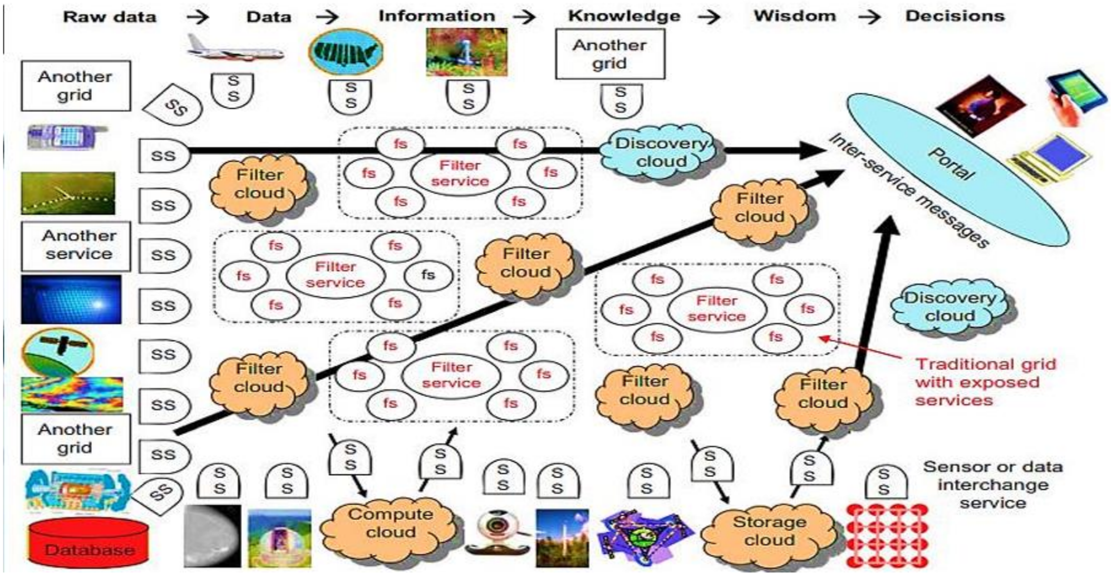
USN											

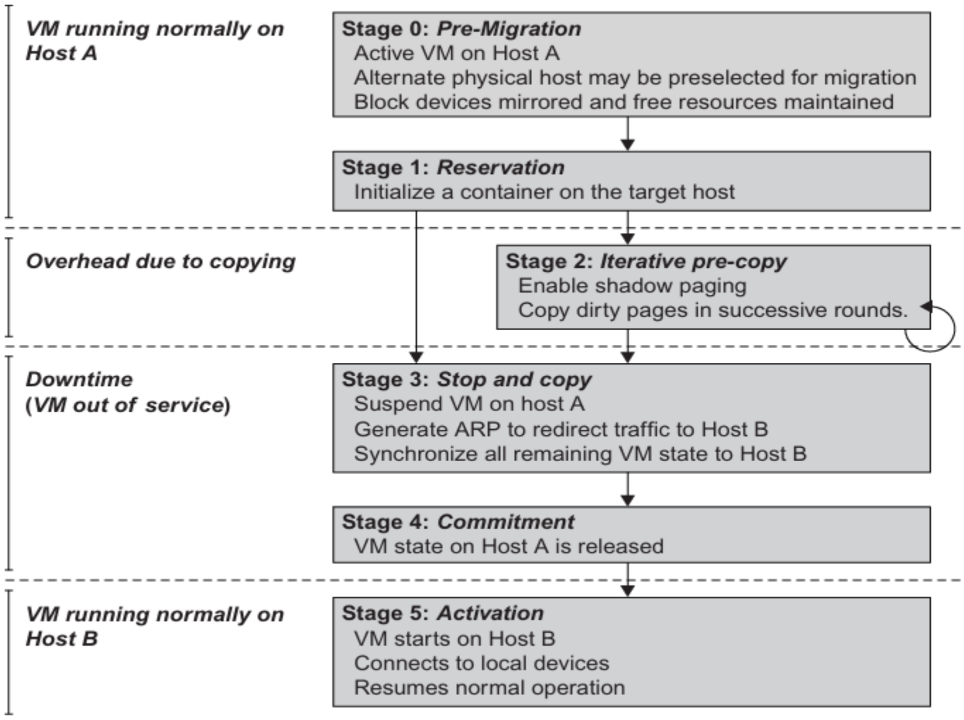
Internal Assessment Test 1 – Mar 2025

Sub:	**Cloud Computing and Security**				Sub Code:	**BIS613D**	Branch:	**ISE**
Date:	**27/03/2025**	Duration:	**90 min**	Max Marks: **50**	Sem/Sec:	**VI/ A, B & C**		**OBE**

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1.	**With a neat diagram, explain the Evolution of SOA in detail.** **Scheme : Definition + explanation + Diagram – 2+5+3 Marks** **Solution :** 1.SOA and Distributed Systems: SOA is applicable for building complex systems like grids, clouds, interclouds (grids of clouds), and systems of systems, supporting interoperability across diverse services. 2. Sensor Services (SS): Sensors (e.g., ZigBee, Bluetooth, WiFi, GPS, wireless phones) provide raw data through sensor services, which are crucial in data collection for various systems. 3. Data Collection and Interaction: Sensor services (SS) interact with various computing entities like small or large computers, grids, and cloud services to manage and process collected data. 4. Clouds and Services: The collected data is processed across different types of clouds such as compute, storage, filter, and discovery clouds, each playing a specific role in managing data and services. 5. Filter Services (FS): Filter services (FS) are used to process and clean raw data, ensuring that only relevant data is passed on to respond to specific requests from web services, grids, or the cloud. 	[10]	1	L2
2.	**Illustrate the steps involved in Live migration process of a VM from one host to another.** **Scheme : Definition + explanation + Diagram – 2+5+3 Marks** **Solution :** A VM can be in one of the following four states. • An inactive state is defined by the virtualization platform, under which the VM is not enabled. • An active state refers to a VM that has been instantiated at the virtualization platform	[10]	2	L2

to perform a real task.

- A paused state corresponds to a VM that has been instantiated but disabled to process a task or paused in a waiting state.
- A VM enters the suspended state if its machine file and virtual resources are stored back to the disk.



Steps 0 and 1: Start migration. This step makes preparations for the migration, including determining the migrating VM and the destination host. Although users could manually make a VM migrate to an appointed host, in most circumstances, the migration is automatically started by strategies such as load balancing and server consolidation.

Steps 2: Transfer memory. Since the whole execution state of the VM is stored in memory, sending the VM's memory to the destination node ensures continuity of the service provided by the VM. All of the memory data is transferred in the first round, and then the migration controller recopies the memory data which is changed in the last round. These steps keep iterating until the dirty portion of the memory is small enough to handle the final copy. Although precopying memory is performed iteratively, the execution of programs is not obviously interrupted.

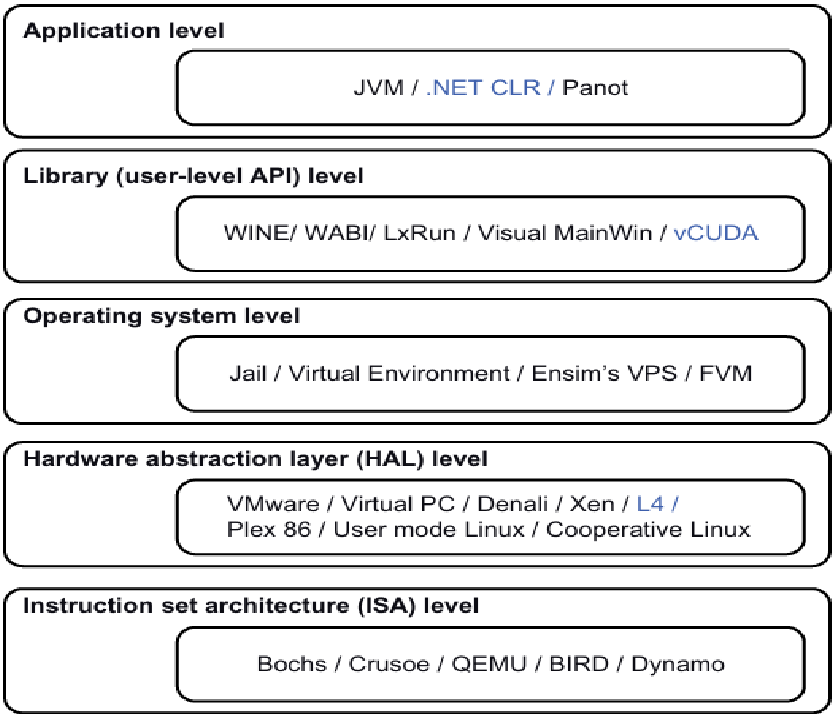
Step 3: Suspend the VM and copy the last portion of the data. The migrating VM's execution is suspended when the last round's memory data is transferred. Other nonmemory data such as CPU and network states should be sent as well. During this step, the VM is stopped and its applications will no longer run. This "service unavailable" time is called the "downtime" of migration, which should be as short as possible so that it can be negligible to users.

Steps 4 and 5: Commit and activate the new host. After all the needed data is copied, on the destination host, the VM reloads the states and recovers the execution of programs in it, and the service provided by this VM continues. Then the network connection is redirected to the new VM and the dependency to the source host is cleared. The whole migration process finishes by removing the original VM from the source host.

No.	Question	Marks	CO	Level
3.	**Explain the mechanism of Virtualization ranging from hardware to applications in five abstraction levels.**	[10]	1	L2

Scheme : Definition + explanation with Diagram for each – 3+2+3+2 Marks

Solution :



- At the ISA level, virtualization is performed by emulating a given ISA by the ISA of the host machine. For example, MIPS binary code can run on an x86-based host machine with the help of ISA emulation.
- Hardware-level virtualization is performed right on top of the bare hardware. On the one hand, this approach generates a virtual hardware environment for a VM.
- OS-level virtualization creates isolated containers on a single physical server and the OS instances to utilize the hard ware and software in data centers.
- Most applications use APIs exported by user-level libraries rather than using lengthy system calls by the OS. Since most systems provide well-documented APIs, such an interface becomes another candidate for virtualization.
- application-level virtualization is also known as process-level virtualization. The most popular approach is to deploy high level language (HLL) VMs. In this scenario, the virtualization layer sits as an application program on top of the operating system, and the layer exports an abstraction of a VM that can run programs written and compiled to a particular abstract machine definition.

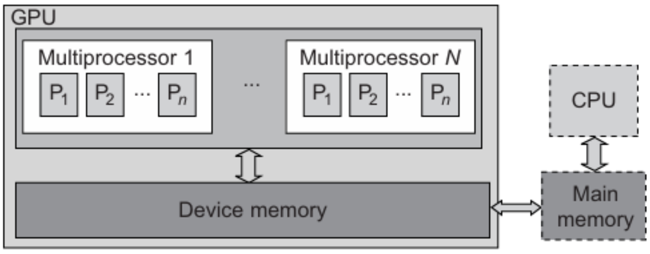
No.	Question	Marks	CO	Level
4.	**a. Explain GPU Programming Model with an example.**	[6]	1	L2

Scheme : Explanation + Example – 3+3 Marks

Solution :

- The CPU is the conventional multicore processor with limited parallelism to exploit.
- The GPU has a many-core architecture that has hundreds of simple processing cores organized as multiprocessors.
- Each core can have one or more threads. Essentially, the CPU's floating-point kernel computation role is largely offloaded to the many-core GPU.
- The CPU instructs the GPU to perform massive data processing. The bandwidth must be matched between the on-board main memory and the on-chip GPU memory.

- This process is carried out in NVIDIA's CUDA programming using the GeForce 8800 or Tesla and Fermi GPUs.

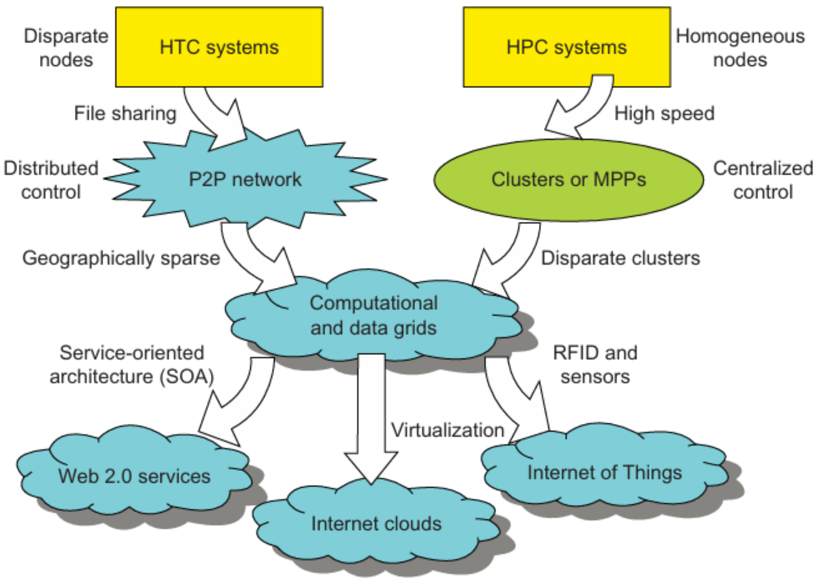


- The NVIDIA Fermi GPU Chip with 512 CUDA Cores In November 2010, three of the five fastest supercomputers in the world (the Tianhe-1a, Nebulae, and Tsubame) used large numbers of GPU chips to accelerate floating-point computations.
- This is a streaming multiprocessor (SM) module. Multiple SMs can be built on a single GPU chip. The Fermi chip has 16 SMs implemented with 3 billion transistors. Each SM comprises up to 512 streaming processors (SPs), known as CUDA cores. The Tesla GPUs used in the Tianhe-1a have a similar architecture, with 448 CUDA cores.

b. Explain the evolution of HPC and HTC systems. [4]

Scheme : Explanation + Diagram – 2+2 Marks

Solution :



- On the HPC side, supercomputers (massively parallel processors or MPPs) are gradually replaced by clusters of cooperative computers out of a desire to share computing resources. The cluster is often a collection of homogeneous compute nodes that are physically connected in close range to one another.
- On the HTC side, peer-to-peer (P2P) networks are formed for distributed file sharing and content delivery applications. A P2P system is built over many client machines.

5. Consider a program for multiplying two large-scale N × N matrices, where N is the matrix size. The sequential multiply time on a single server is T1 = cN3 minutes, where c is a constant determined by the server used. An MPI-code parallel program requires Tn = cN3/n +dN2/n0.5 minutes to complete execution on an n-server cluster system, where d is a constant determined by the MPI version used. Assume the program has a zero sequential bottleneck (α =0). The second term in Tn accounts for the total message-passing overhead experienced by n servers. Answer the following questions for a given cluster configuration with n =64 servers, c =0.8, and d = 0.1. Parts (a, b) have a fixed workload corresponding to the | [10] | 1 | L3

matrix size N = 15,000. Parts (c, d) have a scaled workload associated with an enlarged matrix size N′ = n1/3 N=641/3 × 15,000 = 4 × 15,000 = 60,000. Assume the same cluster configuration to process both workloads. Thus, the system parameters n, c, and d stay unchanged. Running the scaled workload, the overhead also increases with the enlarged matrix size N′.

a. Using Amdahl's law, calculate the speedup of the n-server cluster over a single server.
b. What is the efficiency of the cluster system used in Part (a)?
c. Calculate the speedup in executing the scaled workload for an enlarged N′ × N′ matrix on the same cluster configuration using Gustafson's law.
d. Calculate the efficiency of running the scaled workload in Part (c) on the 64-processor cluster.
e. Compare the above speedup and efficiency results and comment on their implications.

Scheme : Obtaining solution for each carries – 2+1+4+1+2 Marks

Solution :

(a) Speedup using Amdahl's Law:

Amdahl's law is used to predict the theoretical maximum speedup for a given parallelizable portion of the workload. The formula for speedup S_n is:

$$S_n = \frac{T_1}{T_n} = \frac{1}{(1-P) + \frac{P}{n}}$$

Where:

- T_1 is the time for the sequential program (single server).
- T_n is the time for the parallel program on n servers.
- P is the proportion of the workload that is parallelizable.

In this case, it's mentioned that there is zero sequential bottleneck, so $P = 1$.

Thus, the speedup S_n simplifies to:

$$S_n = \frac{1}{(1-1) + \frac{1}{n}} = n$$

So the speedup for $n = 64$ servers is:

$$S_n = 64$$

(b) Efficiency of the Cluster System:

The efficiency E_n of the parallel system is defined as:

$$E_n = \frac{S_n}{n}$$

Substituting the values:

$$E_n = \frac{64}{64} = 1$$

The efficiency is 1, meaning the parallel system is perfectly efficient when there is no overhead or sequential bottleneck (ideal case).

(c) Speedup using Gustafson's Law (for scaled workload):

Gustafson's law is used when the problem size scales with the number of processors. It takes into account the increased workload when more resources are available. The formula for speedup S_n under Gustafson's law is:

$$S_n = n + (1-n) \times \frac{T_1}{T_1 + T_n}$$

Here, since the matrix size increases proportionally with $n^{1/3}$, the scaled problem is effectively $N' = 4 \times N$. This causes the parallelizable work to increase, and so we expect the speedup to be better than the one predicted by Amdahl's law.

For the scaled workload, the matrix size increases from 15,000 to 60,000. The computation time now becomes:

$$T_1' = cN'^3 = 0.8 \times 60,000^3$$

The time for n processors is:

$$T_n' = \frac{cN'^3}{n} + \frac{dN'^2}{n^{0.5}}$$

The speedup S_n is then:

$$S_n = \frac{T_1}{T_n} = \frac{cN'^3}{\frac{cN'^3}{n} + \frac{dN'^2}{n^{0.5}}}$$

Substituting the given values:

$$S_n = \frac{0.8 \times 60,000^3}{\frac{0.8 \times 60,000^3}{64} + \frac{0.1 \times 60,000^2}{64^{0.5}}}$$

Simplifying and computing these terms gives the speedup.

(d) Efficiency for the Scaled Workload:

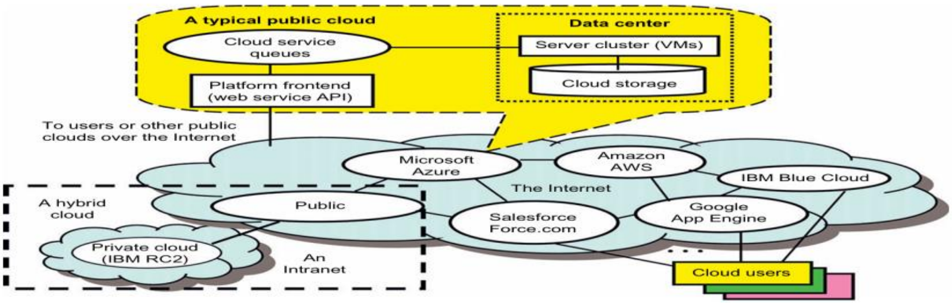
To find the efficiency E_n for the scaled workload:

$$E_n = \frac{S_n}{n}$$

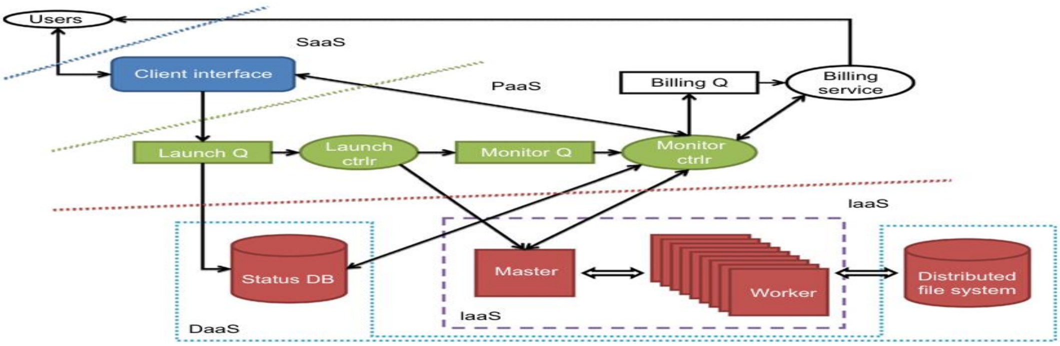
Where S_n is the speedup calculated in part (c).

(e) Comparison of Speedup and Efficiency:

- **Amdahl's law (fixed workload)** predicts limited speedup and perfect efficiency because of the zero sequential bottleneck assumption.
- **Gustafson's law (scaled workload)** predicts a much higher speedup since the problem size increases with the number of processors, allowing better utilization of the system. However, efficiency will likely decrease due to the increased overhead from the message-passing term.

<table>
<tr>
<td>6.</td>
<td>
a. Explain the architecture of a typical public cloud.
Scheme : Explanation + Diagram – 3+2 marks
Solution :


A public cloud is built over the Internet and can be accessed by any user who has paid for the service. Public clouds are owned by service providers and are accessible through a subscription.
A private cloud is built within the domain of an intranet owned by a single organization. Therefore, it is client owned and managed, and its access is limited to the owning clients and their partners. Its deployment was not meant to sell capacity over the Internet through publicly accessible interfaces.
A hybrid cloud is built with both public and private clouds. Private clouds can also support a hybrid cloud model by supplementing local infrastructure with computing capacity from an external public cloud.

</td>
<td>[5]</td>
<td rowspan="2">2</td>
<td rowspan="2">L2</td>
</tr>
<tr>
<td></td>
<td>
b. Illustrate three cloud models at different service levels of the cloud.
Scheme : Explanation + Diagram – 3+2 marks
Solution :


SaaS is applied at the application end using special interfaces by users or clients. At the PaaS layer, the cloud platform must perform billing services and handle job queuing, launching, and monitoring services.
At the bottom layer of the IaaS services, databases, compute instances, the file system, and storage must be provisioned to satisfy user demands.
This model allows users to use virtualized IT resources for computing, storage, and networking. In short, the service is performed by rented cloud infrastructure.
The user can deploy and run his applications over his chosen OS environment. The user does not manage or control the underlying cloud infrastructure, but has control over the OS, storage, deployed applications, and possibly select networking components.
This IaaS model encompasses storage as a service, compute instances as a service, and communication as a service.

</td>
<td>[5]</td>
</tr>
</table>

Faculty Signature CCI Signature HOD Signature