



Organização e Arquitetura de Computadores
Arquitetura de computadores Paralelos

Prof. Marcelo Rabello
 marcelo.rabello@unifg.edu.br

 **UNIFG**
 LAUREATE INTERNATIONAL UNIVERSITIES

1

Objetivos de aprendizagem

1. Distinguir, comparar e classificar as estruturas das máquinas

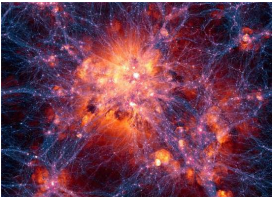


2

Introdução

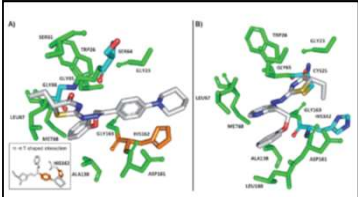
Embora os computadores continuem a ficar cada vez mais rápidos, as demandas impostas a eles estão crescendo no mínimo com a mesma rapidez.

Astrônomos



Ilustris Simulation: Most detailed simulation of our Universe
<https://youtu.be/NjSFR40SY58>
<http://dx.doi.org/doi:10.1038/nature13316>

Farmacêuticos



Structural Design, Synthesis and Structure–Activity Relationships of Thiazolidinones with Enhanced Anti-Trypanosoma cruzi Activity
<http://dx.doi.org/doi:10.1002/cmde.201300354>

Simulation of Ice Crystal Melting
<https://youtu.be/NQhJatCKghE>

3

Introdução

Embora com o aumento das velocidades de clock, a velocidade do circuito não pode aumentar indefinidamente.

Projetistas de computadores de alta tecnologia - Limite da velocidade da luz

Como fazer com que os elétrons e fótons se movam mais rapidamente?



Dissipação de calor - Um computador ou um ar-condicionado?

Diminuição do tamanho dos transistores, até quando?

Transistores com um número pequeno de átomos os efeitos da mecânica quântica pode se tornar um problema.

Máquina com uma única CPU
 Tempo de ciclo de 0,001ns.

Máquina com 1.000 CPUs
 Tempo de ciclo de 1 ns.

4

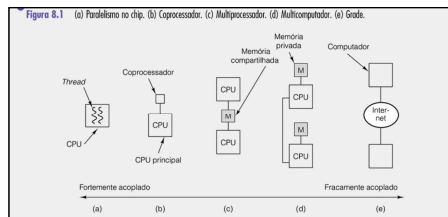
Paralelismo

Pode ser introduzido em vários níveis:

- Chip: Pipeline e projetos superescalares (várias unidades funcionais).
- Coprocessador: Várias CPUs no mesmo chip
- Multiprocessador: Replicar CPUs inteiras.
- Multicomputador (Clusters).
- Computação em Grade.

Fortemente acoplados: Dois elementos de processamento estão perto um do outro.

Fracamente acoplados: Dois elementos longe um do outro, baixa largura de banda e alto atraso.



5

Paralelismo no Chip

Um modo de aumentar a produtividade de um chip é conseguir que ele faça mais coisas ao mesmo tempo. Em outras palavras, explorar o paralelismo.

Modos de paralelismo no nível do chip:

- ▶ No nível da instrução;
- ▶ *Multithreading*;
- ▶ Multiprocessadores em um único chip:

Essas técnicas são bem diferentes, mas cada uma delas ajuda à sua própria maneira.

Conseguir que mais atividades aconteçam ao mesmo tempo.

6

Paralelismo no Chip

- ▶ No nível da instrução:

Paralelismo em nível mais baixo

Emitir múltiplas instruções por ciclo de *clock*.

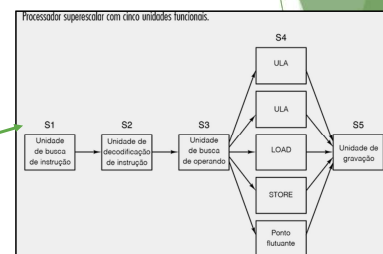
Tipos:

Processadores Superescalares

Processadores que emitem múltiplas instruções (4 ou 6) em um único ciclo de *clock*.

Necessita de várias unidades funcionais para passar todas essas instruções.

Normalmente apresentam apenas 1 *pipeline*.

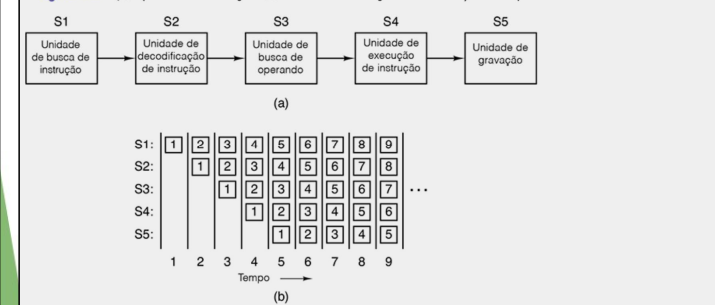


7

Paralelismo no Chip

- ▶ No nível da instrução: Processadores Superescalares

Figura 2.4 (a) Pipeline de cinco estágios, (b) Estado de cada estágio como uma função do tempo. São ilustrados nove ciclos de *clock*.

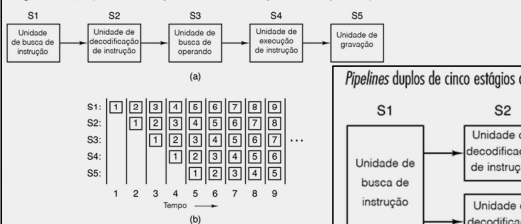


8

Paralelismo no Chip

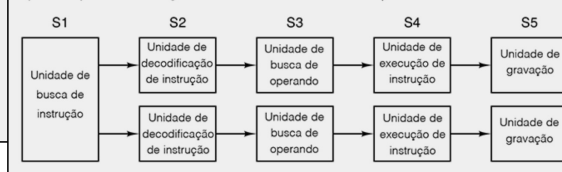
► No nível da instrução: Processadores Superescalares

Figura 2.4 (a) Pipeline de cinco estágios. (b) Estado de cada estágio como uma função do tempo. São ilustrados nove ciclos de clock.



Única unidade de busca de instruções busca pares de instruções ao mesmo tempo e coloca cada uma delas em seu próprio *pipeline*, com sua própria ULA.

Pipelines duplos de cinco estágios com uma unidade de busca de instrução em comum.



Para poder executar em paralelo, as duas instruções não devem ter conflito de utilização de recursos e nenhuma deve depender do resultado da outra.

9

Paralelismo no Chip

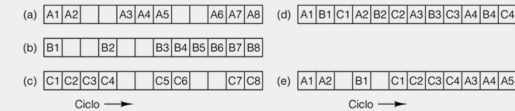
► Multithreading no chip:

Todas as CPUs modernas, com paralelismo (pipeline), têm um problema inerente: quando uma referência à memória encontra uma ausência das caches de nível 1 e nível 2, há uma **longa espera** até que a palavra requisitada (e sua linha de cache associada) **sejam carregadas na cache**, portanto, o pipeline para.

Uma abordagem para lidar com essa situação, denominada **multithreading no chip**, permite que a CPU gerencie múltiplos threads de controle ao mesmo tempo em uma tentativa de mascarar essas protelações.

Em suma, se o thread **1 estiver bloqueado**, a CPU ainda tem **uma chance de executar o thread 2**, de modo a manter o hardware totalmente ocupado.

Figura 8.7 (a)–(c) Três threads. Os retângulos vazios indicam que o thread parou esperando por memória. (d) Multithreading de granulação fina. (e) Multithreading de granulação grossa.



10

Paralelismo no Chip

► Multithreading no chip: Exemplo Prático

No início da década de 200, o Pentium 4 já estava em produção, os arquitetos da intel procuraram vários meios de aumentar sua velocidade sem mudar a interface de programadores(algo que jamais seria aceito).

Modos:

1. Aumentar a velocidade de clock.
2. Colocar duas CPUs em um chip.
3. Adicionar unidades funcionais.
4. Aumentar o comprimento do pipeline.
5. Usar **multithreading**.

- Primeira CPU com **multithreading** da intel: **Xeon**.
- Adicionada ao **Pentium 4**, a partir da versão **3,06 GHz**
- Adicionada também as versões mais rápidas do processador Pentium, incluindo o Core i7.

A intel chamou de **hyperthreading** à implementação de **multithreading**.

Para o sistema operacional, o chip Core i7 com **hyperthreading** parece um processador dual em que ambas as CPUs compartilham em comum uma cache e a memória principal.

Utilizar hardware que, se não fosse por isso, ficaria abandonado.

11

Paralelismo no Chip

► Multiprocessadores com um único chip

Embora o **multithreading** ofereça ganhos em desempenho significativos por um custo modesto, para algumas aplicações é preciso um ganho em desempenho muito maior do que ele pode oferecer.

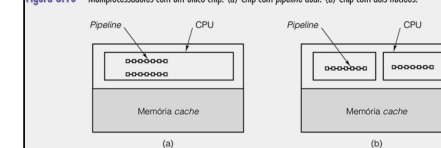
Para conseguir esse desempenho estão sendo desenvolvidos **chips multiprocessadores**.

Há duas áreas de interesse para esses chips que contêm duas ou mais CPUs:

* Servidores de alta tecnologia; * Equipamentos eletrônicos de consumo.

Com os avanços na tecnologia VLSI, agora é possível colocar **duas ou mais CPUs** de grande capacidade em um único chip, visto que essas CPUs em geral **compartilham** a mesma **cache** de nível 2 e **memória principal**, elas se qualificam como um **multiprocessador**.

Figura 8.10 Multiprocessadores com um único chip. (a) Chip com pipeline dual. (b) Chip com dois núcleos.



12

Paralelismo no Chip

► Multiprocessadores com um único chip

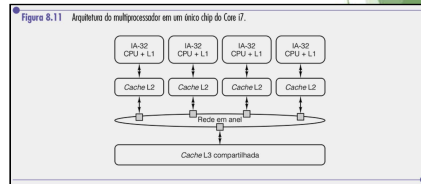
A CPU Core i7 é um processador em um único chip manufacturado com **quatro ou mais núcleos** em uma única pastilha de silício.

Cada processador no Core i7 tem suas **próprias caches L1 privada para instrução e dados**, mais sua **própria cache L2 unificada privada**.

Os processadores são conectados às **caches privadas com conexões ponto a ponto dedicadas**. As **caches L2 se conectam à cache compartilhada L3** usando uma rede em anel.

Rede em anel:

- Oferece um modo de mover pedidos de memória e E/S entre as caches e processadores.
- Executa as verificações necessárias para garantir que cada processador esteja sempre tendo uma visão coerente da memória.



13

Coprocessadores

Agora que já vimos alguns dos modos de conseguir paralelismo no chip, vamos subir um degrau e ver como o computador pode ganhar velocidade com a adição de um segundo processador especializado.

Processadores gráficos: Coprocessadores são usados no tratamento de processamento gráfico de alta resolução, como renderização 3D.

CPUs comuns não são muito boas nas computações maciças necessárias para processar as grandes quantidades de dados requeridas nessas aplicações.

Alguns PCs atuais e a maioria dos PCs futuros serão equipados com GPUs (Graphics Processing Units - unidades de processamento gráfico) para os quais passarão grandes porções do processamento geral

14

Coprocessadores

A regularidade e a estrutura desses programas os tornam alvos especialmente fáceis para **aceleração** por meio de execução paralela.

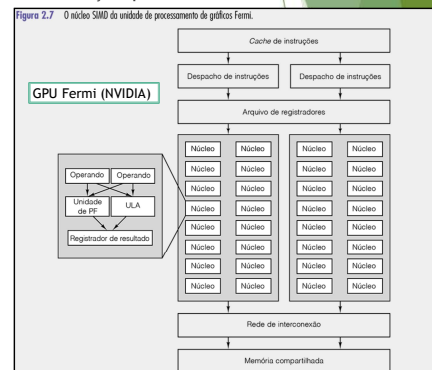
Métodos:

Processadores SIMD e **processadores vetoriais**.

Embora esses dois esquemas guardem notáveis semelhanças na maioria de seus aspectos, por ironia o primeiro deles é considerado um **computador paralelo**, enquanto o segundo é considerado uma **extensão de um processador único**.

Processadores SIMD - Fluxo único de instruções, fluxo múltiplo de dados

GPUs utilizam SIMD para fornecer poder computacional maciço com poucos transistores. Processamento de gráficos é apropriado para SIMD. Maioria dos algoritmos é regular, com operações repetidas sobre pixels, vértices, texturas e arestas.

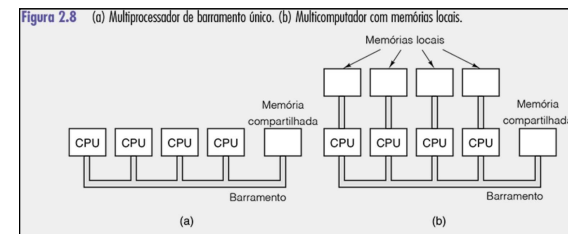


15

Multiprocessadores

Primeiro sistema paralelo com CPUs totalmente independentes, com mais de uma CPU que **compartilha uma memória em comum**.

Os elementos de processamento em um processador SIMD não são CPUs independentes, uma vez que há **uma só unidade de controle compartilhada** por todos eles.



16

Multiprocessadores

Aparecem para o sistema operacional como se tivessem memória compartilhada que pode ser acessada usando instruções comuns **LOAD** e **STORE**.

Programas escritos para um multiprocessador podem acessar qualquer localização na memória sem nada saber sobre a topologia interna ou o esquema de implementação.

Atraente para os programadores.

Não podem ser ampliados para grandes tamanhos.

17

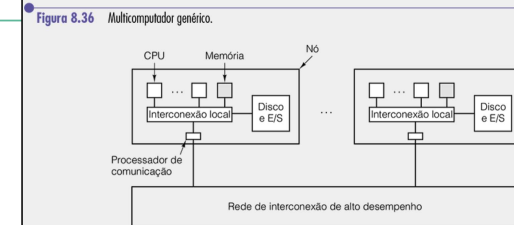
Multicomputadores

Embora seja um tanto fácil construir multiprocessadores com um número modesto de processadores (≤ 256), construir multiprocessadores com um número grande é surpreendentemente difícil.

A dificuldade está em **conectar todos os processadores à memória**.

Multicomputadores: Construção de sistemas que consistissem em grandes números de computadores interconectados, cada um com sua **memória própria e privada, mas nenhuma em comum**.

Costuma-se dizer que as CPUs de um multicomputador são fracamente acopladas, para contrastá-las com as CPUs fortemente acopladas de um multiprocessador.



18

Multicomputadores

Há vários formatos e tamanhos de multicomputadores, portanto, é difícil dar uma taxonomia clara para eles. Não obstante, há dois “estilos” que se destacam: os **MPPs** e os **clusters**.

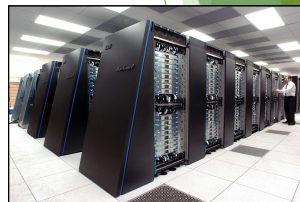
MPPs (processadores maciçamente paralelos):

Imensos supercomputadores de muitos milhões de dólares. Eles são usados em ciências, em engenharia e na indústria para cálculos muito grandes, para tratar números muito grandes de transações por segundo ou para data warehousing (armazenamento e gerenciamento de imensos bancos de dados).

Inicialmente utilizados supercomputadores científicos, mas, agora, a maioria deles é usada em ambientes comerciais.

Sucessores dos poderosos mainframes da década de 1960

BlueGene (IBM) 1999



19

Multicomputadores

MPPs: Utiliza CPUs padronizadas como seus processadores.

Destaques: Utilização de uma rede de interconexão proprietária de desempenho muito alto, projetada para mover mensagens com baixa latência e a alta largura de banda (mensagens pequenas).

Extensivos softwares e bibliotecas proprietárias.

Enorme capacidade de E/S.

Útil quando se têm quantidades maciças de dados a processar (terabytes).

Tolerância à falha.

20

Multicomputadores

Cluster: Consiste em centenas de milhares de PCs ou estações de trabalho conectadas por uma placa de rede disponível no mercado.

A diferença entre um MPP e um cluster é análoga a de um mainframe e um PC: Ambos têm uma CPU, ambos têm RAM, ambos têm discos, ambos têm um sistema operacional e assim por diante.

Os mainframe são mais rápidos (exceto talvez o sistema operacional). No entanto, em termos qualitativos, eles são considerados diferentes e são usados e gerenciados de modo diferente. Essa mesma diferença vale para MPPs em comparação com clusters.



21

Computação em Grade

Muitos dos desafios atuais na ciência, engenharia, indústria, meio ambiente e outras áreas são de grande escala e interdisciplinares.

Resolvê-los requer a experiência, as habilidades, conhecimentos, instalações, softwares e dados de múltiplas organizações e, muitas vezes, em países diferentes.

Alguns exemplos são os seguintes:

1. Cientistas que estão desenvolvendo uma missão para Marte.
2. Um consórcio para construir um produto complexo (por exemplo, uma represa ou uma aeronave).
3. Uma equipe de socorro internacional para coordenar o auxílio prestado após um desastre natural.

Algumas dessas cooperações são de longo prazo, outras de prazos mais curtos, mas todas compartilham a linha comum que é conseguir que organizações individuais, com seus próprios recursos e procedimentos, trabalhem juntas para atingir uma meta comum. Até há pouco tempo, conseguir que organizações diferentes, com sistemas operacionais de computador, bancos de dados e protocolos diferentes, trabalhassem juntas era muito difícil. Contudo, a crescente necessidade de cooperação interorganizacional em larga escala levou ao desenvolvimento de sistemas e tecnologia para conectar computadores muito distantes uns dos outros no que é denominado grade.

considerada um cluster muito grande, internacional, fracamente acoplado e heterogêneo.

22

Computação em Grade

OBJETIVO:

Proporcionar infraestrutura técnica para permitir que um grupo de organizações que compartilham uma mesma meta forme uma organização virtual.

Essa organização virtual tem de ser flexível, com um quadro de associados grande e mutável, permitindo que seus membros trabalhem juntos em áreas que consideram apropriadas e, ao mesmo tempo, permitindo que eles mantenham controle sobre seus próprios recursos em qualquer grau que desejarem.

Com essa finalidade, pesquisadores de grade estão desenvolvendo serviços, ferramentas e protocolos para habilitar o funcionamento dessas organizações virtuais.

A grade é inerentemente multilateral, com muitos participantes de mesmo status.

Ela pode ser contrastada com estruturas de computação existentes.

Figura 8.52 Camadas da grade.

Camada	Função
Aplicação	Aplicações que compartilham recursos gerenciados de modos controlados
Coletiva	Descoberta, corretagem, monitoração e controle de grupos de recursos
De recursos	Acesso seguro e gerenciado a recursos individuais
Base	Recursos físicos: computadores, armazenamento, redes, sensores, programas e dados

23

Atividade Extraclasse

Leitura do artigo "Arquiteturas Superescalares" Santana, M. F. UNICAMP 2010.

<https://www.ic.unicamp.br/~ducatte/mo401/1s2010/T2/100602-t2.pdf>

Leitura do capítulo 8 do livro Organização Estruturada de Computadores. Tenenbaum 2013.

24

Dúvidas? Sugestões?



25

2112 Overture/The temple of Syrinx by Rush

<https://youtu.be/APbogD9uwFU>



*"We are the priests
Of the Temples of Syrinx
Our great computers
Fill the hallowed halls
We are the priests
Of the Temples of Syrinx
All the gifts of life
Are held within our walls"*

26