VM/370 – IBM Virtual Machine

AULA 23 - PCS 3446 - Estudo de Caso - IBM VM-370

Prof. João José Neto

Fonte do material desta apresentação:

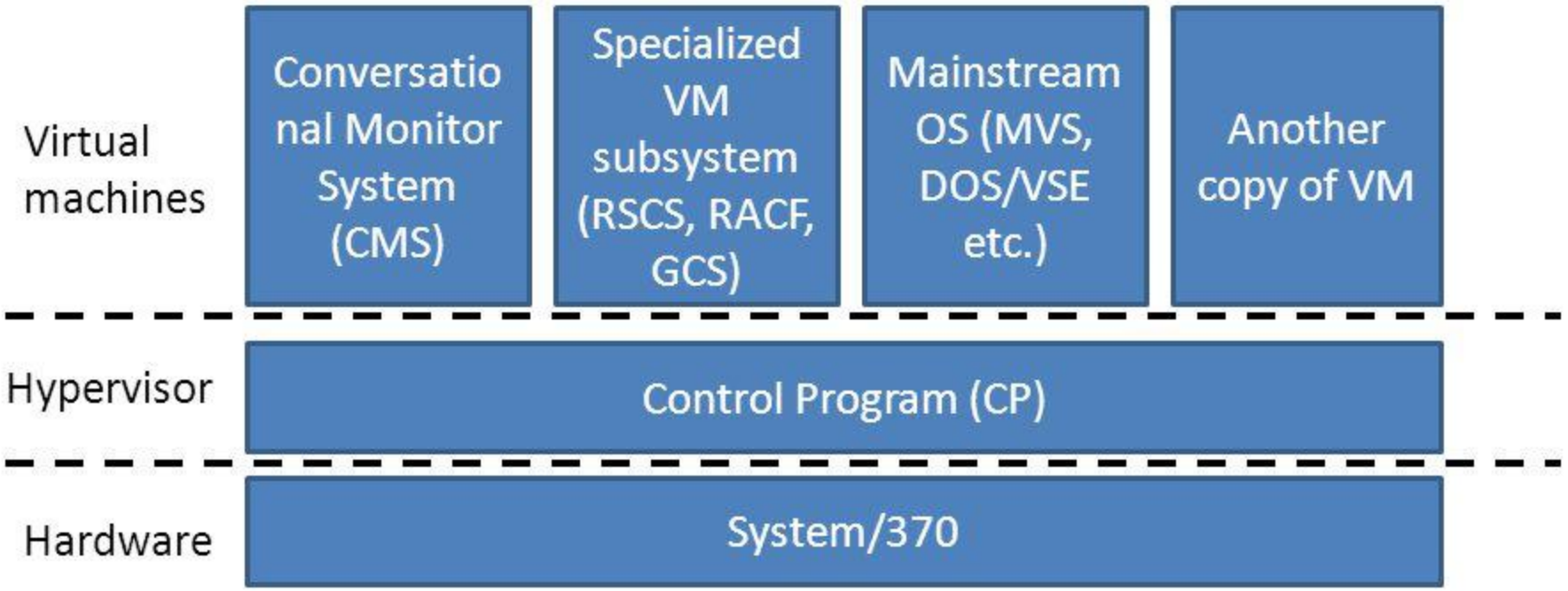
Madnick e Donovan - Operating Systems, cap. 9.5 - McGraw-Hill, 1974

SYSTEM IBM/370 MODEL 165

Introdução

- VM/370 (Virtual Machine/370) é um exemplo de VMM (Virtual Machine Monitor)
- VMM é um tipo de sistema operacional que apenas multiplexa entre os usuários os recursos físicos do sistema, sem quaisquer serviços adicionais
- A máquina estendida é exatamente a mesma que a máquina (hardware) que executa o VMM
- Assim, o VM/370 faz com que o sistema no qual ele é executado tenha para o usuário a aparência de ser na realidade diversos sistemas separados
- Isso se obtém controlando a multiplexação dos recursos físicos do hardware, de forma análoga à da operação de sistemas telefônicos quando multiplexam comunicações usando fios e equipamentos comuns, e posteriormente separando os diálogos individuais, e encaminhando-os aos seus destinatários.

IBM VM/370

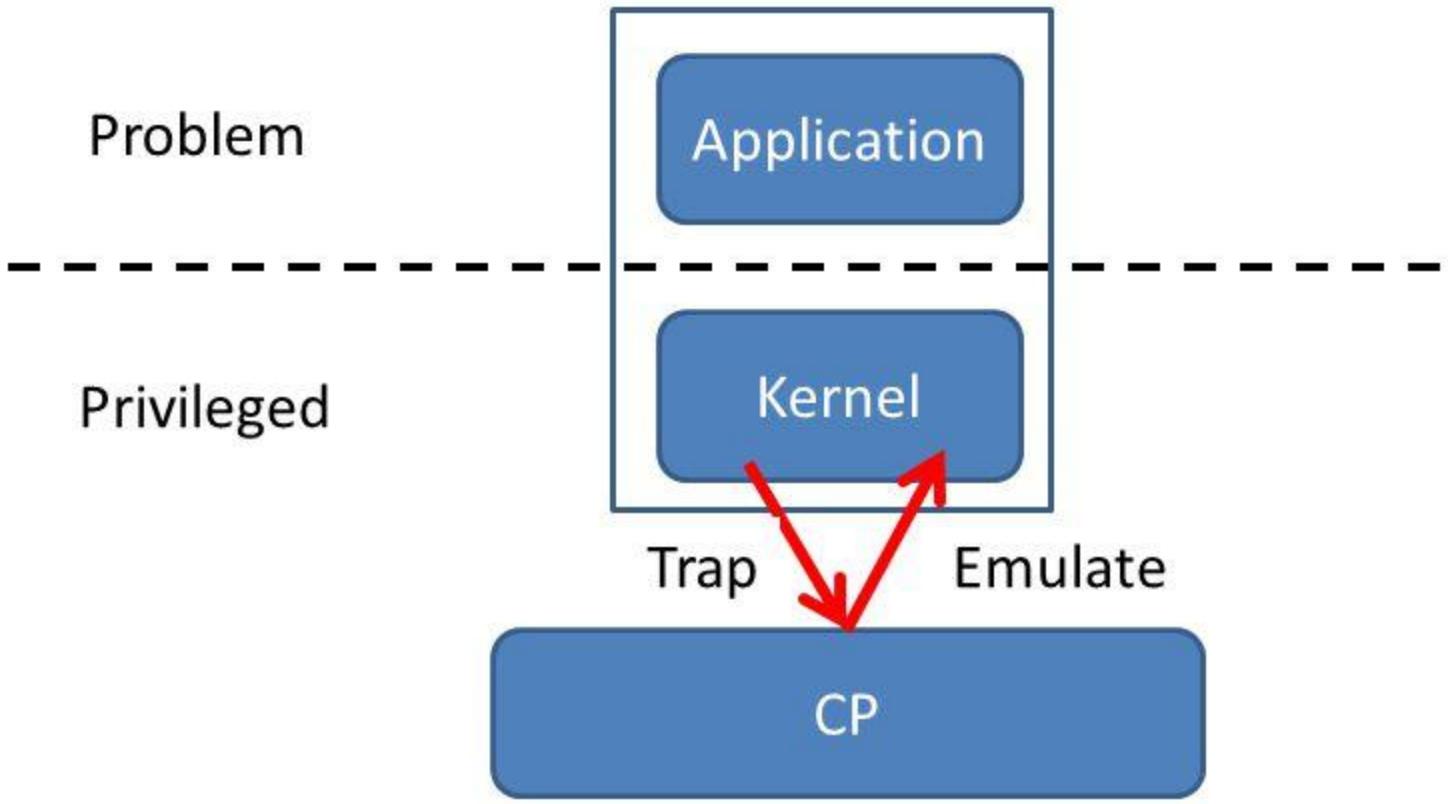

Virtual machines
Conversational Monitor System (CMS)
Specialized VM subsystem (RSCS, RACF, GCS)
Mainstream OS (MVS, DOS/VSE etc.)
Another copy of VM
Hypervisor
Control Program (CP)
Hardware
System/370

A virtualização e o VM/370

- À primeira vista inútil, a ideia de converter uma máquina em outras iguais a ela é na verdade o que permite dar a cada potencial usuário uma máquina (virtual, mas equivalente à máquina real) que seja propriedade exclusivamente sua
- Assim, se pode dispor de diversos /370 onde há um apenas, e cada usuário poderá usá-lo como quiser, escolhendo trabalhar com seu sistema operacional preferido
- Os sistemas que operam no VM/370 são executados em modo usuário, mas se comportam como se estivessem operando em modo supervisor
- Ao executar uma instrução privilegiada, os programas provocam uma interrupção, cujo tratamento é efetuado pelo VM/370. Para o programa, o resultado é o mesmo que ocorreria se a instrução privilegiada tivesse sendo executada pelo próprio hardware.
- As interrupções das instruções privilegiadas constituem a real interface entre os usuários e o VM/370

IBM VM/370

- Technology: trap-and-emulate



Origens

- Começou com o projeto CP-40/CMS, uma pesquisa da IBM Cambridge Scientific Center (IBM CSC)
- Em 1967, atendia até 15 usuários na IBM CSC
- Objetivos dessa pesquisa
  - Estudo de várias técnicas e métodos de timesharing
  - Avaliação de requisitos de hardware para timesharing
  - Implementação de sistema de timesharing para uso local
  - Desenvolvimento de formas de observar a interação entre o sistema operacional e seu hardware hospedeiro

Influências

- Dois desenvolvimentos impactaram a evolução do CP-40/CMS paraa o VVM/370:
  - Problemas de compatibilidade ou com recursos especiais, em instalações usando sistemas OS/360 com vários sistemas operacionais simultâneos
  - Atrasos na disponibilização do sistema IBM Time Sharing System/360 (projetado para o IBM system/360 modelo 67 com hardware de mapeamento dinâmico de endereços)
- Em 1966 uma força tarefa liberou o CP-67/CMS
- O VM/370, anunciado em 1972, foi sucessor do sistema CP-40/CMS e do CP-67/CMS, e pode ser executado em qualquer sistema IBM/370 equipado com o seu hardware de mapeamento dinâmico (Dynamic Address Translation)

Usos e vantagens

- O sistema de hardware virtual apresenta:
  - Concorrência de vários S.O. com diferentes usuários
  - Elimina a necessidade de conversão entre sistemas
  - Permite desenvolver programas para máquinas diversas
  - Permite simular comunicações entre máquinas
  - Permite usar para avaliação de desempenho as interrupções associadas à interceptação de instruções especiais
  - Confiabilidade
  - Segurança e privacidade resultam da isolação entre máquinas virtuais

Simulação dos sistemas /360 e /370

- VM/370 tem 2 componentes utilizadas em conjunto
  - CP (Control Program), o monitor da máquina virtual
  - CMS (Conversational Monitor System), um sistema operacional básico
- System Control Panel – associa a cada máquina virtual um teclado (remoto ou local) e mapeia para esse dispositivo os controles do painel da máquina
- CPU – Executando os programas da máquina virtual em modo usuário, garante-se que as interrupções devidas ocorrerão sempre que instruções privilegiadas forem executadas. Suas rotinas de tratamento simularão as funcionalidades da instrução que causou a interrupção
- Sistema de entrada e saída – toda entrada e saída são simuladas usando as interrupções das instruções privilegiadas.
  - Se o dispositivo físico existir e estiver alocado à máquina virtual, a entrada/saída é realizada diretamente nele
  - Se existir dispositivo similar, mapeiam-se os comandos para simulá-lo
  - Se não, haverá simulação extensiva usando técnica de virtualização de dispositivos similar à de SPOOLing
- Memória – utiliza memória virtual, com alocação dinâmica de páginas requisitadas. Cada máquina virtual tem sua própria memória virtual.

Hardware

- O CP do VM/370 usa o hardware padrão /370 com Dynamic Address Translation
- Modos supervisor e usuário
- Instruções privilegiadas e respectivas interrupções
- Posições especiais de memória são tratadas pelo CP
- Endereços de entrada/saída são tratadas pelo CP
- Alguns modelos disponibilizam microprogramação para o tratamento de interrupções por firmware

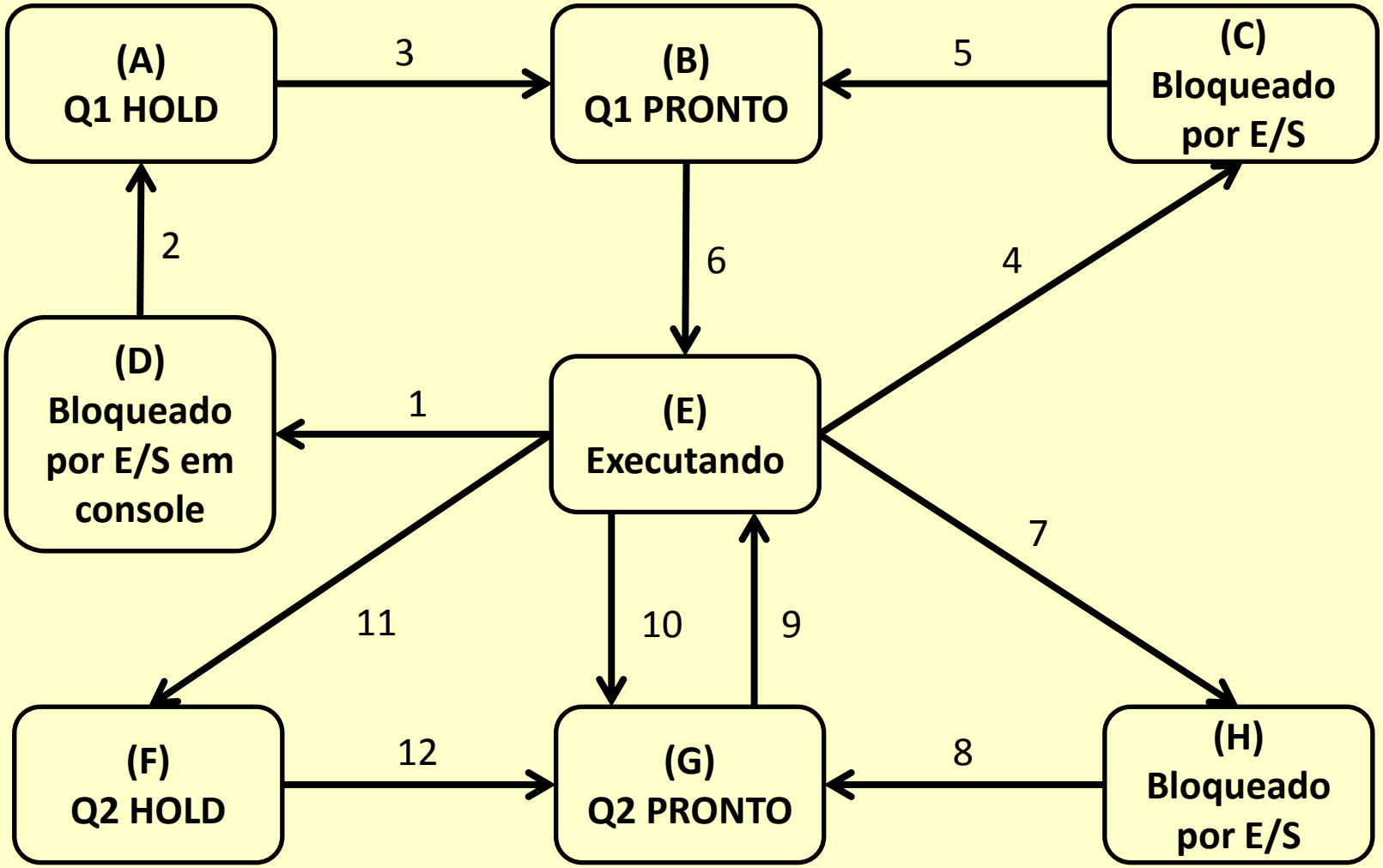
Administração de Memória

- CP usa paginação requisitada.
- A aproximação LRU é a política de substituição de páginas adotada
- CP usa tabelas para descrever as características de cada máquina virtual
- CP dá suporte ao bloqueio (ou não) do uso do hardware do Dynamic Address Translation

Administração de Processador

- Finalidades (similares às do sistema MULTICS)
  - Limitar o número de jobs simultaneamente na multiprogramação para evitar thrashing
  - Dar preferências a jobs interativos
- Os estados dos jobs, controlados pelo CP
  - Os jobs são administrados usando-se duas listas:
    - Q1 – jobs interativos / com muitas operações de entrada/saída
    - Q2 – jobs com muito processamento

Diagrama de estados do CP



Transições do CP

DE	PARA	EVENTO	ESTADO
A	**B**	Q1 não lotado (mais antiga)	A = Q1 HOLD
B	**E**	Escolha por round robin	B = Q1 READY
C	**B**	Término de E/S	C = Bloqueado por E/S
D	**A**	Término de entrada pela console	D = Bloqueado por E/S de console
E	**C**	Início de operação de E/S	E = Executando
E	**D**	Aguardar entrada pela console	
E	**F**	Interrupção de Time Slice	
E	**G**	Job Q2 esgota time slice de 50ms	
E	**H**	Job Q2 Início de operação de E/S	
F	**G**	Q2 não lotado (mais antiga)	F = Q2 HOLD
G	**E**	Só se Q1 ready estiver vazia	G = Q2 READY
H	**G**	Término de operação de E/S	H = Bloqueado por E/S

Observações

- Transição E para F
  - Se job veio de Q1: interrupção de time slice
  - Se mais de 400 ms sem atividade de console
  - Se o job veio de Q2: ultrapassou 5s de CPU acumulados
- Transição G para E
  - Selecionar job a partir de Q2, na seguinte ordem
    - Mais novo em Q2 acabou de ser incluído
    - Mais antigo I/O-bound (sem time-slice)
    - Mais antigo CPU-bound (com time-slice)

Estratégia adotada

- Os jobs multiprogramados ativos em um dado instante se dividem em duas listas: Q1 e Q2
- Os demais se mantêm em estado de espera, aguardando para entrarem em Q1 e Q2
- Jobs I/O-bound e interativos tendem a migrar para Q1
- Jobs CPU-bound tendem a ir para Q1
- Para evitar thrashing e assegurar boa utilização para o sistema, impõem-se limites para o número de jobs permitidos em Q1 e Q2

Dinâmica

- Um novo job é colocado no estado Q1 HOLD
- Quando houver vaga em Q1 READY, este estado recebe o job mais antigo de Q1 HOLD
- Dá-se forte preferência a jobs com muita entrada/saída:
  - Jobs migram de Q2 HOLD para Q2 READY usando essa mesma lógica
  - Jobs em Q1 READY se revezam em round-robin para executar
  - Jobs em Q2 READY só recebem o processador se não houver jobs em Q1 READY

Alternância entre Q1 e Q2

- Precisando esperar por entrada/saída de console,o job é removido de Q1 e de Q2.
- Concluída a operação da console, o job é introduzido novamente em Q1 HOLD.
- Para um job em Q1 ou Q2 ser colocado em Q2 HOLD é preciso que ocorra uma das condições a seguir
  - Estando em Q1, esgota seu quantum (50ms) sem pedir qualquer tipo de operação de entrada/saída
  - Estando em Q1, acumula 400ms de processamento sem operações de entrada/saída na console
  - Estndo em Q2, acumula 2000ms de processamento sem operações de entrada/saída na console
- O operador pode, adicionalmente, impor um limite ao total de usuários conectados ao sistema

Administração de Dispositivos

- CP emprega 3 estratégias:
  - SPOOLed – para E/S em dispositivos de leitura, impressão, perfuração e similares
  - Dedicados – para E/S em fita ou algum outro dispositivo escolhido
  - Compartilhado – para E/S em disco, disquete ou similar, e para controladores de comunicação

P/ dispositivos dedicados e compartilhados

- Copia CCW para a área do sistema
- Traduzir endereços de CCW para refletir endereços físicos (trazendo para a memória páginas de dados se necessário)
- Traduzir comandos de canal que indicam regiões de dados fora dos limites da página, convertendo-os para a forma de múltiplos comandos encadeados
- Escalar a operação de entrada/saída
- Simular a interrupção de entrada/saída para a máquina virtual quando a entrada/saída se completar
- Liberar as páginas de dados

Para dispositivos virtuais (SPOOLed)

- SPOOLing é uma atividade guiada pela ocorrência de eventos de interrupção
- Terminais são dispositivos semi-SPOOLed, com apenas uma cadeia de CCW de entrada/saída

Administração de Informação

- CP não disponibiliza serviços de gerenciamento de informação, já que o sistema /370 não dispõe de nenhum em seu hardware
- Obtêm-se tais serviços ao escolher um sistema operacional a ser executado em uma das máquinas virtuais do VM/370
- O CMS (conversational monitor system) é um desses possíveis sistemas operacionais, compatível com o VM/370
- Conceitualmente o CMS poderia ser executado, como sistema monousuário, diretamente sobre o sistema/370
- Todavia, ele foi projetado para executar melhor nas máquinas virtuais do VM/370


Disc Pack IBM 2314
14”, 7,5 MB, 156 KB/s, US$175,000

IBM 2314 Direct Access Storage Facility

Merlin 3030 disk drive / pack (US$74,000 a US$87,000)

IBM 3030 data storage

- CMS exige uma configuração system/370 real ou virtual, com dois discos no mínimo
- Em ambiente virtual, os dispositivos dedicados de armazenamento podem ser o 2314 ou o 3330

- Tipicamente, disquetes são simulados em discos hospedeiros grandes, como se fossem dispositivos do mesmo tipo, porém menores
- CMS mantém um arquivo de diretório para cada dispositivo, similar aos VTOCs do /360
- Um dos dispositivos de armazenamento (System Disk) mantém as partes não residentes do CMS, bem como os programas de sistema: montadores, compiladores, ferramentas, etc
- Os demais são usados para armazenar os arquivos permanentes dos usuários

- CP permite dois ou mais dispositivos virtuais de armazenamento no mesmo disco físico
- CMS usa um disco real no sistema, e CMS virtuais usam esse mesmo disco
- CP neste caso pode impedir os usuários de alterar o System Disk (read only)
- Arquivos CMS são descritos por: nome, tipo (FORTRAN, TEXT, LISTING), modo (system disk, permanent disk)

Estrutura do disco

- Um dispositivo de armazenamento com sistema de arquivos CMS apresenta a seguinte estrutura:
  - O dispositivo é formatado em blocos de 800 bytes
  - Alguns blocos são ocupados pelo diretório de arquivos e pelo diretório de alocação de área
  - Um Master Directory Block é utilizado para indicar as posições dos diretórios
  - Em operação normal, esses diretórios são copiados para a memória
  - Os demais blocos são empregados para: (1) File map blocks; (2) File data blocks; (3) Unused blocks

- O Space Allocation Directory é um bit map usado para mapear blocos não ocupados
- Havendo n blocos no dispositivo, esse diretório terá n bits de comprimento
- O i-ésimo desses bits vale 0 ou 1 conforme o i-ésimo bloco esteja livre ou ocupado, respectivamente
- O File Directory contém um elemento para cada arquivo existente no dispositivo de armazenamento

- Cada elemento do File Directory indica: (1) nome do arquivo; (2) comprimento corrente; (3) status
- Adicionalmente, cada elemento especifica o Storage Block Number do file map primário do arquivo
- O Primary File Map tem duas tabelas de números de blocos: uma, para até 60 Data Blocks; outra, para até 40 blocos de File Map secundário
- A estrutura de mapeamento de arquivos do CMS permite um acesso direto a registros do arquivo mais eficiente que o sistema CTSS, usado no sistema de Time Sharing


IBM
370