Universidade Federal do Rio de Janeiro
Pós-Graduação em Informática
DCC/IM - NCE/UFRJc

The future of Microprocessors

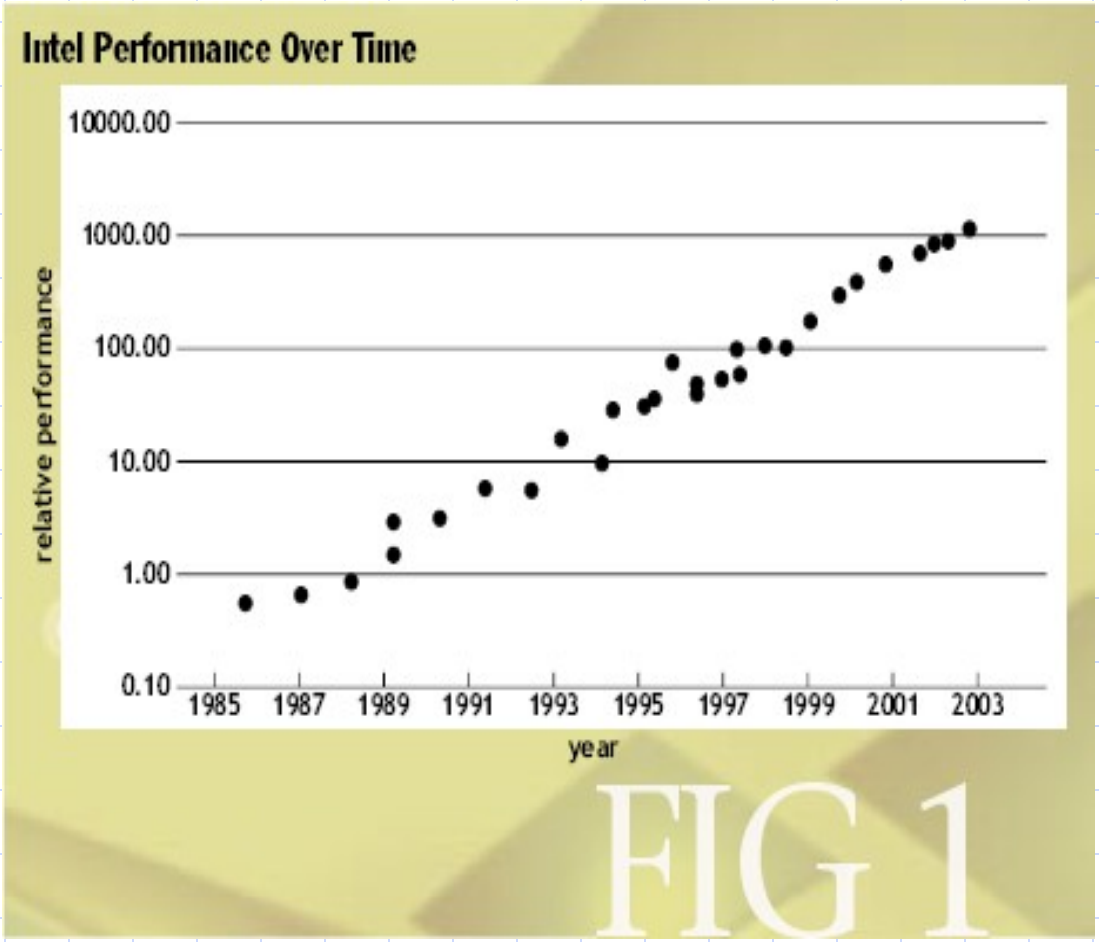
Kunle Olukotun and Lance Hammond, Stanford University

Rafael Viana de Carvalho

Evolução dos Processadores

- O desempenho dos processadores tem aumentado exponencialmente no decorrer dos últimos anos;
  - Transistores menores e mais rápidos
  - Consegue-se atualmente extrair mais paralelismo dos softwares
- Aumento do desempenho de acordo com a lei de Moore:
  - A cada 18 meses a capacidade de processamento dobra

Evolução dos Processadores



Evolução dos Processadores

- A extração do paralelismo nos processadores é virtualmente invisível aos programadores de softwares
- Dentro dos processadores isso resultou em modificações, tais como:
  - Aumento do número de instruções despachadas por ciclo
  - Aumento da frequência do clock mais rápido que o estimado pela Lei de Moore (superpipelining)
- Pipelining permitiu aos projetistas aumentar o clock dividindo a execução das instruções em estágios cada vez menores

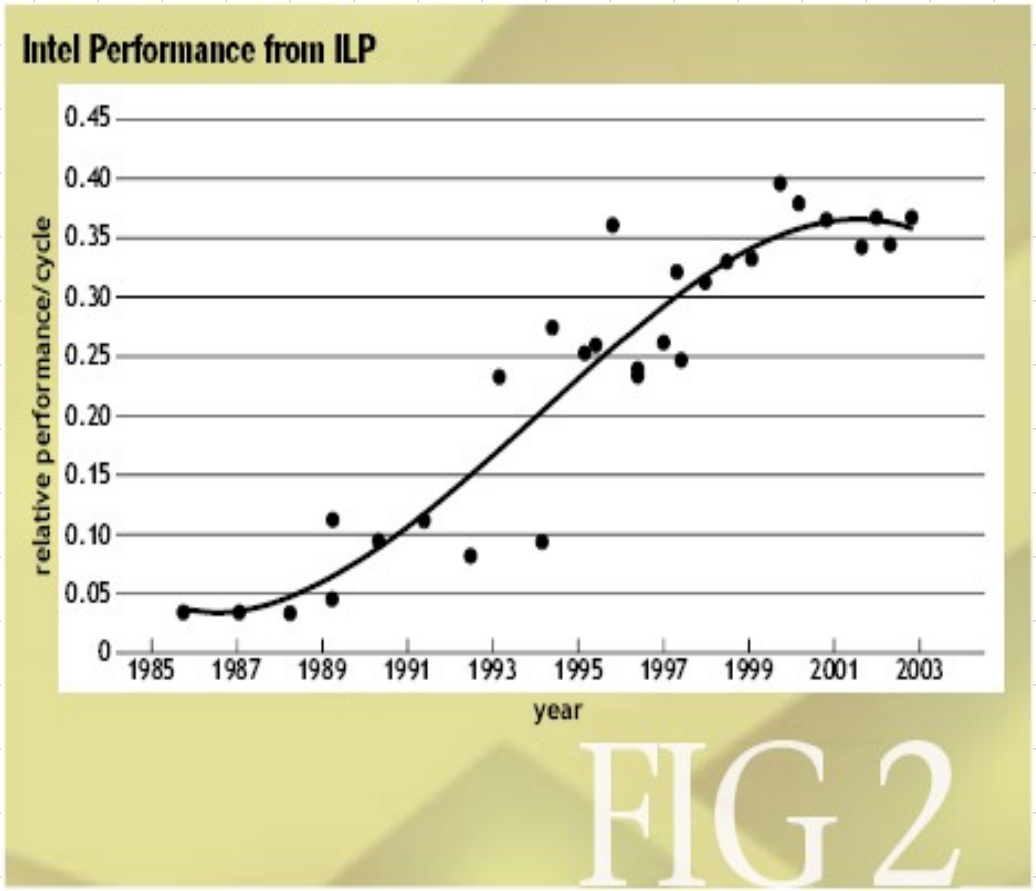
Evolução dos Processadores

- Processadores superescalares foram desenvolvidos com objetivo de executar simultaneamente várias instruções em um mesmo fluxo de execução
  - Examina dinamicamente conjunto de instruções do fluxo
  - Encontra as instruções capazes de serem despachadas simultaneamente (dependências de instruções)
  - Executa as instruções, as vezes até fora da ordem original do programa

Evolução dos Processadores

- O desempenho pode ser melhorado se o programador e/ou compilador ajustar o escalonamento das instruções e o layout dos dados para mapear melhor a arquitetura e o cache
- Códigos antigos executarão corretamente sem modificações mas com desempenho não ótimo caso não haja um reescalonamento do código
- Fluxos típicos de instruções tem uma quantidade de paralelismo limitada
  - Processadores superescalares que despacham mais de quatro instruções por ciclo não apresentam muito ganho adicional na maior parte das aplicações

Evolução da Utilização de Paralelismo nos Processadores

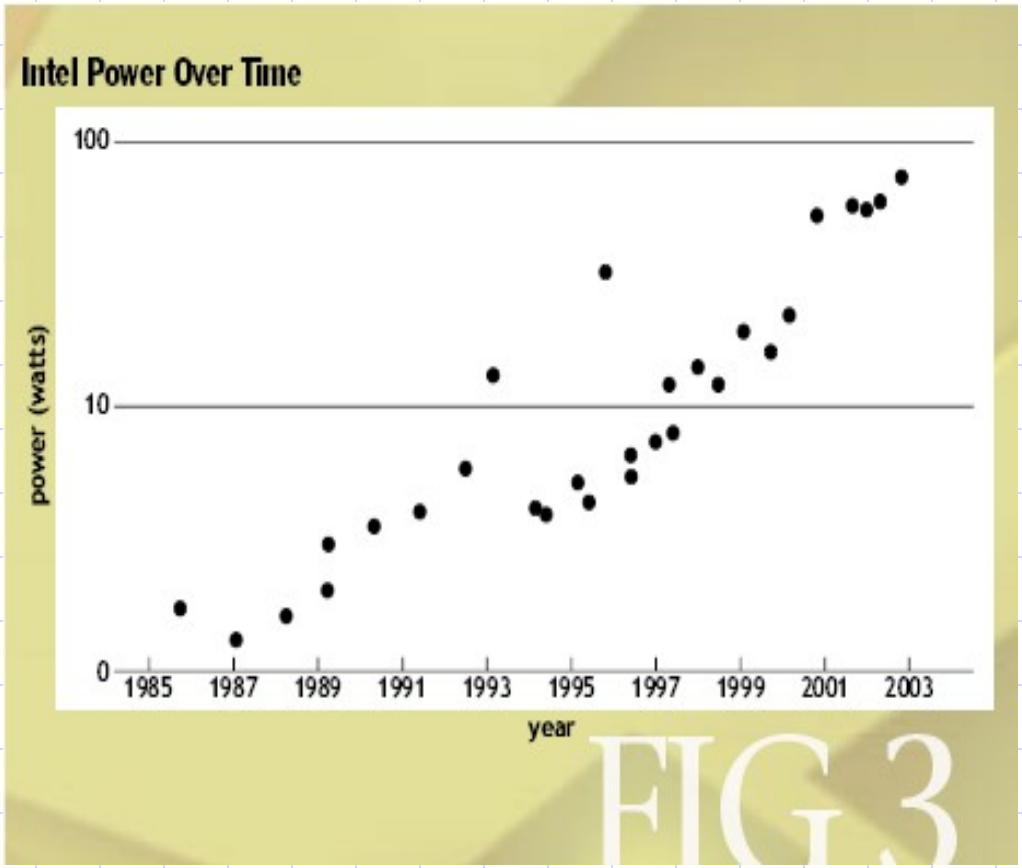


- Eficiência dos processadores Intel em relação a evolução da utilização do paralelismo das instruções no decorrer do tempo. Como podemos ver, a utilização do paralelismo de instruções foi mais explorada nos processadores Intel mais atuais.

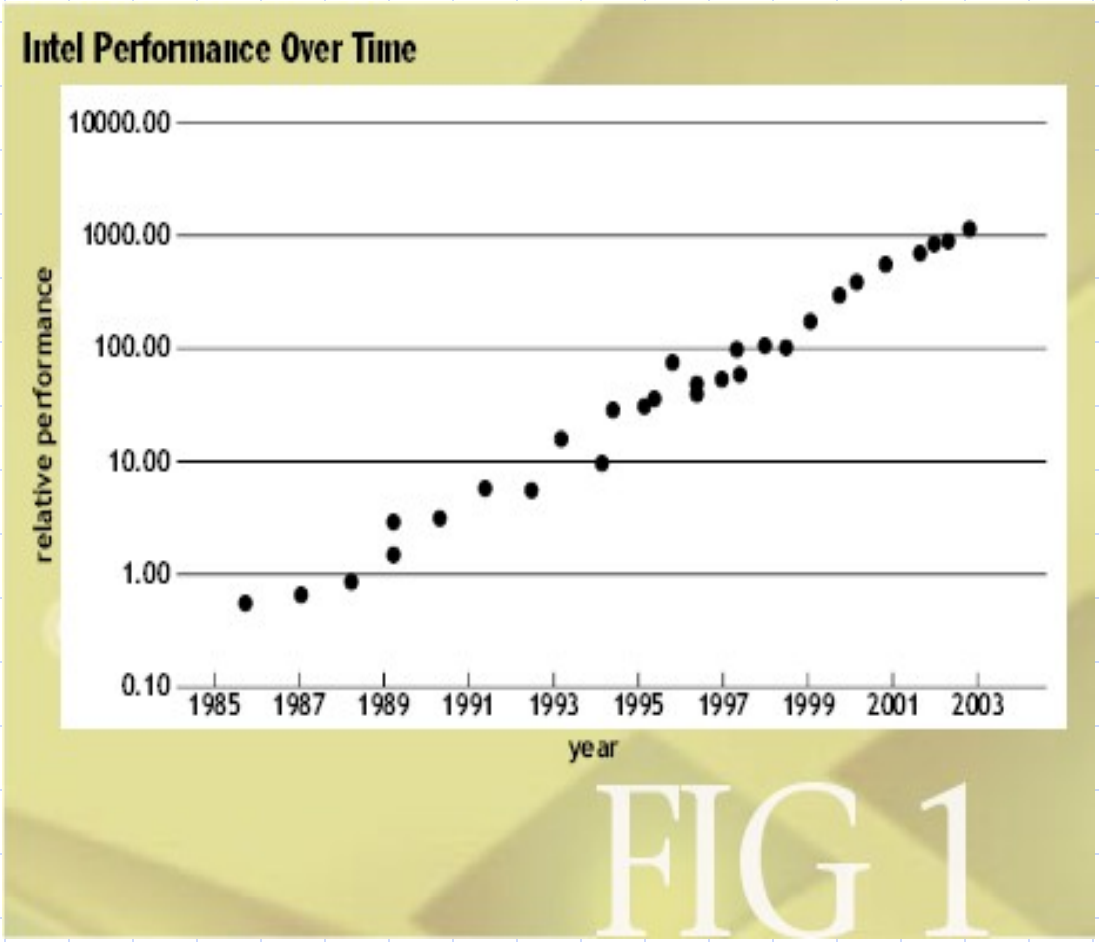
Limitações na Construção de Novos Processadores

- Hoje em dia o progresso no desenvolvimento dos núcleos dos processadores está parado devido a uma limitação física: a potência
- Os processadores superescalares com pipeline das ultimas 2 décadas, dissipavam em torno de 100 watts de potência
- Todos pensavam que a geração de processadores de silício reduziria essa potência, pois quanto menor os transistores, menor a energia requerida para seu funcionamento
- Isso ocorreria se os processadores "encolhessem" em relação ao seu tamanho
- Na prática, os processadores estão utilizando mais transistores em seus núcleos, o que tem exigido mais potência.
- Isso implica no aumento do aquecimento dos processadores e, conseqüentemente, há uma necessidade de melhorar a tecnologia dos sistemas de refrigeração, os quais não tem seguido o crescente desenvolvimento dos processadores.

Limitações na Construção de Novos Processadores



Evolução dos Processadores



Evolução dos Processadores: Multiprocessadores

- Com a Internet, a necessidade de servidores capazes de manipular um grande número de de solicitações de forma rápida na rede, tem aumentado drasticamente.
- Tarefas independentes propagadas entre os vários computadores (servidores) separados
  - Acesso a páginas da web
  - Serviços de arquivo
  - Banco de dados
- Com o aumento das requisições ao longo do tempo, se fez necessário o aumento do número de processadores
  - Mais espaço e refrigeração

Evolução dos Processadores: Multiprocessadores

- Os multiprocessadores consistem em dois ou mais processadores separados, conectados entre si por um barramento, hub ou rede, compartilhando a mesma memória e dispositivos de I/O.
- Todo o sistema pode ser pequeno e usar menos potência que o equivalente a dois processadores convencionais, pois compartilha componentes:
  - Memória
  - HDs
  - Periféricos
  - Suprimento de energia

Evolução dos Processadores: CMPs

- Os data-centers que comportam os servidores possuem limitações de:
  - Espaço físico
  - Estrutura para fiação
  - Refrigeração
- Quanto mais processadores são adicionados maior a necessidade de expansão, com os multiprocessadores (CMPs), a idéia de compartilhamento dos recursos diminuem essa necessidade, além de uma redução no consumo de energia
- O primeiro objetivo dos CMPs era de reduzir o espaço utilizado nas torres dos servidores, implementando para isso dois ou mais processadores superescalares juntos em uma única unidade
- A performance por unidade de volume foi incrementada

Evolução dos Processadores: CMPs

- Percebeu-se uma pequena economia na energia pois os processadores unitários (convencionais) podiam compartilhar uma única conexão com o resto do sistema, reduzindo a quantidade de comunicação rápida necessária nas infra-estruturas;
- A AMD e a Intel desenvolveram processadores que compartilhavam não só recursos de hardware mas também recursos da interface dos núcleos dos microprocessadores

Evolução dos microprocessadores: CMPs

- Quando um CMP substitui um uniprocessador, é possível alcançar um throughput igual ou melhor nas cargas de trabalho de servidores (Server-oriented) com apenas metade da velocidade do clock original
  - Cada requisição pode levar até duas vezes mais tempo para ser processada devido a redução da freqüência de clock
  - Mas a perda será pequena, pois a característica desse tipo de aplicação é memory-bound ou I/O-bound e não CPU-bound
  - Já que duas requisições poderão ser processadas simultaneamente, o throughput será igual ou melhor, a menos que haja grande contenção na memória ou no disco
- Mesmo que a performance não seja tão superior, ainda há vantagens em se utilizar os CMPs:
  - Uma redução na freqüência permite projetar sistema com uma redução linear na tensão de alimentação
  - A potência necessária para obter a performance original é bem menor, geralmente a metade (a potência é proporcional ao quadrado da voltagem)

Evolução dos microprocessadores: CMPs

$$P \approx C \cdot V^2 \cdot f$$

CMPs – Melhorias no throughput

- Para as cargas de trabalho orientadas para o throughput, pode conseguir mais potencia/performance e performance/área do chip, levando-se ao extremo o conceito que a latência não é importante, e construindo-se um CMP com muitos núcleos pequenos ao invés de poucos núcleos grandes
- Nesse caso a utilização de processadores superescalares não se faz necessário pois em um servidor é mais importante realizar várias tarefas ao mesmo tempo e manter os processadores sempre ocupados do que realizar poucas em um tempo curto e ficar com processadores ociosos
- Com isso pode-se trocar um CMP com poucos processadores superescalares por um CMP com muitos processadores escalares. Essa experiência foi feita com sucesso servidor Sun Niagara com seus oito processadores escalares SPARC ao invés de um par de processadores superescalares UltraSPARC

CMPs – Melhorias na Latência

- A performance de várias aplicações é medida em termos da latência de execução de cada tarefa individual ao invés do alto troughput global de algumas tarefas não relacionadas
- Os usuários de desktop estão preocupados com que seus computadores respondam a seus comandos o mais rápido possível
- Essa situação tem mudado lentamente, já que várias aplicações tem sido escritas com tarefas que rodam em segundo plano
- Usuários mais técnicos estão mais interessados em quanto tempo o computador irá demorar para executar uma única aplicação do que executar muitas em paralelo
- Multiprocessadores podem executar esses tipos de aplicações mais rapidamente, porém isto requer um esforço por parte dos programadores para quebrar cada thread com latência longa de execução em vários pedaços de threads menores que possam ser executadas em vários processadores em paralelo
- Historicamente, a comunicação entre processadores é geralmente lenta em relação a velocidade individual de cada processador

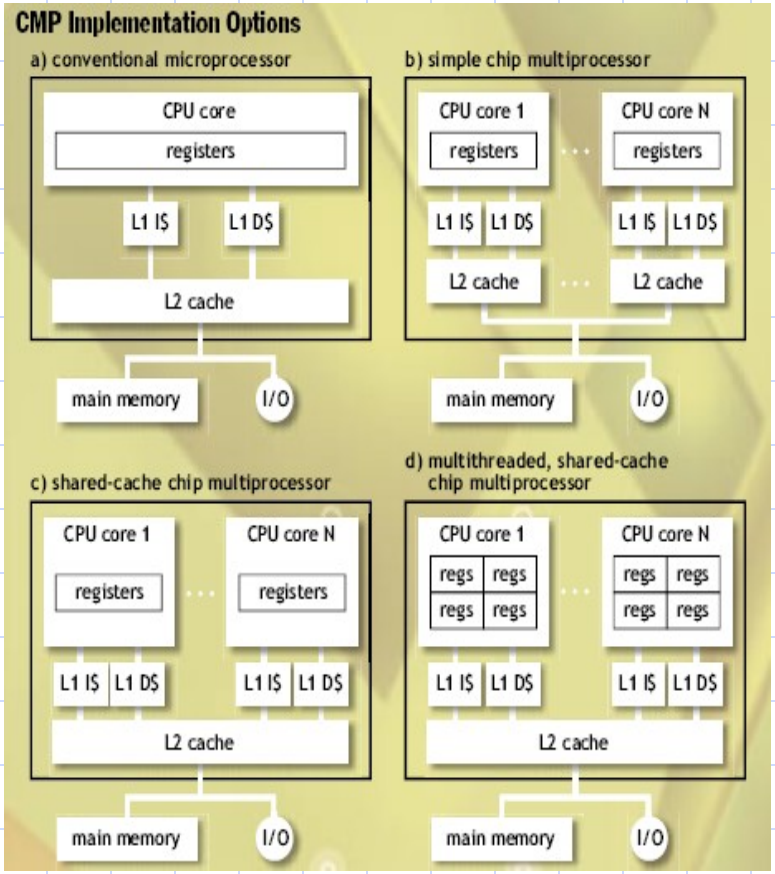
CMPs – Multithreads

- Convenientemente, a transição do sistema de múltiplos chips para o chip com multiprocessadores proporcionou uma simplificação ao tradicional problema de paralelização dos programas
- Primeiramente era necessário minimizar a comunicação entre threads independentes para um nível extremamente baixo, pois cada comunicação poderia requerer centenas ou até mesmo milhares de ciclos de processador
- Qualquer CMP com memória cache compartilhada, cada evento de comunicação toma poucos ciclos do processador. Com a latência dessa forma, os atrasos de comunicação tem muito menos impacto na performance do sistema
- Programadores ainda devem dividir seus trabalhos em threads paralelas, porém não precisam se preocupar tanto com a independência dessas threads, já que o custo de comunicação e relativamente barato

CMPs – Multithreads

- Threads paralelas podem ser bem menores e ainda continuar efetivas
  - Paralelismo pode ser extraído utilizando threads que duram de centenas a poucos milhares de ciclos com esse sistema
  - substituindo os milhões de ciclos necessários para executar threads longas em máquinas convencionais paralelas
- Pesquisas tem mostrado que a paralelização das aplicações pode ser feita de forma mais fácil com diversos esquemas envolvendo a adição de *transactional hardware* para o CMP
  - Esse sistema adiciona buffers lógicos que permite às threads tentarem ser executadas em paralelo
  - Determina dinamicamente em tempo de execução se elas estão em paralelo
  - Se não houver interdependência entre threads detectada em tempo de execução, então as threads serão completadas normalmente, se a dependência existir, então os buffers de algumas threads são esvaziados reiniciando-as e, dinamicamente, serializando-as no processo

Evolução da utilização de paralelismo nos processadore Intel



- Evolução dos microprocessadores em relação ao seu núcleo

CMPs – Vantagens

- CMPs não necessita de um esforço muito grande de engenharia para cada geração dos processadores
  - Cada família de processadores requer apenas "selar" cópias adicionais do núcleo do processador fazendo apenas algumas modificações de conexões lógicas
  - Não é necessário redesenhar a lógica do núcleo dos processadores
- O projeto das placas do sistema necessita de uma menor mudança em relação às gerações dos CMPs
  - Os modelos de CMPs mantêm a sua essência de acordo com sua evolução, aumentando praticamente somente o número de núcleos presente nos chips
  - A única real diferença é que as placas necessitam tratar a alta largura de banda de I/O exigida pelos CMPs

CMPs – **Vantagens**

- O throughput computacional pois os CMPs provê bons resultados na relação potência/performance sem nenhuma modificação nos softwares;
  - Graças ao grande número de threads independentes obitido pelas aplicações multithreads existentes
- A latência crítica computacional, para isso é necessário paralelisar a maioria das latencias criticas de software em multiplas threads para se obter uma vantagem real dos chips CMPs
  - Os CMPs tornam esse processo fácil devido a latência curta da comunicação entre os processadores