Memory Organization

Agenda

- Basics of Memory Technologies
- Memory Hierarchy and Cache
- Types of Memory Architectures
  - SMP
  - NUMA
  - Cluster
- Summary

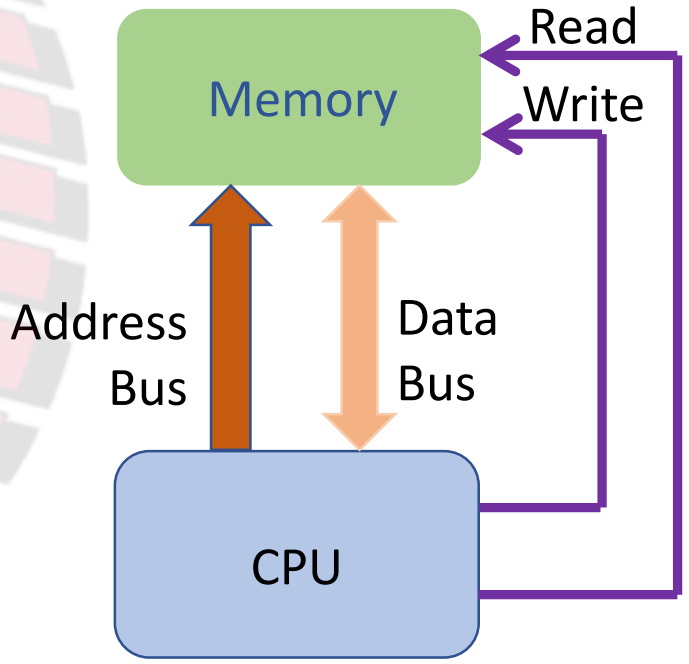
Basics of Memory Technologies

Memory technologies primer

Characteristics	Static RAM	Dynamic RAM	Flash Memory	Hard Disk
Basic storage element	Flip-Flop	Capacitor	Transistor	Magnetic material
Data retention	Volatile	Volatile	Non-Volatile	Non-Volatile
Read speed	> 10 nanoSec	~ 40 nanoSec	Slower than RAMs **Faster than Hard Disk**	Very slow, 10s of miliSeconds
Write speed	Same as read	Same as read	Slower than read	Same as read
Advantage	**Fastest access**	**Best Density and capacity**	Retention	**High Capacity**
Disadvantage	Density, power	Speed	Slower writes	Physical size, speed, power
Preferred use	**Cache memory**	**Main memory**	BIOS, **Solid State Drives**	**Secondary storage**

How CPU access memory

- Read: A LOAD instruction is decoded
  - Operand for LOAD is address
- CPU puts the address on the Address Bus
  - Simultaneously asserts Read signal
- Memory responds with data from the specified address and put on the Data Bus
- CPU accepts the data
- Write: STORE instruction specifies address and data
- CPU address and data on bus, asserts Write signal
- Memory stores the data at specified address location



Memory Requirements

- Programmers want unlimited amount of fast memory
- Realities
  - Fast memory is costly
  - Programs tend to follow principle of locality
    - Temporal: same location is repeatedly accessed
    - Spatial: Close by locations are accessed soon
- Hardware aids programmer by creating an illusion of unlimited fast memory

Processor and memory speeds

- When microprocessors were introduced, standard memory parts were faster than contemporary microprocessors
- Over a period
  - Microprocessors have become faster
  - Memories have become little faster while having higher capacity
- The result - memories are not able to keep up with the speed achieved by microprocessors

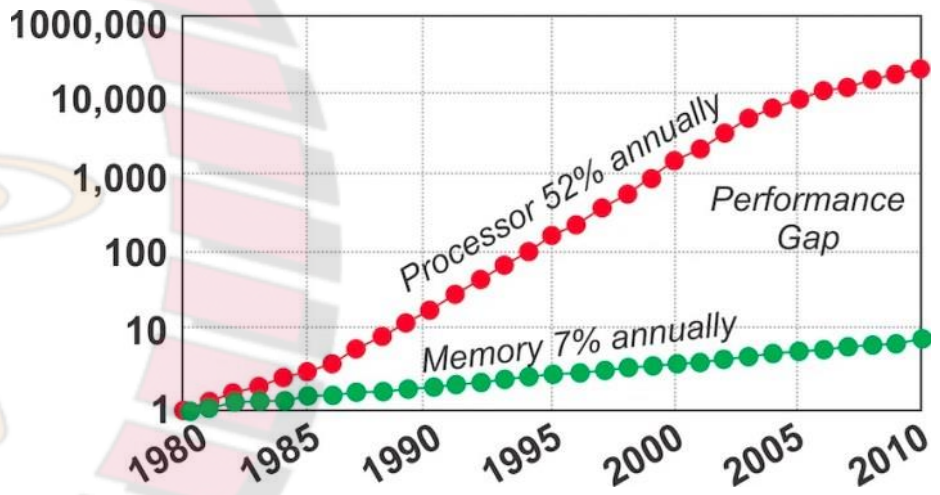


Image source: www.rankred.com

How to bridge the performance gap ?

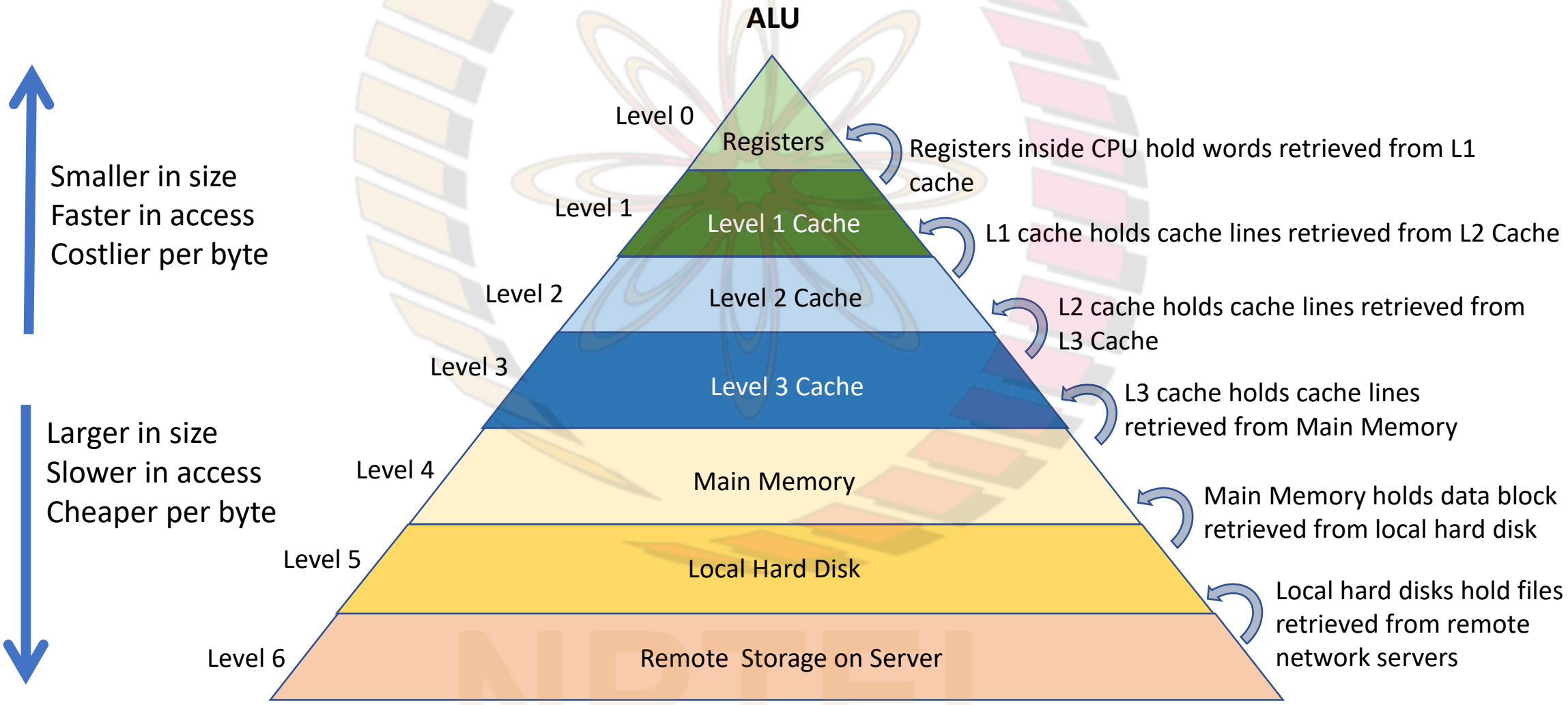
- Interleaving the memory
  - Odd address in one bank and even address in another bank
  - Results in overlaping memory delays and speedup the operations
- Wider memory paths
  - More data is fetched per access cycle
  - Pentium is a 32 bit CPU, whereas memory data path is 64 bit
- Cache memory system
  - Faster memory to serve frequently used data and code

Memory Hierarchy and Cache

Memory Hierarchy

- Takes advantage of principle of locality
- Memory subsystem consists of multiple levels with different speeds and sizes.
- Small fast memory is put close to processor and slower away
- The goal
  - Provide maximum memory which is cheapest
  - Provide access at the speed offered by the fastest
- Data is copied only between two adjacent levels
- If data is found in level closest to CPU, it’s a hit
- Hit rate is a measure of the performance

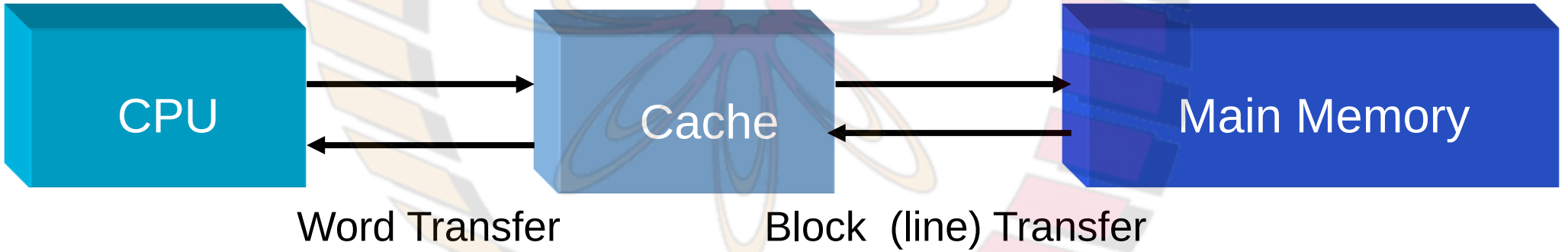
Memory Hierarchy



Cache

Cache: a safe place for hiding or storing things

-Webster's New world dictionary



- Cache contains a copy of portions of main memory
- When CPU attempts to read a word, cache is checked first. If found, it is presented, else a line of words is brought from main memory to the cache

Locality

- Cache memories work on the phenomenon of locality of reference
- Temporal Locality

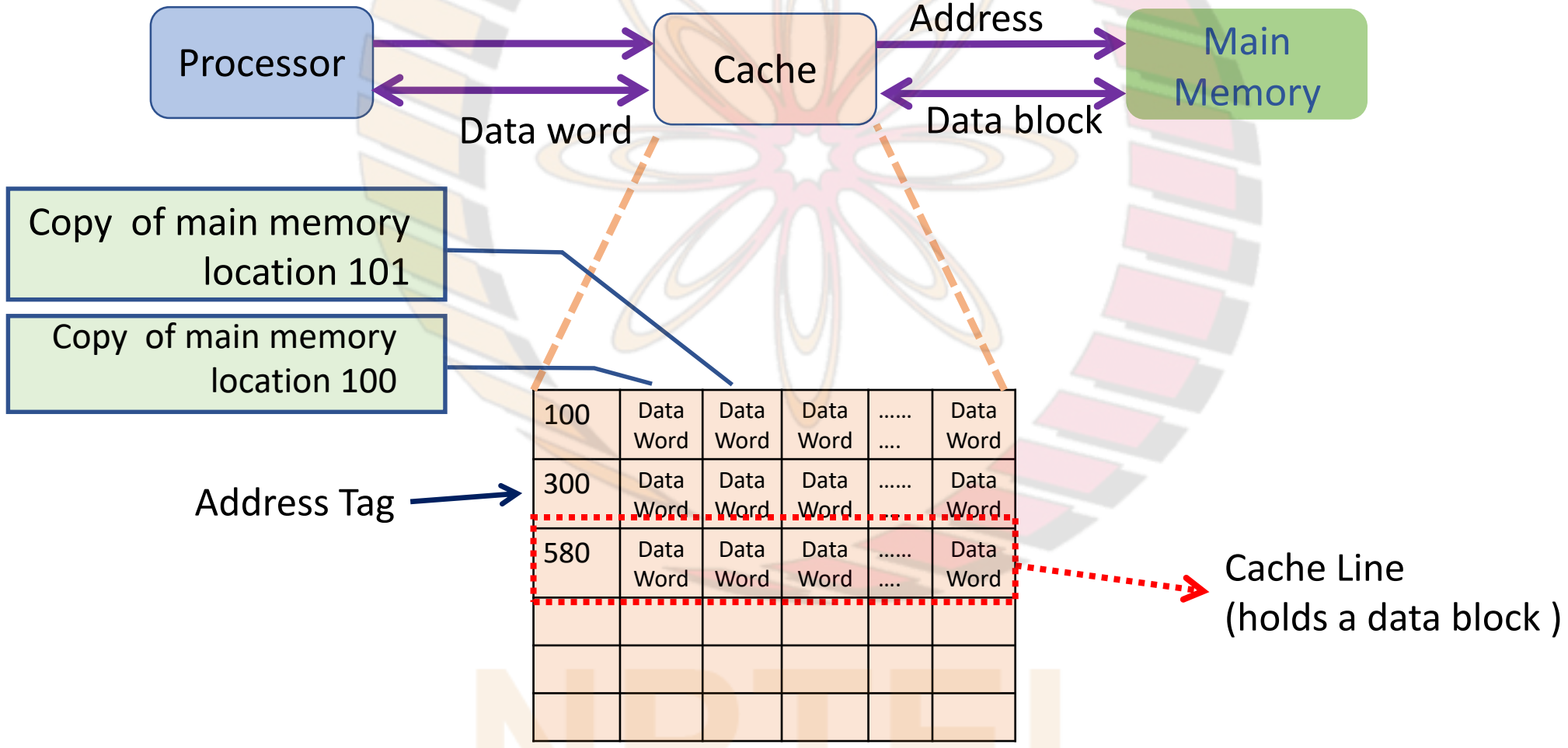
  If a memory location is referenced by the processor, then it is very likely that the same memory location is going to be referenced again in the near future.
- Spatial Locality

  If a memory location is referenced by the processor, then it is very likely that a nearby location is going to be referenced in the near future.

Basic Architecture of Cache

- A cache consists of 3 main parts
  - Directory store
    - The cache must know where the information in cache originated from
    - Each directory entry is called as cache-tag
  - Data section
    - Holds the data copied from main memory
  - Status information
    - Valid bid indicates live data
    - Dirty bit indicates non-coherency

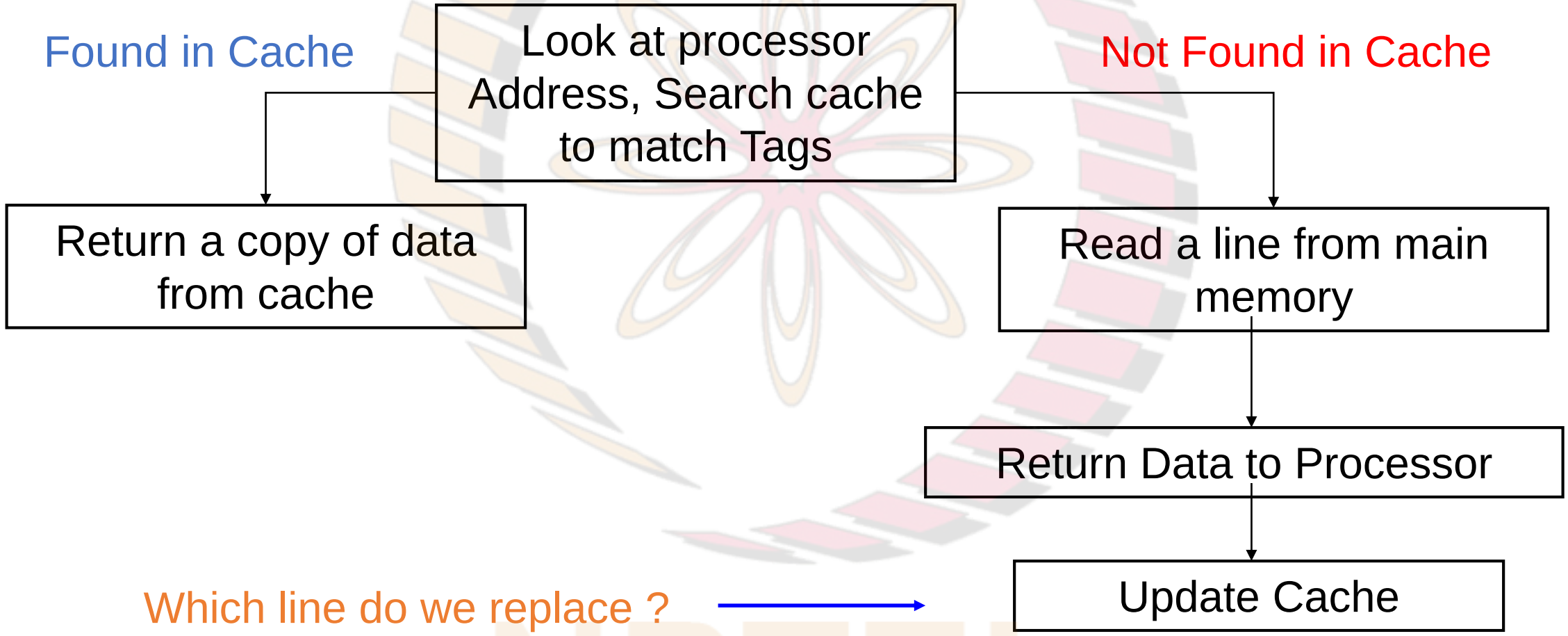
Inside a Cache



Cache Controller

- It is the hardware that copies information from main memory to cache memory automatically.
- The operation is transparent to CPU and the software
- It intercepts reads and writes before passing them on to the memory controller.
- The relationship between main memory and cache is called as mapping

Cache Read Algorithm



Cache Mapping

- Schemes for Mapping between main memory locations and cache
  - Direct mapped
  - N-way set-associative (n=2,4)
  - Fully set-associative

Cache Efficiency

- Cache hit rate is the term use to characterize cache efficiency of a program

$$\text{Hit rate} = \frac{\text{cache hits}}{\text{memory requests}} \times 100$$

- Hit rates can measure read hits, write hits or both

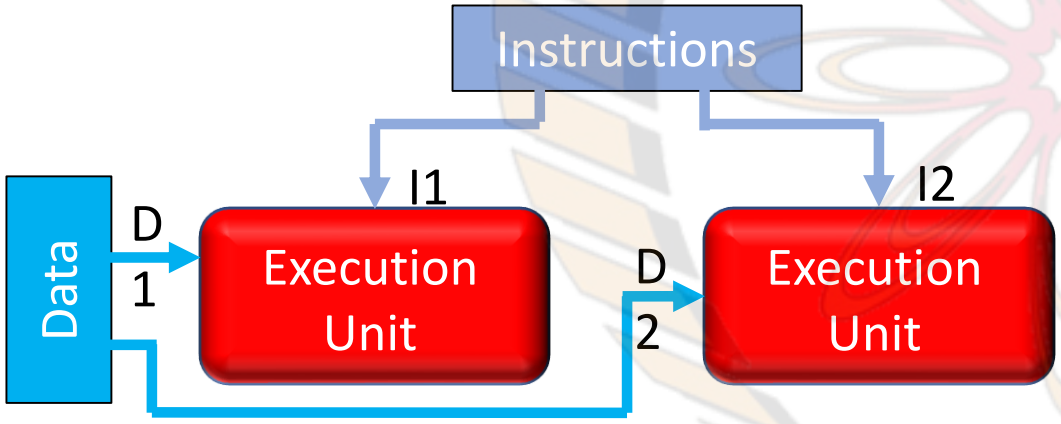
- Other performance measurement terms are
  - Hit time
  - Miss penalty

Cache Coherency Problem

- It arises when the data in the cache is different as compared to the data in the main memory. Stale data.
- This problem is of concern only when there is more than one entity using the memory. e.g. multiprocessor systems
- To avoid the problem write-through policy is used.
- The cache controller snoops the memory bus for memory operations and updates the cache whenever required. This is called as snooping.

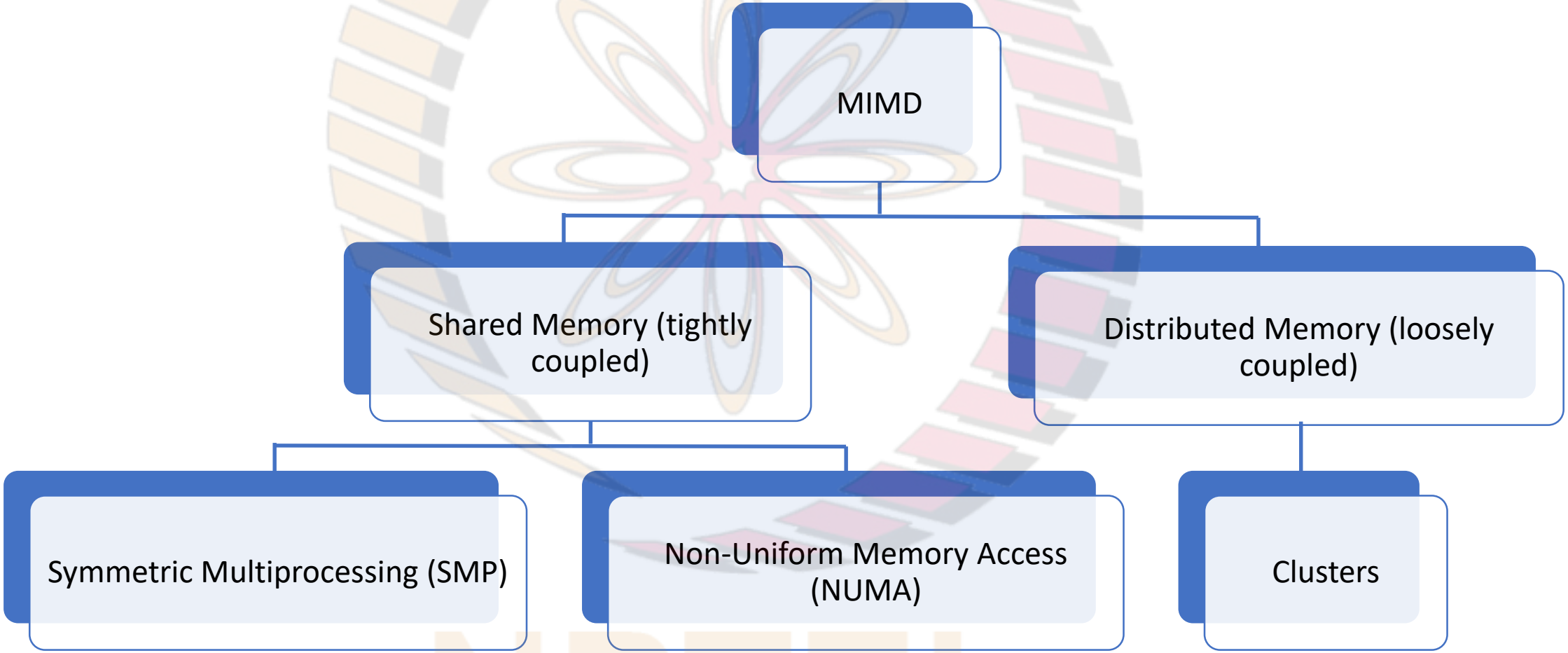
Types of MIMD Architectures

Recall - Multiple Instruction Multiple Data



- Multiple independent processors concurrently executing different instructions on their respective data
- Most of the parallel computers are of this type
- Two varieties of implementation
  - Shared Memory
  - Distributed Memory

Multiple Instruction Multiple Data (MIMD) Systems

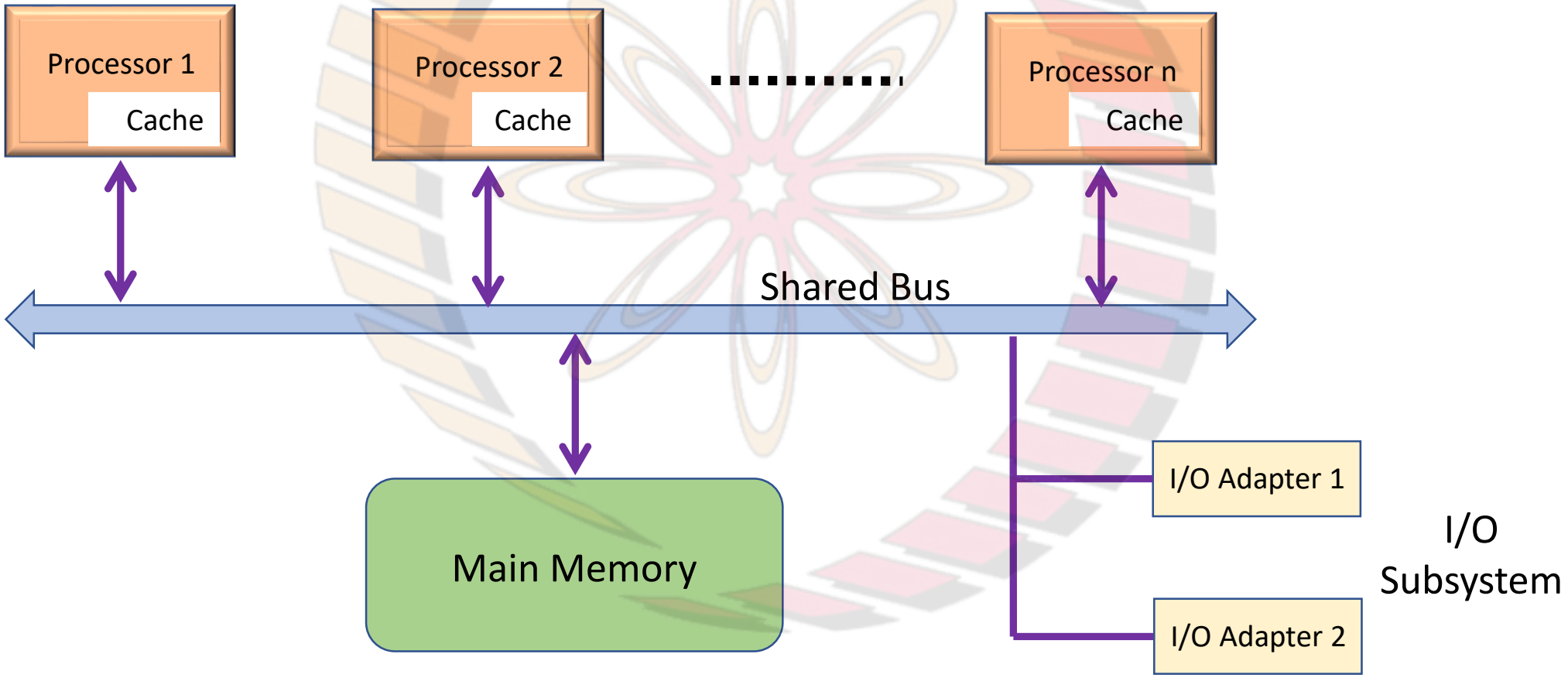


SMP Architecture

Symmetric MultiProcessing (SMP)

- A standalone computer having following characteristics
  - Two or more processors with comparable compute power
  - All processors share the same memory and I/O facilities
  - Processors are interconnected by a bus or other interconnection scheme
    - Such that memory access time for all processors is almost the same
  - All processors have a path to access all I/O devices
  - All processors can perform the same function
- The computer is controlled by a integrated operating system that allows assignment of processors to programs or tasks or threads

Block diagram of SMP

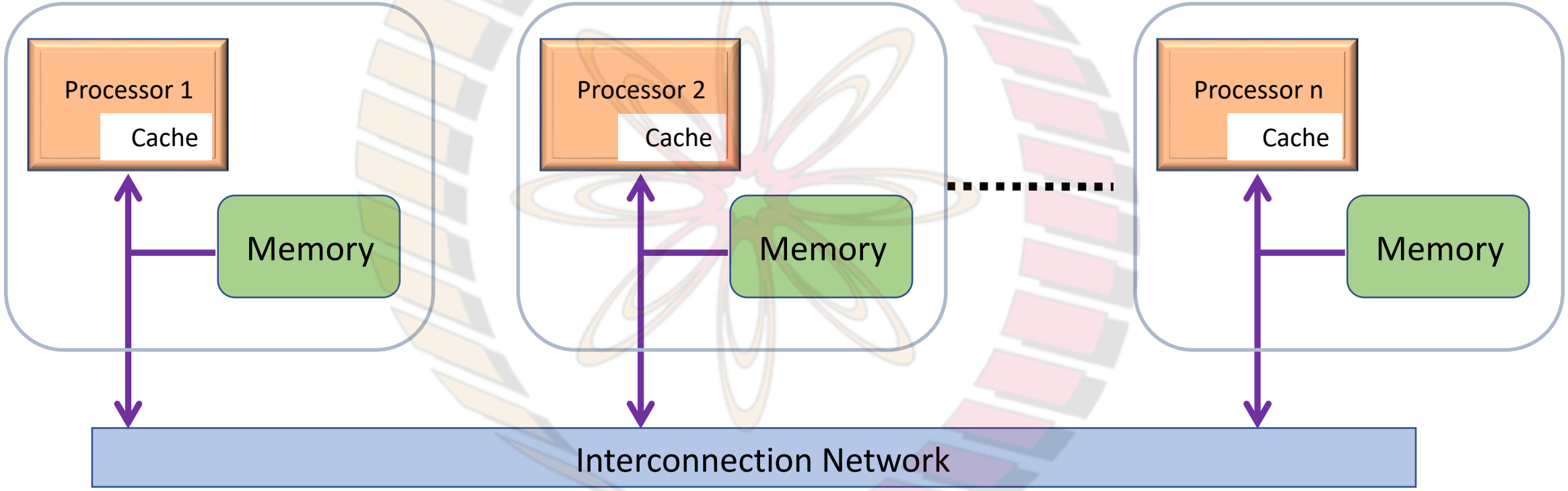


NUMA Architecture

Non-Uniform Memory Access (NUMA)

- Uniform Memory Access (UMA)
  - All processors have access to all parts of main memory
  - Access time for a processor to all parts of main memory is same
  - Access time experienced by different processors are same
  - SMP is a type of UMA architecture
- Non-Uniform Memory Access
  - All processors have access to all parts of main memory
  - Access time for a processor differs, based on which part of memory is being accessed
  - This is true for all processors
  - For each processor, which region is fast and which is slow is different

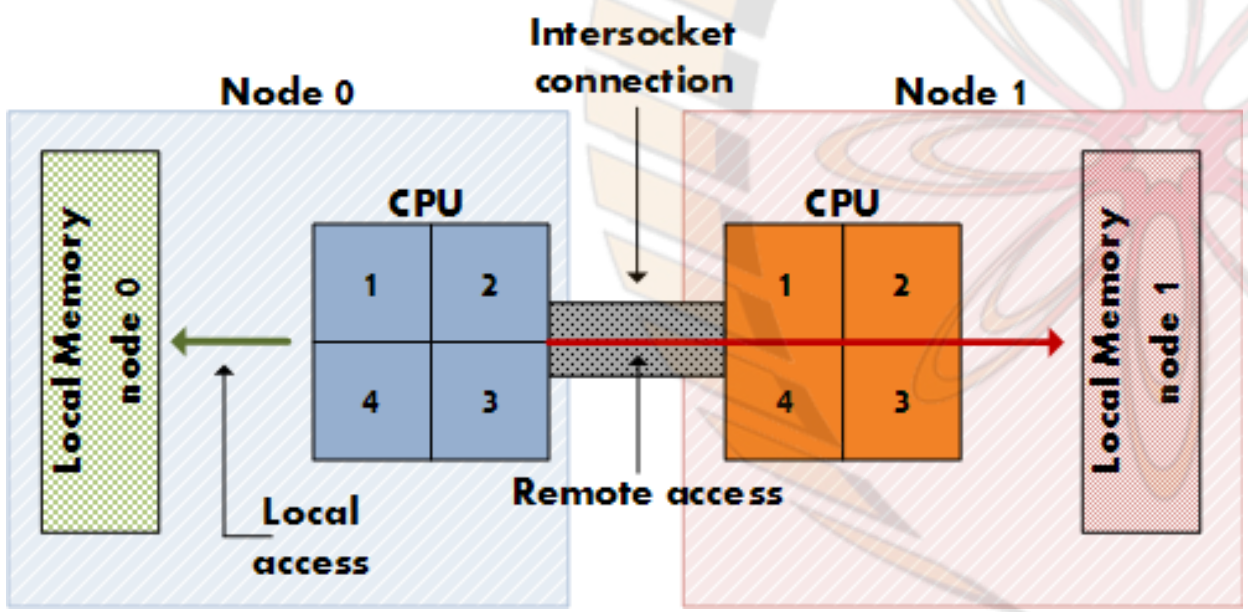
Non-Uniform Memory Architecture

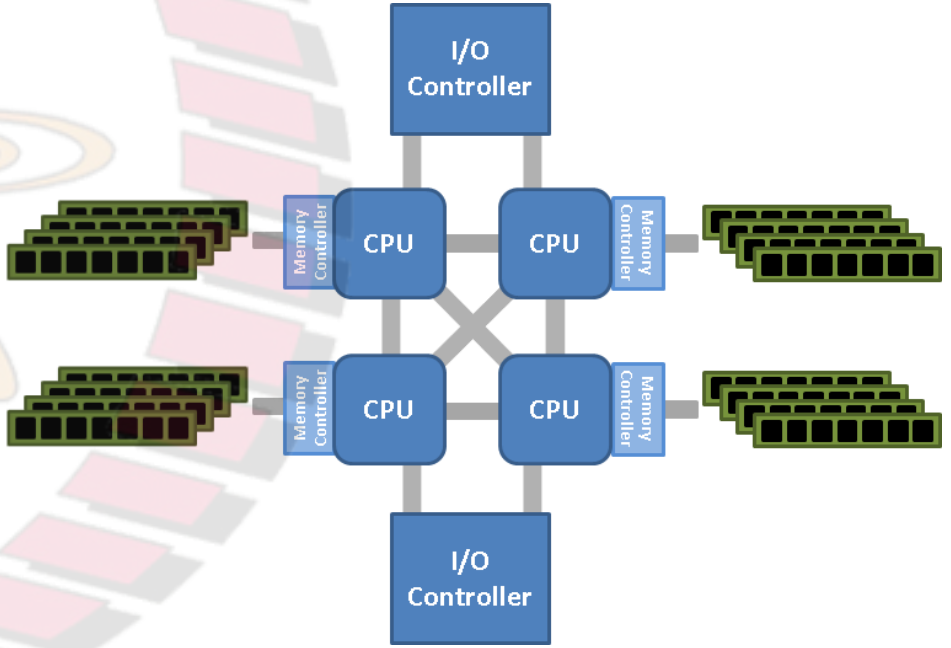


Why NUMA is needed?

- SMP has a practical limit to number of processors and memory which could be used (Electrical/PCB design constraints)
- Typically 64 processors maximum, most expensive machines around
- NUMA machines offer the same advantages as SMP
  - Flat memory addressing model
  - Single OS image
- While the physical restrictions limit extends to hundreds to thousands of physical processors/cores
- Example: SGI's Altix ICE 8400 (> 64K cores)

NUMA implementation



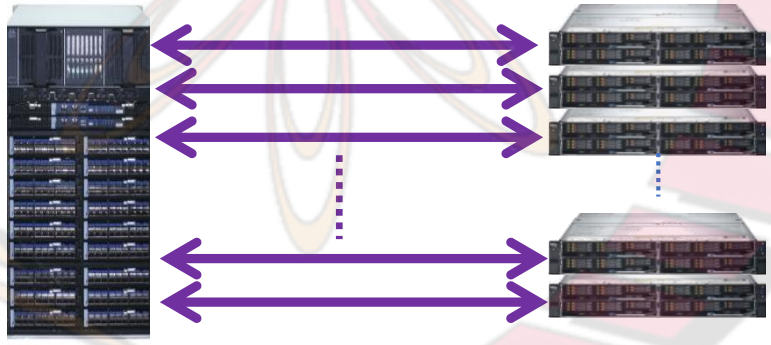


Images Source: Intel Corporation

Cluster

Cluster

- Definition : A group of interconnected, standalone commodity computers that work together as a unified computing resource



Benefits of Cluster

- Extreme Scalability: It is possible to have cluster with large number of nodes. Clusters can have thousands of machine, each one of which is a multiprocessor
- Incremental Scalability: It is possible to design and configure a cluster in such a way that the number of nodes can be increased as need grows
- High Availability: Failure of a node does not affect overall loss of service
    - Software can handle fault tolerance
- Superior price/performance: Use of commodity building block results in equal compute power of a single large machine, at a much lower cost

Attributes of HPC network

- Objective is to move data from memory of one node to memory of another node
- Bandwidth (or throughput)
  - Defined as amount of data transferred in unit time
  - Depends upon the raw data rate at physical layer
- Latency
  - Time taken for a packet to reach destination from source
  - Useful for lots of small messages travelling in the system
    - Processor waits for control/sync messages between processes
- Intelligence in network: Offloading networking stack in hardware, freeing up host processor for doing computations

Summary

- Hierarchical memory system helps in bridging the gap between faster CPUs and relatively slower memories
- Cache memories play a significant role in reducing memory latencies.
- Cache memory subsystem work on principle of locality
- Main Memory for large systems can be implemented in multiple ways
  - Tightly Coupled (SMP and NUMA)
  - Loosely Coupled (Clustering)
- Clustering is most commonly used architecture due to its advantages of scalability and cost-effectiveness

Thank You