1. Porque em algumas situações não é possível manter todos os processos da máquina na memória e por isso, os processos são enviados para uma área no disco reservada para este mecanismo (Swap), isto costuma ser um processo lento, apesar de fazer com que a máquina execute mais processos, já que gera um grande tempo de execução para os programas serem copiados do disco para a memória e vice-versa.
2. A estratégia seria selecionar um bloco de linha e coluna da matriz, pois assim assumimos que estes valores ficaram em cache para efetuar a operação de multiplicação por pelo menos 3 loops.
3. Este conceito diz a respeito à otimização; o programador deve sempre pensar em utilizar a memória cache, já que é a mais próxima do processador, portanto não é necessário efetuar a busca de dados na memória. Para que essa otimização ocorra, o programador deve verificar quais são as variáveis ou dados mais utilizados por determinado programa e fazer com que fiquem em cache, pois é uma memória com pouca capacidade de armazenamento, portanto os dados devem sem selecionados com lógica.