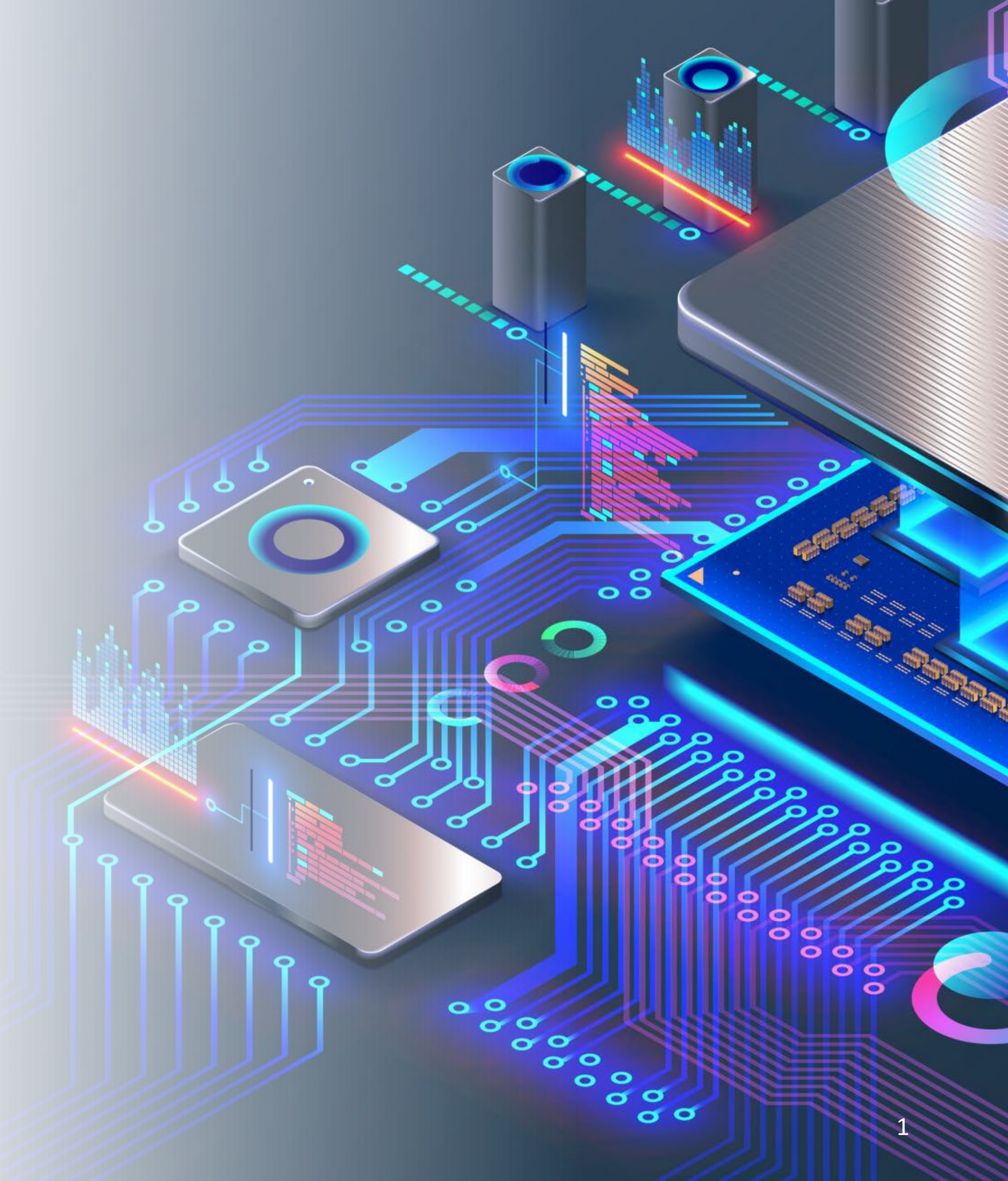


Architecture

KT44502 DISTRIBUTED SYSTEM

Semester 1 2022/2023

Content

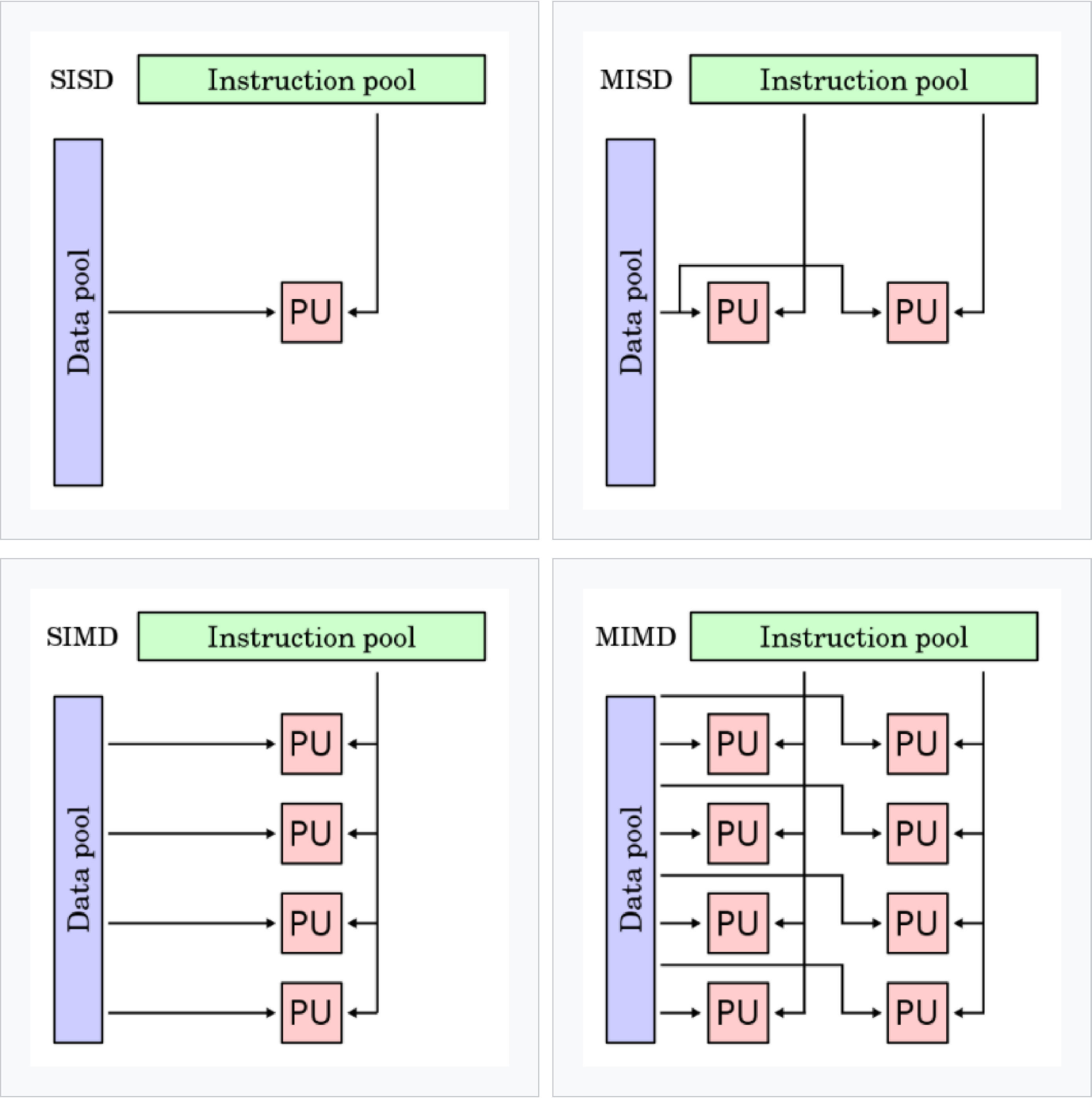
- Parallel Architecture
- Flynn’s Taxonomy
- Memory Organisation

Parallel Architecture

- Why is a 32-bit computer faster than an 8-bit? Parallelism.
    - 32-bit computer can do it in one step, handling each of the 4 bytes within the 32-bit numbers in parallel.
    - That’s why the history of computing has seen us move from 8- to 16-, 32-, and now 64-bit architectures.
- The total amount of benefit we’ll see from this kind of parallelism has its limits, though, which is why we’re unlikely to see 128-bit computers soon.

Classifying Parallel Systems – Flynn's Taxonomy

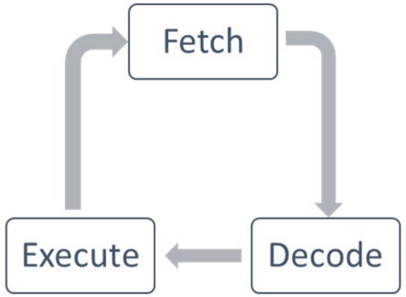
- Distinguishes multi-processor computer architectures along the two independent dimensions
  - ***Instruction*** and ***Data***
  - Each dimension can have one state: ***Single*** or ***Multiple***
- SISD: Single Instruction, Single Data
  - Serial (non-parallel) machine
- SIMD: Single Instruction, Multiple Data
  - Processor arrays and vector machines
- MISD: Multiple Instruction, Single Data (weird)
- MIMD: Multiple Instruction, Multiple Data
  - Most common parallel computer systems



The Flynn-Johnson classification of computer systems.

SISD: Single Instruction, Single Data

- The SISD computing system is a uniprocessor machine. It executes a single instruction that operates on a single data stream.
- In SISD, machine instructions are processed sequentially.
- In a clock cycle, the CPU executes the following operations:
  - Fetch: The CPU fetches the data and instructions from a memory area, which is called a register.
  - Decode: The CPU decodes the instructions.
  - Execute: The instruction is carried out on the data. The result of the operation is stored in another register



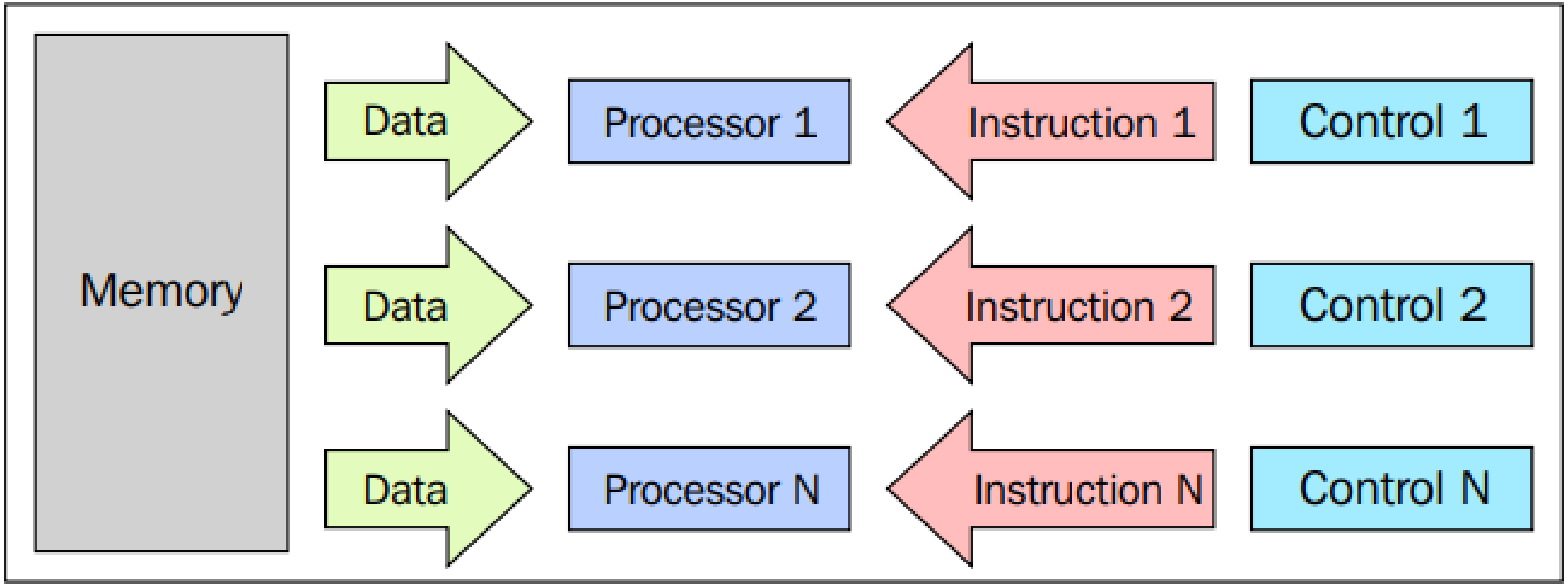
SISD: Single Instruction, Single Data

- The main elements of these architectures are:
  - Central memory unit: This is used to store both instructions and program data.
  - CPU: This is used to get the instruction and/or data from the memory unit, which decodes the instructions and sequentially implements them.
  - I/O system: This refers to the input data and output data of the program.

MISD: Multiple instruction, single data

- In this model, n processors, each with their own control unit, share a single memory unit.
- In each clock cycle, the data received from the memory is processed by all processors simultaneously, each in accordance with the instructions received from its control unit.
- The parallelism (instruction-level parallelism) is obtained by performing several operations on the same piece of data.
- MISD computers are more of an intellectual exercise than a practical configuration.

MISD: Multiple instruction, single data



The MISD architecture scheme

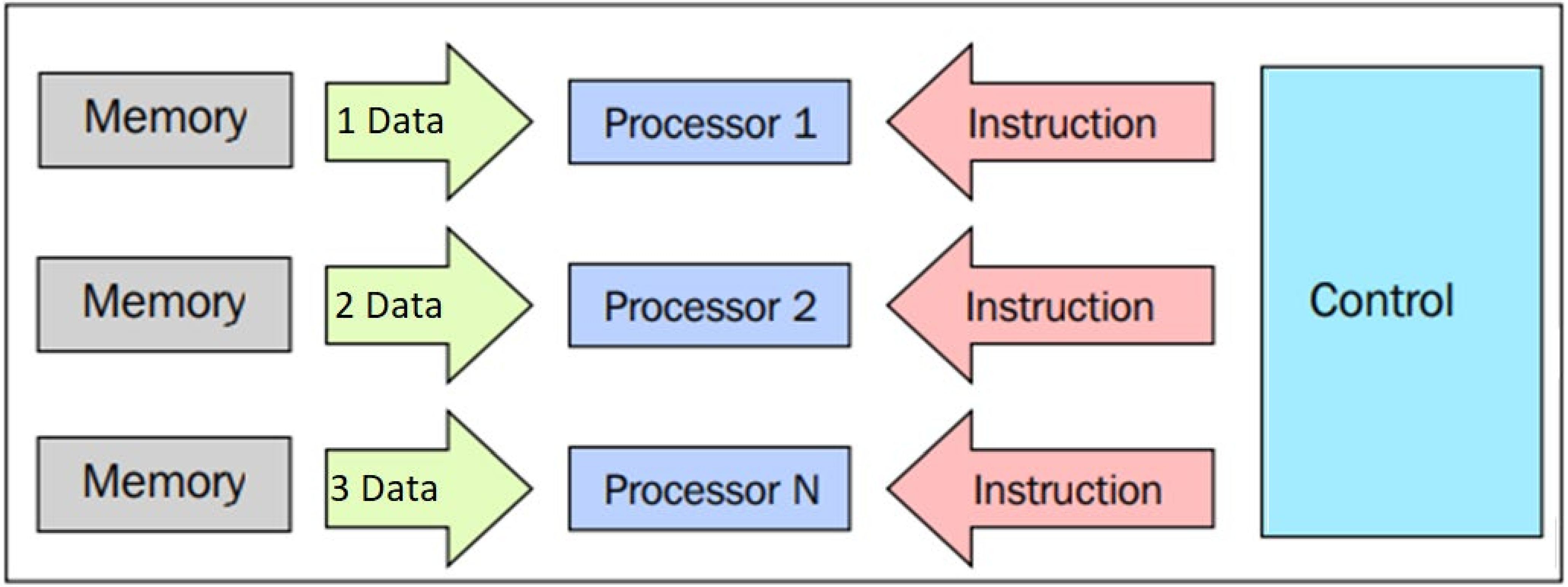
SIMD: Single instruction, multiple data

- A SIMD computer consists of n identical processors, each with its own local memory, where it is possible to store data.
- The processors work simultaneously on each step and execute the same instruction, but on different data elements.
- The algorithms for these computers are relatively easy to design, analyze, and implement.

SIMD: Single instruction, multiple data

- Numerous problems covering a wide range of applications can be solved by parallel algorithms on SIMD computers.
- Another interesting feature is that the algorithms for these computers are relatively easy to design, analyze, and implement.
- The limit is that only the problems that can be divided into a number of subproblems can be addressed with the SIMD computer.

SIMD: Single instruction, multiple data



MIMD: Multiple instruction, multiple data

- A MIMD computer consists of n processors, n instruction streams, and n data streams.
- Each processor has its own control unit and local memory.
- Each processor operates under the control of a flow of instructions issued by its own control unit.
- Processors can run different programs on different data, solving subproblems that are different and can be a part of a single larger problem.

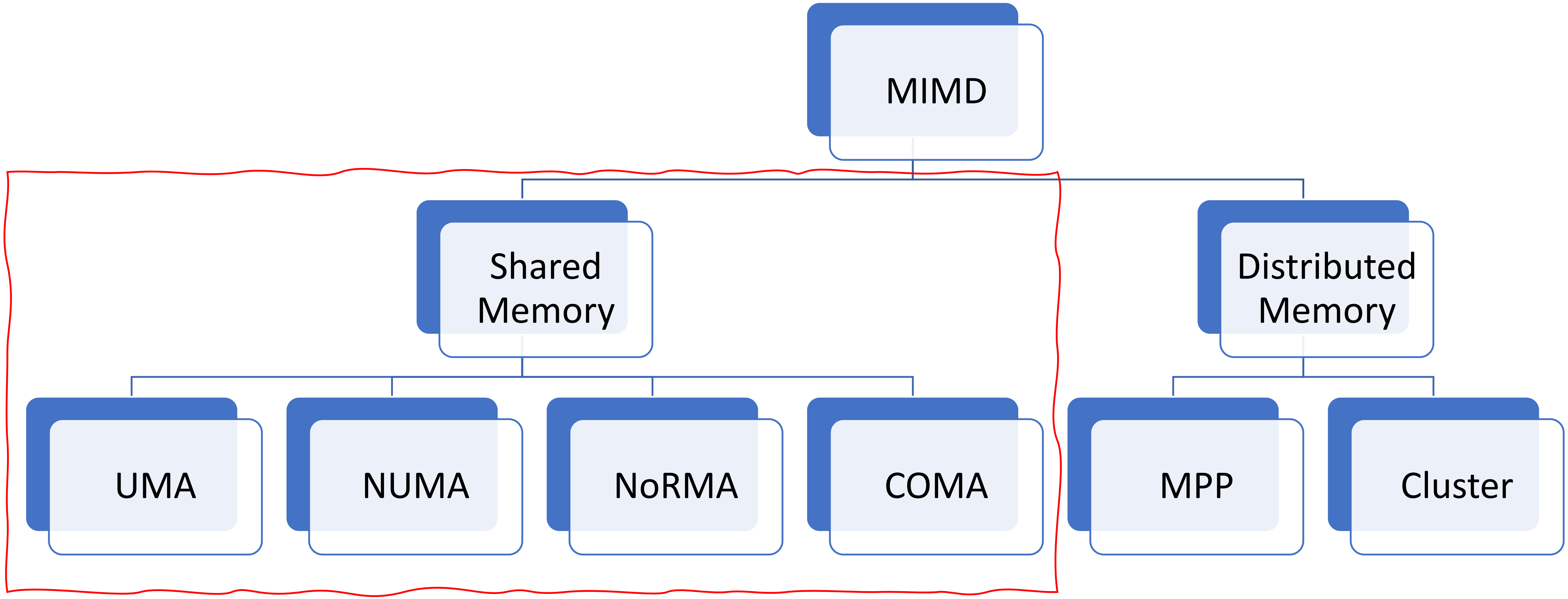
MIMD: Multiple instruction, multiple data

- MIMD, architecture is achieved with the help of the parallelism level with threads and/or processes. This also means that the processors usually operate asynchronously.
- The computers in this class are used to solve those problems that do not have a regular structure that is required by the model SIMD.
- MIMD architecture is applied to many PCs, supercomputers, and computer networks.

Memory Organisation

- Memory organization: the way in which the data is accessed.
- The main problem that must be overcome to make the response time of the memory compatible with the speed of the processor.
- When the processor starts transferring data, the memory will remain occupied for the entire time of the memory cycle.
- MIMD can be divided into machines that have a shared memory and those that have a distributed memory.

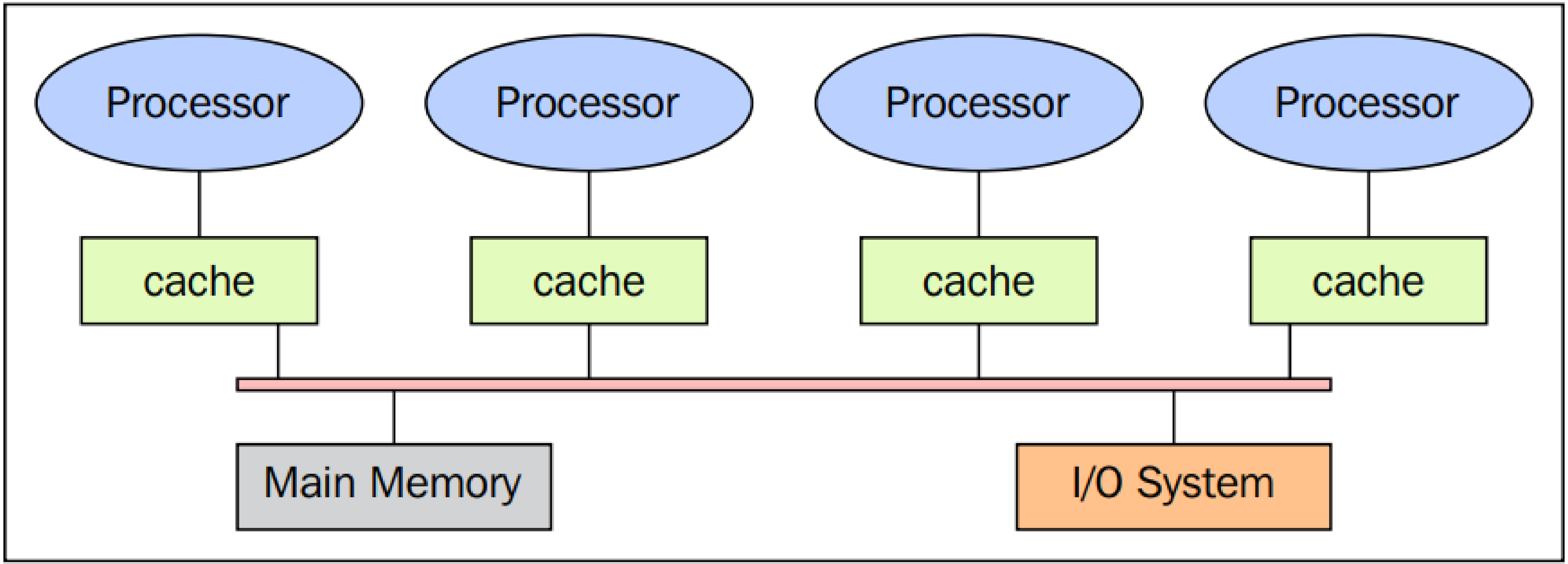
Memory Organisation



Shared Memory

- The memory is the same for all processors.
- The synchronization is made possible by controlling the access of processors to the shared memory.
- A shared memory location must not be changed from a task while another task accesses it.
- Sharing data is fast.
- The memory access in shared memory systems is as follows:
  - Classified as Uniform Memory Access (UMA).
  - Non-Uniform Memory Access (NUMA).
  - No Remote Memory Access (NoRMA).
  - Cache-Only Memory Architecture (COMA).

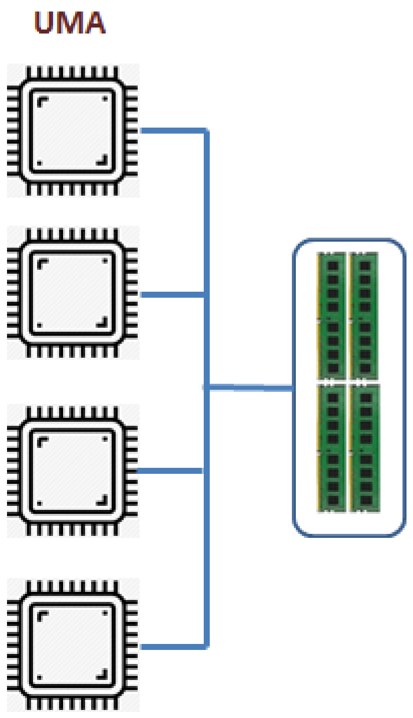
Shared Memory



The shared memory architecture schema

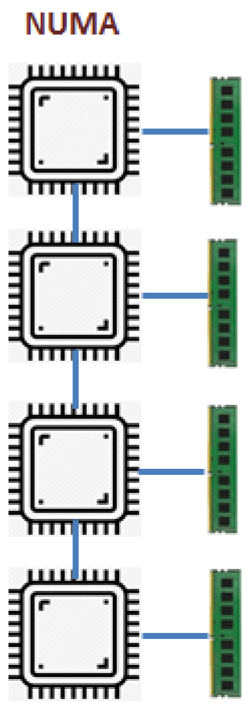
Uniform Memory Access (UMA)

- The fundamental characteristic of this system is the access time to the memory that is constant for each processor and for any area of memory.
- Also called as symmetric multiprocessor (SMP).
- Simple to implement, but hard to scalable.
- Memory access is slow.



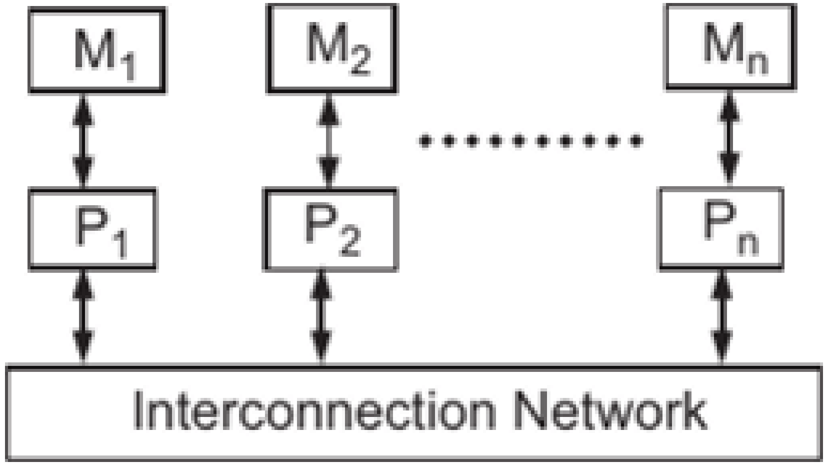
Non-Uniform Memory Access (NUMA)

- The memory area is divided into a high-speed access area that is assigned to each processor.
- A common area for the data exchange, with slower access.
- Processors may have local cache memories.
- The collection of local memories (LM) forms a global address space accessible by all processors.
- Easy to scalable, but complex to develop.
- Memory access is faster than UMA memory



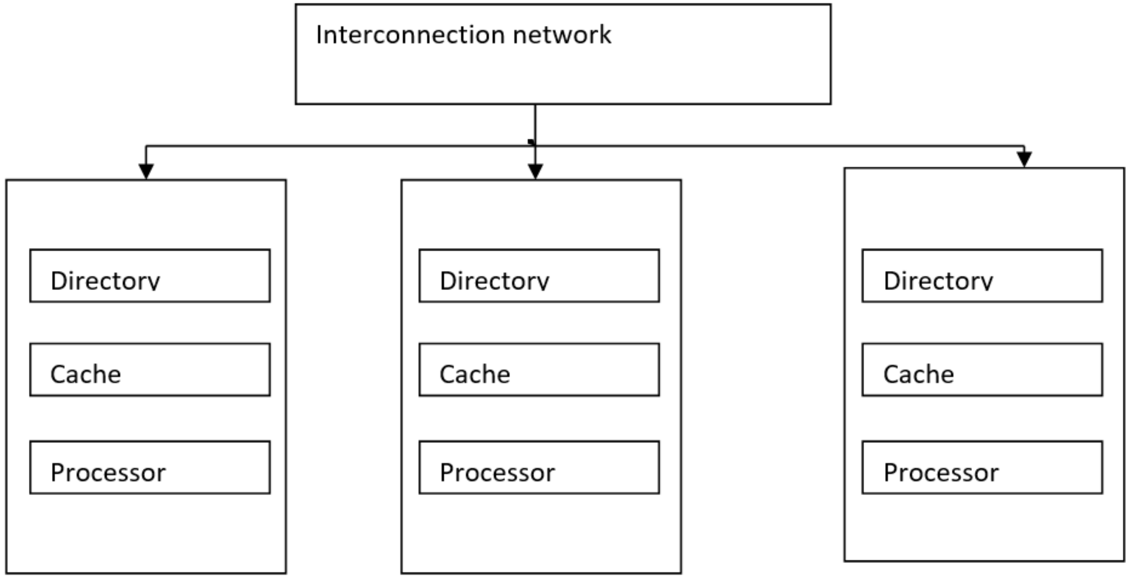
No Remote Memory Access (NoRMA)

- The memory is physically distributed among the processors (local memory).
- All local memories are private and can only access the local processor.
- The communication between the processors is the interconnection network to other processors used for exchanging messages, which is known as the message-passing protocol.

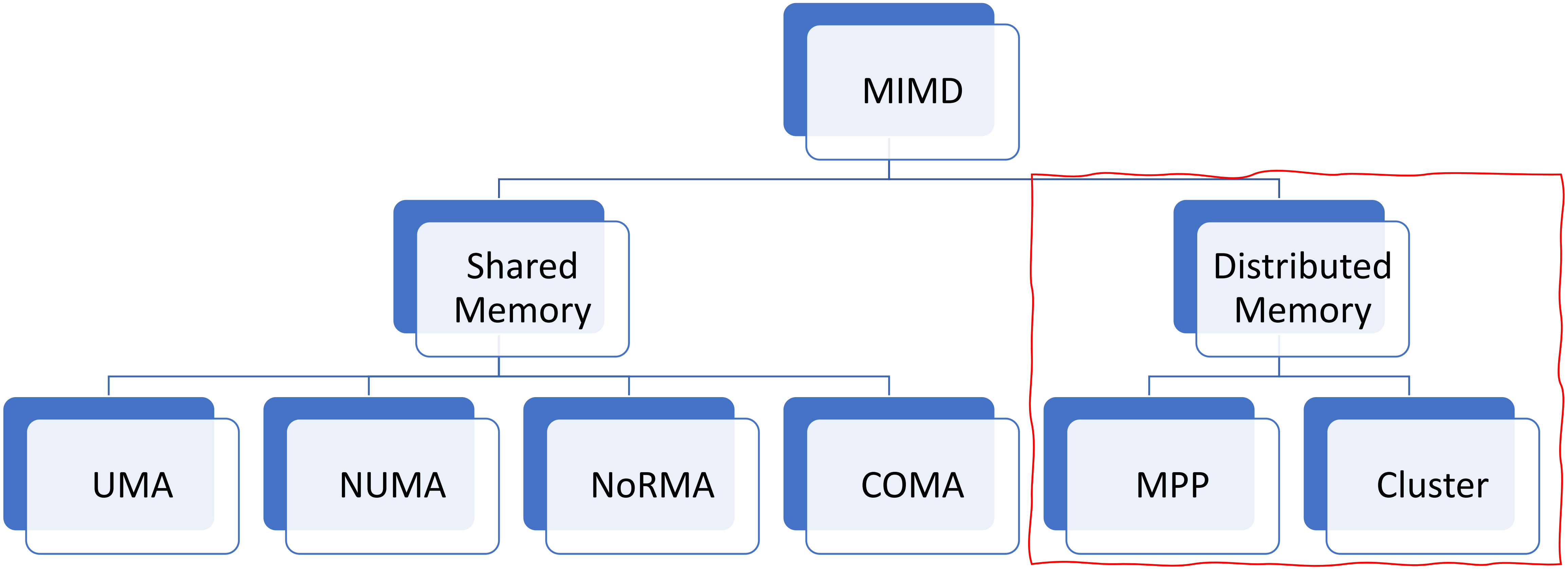


Cache-Only Memory Architecture (COMA):

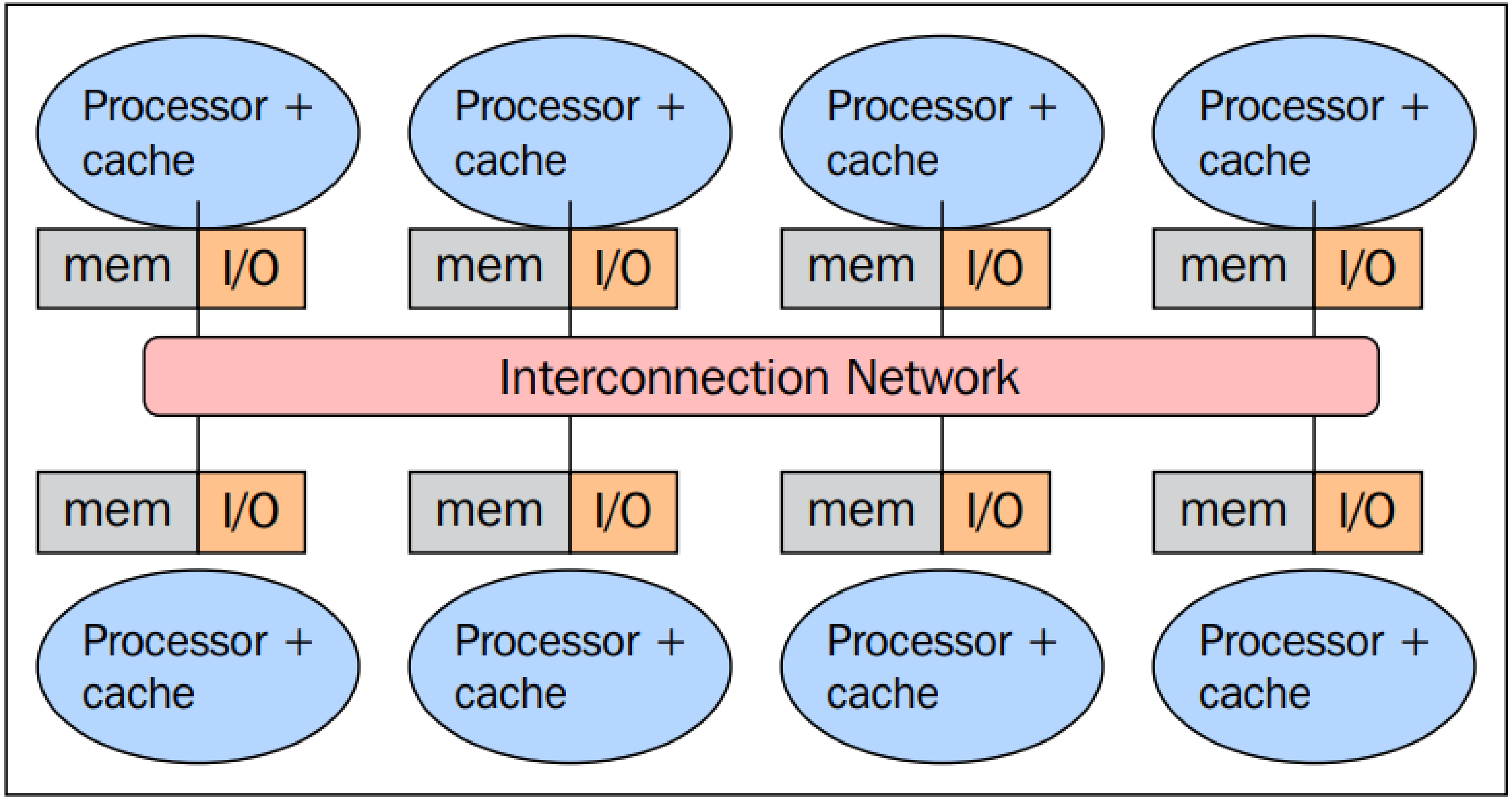
- These systems are equipped with only cached memories.
- This architecture removes duplicates and keeps only the cached memories; the memory is physically distributed among the processor.
- All local memories are private and can only access the local processor.
- The communication between the processors is also through the message-passing protocol.



Memory Organisation



Distributed Memory



The distributed memory architecture scheme

Distributed Memory Factures

- Memory is physically distributed between processors; each local memory is directly accessible only by its processor.
- Synchronization is achieved by moving data between processors (communication).
- The subdivision of data in the local memories affects the performance of the machine.
- The message-passing protocol is used.

Advantages of Distributed Memory Architecture

- No conflicts at the level of the communication bus or switch.
- No intrinsic limit to the number of processors
- The size of the system is only limited by the network used to connect the processors.
- No problems of cache coherency

Disadvantages of Distributed Memory Architecture

- The communication between processors is more difficult to implement.
- If a processor requires data in the memory of another processor, the two processors should necessarily exchange message via the message-passing protocol.
    - To build and send a message one processor to another takes time.
    - Any processor should be stopped in order to manage the messages received from other processors.
- A program designed to work on a distributed memory machine must be organized.

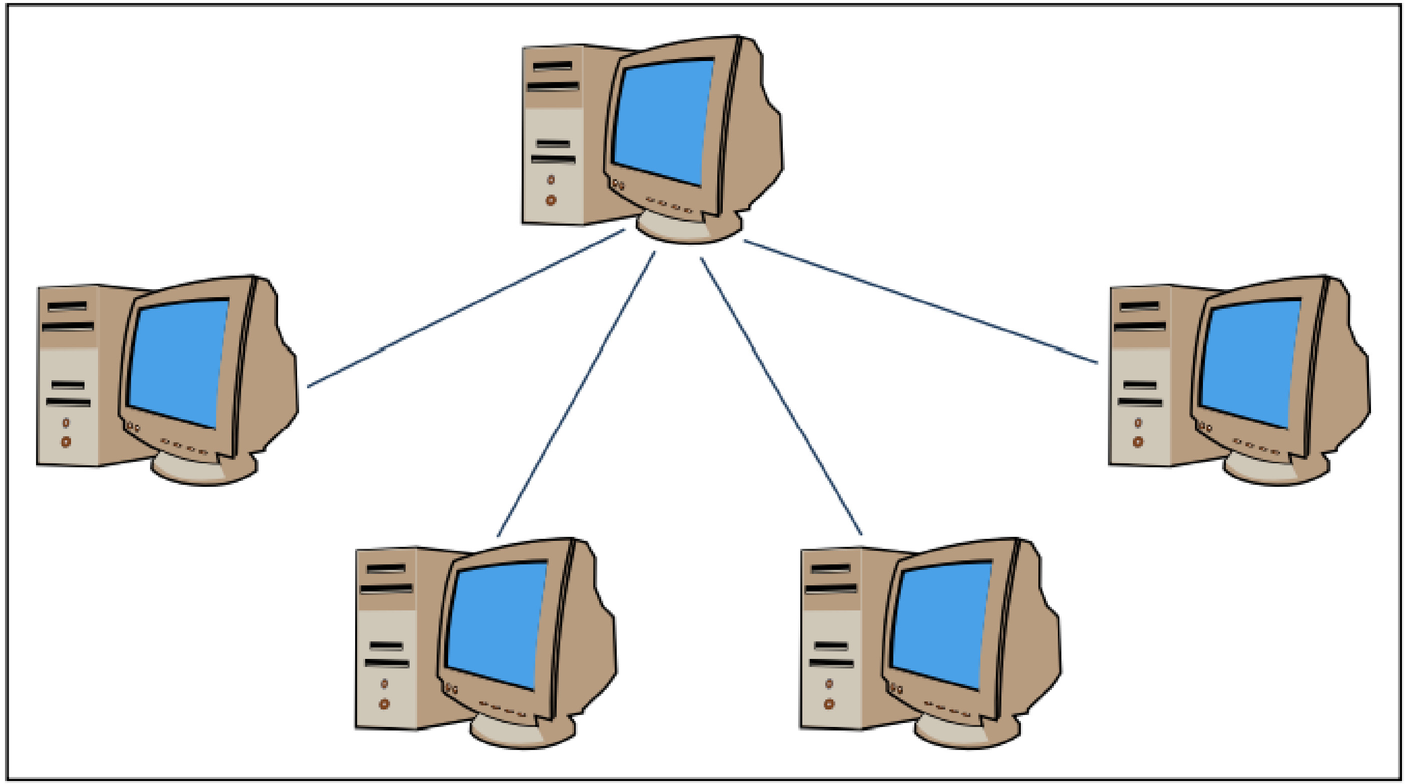
Massively Parallel Processing (MPP)

- MPP is a single computer with many networked processors.
- Each CPU contains its own memory and copy of the operating system and application
- The fastest computers in the world are based on these architectures.
- Example systems: Earth Simulator, Blue Gene, ASCI White, ASCI Red, and ASCI Purple and Red Storm.

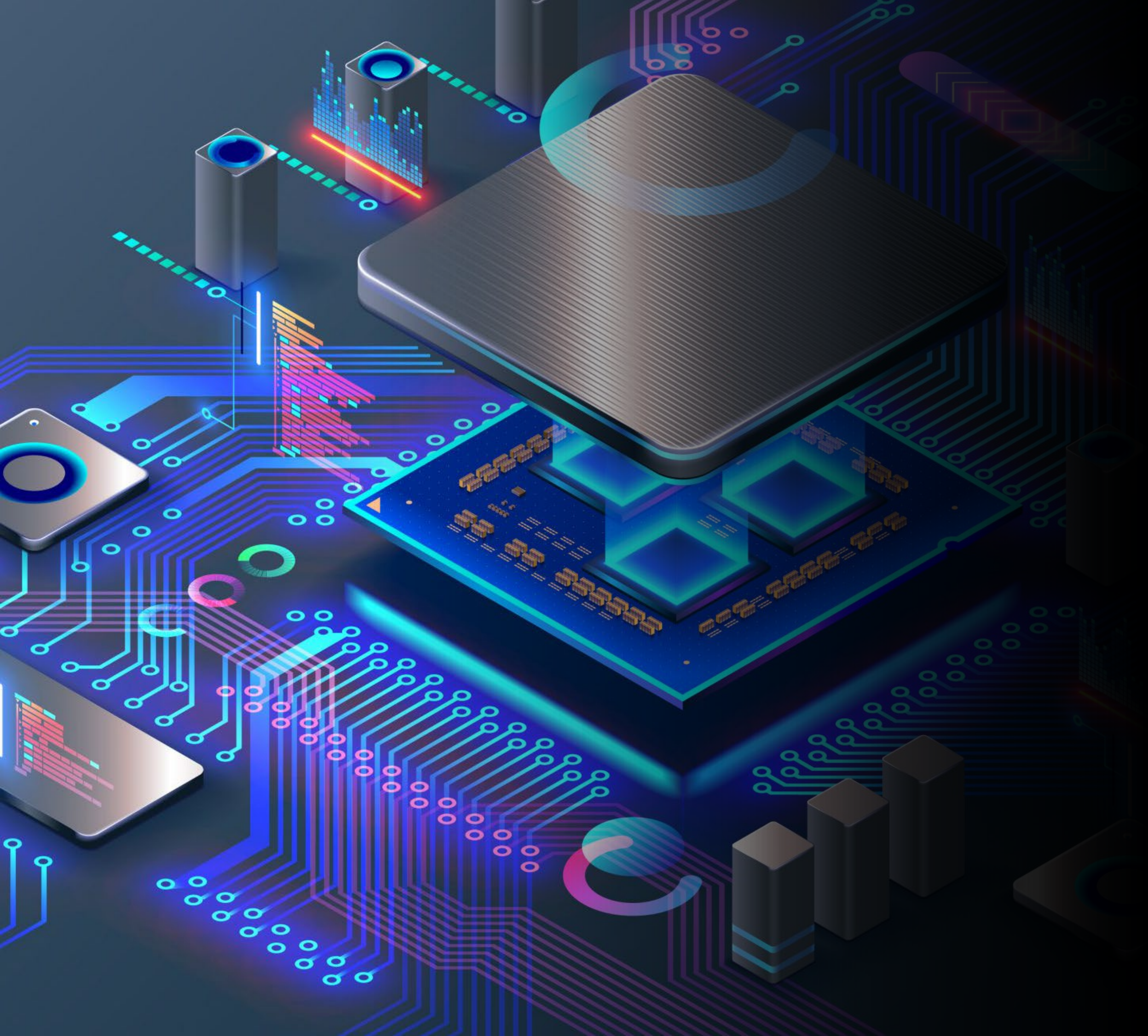
IBM Blue

Workstation Clusters

- Workstation Clusters are processing systems that based on classical computers that are connected by communication networks.
- A single computing takes part in the cluster.
- The cluster is fully transparent.



An example of a cluster of workstation architecture



Thank You

Next: Parallel Computer Architecture