Introduction

The advances expected in memory technology in the '70s will have a significant impact on computer system organization. While improvements can be expected in all phases of memory technology, three specific advances have been selected and some of the implications of these advances will be described. The three memory technology advances to be considered are:

- Associative memories will be available at 2 to 5 times the cost of random-access memories.
- The cost per bit of small-capacity random-access memories will be greatly reduced.
- Mass memories will be available that allow access to a block in a few microseconds.

some effects of
ADVANCES IN MEMORY SYSTEM TECHNOLOGY ON COMPUTER ORGANIZATION

D. C. Gunderson
Honeywell Inc.
St. Paul, Minnesota

* D. C. Gunderson, Honeywell Inc., 2345 Walnut Street, St. Paul, Minnesota 55113.
Associative Memories in Future Computer Systems

After many years of failing to live up to expectations, associative memories finally appear to be ready to make an impact on computer system design. This is primarily due to the advances made in semiconductor technology. A number of semiconductor associative memory designs have recently been described in the literature. [1,2] Using large-scale integration (LSI), associative memories can be projected to 2 to 5 times the cost of random-access memories in the mid '70s. We might expect these memories to fall into two classes according to speed:

- 30 to 50 nsec (For each of the basic operations of read, write, or equality search.)
- 300 to 500 nsec (For each of the basic operations of read, write, or equality search.)

This is because some applications require a speed roughly 10 times as fast as the main memory, while others can be satisfied with associative memory operation times approximately the same as the main memory access time.

Small associative memories, on the order of 50,000 bits or less, with speeds such as defined above will become a significant part of the memory hierarchy in computers of the '70s. Two specific areas of application will be described. These are the associative cache memory, and associative memories as part of a "hardware executive."

Associative Cache Memory

The use of a high-speed small cache (buffer) memory between the processor and main memory (as shown in Fig. 1) has been described by a number of authors.[3,4] Figure 1 shows the cache memory concept being used in a multiprocessor configuration simply because the need for reducing main memory accesses is greater in a multiprocessor, where several processors share a common main memory, than in a single processor system. The cache memory must be associative since main-memory addresses are used to retrieve words from the much smaller cache memory. Briefly, the operating procedure is as follows. Each time a word is obtained from main memory for use by the processor it is written into the associative memory, as is the main memory address. This is done because it is anticipated that the word may be needed again soon. Also prior to going to main memory, the processor will perform a search in the cache memory (using the main memory address as a search word) to see if the desired word is in the cache. If it is, an appreciable amount of time is saved because the search and read from the associative memory is several times faster than the main-memory access time. Results of computations that are to be written into memory are also written into the cache memory rather than into the main memory. This produces a time saving if the result is simply a temporary result that is to be used again soon. If it is a final output of a computation, it will eventually be bumped out of the cache and it will then be written back into the main memory.

The cache memory operation described above is strictly a "look behind" mode of operation. A surprisingly large percentage of accesses are from the cache memory rather than from the main memory for a relatively small-sized cache memory. For example simulation results described in Reference 3 show that for a cache of only 128 words, an overall speed gain of three is realized. An additional improvement can be realized by adding some "look-ahead" capability. Simulation results described in Reference 4 showed that from 4 to 8 words is the appropriate size block to be moved from main memory to the cache memory whenever a main-memory access is required.

The cache memory application is one where AMs will be used extensively. The requirement would call for a capacity of a few hundred words with an operation time of approximately 50 nsec. This would give the cache memory a 100-nsec access time since each access would require a search followed by a read operation.
**Hardware Executive**

Another area of application where associative memories are receiving considerable attention is "hardware executives." A diagram of a "hardware executive" in a multiprocessor system is shown in Fig. 2.

![Diagram of Hardware Executive](image)

By using hardware in the form of associative memories to store files such as the "page swapping" file and the "task" file, the software executive routines can be greatly simplified and executive functions can be performed much more rapidly. Thus, this can lead to more efficient and more flexible computer systems. While the use of such hardware can be effective in virtually any type of computer system, the greatest need and the most likely area of successful application is the multiprocessor.

Two of a number of executive functions to which AMs can be applied are shown:
- Control of paging operations
- Task assignment or scheduling

The page control function is that of controlling the flow of pages (blocks of instructions or data) between main memory and mass memory. In such a system, pages are brought in on an "as-needed basis" and completed pages are bumped out of main memory when they are no longer needed. Particularly when the system is operating in a time-sharing or multiprogramming mode, the page-swapping requirements become very complex and page-swapping decision making can become a real burden. An associative memory for page control might be used to store the identity or description of each page for which there is some need (a need of having it in main memory). Depending on criteria such as program priority, priority of the user, and schedules, a priority can be generated for each of these pages. With each word of AM used to store the description of one page, an operation that can be performed quickly in the AM is a comparison of page priorities to determine which pages are to be in main memory. Using an AM in this way is not only effective for the currently-used "demand" paging schemes, but it might also make it feasible to go to more elaborate schemes such as "look-ahead" paging which would allow high-priority programs to be handled more effectively.

The task assignment function is another case where a rapid comparison of priorities is required. In this case it can be assumed that each word of AM contains the description of a task or program segment and that each of these tasks has a priority associated with it which is dependent on the same things that page priority was dependent on. Thus each time a processor completes its current assignment, it will search through the task file to obtain the highest priority task ready to be executed. Again the associative memory is simply a convenient place to keep a file which is constantly changing and from which items must be retrieved rapidly in a particular order.

The speed requirement in the executive applications calls for an AM with 300 to 500 nsec operation time. Also simulation investigations\[^5,6\] in both the task assignment and page control applications have shown that on the order of 500 words is sufficient even for the very largest systems.
The Effects Due to Small-Capacity RAMs at Low Cost

The semiconductor technology is also making it possible to mechanize very small random-access memories at a considerably lower cost than was possible using magnetic technology. This is likely to have an impact on advanced computer organization in a number of ways. One of these is that it will make parallel-processor (array-processor) organizations more popular; another is that it will favor the distributed computer system organization as compared with the centralized one.

The parallel processor organization is shown in Fig. 3. A single-control unit which includes program memory controls a column or array of processing elements, each of which contains arithmetic capability and a small amount of data memory. This type of organization has previously been described as the single-instruction stream, multiple-data-stream type. In some systems the amount of data memory desired in each processing element may be as little as 16 words. In general the need is for tens of words or hundreds of words which will be mechanized for pennies per bit using the semiconductor technology.

This will make the parallel processor a much more cost-effective approach for solving those problems that have multiple data sets on which the same computational operation is to be performed. Problems such as radar data processing, signal processing, air traffic control, and data acquisition are receiving a lot of attention in the application of parallel-processing techniques.\(^7,8\)

The availability of low-cost buffer memories will be the main cause of a trend toward decentralized computer systems. The "more computing power for the dollar" argument that has made computer systems centralized in the past, will simply not hold to nearly as great an extent in the '70s. Low-cost buffer memories, low-cost read-only memories, and lower-cost special-purpose logic all contribute to this. The main incentive on moving toward less centralization is the reduced data flow in the system. This is particularly important in avionics systems, where high-data flow means more interconnection and more wires. It is also important in ground-based time-shared computer systems where reduced data flow reduces the communications costs.

---


The Implication of Rapid-Access Mass Memories

Mass memory improvements will also have a significant effect on computer system design in the '70s. Solid-state mass memories will be available in the $10^4$ to $10^6$ bit capacity range that will have greatly improved access times and transfer rates compared with current mass memories. Access times in microseconds will be possible and word transfer rates will be as high as 10 megahertz. Such memories will replace rotating mass memories in military and space systems. However, since it will be some time before they will be able to compete with disc files from a cost point of view, they may simply represent another memory in the hierarchy in commercial systems.

One way a solid-state block oriented random-access memory (called BORAM in the past) may be used in a multiprocessor computer system is shown in Fig. 4. In this case the BORAM has virtually replaced the random-access memory for program storage. Features of this system are as follows:

- All programs and large data files are stored in the BORAM.
- Program segments are block transferred from the BORAM to a small random-access memory in each processor for execution. The transfer of program segments from the BORAM to the processor memory is fast because BORAM access time is in microseconds and the word transfer rate is at 10 megahertz. This implies fast memory (perhaps semiconductor) in each processor. This kind of speed in both access time and transfer rate will allow the program segments to be relatively short (perhaps 2000 words).
- A centralized data memory is provided for storage of results that are generated from one program segment and which must serve as inputs to another.

This type of multiprocessor system organization can be expected to have cost advantages compared to a conventional multiprocessor. Cost savings would occur in two areas: reduction in the amount of random-access memory in the system, and a simplified switching requirement between processors and memory. This organization is similar to one of the versions of the Navy's Advanced Avionics Digital Computer now under development.[9]

Summary

A few of the memory system advances expected to have an impact on computer system design in the '70s have been considered. Advances in associative memories, random-access memories, and mass memories were identified. These advances give the computer system designer a new set of memory building blocks with which to work. A number of alternatives to computer system organization using their new building blocks are suggested.

REFERENCES