IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS
Sydney, Australia 10-14 JUNE 2014
COMMUNICATIONS: CENTREPOINT OF THE DIGITAL ECONOMY
IEEE
IEEE COMMUNICATIONS SOCIETY
CONNECT WITH IEEE ICC:

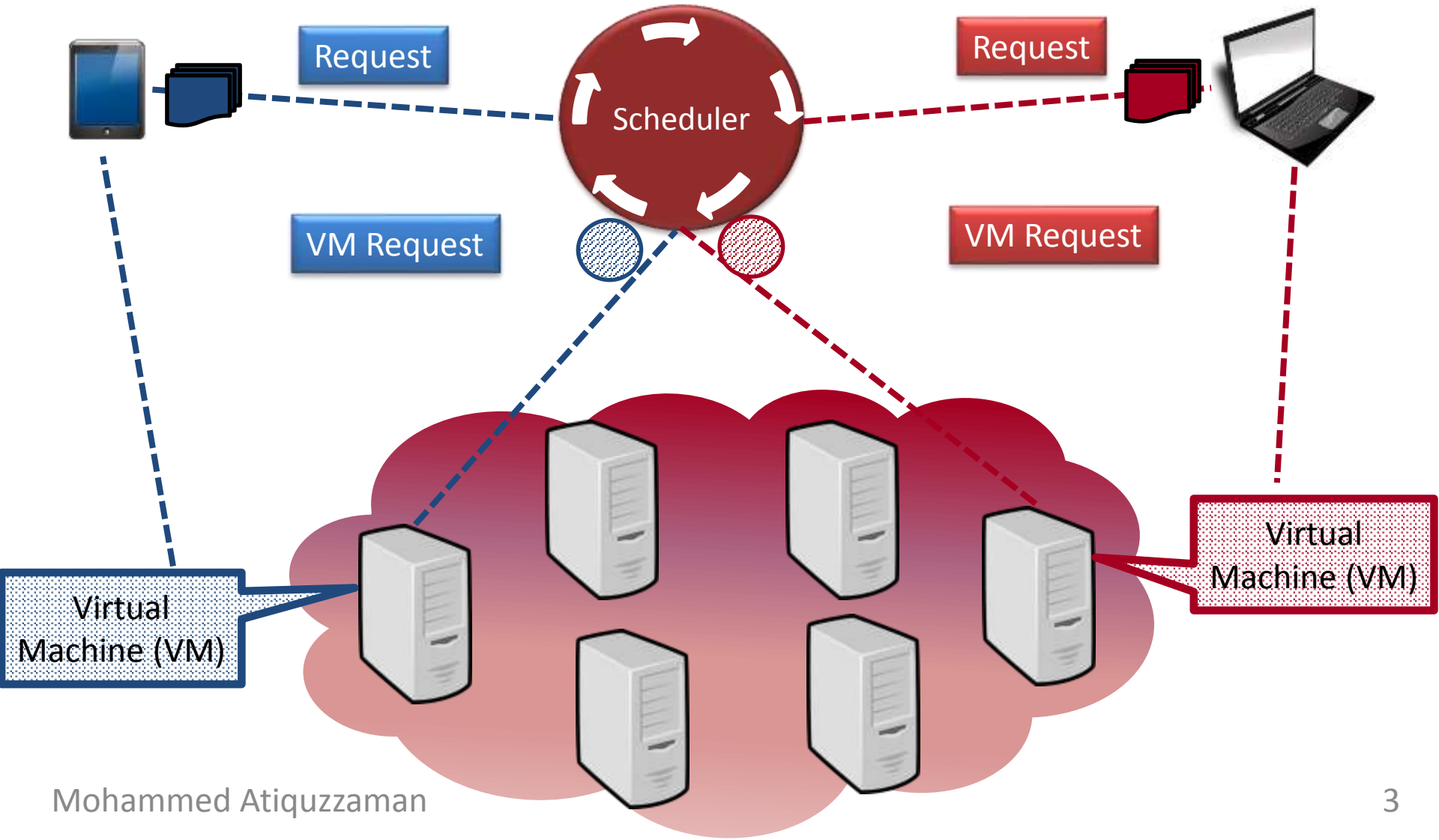
h-DDSS: Heterogeneous Dynamic Dedicated Servers Scheduling in Cloud Computing

Husnu Saner Narman
Md. Shohrab Hossain
Mohammed Atiquzzaman

School of Computer Science
University of Oklahoma, USA.
atiq@ou.edu
www.cs.ou.edu/~atiq

June 2014

What is Cloud Computing



Why Cloud Computing

- Simplicity
  - No need to set up software/hardware
- Flexibility
  - Easily extending memory/CPU capacity
- Maintenance
  - IT services
- Time and energy
  - No time or extra effort for desired environment
- Pay as you go
  - No need to pay for unused hardware or software

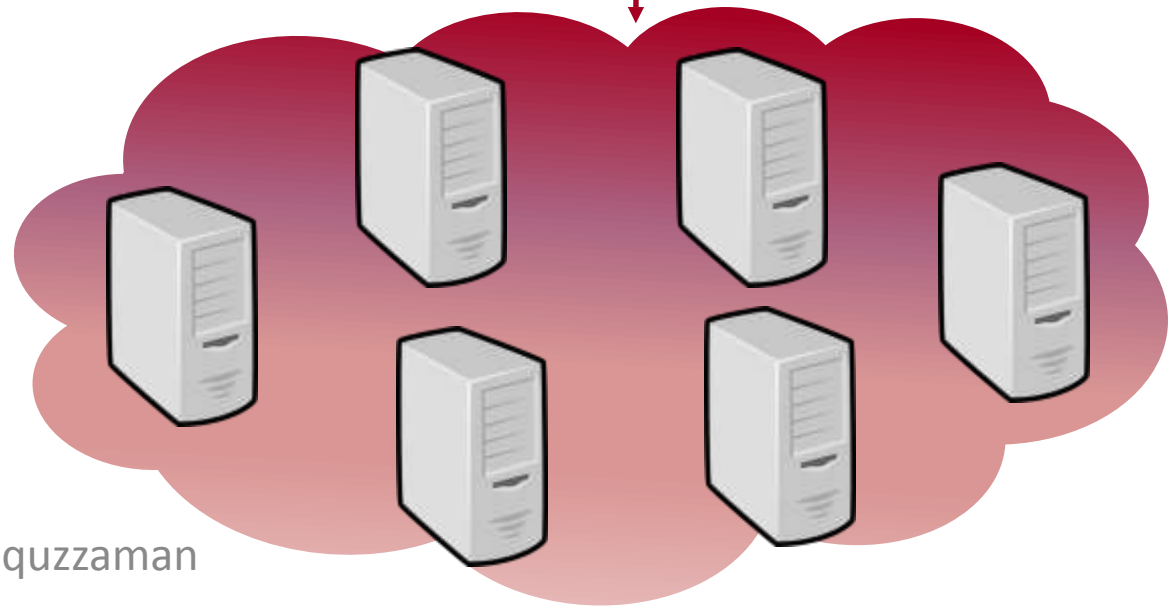
What is Cloud Scheduling

3. Assign VM to customer

1. Request

Scheduler

2. Find the best appropriate machine to create VM.



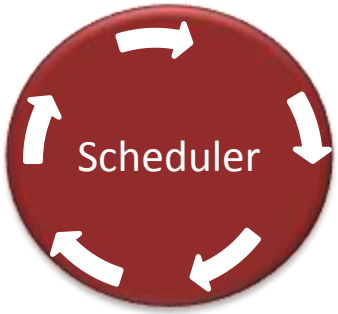
Customer Type

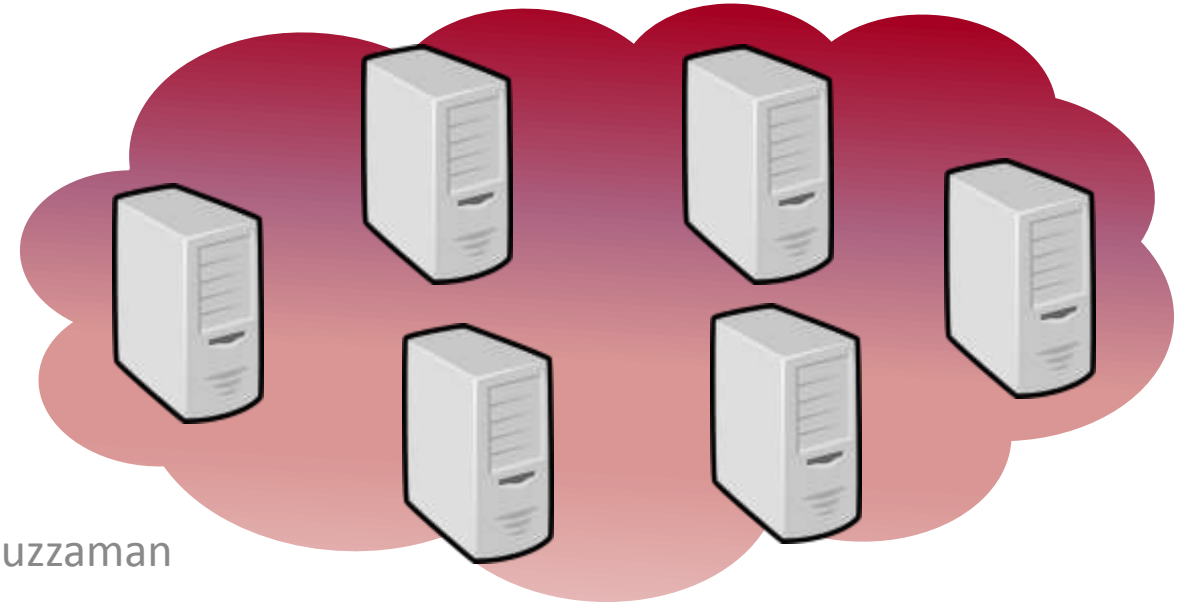
- Different customers classes?
  - Paid and non-paid
- Customer requirements
  - Desired Platform based on Service Level Agreement
- How to satisfy different customer classes?
  - Reserve servers for each customer types
    - Dedicated Servers Scheduling
  - Priority
    - High or Low

Customer Priority

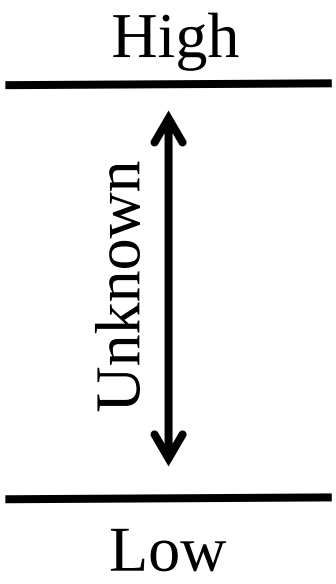
Non-paid (Low Priority)

Paid (High Priority)

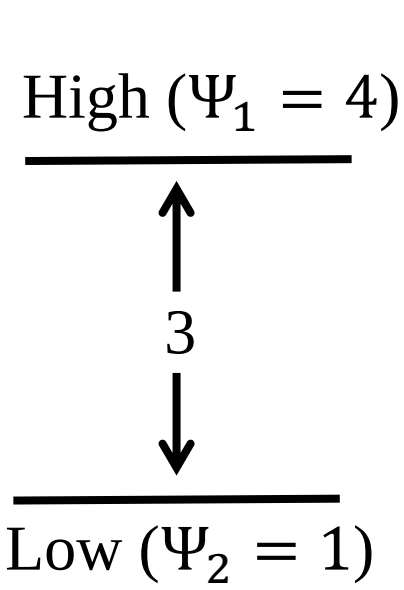


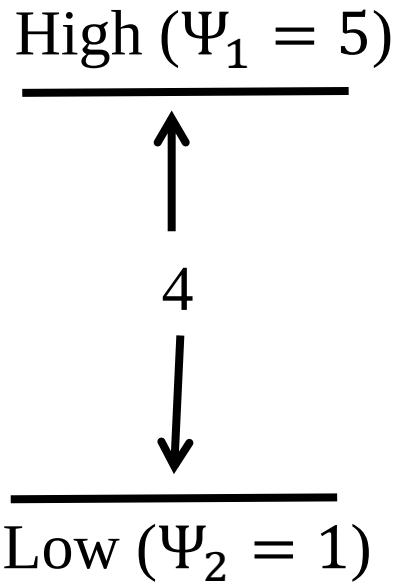


Priority Level



Without priority level in queuing theory





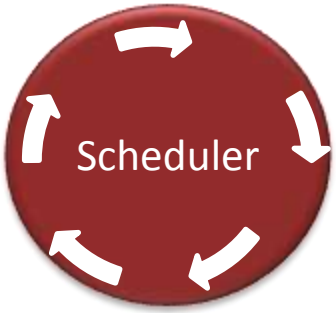
With priority level in cloud computing

Reserved Servers

Non-paid

Paid

How many servers are needed for each group of customers?



Non-paid Customer Servers

Paid Customer Servers

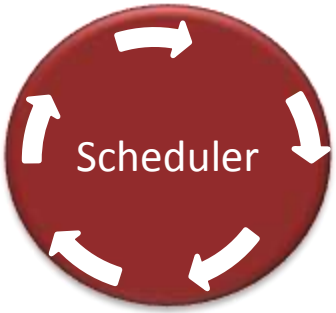
Dedicated Servers Scheduling

Non-paid

What happen when one type of customer arrival increases?

DSS: No update of number of servers for each group.

Paid



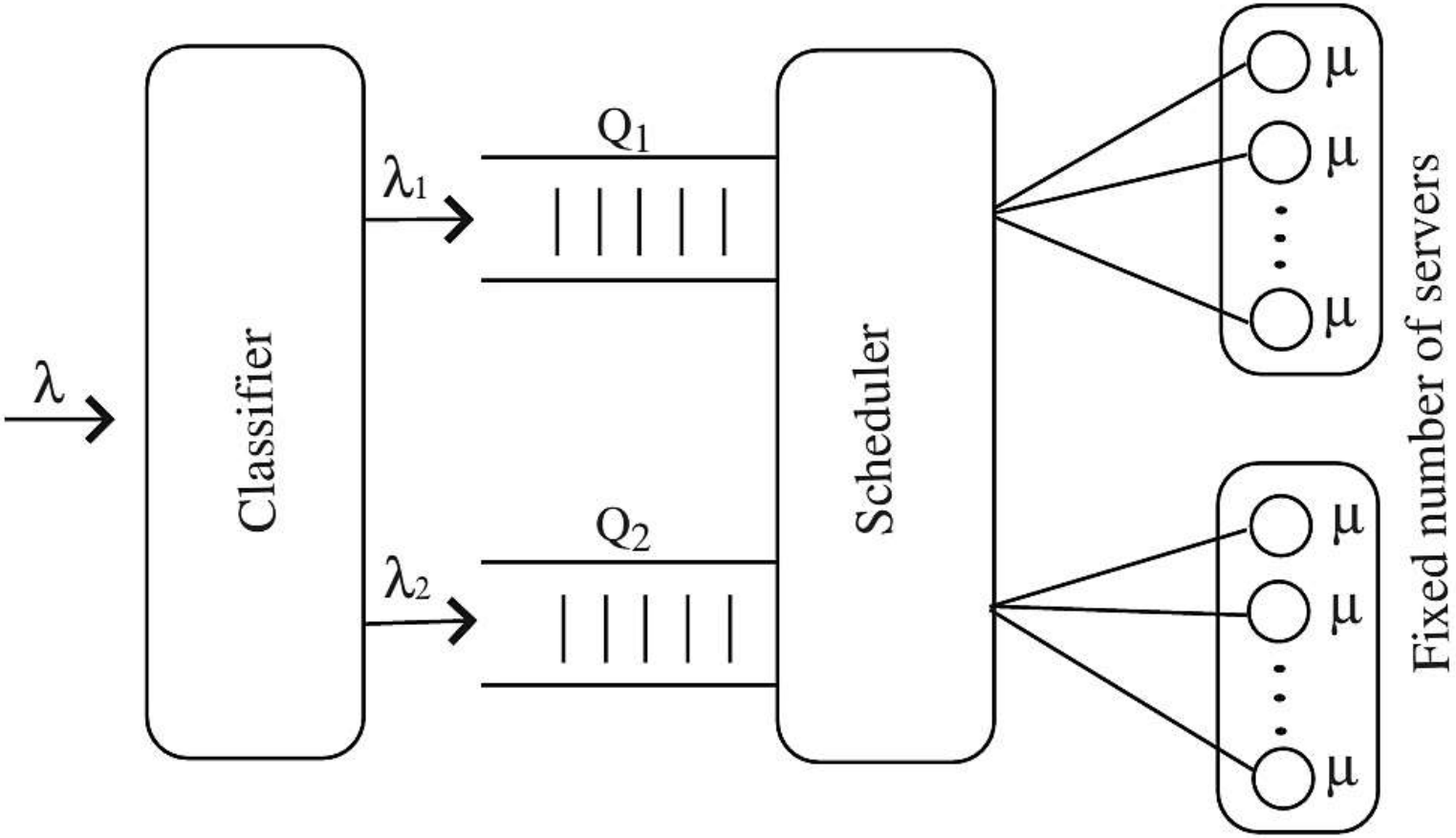
Assumption

Servers are homogeneous

Non-paid Customer Servers

Paid Customer Servers

Dedicated Servers Scheduling



Problems with DSS

- Does not dynamically update number of servers for each group
  - If arrival rate changes
  - If priority level changes
- Servers are homogeneous (Unrealistic)

Dynamic Dedicated Server Scheduling (DDSS)

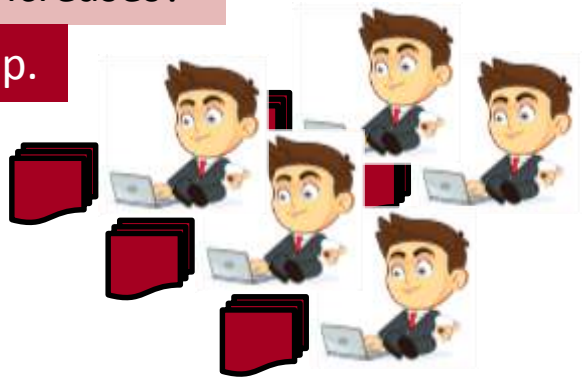
Dynamic Dedicated Servers Scheduling

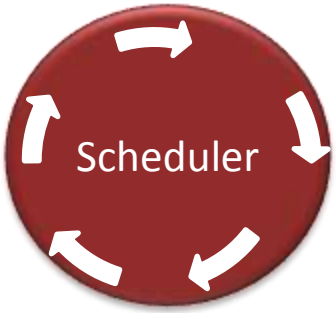
Non-paid

What happen when one type of customer arrival increases?

DDSS: Updating number of servers for each group.

Paid





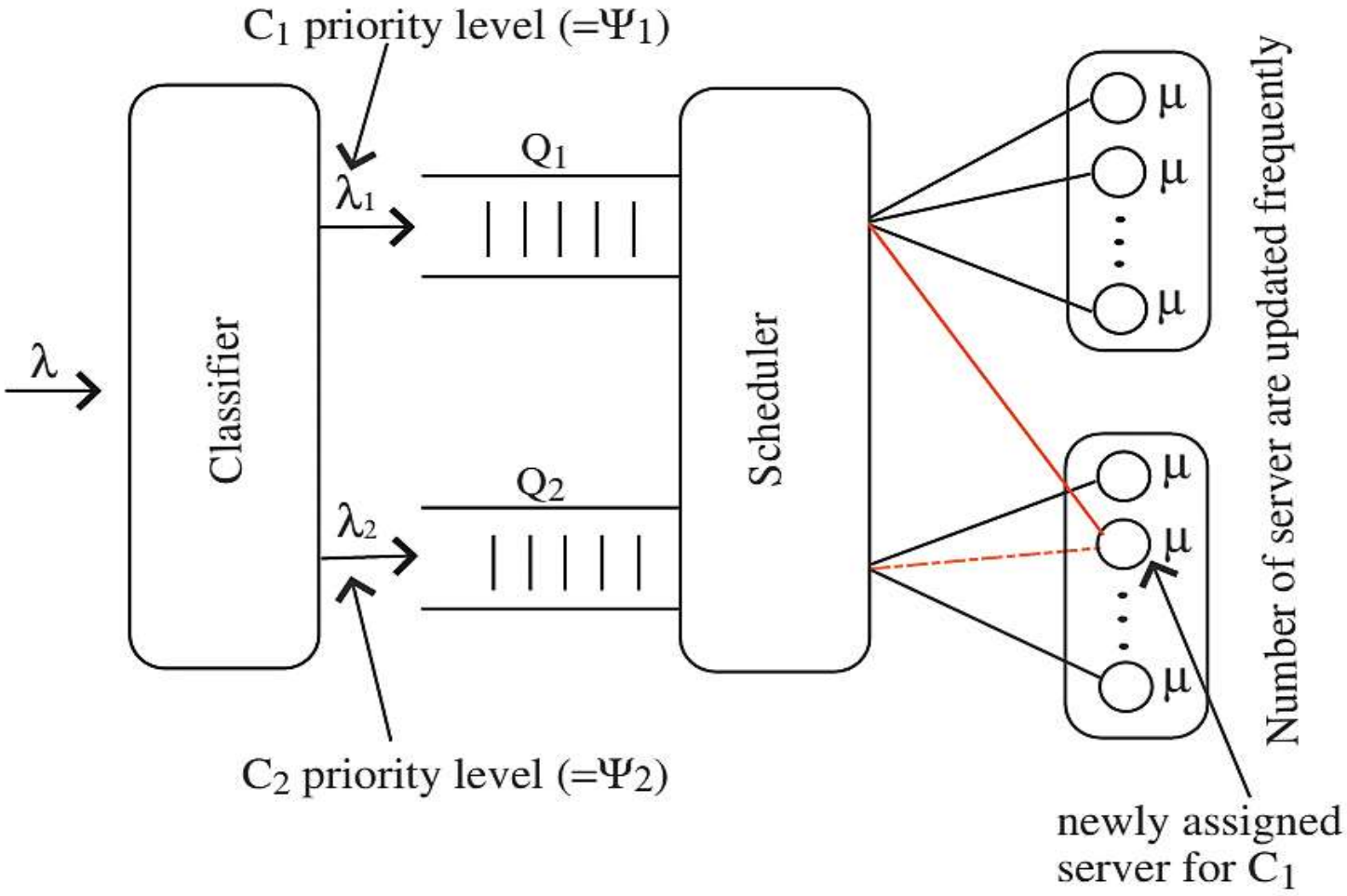
Assumption

Servers are homogeneous

Non-paid Customer Servers

Paid Costumer Servers

Dynamic Dedicated Servers Scheduling



Problems with DDSS

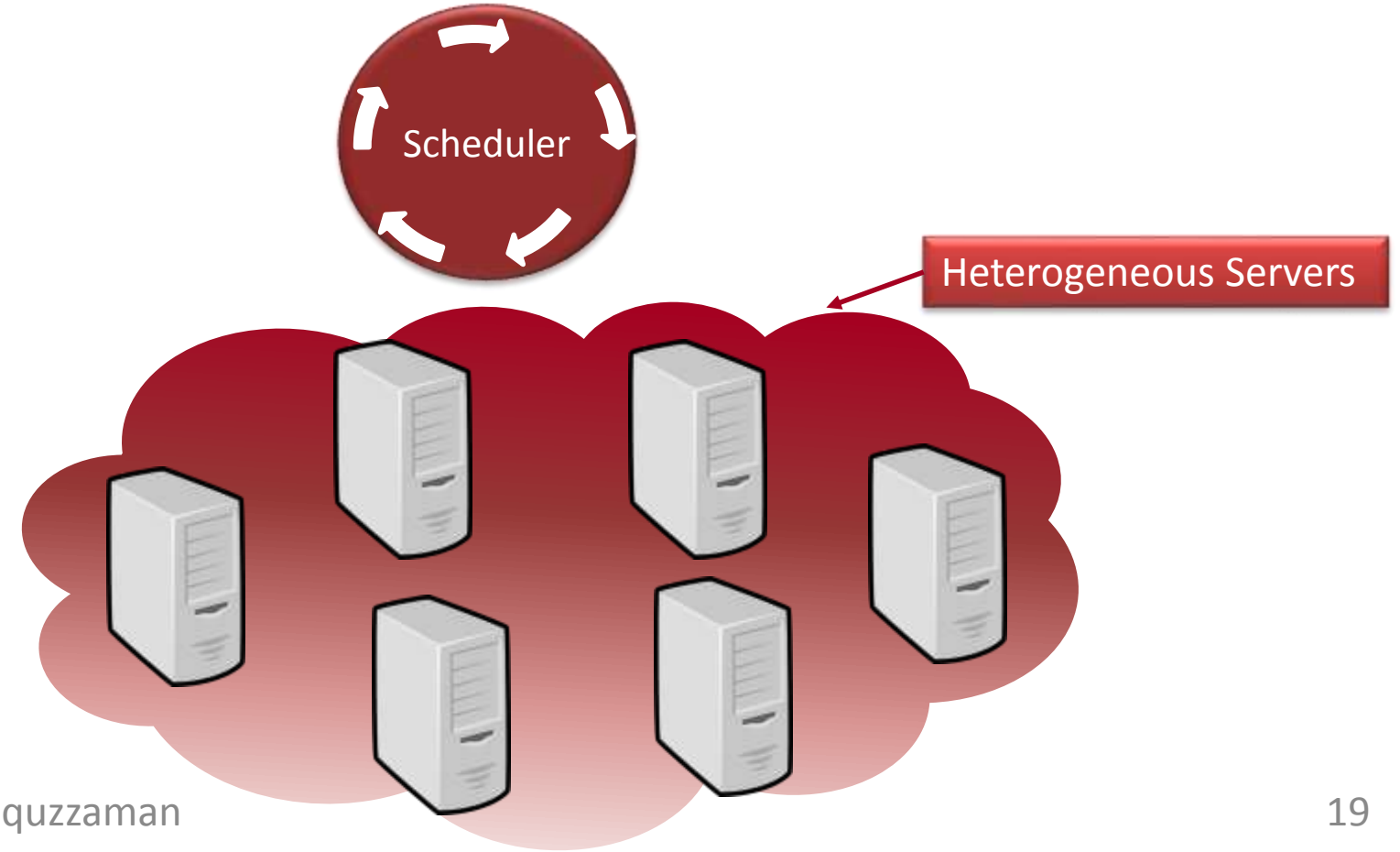
- Servers are homogeneous (Unrealistic)

Heterogeneous Dynamic Dedicated Server Scheduling (h-DDSS)

Why Heterogeneous

- Failed or misbehaved servers of a multi-server system are replaced by new and more powerful ones

Heterogeneous Servers



Objective

- Improve performance of cloud systems for heterogeneous servers
  - Allowing heterogeneous servers to be dynamically allocated to customer classes based on
    - Priority level.
    - Arrival rate.

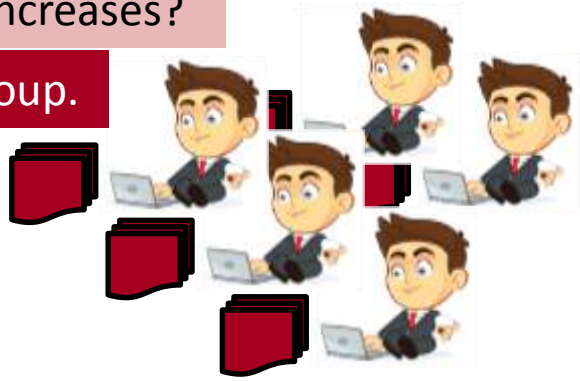
Contribution

- Propose Heterogeneous Dynamic Dedicated Servers Scheduling.
- Develop Analytical Model to evaluate performance
  - Average occupancy
  - Drop rate
  - Average delay
  - Throughput
- Comparing performance of
  - Heterogeneous Dynamic Dedicated Servers Scheduling
  - Dynamic Dedicated Servers Scheduling.

Heterogeneous Dynamic Dedicated Servers Scheduling

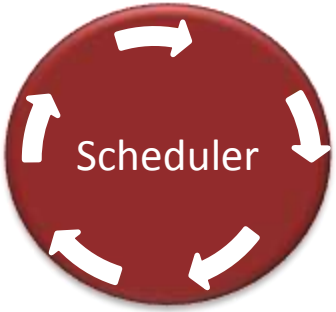
Non-paid

Paid



What happen when one type of customer arrival increases?

h-DDSS: Updating number of servers for each group.



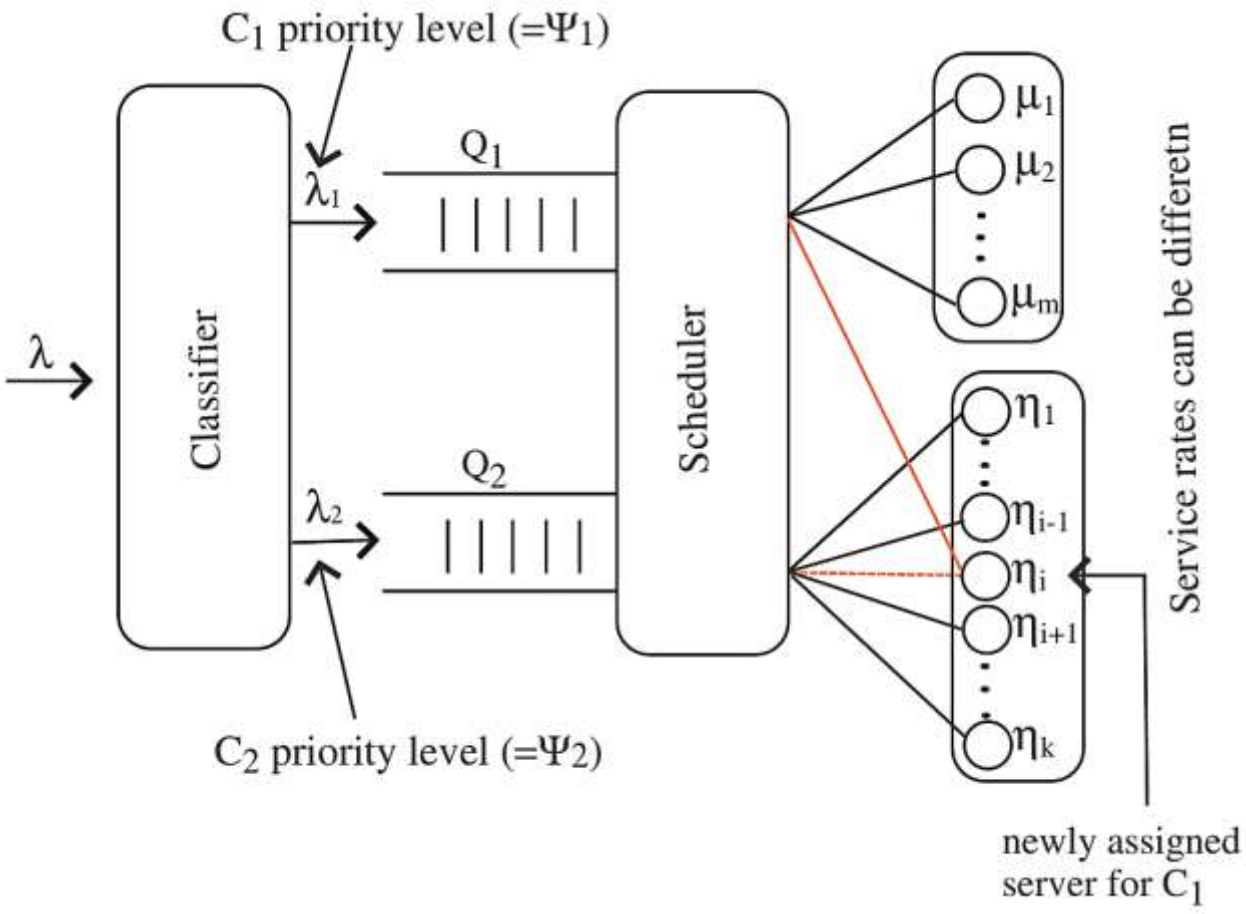
Assumption

Servers are heterogeneous (Realistic)

Non-paid Customer Servers

Paid Customer Servers

Heterogeneous Dynamic Dedicated Servers Scheduling



Dynamic Approach

μ_{total}: Total service rate of servers

Ψ_1: Priority level of C_1 customers

λ_1: Arrival rate of C_1 customers

$$\mu'_{tm} = \left\lfloor \frac{\mu_{total}\Psi_1\lambda_1}{\Psi_1\lambda_1 + \Psi_2\lambda_2} \right\rfloor$$

λ_2: Arrival rate of C_2 customers

μ_{tm}: Total service rate assigned for C_1 customers

$$\mu_{tm} = \sum_{i=1}^{m} \mu_i \approx \mu'_{tm}$$

Ψ_2: Priority level of C_2 customers

μ_i: Service rate of i server

η_{tk}: Total service rate assigned for C_2 customers

$$\eta_{tk} = \mu_{total} - \mu_{tm}$$

$$\mu'_{tm1} = \left\lfloor \frac{\mu_{total}\Psi_1\lambda_1}{\Psi_1\lambda_1 + \Psi_2\lambda_2 + \ldots + \Psi_r\lambda_r} \right\rfloor$$

This formula can be used for *r* number customer types.

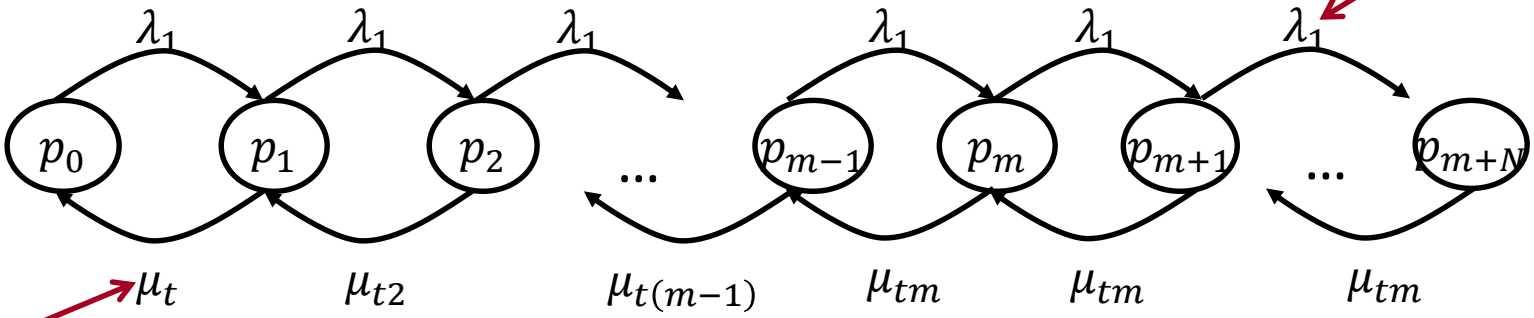
Modeling Assumptions

- System is under heavy traffic flows.
- Arrivals follow Poisson distribution, and service times for customers are exponentially distributed.
- Type of queue discipline used in the analysis is FIFO.
- Service rate of all servers can be different.

Analytical Model

- Only C_1 customers performance metric developed.
- Markov Chain Model :

λ_1: Arrival rate of C_1 customers



$$\mu_{ti} = \sum_{j=1}^{i} \mu_j$$

$$p_i = \begin{cases} p_0 \dfrac{\lambda_1^i}{\prod_{j=1}^{i} \mu_{tj}} & i \le m \\ p_0 \dfrac{\mu_{tm}^m \rho^i}{\prod_{j=1}^{m} \mu_{tj}} & m < i \le m+N \end{cases}$$

m: number of servers for C_1 customers

p_i: Probability of i C_1 customer in the system

N: Queue size

$$\rho = \frac{\lambda_1}{\mu_{tm}}$$

Performance

- Drop Probability : $D = p_0 \frac{\mu_{tm}^{m+N} \rho^{m+N}}{\prod_{j=1}^{m} \mu_{tj}}$

- Throughput: $\gamma = \lambda_1(1 - D)$

- Occupancy: $n = \begin{cases} p_0 \frac{\mu_{tm}^{m}}{\prod_{i=1}^{m} \mu_{ti}} \rho^{m+1} \left(\frac{1-(N+1)\rho^N + N\rho^{N+1}}{(1-\rho)^2} \right) & \rho \neq 1 \\ p_0 \frac{\mu_{tm}^{m}}{\prod_{i=1}^{m} \mu_{ti}} \left(\frac{N(N+1)}{2} \right) & \rho = 1 \end{cases}$

- Delay: $\delta = \frac{n}{\gamma}$

Drop probability

Rate of dropped customers from the systems buffer.

Number of customers served in the systems.

Delay

Average waiting time of a customer in the systems buffer.

Results

- We have used discrete event simulation to implement by following $M/M_i/N/N$ and proposed scheduling.
- Each queue holds 30 customers.
- We ran simulation with 20000 customers for each arrival rate.
- We show h-DDSS with Fastest Server First (FSF) and Slowest Server First (SSF) to compare best and worst performance.

Traffic Arrival Rates

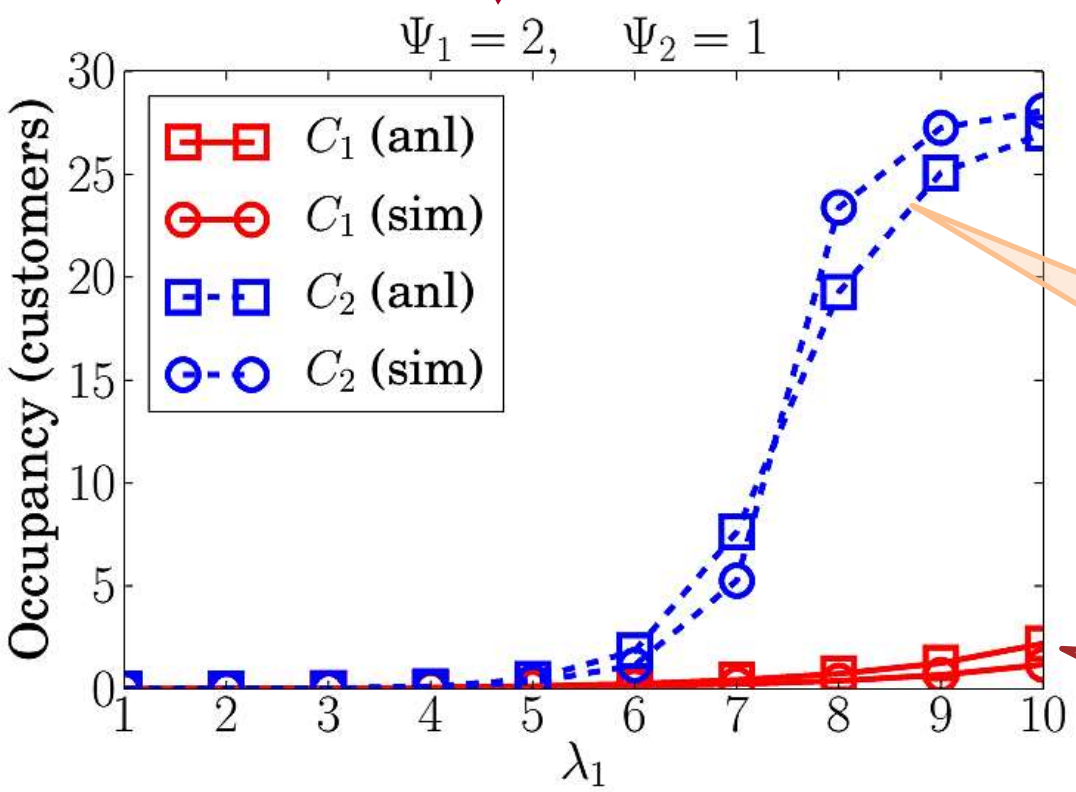
- Simulations were carried out with increased arrival rates of all types of customers to observe the impact of heavy traffic on the system.

- Customer arrival rates at different trials:

λ_1={1, 2, 3, 4, 5, 6, 7, 8, 9, 10},
λ_2={2, 4, 6, 8, 10, 12, 14, 16, 18, 20},
Ψ_1={2 ,3}, Ψ_2={1} and
$\mu = 1, 2, \ldots 7$ for heterogeneous servers and
$\mu = 4$, for homogeneous servers with 7 servers.

Validation of Analytic Formulas: Occupancy

Ψ_1 - Priority level of C_1 customers
Ψ_2 - Priority level of C_2 customers



Occupancy

Number of customers in the systems buffer.

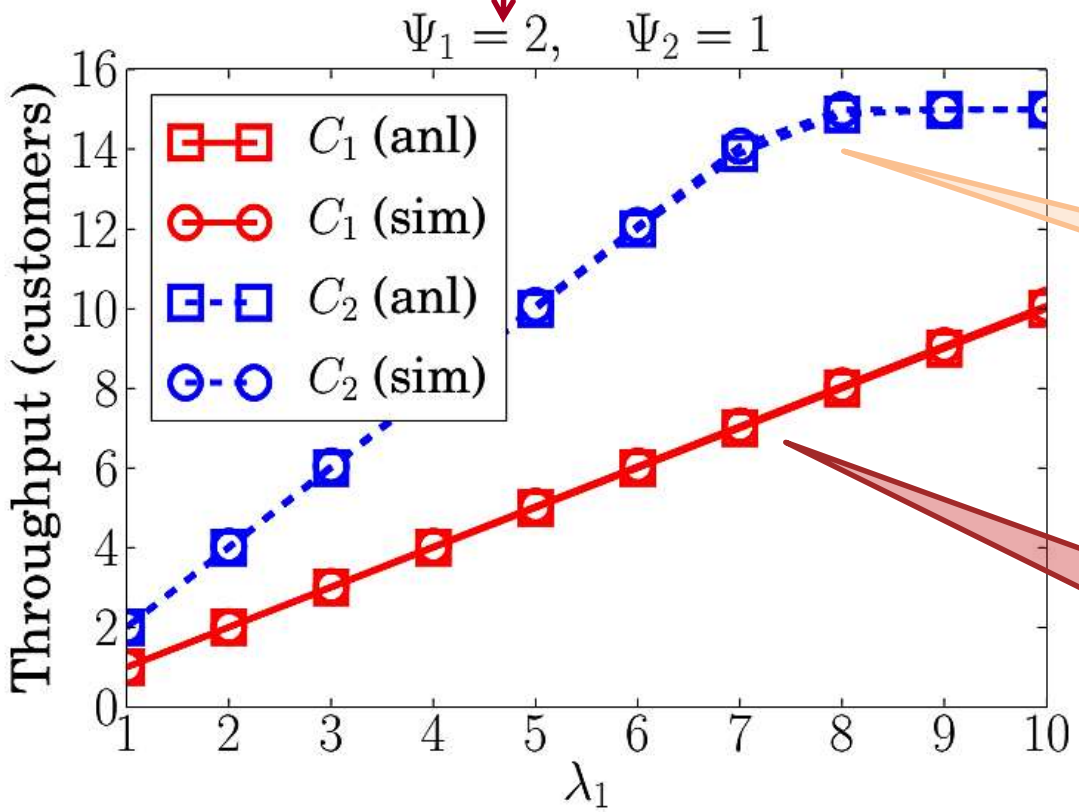
Occupancy of C_2 for analytical and simulation matches.

Occupancy of C_1 for analytical and simulation closely matches.

Occupancy model matches with simulation.

Validation of Analytic Formulas: Throughput

Ψ_1 - Priority level of C_1 customers
Ψ_2 - Priority level of C_2 customers



Throughput

Number of customers are served in the systems.

Throughput of C_2 for analytical and simulation closely matches.

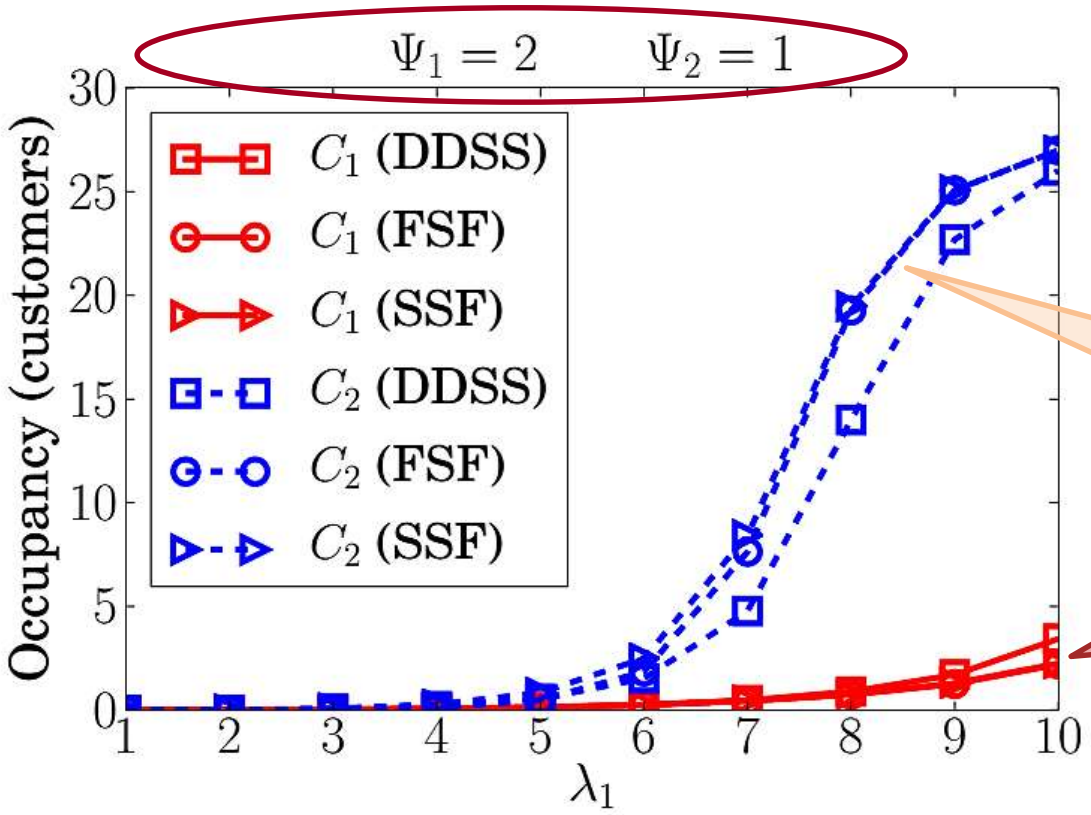
Throughput of C_1 for analytical and simulation closely matches.

Throughput model matches with simulation.

h-DDSS vs DDSS

h-DDSS is heterogeneous.

DDSS is homogeneous.



Objective

We would like to see effects of priority level $\Psi_1 = 2$ on occupancy.

Occupancy of C_2 for DDSS is lower than occupancy of C_2 for h-DDSS.

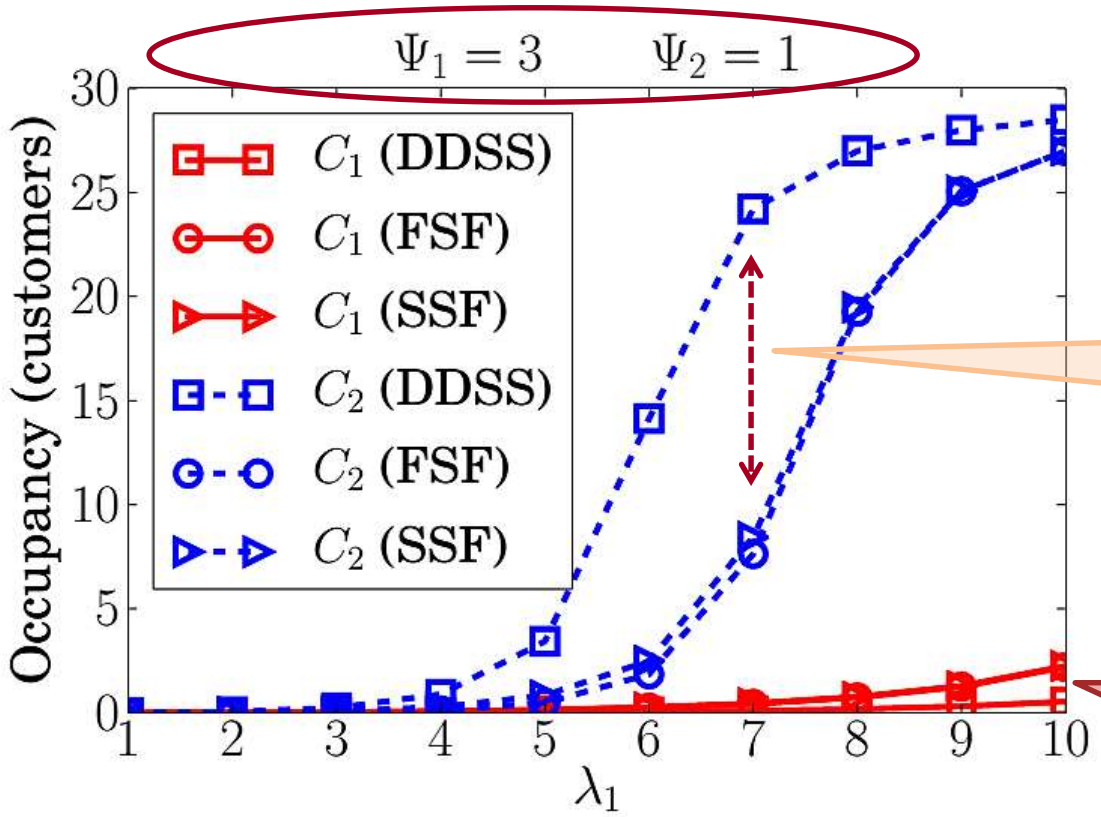
Occupancy of C_1 for DDSS and h-DDSS are same.

DDSS shows better occupancy than h-DDSS for these priority levels.

h-DDSS is heterogeneous.

h-DDSS vs DDSS

DDSS is homogeneous.



Objective

We would like to see effects of priority level $\Psi_1 = 3$ on occupancy.

Occupancy of C_2 for DDSS is higher than occupancy of C_2 for h-DDSS.

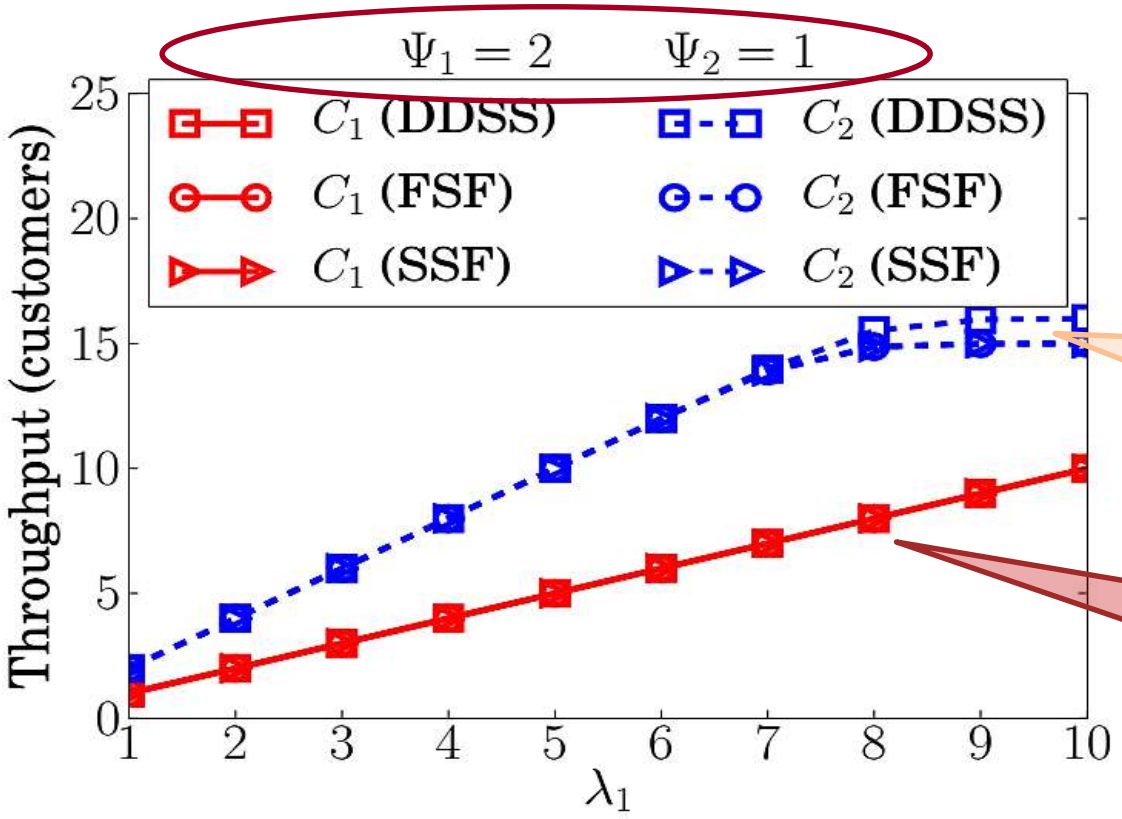
Occupancy of C_1 for DDSS and h-DDSS shows small differences.

h-DDSS shows better occupancy than DDSS for these priority levels.

h-DDSS vs DDSS

h-DDSS is heterogeneous.

DDSS is homogeneous.



Objective

We would like to see effects of priority level $\Psi_1 = 2$ on throughput.

Throughput of C_2 for DDSS is higher than throughput of C_2 for h-DDSS.

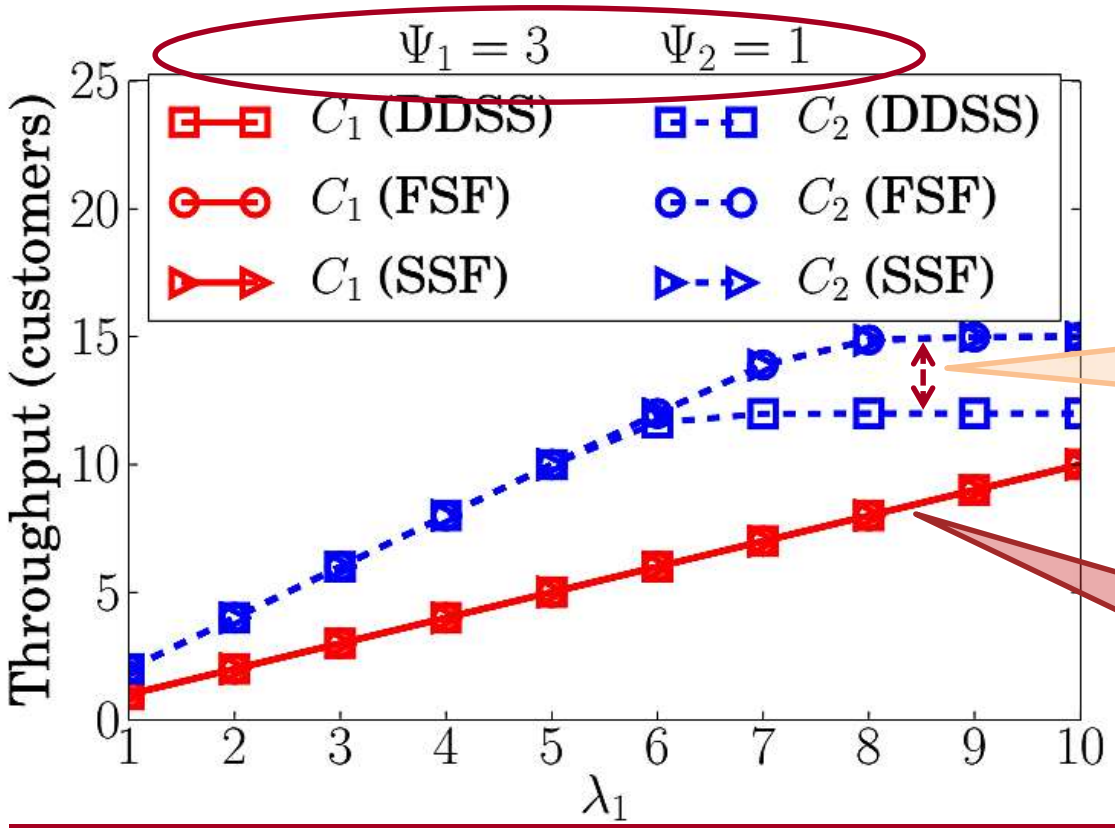
Throughput of C_1 for DDSS and h-DDSS are same.

DDSS shows better throughput than h-DDSS for these priority levels.

h-DDSS vs DDSS

h-DDSS is heterogeneous.

DDSS is homogeneous.



Objective

We would like to see effects of priority level $\Psi_1 = 3$ on throughput.

Throughput of C_2 for DDSS is lower than throughput of C_2 for h-DDSS.

Throughput of C_1 for DDSS and h-DDSS are same.

h-DDSS shows better throughput than DDSS for these priority levels.

Summary of Results

- Priority levels do not affect the performance of DDSS and h-DDSS under low traffic.
- Under heavy traffic, priority levels have a significant impact on the class performances of DDSS.
- Under heavy traffic, performances of FSF and SSF in h-DDSS are same while FSF is better for low traffic arrivals.
- h-DDSS can be more efficient than DDSS for selected class priority levels

Conclusion

- We have proposed a novel scheduling algorithm for cloud computing considering priority, arrival rate and heterogeneous servers.
- Performance metrics of the proposed cloud computing system are presented through different cases.
- h-DDSS and DDSS are compared under different priority levels.
- Proposed scheduling algorithm can help Cloud Computing with homogenous and heterogeneous servers systems have higher throughput and be more balanced.

Thank You

http://cs.ou.edu/~atiq

atiq@ou.edu