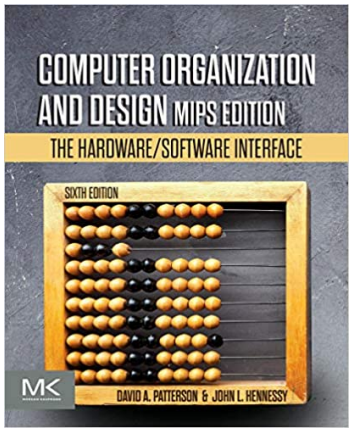
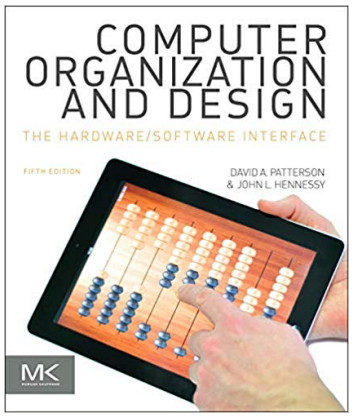
Aula 02 – Arquitetura de Von Neumann

Prof. Dr. Clodoaldo A. de Moraes Lima

Material baseado no livro "Patterson, David A., Hennessy, J. L. - Computer Organization And Design: The Hardware/Software Interface"





Evoluções nos processadores para obter um maior desempenho

Aumento do clock (overclocking)

permite executar mais instruções por segundo, pois é o clock que determina o ritmo de execução das instruções e transferências dos dados. Entretanto, o aumento de clock gera mais calor e maior consumo de energia, além de outros fatores como o atraso na comunicação devido à resistência dos componentes e a capacidade de propagação de ondas eletromagnéticas sobre a superfície dos componentes.

Aumento no número de bits da CPU

aumentar o número de bits de dados num processador permite aumentar a capacidade de armazenamento, transporte e processamento de dados na CPU. Na maioria dos processadores atuais, tais circuitos operam com 64 bits de cada vez.

Evoluções nos processadores para obter um maior desempenho

Aumento na capacidade de endereçamento

aumentar a capacidade de endereçamento de memória não está exatamente relacionado com o desempenho, e sim, com a capacidade de manipular grandes quantidades de dados, aumentando o volume de dados que pode ser processado.

Utilização de memória cache

como o desempenho da memória principal (RAM) é bem inferior ao desempenho da CPU, foi necessário criar uma hierarquia de memória com uma memória cache implementada normalmente na própria CPU. Essa memória armazena uma cópia das instruções e dados recentemente usados e próximos aos recentemente usados. Desta forma, quando a CPU precisar acessar os dados verifica primeiro se a cópia que está na cache contém os dados necessários, minimizando o acesso à memória RAM.

Evoluções nos processadores para obter um maior desempenho

Utilização de pipelines

a técnica de pipelines permite que várias instruções sejam sobrepostas na execução dentro do processador. Uma instrução é decomposta em várias e distintas tarefas e cada uma delas é executada por diferentes partes do hardware simultaneamente. Isso permite que, enquanto uma instrução está sendo buscada na memória, outra instrução esteja sendo decodificada e outra ou outras estejam em execução, no mesmo ciclo de clock.

Utilização de arquitetura escalar e superescalar

no processamento de dados escalares, são necessários vários ciclos para realizar as operações sobre os dados. Os processadores escalares operam sobre um dado de cada vez e se for preciso fazer a mesma operação em mil elementos a CPU precisa repetir a operação mil vezes. Na arquitetura superescalar, vários pipelines são construídos pela replicação de recursos da execução, possibilitando a execução simultânea das instruções em pipelines paralelos, reduzindo o número de ciclos necessários

Evoluções nos processadores para obter um maior desempenho

Utilização de arquitetura vetorial

possui uma grande capacidade de executar cálculos simultâneos sobre um conjunto de dados. No interior desse tipo de processador há dezenas, centenas ou milhares de unidades especificamente dedicadas a cálculos, capazes de operar simultaneamente. Desta forma, quando um programa efetua certa operação sobre todos os dois mil elementos de um vetor e o processador dispõe de, por exemplo, duzentas unidades capazes de efetuar cálculos, as duas mil operações são distribuídas pelas duzentas unidades internas e todo o trabalho é realizado em um centésimo do tempo gasto para efetuar a mesma operação usando uma CPU convencional.

Evoluções nos processadores para obter um maior desempenho

Utilização de arquitetura VLIW (Very Long Instruction Word)

tira proveito do paralelismo em nível de instrução, pois executa um grupo de instruções ao mesmo tempo. Um compilador garante que as instruções a serem processadas não tenham dependências entre si, permitindo a execução ao mesmo tempo, sem perda de lógica do processamento. A abordagem VLIW depende dos próprios programas que fornecem todas as decisões em relação às instruções e como elas devem ser executadas simultaneamente. O processador Intel's Itanium IA-64 EPIC usado em servidores é um exemplo do uso de VLIW.

Evoluções nos processadores para obter um maior desempenho

Utilização de Multithreading Simultâneo (SMT)

os bancos de registradores são replicados para que várias instruções possam compartilhar os recursos dos pipelines. Esta tecnologia é encontrada nos processadores Intel com o nome de hyperthreading e permite simular dois processadores, tornando o sistema mais rápido, quando se usa vários programas ao mesmo tempo. Uma CPU com hyperthreading tem o dobro de registradores, mas apenas uma ULA e uma unidade de controle.

Evoluções nos processadores para obter um maior desempenho

Utilização de multicore

é a combinação de dois ou mais processadores num único chip. É também chamado de chip multiprocessador. Cada processador, também chamado de núcleo ou core, possui todos os componentes de um processador convencional, como registradores, ULA e unidade de controle. Além disso, os chips multicore normalmente incluem caches L1 (em alguns modelos também uma L2) privativas para cada núcleo e caches L2 (ou L3 em alguns modelos) compartilhadas.

Incorporação da Unidade de Processamento Gráfico (GPU) na CPU

transforma a CPU numa APU (Acelerated Processing Unit) ou Unidade de Processamento Acelerada, colocando no mesmo chip a CPU e a GPU, aumentando o desempenho e reduzindo o consumo de energia.

Evolução dos computadores

Primeira Geração (1951-1959)

- Os computadores de primeira geração funcionavam por meio de circuitos e válvulas eletrônicas. Possuíam o uso restrito, além de serem imensos e consumirem muita energia.

Segunda Geração (1959-1965)

- Ainda com dimensões muito grandes, os computadores da segunda geração funcionavam por meio de transistores, os quais substituíram as válvulas que eram maiores e mais lentas.

Evolução dos computadores

Terceira Geração (1965-1975)

- Os computadores da terceira geração funcionavam por circuitos integrados. Esses substituíram os transistores e já apresentavam uma dimensão menor e maior capacidade de processamento.

Quarta Geração (1975-até os dias atuais)

- Com o desenvolvimento da tecnologia da informação, os computadores diminuem de tamanho, aumentaram a velocidade e capacidade de processamento de dados. São incluídos os microprocessadores com gasto cada vez menores de energia.

Evolução dos computadores

Computadores Eletrônicos

- Grande avanço em relação aos seus similares eletromecânicos
- Duas limitações
  - Baixa capacidade de memória
  - Longo tempo de programação

- ENIAC - exigia dias de trabalho, uma vez que várias modifi cações eram necessárias no painel de controle
- Mark I - fácil de reprogramar (troca de fita) porém velocidade de leitura de instruções de unidades mecânicas não era adequada à velocidade de processamento dos computadores eletrônicos

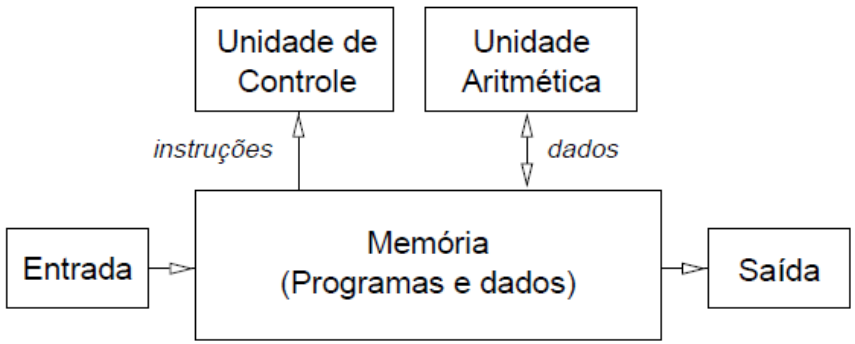
- Marco importante - concepção do conceito de programa armazenado associado ao projeto EDVAC

EDVAC

- Sucessor do ENIAC
- 1K de palavras de 44 bits na memória principal
- 20 K de palavras na memória secundária
- Possuía 4000 válvulas, 1000 diodos
- Velocidade do relógio era de 1MHz
- Projeto concluído em 1952

Programa armazenado

- Atribuído exclusivamente a Von Neuman $\rightarrow$ 101 páginas sobre o projeto EDVAC
  - First Draft of a Report on the EDVAC em junho de 1945, onde o conceito foi descrito pela primeira vez
- Justo seria atribuir a toda equipe do projeto EDVAC, incluindo Mauchly e Eckert (Moore School), von Neumann (Institute for Advanced Study, Princeton), Herman H. Goldstine (inicialmente estava ligado a Marinha) e Arthur W. Burks (filósofo com inclinações matemáticas da University of Michingan)



Programa Armazenado

- Von Neuman, Herman e Arthur tinham uma visão matemática e abstrata, visavam primordialmente á divulgação de resultados
- Mauchly e Eckert - gostariam de obter frutos transformando o EDVAC em produto comercial

Algumas sugestões sobre o conceito de programa armazenado foram apresentadas durante a escola de verão do ENIAC

- Manchester Baby Machine, da Universidade de Manchester (Inglaterra), de junho de 1948, por M. Newman e F. C. Williams
- EDSAC (Electronic Delay Storage Automatic Calculator), da Universidade de Cambridge (Inglaterra), de maio de 1949, por Maurice Wilkes
- BINAC (Binary Automatic Computer), da Eckert-Mauchly Computer Corporation (EMCC) construído sob encomenda da Northrop Aircraft Corporation, operacional em Setembro de 1949;

Programa Armazenado

- UNIVAC (Universal Automatic Computer), da Remington Rand Co. (que incorporou a EMCC), com a primeira unidade operacional em março de 1951;
- Whirlwind, do MIT por Jay Forrester, projetado como o primeiro computador para aplicações tempo-real. O Whirlwind tornou-se a base para projetos de minicomputadores;
- IBM 701, voltado para aplicações científicas (ex-Defense Calculator), foi o primeiro computador eletrônico da IBM (dezembro 1952);
- IBM 650 Magnetic Drum Computer, apresentado como o modelo barato da IBM (US$200K), anunciado em 1953. Essa máquina foi á base para o modelo IBM 1401 (transistorizado, anúncio em outubro de 1959, entrega no início de 1960 a um custo de US$150K).

Manchester Baby Machine

- Primeiro computador de programa armazenado
- Utilizava vários tubos de raios catódicos (CRT) como memória, memória principal de 32 palavras de 32 bits.
- A programação era realizada bit-a-bit por um teclado, com a leitura de resultados também bit-a-bit (de um CRT).
- Tornou-se a base para um computador comercial inglês, o Ferranti Mark I (fevereiro de 1951).
- Posteriormente, Turing juntou-se a essa equipe e desenvolveu uma forma primitiva de linguagem Assembly para essa máquina.

EDSAC

- Utilizava tecnologia de memória por linha de atraso em mercúrio, desenvolvida por William Shockley (Bell Labs) - com 16 tanques de mercúrio, a capacidade era de 256 palavras de 35 bits, ou 512 palavras de 17 bits.
- Foi o primeiro computador de memória armazenado de uso prático. Operava com uma taxa de relógio de 500 KHz.
- A entrada e saída de dados ocorria através de fita de papel. O primeiro programa armazenado foi imprimir quadrados dos primeiros números inteiros.

BINAC

- O BINAC foi projetado como um primeiro passo em direção aos computadores de bordo.
- Era um sistema com processadores duais (redundantes), com 700 válvulas cada e memória de 512 palavras de 31 bits.

UNIVAC

- tinha uma memória de 1000 palavras de 12 dígitos, com uma memória secundária de fi
  tas magnéticas com capacidade de 128 caracteres por polegada.
- A primeira unidade foi desenvolvida sob encomenda do Census Bureau norte-americano.
- Dominou o mercado de computadores na primeira metade dos anos 1950.

Whirlwind

- Foi desenvolvido por Forrester e equipe para o US Navy's Office of Research and Inventions.
- A origem do projeto estava baseada em um simulador de voo universal, com velocidade de operação adequada a aplicaçõoes de tempo real (500K adições ou 50K multiplicações por segundo).
- O projeto foi iniciado em setembro de 1943, tornando-se o computador operacional em 1951.
- Introduziu a tecnologia de memória a núcleos de ferrite (tempo de acesso de 9μs), em 1953, em substituição da memória CRT original de 2048 palavras de 16 bits.
- O custo total do projeto superou vários milhões de dólares.

IBM 701

- O IBM 701 estava disponível com memórias CRT de 2048 ou 4096 palavras de 36 bits.
- O IBM 702 Electronic Data Processing Machine estava voltado para aplicações comerciais , tendo sido anunciado em setembro de 1953 e entregue no início de 1955.

IAS

- Um ano após o término da segunda guerra mundial, Neumann escreveu um relatório sobre o Computador IAS, que posteriormente cunharia o termo Arquitetura de Von Neumann.
- Von Neumann iniciou gestões para a construção de outro computador que seria utilizado para aplicações científi
  cas em geral
- Von Neuman convenceu a direção do Instituto de Estudos Avançados de Princeton abrigar o projeto
- A RCA acabava de estabelecer um laboratório de pesquisa na Universidade de Princeton
- RCA iniciou a construção de tubos iconoscópicos semelhante aos tubos de televisão
- IAS recebeu apoio do Exercito e da Marinha americana

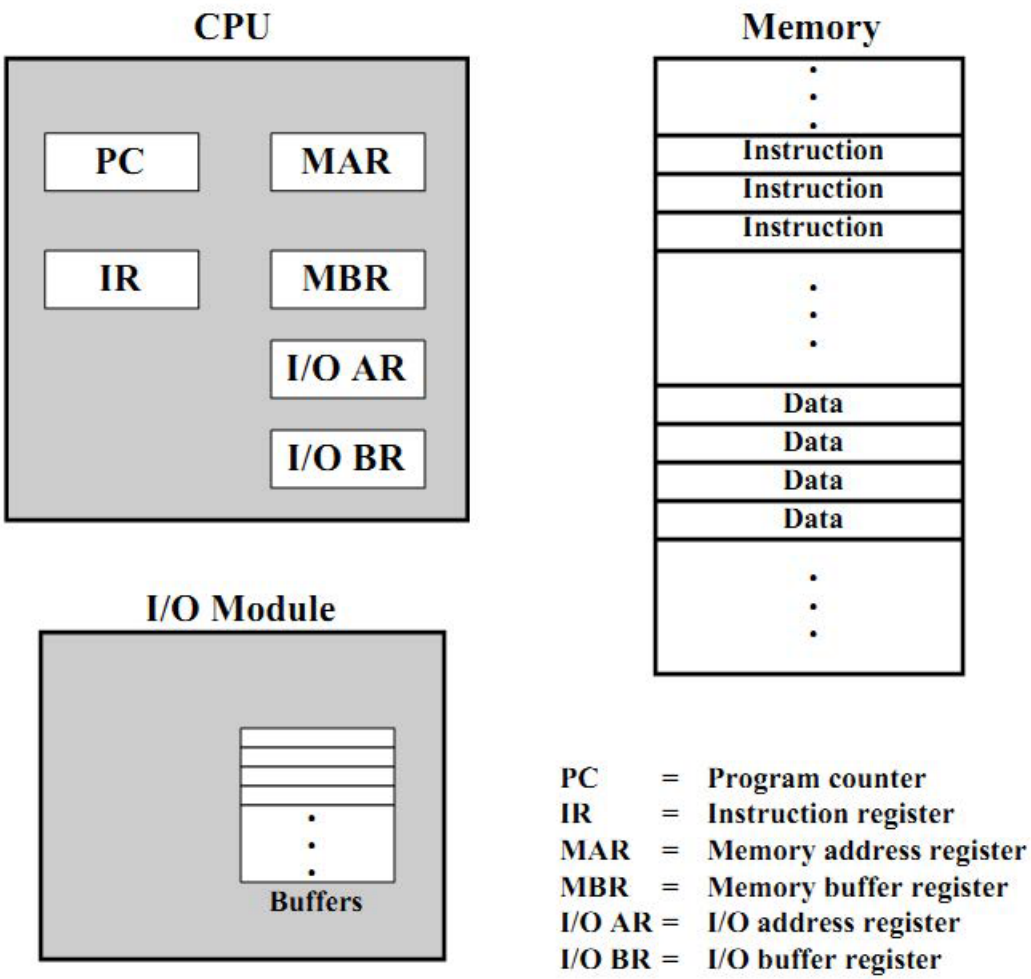
IAS

- O projeto lógico é apresentado na primeira parte escrita por Burks, Goldstine e von Neumann, intitulada "Preliminary of the Logical Design of an Electronic Computing Instrument".
- As operações aritméticas são discutidas em grande detalhe, incluindo problemas de arredondamento.
- Tendo em vista as características da memória, as operaçõoes sobre os 40 bits seriam executadas em paralelo
- Há uma demonstração de que a operação de soma de dois números de 40 bits produziria, em média, cinco "vai um".
- Discussão completa de mecanismos de entrada e saída

IAS

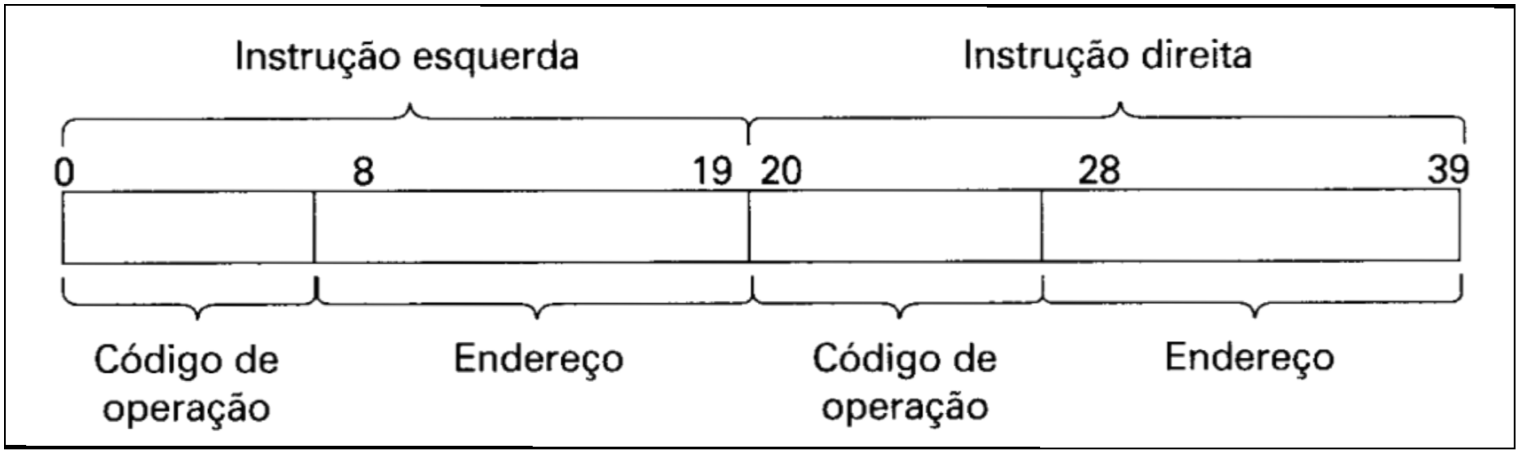
- Observa a necessidade de realoção de instruções para que possam ocupar quaisquer parte da memória
- Problema de dar inicio no sistema a partir de um dispositivo de entrada.
- Discutida a utilização de redundância para a detecção de falhas nas unidade lógicas e outros dispositivos

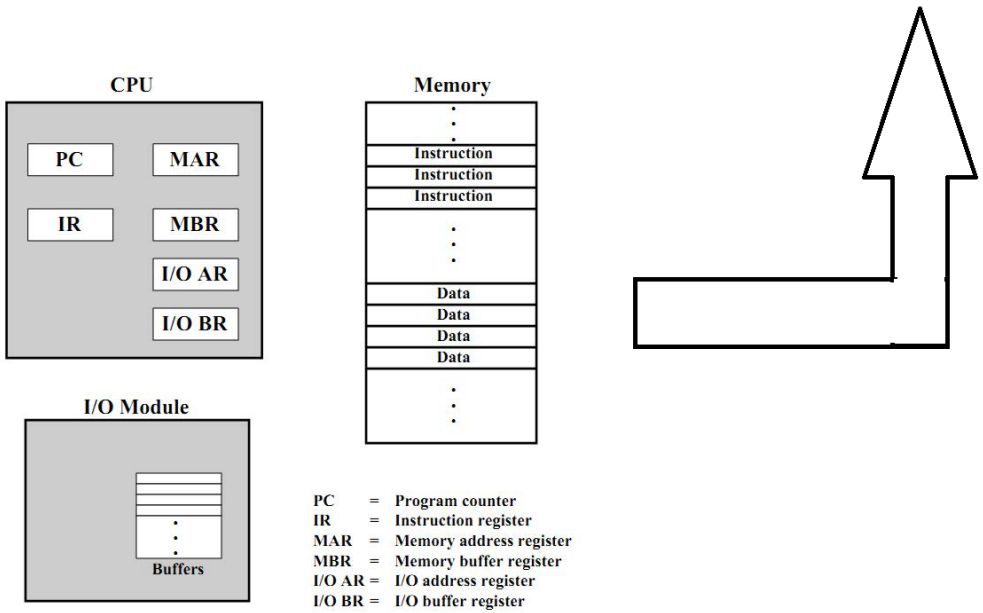
Registradores Especiais



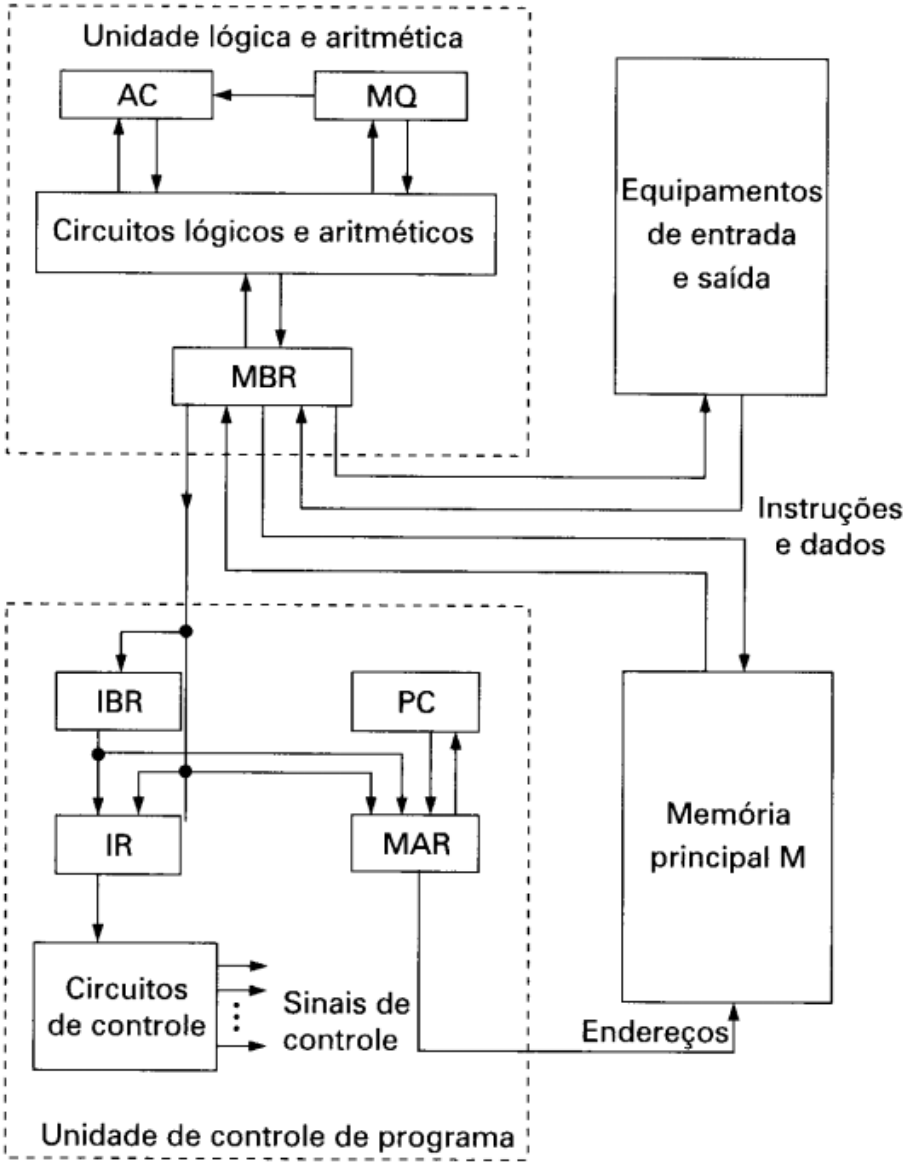
- MAR : especifica o endereço de memória da próxima instrução.
- MBR: contém o valor a ser gravado na memória ou recebido da memória.
- I/O AR: registrador de endereçamento de E/S.
- I/O BR: usado na troca de dados entre módulos de E/S e a CPU.

Registradores Especiais



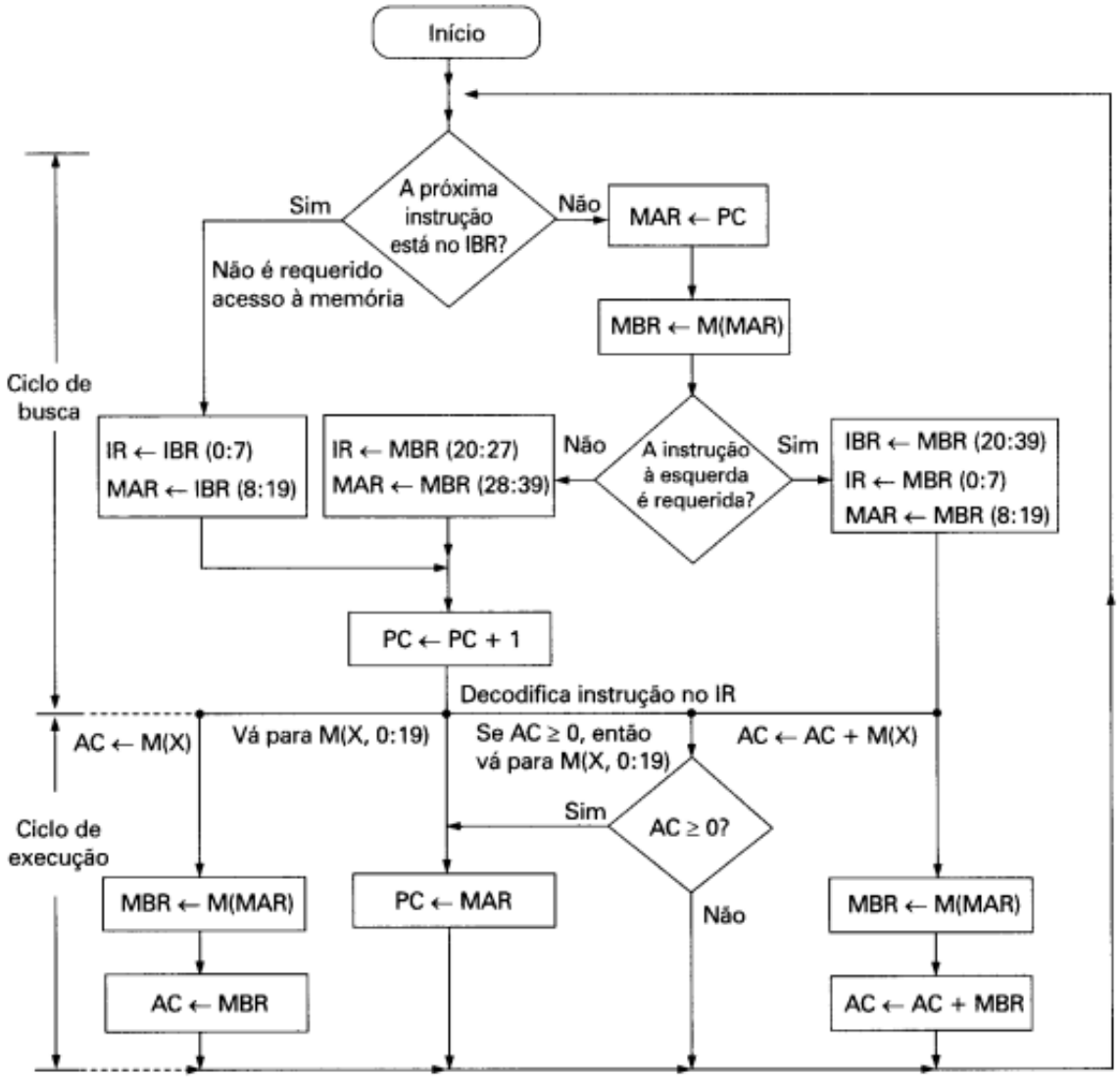


Registradores Especiais



Estrutura detalhada do IAS.

Registradores Especiais



M(X) = conteúdo da posição de memória cujo endereço é X

(X : Y) = bits X a Y

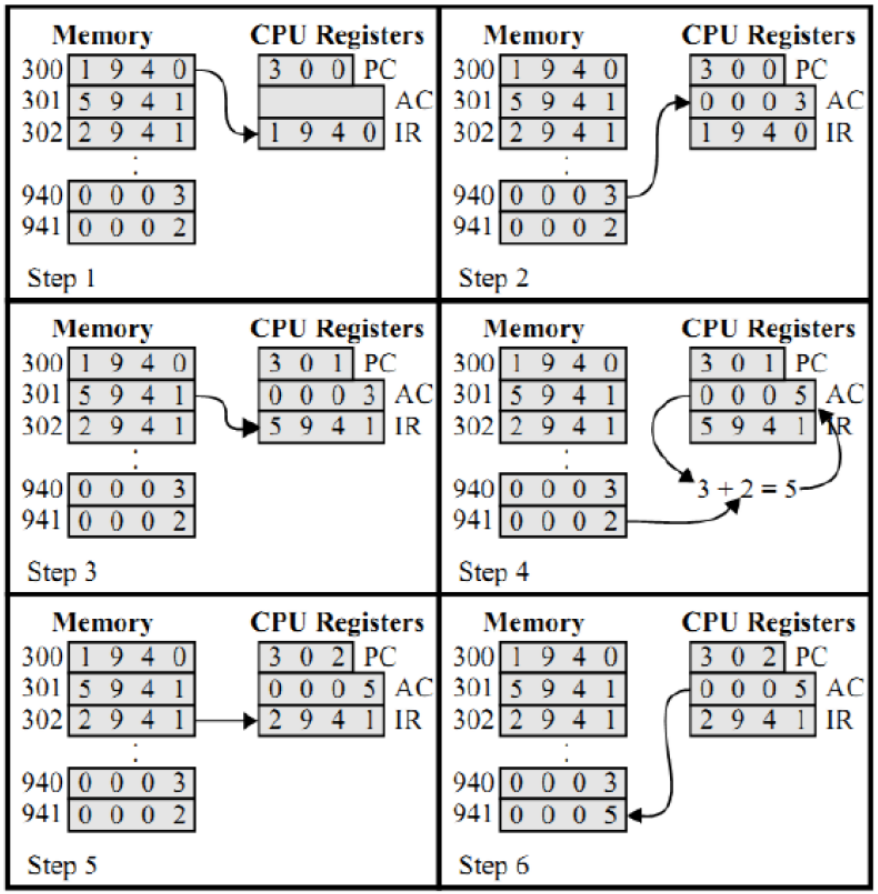
Registradores Especiais

Lista de Códigos de Operação:

0001: AC ← M(X)

0010: M(X) ← AC

0101: AC ← AC + M(X)



Tipos de Arquitetura

Arquitetura de Von Neumann

- Conceito de programa armazenado;
  - Dados e instruções armazenados em uma única memória de leitura e escrita.
- Endereçamento da memória por posição e não pelo tipo;
- Execução sequencial de instruções; e
- Unico caminho entre memória e CPU.

Arquitetura de Harvard

- Variação da arquitetura de Von Neumann.
- Barramentos separados para instruções e dados.
- Termo originado dos computadores Mark I a Mark IV
- Memórias separadas para dados e instruções

Outras Arquiteturas

Máquinas Paralelas

- Várias unidades de processamento executando programas de forma cooperativa.
- Podem ser controladas de forma centralizada ou não

Exemplos de arquiteturas não Von Neumann

- Máquinas de Fluxo de Dados
  - Realizam operações de acordo com a disponibilidade dos dados envolvidos
  - A Memória armazena um conjunto de instruções no formato conhecido como "tokens":operação, operandos, destino
  - Não existe controle da memória a ser lida
  - A execução das instruções (tokens) ocorre quando os operandos estiverem disponíveis

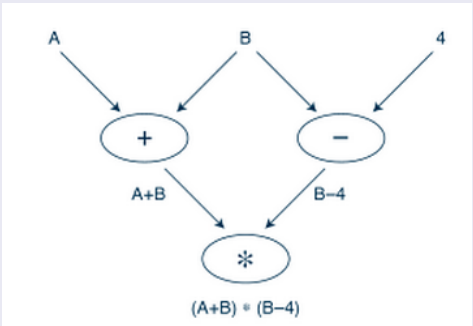


Figura: Grafo de fluxo de dados calculando N = (A + B) * (B - 4)

Outras Arquiteturas

Máquinas de Fluxo de Dados

- A figura acima é um exemplo de uma arquitetura de fuxo de dados estática na qual as unidades fuem através do grafo de forma semelhante aos estagios de pipeline.
- Na arquitetura de fluxo de dados dinâmica, as unidades são etiquedas com informação de contexto e são armazenadas em uma memória.
- Durante cada ciclo de relógio,a memória é pesquisada em busca de um conjunto completo de unidades de entrada para acender um nó.

Outras Arquiteturas

Máquinas de Fluxo de Dados

- Os nó acendem somente quando encontram um conjunto completo de unidades de entrada dentro do mesmo contexto.
- Programa para máquinas de fluxo de dados devem ser escritos em linguagens que são especificamente projetadas para este tipo de arquitetura; esta incluem VAL, Id, SISAL e LUCID.
- A compilação de um programa de fuxo de dados resulta em um grafo de fluxo de dados muito semelhante a figura acima

Outras Arquiteturas

Computadores de array sistólicos

- Derivam seu nome de uma analogia sobre como o sangue flui de forma ritmada através do coração biológico
- Eles são uma rede de elementos de processamento que processam dados de forma ritmada por meio de sua circulação pelo sistema
- Incorporam grandes arrays de processadores simples que usam pipelines de vetor para fuxo de dados (veja figura)
- Desde a sua introdução na década de 1970, eles tem tido um impacto significativo na computação de próposito especial.
- Um processador de array sistólico bem conhecido é o CMUiWrap, que foi fabricado pela Intel em 1990.
- Este sistema consiste de um array linear de processadores conectados por um barramento de dados bidimensional

Outras Arquiteturas

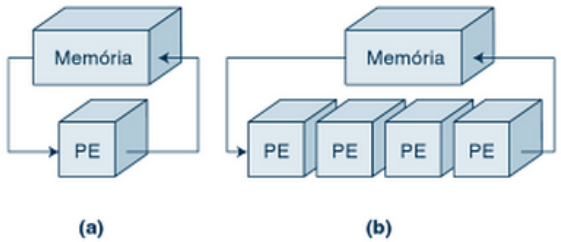
Computadores de array sistólicos

- Arrays sistólicos adotam um alto grau de paralelismo (por meio do pipeline) e podem sustentar uma vazão muito alta.
- As conexões são geralmente curtas e o projeto é simples e, portanto, altamente escalável
- Tendem a ser robustos, altamente compactos, eficientes e baratos para produzir
- Por outro lado, eles são altamente especializados e, portanto, inflexíveis quanto aos tipos e aos tamanhos dos problemas que podem resolver.
- Um exemplo de uso de arrays sistólicos pode ser encontrado na avaliação polinomial. Para avaliar o polinômio $y = a_0 + a_1x + a_2x^2 + \cdots + a_kx^k$, podemos usar a regra de Horner: $y = (((a_kx + a_{k-1}) * x + a_{k-2}) * x + a_{k-3}) * x + \cdots a_1) * x + a_0$

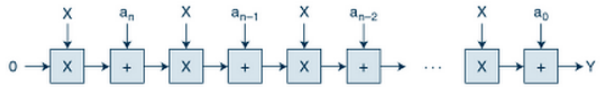
Outras Arquiteturas

Computadores de array sistólicos

- Um array sistólico linear, no qual os processadores são dispostos em pares, pode ser usado para avaliar um polinômio usando a Regra de Horner, como mostrado figura abaixo
- Arrays sistólicos são geralmente usados para tarefas repetitivas, incluindo transformadas de Fourier, processamento de imagens, compressão de dados, problemas de menor caminho, ordenação, processamento de sinais, etc.
- São adequados para problemas computacionais que permitem uma solução paralela usando um grande número de elementos simples de processamento.



a) Um elemento simples de processamento (PE).
b) Um processador de array sistólico.



Usando um array sistólico para avaliar um polinômio.

Outras Arquiteturas

Computação Fotônica

- Todas as arquiteturas clássicas de computadores apresentadas até aqui têm um aspecto em comum: todas usam lógica booleana.
- A lei de Moore, que declara que o número de transitores em um único chip dobra a cada 18 meses, não pode se aplicar para sempre
- As leis da física sugerem que eventualmente os transitores vão se tornar tão finos que a distâncias entre eles vão permitir que elétrons saltem de um para outro, causando curtos circuitos fatais.
- Uma possível resposta é a computação ótica ou fotônica. Em vez de usar elétrons para realizar a lógica em um computador, computadores óticos usam fotons de luz de laser.

Outras Arquiteturas

Computação Fotônica

- A velocidade da luz em circuitos fotônicos pode se aproximar da velocidade da luz no vácuo, com a vantagem adicional de não ter dissipação de calor
- O fato de que fachos de luz podem trafegar em paralelo pode sugerir um aumento adicional na velocidade e na performance
- Muitas pessoas acreditam que a computação ótica será reservada apenas para aplicações de próposito especial.

Outras Arquiteturas

Computação Quântica

- Enquanto computadores clássicos usam bits que são ligados ou desligados, computadores quânticos usam quantum bits (qubits) que podem estar em diversos estados simultaneamente.
- Da física, sabemos que um campo magnético, um elétron pode estar em dois estados possíveis: o giro pode estar alinhado com o campo ou oposto ao campo.
- Quando medimos este giro, vemos que o elétron está em um desse dois estados.
- Entretanto, é possível que a partícula esteja em uma superposição de dois estados, com ambos existingo simultaneamente.
- Se temos um registrados de três bits formando por qubits, este registrado pode conter qualquer um dos números de 0 a 7 simultaneamente, por que cada qubit pode estar em uma superposição de estados.

Outras Arquiteturas

Computação Quântica

- Para obter um único valor de saída, temos que medir o qubit
- Portanto, o processamento com registradores de 3 qubits pode realizar cálculos usando simultaneamente todos os o valores possíveis, trabalhando com oito cálculos ao mesmo tempo, resultado e paralelismo quanticos.
- Em teória, um computador quântico poderia realizar inúmeras operações em paralelo usando uma única CPU.
- Além de serem cerca de um bilhão de vezes mais rápido do que os seus parentes de silício, computadores quânticos podem, teoricamente, operar sem energia.
- Computadores quânticos podem realizar as tarefas cotidianas feitas por máquinas clássicas, mas eles mostram a sua superioridade somente em aplicações que exploram o paralelismo quântico.

Outras Arquiteturas

Computação Biológica

- Usa componentes de organismos vivos em vez daqueles de silício ionorgânico.
- Um projeto assim é o *leech-ulator*, um computador criado por cientistas americanos que é feito de sanguessugas (leeches).
- Um outro exemplo é a computação DNA, que usa DNA como software e enzimas como Hardware.
- O DNA pode ser replicado e programado para realizar tarefas maciçamente paralelas, das quais uma das primeiras primeiras foi o problema do caixeiro-viajante, sendo o paralelismo limitado somente pelo número de espirais de DNA.

Outras Arquiteturas

Computação Biológica

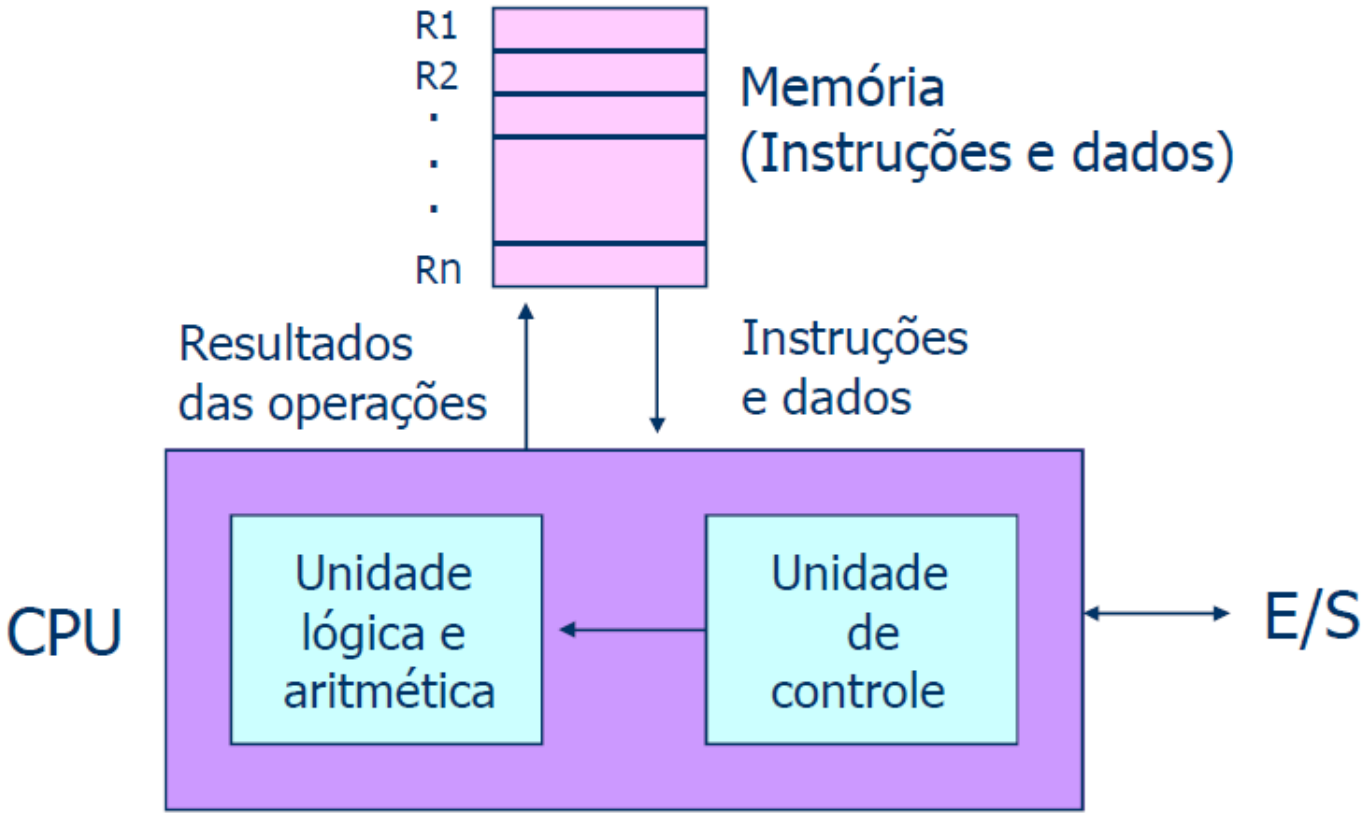
- Computadores DNA (também chamados de computadores moleculares) são basicamente coleções de espirais de DNA especialmente selecionadas para testar todas as soluções de uma só vez e dar como saída a resposta correta.
- Os cientistas também estão experimentando certas bactérias que podem ligar e desligar genes de maneiras previsíveis.
- Pesquisadores já programaram com sucesso a bactéria E. Coli para emitir luz fluorescente vermelha ou verde (ou seja, 0 ou 1).

Outras Arquiteturas

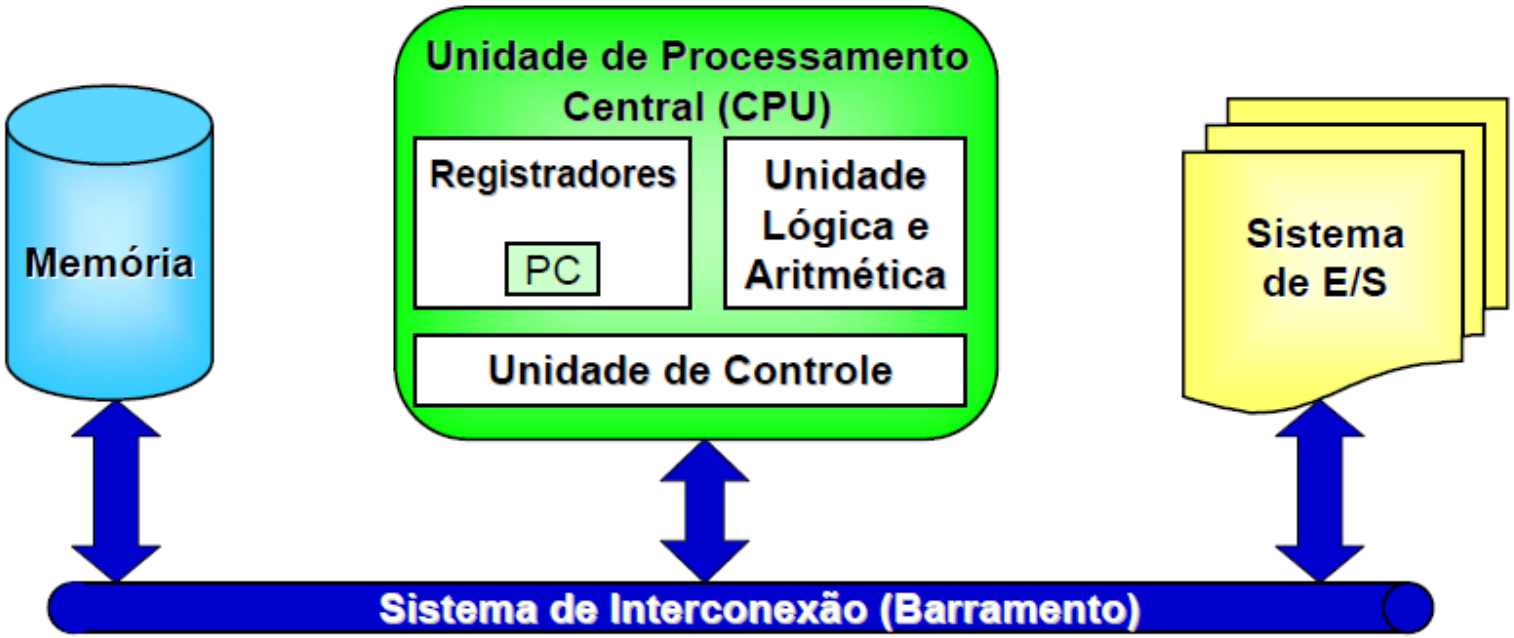
Exemplos de arquiteturas não Von Neumann

- Redes Neurais Artificiais
  - Não executam instruções de um programa
  - Resultados são gerados a partir de respostas a estímulos de entrada

Arquitetura de Von Neumann



Componentes estruturais (computadores atuais)



Arquitetura de Von Neumann

- CPU
  - "Cérebro" do computador
  - Busca, interpreta e executa as instruções
  - Controla os demais componentes
- Memória
  - Armazenamento de dados e instruções
- Sistema de E/S
  - Comunicação externa (ambiente operacional)
- Sistema de interconexão
  - Comunicação interna (entre os componentes)

Componentes do Computador

Processadores

- CPU, Controladores e coprocessadores
- Possuem conjunto de instruções operando sobre instruções e dados organizados em palavras
  - CPU: instruções de propósito geral
  - Coprocessadores: instruções especializadas

Memórias

- 2 subsistemas: memória interna e memória externa

Arquitetura de Von Neumann

Memória

- Armazena instruções
- Armazena dados iniciais e intermediários
- Armazena dados finais
- Byte (Binary Term) - Unidade básica de tratamento de informação
- Cada byte possui 8 bits
- Uma "palavra"(word) é constituída de grupos de 2, 4,6 ou 8 bits.
- Atualmente há palavras de 64, 128 bits (superior)

Memória

BIT

- Binary DigIT
- Sistema biestável:
  - Lâmpada
  - Válvula
  - Armazena dados iniciais e intermediários
- No computador são representados por 0 e 1 (sistema de numeração de base 2)
- Um caracter é representado por 8 bits (byte)
  - 00010110 = A
  - 00010111 = B

Tipos de Memória

Existem basicamente 2 tipos de memórias

ROM (Read Only Memory): É uma memória não volátil utilizada para armazenar Firmwares de placas mãe, DVD player, CD-RW, Placas de Rede, Modens ADSL, etc.

- ROM: Programável por "mascaras"na fábrica do chip.
- PROM: Programável pelo usuário, uma única vez.
- EPROM: "Erasable PROM- Programação pode ser desfeita por UV e refeita pelo usuário.
- EEPROM ou E2PROM: "Eletrically Erasable PROM- Substitui o método UV por um outro processo elétrico que "zera"a memória.
- FLASH: Programação depois de inserida no produto (equipamento).

Tipos de Memória

RAM(Random Access Memory)

- Memória de Acesso Randômico.
- É uma memória volátil ou seja ao desligar o computador o seu conteúdo é perdido.
- Existem vários tipos de memória RAM:
  - SDRAM
  - DDR2 (Double Data Rate), DDR3, DDR4

Tipos de Memória

Memória Cache

- São memórias ultra-rápidas que são usadas em quantidades pequenas, existem 3 níveis L1, L2 e L3.

Gargalo de von Neumann

- Tráfego intenso no barramento do sistema
  - Principal rota de informação: CPU e memória (ponto crítico)
  - Constante fluxo de dados e instruções
- Gera desperdício de tempo (CPU em espera)
- Agrava-se gradativamente pelo aumento do gap de velocidade entre a memória e a CPU