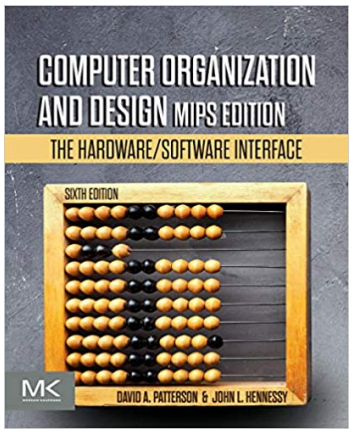
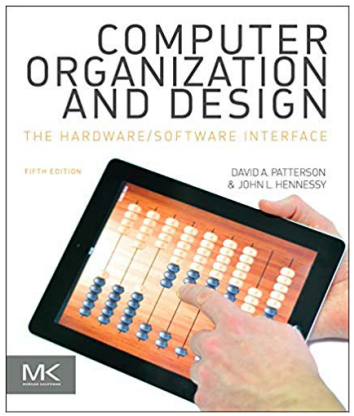
Aula 01 –Introdução

Prof. Dr. Clodoaldo Aparecido de Moraes Lima





Apresentação da Disciplina

Sumário

Introdução

Objetivos

- Introduzir os conceitos de Arquitetura de Von Neuman e os aprimoramentos que esta arquitetura vem experimentando.
- Fornecer contato com os principais componentes de internos de um processador

Introdução

Conteúdo

- Arquitetura de Von Neuman.
- Introdução às Arquiteturas RISC e CISC.
- Técnicas de Pipeline com exemplos em linguagem de montagem
- Processadores Superescalares. Processadores vetoriais.
- Arquiteturas paralelas.
- Processamento em Unidade de Processamento Gráfico (GPU).
- Fluxo de dados nas arquiteturas apresentadas.
- Análise de Arquiteturas Comerciais.

Programa e conteúdos

- Aula 1 - Apresentação da disciplina: Programa, Cronograma, Avaliações
- Aula 2 - Introdução a Arq de Computador - Arq. de Von Neuman
- Aula 3 - Linguagem de montangem - Revisão
- Aula 4 - Avaliando e compreendendo o desempenho
- Aula 5 - O processador: caminho de dados e controle
- Aula 6 - Melhorando o desempenho com o Pipelining
- Aula 7 - Paralelismo no nível de dados
- Aula 9 - Paralelismo no nível de instrução
- Aula 10 - Paralelismo no nível de thread
- Aula 11 - Prova P1 (14/10)

Programa e conteúdos

- Aula 12 - Arquitetura RISC
- Aula 13 - Arquitetura Modernas: RISC x CISC
- Aula 14 - Topologia de Rede de multiprocessadores
- Aula 15 - Arquiteturas Avançadas: Super-Escalar, Vetorial.
- Aula 16 - Arquiteturas Paralelas
- Aula 17 - Processamento em GPU
- Aula 18 - Análise de Arquiteturas Comerciais
- Aula 19 - Prova P2 (06/12)
- Aula 20 - Prova Substitutiva (09/12)
- REC

Critérios de Avaliação

Avaliação

- Frequência Mínima: 70%
- 2 Trabalhos (EP1, EP2)
- Média do trabalho (MP)
- Se EP1¿5.0 & EP2¿5.0 MT = (EP1+EP2)/2
- Else MT = minimo(EP1,EP2)
- 2 Provas Individual (P1, P2)
- Média das Provas (MP)
- MP = (P1+P2)/2
- Média Final (MF)

Critérios de Avaliação

Avaliação

- MF = 0.8*MP + 0.2*MT
- Se MF $\geq$ 5.0 então Aluno APROVADO
- Senão Se MF $\geq$ 3.0 então Aluno em Recuperação
- Média de Recuperação (MR)
- MR = (REC + MF)/2
- Senão Aluno REPROVADO

Bibliografia

Básica

- PATTERSON, D.A.; HENNESSY, J.L. Computer Organization and Design MIPS Edition: The Hardware/Software Interface, Morgan Kaufmann, 6a edition, 2021
- PATTERSON, D.A.; HENNESSY, J.L. Computer Organization and Design: The Hardware/Software Interface, Morgan Kaufmann, 5a edition, 2014
- STALLINGS, W. Arquitetura e Organização de Computadores, Prentice Hall, 5a. ed., 2002.

Bibliografia

Complementar

- HENNESSY, J.; PATTERSON, D. Computer Architecture: A Quantitative Approach, MK, 5a edition, 2011.
- TANENBAUM, A.S. Structured Computer Organization, Prentice Hall, 4th ed, 1999.
- CHAN, P.K.; MOURAD, S. Digital Design Using FPGAs. Prentice Hall, 1994.
- WAKERLY, J.F. Digital Design - Principles & Practices. 3a Ed., Prentice Hall, 2000.
- MANO, M.M. Computer System Architecture, Prentice-Hall, 1993.

Ao termino do curso, espera-se

- Como os programas escritos em uma linguagem de alto nivel, como C ou Java, são traduzidos para a linguagem de máquina e como o hardware executa os programas resultantes?
  - Compreender esses conceitos forma o alicerce para entender os aspectos do hardware e software que afetam o desempenho dos programas.
- O que é a interface entre o software e o hardware, e como o software instrui o hardware a realizar as funções necessárias?
  - Esses conceitos são vitais para entender como escrever muitos tipos de software.
- O que determina o desempenho de um programa e como um programador pode melhorar o desempenho?
  - Como veremos, isto depende do programa original, da tradução desse programa para a linguagem do computador e da eficiência do hardware em executar o programa.

Ao termino do curso, espera-se

- Que técnicas podem ser usadas pelos projetistas de hardware para melhorar o desempenho?
  - Este curso apresentará os conceitos básicos do projeto de um computador moderno.

Arquitetura versus Organização

Arquitetura

- Refere-se a atributos que tem impactos diretos sobre a execução lógica de um programa. Esses atributos são:
  - conjunto de instruções,
  - numero de bits que representa um determinando dado,
  - mecanismos de entrada e saída,entre outros.
- Lida com o funcionamento do Sistema Computacional.
- Corresponde aos aspectos visíveis a um programador em linguagem de máquina, tais como repertório de instruções, número de bits utilizado para representar vários tipos de dados, mecanismo de E/S e modos de endereçamento.

Arquitetura versus Organização

Organização

- Refere-se as unidades operacionais e suas interconexões. Os atributos que representa a organização de um computador são:
  - detalhes de hardware tais como sinais de controle,
  - interfaces entre computadores e periféricos,
  - tecnologias de memórias utilizadas.
- Diz respeito às unidades operacionais (UCP, unidade de memória, barramentos, sinais de controle, etc) necessárias para implementar as especificações de uma arquitetura. A organização é em geral transparente ao programador.

Sete Grandes Idéias

Uso de abstração para simplificar o projeto

usar abstrações para representar o projeto em diferentes níveis de representação; os detalhes de nível inferior são ocultados para oferecer um modelo mais simples em níveis superiores.

Fazendo o caso comum mais rápido

Tornar o caso comum rápido tenderá a melhorar melhor o desempenho do que otimizar o caso raro. Ironicamente, o caso comum é muitas vezes mais simples do que o caso raro e, portanto, é muitas vezes mais fácil de melhorar.

Melhora da Performance via paralelismo

Desde os primórdios da computação, os arquitetos de computadores oferecem projetos que obtêm mais desempenho ao realizar operações em paralelo.

Sete Grandes Idéias

Melhora da Performance via pipeline

Para combater o fogo, os habitantes de uma cidade formam uma corrente humana para transportar uma fonte de água até o fogo, pois poderiam mover baldes pela corrente com muito mais rapidez, em vez de indivíduos correndo para frente e para trás.

Melhora da Performance via predição

Em alguns casos, pode ser mais rápido, em média, adivinhar e começar a trabalhar, em vez de esperar até ter certeza, assumindo que o mecanismo para se recuperar de uma previsão errada não seja muito caro e que sua previsão seja relativamente precisa.

Sete Grandes Idéias

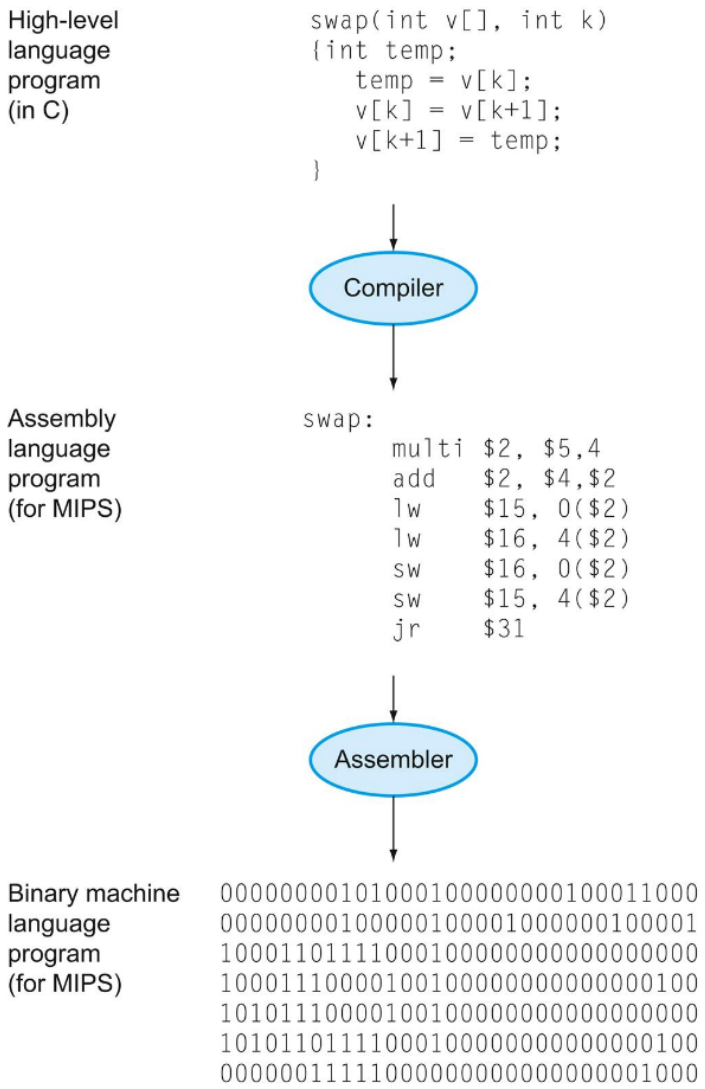
Hierarquia de memória

Os programadores desejam que a memória seja rápida, grande e barata, pois a velocidade da memória geralmente molda o desempenho, a capacidade limita o tamanho dos problemas que podem ser resolvidos e o custo da memória hoje costuma representar a maior parte do custo do computador.

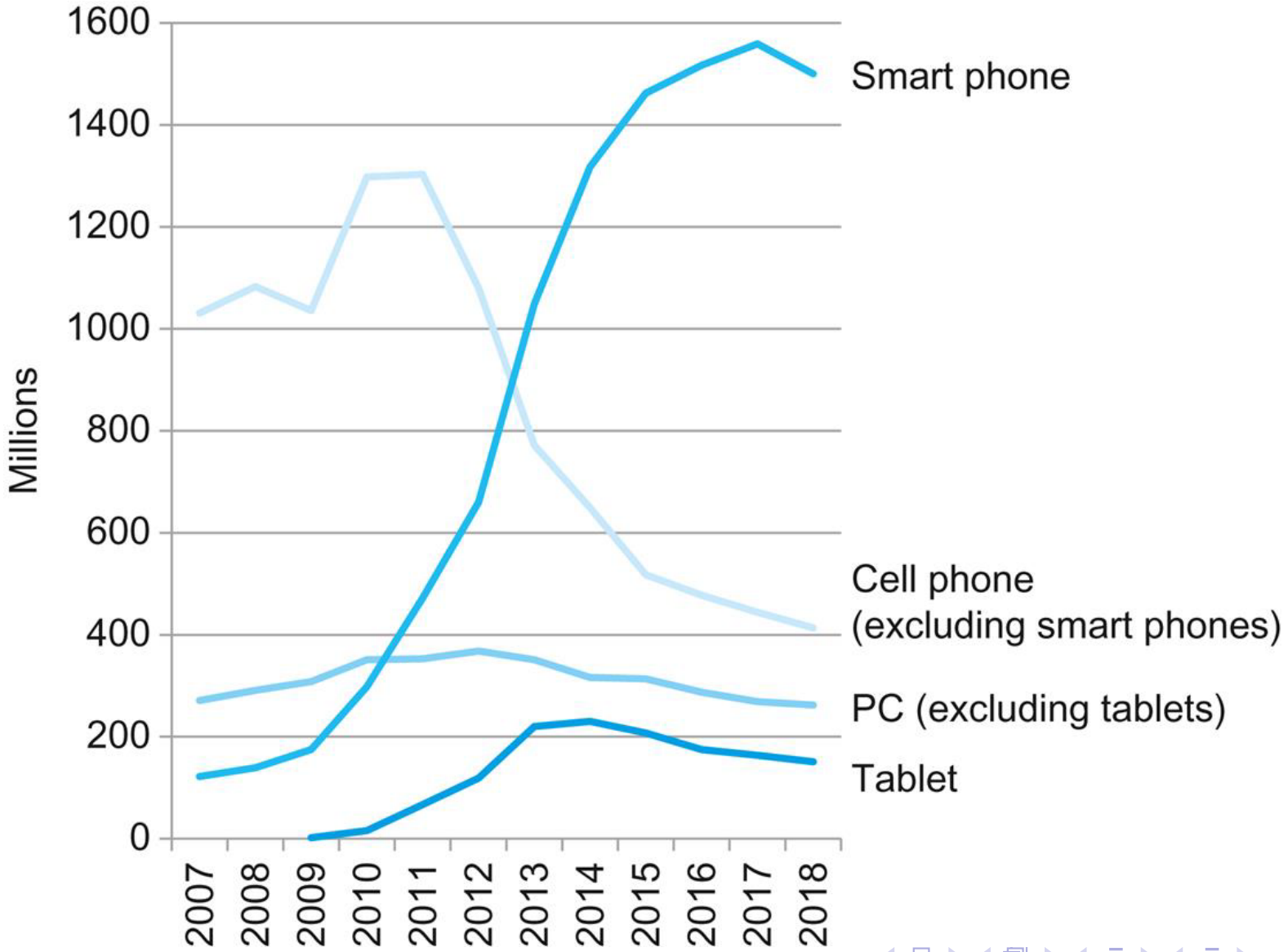
Confiabilidade via Redundância

Os computadores não precisam apenas ser rápidos; eles precisam ser confiáveis. Como qualquer dispositivo físico pode falhar, tornamos os sistemas confiáveis ao incluir componentes redundantes que podem assumir o controle quando ocorre uma falha e ajudar a detectar falhas.

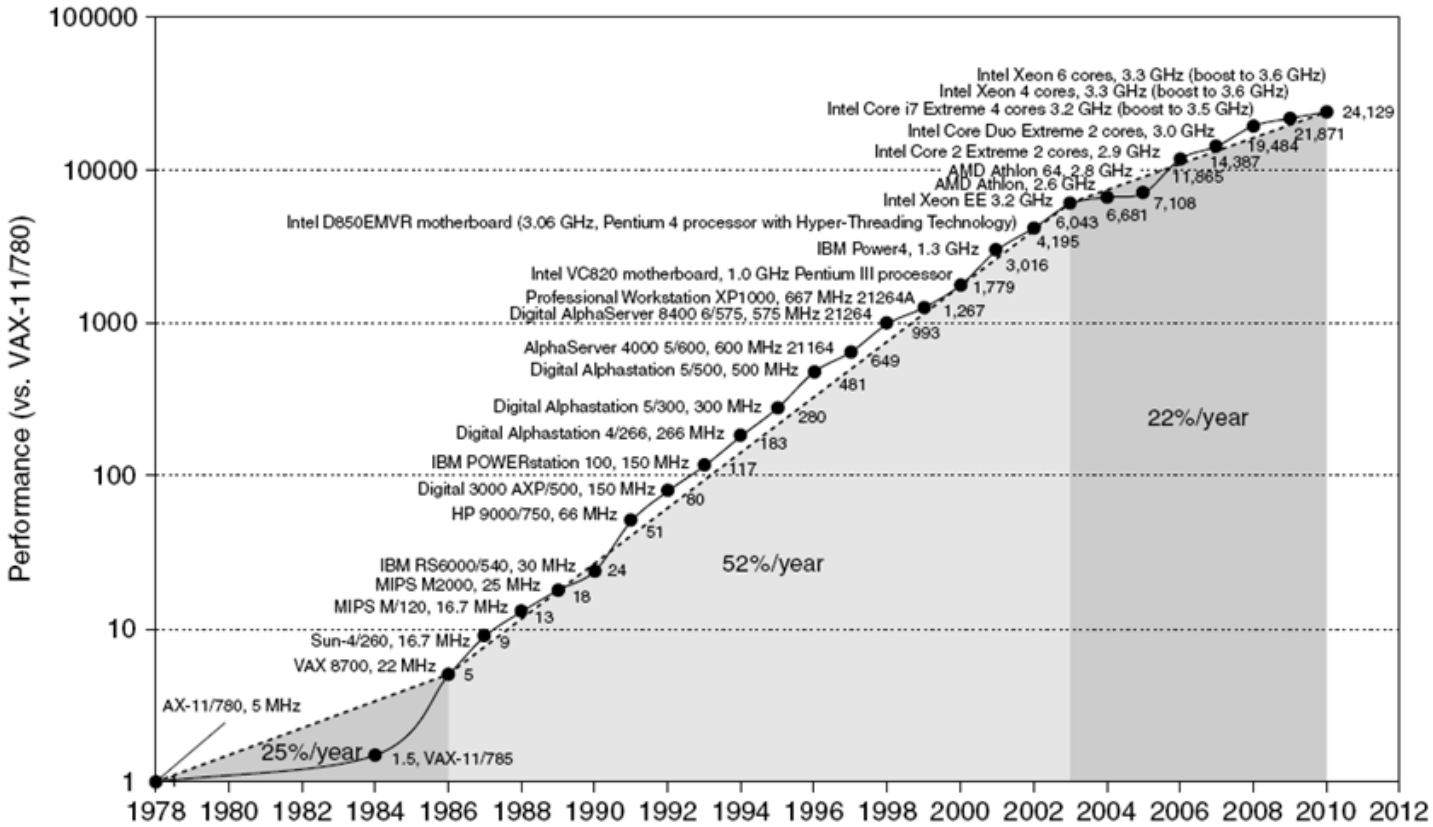
Linguagem programação



Era PosPC


1600
1400
1200
1000
800
600
400
200
0
Millions
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Smart phone
Cell phone
(excluding smart phones)
PC (excluding tablets)
Tablet

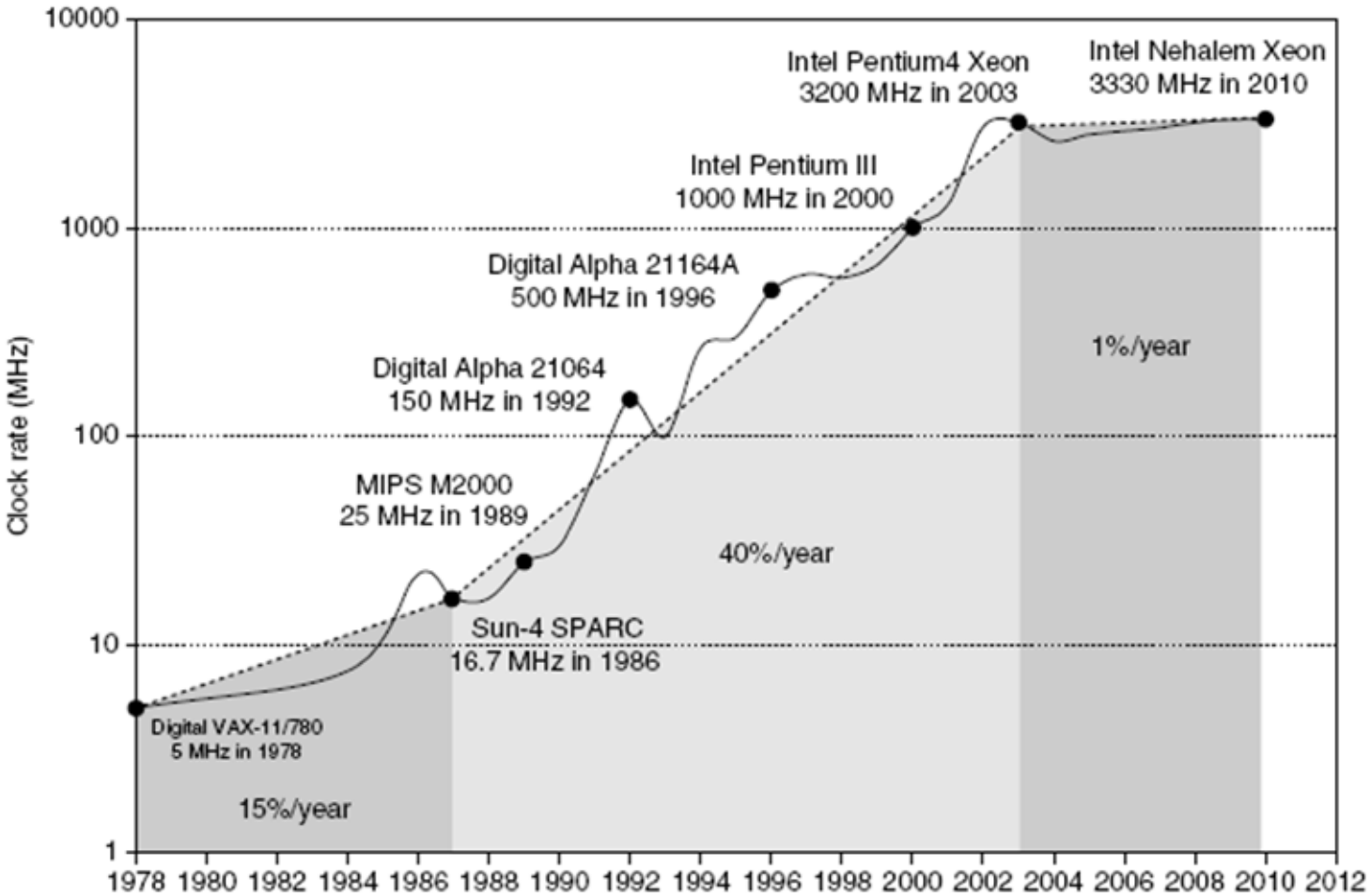
Desempenho histórico do microprocessador



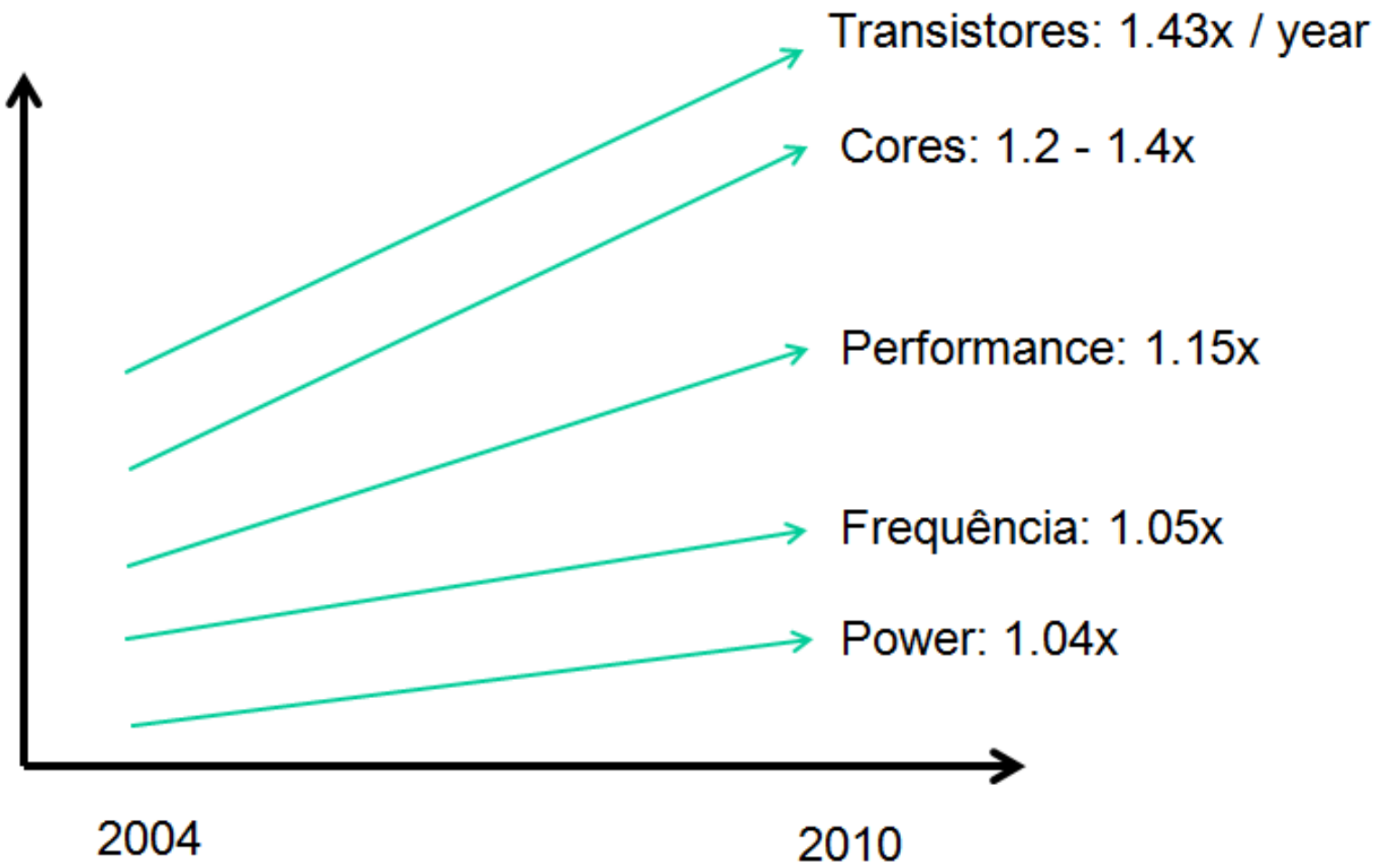
Pontos a serem observados

- O crescimento de 52% ao ano é por causa da velocidade de clock cada vez mais rápido e inovações arquitônicas (levou a uma velocidade 25x mais alta)
- O aumento da velocidade de clock caiu para 1% ao ano nos últimos anos
- O crescimento de 22% inclui a paralelização de múltiplos núcleos
- Lei de Moore: O número de transistores em um chip dobra a cada 18-24 meses

Aumento da velocidade de clock



Aumento da velocidade de clock



Tendências de Tecnologia do processador

- Densidade dos transistores aumenta em 35% por ano e tamanho do núcleo aumenta em 10-20% por ano ... mais funcionalidades
- Velocidade do transistor melhora linearmente com o tamanho (equação complexa que envolve tensões, resistências, capacitâncias)....pode levar a melhorias de velocidade de clock!
- Atrasos no fio não diminui no mesmo ritmo que atrasos na lógica
- Barreira da energia: não é possível executar de forma consistente em frequências mais altas sem atingir limites potência /térrmicos (Modo Turbo pode causar aumentos de frequência ocasionais)

O que ajuda na performance?

Sem aumentar a velocidade de clock

- Em um ciclo de clock, pode haver mais trabalho - uma vez que os transistores são mais rápido, e mais eficientes em termos de energia, e pode haver vários deles
- Melhora na arquitetura: encontrar mais paralelismo em uma thread, melhor previsão de desvios, melhores políticas de cache, melhor organizações de memória, mais paralelismo no nível de thread, etc

Para onde vamos

- Melhorias na velocidade do clock estão diminuindo devido a restrições de energia
- Difícil otimizar ainda mais um único core para melhorar o desempenho
- Multi-núcleos: cada nova geração de processadores vai acomodar mais núcleos
- Precisa de melhores modelos de programação e eficiente execução de aplicações multi-thread
- Precisa de melhor hierarquias de memória
- Precisa de uma maior eficiência energética
- Em alguns domínios, núcleos menos potente são atraentes

Tendência no consumo de energia

- Potência Dyn (dynpower) $\approx$ atividade x capacitância x voltage x frequência
- Capacitância por transistor e a tensão estão diminuindo, mas o número de transistores estão aumentando a um ritmo mais rápido; portanto, frequência de clock deve ser mantida constante
- Fuga de energia também estão aumentando; é uma função do número de transistor, corrente de fuga e tensão de alimentação
- Consumo de energia já está entre 100-500W em processadores de alto desempenho atuais
- Energia = Potência x tempo = (dynpower + lkgpower) x tempo

Potência x Energia

- Energia é uma métrica real: ela nos diz o verdadeiro "custo" na execução de uma tarefa fixa.
- Potência (energia/tempo) implica em restrições, só pode ser bastante rápido até a potência máxima fornecida ou aplicar algum tipo de resfriamento
- Se um processador A comsone 1,2x a potência do processador B, mas termina a tarefa em 30% menos tempo, a sua energia relativa é de 1:2x0:7 = 0; 84;
- Processador A é melhor, assumindo que 1.2x de energia pode ser fornecida pelo sistema

Reduzindo a Potência e a Energia

- Desligar os transistores que estão inativos (reduz o vazamento)
- Projetar o caso típico e desacelerar quando a atividade exceder um limiar
- DFS: escalonamento dinâmico da frequência - reduz frequência e potência dinâmica, mas prejudica a energia
- DVFS: escalonamento dinâmico da tensão e frequência - reduzir a tensão e frequência por (digamos) 10%; pode deixar um programa mais lento (digamos) em 8%, reduz a potência dinâmica em 27%, reduz a potência total (digamos) por 23%, reduzindo a energia total em 17%
- Nota: a queda de tensão transistor mais lento – queda na frequência

Outras Tendências Tecnológica

- DRAM aumenta densidade em 40-60% por ano, a latência tem sido reduzida em 33% em 10 anos, largura de banda melhora duas vezes mais rápido que a latência diminui
- Densidade do disco melhora em 100% a cada ano, a latência melhora de forma similar na DRAM
- Surgimento de tecnologias NVRAM podem fornecer uma ponte entre DRAM e unidades de disco rígido
- Além disso, crescente preocuação com a confiabilidade (transistores menores, operando a baixas voltagens, e muitos deles)

Confiabilidade e Disponibilidade

Um sistema alterna entre

- Realização de serviços: serviço corresponda as especificações
- Interrupção do serviço: serviços desvia das especificações

- A alternância é causada por falhas e restaurações
- Confiabilidade mede a realização de um serviço de forma contínua e é normalmente expressa como tempo médio até a falha (MTTF)
- Disponibilidade mede a fração de tempo que os serviços corresponde as especificações, expressa como MTTF / (MTTF + TMPR)

Custos

- O custo é determinado por vários fatores: volume, rendimento, maturidade fabricação, etapas de processamento, etc
- Importante: área do chip
- Pequena área – mais chips por wafer
- Pequena área – um defeito nos leva a descartar uma pequena área do chips, ou seja, o rendimento sobe
- De um modo geral, a metade da área – um terço do custo

Tecnologia para construção de Processadores e Memória

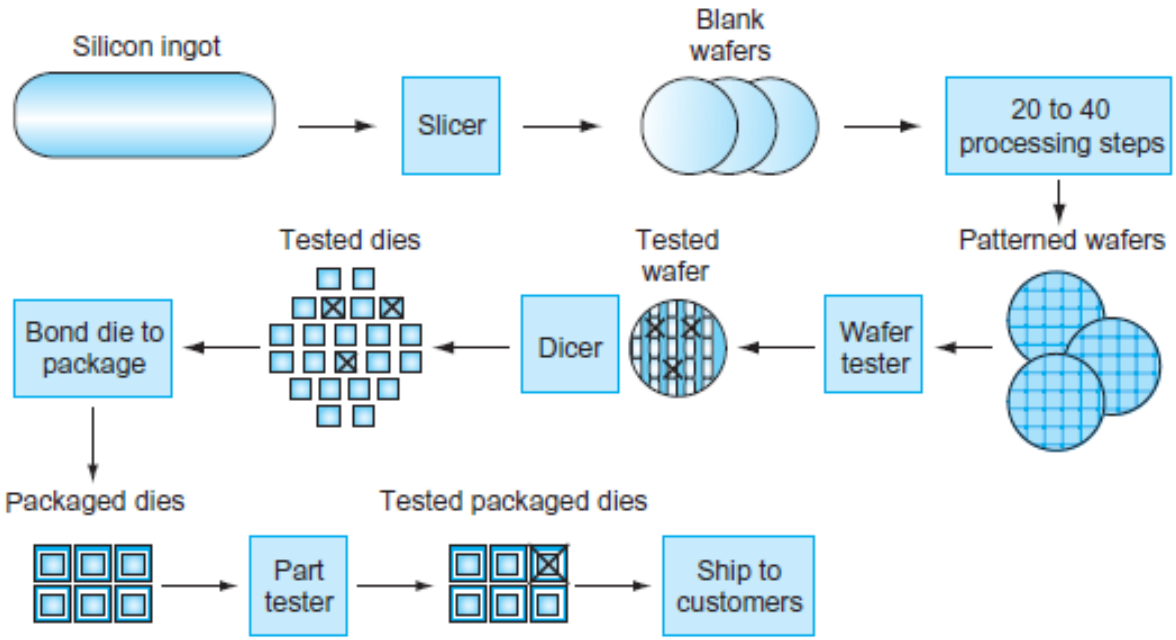
- Transistor é simplesmente uma chave liga/desliga controlada por eletricidade.
- Circuito integrado é a combinação de dezenas a centenas de transitores em um único chip
- VLSI (Escala Muito grande) usado para descrever o aumento no número de transitores de centenas para milhões
- Semicondutor é o material ou substância que não é bom condutor de eletricidade

Tecnologia para construção de Processadores e Memória

Com um processo químico especial, é possível acrescentar ao silício materiais que permitem minúsculas áreas se transformem em uma entre três dispositivos

- Excelentes condutores de eletricidade
- Excelentes isolantes de eletricidade
- Áreas que podem conduzir ou isolar sob condições especiais

Tecnologia para construção de Processadores e Memória



Tecnologia para construção de Processadores e Memória

- Após o lingote de silício serem fatiados, os wafers virgens passam por 20 a 40 passos para criar wafers com padrões
- Esses wafers com padrões são testados com um testador de wafers e é criado um mapa das partes boas
- Os wafer são divididos em dies (moldes)
- Esses dies bons são soldados e encapsulados
- Novamente são testados antes de serem remetidos para os clientes