## Introdução à Organização Básica e Arquitetura de Computadores

## Objetivo

Tratar aspectos técnicos da arquitetura e organização de computadores para permitir uma visão ampla dos componentes e técnicas dessa área que influenciam diretamente no desenvolvimento de programas de computadores de qualidade. Permitir ao aluno conhecer os recursos que os computadores oferecem para o desenvolvimento e execução de programas.

## Arquitetura x Organização de Computadores

Na literatura, ao se descrever um sistema de computação, é feita uma distinção entre os termos arquitetura e organização do computador.

O termo arquitetura de um computador refere-se aos atributos de um sistema que são visíveis para o programador ou, em outras palavras, aos atributos que têm impacto direto sobre a execução lógica de um programa. *Convém lembrar que o termo arquitetura também significa projeto.* 

Já o termo organização de um computador diz respeito às unidades operacionais e suas interconexões que efetivam as especificações de sua arquitetura, ou seja, como as características da arquitetura são implementadas.

#### Por atributos de arquitetura, podemos entender:

- O conjunto de instruções;
- O número de bits usados para representar os tipos de dados;
- Os mecanismos de E/S. etc.

#### Enquanto os atributos da organização:

São transparentes aos usuários e incluem detalhes de hardware como, por exemplo: sinais de controle, tecnologia de memória utilizada, etc.

Especificar se um computador deve ou não ter uma instrução de multiplicação constitui uma decisão de projeto da arquitetura.

Por outro lado, definir se essa instrução será implementada por uma unidade específica de multiplicação ou por um mecanismo que utiliza repetidamente sua unidade de soma é uma decisão de projeto de sua organização.

## Por exemplo

Todo microprocessador Intel da família x86 compartilha a mesma arquitetura básica. Isto garante que eles sejam compatíveis. No entanto, a organização difere de uma versão para a outra. Ou seja, uma organização deve ser projetada para implementar uma especificação particular de arquitetura.

### Estrutura e Função dos Computadores

Tanto a estrutura quanto as funções de um computador são muito simples. Basicamente, podemos definir estrutura e função como segue:

Estrutura: é a forma como os componentes se relacionam uns com os outros.

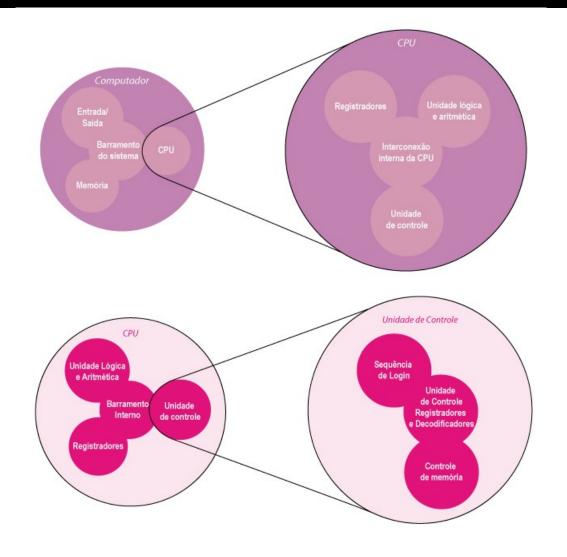
Função: a operação que cada componente individual realiza dentro da organização.

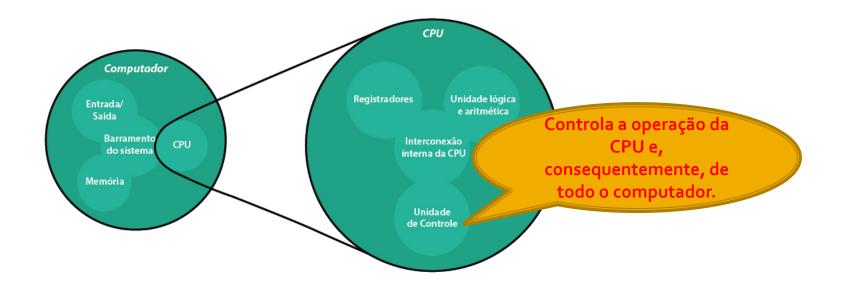
Os computadores são equipamentos eletrônicos digitais que têm como função:

- ✓ Processar dados;
- ✓ Armazenar dados;
- ✓ Comunicar (transferir) dados;
- ✓ Exercer controle.

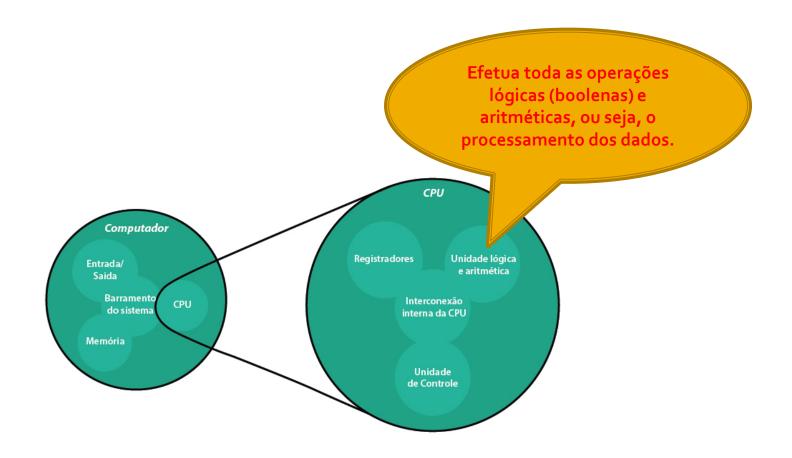
## Para cumprir esses objetivos, eles possuem a seguinte estrutura básica:

- Unidade Central de Processamento (CPU);
  - > Processa os dados através da execução das instruções que compõem os programas;
- Sistema de Armazenamento (Memórias):
  - Armazena dados e instruções (programas). Normalmente um computador possui mais de uma memória, cada uma com sua especificidade;
- Sistema de Entradas e Saídas (E/S):
  - > Transfere dados entre o computador e o usuário, periféricos, outros computadores, ou o ambiente.
- Sistema de Interconexão (Barramentos):
  - Meios de comunicação entre a CPU, memórias e E/S.

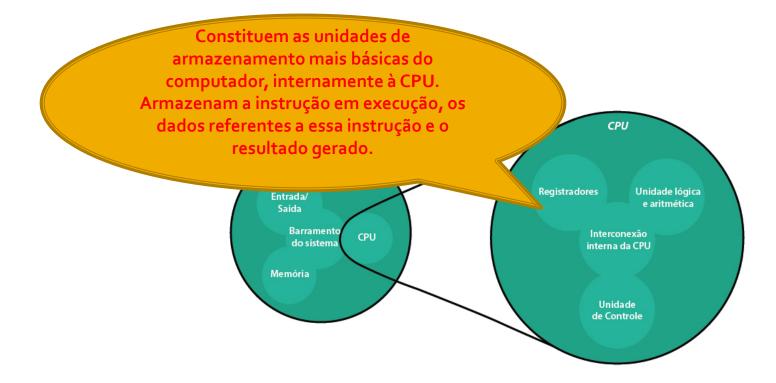




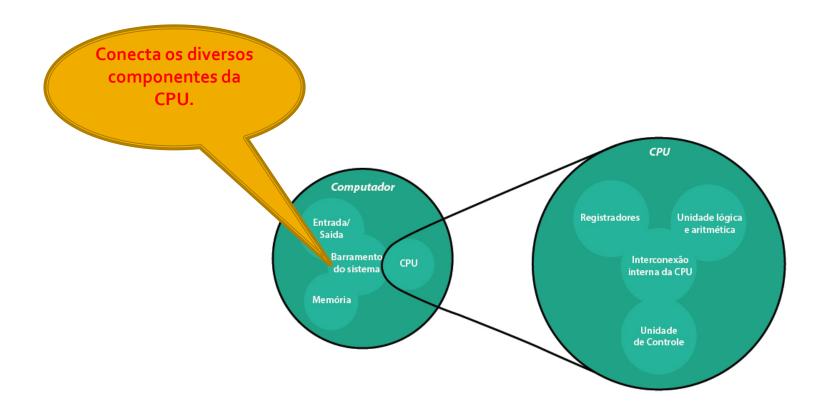
A CPU é o principal componente de um computador. Poderia-se dizer que ela é o "cérebro" do computador.



A CPU é o principal componente de um computador. Poderia-se dizer que ela é o "cérebro" do computador.



A CPU é o principal componente de um computador. Poderia-se dizer que ela é o "cérebro" do computador.



A CPU é o principal componente de um computador. Poderia-se dizer que ela é o "cérebro" do computador.

## Observação:

Os microprocessadores foram criados inicialmente para serem uma CPU em um único circuito integrado porém, é importante ressaltar que atualmente os microprocessadores possuem várias CPUs em um único circuito integrado, além de partes do sistema de armazenamento e do sistema de E/S.

## Estrutura básica do computador; busca e execução de instruções; interrupções; barramentos

Abordaremos agora assuntos referentes aos diversos tipos de barramentos existentes em um computador e ao modo como se comunicam.



#### Barramentos

Os barramentos são portas pelas quais o processador pode comunicar-se com os demais componentes do micro, como a placa de vídeo.

Falando em placa de vídeo, você já percebeu que todas as placas de vídeo modernas são conectadas em slots PCI ou AGP?

E que placas de som e modems antigos quase sempre usam slots ISA? Isso acontece porque placas de som e modems são periféricos relativamente lentos, para os quais o lento barramento ISA já é suficiente.

Porém, as placas de vídeo necessitam de um barramento muito mais rápido, motivo pelo qual utilizam slots PCI ou AGP.

Que tal se agora estudássemos os diferentes tipos de barramento existentes?

#### Barramentos antigos

Os processadores 8088, usados nos micros XT, comunicavam-se com os demais periféricos usando palavras binárias de 8 bits. Para o uso em conjunto com esses processadores, foi criado o ISA de 8 bits.

Esse barramento funciona usando palavras binárias de 8 bits e opera a uma frequência de 8 MHz, permitindo uma passagem de dados a uma velocidade de 8 megabytes por segundo, velocidade muito mais do que suficiente para um processador lento como o 8088.

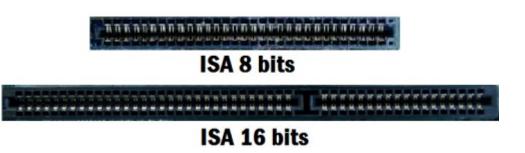
## ISA (Industry Standard Architecture)

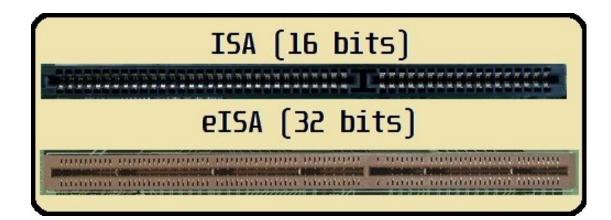
Os processadores 286 se comunicavam com os demais periféricos usando palavras de 16 bits. Para acompanhar essa melhora por parte do processador, foi criada uma extensão para o barramento ISA de 8 bits, formando o ISA de 16 bits. Esse barramento, assim como o processador 286, trabalha com palavras de 16 bits, a uma frequência de 8 MHz, permitindo um barramento total de 16 MB/s.

Os periféricos ISA vem sendo usados desde a época do 286, mas, na verdade, esse padrão já existe desde 1981, ou seja, quase 40 anos de idade!. O ISA é um bom exemplo de padrão obsoleto que foi ficando, ficando, ficando, mesmo depois de terem sido criados barramentos muito mais rápidos, como o PCI.

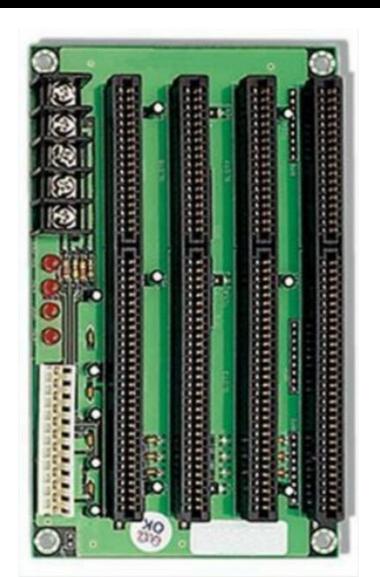
A verdade é que o ISA durou muito tempo porque o barramento de 16 megabytes por segundo permitido por ele é suficiente para acomodar periféricos lentos como modems e placas de som, fazendo com que os fabricantes desses periféricos se acomodassem, e continuassem produzindo periféricos ISA praticamente até hoje.

#### ISA & eISA





Usado para encaixar placas de expansão, com modems, placas de som e placas de vídeo, é um tipo de barramento que está a cair em desuso por ser relativamente lento em relação ás novas tecnologias.



#### MCA (Micro Channel Architecture)

Com o surgimento dos processadores 386, que trabalhavam usando palavras binárias de 32 bits, tornou-se necessária a criação de um barramento mais rápido que o ISA para o uso de periféricos rápidos, como placas de vídeo e discos rígidos. A IBM criou então o MCA, que funcionava com palavras de 32 bits e a uma frequência de 10 MHz, sendo 2,5 vezes mais rápido que o ISA de 16 bits.

Apesar de trazer recursos surpreendentes para a época em que foi lançado, como o bus mastering e suporte ao plug-and-play (foi o primeiro barramento a suportar esses recursos, isso em 1987), o MCA não conseguiu se popularizar por causa de seu alto custo, incompatibilidade com o ISA e, principalmente, por ser uma arquitetura fechada, caindo em desuso com o surgimento do EISA e do VLB.

#### EISA (Extended ISA)

Esse novo barramento foi uma resposta dos demais fabricantes liderados pela Compac ao MCA, criado e patenteado pela IBM.

Com o objetivo de ser compatível com o ISA, o EISA funciona também a 8 MHz, porém, trabalha com palavras binárias de 32 bits, totalizando 32 MB/s de barramento, sendo duas vezes mais rápido do que seu antecessor. O EISA também oferecia suporte a bus mastering e plug-and-play, com eficiência comparável à do MCA.

Uma das grandes preocupações dos fabricantes durante o desenvolvimento do EISA foi manter a compatibilidade com o ISA. O resultado foi um slot com duas linhas de contatos, capaz de acomodar tanto placas EISA quanto placas ISA de 8 ou 16 bits.

#### VLB (VESA Local Bus)

Lançado em 1993 pela Video Electronics Standards Association, o VLB é muito mais rápido que o EISA ou o MCA, sendo utilizado por placas de vídeo e controladoras de disco, as principais prejudicadas pelos barramentos lentos. Com o VLB, os discos rígidos podiam comunicar-se com o processador usando toda a sua velocidade, e se tornou possível a criação de placas de vídeo muito mais rápidas.

Como antes, existiu a preocupação de manter a compatibilidade com o ISA, de modo que os slots VLB são compostos por três conectores. Os dois primeiros são idênticos a um slot ISA comum, podendo ser encaixada neles uma placa ISA, sendo o terceiro destinado às transferências de dados a altas velocidades permitidas pelo VLB.

As desvantagens do VLB são a falta de suporte a bus mastering e a plug-and-play, além de uma alta taxa de utilização do processador e limitações elétricas, que permitem um máximo de 2 ou 3 slots VLB por máquina.

## PCMCIA (Personal Computer Memory Card International Association)

O PCMCIA é utilizado principalmente em notebooks e handhelds em que, na maioria das vezes, é o único meio de conectar placas de expansão.

A principal vantagem desses dispositivos é o tamanho: todos possuem dimensões um pouco menores que as de um cartão de crédito, apenas mais espessos.

Atualmente é possível encontrar praticamente qualquer tipo de dispositivo na forma dessas placas: modems, placas de som, placas de rede, placas decodificadoras de DVD, cartões de memórias SRAM e memórias flash e, até mesmo, discos rígidos removíveis.

#### AMR (Audio Modem Riser)

Esse é um padrão de barramento que permite o encaixe de placas de som e modems controlados via software. O slot AMR se parece com um slot AGP, mas tem apenas 1/3 do tamanho deste. O objetivo é permitir a criação de componentes extremamente baratos para serem usados em micros de baixo custo.

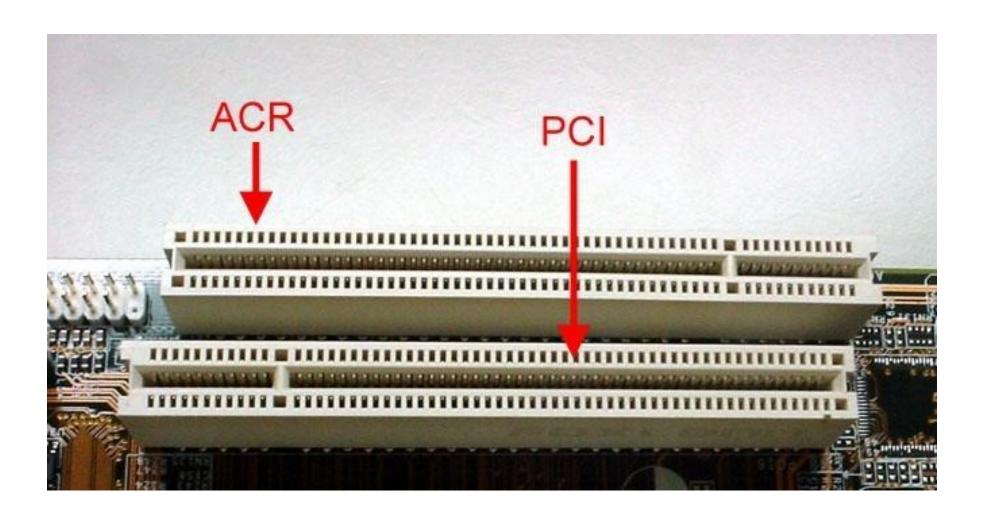
A vantagem é o preço, já que uma placa de som ou modem AMR não custa mais de 5 ou 7 dólares para o fabricante (um pouco mais para o consumidor final, naturalmente). A desvantagem, por sua vez, é o fato desses componentes serem controlados via software, o que consome recursos do processador principal, tornando o micro mais lento.

## ACR (Advanced Communications Riser)

O ACR é um padrão desenvolvido por uma associação de vários fabricantes, que inclui a AMD, Lucent, Motorola, 3Com, Nvidia, Texas Instruments e Via. Os slots ACR se parecem com um slot PCI invertido; na verdade os fabricantes optaram por aproveitar o mesmo encaixe para cortar custos, mas as semelhanças param por aí, já que foram mudadas a posição e a sinalização elétrica do slot.

Os slots ACR são risers para a conexão de placas de som e modems de baixo custo, assim como os slots AMR. Muitas placas atuais trazem um slot ACR, mas os fabricantes evitam desenvolver placas com dois ou mais slots ACR para não diminuir o número de slots PCI da placa.

A principal vantagem do ACR sobre o AMR é que, enquanto o AMR permite que o riser inclua apenas modem e placa de som, no ACR o riser pode conter praticamente todos os tipos de dispositivo, desde modems e placas de som baratas, controlados via software, até placas de rede, modems ADSL ou ISDN, placas de som e modems controlados via hardware etc.

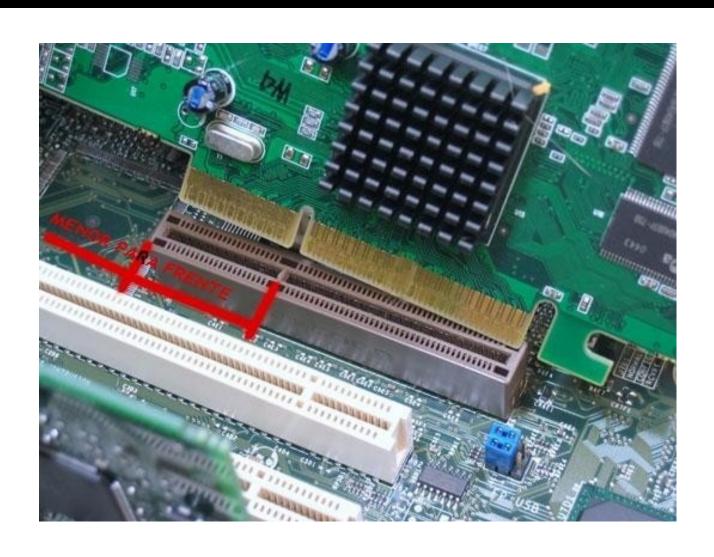


## AGP (Accelerated Graphics Port)

O AGP é um barramento feito sob medida para as placas de vídeo. O AGP foi criado com base nas especificações do PCI 2.1 e opera ao dobro da velocidade do PCI, ou seja, 66 MHz, permitindo transferências de dados a 266 MB/s, contra apenas 133 MB/s possíveis pelo barramento PCI.

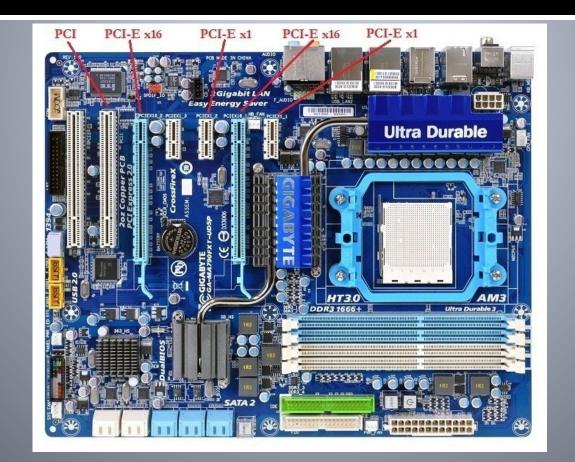
Além da velocidade, o AGP permite que uma placa de vídeo possa acessar diretamente a memória RAM para armazenar texturas. Esse é um recurso muito utilizado em placas 3D, que usa a memória RAM para armazenar as texturas que são aplicadas sobre os polígonos que compõem a imagem tridimensional.

## AGP (Accelerated Graphics Port)



#### Barramentos mais recentes

Atualmente, os computadores possuem barramentos mais rápidos capazes de suportar vários equipamentos em uma mesma porta. Os principais barramentos utilizados hoje são:



## PCI (Peripheral Component Interconnect)

Criado pela Intel, o PCI é tão rápido quanto o VLB, porém mais barato e muito mais versátil. Outra vantagem é que, ao contrário do VLB, ele não é controlado pelo processador, e sim por uma controladora dedicada, incluída no chipset. Além de diminuir a utilização do processador, isso permite que o PCI seja empregado com qualquer processador, sem qualquer tipo de modificação.

Atualmente, todos os periféricos rápidos, placas de vídeo e controladoras de disco usam quase obrigatoriamente o barramento PCI. Componentes mais lentos, como placas de som e modems, ainda podem ser encontrados em versões ISA, apesar de cada vez mais acharmos esses componentes em versões PCI.

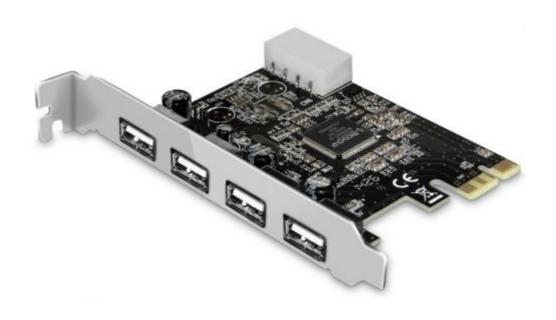
#### USB (Universal Serial Bus)

O USB é um padrão para a conexão de periféricos externos. Suas principais armas são a facilidade de uso e a possibilidade de se conectar vários periféricos a uma única porta USB.

É o primeiro barramento para micros PC realmente plug-and-play. Podemos conectar periféricos mesmo com o micro ligado, bastando fornecer o driver do dispositivo para que tudo funcione sem ser necessário nem mesmo reinicializar o micro. A controladora USB também é suficientemente inteligente para perceber a desconexão de um periférico.

Podemos conectar até 127 periféricos em filaa uma única saída USB, ou seja, conectando o primeiro periférico à saída USB da placa-mãe e conectando os demais a ele. O USB em sua versão 2.0 possui uma taxa de transferência que pode chegar a 480 Mbps.

## USB (Universal Serial Bus)





#### IEEE 1394 (FireWire)

O FireWire (também conhecido como i.Link, IEEE 1394 ou High Performance Serial Bus/HPSB) é uma interface serial para computadores pessoais e aparelhos digitais de áudio e vídeo que oferece comunicações de alta velocidade e serviços de dados em tempo real.

O FireWire pode ser considerado uma tecnologia sucessora da quase obsoleta interface paralela SCSI. Podemos conectar até 63 periféricos em fila a uma única saída FireWire, ou seja, conectando o primeiro periférico à saída FireWire da placa-mãe e conectando os demais a ele.

O FireWire é uma tecnologia de entrada/saída de dados em alta velocidade para conexão de dispositivos digitais, desde camcorders e câmeras digitais, até computadores portáteis e desktops. Amplamente adotado por fabricantes de periféricos digitais como Sony, Canon, JVC e Kodak, o FireWire tornou-se um padrão estabelecido na indústria tanto por consumidores como por profissionais.

#### IEEE 1394 (FireWire)

Desde 1995, um grande número de camcorders digitais modernos inclui essa ligação, assim como os computadores Macintosh e PCs da Sony, para uso profissional ou pessoal de áudio/vídeo. O FireWire também foi usado no iPod da Apple durante algum tempo, o que permitia que as novas músicas pudessem ser carregadas em apenas alguns segundos, recarregando simultaneamente a bateria com a utilização de um único cabo.

Os modelos mais recentes, porém, como o iPod nano e o novo iPod de quinta geração, já não utilizam uma conexão FireWire (apenas USB 2.0).

## Pedido de interrupção (IRQ)

Nos micros PC, existe um recurso chamado de pedido de interrupção. A função dos pedidos de interrupção é permitir que os vários dispositivos do micro façam solicitações ao processador.

Existem 16 canais de interrupção, chamados de IRQ ("interrupt request", ou "pedido de interrupção"), que são como cordas que um dispositivo pode puxar para dizer que tem algo para o processador. Quando solicitado, o processador para tudo o que estiver fazendo para dar atenção ao periférico que está chamando, continuando seu trabalho após atendê-lo.

Dois dispositivos não podem compartilhar a mesma interrupção, caso contrário teríamos um conflito de hardware. Isso acontece porque, nesse caso, o processador não saberá qual dispositivo o está chamando, causando os mais diversos tipos de mau funcionamento dos dispositivos envolvidos.

# Visão geral do sistema de armazenamento e hierarquia de memória

Conhecer os dispositivos de armazenamento por meio do conceito e dos tipos de memórias utilizadas no computador.

# Subsistemas de memória

Memória é o dispositivo físico capaz de armazenar algum tipo de conteúdo para uso no computador. Esse conteúdo pode ser:

- Instruções
- Dados
- Resultados de processamento intermediário
- Resultados de processamento final (informação)

Apesar de parecer simples como conceito, a memória de um computador exibe, talvez, a mais vasta gama de: tipos, tecnologia, organização, rendimento e custos, dentre todas as especificidades de um sistema de computação.

## Memória interna e memória externa

Nenhuma tecnologia pode ser considerada ótima em termos da satisfação dos requisitos de memória de um sistema de computação. Como consequência, o sistema de computação típico está equipado com uma hierarquia de subsistemas de memória, alguns internos ao sistema (diretamente acessíveis ao processador) e outros externos (acessíveis ao processador por meio de um módulo de E/S).

A memória interna (também chamada de primária) é responsável pelo armazenamento temporário de dados utilizados no processamento. É essencial para que ocorra o processamento. Corresponde aos registradores, cache e memória principal.

Já a memória externa (secundária) armazena dados não voláteis em dispositivos periféricos acessíveis por meio de controladores de E/S. Alguns exemplos: disco rígido, CD-rom etc.

#### Memória interna e memória externa

Na base da pirâmide que representa a hierarquia (apresentada em item posterior) de memória em um sistema de computação encontra-se um tipo de memória com maior capacidade de armazenamento do que os outros tipos já descritos, menor custo por byte armazenado e com tempos de acesso também superiores aos outros tipos.

Essa memória, denominada memória secundária, memória auxiliar ou memória de massa, tem por objetivo garantir um armazenamento mais permanente a toda a estrutura de dados e programas do usuário, razão pela qual deve naturalmente possuir maior capacidade que a memória principal.

#### Memória interna e memória externa

A memória secundária de um sistema de computação pode ser constituída por diferentes tipos de dispositivos, alguns diretamente ligados ao sistema para acesso imediato (discos rígidos, por exemplo), e outros que podem ser conectados quando desejado (como os disquetes, fitas de armazenamento, CD-ROM etc.), cuja informação armazenada se torna diretamente conectada e disponível para o específico disquete ou fita que estiver inserido no elemento de leitura/escrita (drive ou acionador), enquanto os demais ficam disponíveis (off-line) para acesso manual pelo usuário.

# Capacidades

Por causa da simplicidade de projeto e construção, acarretando a redução de seu custo e maior confiabilidade, os circuitos eletrônicos que formam os computadores digitais atuais são capazes de distinguir apenas dois níveis de tensão – computadores digitais binários.

Esses sinais elétricos são tensões que assumem dois diferentes valores: um valor positivo (hoje, nos PCs, cerca de +3V) para representar o valor binário 1 e um valor aproximado a oV para representar o valor binário o.

Na realidade, esses valores não são absolutos, e sim faixas de valores com uma margem de tolerância (entre + 2,5 e + 3,5 V para 1 e entre o e + 0,5 V para 0). A lógica que permite aos computadores operar baseados nesses dois valores é chamada Álgebra de Boole, em homenagem ao matemático inglês George Boole (1815–1864).

### Unidades de transferência e armazenamento

As unidades de transferência de dados são medidas da seguinte forma:

- Em memórias primárias: é o número de bits que podem ser lidos/escritos de cada vez (palavra).
- Em memórias secundárias: é a quantidade de blocos de dados lidos/escritos por vez.

Um termo comumente empregado é a palavra, que representa a unidade de informação da CPU e da memória principal.

O conceito mais usado define palavra como sendo a capacidade de manipulação de bits do núcleo do computador (CPU e MP).

Pressupõe-se aqui que todos os elementos do núcleo do computador (que incluem o tamanho da ULA, registradores da CPU e o barramento de dados) tenham a mesma largura (processem simultaneamente o mesmo número de bits), o que nem sempre acontece. Muitas vezes encontram-se computadores cujo tamanho da ULA e dos registradores não é o mesmo dos barramentos.

Dessa forma, encontram-se especificações de computadores de 64 bits, mesmo quando seu barramento de dados é de 32 bits.

Concluindo, deve-se analisar caso a caso, porque a simples menção ao tamanho da palavra não é uma terminologia que permita definir de forma conclusiva a arquitetura do computador.

Quanto ao armazenamento, os dados apresentam outra forma. Em geral, o termo "célula" é usado para definir a unidade de armazenamento, e o termo "palavra" para definir a unidade de transferência e processamento, significando na prática quantos bits o computador movimenta e processa em cada operação. Não confundir: célula não é sinônimo de palavra, embora em algumas máquinas a palavra seja igual à célula.

# Hierarquia de memórias

As restrições no projeto de memória de um computador podem ser sumarizadas em três questões: Qual sua capacidade de armazenamento? Qual sua velocidade de acesso? Qual seu preço por bit armazenado?

Existe uma variedade de tecnologias usadas para implementar os sistemas de memória. Nesse espectro de tecnologias, podem ser consideradas as seguintes relações:

- Tempo de acesso mais curto, maior o custo por bit.
- Maior capacidade, menor o custo por bit.
- Maior capacidade, maior o tempo de acesso.

Um projetista gostaria de usar as tecnologias de memória que disponibilizam uma elevada capacidade de memória, tanto porque a capacidade é necessária como pelo baixo custo por bit. Contudo, para atingir os requisitos de rendimento, necessita usar memórias dispendiosas, de relativamente baixa capacidade de armazenamento e baixo tempo de acesso.

### Hierarquia de memórias

A solução desse dilema não está dependente de um simples componente de memória ou da tecnologia, mas do emprego de uma hierarquia de memória.

- Redução do custo/bit
- Aumento de capacidade
- Aumento do tempo de acesso
- Redução da frequência dos acessos à memória pelo processador

Para que um dado localizado na memória virtual possa ser utilizado no processamento é preciso que passe por cada nível da hierarquia. A razão principal é que o custo por bit de uma tecnologia de memória é geralmente proporcional à sua velocidade. Memórias rápidas, como SRAM, tendem a ter alto custo por bit, tornando proibitivamente caro construir a memória de um computador totalmente com esses dispositivos.

Na hierarquia, os níveis mais próximos ao processador, como a cache, contêm uma quantidade relativamente pequena de memória, que é implementada numa tecnologia de memória rápida, de modo a fornecer um baixo tempo de acesso.

O objetivo de uso da hierarquia de memória é manter os dados que serão mais referenciados por um programa nos níveis superiores, de modo que a maioria das solicitações à memória possa ser tratada no(s) nível(is) superior(es).

Por essa razão, o nível mais alto (cache) é geralmente implementado utilizando SRAM, com o emprego de blocos pequenos (normalmente entre 32 e 128 bytes) controlados por hardware. Já a memória principal é geralmente construída com DRAM, com o tamanho de blocos grandes (vários kilobytes) e controle pelo sistema operacional.

E, finalmente, a memória virtual é usualmente implementada utilizando discos que contém todos os dados do sistema de memória.

Quando uma solicitação de memória é enviada para a hierarquia, o nível mais alto é verificado para ver se contém o endereço.

Se for assim, a solicitação é completada. Caso contrário, o próximo nível mais baixo é verificado, com o processo sendo repetido até que o dado seja encontrado ou o nível mais baixo da hierarquia seja atingido, no qual se tem a garantia de que o dado está contido.

Com isso, um bloco de posições sequenciais contendo o endereço referido é copiado do primeiro nível que contém aquele endereço para todos os níveis acima dele.

## As razões para a utilização de blocos são:

- Muitas tecnologias de armazenamento (DRAM modo paginado, discos rígidos) permitem que várias palavras sequenciais de dados sejam lidas e escritas em menos tempo que um número igual de palavras localizadas aleatoriamente, tornando mais rápido transferir um bloco de vários bytes de dados que buscar cada byte individualmente.
- A maioria dos programas apresenta localidade de referência: as referências à memória que ocorrem próximas no tempo tendem a ter endereços que são próximos uns aos outros, fazendo com que seja provável que outros endereços dentro do bloco sejam referenciados em breve, após um primeiro acesso a um endereço no bloco.

## Referências

- STALLINGS, Willian. Arquitetura e organização de computadores. 5. ed. Prentice Hall. São Paulo, 2006.
- TANENBAUM. Andrew S. Organização estruturada de computadores. 5. ed. Rio de Janeiro: LTC, 2007.
- MACHADO, Francis B.; MAIA, Luiz P. Arquitetura de sistemas operacionais. 4. ed.
  Rio de Janeiro: LTC, 2007.
- WEBER, Raul Fernando. Arquitetura de computadores pessoais. 2. ed. Porto Alegre: Sagra Luzzatto, 2003.
- Fundamentos de arquitetura de computadores. 3. ed. Porto Alegre: Sagra Luzzatto, 2004.