UNIT - IV

Cloud Resource management and Scheduling:

⇒ policies and mechanisms for resource management

⇒ Applications of Control Theory to task Scheduling on a cloud.

⇒ Stability of a Two level Resource Allocation Architecture.

⇒ Feedback Control Based on Dynamic thresholds

⇒ Coordination of Specialized Automic Performance managers.

⇒ Resource Bundling

⇒ Scheduling Algorithms for computing Cloud - Fair Queuing

⇒ Start Time Fair Queuing.

Policies and mechanisms for resource management

⇒ A Policy typically refers to the principal guiding decisions.

⇒ whereas mechanisms represent the means to implement policies.

⇒ Separation of policies from mechanisms is a guiding principle in computer science.

⇒ Cloud resource management policies can be loosely grouped into five classes:

1. Admission control
2. capacity allocation
3. Load balancing
4. Energy optimization
5. Quality-of-Service (QoS) guarantees.

⇒ The explicit goal of an admission control policy is to prevent the system from accepting workload in violation of high-level system policies.

Ex: A system may not accept an additional workload that would prevent it from completing work already in progress.

⇒ capacity allocation means to allocate resources for individual instances.

⇒ An instance is an Activation of a service.

⇒ Load balancing and energy optimization policies encounter the same difficulties as the one.

⇒ Load balancing and energy optimization are correlated and affect the cost of providing the service.

⇒ Quality of service is that aspect of resource management that is probably the most difficult to address and, at the same time, possibly the most critical to the future of cloud computing.

⇒ The four basic mechanisms for the implementation of resource management policies are:

1. control theory:

Control theory uses the feedback to guarantee system stability and predict transient behavior

2. machine Learning:

A major advantage of machine learning technique is that they do not need a performance model of the system

3. utility-based:

utility-based approaches require a performance model and a mechanism to correlate user-level performance with cost.

4. Market oriented/ economic mechanism:-

Such mechanisms do not require a model of the system.

Applications of control theory to task scheduling on a cloud:-

⇒ control theory has been used to design adaptive resource management for many classes of applications, including power management, task scheduling, QoS Adaptation in web servers, and Load balancing.

⇒ the classical feedback control method are used in all these cases to regulate the key operating parameters of the system based on measurement of the system output.

⇒ the feedback control in these methods assumes a linear time-invariant system model and a closed-loop controller.

⇒ The main components of a Control System:

* The Inputs:

The offered workload and the policies for admission control, the capacity allocation, the load balancing.

* The Control System components:

The sensors used to estimate relevant measures of performance and controllers.

* The outputs:

The resource allocations to the individual applications.

⇒ Feedback and Stability:

* Control granularity the level of detail of the information used to control the system.

⇒ The controllers use the feedback provided by sensors to stabilize the system.

⇒ Stability is related to the change of the output disturbance

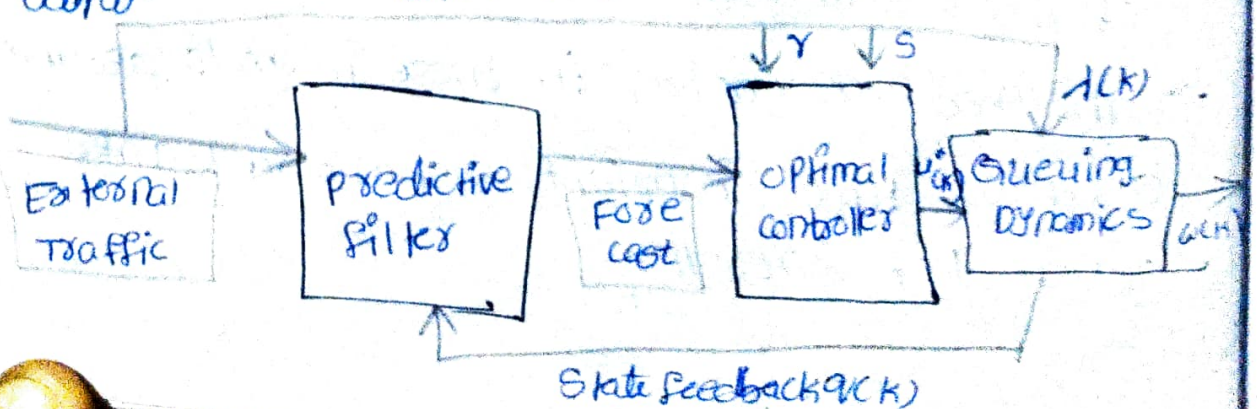


Fig: The structure of a cloud controller

⇒ the controller uses the feedback regarding the current state and the estimation of the future disturbance due to environment to compute the optimal inputs over a finite horizon.

⇒ γ and s are the weighting factors of the performance index.

Task scheduling in the cloud computing:-

⇒ cloud computing is made up of large number of datacenters that house numerous physical machines (host).

⇒ Each host runs several virtual machines (VMs) that are in charge of executing user tasks with varying quality of services (QoS).

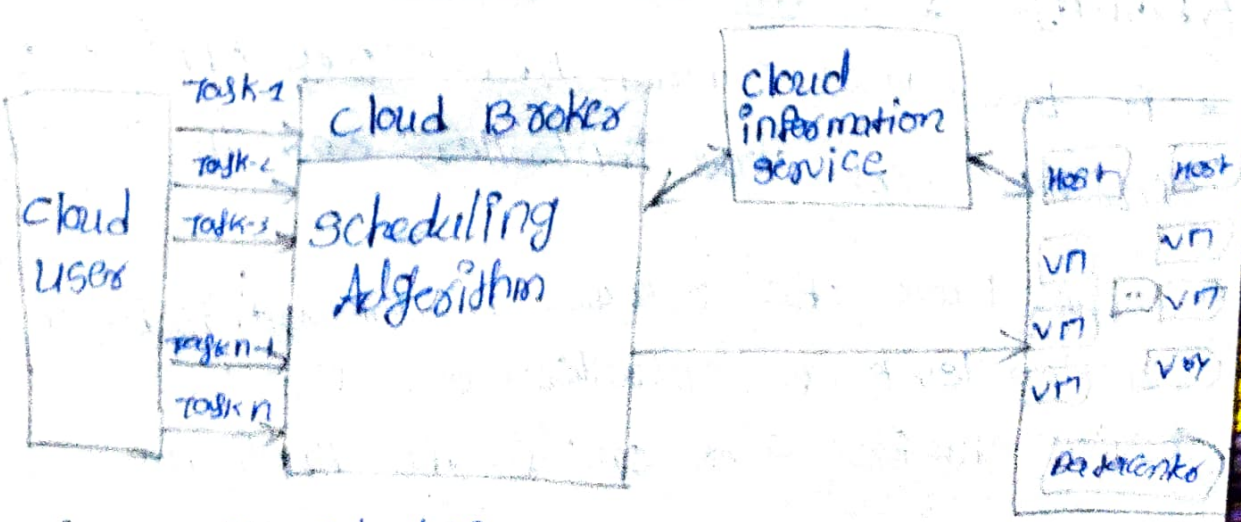


Fig: Task Scheduling process in the cloud computing

⇒ Supposing that there are n tasks

$T = T_1, T_2, T_3 \ldots T_n$, which are executed using 'm' virtual machines $VM = VM_1, VM_2, VM_3 \ldots, VM_m$.

⇒ These tasks have various lengths and the VMs are heterogenous in terms of bandwidth, RAM and CPU Time.

⇒ The cloud broker makes a request to the cloud information service, to obtain details about the services.

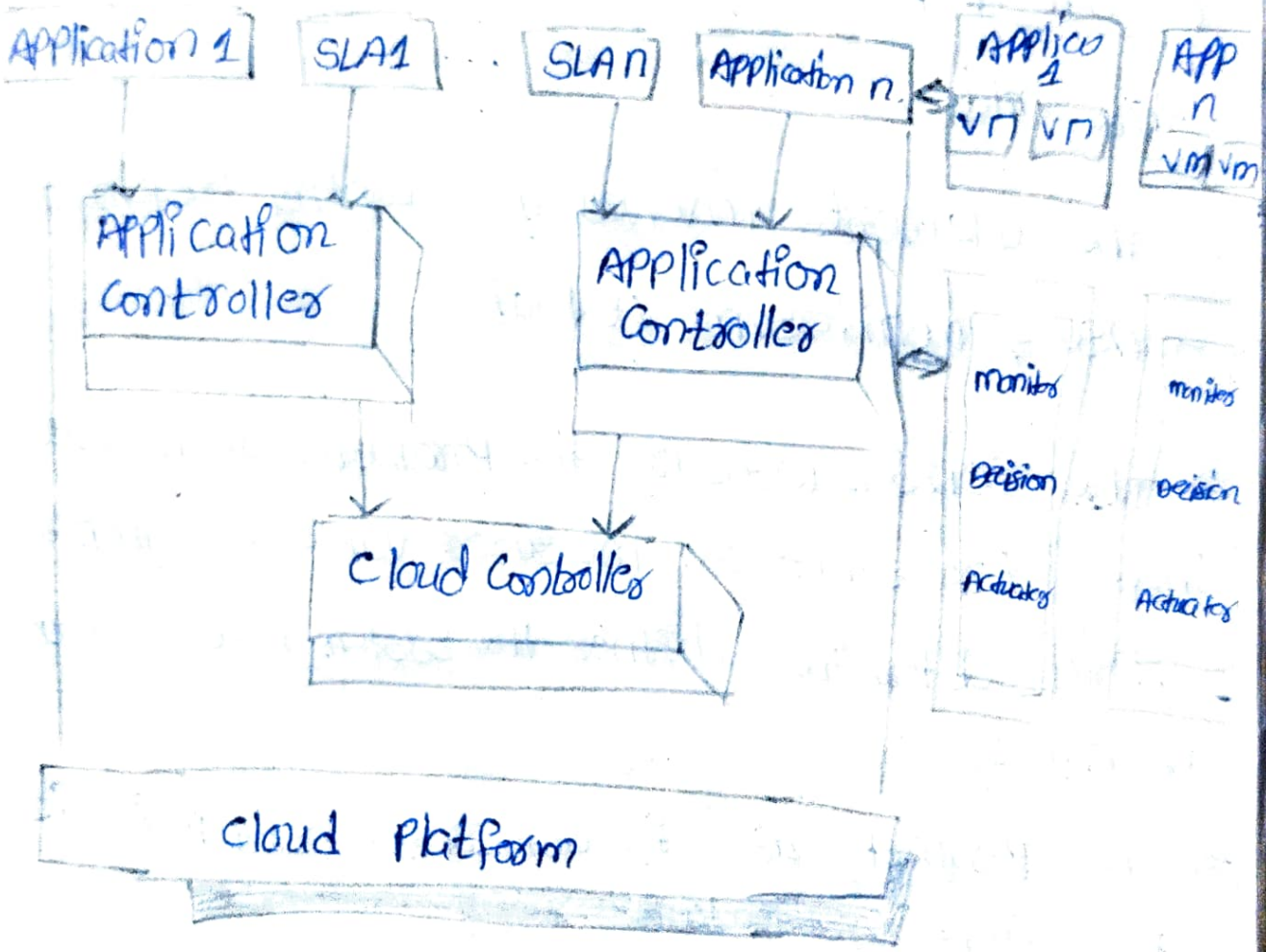
⇒ The cloud broker is the key element of the task scheduling process.

⇒ The number of tasks entered It can follow FCFS principle.

Stability of a two level Resource Allocate Architecture

⇒ A two-level control Architecture Application controllers and cloud controllers work in concert.

⇒ The Automatic resource management is based on two levels of controllers, one for the service provider and one for the Application.



⇒ Two types of policies are used in Autonomic systems:
(i) threshold - based Policies and
(ii) sequential decision Policies

⇒ The main components of a control system are: the inputs, the control system components, and the outputs.

⇒ The Actions consist of allocation/ deallocation of one or more virtual machines

Feedback ~~based~~ Control based on Dynamic thresholds :-

⇒ The elements involved in a control system are sensors, monitors and Actuators.

⇒ The sensors measure the parameter of interest then transmit the measured values to a monitor.

⇒ which determine whether the system behaviour must be changed

⇒ The parameter used for admission control policy in the current system load

Ex:-

when a threshold 80% is reached the cloud stops accepting additional load.

Thresholds :

⇒ A threshold is the value of a parameter related to the state of a system that triggers a change in the system behavior.

⇒ thresholds are used in control theory to keep critical parameters of a system in a predefined range.

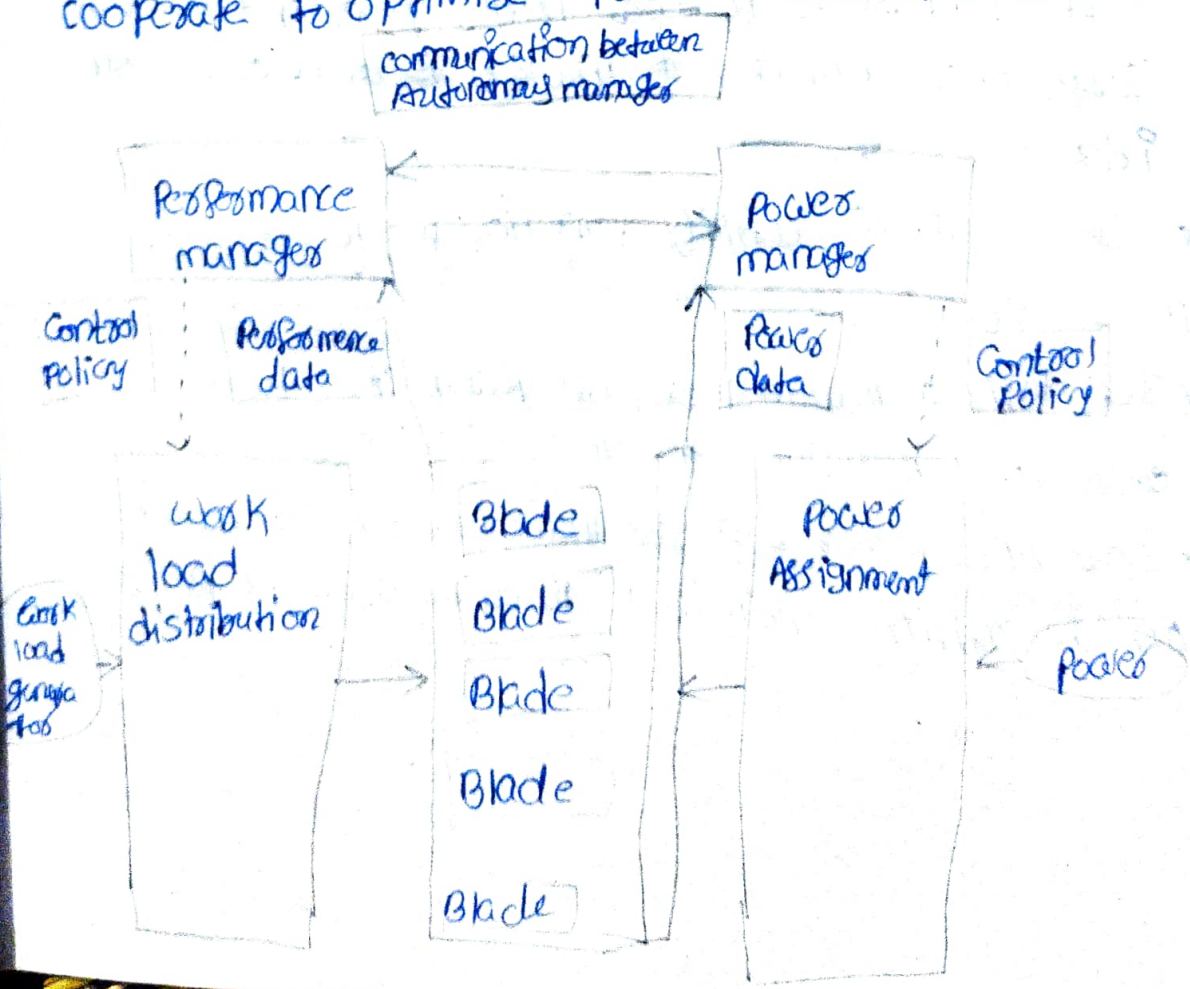
⇒ The Threshold can be represented in a static and dynamic.

⇒ The essence of the proportional thresholding is captured by the following algorithm:

1. Compute the integral value of the high and the low thresholds
2. Request additional VMs when the average value of the CPU utilization.
3. Release a VM when the average value of the CPU utilization.

Coordination of specilized Autonic Performance manager:-

⇒ Can specilized Autonic Performance managers cooperate to optimize power consumption.



⇒ Virtually all modern processors support dynamic voltage scaling (DVS) as a mechanism for energy saving.

⇒ The Power management controls the CPU frequency and, thus the rate of instruction execution.

⇒ For some compute-intensive workloads the performance decreases linearly with the CPU clock frequency.

⇒ The approach to coordinating Power and Performance management in [187] is based on several ideas:

* Use a joint utility function for Power and Performance.
* Identify a minimal set of parameters to be exchanged between the two managers
* set up a Power cap for individual systems based on the utility-optimized Power management policy.

Resource Bundling :-

⇒ Resources in a cloud are allocated in bundles.

⇒ Users get maximum benefit from a specific combination of resources: CPU cycles, main memory, disk space, network bandwidth and so on.

⇒ Resource bundling complicates traditional resource allocation models and has generated an interest in economic models and, in particular in Auction algorithms.

⇒ The bidding process aims to optimize an objective function $f(x, P)$.

⇒ In the context of cloud computing, an Auction is the allocation of resources to the highest bidder

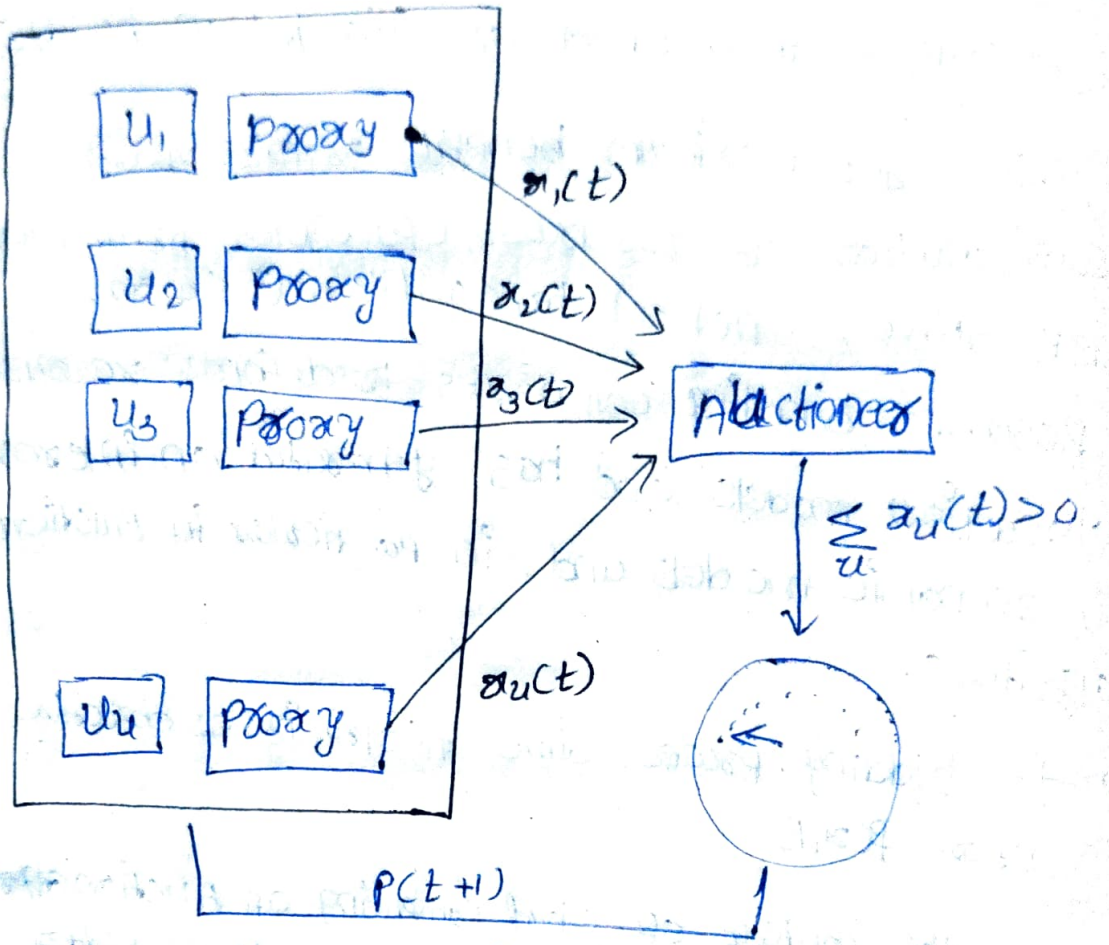
combinatorial auctions for cloud resources :-

⇒ Simultaneous clock Auction, clock Proxy Auction and ascending clock auction are known combinatorial Auctions.

⇒ Users provide bids for desirable bundles and the price they are willing to pay.

⇒ Prices and allocation are set as a result of an Auction.

⇒ Ascending clock Auction (ASCA) - the current price for each resource is represented by a clock.



⇒ The schematics of the ASCA algorithms, to allow for a single round Auction users are represent Processes At time $x_u(t)$.

Pricing and Allocation Algorithms:-

⇒ A Pricing and Allocation Algorithms partitions the set of users in two disjoint sets, winners and losers.

⇒ Desirable Properties of a pricing Algorithm:

- Be computational tractable
- scale well
- Be objective
- Be fair
- The unit prices
- The supply and the demand

Scheduling Algorithms for cloud computing:-

⇒ Scheduling means which is responsible for resource sharing at several levels

* A server can be shared among several virtual machines.
* A virtual machine could support several Applications
* An Application may consist of multiple threads

⇒ A Scheduling Algorithm should be efficient, fair and starvation-free

⇒ The objective of a scheduler:

Batch System: maximize throughput and minimize turnaround time

Real-time system: meet the deadlines and Predictable

⇒ common algorithms for best effort applications:

* Round-Robin
* First-Come-First-Serve
* Shortest-Job-First
* priority Algorithm.

⇒ multimedia Applications have soft real-time constraints

⇒ Real-time Applications have hard real-time constraints.

Fair queuing :-

⇒ computing and communication on a cloud are intimately related.

⇒ The multiple packets are arriving At the same time It uses FCFS scheduling Algorithm.

⇒ The Advantage of the FCFS algorithm is a simple management of the three quantities : bandwidth, timing and buffer space.

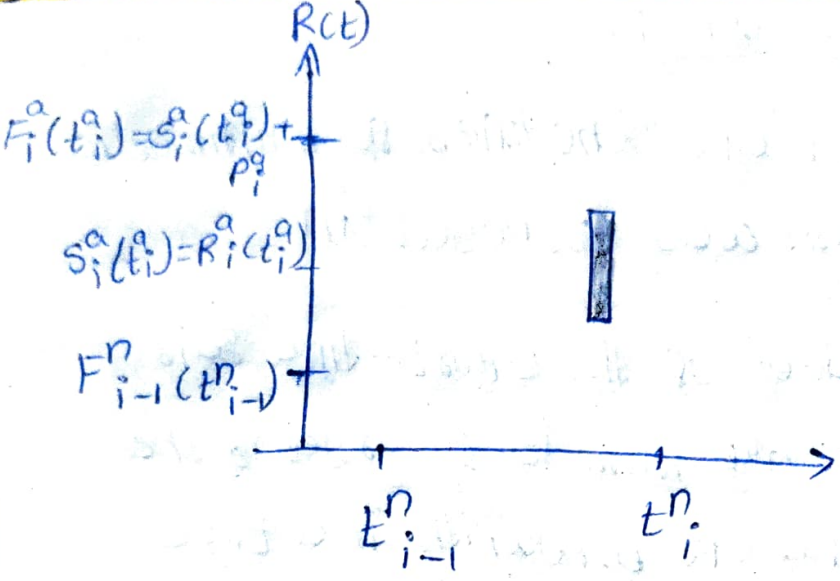
⇒ But the FCFS Algorithm does not guarantee fairness.

⇒ A fair queuing Algorithm Proposed in ~~request~~ requires that separate queues, one per flow

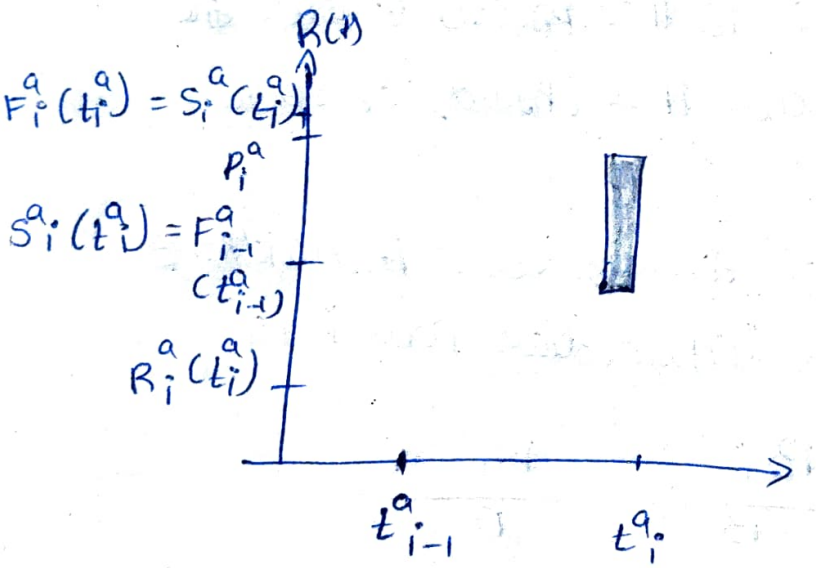
⇒ this algorithm guarantees the fairness of buffer space management, but does not guarantee fairness of bandwidth allocation.

⇒ The fair queing algorithm proposes a solution to this Problem, First it introduces a bit-by-bit round-robin Strategy

⇒ A Fair allocation of the bandwidth does not have an effect on the timing of the Transmission.



(a)



(b)

Fig :-

Transmission of a packet 'i' of flow a arriving at the t_i^a of size P_i^a bits.

Start-time fair queuing:-

⇒ A hierarchical CPU scheduler for multimedia operating system was proposed in.

⇒ The basic idea of the start-time fair queuing (SFQ) algorithm is to organize the consumers of the CPU bandwidth in a tree structure.

⇒ The root node is the processor and the leaves of this are the threads of each application.

⇒ The fraction of the processor bandwidth B allocated to the intermediate node i is

$$\frac{B_i}{B} = \frac{w_i}{n}$$

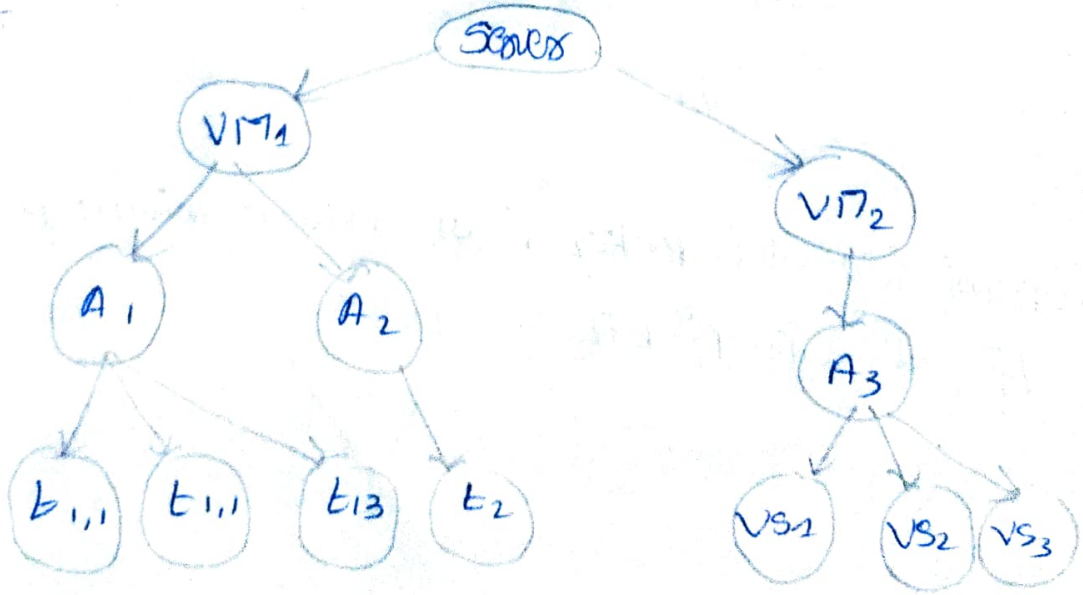


Fig: The SFQ tree for scheduling when two virtual machines VM_1 and VM_2 run on a Powered Server.

The Age of Internet Computing:

1.1 The platform evolution:-

⇒ It describes how the interplay of technical Architecture and governance guides the evolution of the platform.

⇒ each generation has a new features to develop

⇒ The general computing trend is to leverage shared web resources and massive amounts of data over the Internet.

⇒ on the HPC side, Supercomputers are gradually replaced by Clusters of cooperative computers out of a desire to share computing resource.

⇒ The Clusters is a collection of homogeneous compute nodes that are physically connected in close range.

⇒ on the HTC side, peer-to-peer networks are formed for distributed file sharing and content delivery Applications.

⇒ clustering and P2P technologies lead to the development of computational grids or data grids.

⇒ P2P build many client machine.

1.6. Distributed System Families:-

⇒

⇒ Technologies for building P2P network and Networks of clusters have been designed to consider wide area computing infrastructure known as computational grids or data grids

2.