

---

# Arquiteturas de Computadores

## Multithreading

Fontes dos slides: *Curso Advanced Computer Architectures*

Prof. Marco Ferretti - Università di Pavia

(copyright Marco Ferretti )

---

---

# Intel Multithreading

- Multithreading foi introduzida pela Intel no processador Xeon em 2002, e, depois no 3,06 GHz Pentium 4, com o nome **hyperthreading**. Na realidade hyperthreading suporta somente duas threads no modo SMT.
  - Benchmarks da Intel sugerem um aumento do desempenho da CPU em 25% -- 30%.
-

---

# Intel Multithreading

- Para o sistema operacional, um processador multithread é realmente um processador duplo, com 2 CPUs compartilhando caches e RAM: se 2 aplicações podem executar de forma independente e compartilham o mesmo espaço de endereçamento, elas podem ser executadas em paralelo em duas threads.
  - Um editor de edição de filme pode usar diferentes filtros em cada frame. O código pode ser estruturado com duas threads, que processam frames pares/ímpares, e que executam em paralelo.
-

---

# Intel Multithreading

- Como duas threads podem usar a CPU concorrentemente, é necessário projetar uma estratégia que permita que as 2 threads usem efetivamente os recursos da CPU.
  - Intel usa 4 estratégias para compartilhar recursos entre threads
  - **Replicação**
    - Alguns recursos tem que ser duplicados: 2 contadores de instruções (PC) e registradores (PC). Esta replicação causa aumento de 5% na área do processador.
-

---

# Intel Multithreading

- **Particionamento**
    - Alguns recursos de hardware são particionados entre as threads. Cada thread pode usar metade dos recursos. Se aplica aos buffers para instruções de LOAD e STORE e ao ROB (retirement queue na terminologia INTEL)
    - O particionamento pode reduzir a utilização dos recursos particionados, quando uma thread não usa sua parte, que poderia ser usada por outra thread
-

---

# Intel Multithreading

- **Compartilhamento**

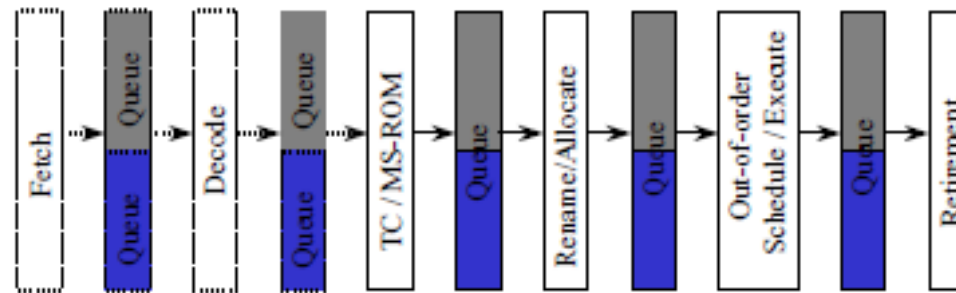
- O recurso de hardware é completamente compartilhado. A primeira thread que pega o recurso o utiliza e a outra thread espera.
  - Este tipo de gerenciamento de recurso resolve o problema devido a um recurso que não esteja sendo utilizado, caso uma thread não precise dele, pois ele pode ser alocado a uma segunda thread. Mas por outro lado uma thread pode ser desacelerada porque um recurso estará completamente alocado a outra thread
  - Por essa razão, os únicos recursos completamente compartilhados na arquitetura Intel são aqueles que estão disponíveis em grandes quantidades: para eles é pouco provável que aconteça uma situação de “starvation” , e.g., linhas de cache.
-

---

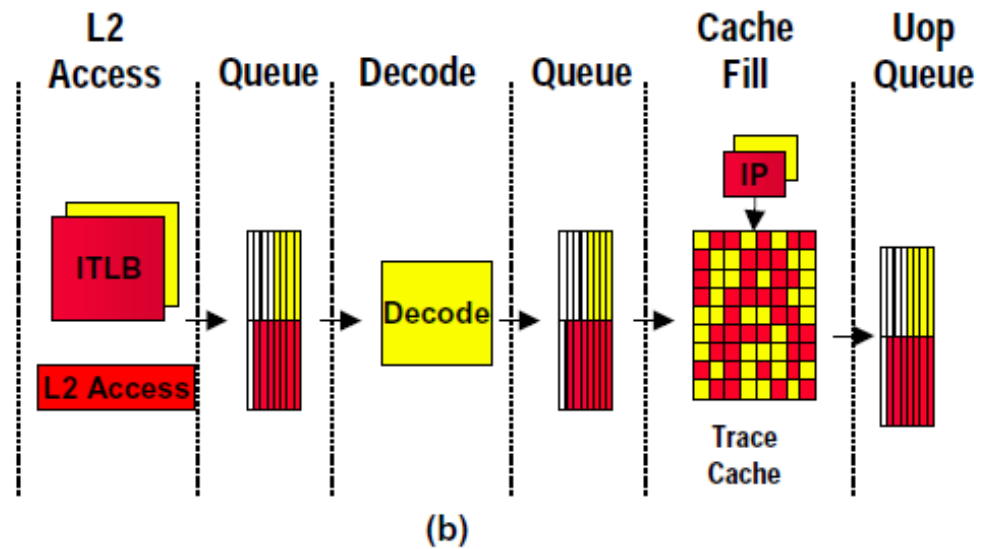
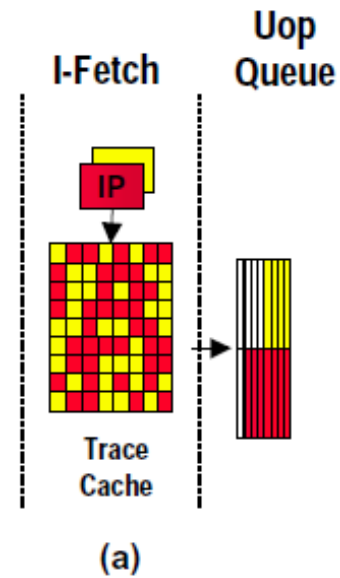
# Intel Multithreading

- **Compartilhamento limitado:**
    - Uma thread pode usar um recurso dinamicamente, até uma determinada porcentagem, de modo que uma parte permaneça disponível para outra tarefa
    - O escalonador que despacha instruções para as estações de reserva usa esta política
-

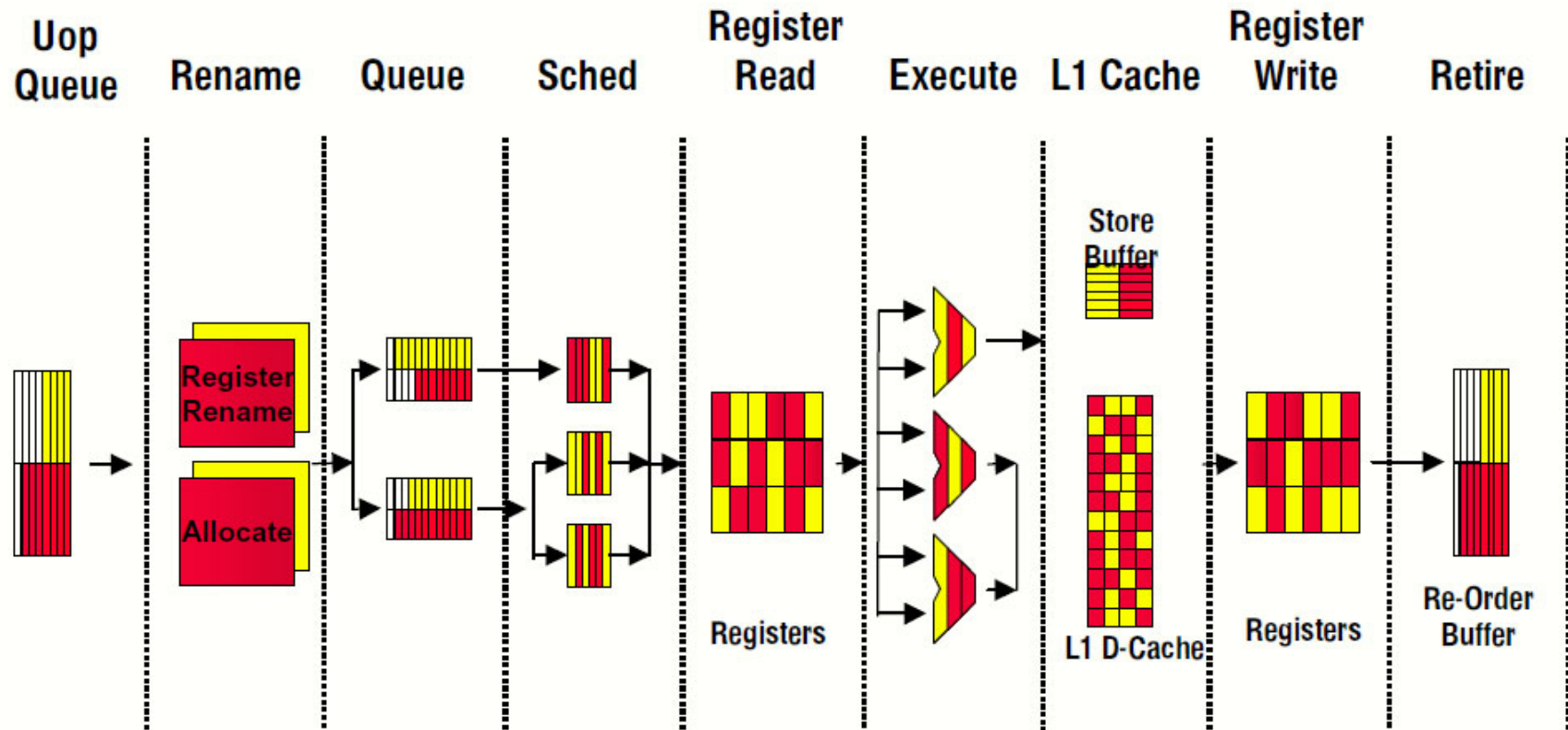
# Intel Multithreading



# Intel Multithreading

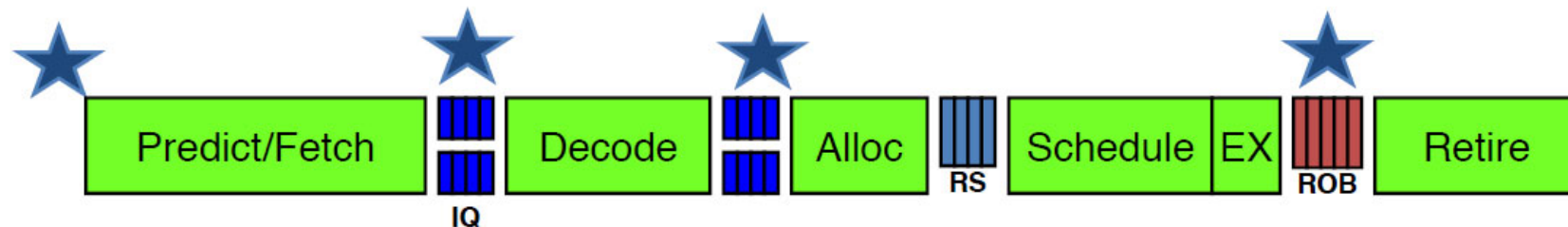


# Intel Multithreading



# Intel Multithreading

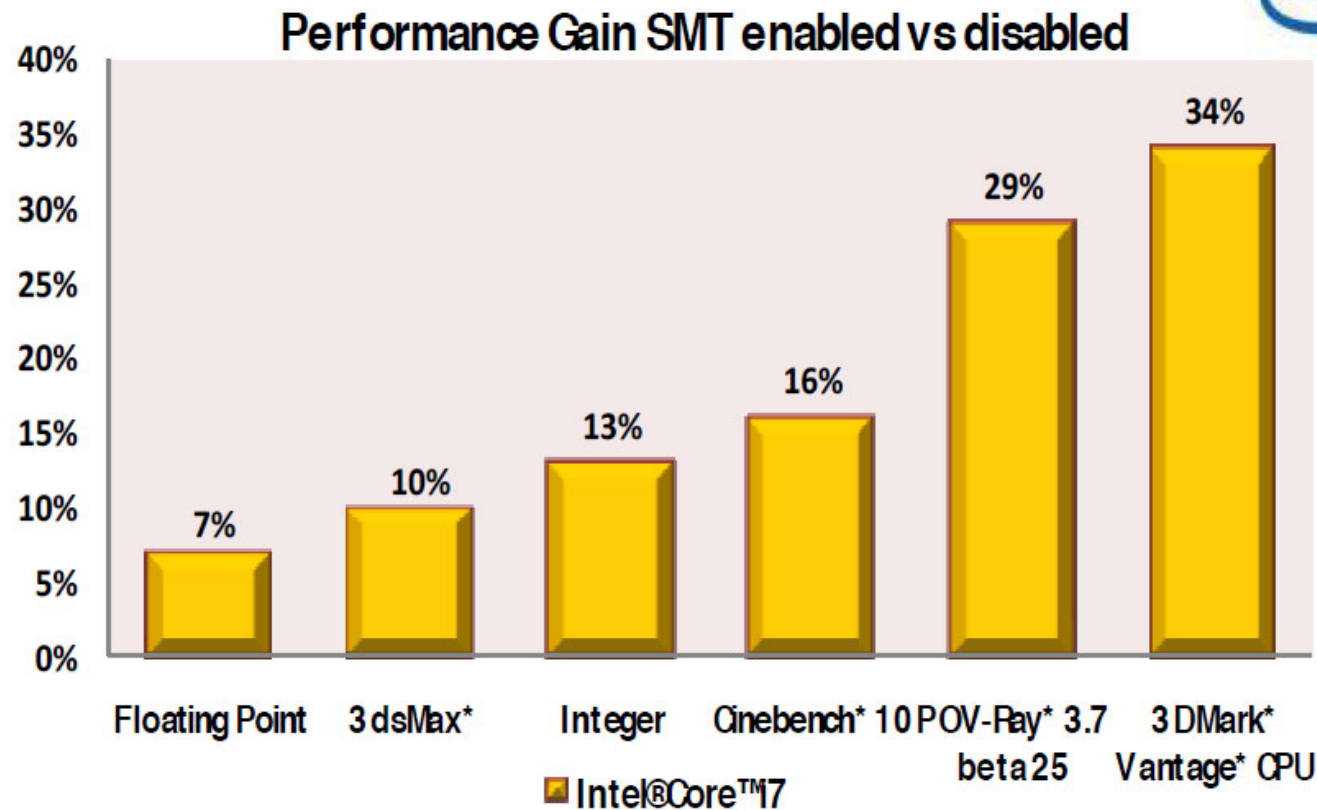
- O pipeline possui vários pontos onde a arquitetura pode escolher para trabalhar com uma thread ou 2



- Seleciona thread para a qual vai buscar instrução
- Seleciona instrução para decodificar
- Seleciona u-operação para alocar
- Seleciona instrução para retirar
- Pontos de seleção adicionais no pipeline de memória para escalonar entradas do buffer de ordenação da memória

# Intel Multithreading

## SMT Performance Chart



---

# Intel Multithreading

- Aplicação de compartilhamento limitado aplicado em tempo de execução requer monitoramento em tempo de execução dos recursos utilizados; é necessário hardware adicional e ocorrem overheads adicionais
  - Compartilhamento completo pode trazer problemas especialmente com a memória cache.
    - Compartilhar linhas de cache faz com que o gerenciamento da cache seja simples, mas o que corre se as duas threads precisam cada uma de  $\frac{3}{4}$  das linhas de cache para executar rapidamente?
      - Um número alto de cache miss, que não ocorreria se uma única thread estivesse sendo executada
-

# Intel Multithreading

- Intel temporariamente parou de utilizar tecnologia hyper-threading em processadores dual-core
- Hyper-threading foi re-introduzido em 2008 em processadores Nehalem
- Outros processadores que usam multithreading SMT:
  - IBM Power 5 (2004) dual core com 2 SMT;
  - IBM Power 7 (2010) 2-8 core com 4 SMT
  - UltraSPARC T2 (*Niagara*) and T3 (*Rainbow Fall*) possui multithreading de granulosidade fina com 8 threads por core