Resumo ARQ2:

Processadores:

Componentes: Processador, Memória, Interconexão, Sistema de entrada e saída e Dispositivos E/S;

Operação básica: busca, decodificação e execução;

Objetivo: Aumentar o número de instruções completadas por segundo;

Operando da ULA é de 64 bits;

ISA: Usado para descrever os atributos de um sistema visto pelo programador;

Intel 80286: Memória virtual segmentada (apenas no modo protegido);

Intel 80386: Introdução da memória virtual segmentada sempre e paginada opcional, ambas no modo protegido;

Memória Virtual:

Memory Management Unit (MMU): Faz o mapeamento utilizando tabelas de tradução de endereços gerenciadas pelo SO;

Tamanho constante (paginação) ou variável (segmentação);

Mapeador transforma endereço virtual em endereço físico;

Se o bloco está na memória, calcula o endereço físico. Caso contrário, busca o bloco da memória secundária;

Segmentada:

Se o endereço está na memória principal e fora do segmento, gera erro;

Se existe espaço suficiente na memória principal, carregar o segmento na “melhor posição” e atualizar o descritor. Caso não tenha, retirar um (ou mais) segmentos, carregar novo segmento e atualizar descritores;

Registrador de segmento:

RPL (bits 0 e 1) - Requested Privilege Level – varia entre 0 (maior privilégio) e 3 (menor privilégio);

TI (bit 2) - Table Indicator – seleciona a tabela global (GDT) ou a local (LDT);

Índice (bits 3 a 15) - Seleciona um dos 8192 descritores – o índice é multiplicado por 8 e somado ao endereço inicial da GDT ou da LDT;

Basic Flat Model:

Aplicativos e SO tem acesso a um espaço de memória contíguo;

No mínimo dois descritores: um para o SO e outro para aplicativos;

Descritores podem apontar para a mesma região de memória;

Protected Flat Model:

Idêntico ao modelo básico, mas segmentos somente apontam para regiões onde existe memória física;

Multi-Segmented Model:

Cada programa recebe seus segmentos de memória, com acesso individual ou compartilhado;

Paginada:

TLB:

- Pequena memória associativa que retém os últimos e mais frequentes endereços de página acessados;
- Uma pequena cache de endereços físicos;
- Diminui a frequência de acessos adicionais à memória RAM para buscar endereços de página na estrutura de mapeamento;

Cache:

Lançada no Intel 80486;

Maior custo, menor tamanho e maior velocidade;

Memória principal só é acessada se conteúdo não se encontra na cache;

A memória associativa armazena parte dos endereços e ajuda a identifcar os blocos presentes;

Possui vários níveis;

Princípio da localidade espacial: É alta a probabilidade de o acesso a um endereço de memória ser seguido do acesso a endereço próximo;

Trabalhar com blocos de dados de tamanho fixo;

Número do bloco da RAM:

- Endereço / tam_da_página;
- Tam_da_pagina = 2^n; n = número de bits para endereçar o dado dentro do bloco;

Ex:

Ram = 16 bytes (4 bits de endereço);

Cache = 8 bytes (2 bytes/bloco)

Line size (tamanho da página) = 2 bytes;

End 13 -> 13/2 = 6 (bloco da RAM);

Mapeamento direto:

RAM = 256 bytes = 2^8;

Cache = 16 bytes = 4 bytes/bloco = 4 blocos -> 2 bits para representar o bloco + 2 para o offset;

Endereço na cache= 8 bits -> 4 da Tag, 2 do bloco e 2 do offset;

Blocos da RAM no mapeamento direto estarão sempre no mesmo bloco da cache;

Bloco Ram = <end_RAM> DIV 2^w

Bloco Cache = Bloco_RAM MOD 2^b

Offset = <end_RAM> MOD 2^w

Mapeamento associativo:

Totalmente associativo: memória com muitos comparadores simultâneos (alto custo);

O bloco da RAM pode estar em qualquer bloco da cache;

Bloco Ram = <end_RAM> DIV 2^w

Bloco Cache = Bloco_RAM MOD 2^s

Offset = <end_RAM> MOD 2^w

Pipeline:

Paralelismo temporal:

- Usa o tempo em que o estágio ficaria sem utilização;
- Dependência de dados e Dependência de controle:
    - O resultado da instrução anterior é usado na instrução atual;
    - Desvios absolutos e condicionais;
    - Faz reiniciar o pipeline;

Superescalaridade:

- Capacidade de finalizar mais de uma instrução no mesmo ciclo de relógio;
- Havendo dependência de dados, não há paralelização;
- Ordem das instruções não é alterada;
- Problema principal: Dependência de dados;

Pentium Pro:

- Execução fora de sequência
- Execução especulativa

- Dependência de dados RAW (leitura depois de escrita) -> Execução fora de sequência;
- Renomeação de registradores -> Resolve falsas dependências WAW e WAR;

MMX – SIMD:

- Operações de movimentação de pontos na tela e tratamento de som utilizam vetores de bytes ou palavras;
- Operações em ponto fixo;
- SSE: Extensão de operações em ponto flutuante;
- SSID: Single Instruction Single Data – Cada instrução atua sobre os dados individuais;
- SIMD: Single Instruction Multiple Data – Modelo de paralelismo onde cada instrução atua sobre um conjunto de dados simultâneo;

Thread:

- SO multitarefa são capazes de executar mais de um programa ao mesmo tempo;
- Programas em execução são chamados de processes, e cada processo pode ter uma ou mais linhas de execução independentes, chamadas de threads;
- Single-thread/single-core:
    - O SO pode alocar apenas um thread por vez;
    - Para trocar de thread, o SO precisa salvar o estado arquitetural do processador e carregar o estado arquitetural da próxima thread;
- TLP (Thread-Level Parallelism):
    - Processador multi-thread: mais de uma thread ativa no mesmo core;
    - Processador multi-core: mais de uma thread ativa em cores diferentes;
- Single-Thread:
    - Uma única linha de execução no pipeline e entre as unidades funcionais;
- Multi-Thread:
    - Em diferentes ciclos de execução, o pipeline e as unidades funcionais podem processar instruções de diferentes threads;
    - Em um mesmo ciclo, todas as instruções são da mesma thread;
- SMT – Simultaneous Multi-Thread:
    - Combina superescalaridade com multi-thread;
    - No mesmo ciclo de execução as diferentes unidades funcionais podem processar instruções de diferentes threads;
    - A implementação da Intel para SMT chama-se Hyper-threading;

Memória:

RAM (Volátil):

- Estática (SRAM) - Cache
- Dinâmica (DRAM) - Refresh

ROM (não volátil)

OPERAÇÃO Memória Dinâmica:

- Endereço -> Controlador de Memória -> Bits mais significativos (Decodificador de linhas (RAS)), bits menos significativos (Decodificador de colunas ((CAS));
- Sinal de linhas é ativo para encontrar a linha onde está o dado. Logo após, o sinal de coluna é ativo para encontrar a coluna do dado. Com isso, o dado é encontrado na memória;
- FAST-PAGE: Leitura das colunas de uma linha;
- SDRAM:
    - Sinal de clock adicionado à interface da memória;
    - Fornece os endereços em sequência;
    - Dados em rajada;

DDR – SDRAM:

- Double Data Rate;
- Dois acessos a cada ciclo de relógio;
- Número de bytes * 2 * Clock;
- Pinos diferentes em cada geração;

O controlador de memória envia os comandos a cada ciclo de relógio;

Latência = Número de ciclo para realizar operação de memória;

Vazão = Capacidade máxima de dados por segundo;

Latência = 1/Vazão;

Rank:

- Conjunto de dispositivos DRAM que operam em paralelo;
- Barramentos são compartilhados entre os ranks;
- Cada rank opera de forma independente;

Bank:

- Conjunto de matrizes de células que operam em paralelo;
- Cada chip DRAM tem vários banks que operam de forma independente;
- Apenas um bank provê os dados por vez;

INTERRUPÇÃO:

Controle direto do processador:

- Polling – Processador verifica ativamente o estado do dispositivo E/S;
- Interrupção - E/S controlada diretamente pelo processador que é informado quando ocorre o evento;

Acesso direto à memória DMA:

- Processador divide o barramento com DMA;
- DMA realiza a operação de leitura enquanto o processador continua executando o programa;

Os tipos não são mutuamente exclusivos;

- Interrupção:
    - Rotina de tratamento de interrupção -> vetor de interrupções;
    - Tabela de 256 posições;
    - A identificação da origem da interrupção é o índice da tabela;
- DMA:
    - 4 canais de DMA independentes;
    - O tamanho máximo do bloco a ser transferido é de 64KB;
    - Transferências exigem dois canais de DMA;
    - O canal 0 deve participar;
    - Atualmente os PCs emulam o funcionamento dos 8 canais de DMA (chipset);
    - Permitem que todos os canais manipulem dados de 16 bits;
- Refresh de Memória;

ENTRADA E SAÍDA:

- Barramentos:
    - Meio de comunicação compartilhado entre vários componentes;
    - Disciplina de acesso;
- Mestre/Escravo:
    - Mestre inicia a transferência e escravo responde;
    - Direção: origem/destino;
    - Multiplexação temporal;
- Síncrono vs. Assíncrono:
    - Ponto de vista de existência de relógio comum;
- Strobe vs. Handshake:
    - Eventos por contagem de tempo ou sinalização específica;
- Barramento ISA:
    - 20 bits de endereçamento;
    - Controle com strobe/handshake;
    - Protocolo do barramento determina a máxima taxa de transferência;
- Hierarquia de barramentos:
    - Ponte:
        - Conecta barramentos de diversas vazões;
        - Atua com árbitro;
        - Atua como mestre no barramento onde o processador não está ligado;
- PCI Express:
    - Comunicação serial;
    - Um barramento para todo o tipo de periférico;

- Rede ponto a ponto;
- Transmissão de pacotes;
- Transmissão de 2 gigabauds (2,5GT/s);
- Conector pode ter até 32 vias;
- Dados, controles e interrupções transitam pelas mesmas vias;
- Codificação 8b/10b -> 80% eficiência;
- Vazão máxima de 250 MB/s por via em ambas as direções
- Camada de enlace de dados -> sequenciamento de pacotes;
- Camada de transações -> exige buffers de tamanho adequado nos dispositivos;