Lista de Exercícios

1. Boa parte do crescimento de desempenho dos microprocessadores nos últimos anos se deve principalmente ao uso de paralelismo de instruções. Explique o que é paralelismo de instruções e indique por que o crescimento de desempenho baseado exclusivamente nesse tipo de paralelismo chegou ao fim.

2. O que é **hierarquia de memória** e qual sua importância nas arquiteturas atuais?

3. O que é **localidade**? Por que ela é importante para o desempenho de sistemas computacionais? Quais novos elementos de localidade existem em processamento paralelo?

4. Considere um processador com freqüência de operação de 2GHz e um sistema de memória com um nível de cache, sendo que o tempo de acesso ao cache é de 2 ciclos, enquanto o tempo de acesso à memória é de 200 ciclos. O *hit ratio* de um programa é a fração média de acessos à memória que são servidas pelo cache. Suponha um problema que precisa fazer um total de 100 milhões de acessos à memória. Calcule qual será o tempo gasto com acessos à memória se o *hit ratio* for cada um dos seguintes valores: 0.5, 0.75, 0.8, 0.9, 0.95 e 0.98.

5. Uma característica importante para um sistema paralelo é sua **escalabilidade**. Explique o que é escalabilidade e por que ela é importante.

6. O que é **comunicação**? Por que ela é importante em processamento paralelo.

7. Por que razão o uso de chaves de interconexão *crossbar* era adequado para sistemas antigos mas não é adequado para sistemas modernos?

8. Quais as vantagens e desvantagens das chaves multiestágio em relação às chaves *crossbar*? E em relação a dutos? Não esqueça de levar em consideração em seus comentários o problema do escalamento.

9. Qual a diferença entre redes de interconexão estáticas e dinâmicas?

10. Na avaliação de redes estáticas, são usados os conceitos abaixo. Explique o que é cada um deles e qual sua importância:
    (a) Diâmetro
    (b) Largura de bissecção
    (c) Custo

11. Em termos de diâmetro, largura de bissecção e custo, compare as vantagens e desvantagens das seguintes topologias de redes estáticas:
    (a) Completamente conectada
    (b) Estrela
    (c) Árvore binária
    (d) Toro (malha com ligações nas bordas) 2D
    (e) Hipercubo

12. As máquinas de memória compartilhada evoluiram recentemente para arquiteturas **NUMA**. Explique o que é uma arquitetura NUMA e quais suas vantagens em relação a arquiteturas SMP considerando as tecnologias atuais.

13. O que é coerência de cache, e porque isso é importante em sistemas de memória compartilhada?

14. O que é o esquema de *snooping* para coerência de cache e porque ele não é utilizado em máquinas paralelas com muitos processadores?

15. Três métricas comuns de desempenho de operações são: **latência**, **largura de banda** e **custo**. Explique o significado desses termos e as relações entre os mesmos.

16. O *speedup absoluto* devido a paralelismo em P processadores $S(P)$ é definido como a razão entre o tempo de execução, T_s, do programa seqüencial e o tempo de execução, $T_p(P)$, do programa paralelo em P processadores, $S(P) = T_s/T_p(P)$. Suponha que um programa tenha uma quantidade fixa de trabalho a executar. Dessa quantidade, uma fração α (calculada em termos de tempo de execução) é estritamente seqüencial (não pode ser paralelizada) enquanto que o restante $1 - \alpha$ pode ser perfeitamente paralelizado (pode ser dividido igualmente entre P processadores, para qualquer valor de P, sem acrescentar custos adicionais). Derive uma fórmula para o *speedup* desse programa em função de α e P. Use essa fórmula para encontrar um limite superior para o *speedup* quando se dispõe de processadores à vontade. (Este resultado é a chamada *Lei de Amdahl*.)

17. Considere um *pipeline* de 10 estágios com tempo de ciclo de $10ns$. Suponha que uma aplicação alterna a execução entre duas fases, numa das quais ela envia m operações consecutivas para serem executadas no *pipeline* e na outra ela não usa o pipeline por um tempo T (dado em ns). Encontre uma expressão para o tempo de execução do programa que executa N dessas fases.

18. Para mover uma mensagem de n bytes através de H elos de ligação numa rede de armazenamento e expedição toma-se um tempo dado por $H\frac{n}{W} + (H-1)R$, onde W é a largura de banda por elo e R é o atraso de roteamento por elo. Se a rede usar o método de *wormhole routing* este tempo fica $\frac{n}{W} + (H-1)R$. Compare tempos mínimos, máximos e médios para a transmissão de uma mensagem de 64 bytes em uma malha 8×8 nos dois tipos de roteamento. Considere $W = 40MB/s$ e $R = 250ns$. Faça o mesmo para uma mensagem de 128 bytes. *(Dica: O tempo mínimo é o tempo para transmissão entre vizinho; o tempo máximo é o tempo para transmissão entre os dois nós mais distantes na rede e o tempo médio é o tempo para uma distância média na rede.)*

19. Suponha que uma linha de *cache* de 32 bytes deva ser transmitida a outro processador pela rede. Suponha ainda que o tempo de partida é de $2\mu s$ e os dados podem ser transmitidos a $100MB/s$. Qual a latência total da operação remota?

20. Suponha uma máquina com tempo de partida de $100\mu s$ e uma largura de banda assintótica de $80MB/s$. Para que tamanho de mensagem a largura de banda efetiva é a metade da largura de banda de pico? (A largura de banda efetiva é determinada pelo tamanho da mensagem dividido pelo tempo tomado em sua transmissão.)