CPU to RAM Connection

- CPU to RAM connection is more important.
- Different architectures have diverse strategies to implement this connection, which have their own consequences.

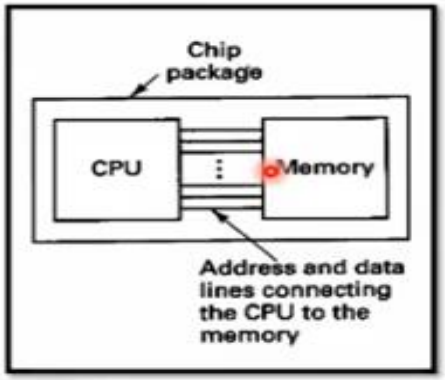
Multiprocessor Architectures && RAM to CPU Connection Strategies

1. On-Chip Memory
2. Bus-Based Multiprocessors
3. Ring-Based Multiprocessors
4. Switched Multiprocessors
5. NUMA Multiprocessors

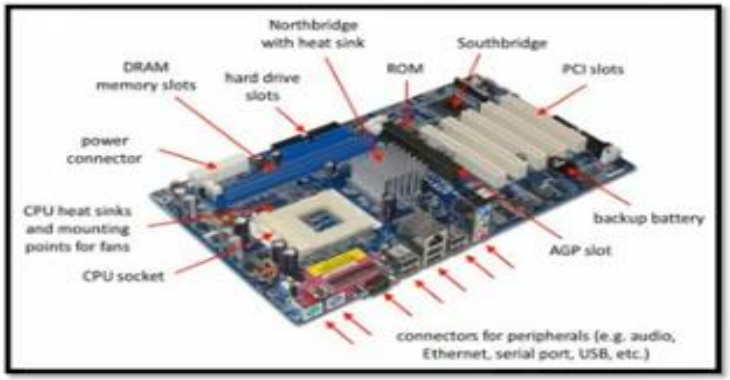
Connection Types cont...

On-Chip Memory

- Although most computers have ***external*** (to CPU) ***memory chips***, whereas ***self-contained chips containing a CPU and all the memory*** also exists.



self-contained chips

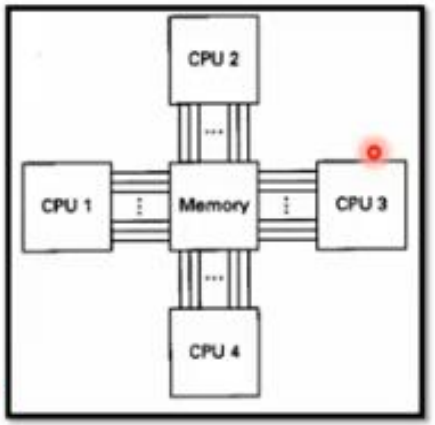


external (to CPU) ***memory chips***

- Self-contained chips are **costly enough**, and are widely used in cars, appliances, and even toys.
- In this design the CPU portion of the chip has **address**, **data** and **control lines** that **directly connect to memory portion**.

Connection Types cont...
On-Chip Memory

- The simple **extension** of self-contained chips are given here:
- While it is **difficult** to construct chip like this, it would be **complicated**, **expensive**, and highly unusual.
- So the different approaches should be used.



Shared Memory
multi-processor
self-contained chips

Connection Types cont...

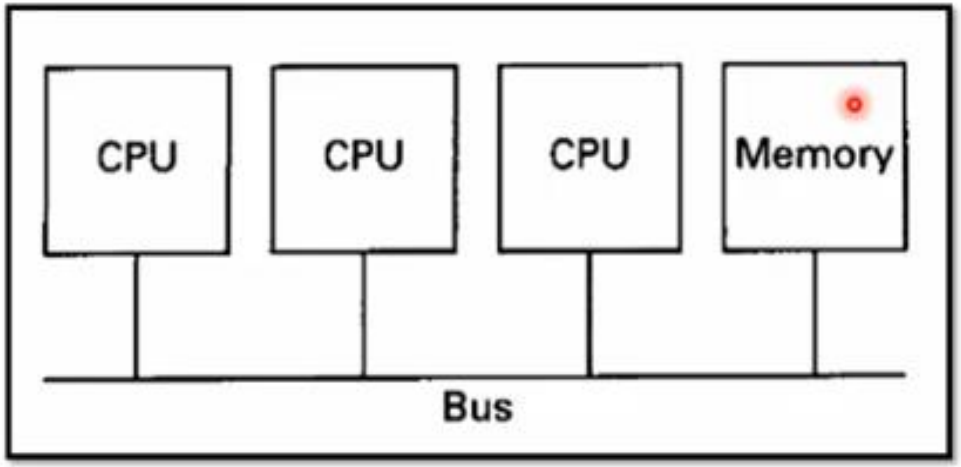
Bus-Based Multiprocessors

- In the on-chip memory the buses were also part of the chip.
- In the most systems **buses** are **external** and used to connect circuit boards containing CPU, Memories and I/O controllers.
- In a desktop or laptop systems, such external buses are used.
- A simple but practical way to build a multiprocessor system, is to **base it on a bus** to which **more than one CPU is connected**.

Connection Types cont...

Bus-Based Multiprocessors

- When any **CPU wants to read a word from memory**, it **puts the address of the word it wants on the bus** and **add a control signal to indicate**, it wants to read. When the **word is fetched it sent to CPU**.



Connection Types cont...

Bus-Based Multiprocessors

- To prevent two or **more CPUs from trying to access the memory at the same time some arbitrator is needed**. Various types are in practice.
- A **CPU might first have to request** it by declaring a special request line, only **after receiving permission**, it would be allowed to use the bus. The permission granting **process is supervised by arbitrator**.

Connection Types cont...

Ring-Based Multiprocessors

- In Ring-Based Multiprocessor the single memory is divided into **private part** and **shared part**.
- The private part is further divided into regions so that the each machine has a piece of stacks and other unshared data and code.
- The **shared part is common to all the machines** (and distributed over them).

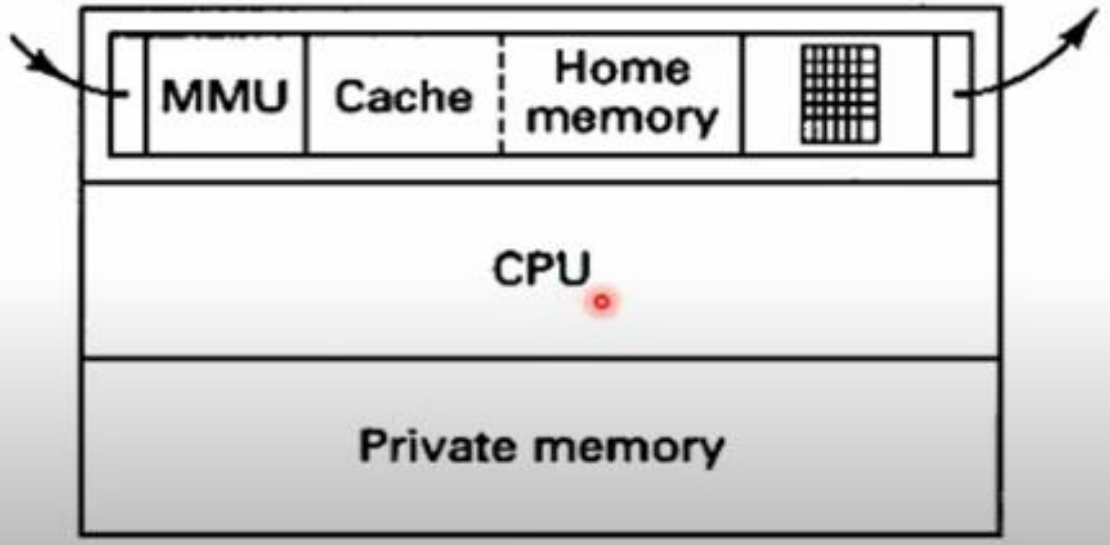
Connection Types cont...

Ring-Based Multiprocessors

- All the machines in the network are **connected through a modified token-passing ring**.
- The ring consist of parallel wires, which together allow **maximum data bits** and **some control bits** to sent and control data.
- The ring interface, **MMU** (Memory Management Unit), **cache** and part of the **memory** are **integrated together in the architecture**.

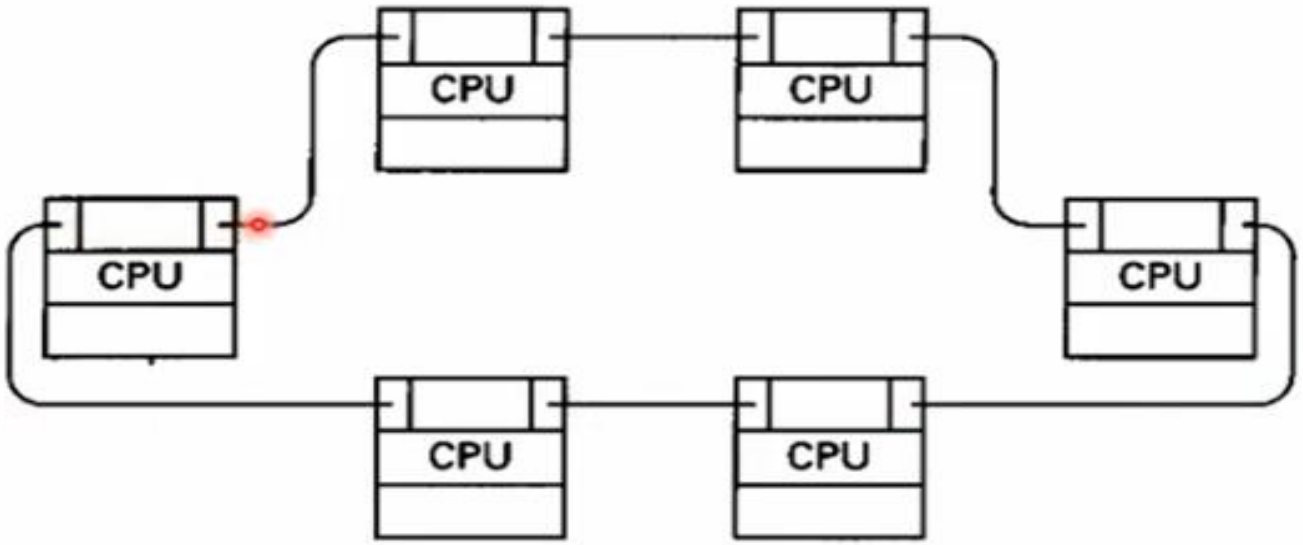
Connection Types cont...

Ring-Based Multiprocessors



Connection Types cont...

Ring-Based Multiprocessors



Connection Types cont...

Switched Multiprocessors

- Although Bus-Based Multiprocessors and Ring-Based Multiprocessors were working for small business (**up around 64 CPUs**) they **do not scale well** with systems **hundreds or thousands of CPUs**.
- As CPUs are added at some point the bus or ring **bandwidth were saturated**.
- Adding additional CPUs does not improve the system performance.

Connection Types cont…

Switched Multiprocessors

- Two approaches can be taken to tackle the problem:

1. **Reduce the amount of communication**.
2. **Increase the communication capacity**.

- On the other hand **communication can't reduced** furthermore **networks are already saturated** so it is hard to increase capacity.
- So a different method is to build the system as a hierarchy.

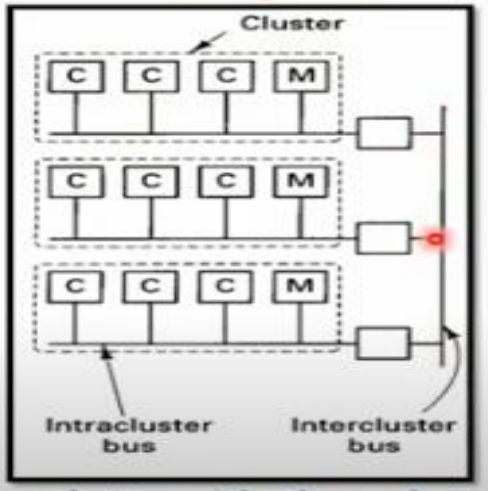
Connection Types cont...

Switched Multiprocessors

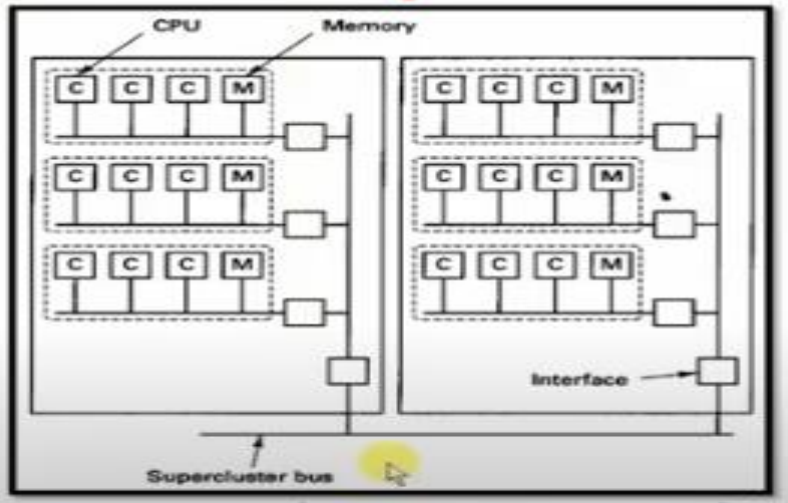
- Continue to **put some CPUs on a single bus**, but now regard this entire unit (**CPU plus bus**) as a **Cluster**.
- Built the system as **Multiple clusters** and **connect the clusters using an inter-cluster bus**, as shown here:

Connection Types cont...

Switched Multiprocessors



Super-cluster with, three clusters connected by a inter-cluster bus.



Two super-clusters are connected by a super-cluster bus.

Connection Types cont...

NUMA Multiprocessors

- It is clear by now that to build large multiprocessor systems is not a simple task, it is expensive as well as complex also.
- One architecture NUMA (Non-Uniform Memory Access) multiprocessor is yet another solution.
- On a NUMA machine, access to a remote memory is much slower than access to a local memory.

Cache Coherence Problem

- Multiple copy of the same data can exist in the different caches simultaneously,
- and if processors allowed to update their own copies freely, an inconsistent view of memory can result.
- Write policies : write back, write through

->In the write back policy only cache is updated and the location marked so that it can be copied later into main memory.

->In the write through policy cache and main memory are updated with every write operation.

Solutions to Cache Coherence

- Hardware Solution : in hardware solution the cache controller specify designed to allow it to monitor all bus requests from CPUs and IOPs.

->Directory protocol :

>it collect & maintain the information about copies of lines reside .

>contain the information about content of various local caches.

>keeping the information up-to-date.

>manage the information which caches copy of which line.

Drawback – only for less buses not large scale system

Snoopy Cache Protocol

->distributed responsibility for maintaining cache coherence among all of the cache controller in the multiprocessor.

Basic Approach: write invalid & write update.

- Write invalid protocol – there can be multiple readers but only one writer at a time, only one cache can write to the line.
- Write update protocol – there can be multiple writer as well as multiple readers.

->when a processor wishes to update a shared line, the word to be distributed to all others, and caches containing that line can update it.