Advanced Computer Architecture (0630561)

Lecture 12

Memory Systems in Pipelined Processors

Prof. Kasim M. Al-Aubidy

Computer Eng. Dept.

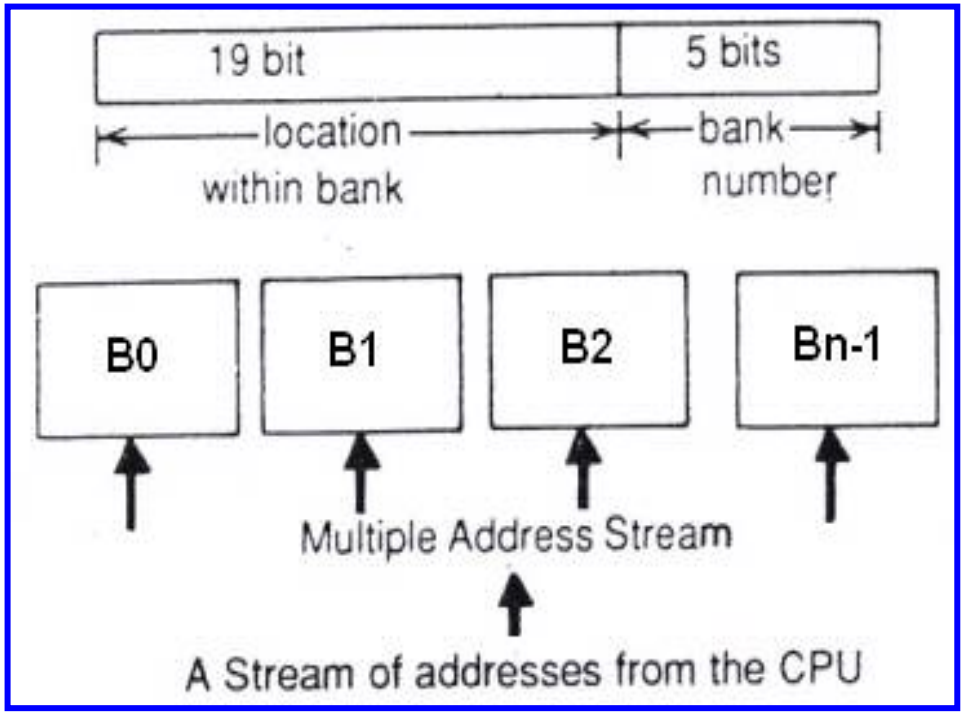
Interleaved Memory:

- In a pipelined processor data is required every processor clock cycle.
- Memory system usually is slower than the processor and may be able ti deliver data every n processor clock cycles. To overcome this limitation, it is necessary to operate n memory units in parallel to maintain the bandwidth match between the processor and memory.
- Bandwidth is defined as a numbers of bits that can be transferred between two units every second.
- The performance gain or speed up of the memory system is defined as:

$$\text{speedup} = \frac{\text{Throughput of the interleaved system}}{\text{Throughput of the single bank system}}$$

A good approximation to the speed up is given by:

$$speedup = \sqrt{(\pi(n/2))} - 0.28$$

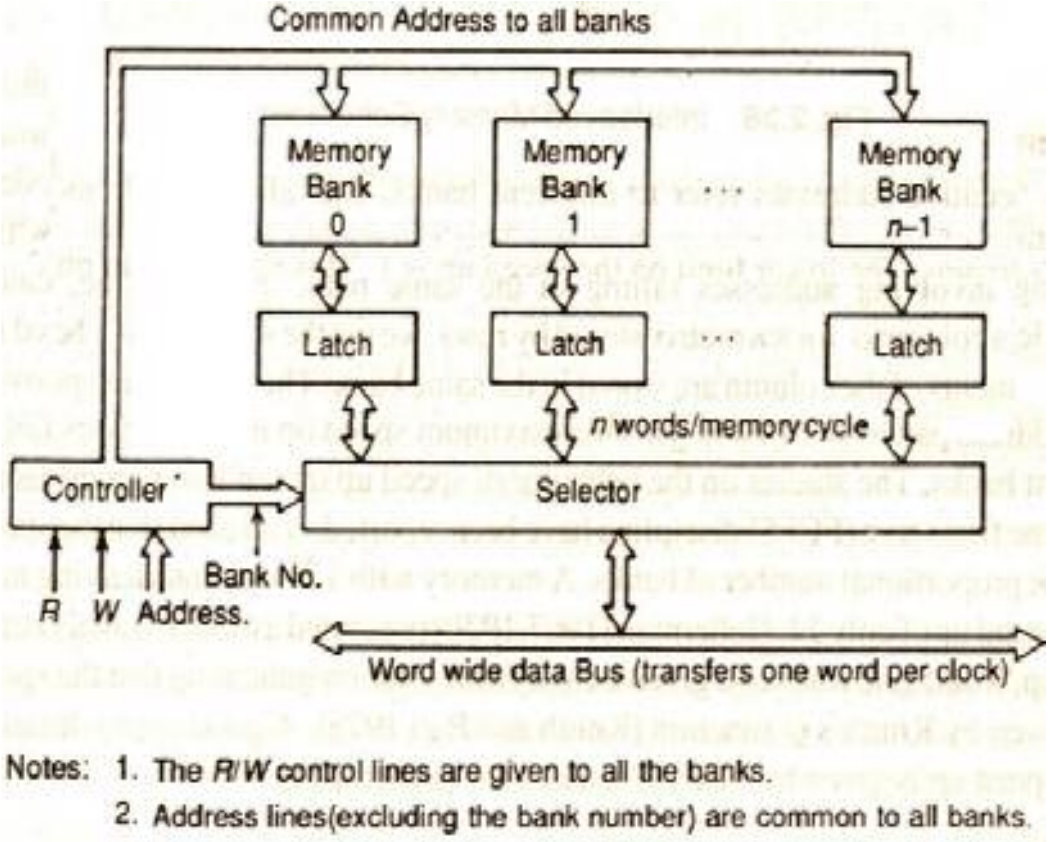


Memory Access:

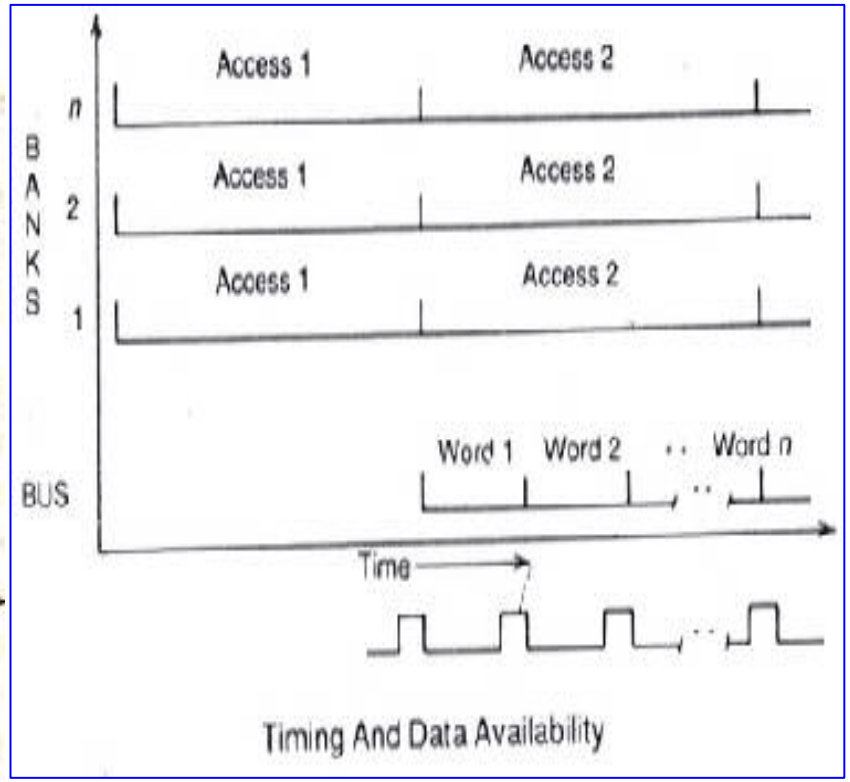
It is possible to organize the memory system on two basic forms:

1. **Synchronous access Organization:**
2. **Asynchronous access Organization:**

Synchronous access Organization:



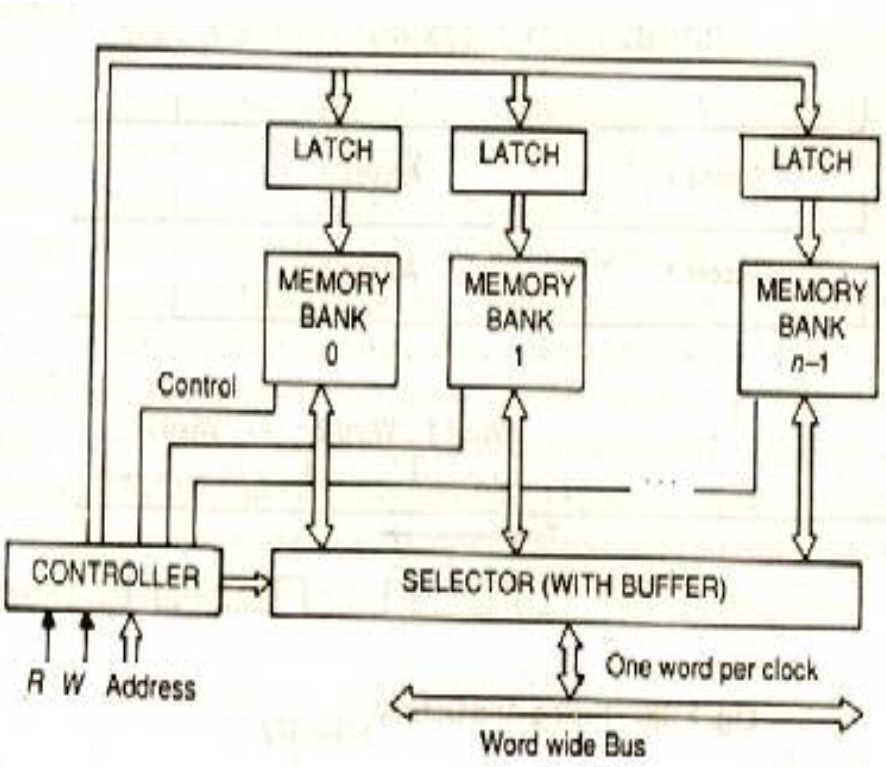
Notes: 1. The R/W control lines are given to all the banks.
2. Address lines(excluding the bank number) are common to all banks.

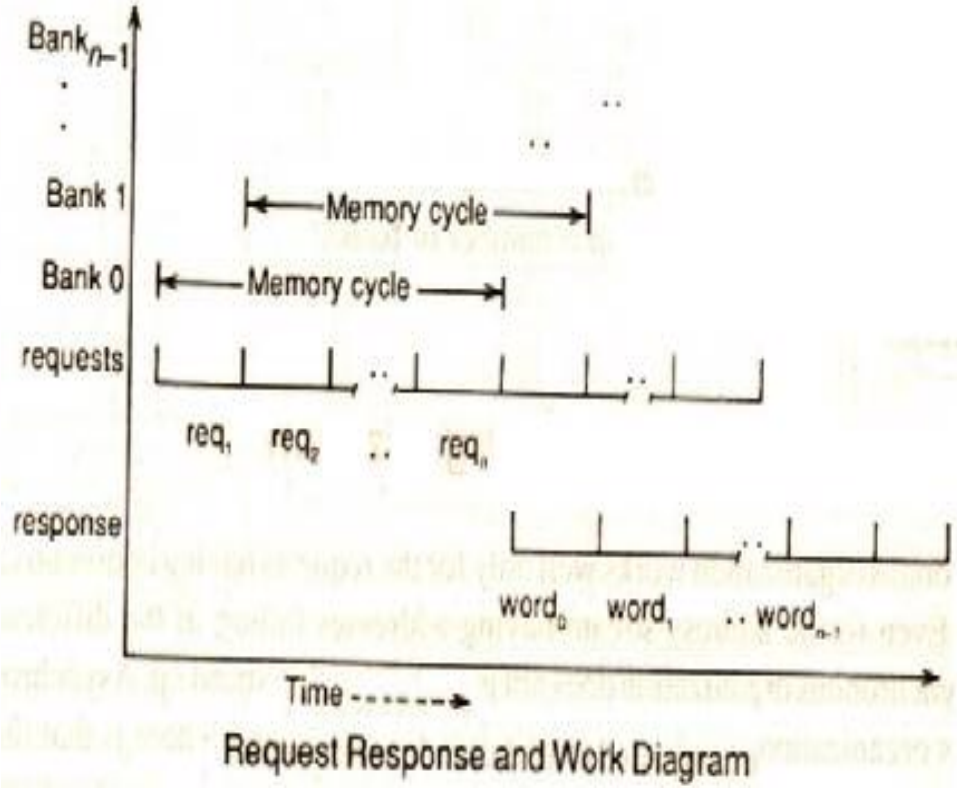


Timing And Data Availability

Asynchronous access Organization:

- It works well only for the requests having consecutive addresses.
- The address latches are provided to all the banks. This allows individual banks to hold addresses of the requests being served allowing them to carry out their memory cycles independently.





Synchronous or Asynchronous Access Organization?

Memory system performance is characterized by the following theorems:

- For addresses spaced at a distance of m, the average data access time (t) per word access in synchronous organization is:

$$t = m.T/n \quad \text{for } m << n$$

$$t = T \quad \text{for } m >> n$$

where T= bank cycle time and n= number of banks.

- The average data access time (t) per element with requests spaced m addresses apart for asynchronous organization is:

$$t = \gcd(m.n).T/n$$, where T= memory cycle.

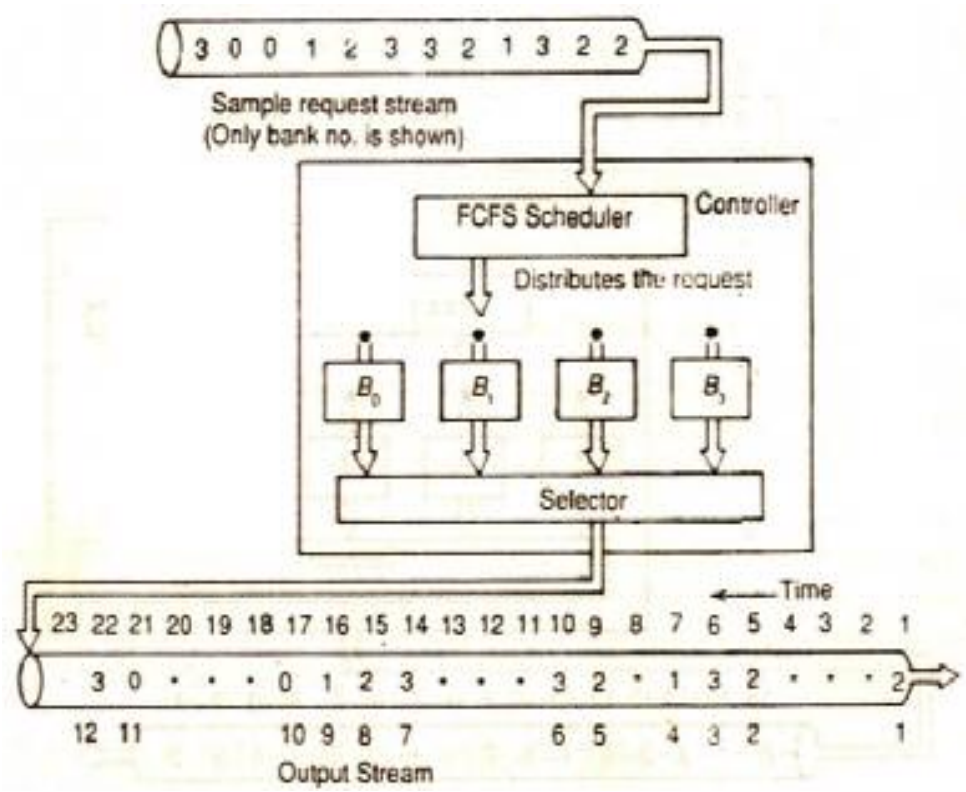
- If n and m are relatively prime then the data output rate is T/n for asynchronous memory organization, where T= memory cycle.

➢ The asynchronous memory system performs better than synchronous memory system for most of the time.

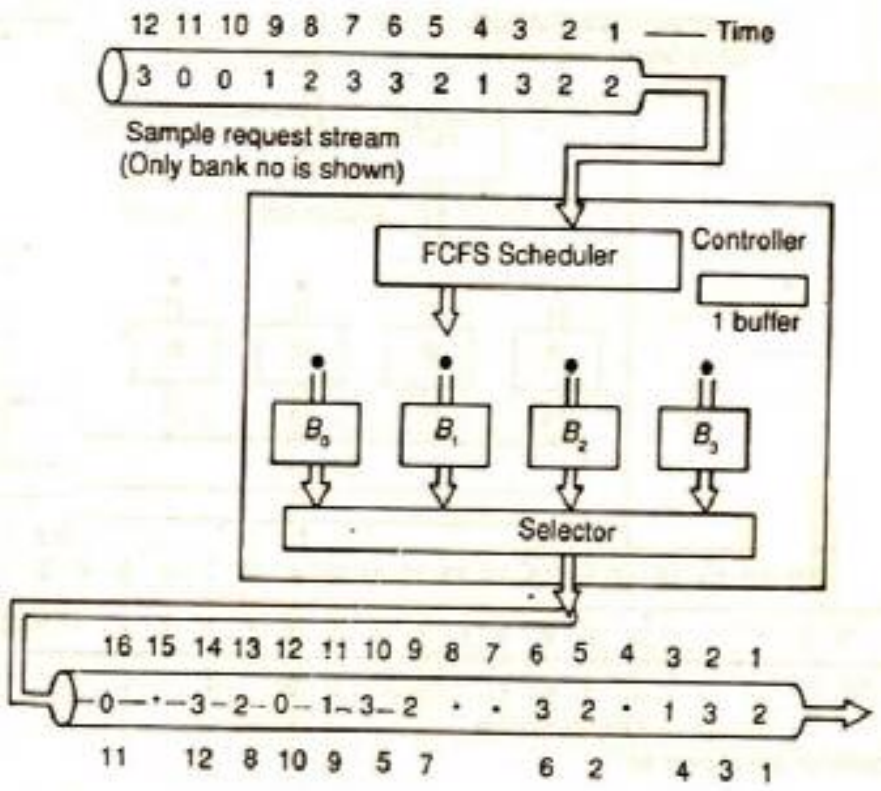
Performance Enhancement:

- Average performance of an interleaved memory grows with the square root of the degree of interleaving.
- The low performance occurs due to the loss of the time slots.

Example: consider a 4-way interleaved memory with FCFS schedule. A simple request queue with 12 memory requests.



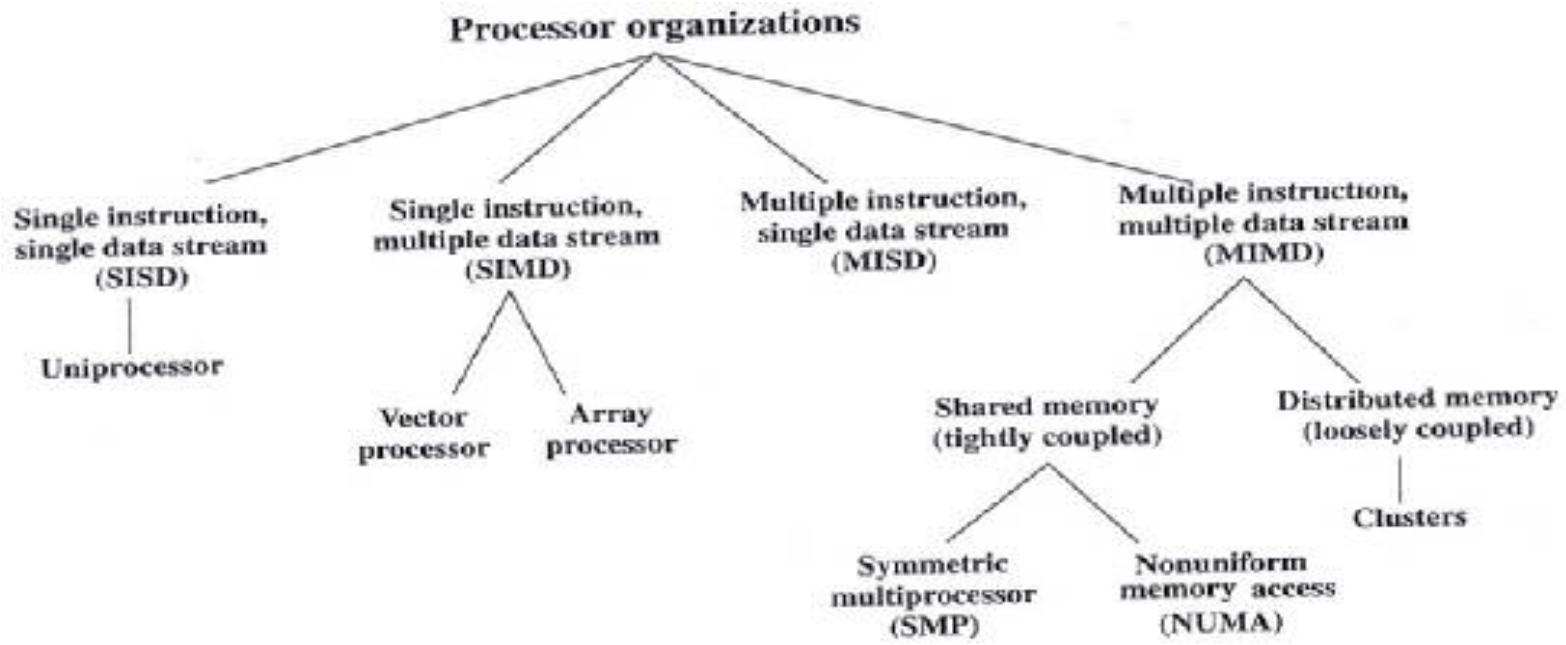
The memory gives 12 outputs in 22 clocks.



The memory gives 12 outputs in 16 clocks.

Multiple Processor Organizatios:

- A traditional way to increase system performance is to use multiple processors that execute in parallel.
- The most common multiple-processor organizations are;
  1. Symmetric multiprocessors (SMPs),
  2. Clusters, and
  3. Nonuniform memory access (NUMA).



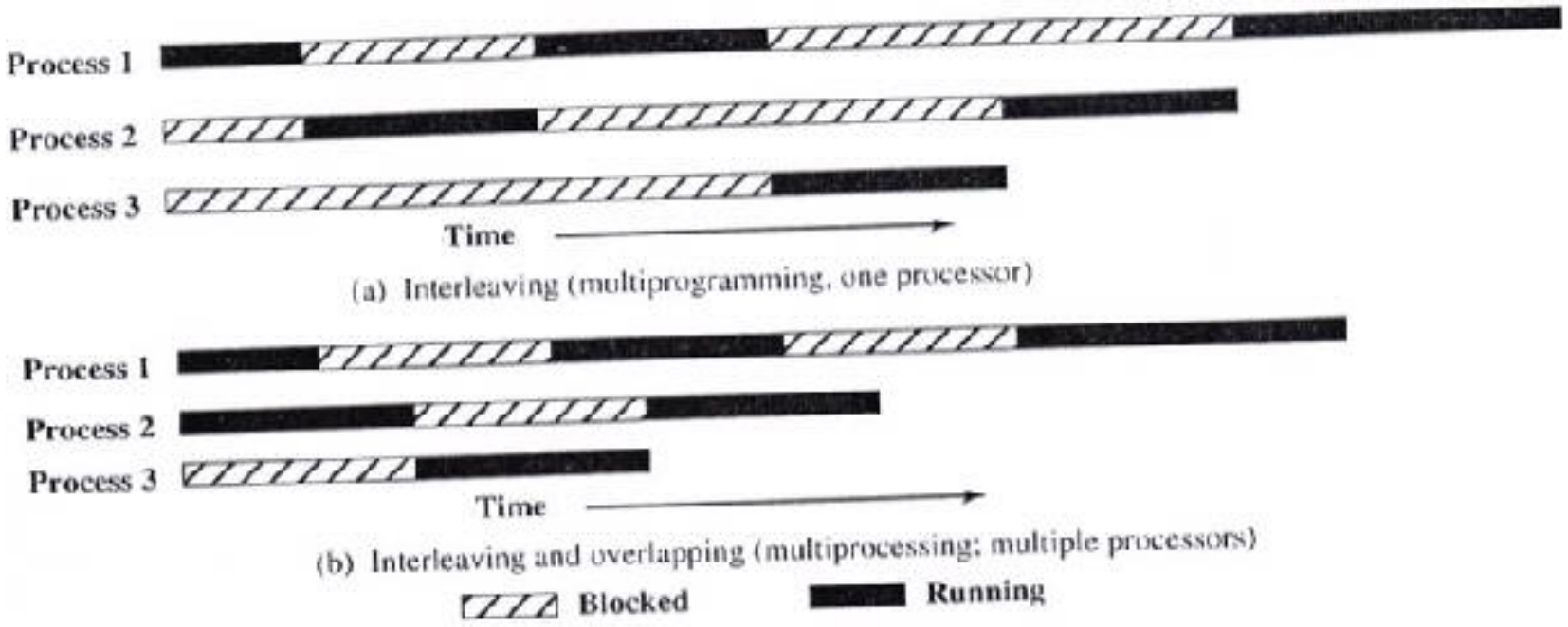
Symmetric Multiprocessors (SMPs):

- The term SMP refers to a computer hardware architecture and also to the OS behavior that reflects that architecture. SMP has the following characteristics:
  1. There are two or more similar processors.
  2. These processors share the same main memory and I/O facilities and are interconnected by a bus or other Ins, such that memory access time is approximately the same for each processor.
  3. All processors share access to I/O devices, either through the same channels or through different channels.
  4. All processors can perform the same functions.
  5. The system is controlled by an integrated OS that provides interaction between processors and their programs at the job, task, file, and data element levels.
  6. The OS of an SMP schedules processes or threads across all of the processors.

SMP Advantages:

- An SMP organization has a number of potential advantages over a uniprocessor organization, including the following:

1. **Performance:** if some portions of the work can be done in parallel, then a system with multiple processors will yield greater performance than one with a single processor of the same type.



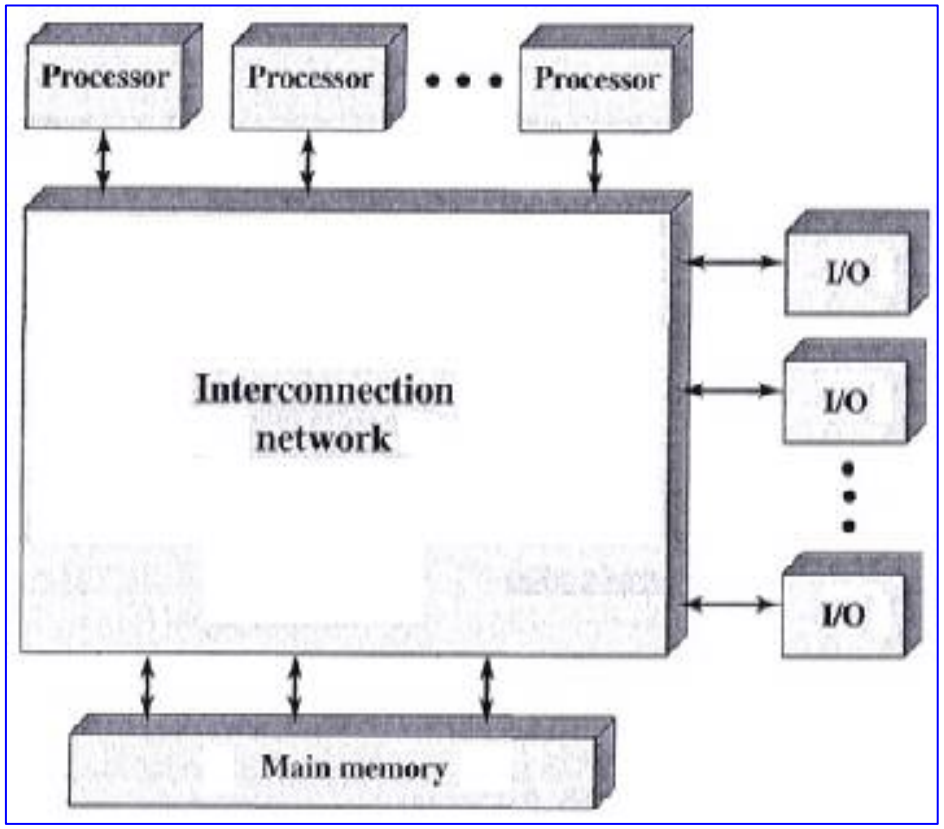
SMP Advantages: (cont.)

2. **Availability:** in an SMP, because all processors can perform the same functions, the failure of a single processor does not halt the system. Instead, the system can continue to function at reduced performance.
3. **Incremental growth:** a user can enhance the system performance by adding an additional processor.
4. **Scaling:** vendors can offer a range of products with different price and performance characteristics based on the number of processors in the system.

Organization of a Multiprocessor System:

1. **Tightly Coupled Multiprocessor:**

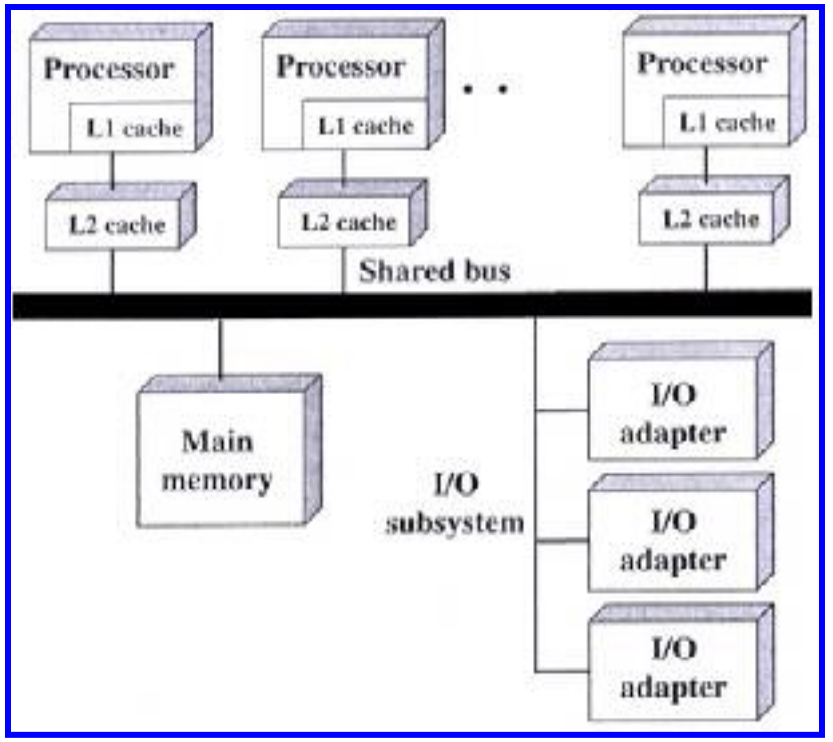
- There are two or more processors.
- Each processor is self-contained, including a control unit, ALU, registers, and one or more levels of cache.
- Each processor has access to a shared main memory and I/O devices through interconnection mechanism.
- The processor can communicate with each other through memory (messages and status information left in common data areas).



2. Time-Shared Bus Organization:

- It is the most common organization for PCs, workstations, and servers.
- It is the simplest mechanism for constructing a multiprocessor system.
- The bus consists of data, address, and control lines.
- To facilitate DMA transfers from I/O processors, the following features are provided:

- **Addressing:** it must be possible to distinguish modules on the bus to determine the source and destination of data.
- **Arbitration:** any I/O module can temporarily function as “master”. A mechanism is provided to arbitrate competing requests for bus control using priority scheme.
- **Time-sharing:** when one module is controlling the bus, other modules are locked out and must suspend operation until bus access is achieved.

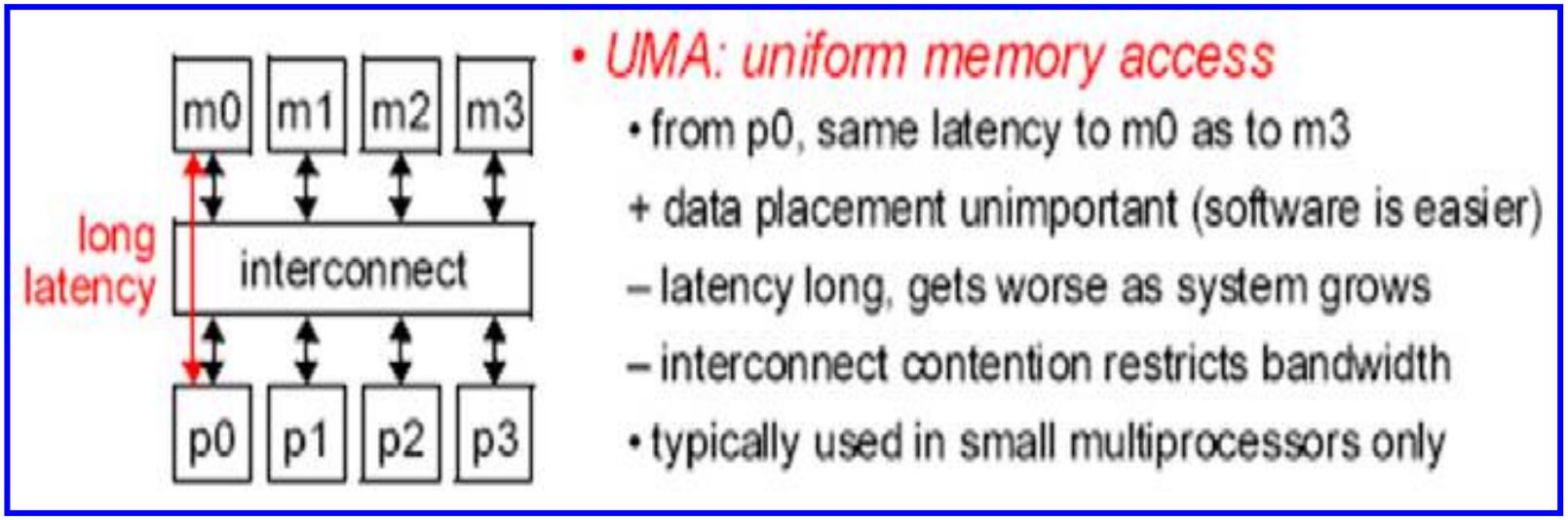


2. Time-Shared Bus Organization: (Cont.)

- The time-shared bus organization has several attractive features:
- **Simplicity:** the physical interface and the addressing, arbitration, and time-sharing logic of each processor remain the same as in a single-processor system.
- **Flexibility:** easy to expand the system by attaching more processors to the bus.
- **Reliability:** the bus is essentially a passive medium, and the failure of any attached device should not cause failure of the whole system.
- The main drawback to the bus organization is performance. All memory references pass through the common bus.
- To improve performance, it is desirable to equip each processor with a cache memory. Typically, workstations and SMPs have two levels of cache, and some processors now employ a third level of cache as well.

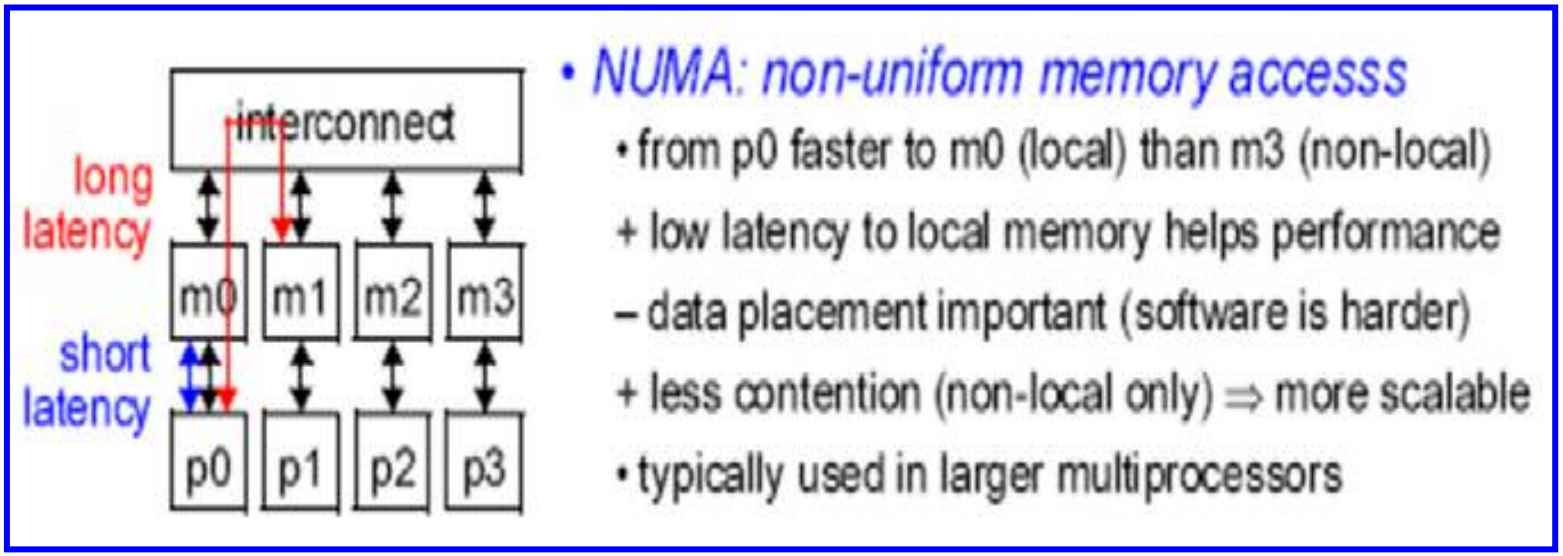
Uniform Memory Access (UMA):

- All processors have access to all parts of main memory using loads and stores.
- The memory access time of a processor to all regions of memory is the same.
- The access times experienced by different processors are the same.



Nonuniform Memory Access (NUMA):

- All processors have access to all parts of main memory using loads and stores.
- The memory access time of a processor differs depending on which region of main memory is accessed.
- For different processors, which memory regions are slower and which are faster differ.



Cache-coherent NUMA (CCNUMA):

- A NUMA system in which cache coherence is maintained among the caches of the various processors.

CCNUMA Organization:

- There are multiple independent nodes, each of which is an SMP organization.
- Each node contains multiple processors, each with its own L1 and L2 caches, plus main memory.