SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME :19IT301 COMPUTER ORGANIZATION AND ARCHITECTURE

II YEAR /III SEMESTER

Unit 5: I/O ORGANIZATION AND PARALLELISM

Topic 8: Introduction to multicore processor

Introduction to multicore processor

- **A multi-core processor is a computer processor on a single integrated circuit with two or more separate processing units called cores, each of which reads and execute program instructions.**
- Processor cores are **individual processing units within the computer's central processing unit** (CPU)

- **Multiprocessor:** A computer system with at least two processors
- **Multicore:** More than one processor available within a single chip

Multi-core processor

A **multi-core processor** is a single computing component with two or more independent actual processing units (called "cores"), which are the units that read and execute program instructions.
Shared memory multiprocessor
Message-passing multiprocessor

Multicore processors

Single address space shared by all processors

Processors coordinate/communicate through shared variables in memory (via loads and stores)

Use of shared data must be coordinated via synchronization primitives (locks) that allow access to data to only one processor at a time

They come in two styles

Uniform memory access (UMA) multiprocessors

Nonuniform memory access (NUMA) multiprocessors

Multicore processors

shared memory multiprocessor(SMP)

A parallel processor with a single address space, implying implicit communication with loads and stores.

A shared memory multiprocessor (SMP) is one that offers the programmer a *single physical address space* across all processors.

uniform memory access (UMA)

A multiprocessor in which accesses to main memory take about the same amount of time no matter which processor requests the access and no matter which word is asked.

nonuniform memory access (NUMA)

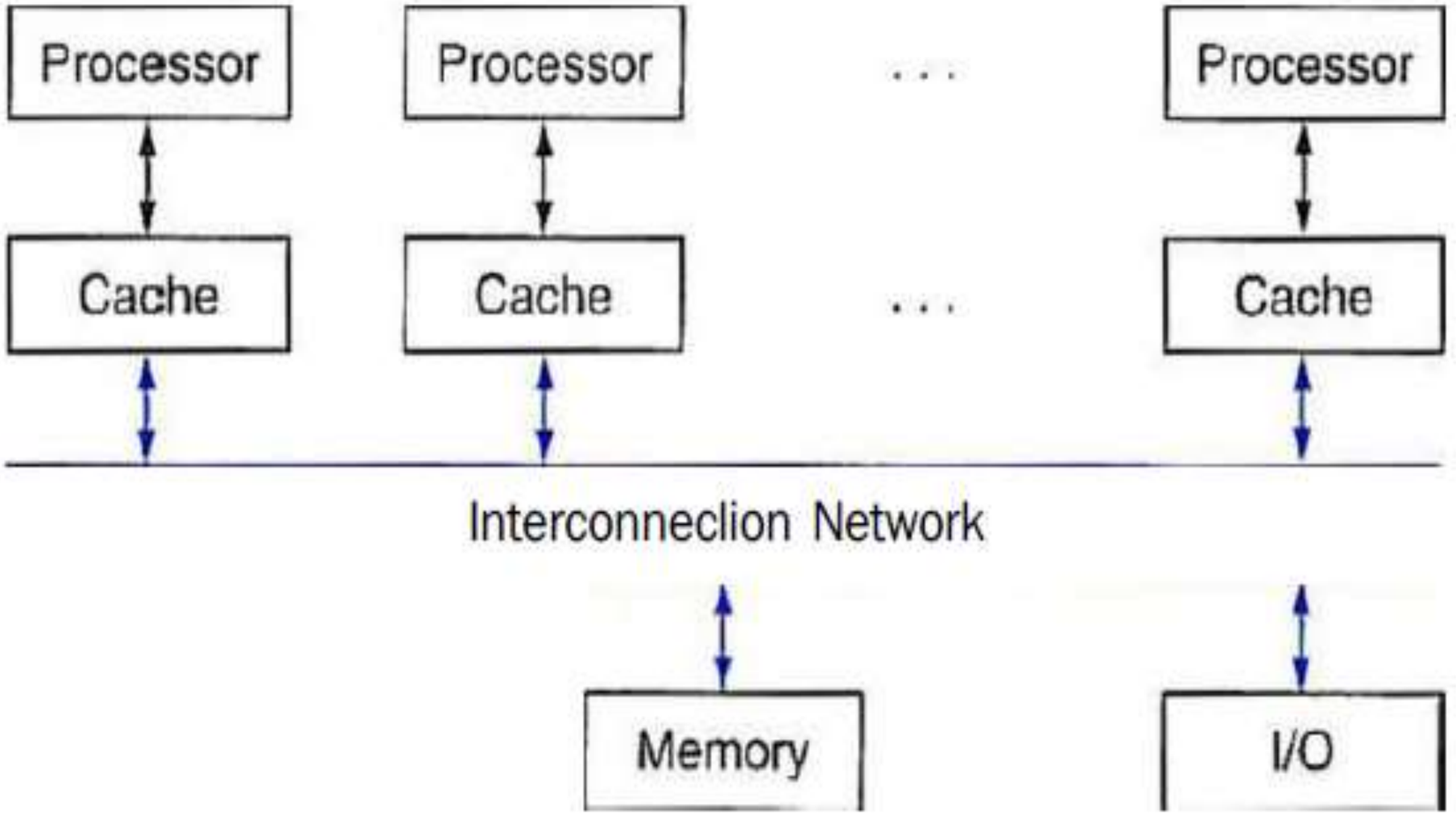
A type of single address space multiprocessor in which some memory accesses are much faster than others depending on which processor asks for which word.

synchronization

The process of coordinating the behaviour of two or more processes, which may be running on different processors.

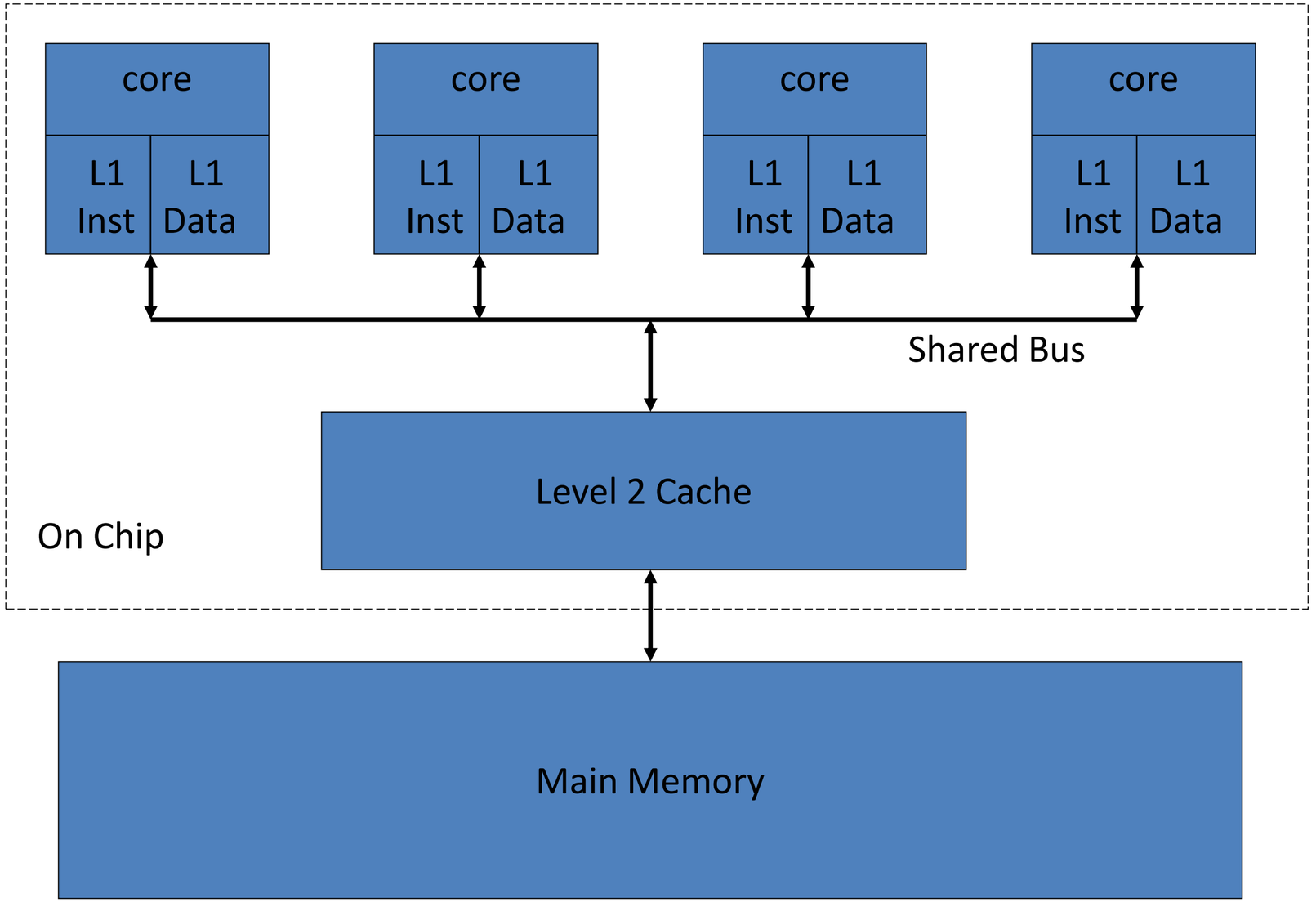
lock

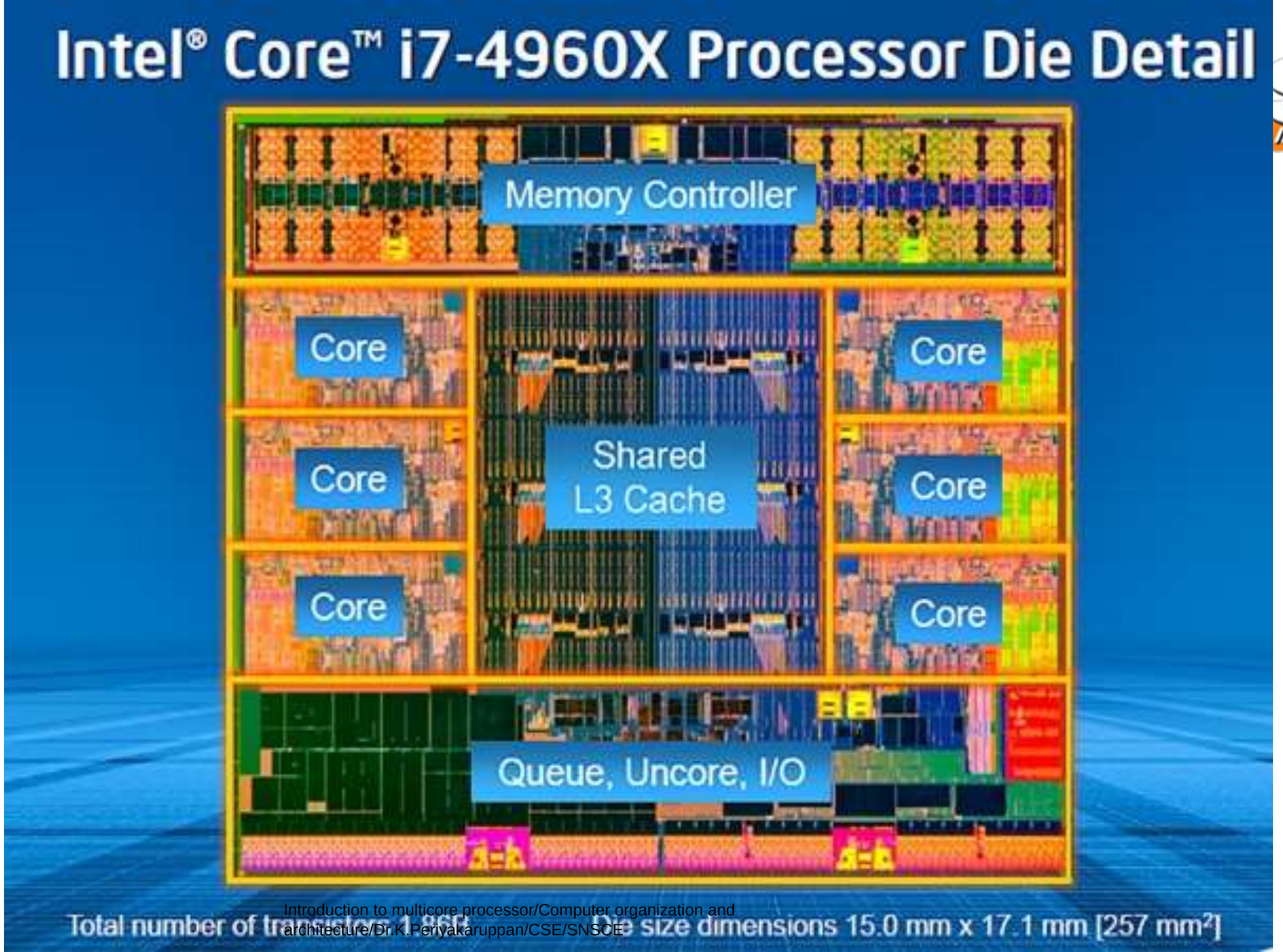
A synchronization device that allows access to data to only one processor at a time.



Classic organization of a shared memory multiprocessor.

Simplified Multi-Core Structure




Intel® Core™ i7-4960X Processor Die Detail
Memory Controller
Core
Core
Core
Core
Core
Core
Shared
L3 Cache
Queue, Uncore, I/O
Total number of transistors and die size dimensions 15.0 mm x 17.1 mm [257 mm²]

Advantages

Signals between different CPUs travel shorter distances, those signals degrade less.

These higher quality signals allow more data to be sent in a given time period since individual signals can be shorter and do not need to be repeated as often.

A dual-core processor uses slightly less power than two coupled single-core processors.

Graphics Processing Units

- Early video cards
  - Frame buffer memory with address generation for video output
- 3D graphics processing
  - Originally high-end computers (e.g., SGI)
  - Moore's Law $\Rightarrow$ lower cost, higher density
  - 3D graphics cards for PCs and game consoles
- Graphics Processing Units
  - Processors oriented to 3D graphics tasks
  - Vertex/pixel processing, shading, texture mapping, rasterization

Introduction to Graphics Processing Units

- A **GPU cluster** is a **computer cluster** in which each node is equipped with a **Graphics Processing Unit** (**GPU**). By harnessing the computational power of modern **GPUs** via General-Purpose **Computing** on **Graphics Processing Units** (GPGPU), very fast calculations can be performed with a **GPU cluster**.
- A graphics processing unit (**GPU**) is a **computer** chip that performs rapid mathematical calculations, primarily for the purpose of rendering images.

- GPU contains a collection of multithreaded SIMD processors; that is, a GPU is a MIMD composed of multithreaded SIMD processors. For example,
- NVIDIA has four implementations of the Fermi architecture at different price points with 7, 11, 14, or 15 multithreaded SIMD processors.

GPU Architectures

- GPU contains a collection of multithreaded SIMD processors; that is, a GPU is a MIMD composed of multithreaded SIMD processors
- Processing is highly data-parallel
  - GPUs are highly multithreaded
  - Use thread switching to hide memory latency
    - Less reliance on multi-level caches
  - Graphics memory is wide and high-bandwidth
- Trend toward general purpose GPUs
  - Heterogeneous CPU/GPU systems
  - CPU for sequential code, GPU for parallel code
- Programming languages/APIs
  - DirectX, OpenGL
  - C for Graphics (Cg), High Level Shader Language (HLSL)
  - Compute Unified Device Architecture (CUDA)

GPU hardware has two levels of hardware schedulers:
1. The *Thread Block Scheduler that assigns blocks of threads to multithreaded SIMD* processors, and
2. the SIMD Thread Scheduler within a SIMD processor, which schedules when SIMD threads should run.

The SIMD instructions of these threads are 32 wide, so each thread of SIMD instructions would compute 32 of the elements of the computation.

Since the thread consists of SIMD instructions, the SIMD processor must have parallel functional units to perform the operation.

We call them *SIMD Lanes, and they are quite similar to the Vector* Lanes .

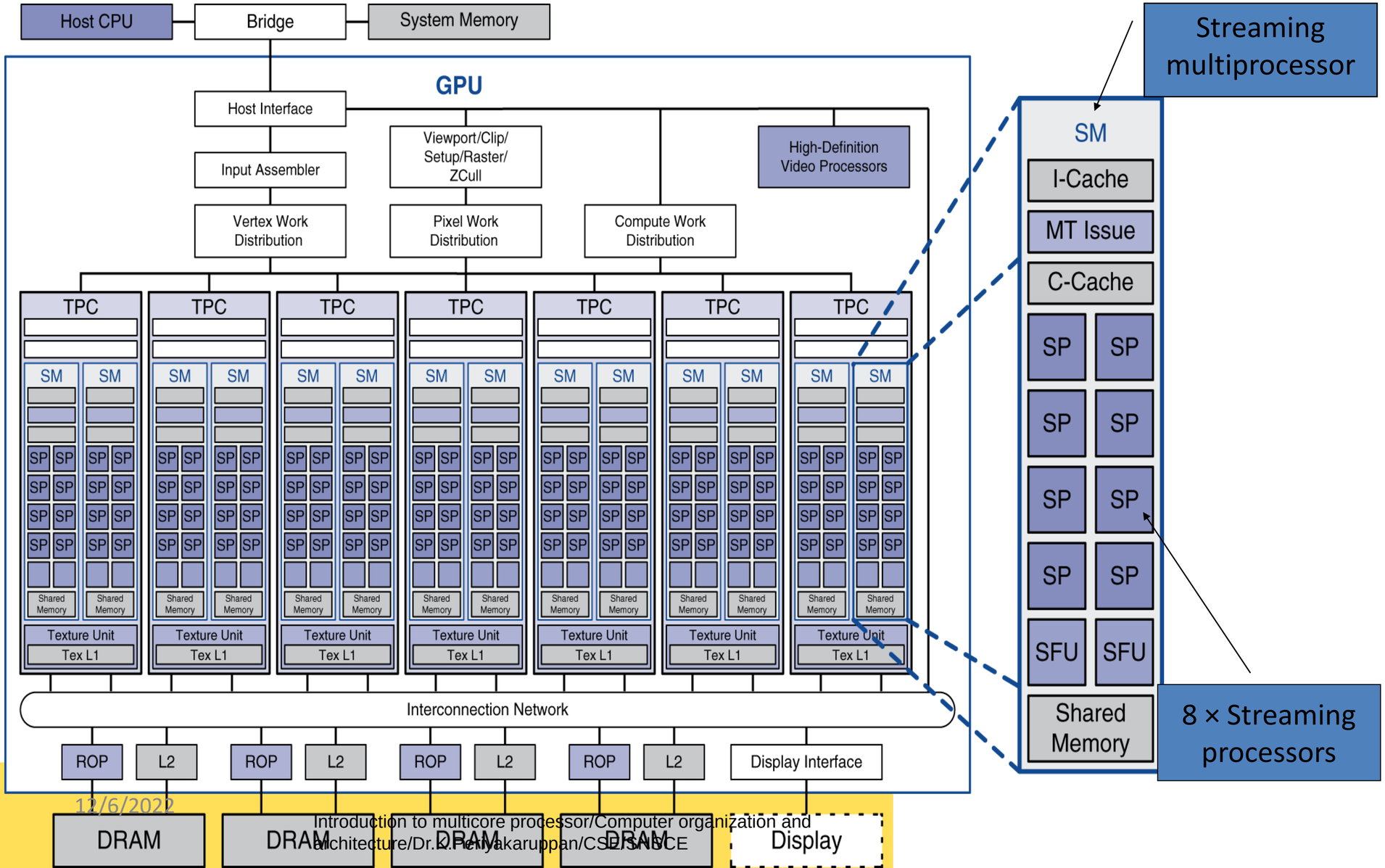
The render output unit(ROP), **ROPs** control antialiasing, when more than one sample is merged into one pixel

TPC – Texture Processing Clusters

Texture / Processor **Cluster**, is a group made up of several SMs, a **texture** unit and some logic control. ... The SP is the real **processing** element that acts on vertex or pixel data. Several TPCs can be grouped in higher level entity called a Streaming Processor Array.

Z-Cull: The elimination of certain objects from the display that should not be rendered because they are behind closer objects(depth test)

Example: NVIDIA Tesla



Assessment

What is Multicore?
What are the advantages of multicore?
What is UMA?
What is GPU?

Reference

1. Carl Hamacher, Zvonko Vranesic and Safwat Zaky, “Computer Organization”, McGraw-Hill, 6th Edition 2012.