

a. 最有效的方法為使 write buffer entry 大小 = L2 cache 一次能接收的 data 量, 這樣可以使每次傳輸都充分利用 L2 cache 的 write data bus, 避免效率低。

ans. 16 B wide

b. 64 bits = 8 B

non-merging: 每次 8 B stores 就會進入 write buffer 作為獨立的 entry, 並觸發 L2 cache 寫入 8 B (浪費 8 B)

merging: 兩個 8 B stores 存到一個 16 B block, 並觸發 L2 cache 寫入 16 B, 減少對 L2 cache 的壓力。

可以發現要寫 2 個 8 B 時, non-merging 要 2 次 L2 cache 寫入, merging 只需 1 次 L2 cache 即可, 因此 speedup = 2

c. blocking cache 時, L1 miss 會 stall the processor, 因此需要更多 write buffer 來儲存 pending writes. 如果 buffer 太小容易 overflow, 導致進一步在寫入時 stall。

non-blocking 下, 可以 handle cache misses 而不用 stall the processor, 讓 write buffer 可以有更好的利用。

可以得知, blocking 下 write buffer 壓力較大, 因此所需較多 non-blocking 則不需那麼多的 write buffers。

更: c 的 blocking write buffer 不需改變