Cache Memories

- The cache is a small and very fast memory, interposed between the processor and the main memory.
- Its purpose is to make the main memory appear to the processor to be much faster than it actually is.
- The effectiveness of this approach is based on a property of computer programs called locality of reference.
- Analysis of programs shows that most of their execution time is spent in routines in which many instructions are executed repeatedly.
- These instructions may constitute a simple loop, nested loops, or a few procedures that repeatedly call each other.
- The actual detailed pattern of instruction sequencing is not important—the point is that many instructions in localized areas of the program are executed repeatedly during some time period.
- This behavior manifests itself in two ways: temporal and spatial. The first means that a recently executed instruction is likely to be executed again very soon.
- The spatial aspect means that instructions close to a recently executed instruction are also likely to be executed soon.

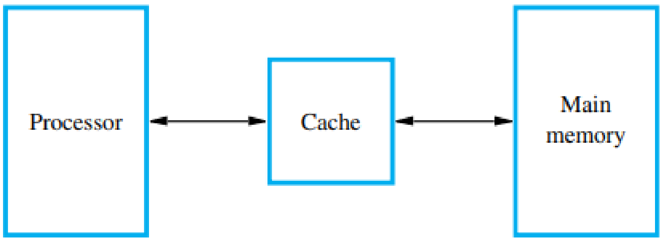

Figure 8.15 Use of a cache memory.

Conceptually, operation of a cache memory is very simple.

- The memory control circuitry is designed to take advantage of the property of locality of reference.
- Temporal locality suggests that whenever an information item, instruction or data, is first needed, this item should be brought into the cache, because it is likely to be needed again soon. Spatial locality suggests that instead of fetching just one item from

the main memory to the cache, it is useful to fetch several items that are located at adjacent addresses as well.

- The term cache block refers to a set of contiguous address locations of some size. Another term that is often used to refer to a cache block is a cache line.
- Consider the arrangement in Figure 8.15. When the processor issues a Read request, the contents of a block of memory words containing the location specified are transferred into the cache.
- Subsequently, when the program references any of the locations in this block, the desired contents are read directly from the cache. Usually, the cache memory can store a reasonable number of blocks at any given time, but this number is small compared to the total number of blocks in the main memory.
- The correspondence between the main memory blocks and those in the cache is specified by a mapping function. When the cache is full and a memory word (instruction or data) that is not in the cache is referenced, the cache control hardware must decide which block should be removed to create space for the new block that contains the referenced word.
- The collection of rules for making this decision constitutes the cache’s replacement algorithm.

Cache Hits

- The processor does not need to know explicitly about the existence of the cache. It simply issues Read and Write requests using addresses that refer to locations in the memory.
- The cache control circuitry determines whether the requested word currently exists in the cache.
- If it does, the Read or Write operation is performed on the appropriate cache location. In this case, a read or write hit is said to have occurred. The main memory is not involved when there is a cache hit in a Read operation.
- For a Write operation, the system can proceed in one of two ways. In the first technique, called the write-through protocol, both the cache location and the main memory location are updated.
- The second technique is to update only the cache location and to mark the block containing it with an associated flag bit, often called the dirty or modified bit. The

main memory location of the word is updated later, when the block containing this marked word is removed from the cache to make room for a new block.

- This technique is known as the write-back, or copy-back, protocol.

Cache Misses

- A Read operation for a word that is not in the cache constitutes a Read miss. It causes the block of words containing the requested word to be copied from the main memory into the cache.
- After the entire block is loaded into the cache, the particular word requested is forwarded to the processor. Alternatively, this word may be sent to the processor as soon as it is read from the main memory.
- The latter approach, which is called load-through, or early restart, reduces the processor's waiting time somewhat, at the expense of more complex circuitry. When a Write miss occurs in a computer that uses the write-through protocol, the information is written directly into the main memory.
- For the write-back protocol, the block containing the addressed word is first brought into the cache, and then the desired word in the cache is overwritten with the new information.