

Coherency problem / Cache Write Policies

Coherence between a cache word and its copy in the main memory should be maintained at all times, if at all possible by using a number of policies (techniques) to perform write operations to the main memory blocks while residing in the cache. These policies determine the degree of coherence that can be maintained between cache words and their counterparts in the main memory.

In the following paragraphs, we discuss two main cases: cache write policies upon a cache hit and the cache write policies upon a cache miss. We also discuss the cache read policy upon a cache miss. Cache read upon a cache hit is straightforward.

A) Cache Write Policies Upon a Cache Hit

There are essentially two possible write policies upon a cache hit, These are:

1) write-through :

every write operation to the cache is repeated to the main memory at the same time. The write-through policy maintains (تحفظ) coherence between the cache blocks and their counterparts in the main memory at the expense of the extra time needed to write to the main memory. This leads to an increase in the average access time.

2) write-back policy:

all writes are made to the cache. A write to the main memory is postponed until a replacement is needed. Every cache block is assigned a bit, called the dirty bit, to indicate that at least one write operation has been made to the block while residing in the cache. At replacement time, the dirty bit is checked:

if dirty bit==1 then

the block is written back to the main memory

else

the block is simply overwritten by the incoming block.

The write-back policy eliminates the increase in the average access time. However, coherence is only guaranteed at the time of replacement.

B) Cache Write Policy Upon a Cache Miss

Two main schemes can be used:

1) write-allocate :

the main memory block is brought to the cache and then updated.

2) write-no-allocate:

The missed main memory block is updated while in the main memory and not brought to the cache.

In general, write-through caches use write-no-allocate policy while write-back caches use write-allocate policy.

C) Cache Read Policy Upon a Cache Miss

Two possible strategies can be used:

1) the main memory missed block is brought to the cache while the required word is forwarded immediately to the CPU as soon as it is available.

2) the missed main memory block is entirely stored in the cache and the required word is then forwarded to the CPU.