計算機組織 Lab5 0616223 0616059

1. Stall calculate:

內層 loop_k 有 22 個 cycle, 跑 n 次做 k:1~n 的 A[i][k] * B[k][j]。再 2 個 cycle2 到 end_k, 再 2 個到 loop_j, 再 3 個 cycle 開始 C 的下一個, 跑 p 次, C[i][1~p]。同上再做一次一樣的, 做 C[1~m]。以及一開始的兩個 cycle 做$1, $3 最後 exit 的 nop。

所以總共是 ((((22*n+2+2)+3)*p+2+2)+3)m+2+3。

2. miss penalty of 1a:

if hit : 需要 1(send address) + 2(access single cache) + 1(send word) = 4。

if miss : 需要 1(send addree) + 8(block size/word) * (1(send address) + 100(access memory) + 1(send word) + 2(access single cache)) + 2(access single cache) + 1(send word) = 836。

3. miss penalty of 1b:

if hit : 需要 1(send address) + 2(access single cache) + 1(send word) = 4。

if miss : 需要 1(send addree) + (1(send address) + 100(access memory) + 1(send word) + 2(access single cache))+ 2(access single cache) + 1(send word) = 108。

4. miss penalty of 1c:

if L1 hit : 需要 1(send address) + 1(access L1 cache) + 1(send word) = 3。

if L2 hit L1 miss : 需要 1(send address) + 4(block size/word) * (1(send address) + 10(access L2 cache) + 1(send word) + 1(access L1 cache))+ 1(access L1 cache) + 1(send word) =55。

if memory hit L1 miss L2 miss : 需要 1(send address) + 32 (block size/word) * (1(send address) + 100(access memory) + 1(send word) + 10(access L2 cache)) + 4(block size/word) * (1(send address) + 10(access L2 cache) + 1(send word) + 1(access L1 cache)) = 3639。

5. Result:

a, b 的差別在 bus 的寬度不一樣, 所以傳送 data 速度不一樣, 因此就算他們的 miss rate 一樣, 但因 b 的 miss penalty 比較少, 所以 b 的 efficiency 會比較好。那至於 c, 因為有兩層 cache, 第一層的 block size 是 4 word, 第二層是 32word, 兩個 block size 的不同的原因在於第一層 cache 的主要目的在於要 minimize hit 所花時間, 第二層的主要目的是要將低 miss rate。雖然兩層 cache 的設計在大部分的情況下都可以很有效的減少 access memory 發生的機率, 但由於 L2 跟 memory 之間傳送資料的速度沒有增加, 而 L2 的 block size 相較於 a,b 大的許多, 所以造成 miss penalty 更高, 因此結果不一定會比 a, b 還要好。

從 execution cycle 跟 a b c 三種 memory stall cycle 的數量, 相比之下, 反而不一定是程式執行中最花時間的一部分, 因此提高 CPU 的效能並不一定是提昇電腦效能唯一的辦法, 有時候善用 memory hierarchy 並提高 cache 的 bus 反而能有比提昇 CPU 效能更好的成果。