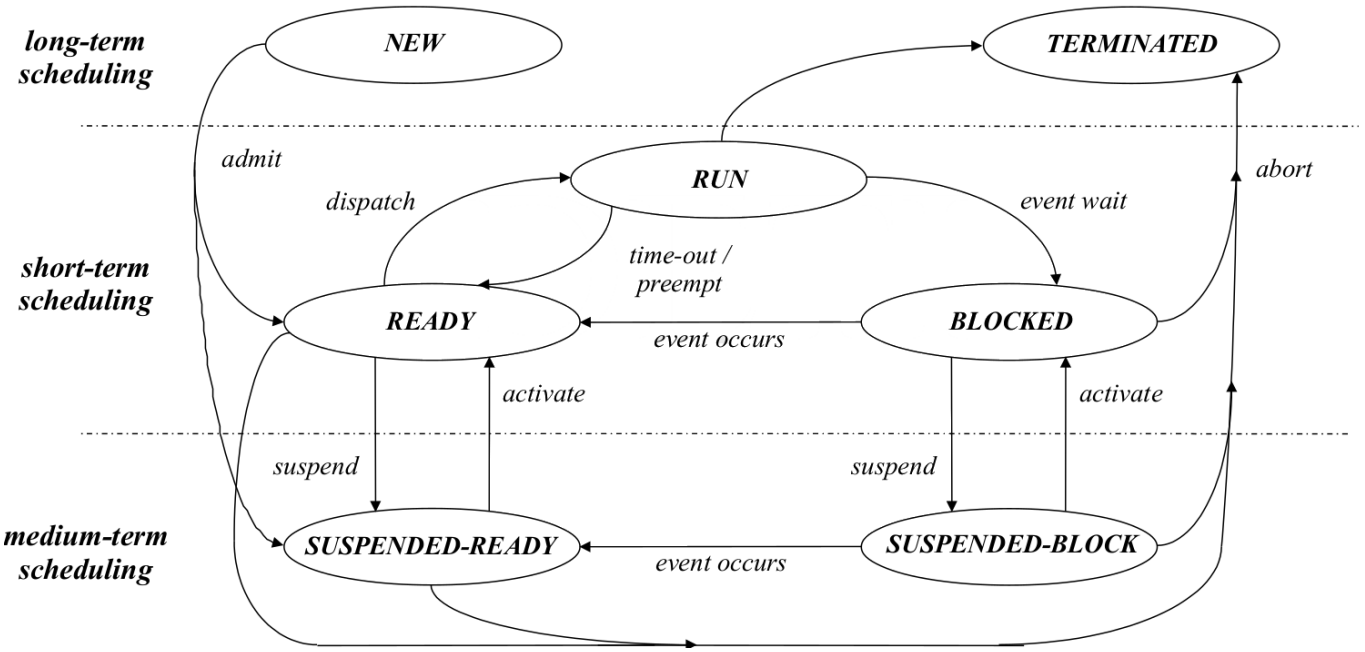


Figure 46: Identificação dos diferentes tipos de schedulers no diagrama de estados dos processos

59.1.1 Long-Term Scheduling

Determina que **programas são admitidos para serem processados**:

- Controla o **grau de multiprogramação** do sistema
- Se um programa do utilizador ou `job` for aceite, torna-se um **processo** e é adicionado à ***queue* de processos `ready` em fila de espera**
  - Em princípio é adicionado à `queue` do **`short`**`-term scheduler`
  - mas também é possível que seja adicionada à `queue` do `medium-term scheduler`
- Pode colocar processos em `suspended ready`, libertando quer a memória quer a fila de processos

59.1.2 Medium Term Scheduling

Gere a `swapping area`

- As decisões de `swap-in` são **controladas pelo grau de multiprogramação**
- As decisões de `swap-in` são **condicionadas pela gestão de memória**

59.1.3 Short-Term Scheduling

Decide qual o **próximo processo a executar**

- É invocado quando existe um evento que:
  - **bloqueia o processo atual**
  - **permite que este seja `preempted`**
- Eventos possíveis são:
  - interrupção de relógio
  - interrupção de I/O
  - `system calls`
  - signal (e.g. através de semáforos)

59.2 Critérios de Scheduling

59.2.1 User oriented

Turnaround Time:

- Intervalo de de tempo entre a submissão de um processo até à sua conclusão
- Inclui:
  - Tempo de execução enquanto o processo tem a posse do CPU
  - Tempo dispendido à espera pelos recursos que precisa (inclui o processador)
- Deve ser minimizado em sistemas `batch`
- É a medida apropriada para um `batch job`

Waiting Time:

- Soma de todos os períodos de tempo em que o processo esteve à espera de ser colocado no estado `ready`
- Deve ser minimizado

Response Time:

- Intervalo de tempo que decorre desde a submissão de um pedido até a resposta começa a ser produzida
- Medida apropriada para sistemas/processos interativo
- Deve ser minimizada para este tipo de sistemas/processos
- O número de processos interativos deve ser maximizado desde que seja garantido um tempo de resposta aceitável

Deadlines:

- Tempo necessário para um processo terminar a sua execução
- Usado em sistemas de tempo real
- A percentagem de `deadlines` atingidas deve ser maximizada, mesmo que isso implique subordinar/reduzir a importância de outras objetivos/parâmetros do sistema

Predictability:

- Quantiza o impacto da carga (de processos) no tempo de resposta dos sistema
- Idealmente, um `job` deve correr no **mesmo intervalo de tempo** e gastar os **mesmos recursos de sistema** independentemente da carga que o sistema possui

59.2.2 System oriented

Fairness:

- Igualdade de tratamento entre todos os processos
- Na ausência de diretivas que condicionem os processos a atender, deve ser efetuada um gestão e partilha justa dos recursos, onde todos os processos são tratados de forma equitativa
- Nenhum processo pode sofrer de `starvation`

Throughput:

- Medida do número de processos completados por unidade de tempo ("taxa de transferência" de processos)
- Mede a quantidade de trabalho a ser executada pelos processos
- Deve ser maximizado
- Depende do tamanho dos processos e da **política de escalonamento**

Processor Utilization:

- Mede a percentagem que o processador está ocupado
- Deve ser maximizada, especialmente em sistemas onde predomina a partilha do processador

Enforcing Priorities:

- Os processos de **maior prioridade** devem ser sempre favorecidos em detrimento de processos menos prioritários

É **impossível favorecer todos os critérios em simultâneo**

Os **critérios a favorecer** dependem da **aplicação específica**

59.3 Preemption & Non-Preemption

Non-preemptive scheduling:

- O processo mantêm o processador até este ser bloqueado ou terminar
- As **transições são sempre por `time-out`**
- Não existe `preempt`
- Típico de sistemas `batch`
  - Não existem deadlines nem limitações temporais restritas a cumprir

Preemptive Scheduling:

- O processo pode **perder o processador devido a eventos externos**
  - esgotou o seu `time-quantum`
  - um processo **mais prioritário** está `ready`
  - Tipico de **sistemas interativos**
    * É preciso garantir que a resposta ocorre em intervalos de tempo limitados
    * É preciso "simular" a ideia de paralelismo no `macro-tempo`
  - Sistemas em tempo real são `preemptive` porque existem `deadlines` **restritas** que precisam de ser cumpridas
  - Nestas situações é importante que um **evento externo** tenha capacidade de libertar o processador

59.4 Scheduling

59.4.1 Favouring Fearness

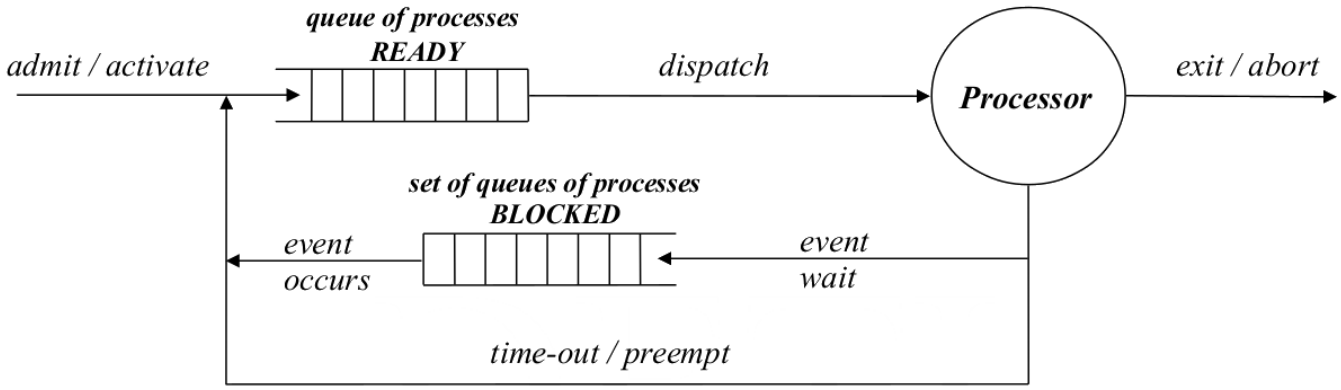


Figure 47: Espaço de endereçamento de um processo em Linux

Todos os processos **são iguais** e são atendidos por **ordem de chegada**

- É implementado usando **FIFOs**
- Pode existir mais do que um processo à espera de eventos externos
- Existe uma fila de espera para cada evento
- Fácil de implementar
- Favorece processos `CPU-bound` em detrimento de processos `I/O-bound`
  - Só necessitam de acesso ao processador, não de recursos externos
  - Se for a vez de um processo `I/O-bound` ser atendido e não possuir os recursos de I/O que precisa tem de voltar para a fila
- Em **sistemas interativos**, o `time-quantum` deve ser escolhido cuidadosamente para obter um bom compromisso entre `fairness` e `response time`

Em função do scheduling pode ser definido como:

- `non-preemptive scheduling` $\implies$ `first come, first-served` (FCFS)
- `preemptive scheduling` $\implies$ `round robin`

59.4.2 Priorities

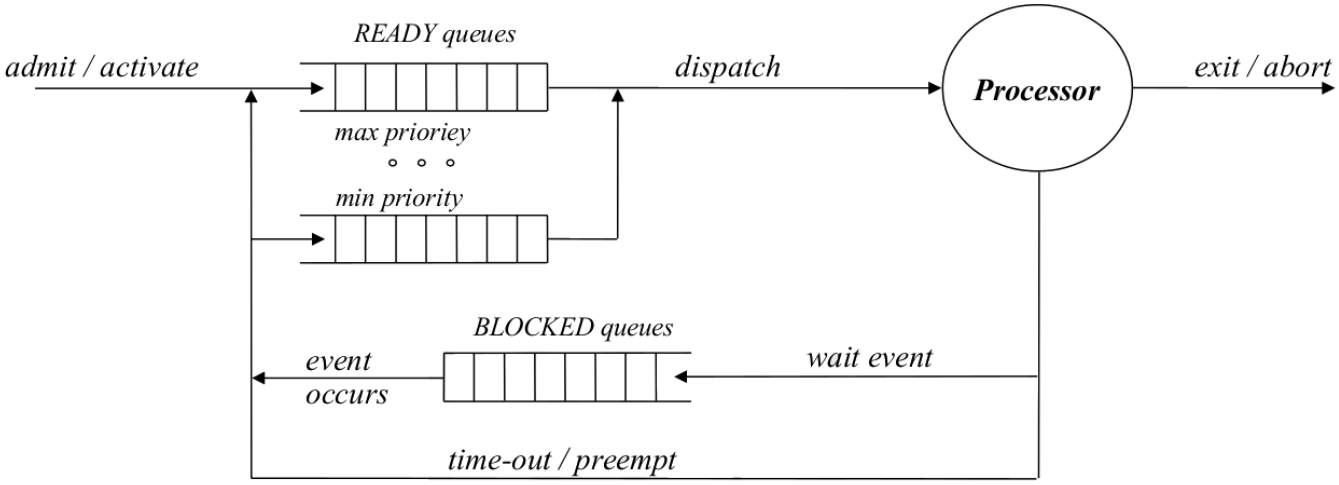


Figure 48: Espaço de endereçamento de um processo em Linux

Segue o princípio de que atribuir a mesma importância a todos os processos pode ser uma solução errada. Um sistema injusto *per se* não é necessariamente mau.

- A **minimização do tempo de resposta** (`response time`) exige que os processos `I/O-bound` sejam **privilegiados**
- Em **sistemas de tempo real**, os processos associados a **eventos/alarmes** e **ações do sistema operativo** sofrem de várias **limitações e exigências temporais**

Para resolver este problema os processos são **agrupados** em grupos de **diferentes prioridades**

- Processos de maior prioridade são executados primeiros
- Processos de menor prioridade podem sofrer `starvation`

Prioridades Estáticas

As prioridades a atribuir a cada processo são determinadas *a priori* de forma **determinística**

- Os processos são **agrupados em classes de prioridade fixa**, de acordo com a sua importância relativa
- Existe risco de os processos menos prioritários sofrerem `starvation`
  - Mas se um **processo de baixa prioridade não é executado** é porque o sistema foi **mal dimensionado**
- É o sistema de `scheduling` mais injusto
- É usado em sistemas de tempo real, para garantir que os processos que são críticos são sempre executados

Alternativamente, pode se fazer:

1. Quando um processo é criado, é lhe **atribuído um dado nível de prioridade**
2. Em `time-out` a prioridade do processo é **decrementada**
3. Na ocorrência de um `wait event` a prioridade é **incrementada**
4. Quando o valor de **prioridade atinge um mínimo**, o valor da prioridade sofre um `reset`
   - É colocada no valor inicial, garantindo que o processo é executado

Previnem-se as situações de `starvation` impedindo que o processo não acaba por ficar com uma prioridade tão baixa que nunca mais consegue ganhar acesso

Prioridades Dinâmicas

- As classes de prioridades estão definidas de forma funcional *a priori*
- A mudança de um processo de classe é efetuada com base na utilização última janela de execução temporal que foi atribuída ao processo

Por exemplo:

- **Prioridade 1:** `terminais`
  - Um processo entra nesta categoria quando se efetua a transição `event occurs` (evento de escrita/leitura de um periférico) quando estava à espera de dados do `standard input device`
- **Prioridade 2:** `generic I/O`
  - Um processo entra nesta categoria quando efetua a transição `event occurs` se estava à espera de dados de **outro tipo de input device** que não o `stdin`
- **Prioridade 3:** `small time quantum`
  - Um processo entra nesta classe quando ocorre um `time-out`
- **Prioridade 4:** `large time quantum`
  - Um processo entra nesta classe após um sucessivo número de `time-outs`
  - São claramente processos `CPU-bound` e o objetivo é atribuir-lhes janelas de execução com grande `time quantum`, mas menos vezes

Shortest job first (SJF) / Shortest process next (SPN)

Em sistemas `batch`, o `turnaround time` deve ser minimizado.

Se forem conhecidas **estimativas do tempo de execução** *a priori*, é possível estabelecer uma **ordem de execução** dos processos que **minimizam o tempo de `turnaround` médio** para um dado grupo de processos

Assumindo que temos N `jobs` e que o tempo de execução de cada um deles é te_n, com $n = 1, 2, ..., N$. O `average turnaround time` é:

$$t_m = te_1 + \frac{N-1}{N} \cdot te_2 + ... + \frac{1}{N} \cdot te_N$$

onde t_m é o `turnaround time` mínimo se os `jobs` forem sorteados por ordem ascendente de tempo de execução (estimado)

Para **sistemas interativos**, podemos usar um sistema semelhante:

- Estimamos a taxa de ocupação da próxima janela de execução baseada na taxa de ocupação das janelas temporais passadas
- Atribuímos o processador ao processo cuja estimativa for a **mais baixa**

Considerando fe_1 como sendo a **estimativa da taxa de ocupação** da primeira janela temporal atribuída a um processo e f_1 a fração de tempo efetivamente ocupada:

- A estimativa da segunda fração de tempo necessária é

$$fe_2 = a \cdot fe_1 + (1-a) \cdot f_1$$

- A estimativa da e-nésima fração de tempo necessária é:

$$fe_N = a \cdot fe_{N-1} + (1-a) \cdot f_{N-1}$$

  Ou alternativamente:

$$a^{N-1} \cdot fe_1 + a^{N-2} \cdot (1-a) \cdot fe_2 + a \cdot (1-a) \cdot fe_{N-2} + (1-a) \cdot fe_{N-1}$$

Com $a \in [0, 1]$, onde a é um coeficiente que representa o peso que a história passada de execução do processo influencia a estimativa do presente

Esta alternativa levanta o problema que que processos `CPU-bound` podem sofrer de `starvation`. Este problema pode ser resolvido contabilizando o tempo que um processo está em espera (`aging`) enquanto está na fila de processos `ready`

Normalizando esse tempo em função do período de execução e denominando-o R, a **prioridade** de um processo pode ser dada por:

$$p = \frac{1 + b \cdot R}{fe_N}$$

onde b é o coeficiente que **controla o peso do aging** na fila de espera dos processos `ready`

59.5 Scheduling Policies

59.5.1 First Come, First Serve (FCFS)

Também conhecido como `First In First Out` (FIFO). O processo mais antigo na fila de espera dos processos `ready` é o primeiro a ser selecionado.

- `Non-preemptive` (em sentido estrito), podendo ser combinado com um esquema de prioridades baixo
- Favorece processos `CPU-bound` em detrimento de processos `I/O-bound`
- Pode resultar num **mau uso** do processador e dos dispositivos de I/O
- Pode ser utilizado com **low priority schemas**

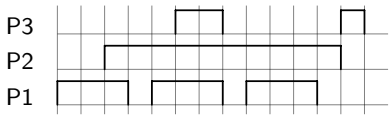


Figure 49: Problema de Scheduling

Usando uma política de `first come first serve`, o resultado do scheduling do processador é:

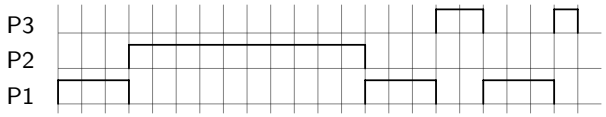


Figure 50: Política FCFS

- O P1 começa a usar o CPU.
- Como é um sistema FCFS, o processo 1 só larga o CPU passado 3 ciclos.
- O processo P2 é o processo seguinte na fila `ready`, e ocupa o CPU durante 10 ciclos.
- Quando P2 termina, P1 é o processo que está à mais tempo à espera, sendo ele que é executado

- Quando P2 abandona voluntariamente o CPU, o processo P1 corre os seus primeiros dois ciclos
- Quando P3 liberta o CPU, o processo P1 termina os últimos 3 ciclos que precisa
- Quando P3 liberta o CPU, o processo P1 como é `I/O-bound` e precisa de 5 ciclos para o dispositivo estar pronto fica mais dois ciclos à espera para puder terminar executando o seu último ciclo

59.5.2 Round-Robin

- `Preemptive`
  - O `scheduler` efetua a gestão baseado num `clock`
  - A cada processo é atribuído um `time-quantum` máximo antes de ser `preempted`
- O processo **mais antigo** em `ready` é o **primeiro a ser selecionado**
  - não são consideradas prioridades
- Efetivo em sistemas `time sharing` com objetivos globais e sistemas que processem transações
- **Favorece** `CPU-bound` em detrimento de processos `I/O-bound`
- Pode resultar num **mau uso de dispositivos I/O**

Na escolha/otimização do `time quantum` existe um **tradeoff:**

- **tempos muito curtos** favorecem a execução de **processos pequenos**
  - estes processos vão ser executados **rapidamente**
- **tempos muito curtos** obrigam a `processing overheads` devido ao `process switching` **intensivo**

Para os processos apresentados acima, o diagrama temporal de utilização do processador, para um `time-quantum` de 3 ciclos é:

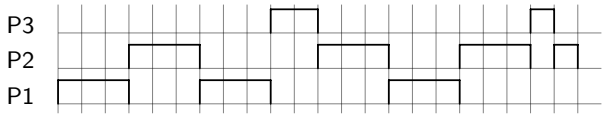


Figure 51: Política Round-Robin

A história de processos em `ready` em fila de espera é: 2, 1, 3, 2, 1, 2, 3, 1

59.5.3 Shortest Process Next (SPN) ou Shortest Job First (SJF)

- `Non-preemptive`
- O process com o `shortest CPU burst time` (menor tempo espectável de utilização do CPU) é o **próximo a ser selecionado**
  - Se vários processos tem o **mesmo tempo de execução** é usado FCFS para desempatar
- Existe um **risco de** **`starvation`** para grandes processos
  - o seu acesso ao CPU pode ser **sucessivamente adiado** se existir "forem existindo" processos com **tempo de execução menor**
- Normalmente é usado em escalonamento de logo prazo, **`long`**`-term scheduling` em sistemas `batch`, porque os utilizadores esperam estimar com precisão o tempo máximo que o processo necessita para ser executado

59.5.4 Linux

No Linux existem 3 classes de prioridades:

1. **FIFO**, `SCHED_FIFO`
   - `real-time threads`, com política de prioridades
   - uma `thread` em execução é `preempted` apenas se um processo de **mais alta prioridade da mesma classe** transita para o estado `ready`
   - uma `thread` em execução pode **voluntariamente abandonar o processador**, executando a primitiva `sched_yeld`
   - dentro da mesma classe de prioridade a política escolhida é `First Come, First Serve` (FCFS)
   - Só o `root` é que pode lançar processos em modo FIFO
2. **Round-Robin real time threads**, `SCHED_RR`
   - `threads` com prioridades com necessidades de execução em tempo real
   - Processos nesta classe de prioridades são `preempted` se o seu `time-quantum` termina
3. **Non real time threads**, `SCHED_OTHER`
   - Só são executadas se não existir nenhuma `thread` com necessidades de execução em tempo real
   - Está associada à processos do utilizador
   - A política de escalonamento tem mudado à medida que a são lançadas novas versões do *kernel*

A **escala de prioridades** varia

- 0 a 99 para `real-time threads`
- 100 a 139 para as restantes

Para lançar uma `thread` (sem necessidades de execução em tempo real) com diferentes prioridades, pode ser usado comando `nice`.

Por *default*, o comando lança uma `thread` com prioridade 120. O comando aceita um `offset` de [-20, +19] para obter a prioridade mínima ou máxima.

Algoritmo Tradicional

- Na classe `SCHED_OTHER` as prioridades são baseadas em **créditos**
- Os créditos do processo em execução são **decrementados** à medida que ocorre uma interrupção do `real time clock`
- O processo é `preempted` quando são atingidos zero créditos
- Quando todos os processos `ready` têm zero créditos, os créditos de **todos os processos** (incluindo os que estão bloqueados) são **recalculados** segundo a fórmula:

$$CPU_j(i) = \frac{CPU_j(i-1)}{2} + PBase_j + nice_j$$

  onde são tido em conta a **história passada de execução do processo** e as **prioridades**
- O `response time` de processos `I/O-bound` é minimizado
- A `starvation` de processos `CPU-bound` é evitada
- Solução **não adequada para múltiplos processadores** e é má se o número de processos é elevado

59.6 Novo Algoritmo

- Os processos na classe `SCHED_OTHER` passam a usar um `completely fair scheduler (CFS)`
- O scheduling é baseado no `vruntime`, *virtual run time*, que mede durante quanto tempo uma `thread` esteve em execução
  - o `virtual run time` está relacionado quer com o **tempo de execução real** (`physical run time`) e a **prioridade** da `thread`
  - Quanto maior a prioridade de um processo, menor o `physical run time`
- O `scheduler` seleciona as `threads` com menor `virtual run time`
  - Uma `thread` com prioridade mais elevada que fique pronta a ser executada pode "forçar" um `preempt` uma `thread` com menor prioridade
    * Assim é possível que uma `thread I/O bound` "forçar" o processador a `preempt` um processo `CPU-bound`
- O algoritmo é implementado com base numa `red-black tree` do processador

60 Introdução à Gestão de Memória

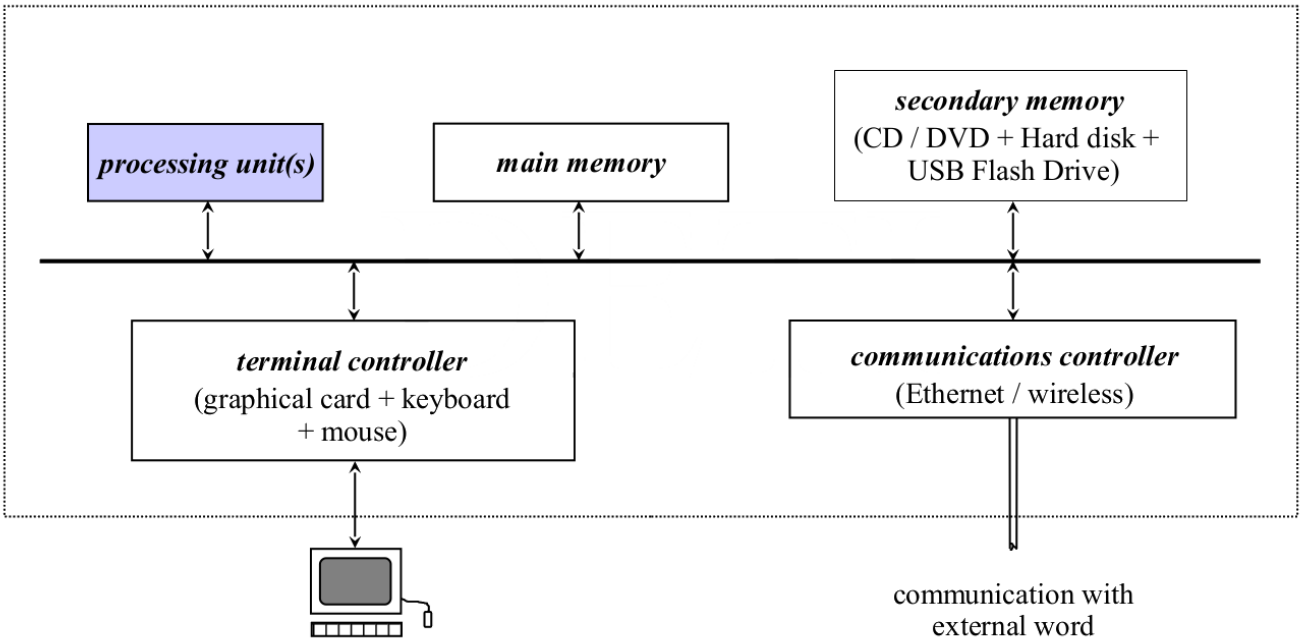


Figure 52: Relembrando o diagrama de um sistema Computacional

60.1 Porquê a gestão de memória

- Para puder executar um programa este tem de residir em **memória principal**
  - As varáveis, instruções, etc. tem de estar na memória principal, pelo menos de forma parcial
- É necessário maximizar a ocupação do processador e minimizar o tempo de resposta (`turn-around time`)
  - Ambiente **multiprogramado**
  - Ocorre **comutação de processos**
  - Existem **vários processos em memória**

> **Lei de Parkinson** *"Os programas tendem a expandir-se ocupando toda a memória disponível"*

Ou seja, apesar de o espaço disponível em memória principal ter aumentado ao longo dos anos, os mesmos problemas mantêm-se.

Supondo que a **fração de ocupação do processador** pode ser modelada de forma simplificada pela expressão

$$\%_{ocupaoCPU} = 1 - p^n$$

onde:

- p: fração de tempo em que um processo está **bloqueado à espera** que as operações de I/O, sincronização, etc terminem
- n: **número de processos** que **coexistem** de forma concorrente e a competir por recursos em memória principal

Supondo $p = 0.8$, temos:

Nº de processos em Memória Principal	% de ocupação do Processador
4	59
8	83
12	93
16	97

De forma mais geral:

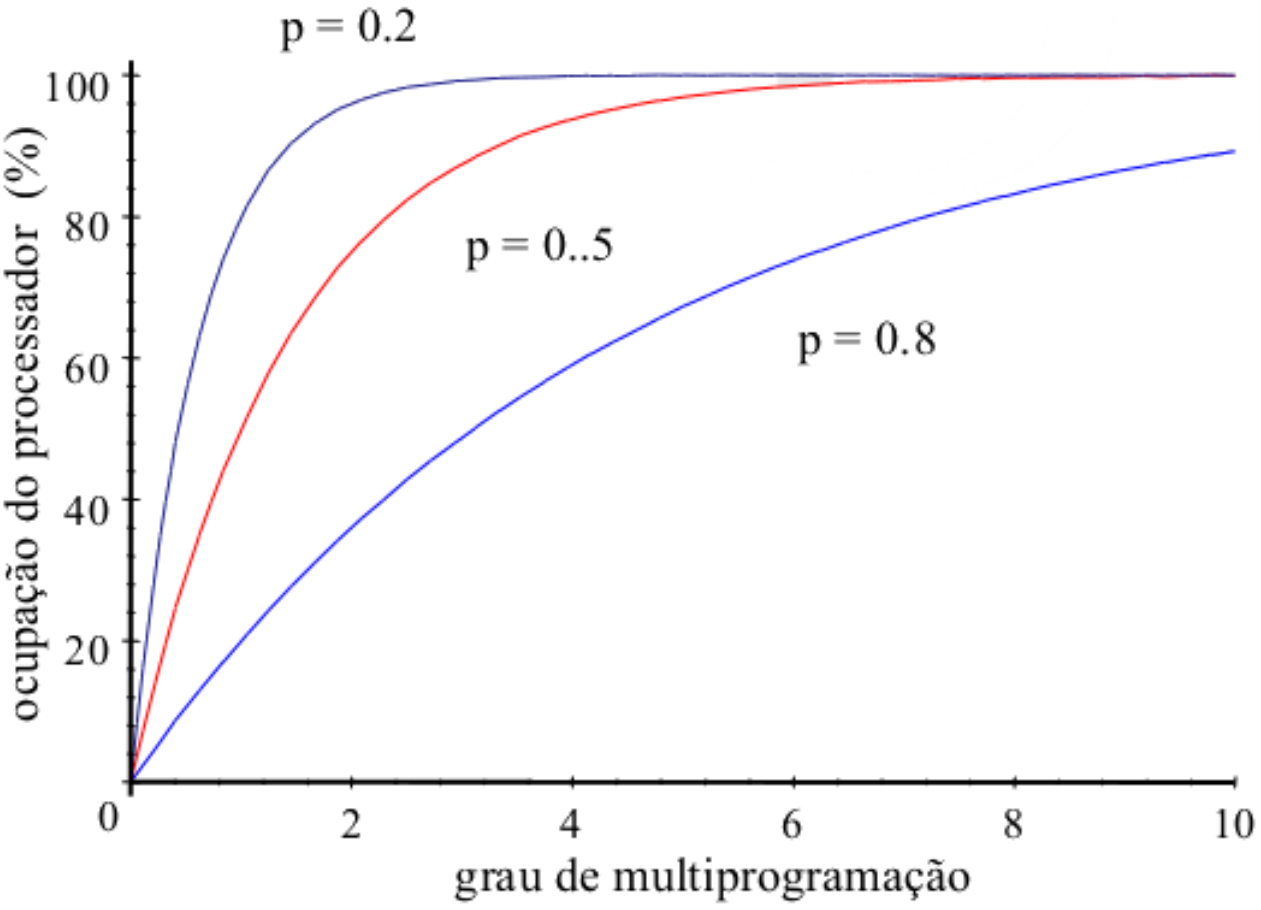


Figure 53: Grau de ocupação do processador em função do número de processos concorrentes residentes em memória principal em simultâneo

O número de processos que devem estar em memória têm de ser otimizados. O número de processos em memória depende do número de processos I/O intensivos (`I/O-bound`) ou CPU intensivos (`CPU-bound`)

60.2 Hierarquia da memória

Para melhorar a eficiência do sistema e reduzir os custos, as memórias devem ser otimizadas para as funções que vão desempenhar:

	Cache	Principal	Secundária
tamanho	pequena (dezenas de KB ou unidades de MB)	tamanho médio (centenas de MB ou unidades de MB)	Grande (dezenas, centenas ou milhares de GB)
velocidade	muito rápida	rápida	lenta
preço	cara	razoável	barata
volátil	✓	✓	x

Table 10: Comparação entre os diferentes tipos de memórias de um sistema computacional

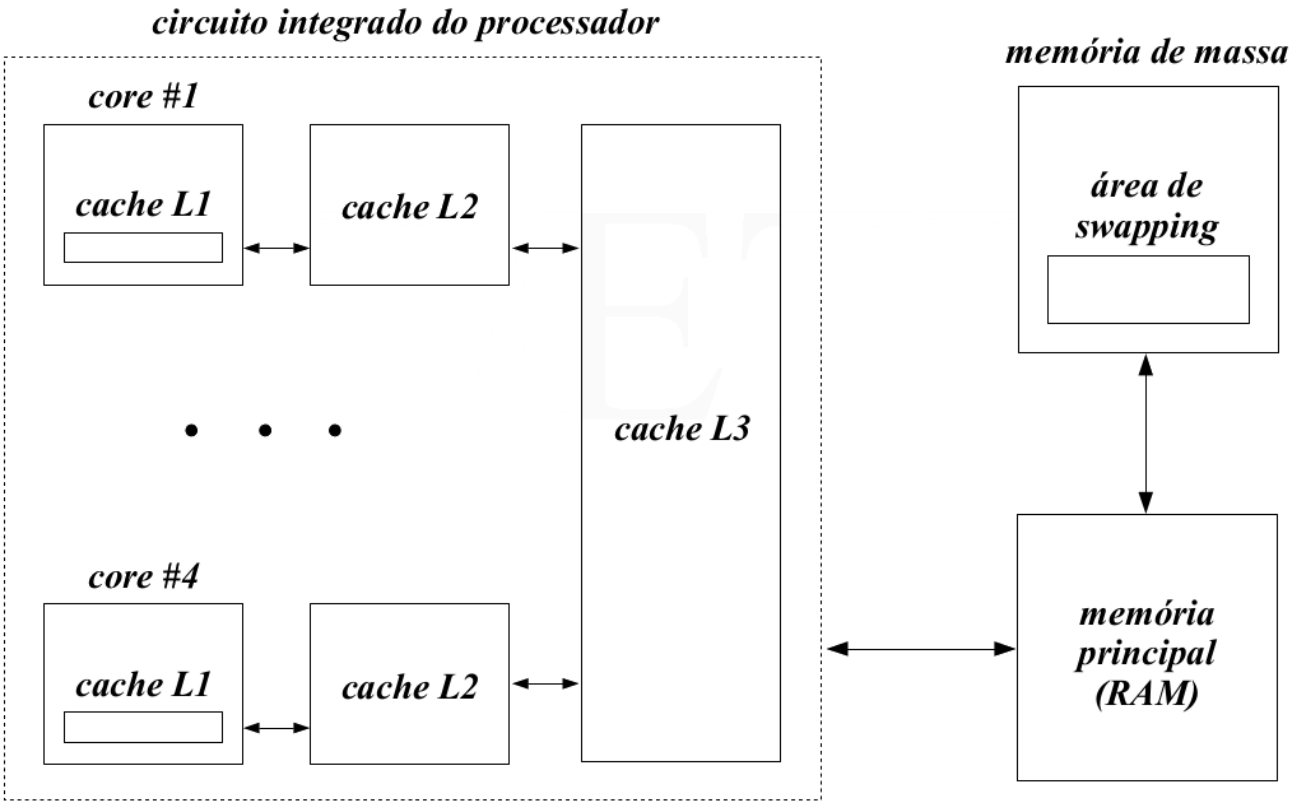


Figure 54: Hierarquia da Memória num sistema de computação

60.2.1 Memória Cache

- Contém uma cópia das **posições** e **operandos** mais frequentemente referenciadas pelo processador num passador próximo
- Existem 3 tipos de `memória cache`
  - **L1:** localizada no **IC**[7] **do processador**
  - **L2** e **L3:** localizadas num **IC autónomo** mas no mesmo substrato que L1

[7] ficheiro em código fonte de compilação separada

- O controlo da transferência de dados de/para a memória principal é feito de modo quase completamente **transparente** ao programador
- É útil devido ao **princípio da localidade de referência**

60.2.2 Memória Secundária

Duas funções principais:

- Armazenar de forma **não volátil** a informação (dados e programas), através de um **sistema de ficheiros** implementado no dispositivo
- **extender a memória principal** para que o tamanho desta não seja limitativo ao número de processos que podem coexistir em memória - `área de wapping`

60.2.3 Princípio da Localidade da Referência

Temporal: > Quando é acedido um endereço de memória (quer seja para r/w uma variável ou para ler uma instrução), a probabilidade de voltar a aceder a esse mesmo endereço de memória é mais elevada do que aceder a outros endereços de memória

Espacial: > Quando é acedido um endereço de memória (quer seja +ara r/w de uma variável ou para ler uma instrução), a probabilidade de aceder a offsets do endereço de memória é mais elevada do que aceder a endereços muito distantes

Estes princípios baseiam-se no facto de que quanto **mais afastada** uma instrução/operando está do endereço atual que o processador está a executar, **menos vezes será referenciado**.

O uso destes princípios no design de software e hardware tem como objetivo diminuir o tempo médio de acesso à referência.

Estes princípios foram derivados da **constatação heurística** do comportamento de um programa em execução. Conclui-se que as referências à memória durante a sua execução tendem a **concentrar-se** em **frações bem definidas do seu espaço de endereçamento** em intervalos de tempo mais ou menos longos

60.3 Gestão da memória num ambiente multiprogramado

- **objetivo:** Tirar partido dos princípios anteriores

A função principal é **controlar a transferência de dados** entre a **memória principal** e a **memória secundária**, garantindo:

- Manter o *track* das partes da **memória principal** que estão **ocupadas** e as partes que estão **livres**
- Reservar secções da memória para as necessidades dos processos
- Libertar secções da memória quando os processos terminam
- Transferir para a `área de swapping` a totalidade/parte do **espaço de endereçamento** de um processo quando a memória principal não consegue guardar todos os processos que coexistem em memória

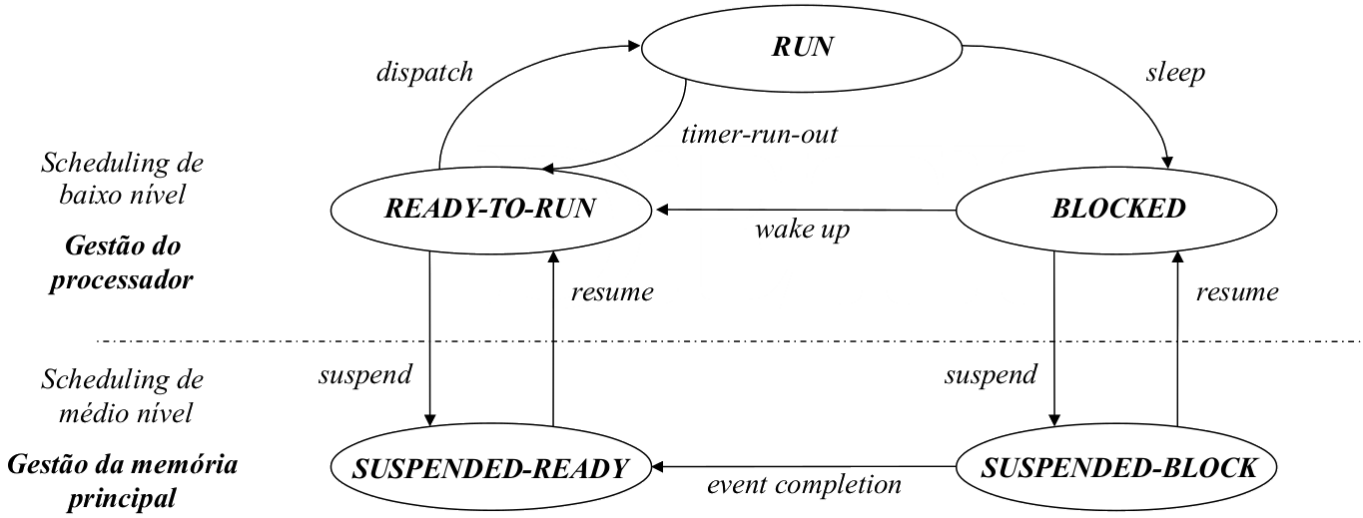


Figure 55: Diagrama da inclusão da gestão de memória com o scheduling de baixo nível do processador

60.4 Espaço de Endereçamento

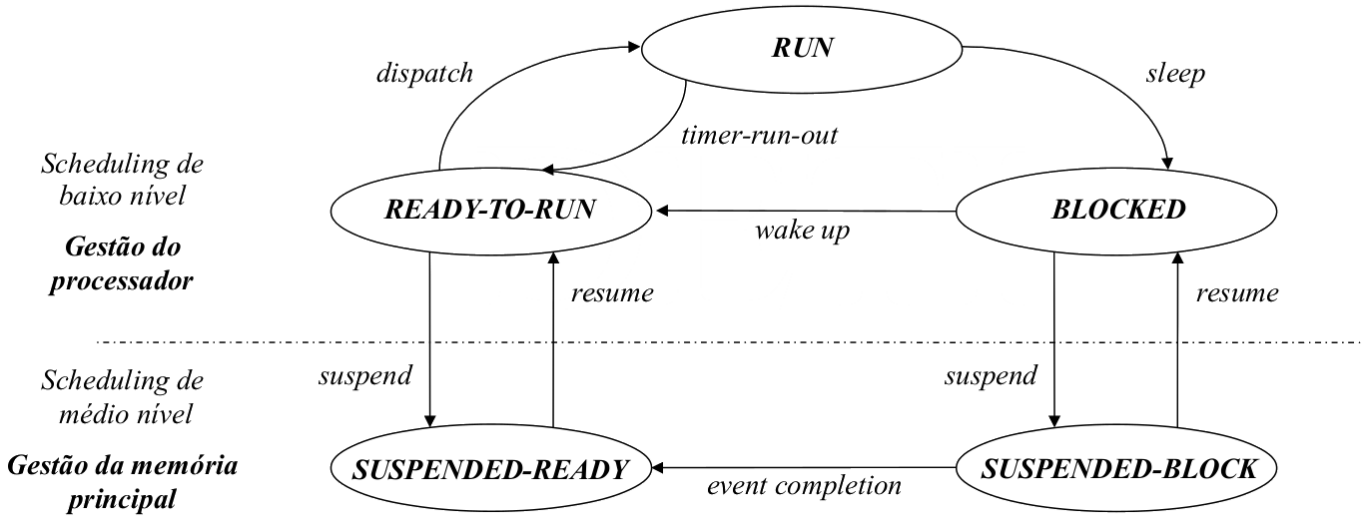


Figure 56: Construção do espaço de endereçamento de um programa após compilação e linkagem

- Os ficheiros `object` (resultantes da compilação), possuem todos os seus endereços das diversas instruções, constantes e variáveis calculados a partir do endereço 0 (início dos endereços do módulo)

Se a linkagem for **estática:**

- Após linkagem, os diferentes ficheiros objeto são reunidos num único ficheiro executável
  - São resolvidas as várias referências externas

  - As bibliotecas de sistema podem não estar incluídas na linkagem para minimizar o tamanho do ficheiro

- O `loader` constrói a **imagem binária do espaço de endereçamento do processo**
  - ficheiro executável + bibliotecas de sistema
  - Resolve todas as dependências externas que não foram incluídas no ficheiro executável no processo de linkagem

Se a linkagem for **dinâmica**, cada referência no código do processo é substituída por um `stub`:

- `stub`: pequeno conjunto de instruções que determina a localização de uma rotina
  - se a rotina estiver em memória principal, executa-a
  - se não estiver, força o seu carregamento para memória principal (e depois executa-a)
- Quando um `stub` é executado **pela primeira vez** obtém a referência para o endereço de memória da rotina
  - Substitui no **código do processo** o seu endereço pelo endereço da rotina
  - Executa a rotina
- Quando a secção de código onde era executado o `stub` é novamente atingida, a **rotina do sistema** é executada **diretamente**

Como vários processos independentes podem executar a mesma biblioteca do sistema, ao todos executarem uma cópia do código **minimiza-se** a ocupação da **memória principal**

Figure 57: Diagrama da divisão do espaço de endereçamento de um programa

- As zonas de de código e definição estática de variáveis têm um **tamanho fixo**
  - Determinado pelo `loader`
- As `zonas de definição dinâmica` e a `stack` podem variar de tamanho ao longo da execução do programa
  - O espaço de memória referente à `zona de definição dinâmica` e à `stack` pode ser usada alternativamente entre eles
  - Quando o espaço para a `stack` cresce e o espaço disponível é esgotado pelo **lado da `stack`**, ocorre um `stack overflow`

60.4.1 Exemplo

Considerando o seguinte código fonte

```c
//ficheiro fonte
#include <stdio.h>
#include <stdlib.h>

int main (void)
{
    printf ("hello, world!\n");
    exit (EXIT_SUCCESS);
}
```

A produção do ficheiro objeto pode ser feita com:

```
gcc -Wall -c hello.c
```

E o executável gerado através de:

```
gcc -o hello hello.o
```

Antes da linkagem temos:

```
>> file hello.o
hello.o: ELF 32-bit LSB relocatable, Intel 80386, version 1 (SYSV), not stripped
```

```
>> objdump -fstr hello.o

hello.o:      file format elf32-i386
architecture: i386, flags 0x00000011:
HAS_RELOC, HAS_SYMS
start address 0x00000000
SYMBOL TABLE:
00000000 l     df *ABS* 00000000 hello.c
00000000 l     d .text            00000000
00000000 l     d .data            00000000
00000000 l     d .bss             00000000
00000000 l     d .rodata          00000000
00000000 l     d .note.GNU-stack 00000000
00000000 l     d .comment         00000000
```

```
00000000 g     F .text  0000002a main
00000000         *UND*  00000000 printf
00000000         *UND*  00000000 exit

RELOCATION RECORDS FOR [.text]:
OFFSET   TYPE              VALUE
00000014 R_386_32          .rodata
00000019 R_386_PC32        printf
00000026 R_386_PC32        exit

Contents of section .rodata:
 0000 68656c6c 6f20776f 726c6421 0a000000   hello world!....

Contents of section .comment:
 0000 00474343 3a202847 4e552920 332e332e   .GCC: (GNU) 3.3.
 0010 3120284d 616e6472 616b6520 4c696e75   1 (Mandrake Linu
 0020 7820392e 3220332e 332e312d 326d646b   x 9.2 3.3.1-2mdk
 0030 2900                                  ).
$
```

Após a linkagem temos:

```
>> file hello
hello: ELF 32-bit LSB executable, Intel 80386, version 1 (SYSV), for GNU/Linux
2.2.5, dynamically linked (uses shared libs), not stripped
```

```
>> objdump -fTR hello

hello:     file format elf32-i386
architecture: i386, flags 0x00000112:
EXEC_P, HAS_SYMS, D_PAGED
start address 0x080482d0
DYNAMIC SYMBOL TABLE:
0804829c      DF *UND*  000000e6 GLIBC_2.0      __libc_start_main
080482ac      DF *UND*  0000002d GLIBC_2.0      printf
080482bc      DF *UND*  000000c8 GLIBC_2.0      exit
080484b4 g    DO .rodata         00000004 Base        _IO_stdin_used
00000000 w    D *UND*  00000000                 __gmon_start__
DYNAMIC RELOCATION RECORDS
OFFSET   TYPE              VALUE
080495cc R_386_GLOB_DAT    __gmon_start__
080495c0 R_386_JUMP_SLOT   __libc_start_main
080495c4 R_386_JUMP_SLOT   printf
080495c8 R_386_JUMP_SLOT   exit
$
```

Podemos verificar que passam a existir mais instruções no ficheiro `object`:

- A função `main` tem de ser executada por alguém…
- É preciso construir o `argv` e o `argc` para passar à `main`
- Implica código adicional

60.4.2 Espaço de endereçamento lógico vs físico

- **espaço de endereçamento lógico:** espaço de endereçamento realocável
- **espaço de endereçamento físico:** região da memória principal onde o processo é carregado para ser executado

Uma vez que a **Imagem binária do espaço de endereçamento de um processo** é mapeada no **espaço lógico** do processo, em sistemas multiprogramados é necessário garantir:

- **mapeamento dinâmico:** capacidade de conversão em `run-time` de um **endereço lógico** num **endereço físico**
  - Passo intermédio para permitir o armazenamento do espaço de endereçamento de um processo em qualquer região da memória principal (incluindo a sua mudança em `run-time`)
- **proteção dinâmica:** impedimento em `run-time` de referenciar endereços que estão localizados fora do espaço de endereçamento do processo (aka, `core dumped`)

61 Arquitecturas de Memória Particionadas

61.1 Arquitectura de partições fixas

- A memória principal (restante) é dividida num **conjunto fixo de partições**
  - Mutualmente exclusivas
  - Não necessariamente iguais
- Cada uma das partições contém o espaço de endereçamento físico de um processo

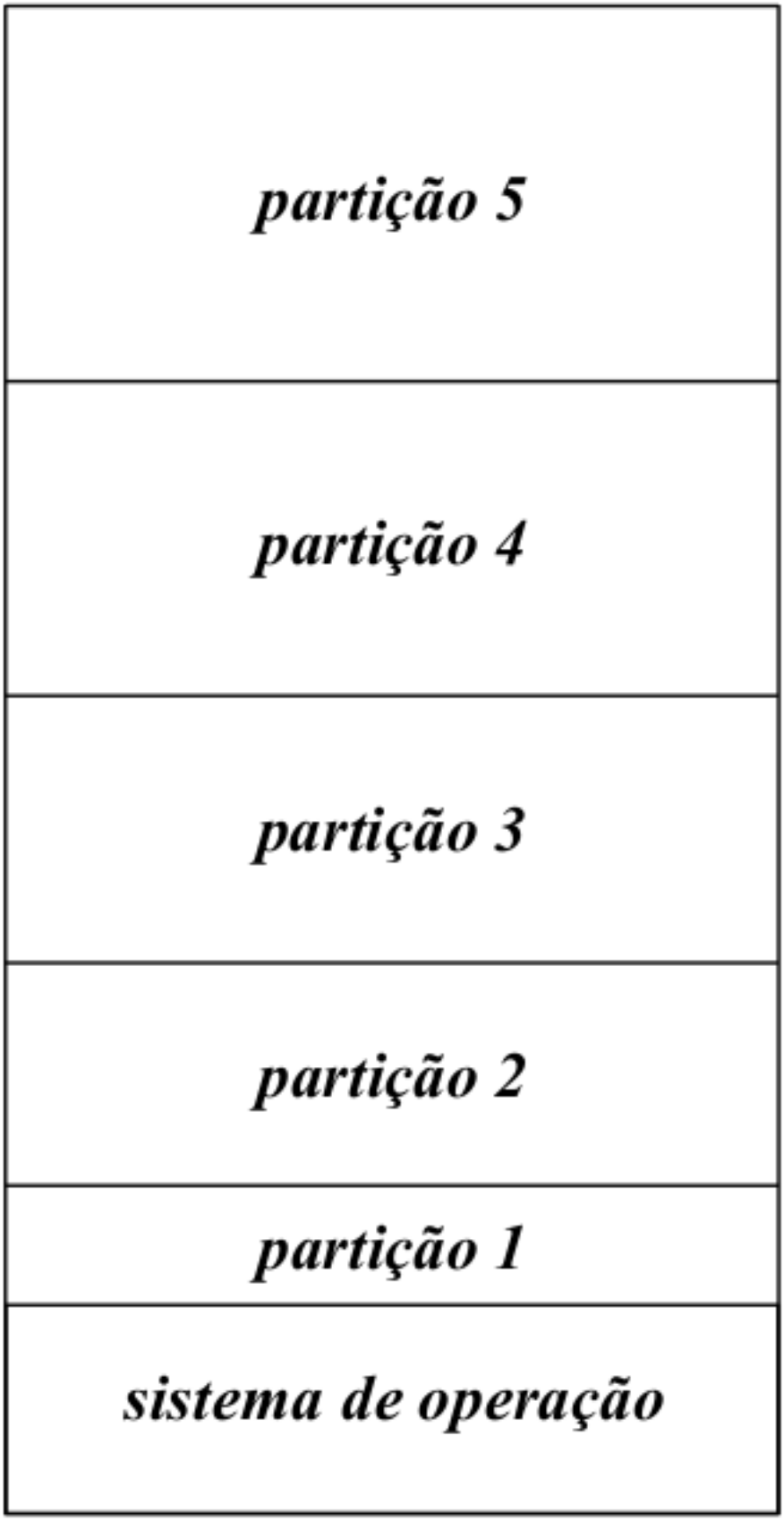


Figure 58: Divisão em partições fixas mutuamente exclusivas com diferentes tamanhos

- A memória principal vai sendo dividida à medida que vai recebendo solicitações
- Podem ser utilizadas diferentes filosofias de escalonamento
  - **Valorização do critério de justiça**
    * Escolher o primeiro processo da fila de espera dos processos `Suspended-Ready` cujo **espaço de endereçamento cabe na partição**
  - **Valorização da ocupação da memória principal**
    * Escolher o primeiro processo da fila de espera dos processos `Suspended-Ready` com o **espaço de endereçamento de tamanho maior** que caiba na partição
    * Corre-se o risco de adiamento indefinido de processos com espaço de endereçamento pequeno (`starvation`)
    * Por isso associa-se um contador a cada processo
      · o contador é incrementado a cada passagem
      · contador > valor pré-definido $\implies$ **processo já não pode ser descartado**
      · Passa-se a aplicar a 1ª regra

61.1.1 Vantagens e Desvantagens

Vantagens:

- `simples de implementar`: não exige hardware ou estruturas de dados especiais para gerir a memória
- `eficiente`: a seleção pode ser feita rapidamente com qualquer das políticas acima

Desvantagens:

- `fragmentação interna da memória principal`: o espaço restante de cada partição que não é alocado pelo processo é desperdiçado
- `política direcionada para certos tipos de aplicações`:
  - O tamanho das partições é fixo
  - A única maneira de ser evita o desperdício de memória é através da adequação do tamanho das partições ao tipo de processos a utilizar
    * número de processos
    * tipo de processos
    * tamanho do seu espaço de endereçamento
  - torna a solução pouco generalizável

61.2 Arquitectura de posições variáveis

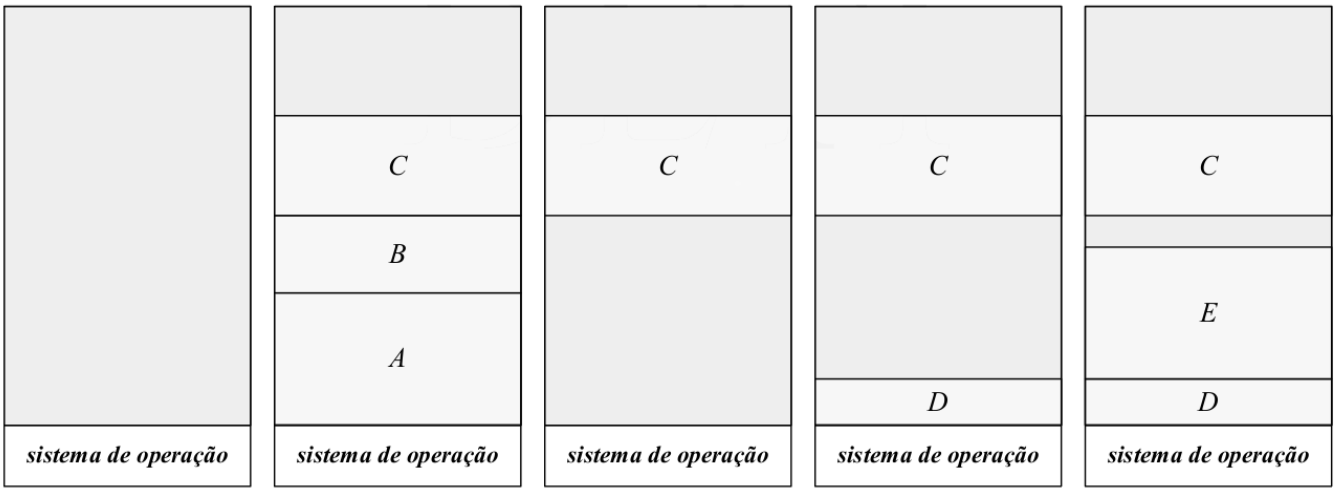


Figure 59: Divisão da memória em partições de tamanho variável

- Toda a parte disponível da memória constitui um bloco único
  - Sucessivamente são alocadas alocadas/atribuídas/reservadas regiões de tamanho suficiente para conter o espaço de endereçamento dos processos que vão sendo criados/`swapped-in`
  - Posteriormente ao processo terminar, os espaços de endereçamento deixam de ser usados e são libertados

61.2.1 Gestão do espaço

Como a memória é reservada dinamicamente, o sistema de operação tem de manter um registo atualizado de:

- **Regiões livres:**
  - regiões ainda disponíveis na memória para armazenar o espaço de endereçamento dos novos processos:
    * criados
    * transferidos da á`rea de swapping`
- **Regiões Ocupadas:**
  - localiza as regiões que foram reservadas para armazenamento do espaço de endereçamento dos processos que residem em memória principal

Usando uma lista biligada (ou simplesmente ligada) para cada região. Estas listas:

- Não vão sempre indicar os espaços livres
- Podem ser alargadas
- A sua gestão em feita em blocos
  - Posso adicionar blocos à lista medida que vou libertando blocos da memória principal
  - Se o bloco a libertar estiver contíguo a um (ou dois) bloco(s) livre(s) tenho de modificar a lista (e não simplesmente introduzir um novo bloco)

Problema: Se a região de memória reservada for exatamente a suficiente para o espaço de armazenamento do processo, existe o risco de `fragmentar` o disco em regiões de memória tão pequenas que não podem ser utilizadas. Para complicar,

estas partições seriam introduzidas na lista de regiões livres tornando a lista mais complexa e aumentando o seu custo de processamento

Solução: A memória principal é dividida em múltiplos de **blocos de tamanho fixo** que constituem a unidade de trabalho para a alocação de partições

61.2.2 Exemplo

Considerando o seguinte diagrama e tabela que mostra a distribuição de 3 processos em memória, juntamente com o espaço ocupado pelo sistema operativo e, temos:

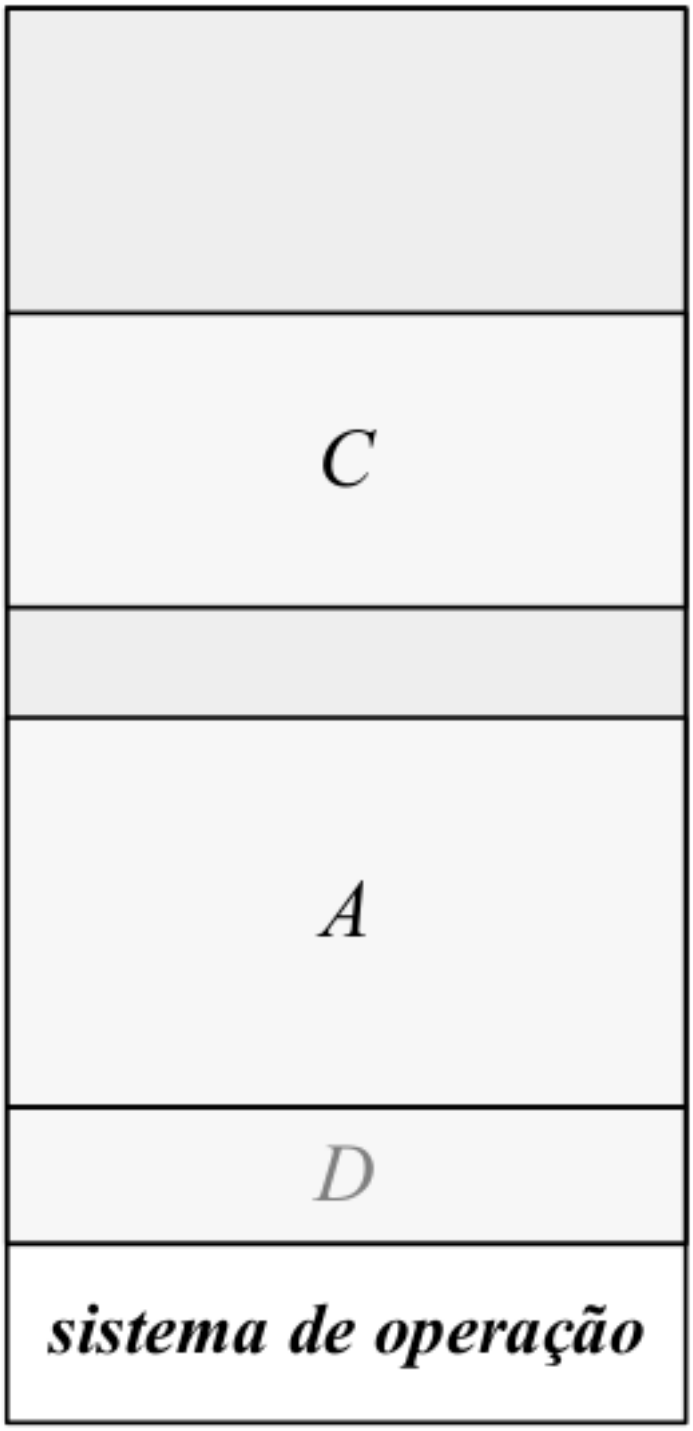


Figure 60: Diagrama da Memória Particionada

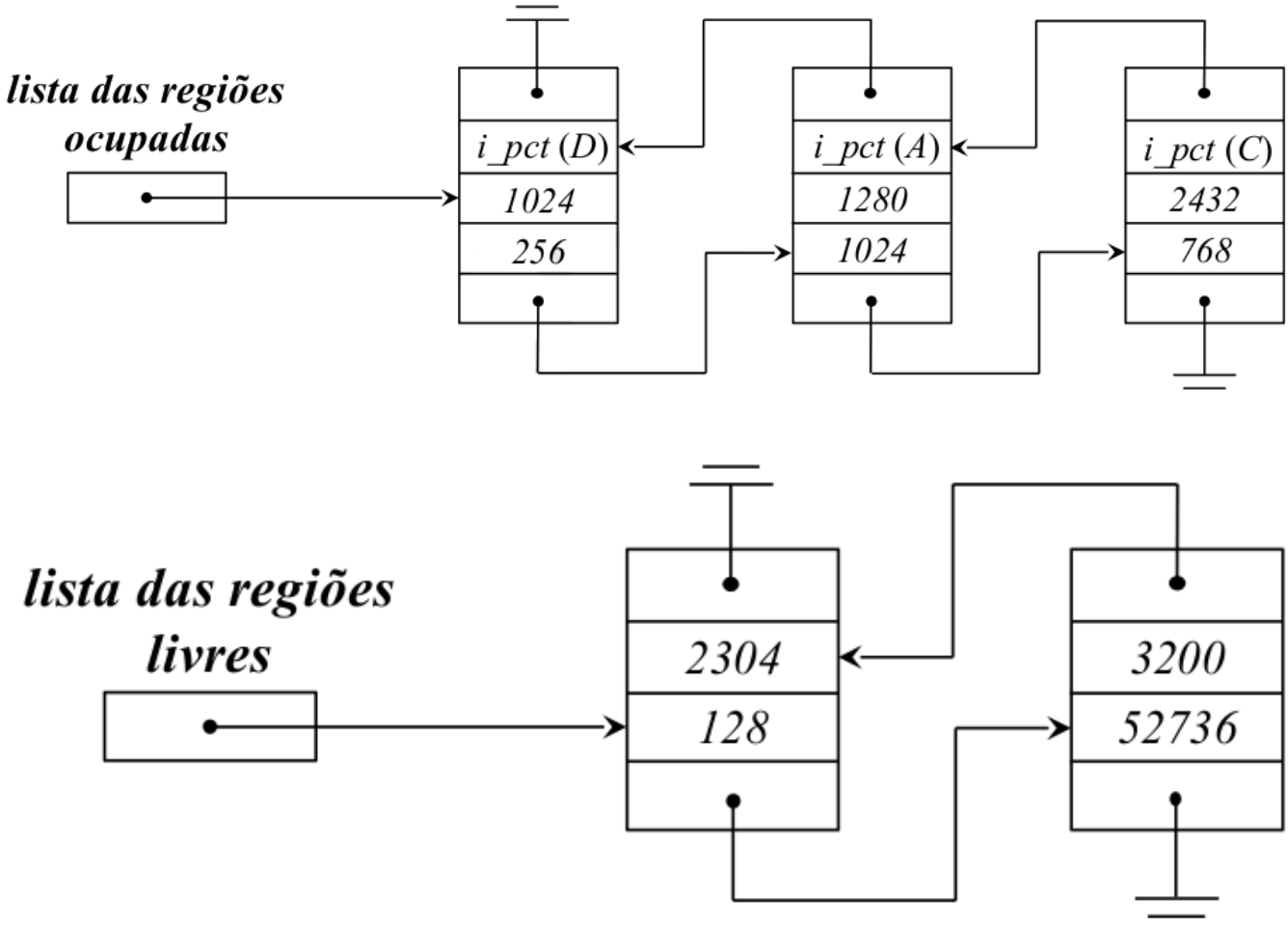
Table 11: Distribuição da ocupação da memória

	Tamanho (bytes)
Memória Principal	256 M

	Tamanho (bytes)
Sistema de Operação	4M
Unidade de reserva [^3]	4K
Processo A	4M
Processo C	3M
Processo D	1M

Tamanho mínimo de uma partição. Todos os endereços em memória são múltiplos da unidade de reserva

A obtenção da `lista das regiões ocupadas` e da `lista das regiões livres` pode ser determinada de forma trivial:



61.2.3 Políticas de Escalonamento

A **valorização do critério de justiça** é a disciplina de escalonamento mais adotada.

- É escolhido o **primeiro** processo da **fila de espera** dos processos `Suspended-Ready` cujo espaço de endereçamento pode ser colocado em memória principal

- O **principal problema** de uma arquitectura de partições variáveis é o `grau de fragmentação externa` que é produzido na memória principal
  - sucessivas **alocações** e **libertações** das partições do espaço de endereçamento dos processos
  - em casos críticos, podemos ter situações em que apesar de **haver memória livre em quantidade suficiente**, ela **não é continua** e não é possível alocar espaço em memória para um novo processo
- A solução passa por efetuar `garbage collection`
  - agrupar todas as posições/partições de memória livres num dos extremos da memória
  - obriga a mudança em memória de todos os processos realocados
  - exige a paragem de todo o processamento
  - se a memória for grande, tem um tempo de execução elevado

Para a escolha da região de memória principal a ser reservada para o armazenamento do processo, dominam as seguintes políticas: 1.`first fit` - A lista de regiões livres é pesquisada desde o princípio - A região de memória a ocupar é a primeira região com tamanho suficiente encontrada 2. `next fit` - Variante do `first fit`, com os mesmos princípios de decisão - No entanto, a pesquisa é iniciada do ponto de paragem da pesquisa anterior 3. `best fit` - A lista de regiões livres é pesquisada na sua totalidade - Escolhe-se a região mais pequena de tamanho igual ou maior ao espaço de endereçamento do processo 4. `worst fit` - A lista de regiões livres é pesquisada na sua totalidade - A região escolhida é a maior região existente

Algumas considerações:

- Uma política que seja boa/rápida a inserir elementos na fila será má/lenta a removê-la
- É difícil encontrar soluções que sejam tão rápidas a inserir como a remover da fila de partições
- Os diferentes métodos possuem diferentes desempenhos:
  - grau e tipo de fragmentação causado
  - eficiência na reserva/libertação do espaço

61.2.4 Vantagens vs Desvantagens

Vantagens:

- `Geral`: O âmbito da sua aplicação é independente do:
  - tipo de processos que vão ser executados
  - número
  - e tamanho do seu espaço de endereçamento (com algumas limitações)
- `pouco complexo`
  - não exige `hardware` especial
  - as estruturas reduzem-se a listas biligadas

Desvantagens:

- `Grande fragmentação externa da memória principal`
  - Uma fração da memória principal acaba por ser desperdiçada
  - É dividida em regiões reduzidas que não são úteis
  - O desperdício de memória pode chegar a um terço (regra dos 50%)
- `Pouco Eficiente`
  - Não é possível desenvolver algoritmos que sejam eficientes quer a:
    * **reservar/alocar espaço**
    * **libertar espaço**

62 Organização da memória real

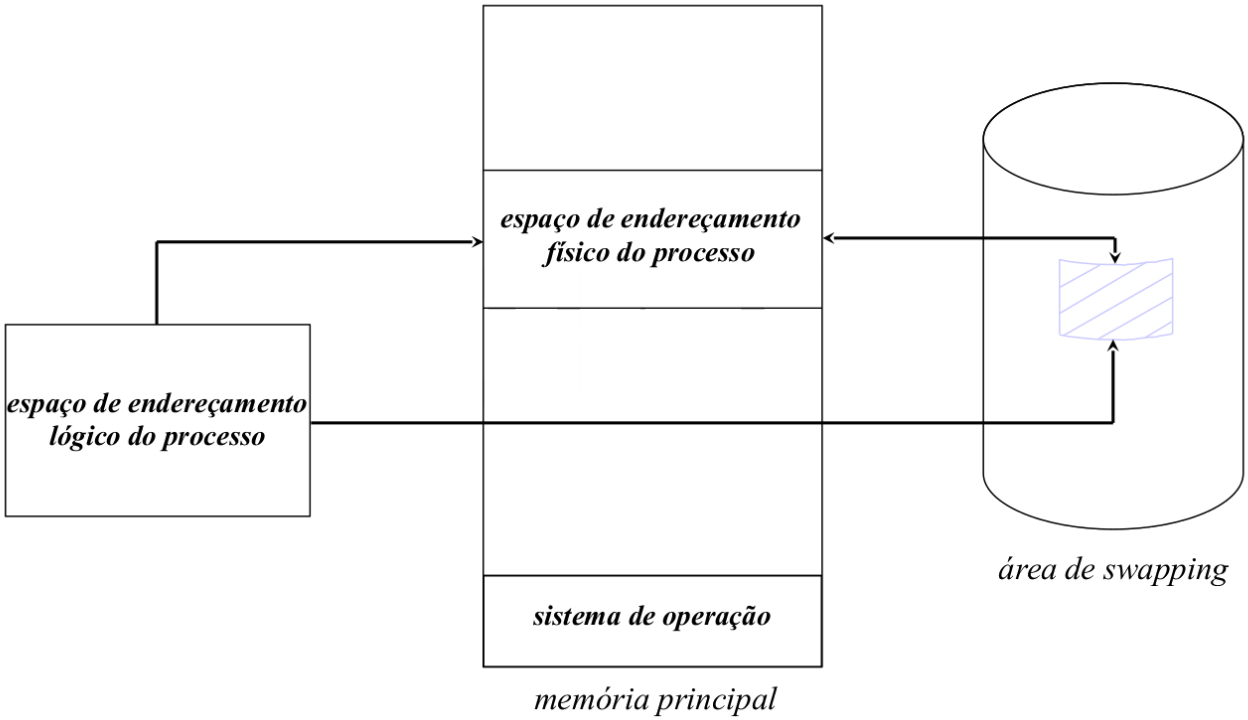


Figure 61: Espaço de endereçamento real de um processo

Existe uma correspondência biunívoca [8] entre o **espaço de endereçamento lógico** de um processo e o **espaço de endereçamento físico** de um processo. Isto implica

- O **espaço de endereçamento de um processo é limitado**
  - O espaço de endereçamento de um processo nunca pode ser superior ao tamanho de memória principal disponível
  - Os mecanismos que o tentem fazer devem ser bloqueados
- O **espaço de endereçamento físico de um processo deve ser contíguo**
  - Não e uma condição estritamente necessária
  - Simplifica e torna mais eficiente se o espaço de endereçamento de um processo for obrigado a ser contíguo
- A **existência de uma área de swapping**
  - Serve como extensão da memória principal
  - Armazena espaços de endereçamentos de processos que não podem residir em memória principal por falta de espaço

[8] De um para um

62.1 Tradução de um endereço lógico num endereço físico

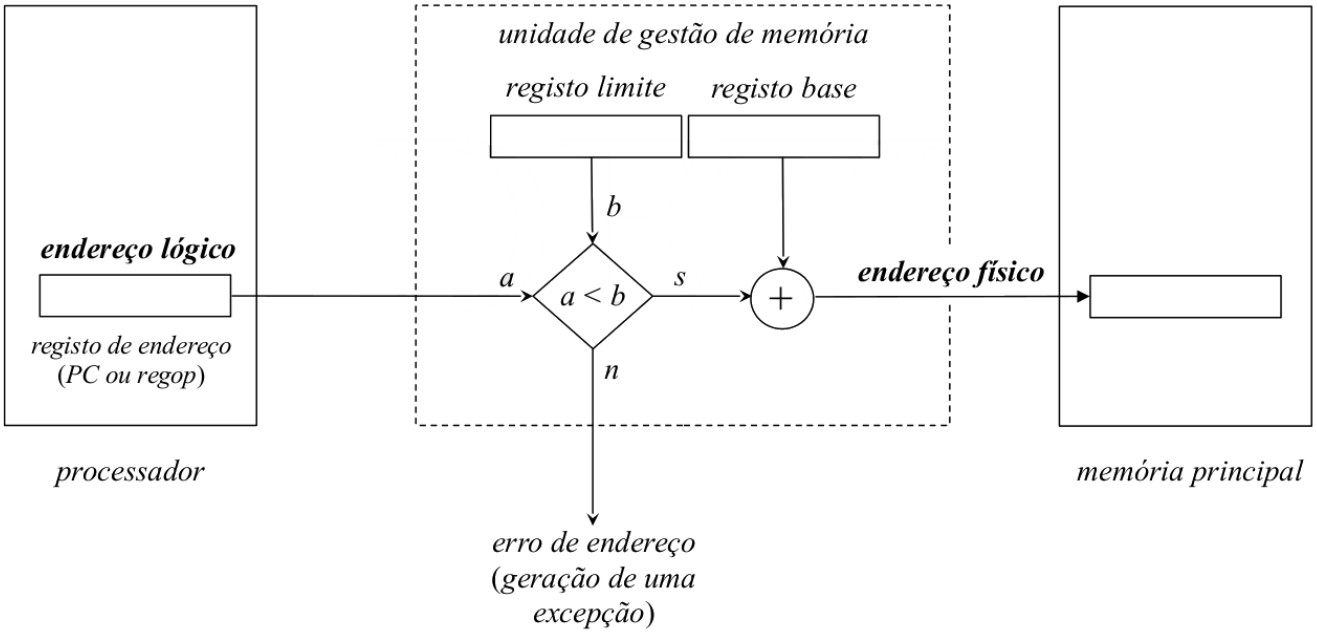


Figure 62: Tradução de um endereço lógico num endereço físico

- `registo base`: endereço do início da região de memória principal onde está alojado o espaço de endereçamento físico do processo
- `registo limite`: tamanho em *bytes* do espaço de endereçamento

Na comutação de processos:

- `dispatch` carrega o registo base e o registo limite da tabela de controlo de processos do processo que vai ser calendarizado para discussão

Sempre que há uma referência à memória, o **endereço lógico** comparado com o **registo limite** e:

- Se for **maior** $\implies$ **referência inválida**
  - Acesso à memória nulo é executado (aka `dummy cycle`)
  - É gerada uma exceção por erro de endereço
- Se for **menor** $\implies$ **referência válida**
  - a referência aponta para dentro do espaço de endereçamento do processo
  - o conteúdo do registo base é adicionado ao endereço lógico para produzir o endereço físico

62.2 Memória real e o ciclo de vida de um processo

Depois de carregado o Sistema Operativo, o que resta da memória principal é usado para conter o espaço de endereçamento dos diferentes processos

62.2.1 Criação de um processo

- O processo está no estado `CREADTED`
- São inicializadas as estruturas de dados destinadas a geri-lo
  - A imagem binária do seu espaço de endereçamento é construída
  - O valor do campo `registo limite` da entrada da tabela de controlo de processos é determinado
- Se houver espaço em memória
  - o espaço de endereçamento do processo é carregado
  - o campo `registo base` é atualizado com o endereço inicial da região reservada
  - o processo transita para o estado `Ready-to-Run` e é colocado na respetiva fila de espera
- Se não houver espaço em memória
  - O processo transita para o estado `Suspended-Ready`
  - O processo é colocado na respetiva fila de espera
  - O seu espaço de endereçamento é armazenado temporariamente na `área de swap`

62.2.2 Ciclo de Vida do processo

- Ao longo da sua execução, o espaço de endereçamento do processo pode ser deslocado temporariamente para a `área de wapping`.
  - `Ready-to-Run` $->$ `Suspended-Ready`
  - `Blocked` $->$ `Suspended-Blocked`
- Sempre que há espaço em memória
  - Um dos processos presentes na fila de espera dos processos `Suspended-Ready` é selecionado
  - O seu espaço de endereçamento é carregado
  - O campo `registo base` da entrada da **tabela de controlo de processos** é atualizada com o endereço inicial da região reservada
  - O processo é colocado na fila de espera `Ready-to-Run`, transitando para esse estado
- Caso a lista de espera `Suspended-Ready` estiver vazia e existirem processos na fila de espera dos processos `Suspended-block`, um desses processos pode ser selecionado
  - À semelhança da transição `Suspended-Ready` para `Ready`, na transição `Suspended-Blocked` para `Blocked` as mesmas inicializações são feitas

62.2.3 Fim de Vida do processo

- O processo transita para o estado `Terminated`
- O seu espaço de endereçamento é transmitido para a `área de swapping` (se não estiver lá), para aguardar o fim das operações

63 Organização da memória virtual

- Num sistema com memória virtual, o `espaço de endereçamento lógico` e o `espaço de endereçamento físico` de um processo estão **totalmente dissociados**

Como consequência:

- O espaço de endereçamento de um processo não está limitado à memória física
  - O espaço de endereçamento virtual é "ilimitado"
  - Podem criar-se mecanismos que permitam a um processo ocupar mais do que a memória principal disponível
- Não continuidade **do** espaço de endereçamento físico
  - O espaço de endereçamento dos processos podem estar dispersos por toda a memória
    - quer os blocos sejam de tamanho fixo ou variável
  - Garante-se uma ocupação mais eficiente do espaço disponível
- Área de swapping
  - Serve como extensão da memória principal
  - Guarda uma imagem atualizada dos espaços de endereçamento dos processos que coexistem de forma concorrente
  - guarda também as variáveis dinamicamente alocadas:
    - stack
    - zona de definição estática

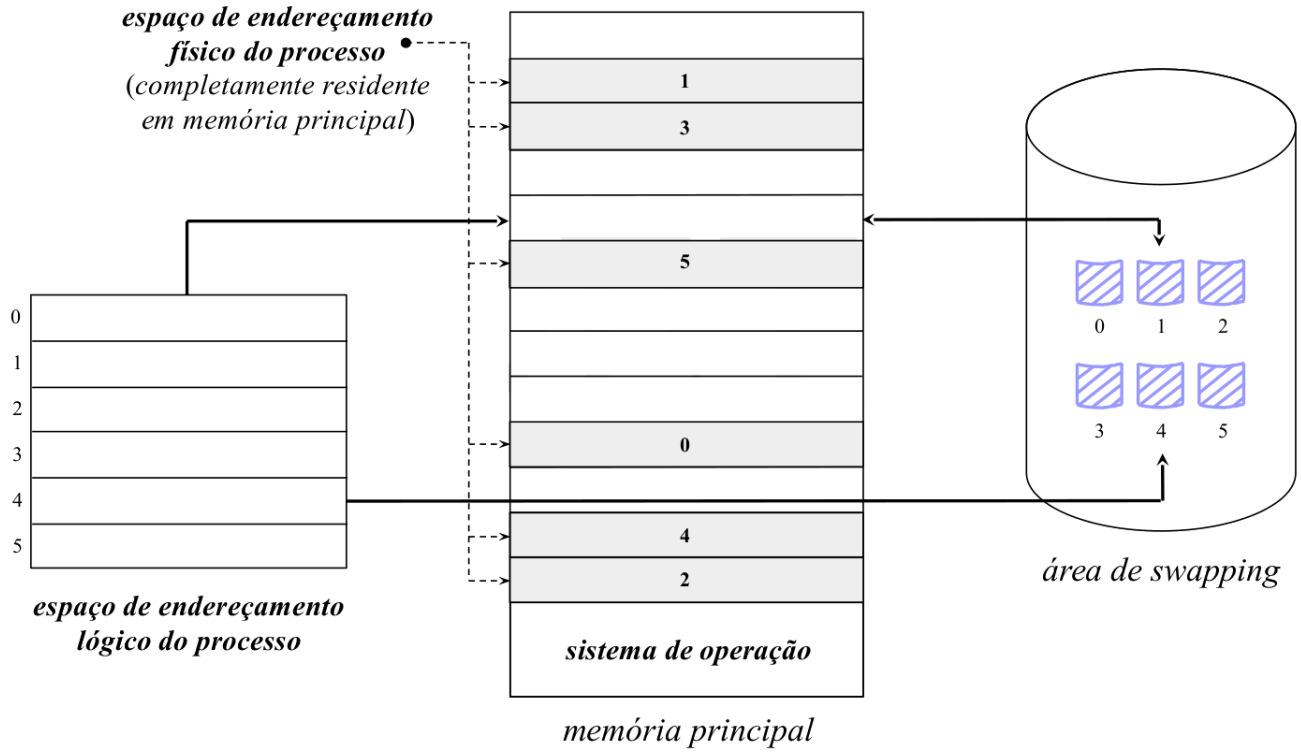


Figure 63: Espaço de endereçamento completamente em memória virtual

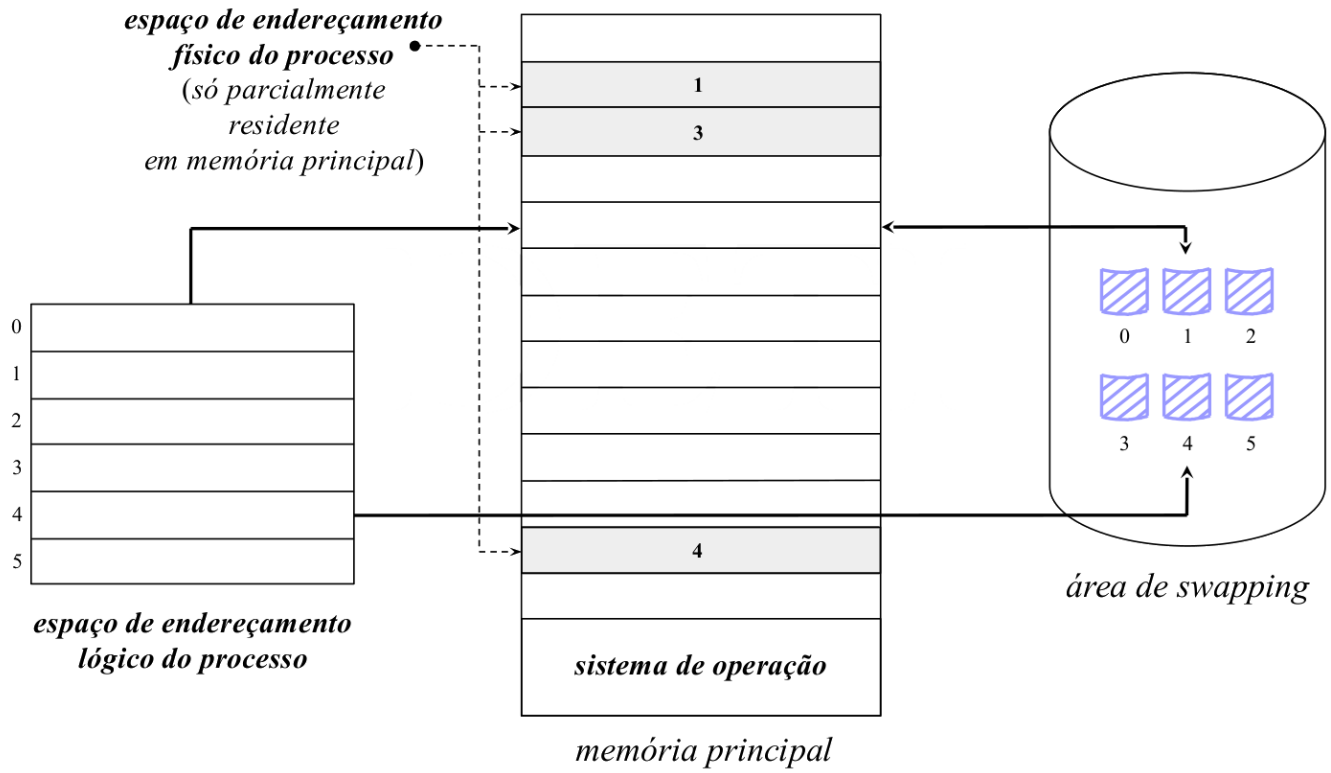


Figure 64: Espaço de endereçamento apenas parcialmente em memória virtual

63.1 Tradução de um endereço lógico num endereço físico

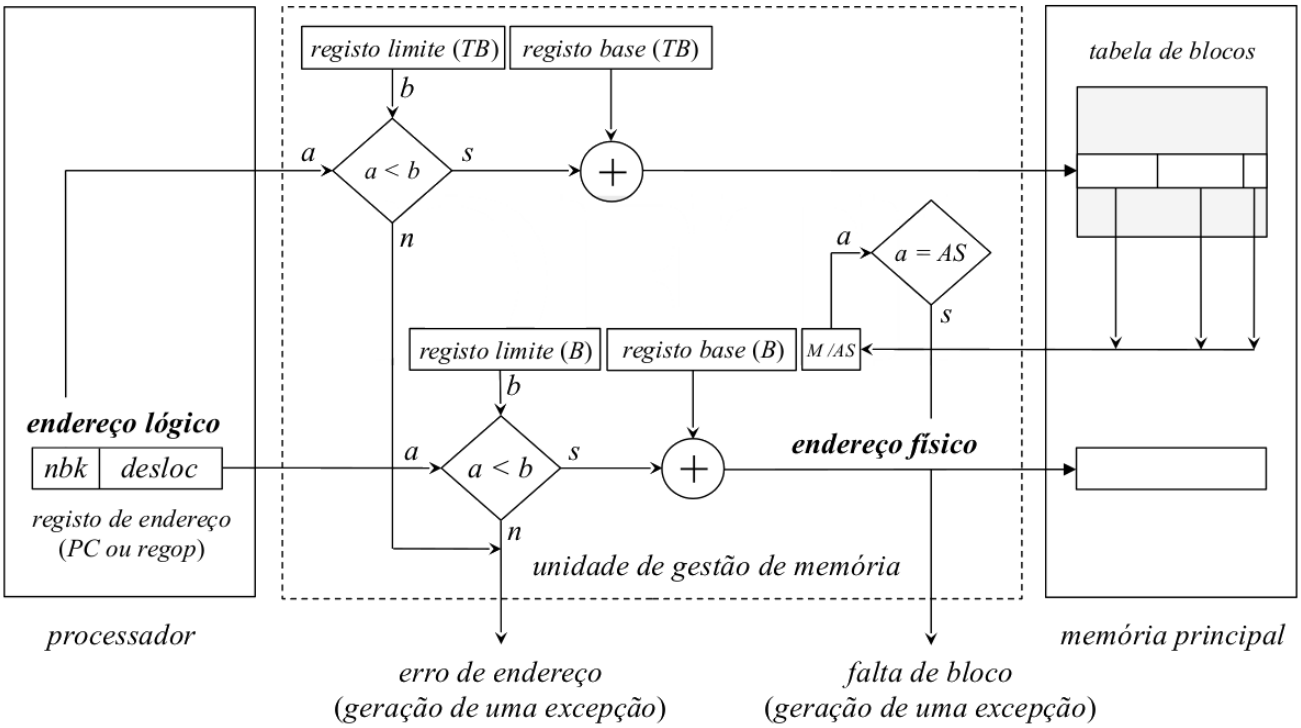


Figure 65: Diagrama de blocos da decomposição de um endereço lógico num endereço físico

O endereço lógico é formado por dois campos:

- **nbk:** identificador de um bloco específico
- **desloc:** localiza uma posição de memória concreta dento do bloco, através do cálculo da distância ao seu erro

A **unidade de gestão de memória** contém dois pares de `registos base` e `limite`

1. Associado com a **tabela de blocos do processo**:
    - descreve a localização dos vários blocos me que o espaço de endereçamento do processo está dividido
2. Descrição de um bloco particular

Quando ocorre uma **comutação de processos** a operação de `dispatch` carrega o `registo base` e o `registo limite` da tabela de blocos com os valores na tabela de controlo de processos (associada com o processo que vai ser calendarizado para execução)

```
- O valor do ‘registo base da tabela de blocos‘ representa o **endereço do início da
    região de memória principal** onde está alojada a tabela de blocos do processo
- O valor do ‘registo limite‘ está relacionado com o número de entradas na tabela
```

63.1.1 Acesso à memória

- Decomposto em 3 fases

1. Campo `ndk` do `endereço lógico` é comparado com o valor do **registo limite da tabela de blocos**
   - $ nbk < registo limite (TB) $
     - nbk é adicionado ao conteúdo do `regiso base da tabela de blocos` para produzir o endereço da entrada da tabela de blocos
   - $ ndk >= registo limite (TB) $
     - A referência é inválida e não é efetuado nenhuma acesso à memória
     - A instrução é interrompida por uma instrução de acesso à memória nulo *(dummy cycle)*
     - Gera-se uma exceção por erro de endereço: `segmentation fault`
2. É efetuada a avaliação do registo `M`/`AS`(*Memória/Área de Swap*)
   - se `M`: os **campos** da entrada da **tabela de blocos** referenciada são transferidos para os **registos respetivos** da unidade de gestão de memória.
   - se `AS`:
     - o bloco não está **atualmente em memória**
       * tem de ser transferido para a instrução puder continuar
     - Instrução finalizada com um acesso à memória nulo
     - Gera-se uma exceção por falta de bloco
       * será responsável por iniciar a transferência dos blocos da swap para a memória principal
     - Processo transita para o estado `blocked`
3. o campo `desloc` (deslocamento) do `endereço lógico` é comparado com o valor do `registo limite (B)` (do bloco)
   - $ desloc < registo limite (B) $ **referência válida
     - A referência é efetuada para dentro do espaço de endereçamento do bloco
     - `desloc` é adicionado ao conteúdo do registo base do bloco para produzir o endereço físico
   - $ desloc >= registo limite (B) $ **referência inválida**
     - é efetuado um acesso à memória nulo *(dummy cycle)*
     - gera-se uma exceção por erro de endereço `segmentation fault`

A gestão do espaço de memória principal usando uma organização de memória virtual possui a vantagem de permitir maior versatilidade, mas também possui o custo associado de que **cada pedido de acesso à memória (r/w) requer dois acessos para puder ser executado**

- **1ª Acesso:**
  - Referencia a entrada da tabela de blocos do processo, usando o campo `nbk` do endereço lógico como endereço do bloco em memória que contém o endereço da posição de memória que se quer ler/escrever
- **2º Acesso:** É feita referência à posição de memória específica (que se deseja efetivamente aceder)
  - O cálculo do seu endereço é efetuado adicionando o campo `desloc` do endereço lógico ao endereço que corresponde ao início do bloco em memória
- A organização em memória virtual causa um **fracionamento do espaço de endereçamento lógico** do processo.
- Os blocos/frações são tratadas dinamicamente como sub espaços de endereçamento **autónomos** numa organização de memória real
  - A memória real pode estar organizada em partições físicas ou em partições variáveis
- A diferença entre uma organização de memória virtual vs uma organização de memória real é que passas a existir a possibilidade de ocorrer o acesso a um bloco que atualmente não reside em memória principal

  - Nestas condições o sistema é capaz de anular a instrução de acesso atual
  - Meter em marcha a sequência de instruções que permite carregar esse bloco para memória principal
  - repetir a instrução assim que o bloco for carregado

Evidentemente, a necessidade do duplo acesso à memória pode ser minimizada tirando partido do **Princípio da localidade da referência:**

- Os acessos tenderão a estar concentrados num conjunto bem definido de blocos durante grandes intervalos de tempo de execução do processo
- A MMU faz caching do conteúdo das entradas da tabela de blocos que forma ultimamente referenciadas, usando uma memória associativa *(*`translation lookaside buffer (TLB)`*)*:
  - Cada acesso passa assim a ser um:
    * `hit`:
      · a entrada está armazenada no processador
      · o acesso é interno
      · não é referenciada memória na 1ª fase
    * `miss`:
      · a entrada não está armazenada na TLB
      · é preciso um acesso externo à memória principal na 1ª fase
- O tempo médio de acesso a uma instrução/operado tende para:
  - Acesso ao TLB + acesso à memória principal

63.2 Ciclo de vida de um processo

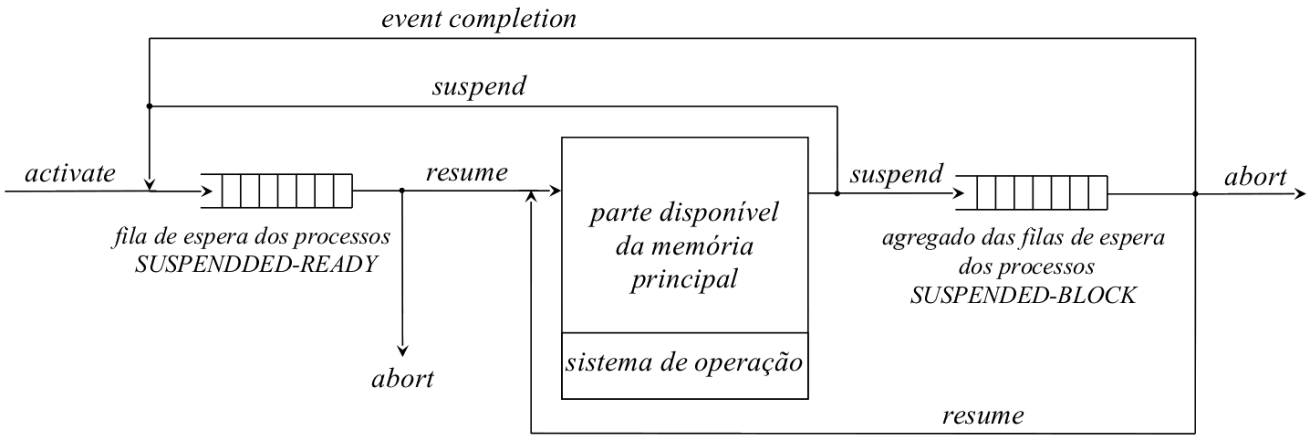


Figure 66: Diagrama de eventos e estrutura de uma organização em memória virtual

63.2.1 Criação de um processo

- estado: `CREATED`
- são inicializadas a estruturas de dados destinadas a geri-lo

  - É construída a imagem binária do seu espaço de endereçamento
  - É transferida para a área de swapping a sua parte variável
  - A tabela de blocos associada é organizada
    * Se existir espaço livre em memória é carregado em memória principal:
      · o 1º bloco de código do processo
      · o bloco da stack
      · as entradas correspondentes da tabela de blocos são atualizadas
      · estado: `READY-TO-RUN`
      · colocado na fila de espera de processos prontos a serem executados
    * Se não existir:
      · estado: `SUSPENDED-READY`
      · colocado na fila de espera de processos suspensos mas prontos a serem transferidos para memória principal e executados
- Os escalonadores tentam sempre garantir que o registo PC, a stack estão em memória e o primeiro bloco de instruções de estão em memória - São as condições principais para que um programa possa ser executado

63.2.2 Ao longo da execução

- Se ocorrer um acesso a um bloco não residente em memória principal:
  - estado passa a `BLOCKED`
    * permanece `BLOCKED` enquanto ocorre a transferência do bloco para memória principal
    * quando a transferência terminar, passa a `READY-TO-RUN`
    * é colocado na fila de espera de processos `READY-TO-RUN`
- Os blocos residentes na memória principal pertencentes a um mesmo processo podem ser `swappeed-in`
  - os seus blocos são "movidos" para a área de swap
  - o estado passa de:
    * `READY-TO-RUN` -> `SUSPENDED-READY`
    * `BLOCKED` -> `SUSPENDED-BLOCK`
- Sempre que há espaço memória:
  - um dos processo presentes na fila de espera `SUSPENDED-READY` é selecionado
  - A tabela de blocos e um grupo de blocos do seu espaço de endereçamento são carregados
  - As entradas correspondentes na tabela são atualizadas com os endereços iniciais das regiões reservadas
  - o processo é colocado na fila de espera de processos `READY-TO-RUN`
- Se a lista `SUSPENDED-READY` estiver vazia e **houver processos na fila de espera** dos processos `SUSPENDED-BLOCK`
  - Pode ser selecionado um destes processos
  - Passa para o estado `BLOCKED` e é inserido na respetiva lista de espera

63.2.3 Término de um processo

- estado: `TERMINATED`
- A imagem do seu espaço de endereçamento residente na área de swapping (ou pelo menos a sua parte variável) é atualizada
  - libertação de todos os blocos existentes em memória principal
  - Aguarda pelo fim das operações

63.3 Exceção por falta de bloco

- A rotina de serviço a esta interrupção/exceção é responsável por em marcha as ações que permitam:
  - a transferência desse bloco da área de swapping para a memória principal
  - a repetição da instrução que produziu a referência
- Estas operações são realizadas de forma totalmente transparente ao utilizador

Se um processo estivesse continuamente a gerar exceções por falta de bloco:

- o ritmo de processamento seria muito lento
- o `throughput` do sistema computacional seria mais baixo

Isto não acontece (pelo menos com tanta frequência devido à hierarquia da memória e principalmente há `Translation Lookaside Bffer` da unidade de memória, usando o princípio da localidade de referência.

Apesar da fração do espaço de endereçamento variar com o tempo, o custo em carregar vários blocos pontualmente para memória, relativamente a outras soluções é reduzido, podendo progredir a execução do processo praticamente sem ocorrerem faltas de bloco.

Situações de falta de página degridem muito a qualidade do sistema

63.3.1 Sequência de instruções

1. Salvaguardar o contexto do processo na entrada correspondente na tabela de controlo de processos
   - estado: `BLOCKED`
   - atualizar o seu $PC [9] para o endereço que produziu a falta de bloco
2. Determinar se existe espaço em memória para carregar o bloco em falta
   - Caso **exista**: selecionar uma região livre
   - Caso **não exista**: selecionar uma região cujo bloco vai ser substituído
     - se tiver sido modificado -> transferi-lo para a á`rea de swapping`
     - atualizar a entrada da tabela de blocos do processo a que o bloco pertence
       * indicar que o bloco já não se encontra em memória (registo `M/AS`)
3. Transferir o bloco em falta da área de swapping para a região selecionada
4. Invocar o escalonador para calendarizar a execução de um dos processos da fila de espera `READY-TO-RUN`
5. Quando a transferência estiver concluída
   - Atualizar a entrada da tabela de blocos do processo
     - indicar que o processo está residente em memória
     - indicar a sua localização
   - estado: `READY-TO-RUN`
   - colocar o processo na fila de espera `READY-TO-RUN`

[9] ficheiro em código fonte de compilação separada

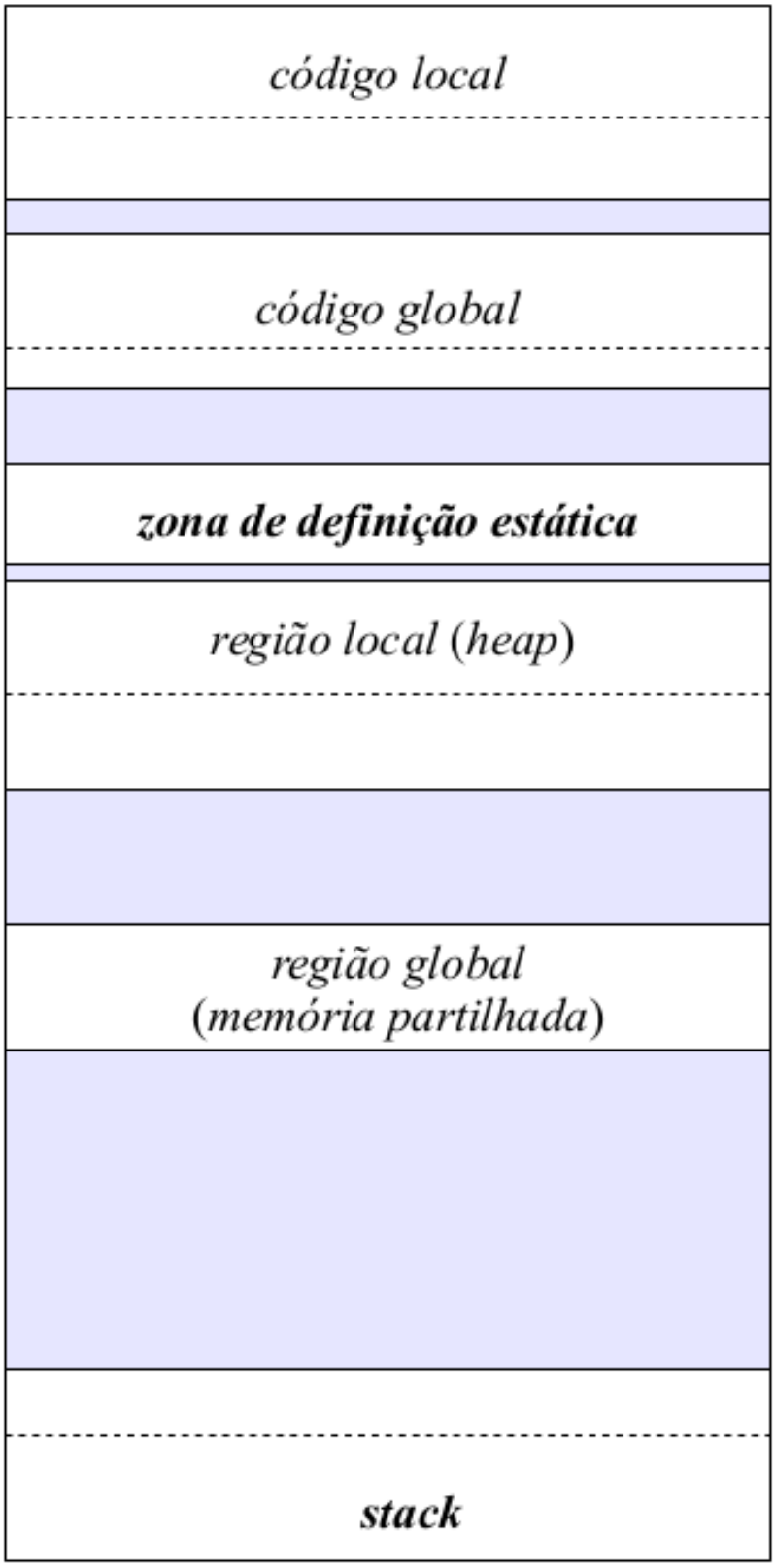


Figure 67: Estrutura de uma organização de memória em arquitectura paginada

Os blocos do espaço de endereçamento do processo passam a ser designados de **páginas**.

- São todos iguais
- Tamanho múltiplo de uma potência de 2
  - Tipicamente 4 ou 8 KB
-

O espaço de endereçamento lógico é endereçado usando:

- **bits mais significativos:** número da página
- **bits menos significativos:** deslocamento

A memória principal é dividida em blocos da **mesma dimensão** que as **páginas**. A estes blocos chamamos **frames**

O `linker` organiza o espaço de endereçamento lógico do processo atribuindo o início de uma nova página a cada uma das regiões funcionalmente distintas:

- código local
- código global
- zona de definição estática
- região local *(heap)*
- região global *(shared memory)*
- `stack` (neste caso, atribuí o fim da página)

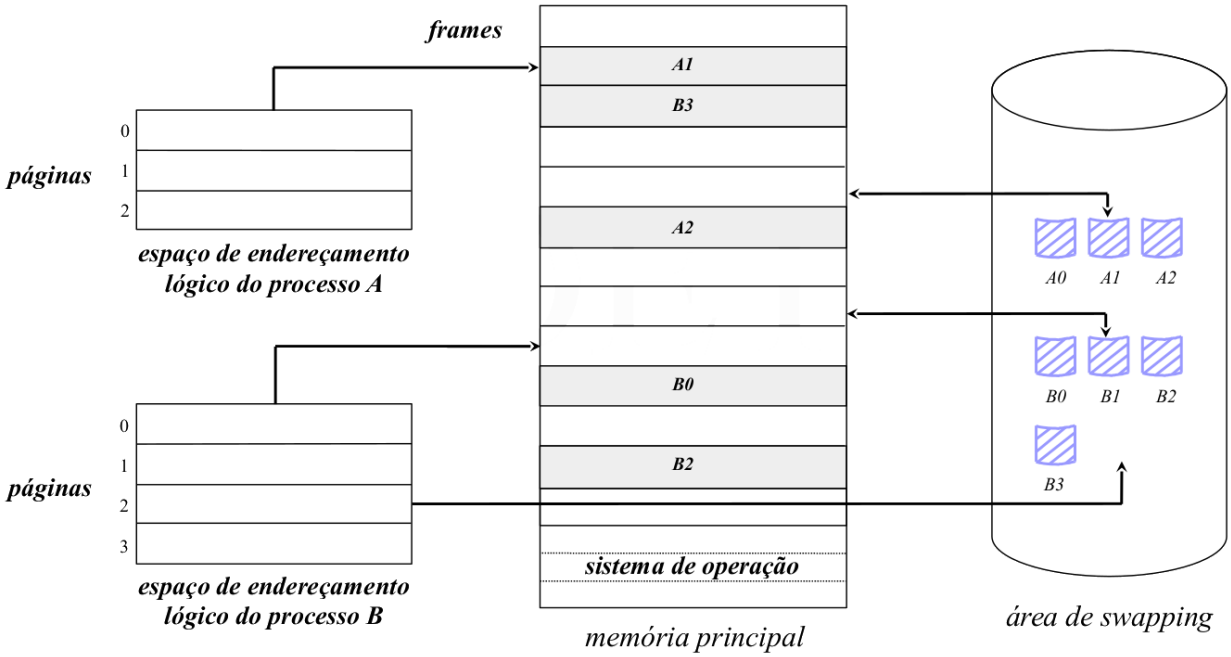


Figure 68: Exemplo da ocupação da memória principal e swap num arquitectura paginada

63.4 Acesso à memória

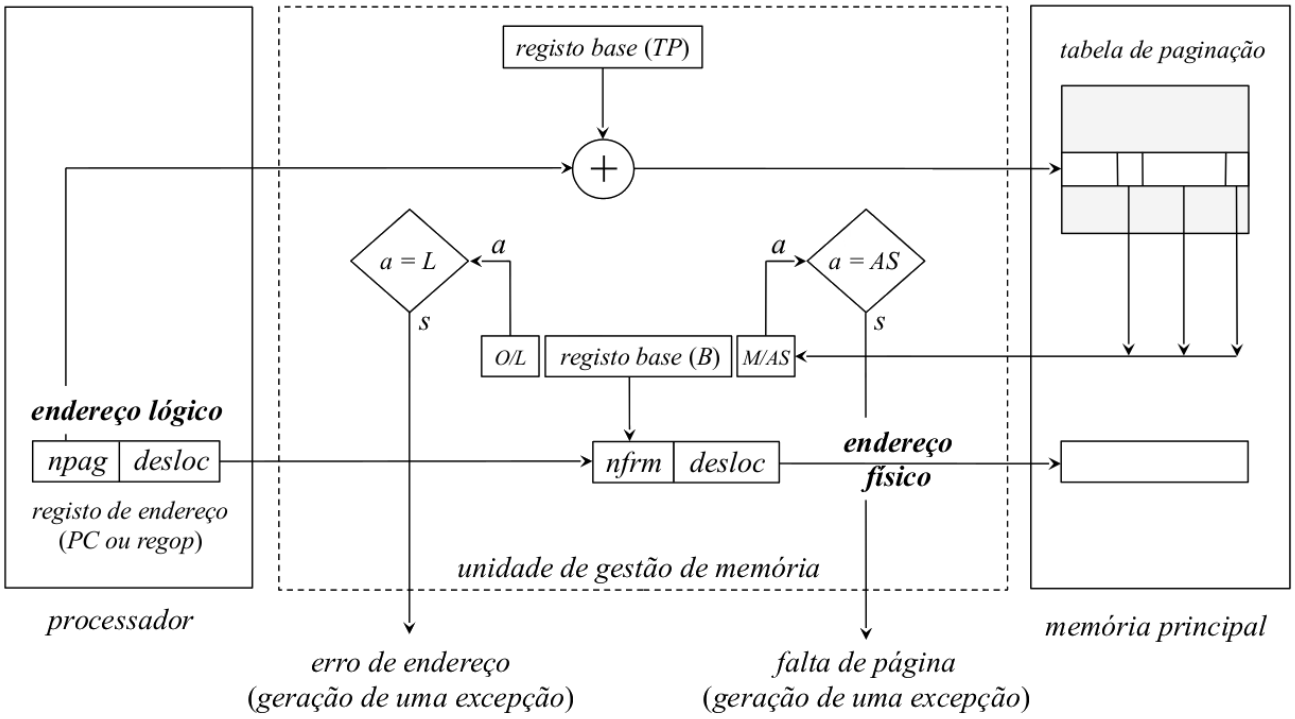


Figure 69: Diagrama de blocos para efetuar o acesso

- Deixa de ser necessário o registo limite na tabela de paginação do processo
  - A Tabela de paginação do processo só precisa do registo base
  - Não é necessário o endereço limite de uma página articular
  - O endereço físico é formado pela concatenação entre os campos:
    * `nfrm` (identifica o frame da memória principal onde a página está localizada)
    * `desloc` (identifica o deslocamento dentro da página/frame
    * endereço físico = nfrm | desloc
  - O endereço lógico é a concatenação de (32 bits):
    * bits MSB: número da página (20 bits)
    * bits LSB: offset dentro da página (12 bits)
  - É possível estruturar o espaço de endereçamento lógico do processo de modo a mapear a totalidade (ou pelo menos uma fração) do espaço de endereçamento do processador
  - Estas frações de espaço podem ser maiores ou iguais ao tamanho da memória principal existente
- Diretamente resultam duas consequências:
  - É possível reservar espaço na zona de definição dinâmica
  - A stack pode atingir a máxima amplitude possível
- As páginas correspondentes à zona de **definição dinâmica** (heap) e à **stack** só são criadas **quando necessário**
  - Permite poupar área de swapping
  - Origina um erro de endereço sempre que se tenta aceder a uma página que ainda não existe
    * Erro de segmento != Falta de página

· Acesso a um endereço invalido != Acesso a um endereço válido mas que não existe em memória de swap

Conteúdo da entrada da tabela de paginação

O/L	*M/AS*	*Ref*	*Mod*	*Perm*	*N. do frame em memória*

- **O/L** (Ocupada/Livre): bit que sinaliza a ocupação ou não desta entrada (a não ocupação significa que ainda não foi reservado espaço na área de swapping para esta página)
- **M/AS** (Memória/Área de Swap): bit que sinaliza se a página está ou não `residente em memória principal`
- **Ref** (Referenciada): bit que sinaliza se a página foi ou não `referenciada para leitura e/ou escrita`
- **Mod** (Modificada): bit que sinaliza se a página foi ou não `referenciada para escrita`
- **Perm** (Permissões): indicação do tipo de acesso permitido
  - `ronly` (read-only)
  - `read`/`write`
  - `rwx` (ler/escrever operandos, executar instruções)
- **Número do frame em memória (nfrm):** localização da página, se residente em memória principal
- **Número do bloco na área de swapping:** localização da página na área de swapping, se lhe foi atribuído espaço

63.5 Vantagens e Desvantagens

63.5.1 Vantagens

- **geral:** o âmbito da aplicação é **independente do tipo de processos** que vão ser executados (número e tamanho do espaço de endereçamento)
- **grande aproveitamento da memória principal:**
  - não conduz à fragmentação externa
  - desfragmentação interna desprezável
- **não exige requisitos especiais de hardware:** a unidade de gestão de memória (MMU) existente nos processadores atuais já vem preparada para a sua implementação
  - Gera as exceções, os processadores é que têm de tratar delas
- As páginas só vão sendo atribuídas ao processos à medida das necessidades

63.5.2 Desvantagens

- **acesso à memória mais longo:**
  - cada acesso à memória transforma-se num duplo acesso devido à consulta prévia da tabela de paginação
  - pode ser minimizado usando a TLB, `translation lookaside buffer` para armazenar as entradas da tabela de paginação recentemente mais referenciadas
- **operacionalidade muito exigente:**
  - a sua implementação exige que o SO possua um conjunto de operações de apoio complexas
  - essas operações têm de ser cuidadosamente estabelecidas para que não existam perdas grandes de eficiência
  - As entradas das tabelas de paginação são mais completas

64 Arquitectura Segmentada

- Divide o espaço de endereçamento lógico do processo em segmentos
  - **Divisão cega**.
  - Não leva em consideração qualquer informação sobre a estrutura do programa
  - Apenas efetua a divisão tendo em contas as regiões do código funcionalmente distintas (distinguidas atrás)
  - Não é possível trabalhar com grupos de segmentos
  - Na prática cada segmento é tratado de forma independente

Tem como consequência:

- A estrutura modular que está na base do desenvolvimento de software complexo **não é tida em conta**
  - **Não é possível usar o princípio da localidade da referência** para minimizar o número de páginas que tenham de estar residentes em memória principal em cada etapa de execução do processo
- A **gestão do espaço disponível** entre a zona de **definição dinâmica** e **stack** torna-se difícil e pouco eficiente
  - É agravada no caso de surgirem em run-time múltiplas regiões de dados partilhados de tamanho variável ou estruturas de dados de crescimento contínuo

Uma solução consiste em desdobrar o espaço de endereçamento lógico do processo. Passamos de um espaço de endereçamento linear único (como na arquitectura paginada) para uma multiplicidade de espaços de endereçamento lineares autónomos definidos na fase de linkagem

- Cada módulo [10] da aplicação irá originar dois espaços de endereçamento autónomos:
  1. código
  2. zona de definição estática:
     - variáveis globais á aplicação (definidas localmente)
     - variáveis localmente globais (internas ao módulo)
- Cada um destes espaços de endereçamento autónomo designa-se por `segmento`
  - possui uma organização em memória virtual
- Os blocos/segmentos podem ser de comprimento variável

[10] ficheiro em código fonte de compilação separada

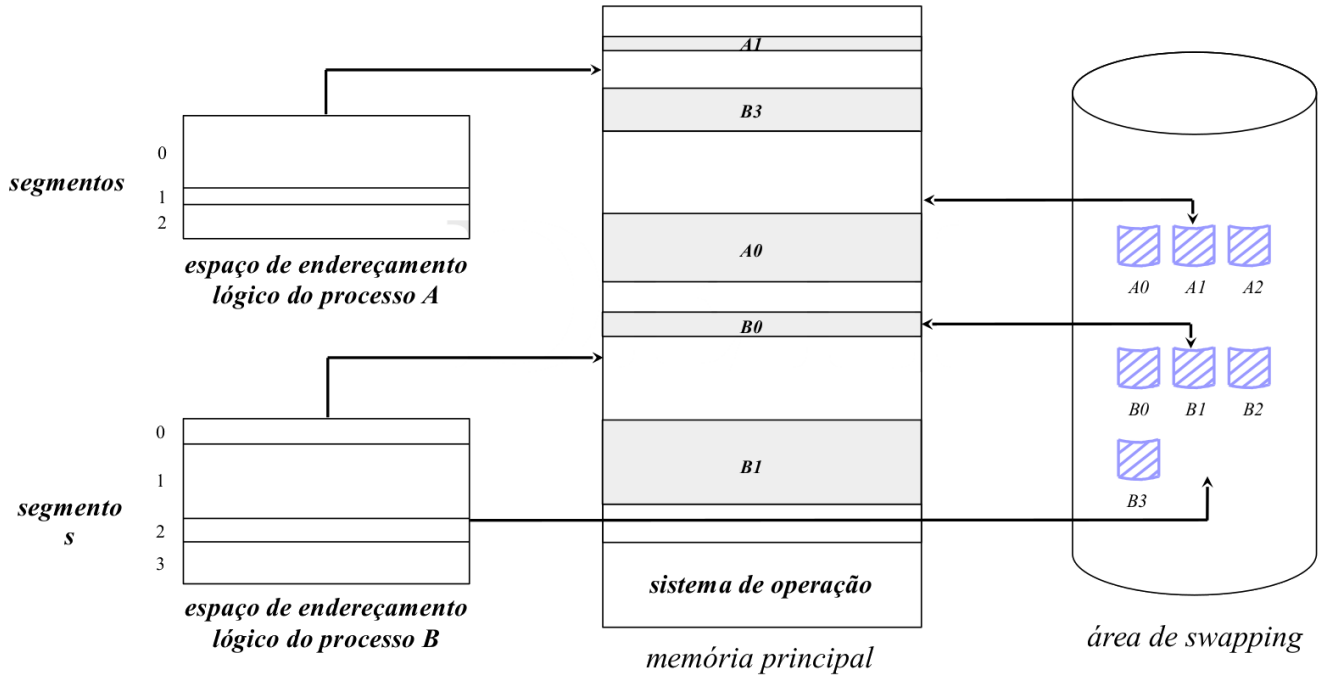


Figure 70: Exemplo de memória segmentada

64.1 Tipos de Segmentos:

- **região de código:** um segmento por cada módulo que contenha código (global ou local)
- **região de definição estática:** um segmento por cada módulo que contenha a definição de variáveis globais à aplicação ou módulo
- **zona de definição dinâmica local (heap):** um segmento
- **zona de definição dinâmica global:** um segmento por cada região de memória de partilha de dados
- **stack:** um segmento

64.2 Tradução de um endereço lógico num endereço físico

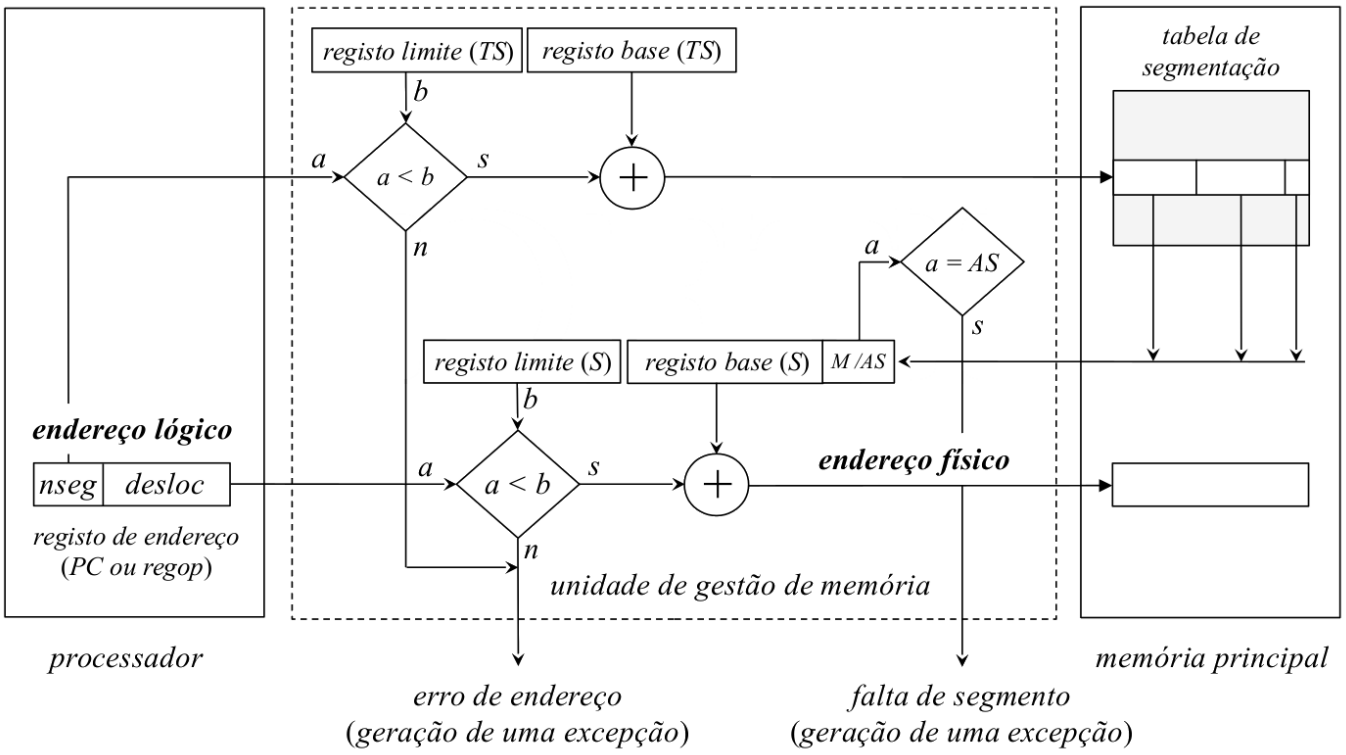


Figure 71: Diagrama de blocos da operação de tradução de um endereço lógico num endereço físico

64.3 Conclusão

A arquitectura segmentada, na sua versão pura, possui pouco interesse prático. Ao tratar a memória principal como um espaço contínuo, exige que sejam aplicadas técnicas de reserva de espaço para carregamento de um segmento de memória.

Estas técnicas assemelham-se ao que acontece numa estrutura de memória real com partições variáveis

Como consequência, existe uma grande **desfragmentação externa da memória principal**, resultando no **desperdício de espaço**

Coloca-se outro problema referente a segmentos de dados de crescimento contínuo:

- pode ser necessário efetuar um acréscimo de espaço ao tamanho do segmento mas este poderá não ser realizado na sua localização presente
  - obriga à transferência total para outra região de endereçamento de memória
  - no caso limite não existe memória disponível para essa expansão
    * o processo é bloqueado/suspenso
    * o seu segmento ou a totalidade do espaço de endereçamento são movidos para a área de swapping

65 Arquitectura Segmentada/Paginada

- Arquitectura mista que combina as características desejáveis das duas arquitecturas anteriores:

- O **espaço de endereçamento lógico** é dividido em segmentos, com a atribuição na fa se de linkagem de múltiplos espaços de endereçamento autónomos
- Cada um dos **segmentos** (espaços de endereçamento lineares) é divido em **páginas**
  * Origina um mecanismo de carregamento de blocos em memória principal com todas as características da arquitectura paginada

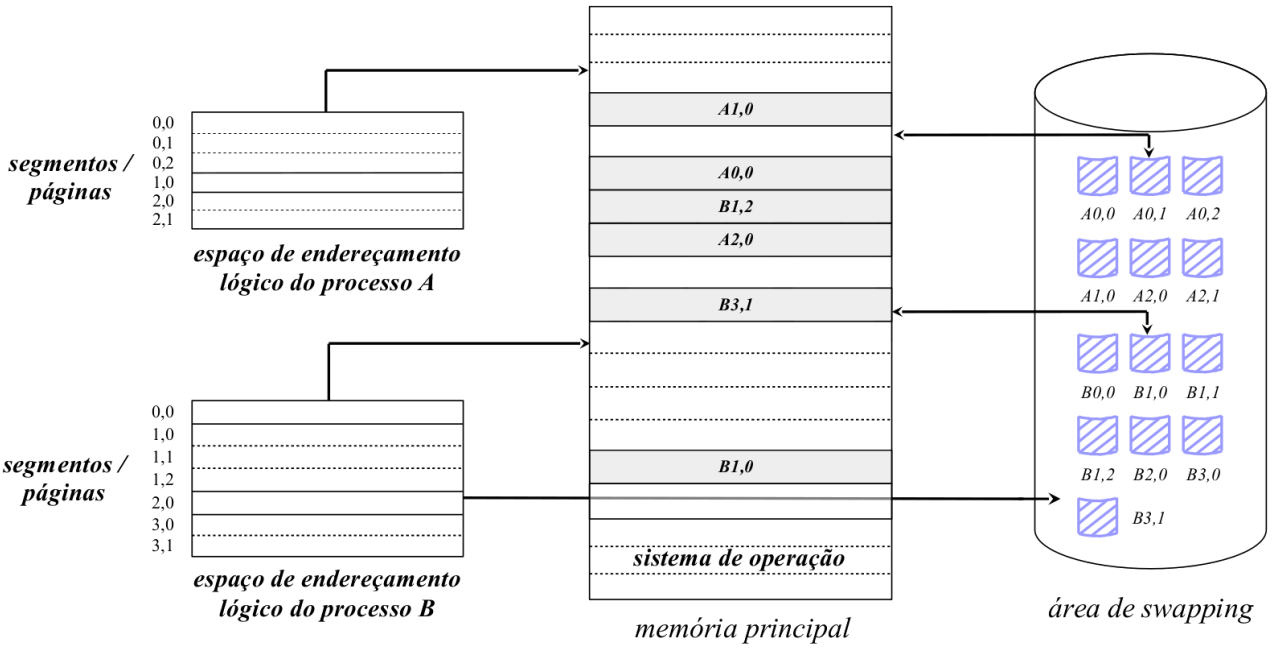


Figure 72: Estrutura de uma arquitectura segmento-paginada

65.1 Tradução de um endereço lógico num endereço físico numa arquitectura segmento-paginada

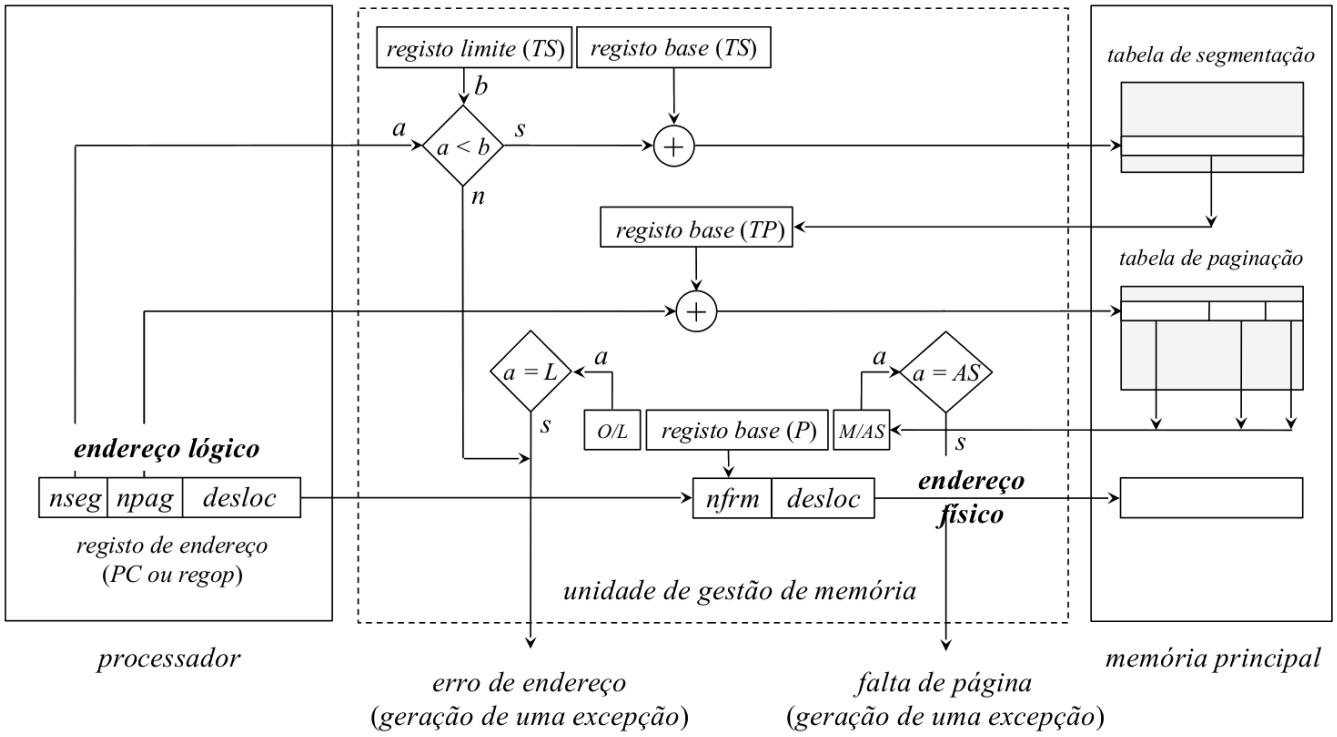


Figure 73: Tradução de um endereço lógico num endereço físico numa arquitectura segmento-paginada

- O endereço lógico passa a possuir 3 campos:
  1. `nseg`: número do segmento
  2. `npah`: identifica a página no segmento
  3. `desloc`: localiza uma posição de memória concreta dentro da página (offset)
- A unidade de gestão de memória contém três registos base e um registo limite associados:
  - **endereço da tabela de segmentação** do processo: registo base (TS)
  - **número de entradas na tabela de segmentação**: registo limite
  - **endereço da tabela de paginação do segmento** que está a ser referenciado: registo base (TP)
  - **frame da memória principal** onde está localizada a página: registo base (P)
- Cada acesso à memória transforma-se em **3 acessos**:
  1. Referencio a **tabela de segmentação do processo** associada com o segmento descrito no campo `nseg` do endereço lógico para obter o **endereço da tabela de paginação do segmento**
  2. Referencio a entrada da **tabela de paginação do segmento** associada com a página descrita no campo `npag` do endereço lógico para obter o `frame` da memória principal onde está localizada a página
  3. Referencio a posição de memória pretendida, concatenando o `nfrm` com o campo `desloc`

Perm	*Endereço em memória da tabela de paginação do segmento*

Figure 74: Conteúdo de cada entrada da tabela de segmentação

O/L	*M/AS*	*Ref*	*Mod*	*N. do frame em memória*	*N. do bloco na área de swapping*

Figure 75: Conteúdo de cada entrada da tabela de paginação de cada segmento

O campo `Perm` é deslocado para a entrada que descreve o segmento. O aceso pode ser tratado de maneira global, sendo as permissões aplicadas ao segmento.

Do ponto de vista do processo passam a ser precisas várias tabelas de paginação e várias tabelas de segmentação, sendo que cada tabela de segmentação pode ter mais que uma tabela de paginação.

65.2 Vantagens vs Desvantagens

65.2.1 Vantagens

- **geral:** pode ser aplicado independentemente do tipo de processos que vão ser executados. quer em número como em tamanho do seu espaço de endereçamento
- **grande aproveitamento da memória principal**:
  - não conduz à fragmentação externa da memória
  - fragmentação interna +e desprezável
- **gestão mais eficiente da memória no que respeita a regiões de crescimento dinâmico**
- minimização do número de páginas que têm de estar residentes em memória principal em cada execução do processo

65.2.2 Desvantagens

- **Exige requisitos especiais de hardware**
  - Nem todos os processadores atuais de uso geral estão preparados para a sua implementação
- **Acesso à memória mais longo**
  - Cada acesso à memória é um **triplo acesso**
  - Pode ser minimizado se a unidade de gestão de memória contiver um **TLB** `translation lookaside buffer`
    * Seria usado para armazenamento das entradas da tabela de paginação recentemente referenciadas no segmento
- **Operacionalmente muito exigente**
  - A sua implementação por parte do SO é mais exigente do que a arquitectura paginada

66 Políticas de Substituição de páginas em memória

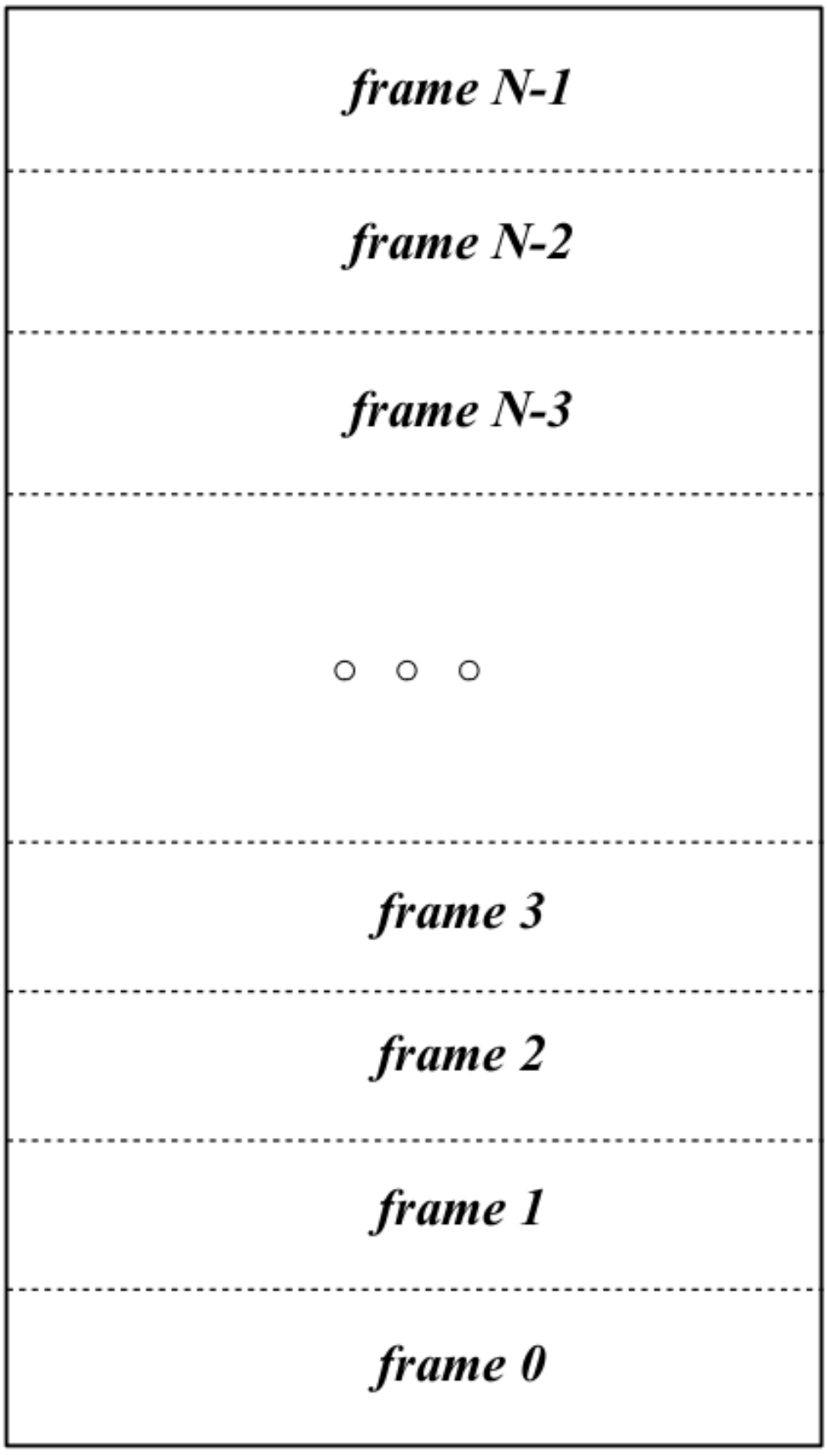


Figure 76: Divisão da memória em frames

Numa arquitectura paginada ou segmentada/paginada a memória principal é vista como dividida operacionalmente me frames do tamanho de cada página

- Cada `frame` vai permitir o armazenamento do conteúdo de uma página do espaço de endereçamento lógico de um processo
- As páginas podem estar em dois estados diferentes:
  - `locked`: **Não podems er removidas de memória**
    * páginas associadas com o *kernel* do SO
    * *buffer cache* do sistema de ficheiros
    * *memory mapped file*
    * *memory mapped variables*
    * *memory mapped IO*
  - `unlocked`: **podem ser removidas de memória**
    * páginas associadas aos processos convencionais

Os `frames` estão associados em listas biligadas:

- se ocupados e associados a páginas `unlocked` $\implies$ `frames` passíveis de substituição.
- `frames` livres

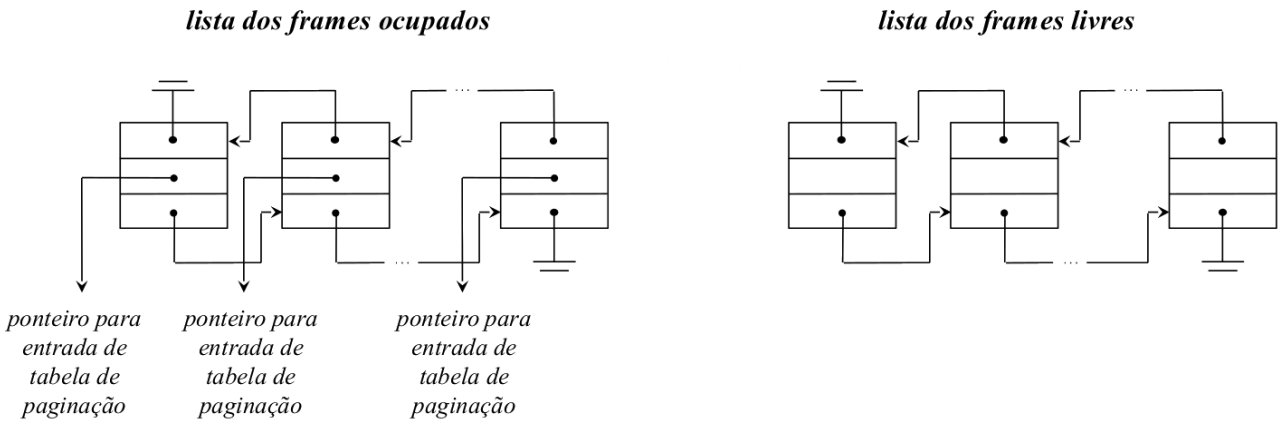


Figure 77: Exemplos do estado das listas biligadas

O tipo de memória implementado pela **lista dos frames** ocupados depende do **algoritmo de substituição utilizado**

Quando ocorre uma **falta de página** (programa tenta aceder a uma dada página que não está em memória):

- a situação mais provável é a lista dos `frames livres` estar vazia
  - torna-se necessário selecionar um `frame` para substituição da lista dos `frames ocupados`
  - alternativamente, pode manter-se sempre na `lista dos frames livres` alguns `frames`
    * usa-se um deles para carregar a página em falta
    * procede-se de seguida **substituição de um `frame ocupado`**
    * é o **método mais eficiente** (as operações decorrem em paralelo)

É necessário ir mudando dinamixamente as páginas dos vários processos que vão existindo em memória. Se assumirmos ainda que os processos em execução ocupam toda a memória disponível, se um dos processos que está em execução precisar de aceder a um bloco que ainda não está em memória, como faço?

O problema que se coloca é: **Que frame escolher para a substituição?**. Em teoria deve ser um `frame` que:

- não irá ser mais referenciado
- ou a sê-lo, sê-lo-á o mais tarde possível

A condição anterior enuncia o **Princípio da Otimalidade**. Ao aplicar estes critérios na escolha da página **Minimiza-se** a ocorrência de outras **faltas de página**

PROBLEMA: o princípio da otimalidade é **não-causal**. Não pode ser diretamente implementado.

Objetivo: Encontrar estratégias de substituição que sejam realizáveis e que ao mesmo tempo, se aproximem tanto quanto possível do princípio da otimalidade

66.1 Algoritmo LRU - Least Recently Used

- Visa encontrar o `frame` que não é referenciado à mais tempo
- Assumo que cada processo vai usar às paginas que usou à menos tempo
- Partindo do **princípio da localidade de referência**, se um `frame` não é referenciado há muito tempo, é fortemente provável que não venha a ser referenciado num futuro próximo
- Cada referência à memória precisa de ser sinalizada com o instante da sua ocorrência (conteúdo de um `timer` ou um contador)
  - Preciso de ordenar cronologicamente quantas páginas possuo em memória
  - Como é pouco provável que a unidade de gestão de memória possua a capacidade de o fazer, será necessário *hardware* especializado
  - Ou então tenho de ir à memória ler a lista em cada pedido de acesso à memória
- Sempre que ocorre uma **falta de página**, a **lista biligada** dos `frames` ocupados tem de ser percorrida para determinar qual o `frame` que foi acedido à mais tempo
  - HEAD: página acedida à menos tempo
    - tem de ser atualizada a cada acesso à memória
  - TAIL: página acedida à mais tempo

Possui um **custo de implementação elevado e pouco eficiente**

66.1.1 Algoritmo NRU - Not Recently used

- Aproximação menos exigente e relativamente eficiente do LRU.
- Usa os bits `Ref` e `Mod` que são processados tipicamente por uma unidade de gestão de memória convencional:
  - Sempre que uma página é acedida para leitura, o campo Ref é colocado a 1
  - Sempre que uma página é acedida para escrita, o campo Ref e Mod são colocados a 1
- Periodicamente o SO percorre a `lista dos frames ocupados` e coloca a **zero o bit `Ref`**
- Quando ocorre uma falta de página os `frames ocupados` enquadram-se numa das classes seguintes

Classes	Ref	Mod
classe 0	0	0
classe 1	0	1
classe 2	1	0
classe 3	1	1

A seleção da página a substituir será feita entre aquelas pertencentes à classe de ordem mais baixa existente atualmente na lista dos frames ocupados

66.2 Algoritmo FIFO - First In, First Out

- Critério baseado no tempo de estadia das páginas em memória principal
- Baseia-se no pressuposto que **quanto mais tempo as páginas residirem em memória, menos provável será que elas sejam referenciadas a seguir**
- A `lista dos frames` ocupados está organizada num FIFO que espelha a **ordem de carregamento** das páginas correspondentes em memória principal

Quando ocorre uma **falta de página:**

- retira-se do FIFO o elemento correspondente à **página há mais tempo em memória**

Algoritmo extremamente falível! Por exemplo:

- Páginas associadas com o código de um editor de texto
- compilador
- bibliotecas do sistema

66.3 Algoritmo da Segunda Oportunidade

- A lista dos `frames ocupados` está organizada num FIFO que espelha a ordem de carregamento das páginas correspondentes em memória principal
- Quando ocorre uma **falta de página:**
  - retira-se do FIFO o **elemento correspondente à página há mais tempo em memória**
  - se o seu bit `Ref` estiver a **zero** $\implies$ a página é escolhida para **substituição**
  - caso contrário, coloca-se o seu bit `Ref` a **zero**
    * o nó é reintroduzido no fim da FIFO
    * o processo repete-se

66.4 Algoritmo do relógio

- Segue a estratégia subjacente ao algoritmo da segunda oportunidade
- Torna a mais eficiente implementando a FIFO numa lista circular
- As operações `fifoIn` e `fifoOut` correspondem a incrementos de um ponteiro

Quando ocorre uma **falta de página:**

- Enquanto o bit `Ref` do `frame` pelo ponteiro **não for zero**:
  - O bit `Ref` é colocado a zero
  - O ponteiro avança uma posição
- A página apontada é escolhida para substituição
- O ponteiro avança uma posição

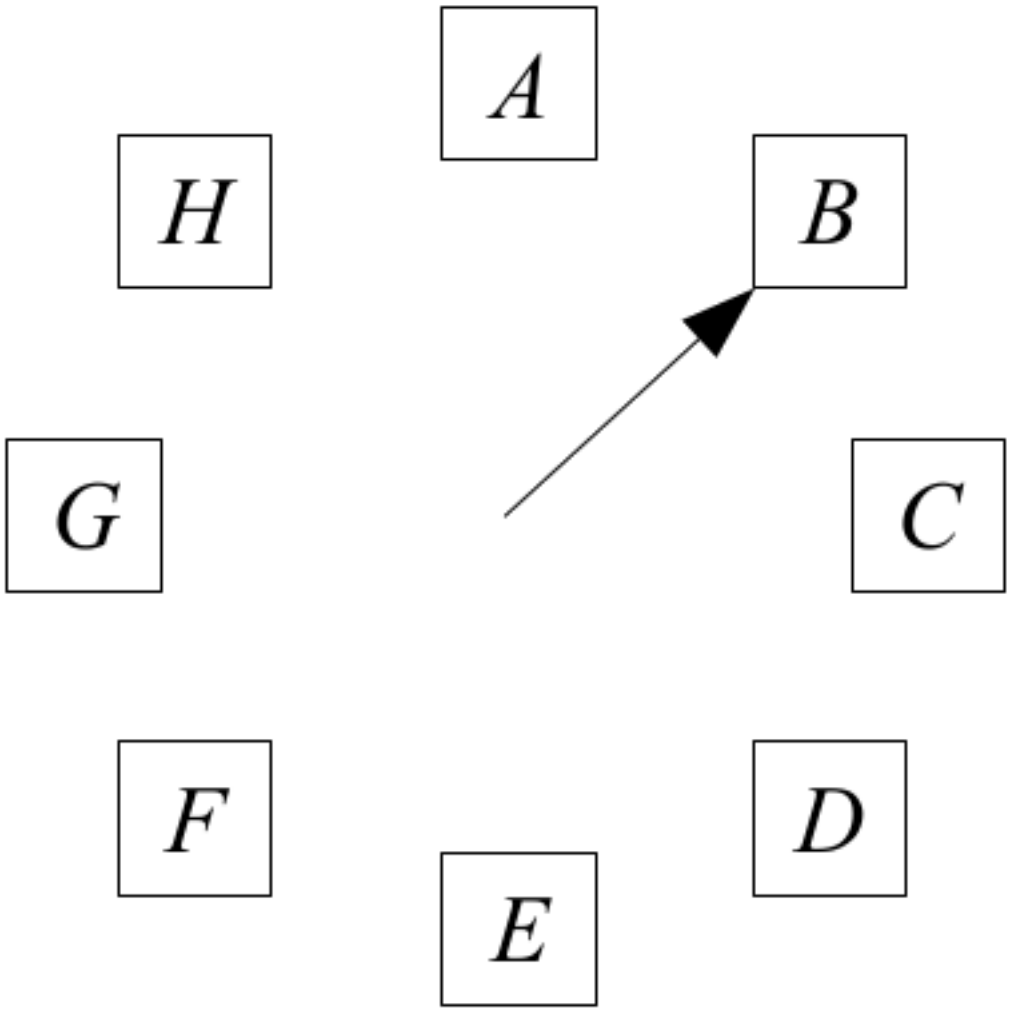


Figure 78: Algoritmo do Relógio

67 Working set

- Quando um processo é colocado pela primeira vez na fila de espera dos processos `READY-TO-RUN`, só a 1ª e última página do seu espaço de endereçamento (início do código e *stack*, respetivamente) é que são carregadas em memória
- Quando o processador for atribuído ao processo, suceder-se-ão inicialmente várias **faltas de página** a um ritmo rápido, porque não possui as páginas necessárias à sua execução em memória principal
- De seguida o número de flatas de página diminui e o processo entra numa fase da execução sem faltas de página
- Dado o princípio da localidade da referência, um processo vai aceder às mesmas variáveis e instruções.
- Assim, todas as páginas associadas à fração do espaço de endereçamento que o processo está atualmente a referenciar já estão todas presentes em memória principal

working set: conjunto de páginas é designado o working set do processo

Ao longo do tempo o `working set` do processo vai variar, não só no que respeita ao número, mas também às páginas concretas que o definem

Se o working set não consegue estar todo em memória vão ocorrer muitas **faltas de página** e o ritmo de execução será muito lento. Ocorre `trashning`:

- `frames` do working set a passarem para da memória para a swap
- `frames` do `working set` na swap a serem passados para a memória

Se não correr `trashing`, o processo alterna entre períodos curtos que sofrerá muitas **faltas de páginas** e períodos longos quase sem **faltas de página**

O **objetivo prioritário** de qualquer política de substituição é garantir que mantem sempre o `working set` do processo em memória principal

Uma estratégia consiste em atribuir novos `frames`ao process sempre que este se encontre num período elevado de **faltas de página** e retirar-lhe `frames` quando as ocorrência de faltas de página baixar.

68 Demand paging vs prepaging

Quando um processo é introduzido na fila de espera dos processos `READY-TO-RUN` pela 1ª vez ou em resultado de uma suspensão, é preciso decidir que páginas colocar em memória principal.

- `Demand paging`: Estratégia minimalista e menos eficiente
  - nenhuma página é colocada
  - o mecanismos de geração de **faltas de página** é que é responsável por formar o `working set` do processo
- `prepaging`: estratégia mais eficiente
  - procura-se adivinhar o `working set` do processo para minimizar a geração de faltas de página
  - na 1ª vez são colocadas as primeiras duas páginas atrás referidas
  - nas vezes seguintes, são colocadas o conjunto de páginas que residiam em memória no **momento em que o processo foi suspenso**

68.1 Substituição global vs substituição local

Qual o âmbito da aplicação dos algoritmos de substituição?

- **local**: a escolha é efetuada entre o conjunto de `frames` de um processo
- **global**: a escolha é efetuada entre o conjunto de todos os `frames` que continuem a lista de `frames` ocupados

É preferível o âmbito de aplicação global:

- Penso em cada processo globalmente
  - É mais fácil gerir `working sets` que mudam de dimensão
  - Permite-me suportar grandes variações de `working set` dos processos sem resultar desperdício de memória ou `trashing`
  - Desde que os processos não for em demasia, é possível minimizar o `thrashing`
    * Se a soma dos `working sets` de todos os processos é superior ao número de `frames unlocked` disponíveis em memória principal, entro em `thrashing`
    * A solução passa por ir suspendendo processos até que o `trashing` desapareça