Multi Heterogeneous Queueing Server System

General Exam Oral Examination

Fall 2012

prepared by

Husnu Saner Narman

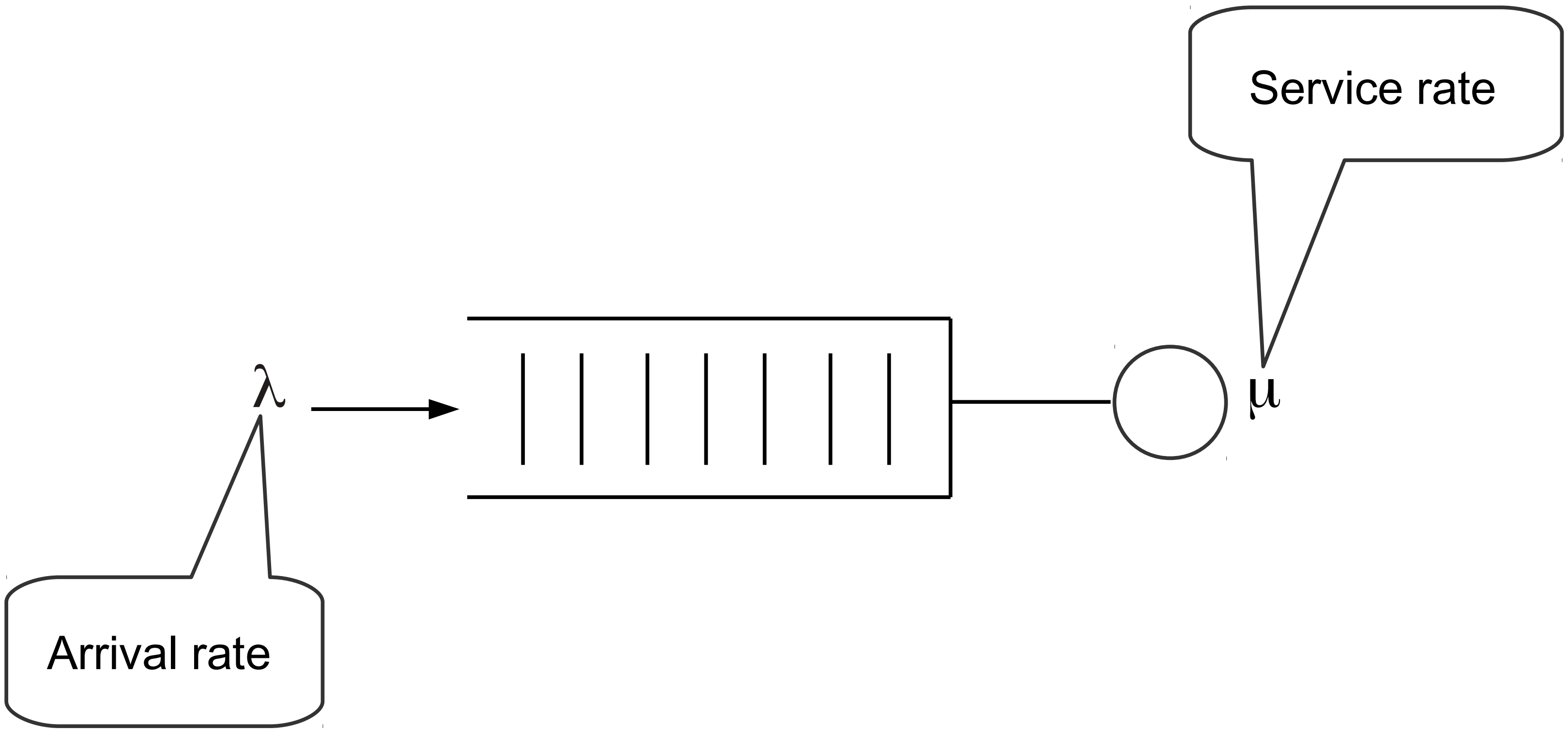
Content

- Motivation
- Contribution
- Multi Heterogeneous System First Model
  - Analysis of First Model
- Multi Heterogeneous System Second Model
  - Analysis of Second Model
- Conclusion
- References

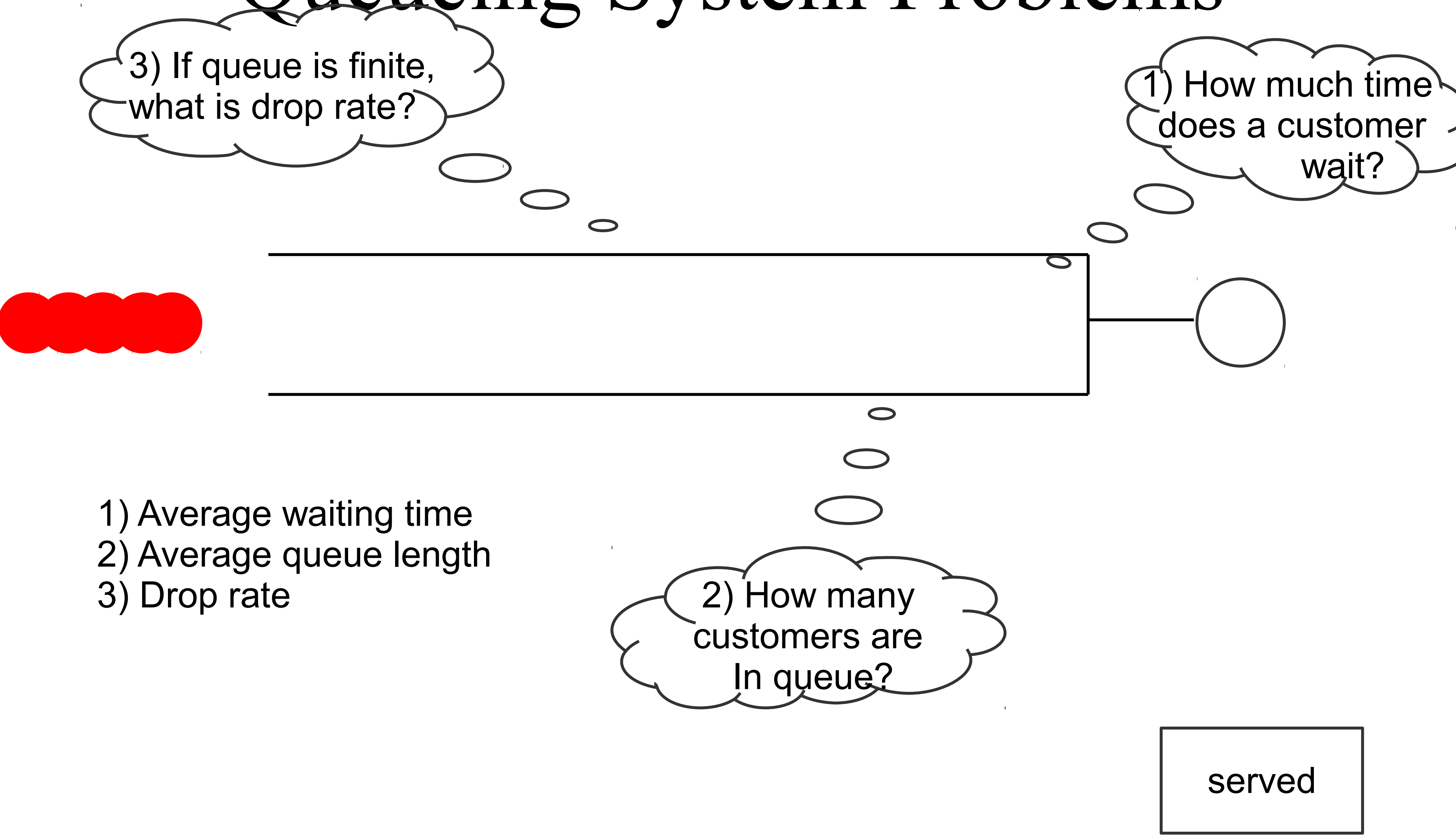
Content

- Motivation
- Contribution
- Multi Heterogeneous System First Model
  - Analysis of First Model
- Multi Heterogeneous System Second Model
  - Analysis of Second Model
- Conclusion
- References

Queueing System

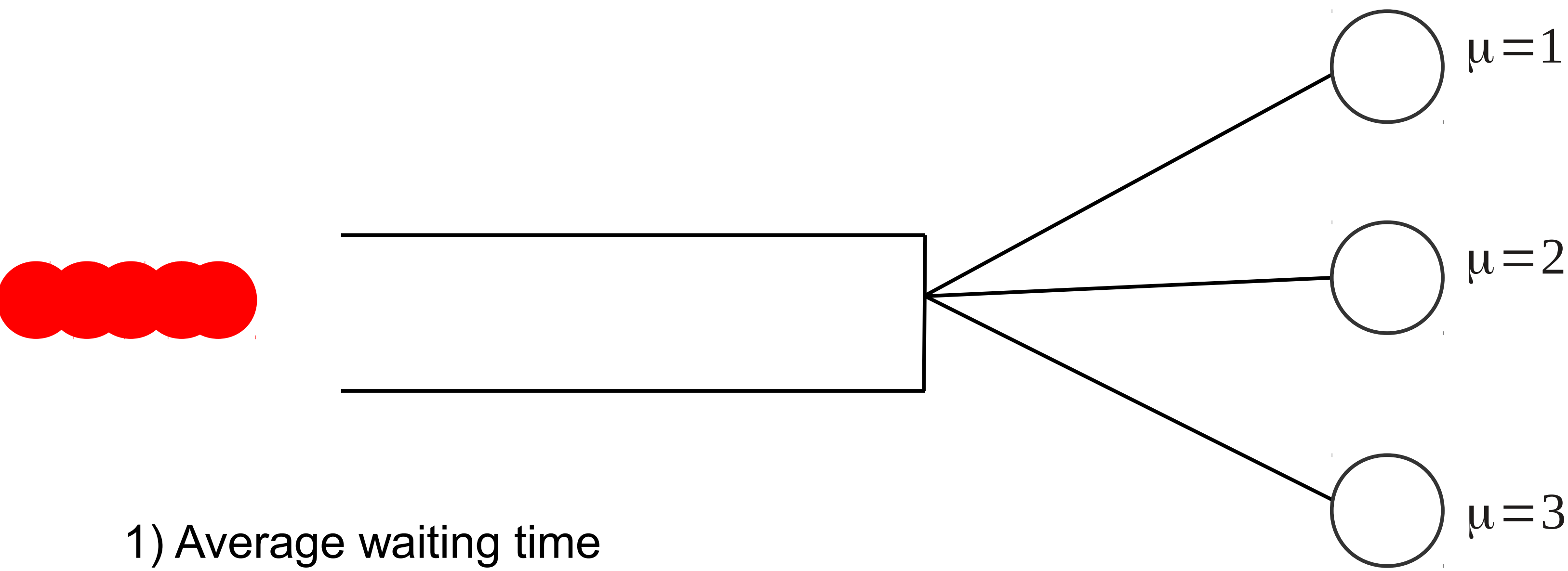


Queueing System Problems



1) Average waiting time
2) Average queue length
3) Drop rate

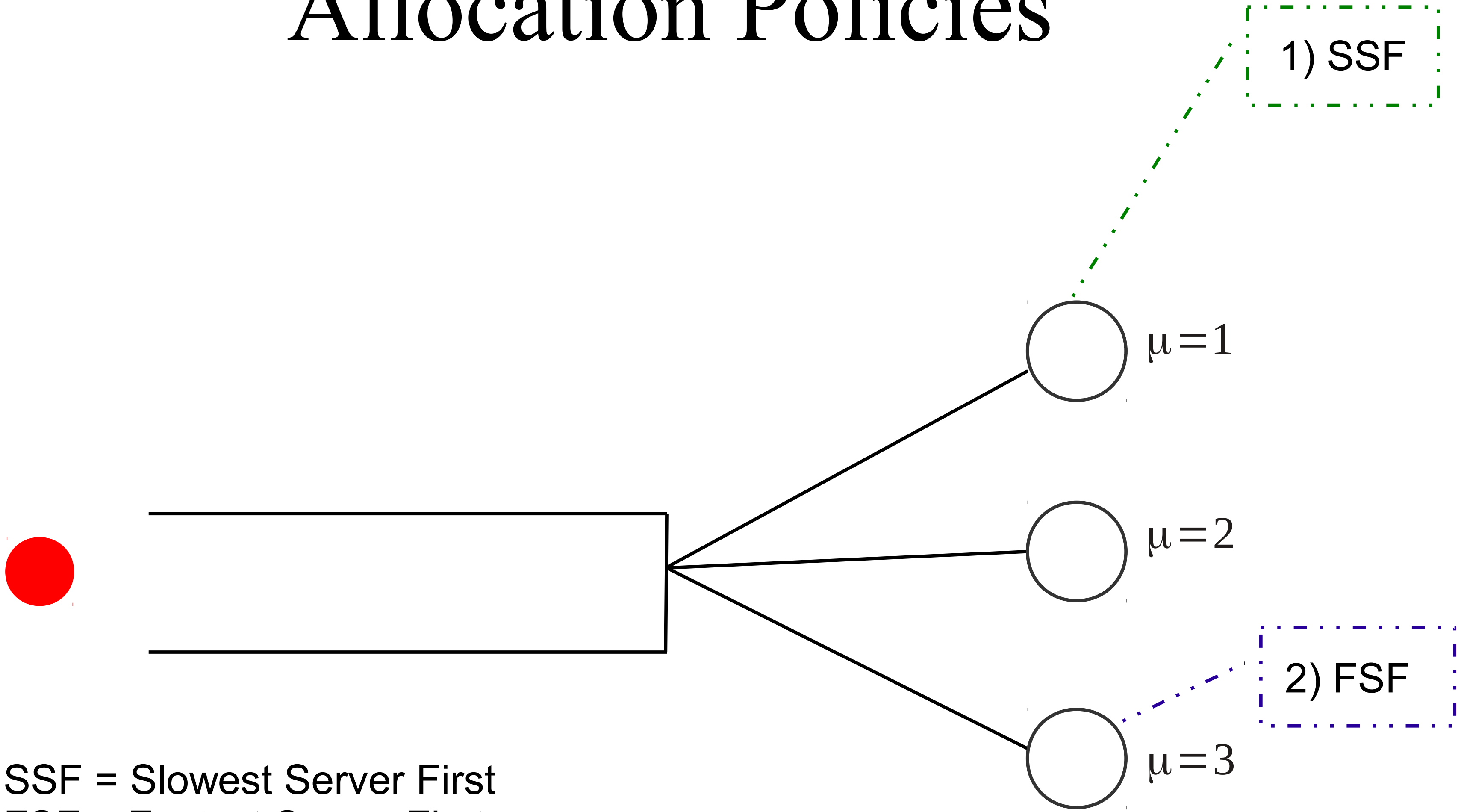
Heterogeneous Multi Server Queueing System



1) Average waiting time
2) Average queue length
3) Drop rate

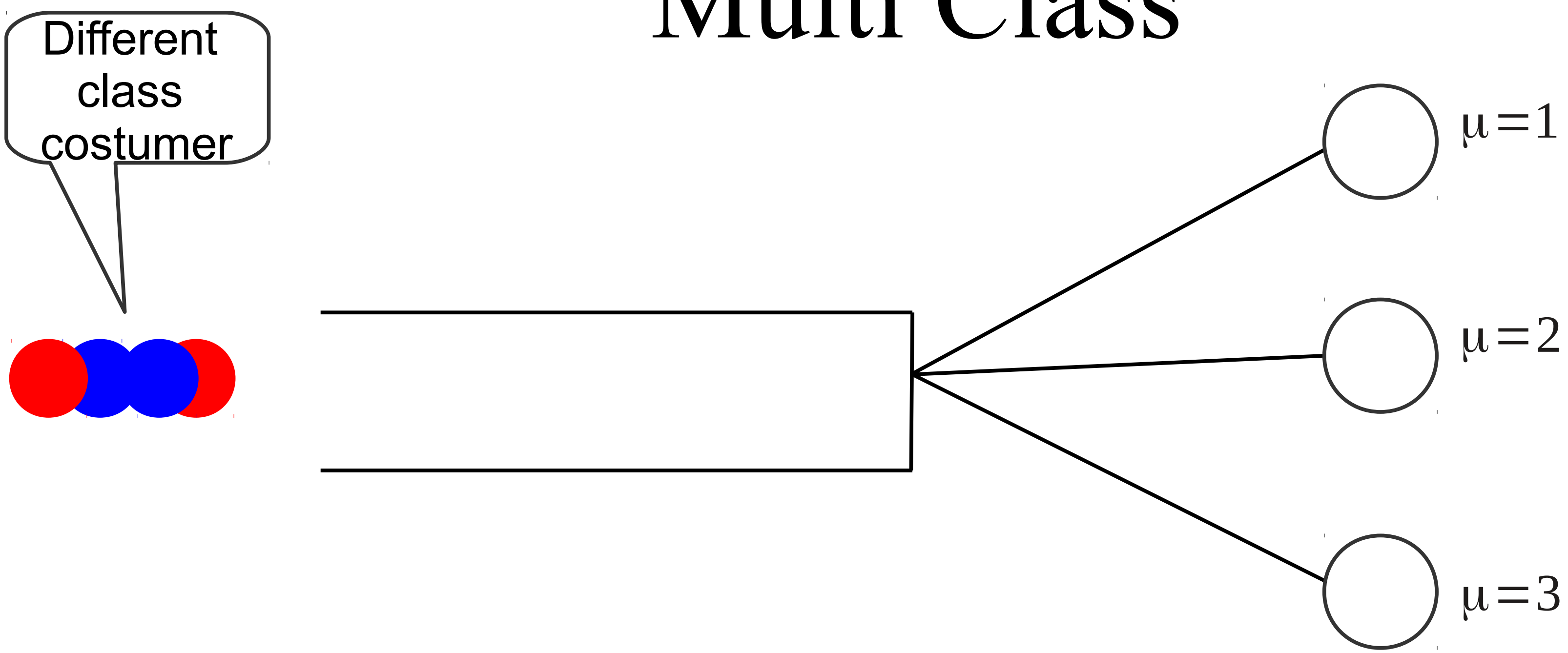
served

Heterogeneous Multi Server Allocation Policies



1) SSF = Slowest Server First
2) FSF = Fastest Server First
3) RCS = Randomly Chosen Server

Heterogeneous Multi Server Multi Class



1) Priority = Which type of customer is served first?
2) Flexibility = Which type of customer will be served by which server?

Heterogeneous Multi Server Queueing System

- Performance metrics:
  - Average waiting time
  - Average queue length
  - Drop rate (for finite queue)
- Allocation Policies
  - FSF, SSF, RCS ...
- Priority in Multi Class Multi Server System
- Flexibility in Multi Class Multi Server System

Content

- Motivation
- Contribution
- Multi Heterogeneous System First Model
  - Analysis of First Model
- Multi Heterogeneous System Second Model
  - Analysis of Second Model
- Conclusion
- References

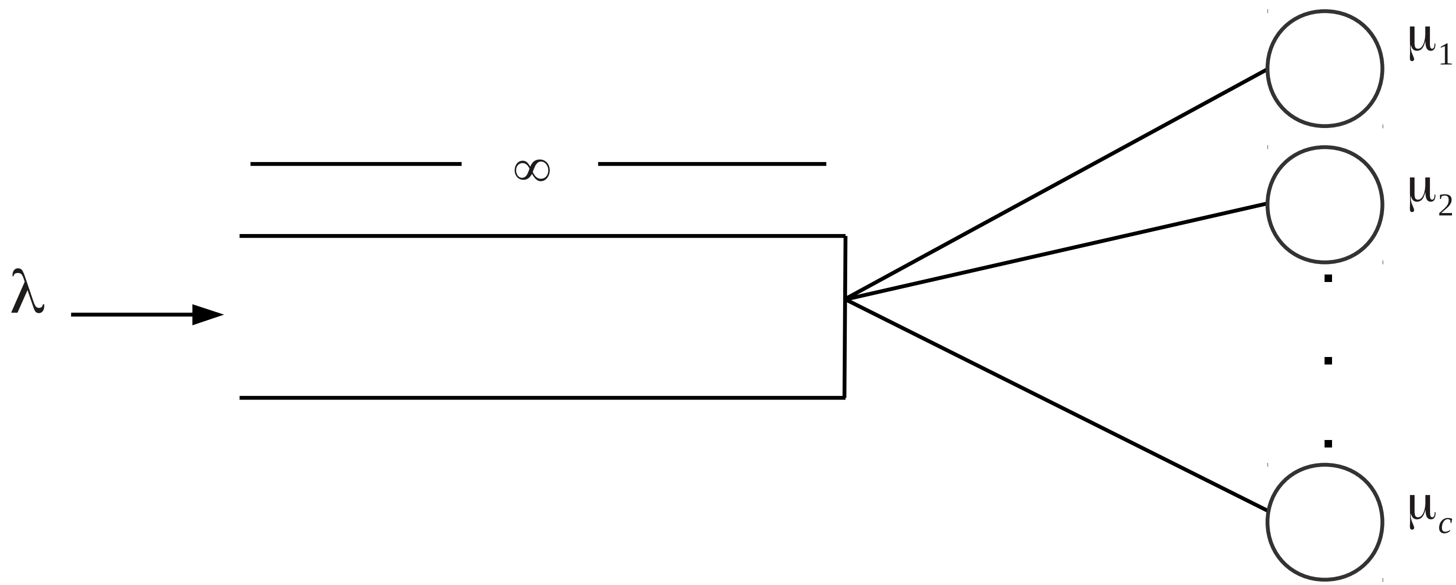
Contribution of Works

- Upper and Lower bound average waiting time for Multi Heterogeneous Single Queue System
- Upper and Lower bound average queue length for Multi Heterogeneous Single Queue System
- Lower bound queue drop rate for Multi Heterogeneous Single Finite Queue System
- Developed performance approximations are better than homogeneous approximation

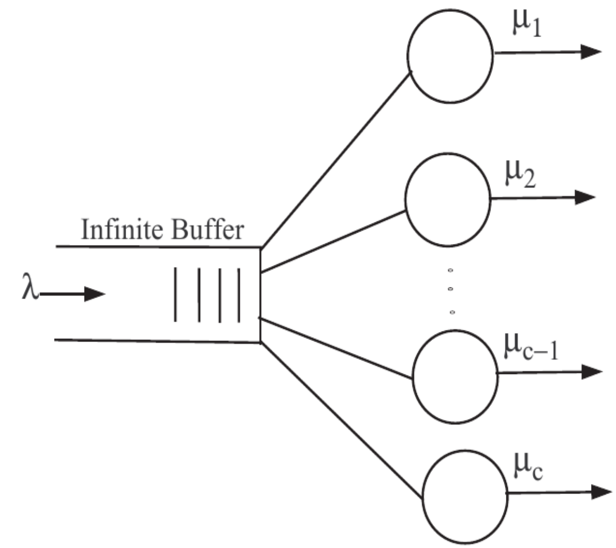
Content

- Motivation
- Contribution
- Multi Heterogeneous System First Model
  - Analysis of First Model
- Multi Heterogeneous System Second Model
  - Analysis of Second Model
- Conclusion
- References

Multi Heterogeneous System First Model ($M/M_i/c$)



Assumption for First Model

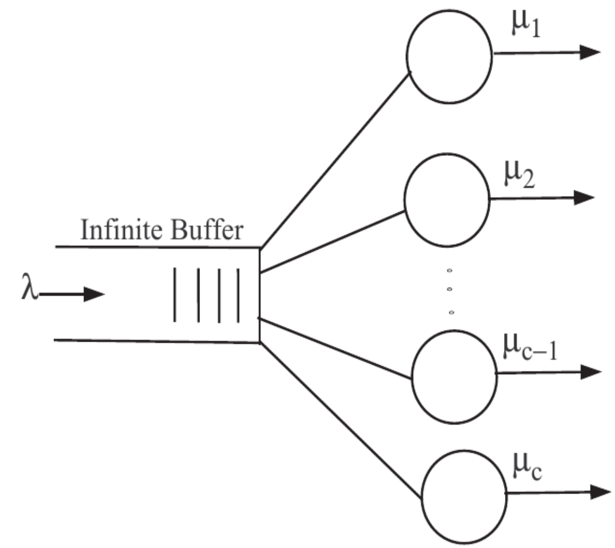


- $\mu_1 \leqslant \mu_2 \leqslant ..... \leqslant \mu_c$
- Poisson distribution with rate $\boldsymbol{\lambda}$,
- Exponential distribution with rate $\boldsymbol{\mu_i}$,
- ***c*** number of servers,
- Slowest Server First (SSF) allocation
- Because of SSF, upper bound performance metrics

Content

- Motivation
- Contribution
- Multi Heterogeneous System First Model
  - Analysis of First Model
- Multi Heterogeneous System Second Model
  - Analysis of Second Model
- Conclusion
- References

Notations



λ Job arrival rate,

μ_i Service rate of ith server,

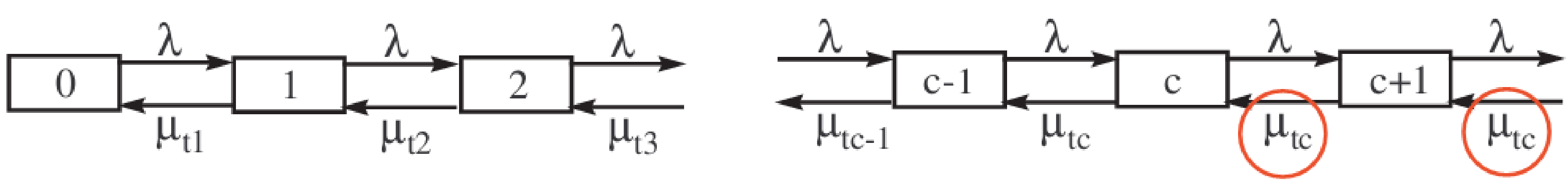
p_i State probability of ith state,

E_n^u Upper bound average queue length of the system,

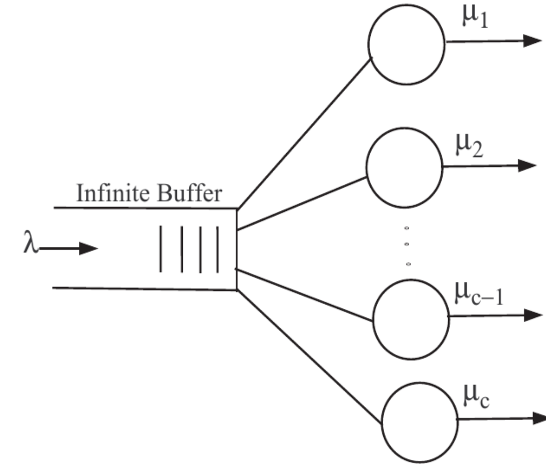
E_T^u Upper bound average waiting time of the system,

μ_{ti} Total service rates until ith server, or $\mu_{ti} = \sum_{i=1}^{c} \mu_i$

ρ Utilization of the system, or $\rho = \lambda / \mu_{tc}$



Probability of States



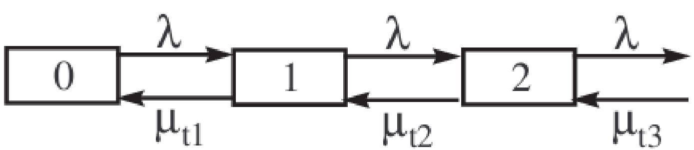
$$\lambda p_0 = \mu_{t1} p_1 \Leftrightarrow p_1 = p_0 \frac{\lambda}{\mu_{t1}}$$

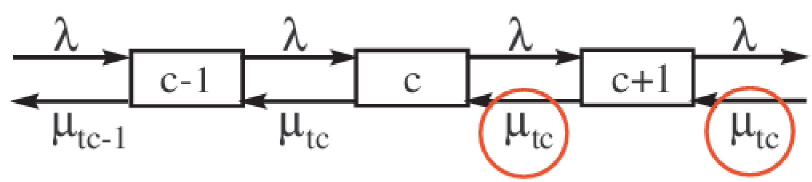
$$\lambda p_1 = \mu_{t2} p_2 \Leftrightarrow p_2 = p_0 \frac{\lambda^2}{\mu_{t1}\mu_{t2}}$$

$$\lambda p_{c-1} = \mu_{tc} p_c \Leftrightarrow p_c = p_0 \frac{\lambda^c}{\mu_{t1}\mu_{t2}\dots\mu_{tc}}$$

$$\lambda p_{i-1} = \mu_{tc} p_i \Leftrightarrow p_i = p_0 \frac{\lambda^i}{\mu_{t1}\mu_{t2}\dots\mu_{tc}\mu_{tc}^{i-c}} = \frac{\lambda^i}{\mu_{tc}^{i-c}\prod_{j=1}^{c}\mu_{tj}} \quad \text{where} \quad i > c$$

$$\sum_{i=0}^{\infty} p_i = 1 \quad \text{and} \quad \rho < 1$$

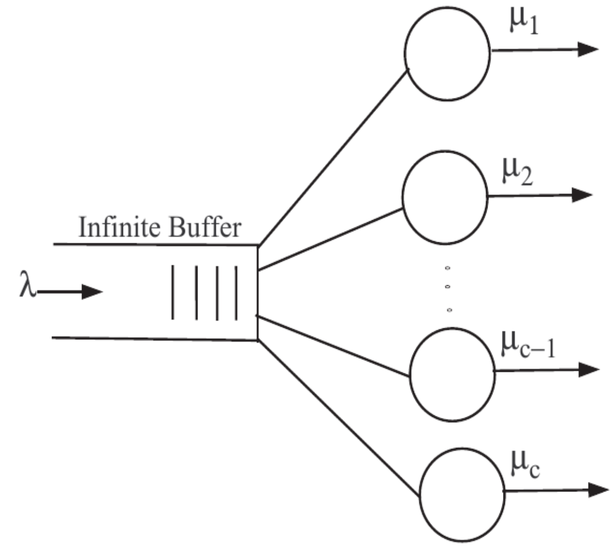




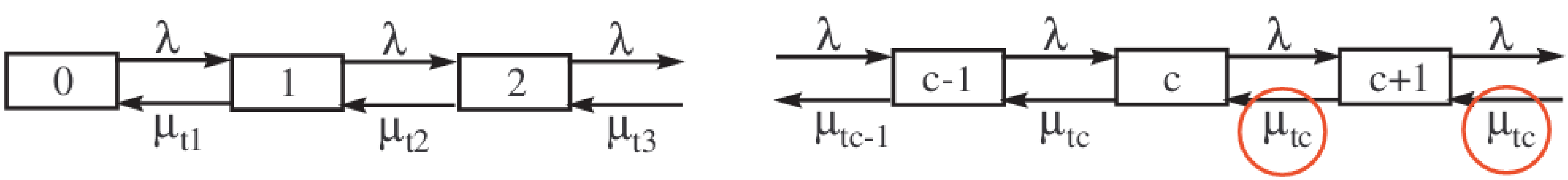
Probability of States

$$p_0 = \frac{1}{1 + \sum_{j=1}^{c-1} \left(\frac{\lambda^j}{\prod_{i=1}^{j} \mu_{ti}} \right) + \left(\frac{\lambda^c}{(1-\rho) \prod_{i=1}^{c} \mu_{ti}} \right)}$$

$$\rho = \frac{\lambda}{\mu_{tc}} < 1$$



where $\rho = \mu_{tc} / \lambda$



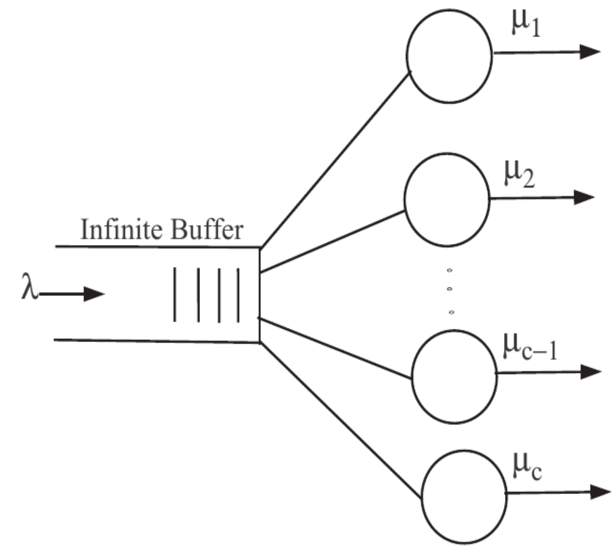
Average Queue Length and Waiting Time

General formula of average queue length for single server

$$E_n = \sum_{i=1}^{\infty} i p_i$$

Modified formula of average queue length for multi servers

$$E_n^u = \sum_{i=c+1}^{\infty} (i-c)\, p_i$$



$$E_n^u = p_0 \sum_{i=c+1}^{\infty} \frac{(i-c)\lambda^i}{\mu_{tc}^{i-c} \prod_{j=1}^{c} \mu_{tj}}$$ because $$p_i = \frac{\lambda^i}{\mu_{tc}^{i-c} \prod_{j=1}^{c} \mu_{tj}}$$

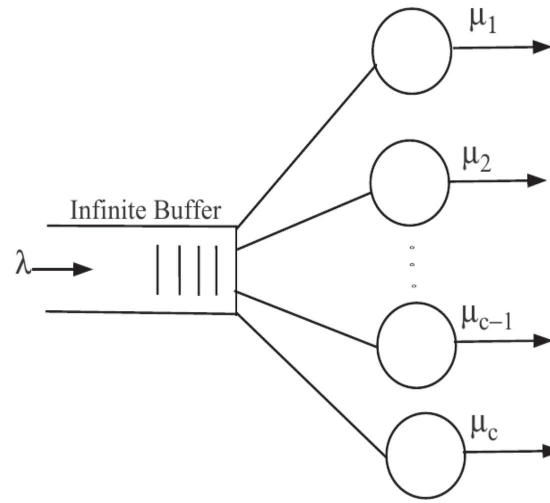
Average Queue Length and Waiting Time

$$\mu_{tc} = \sum_{i=1}^{c} \mu_i$$

Upper bound average queue length

$$E_n^u = p_0 \frac{\mu_{tc}^c}{\prod_{i=1}^{c} \mu_{ti}} \frac{\rho^{c+1}}{(1-\rho)^2}$$

$$\rho = \frac{\lambda}{\mu_{tc}} < 1$$



From Little's Law

Upper bound average waiting time

$$E_T^u = \frac{E_n^u}{\lambda} = p_0 \frac{\mu_{tc}^c}{\prod_{i=1}^{c} \mu_{ti}} \frac{\rho^{c+1}}{(1-\rho)^2} \frac{1}{\lambda}$$

Arrival rate

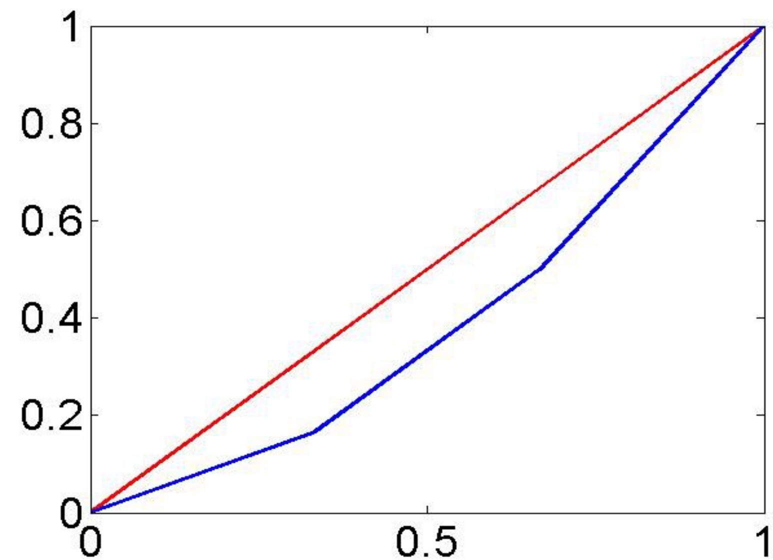
Result for First Model

- Arrival rate
- Number of servers
- Heterogeneity level of server rates
  - Gini Index:

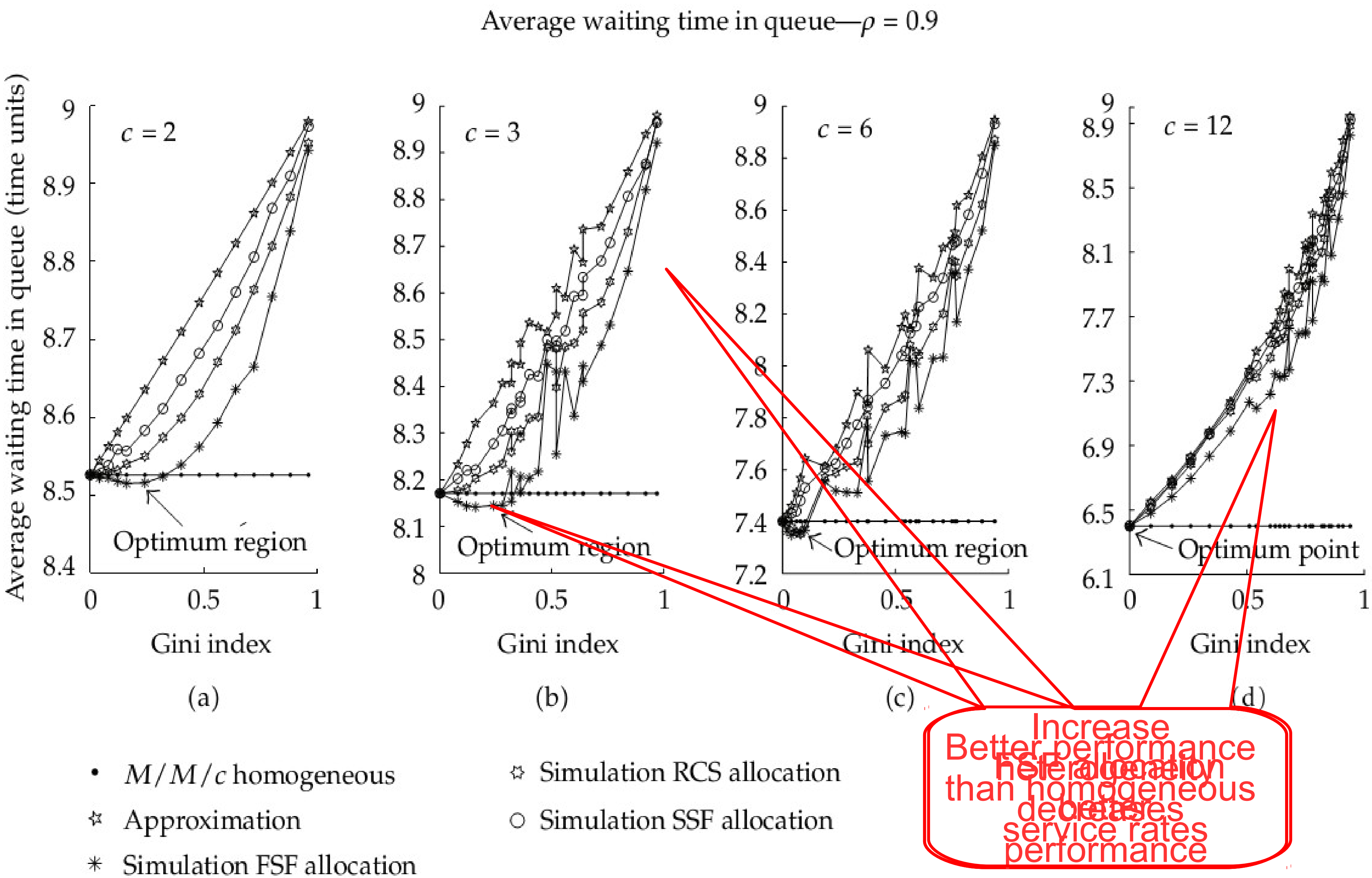
$\mu_1 = 1, \mu_2 = 2, \mu_3 = 3$

$(0,0) \quad (1/3, 1/6) \quad (2/3, 3/6) \quad (3/3, 6/6)$

$(0,0) \quad (1/3, 1/6) \quad (2/3, 1/2) \quad (1,1)$



Waiting Time for First Model

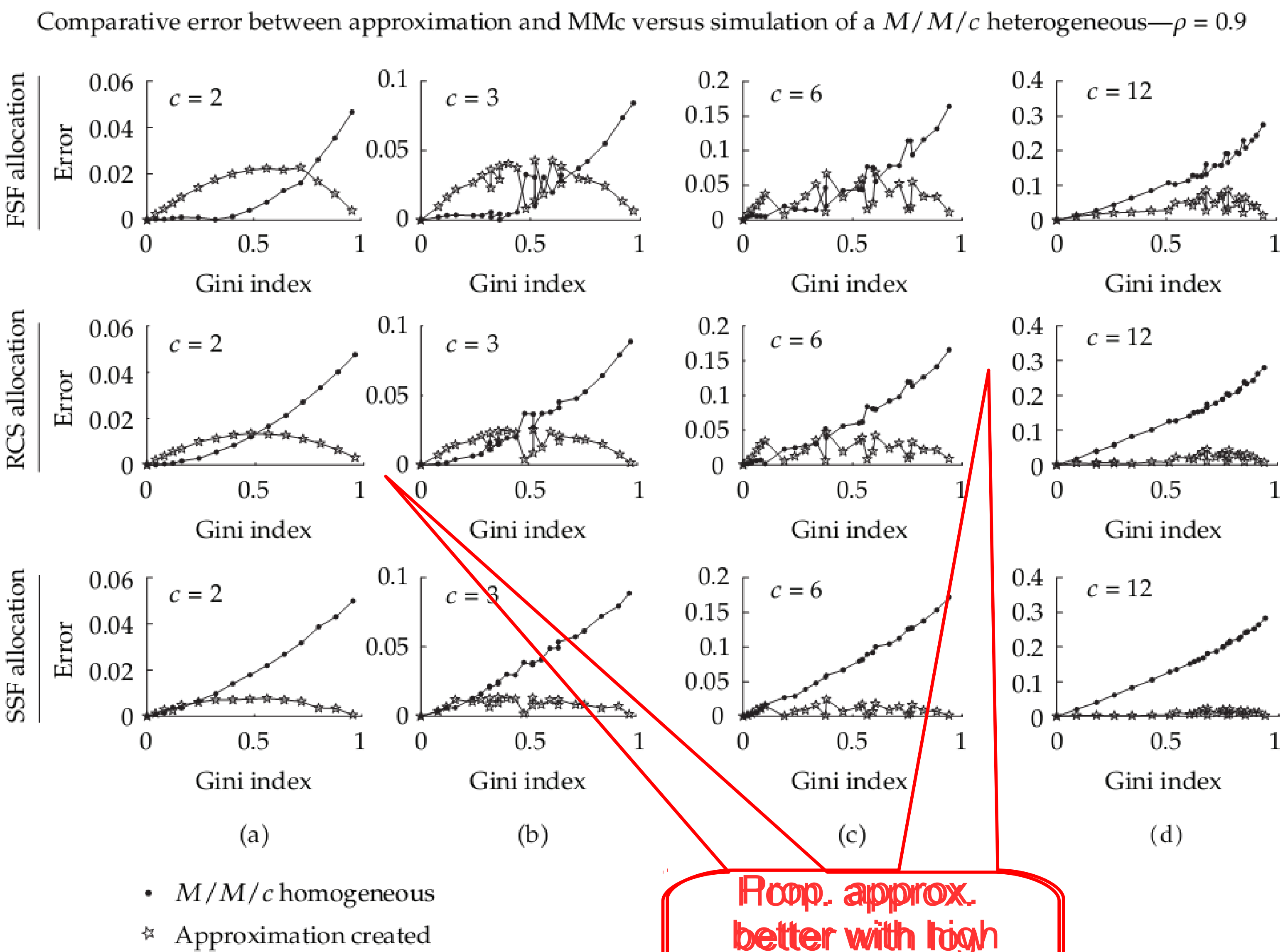


Error Comparison

- Error Rate Formula
- Which M/M/c or developed formula is better

$$error = \frac{\left| x_{simulation} - x_{calculation} \right|}{x_{simulation}}$$

Error Figures of Different Allocations for First Model

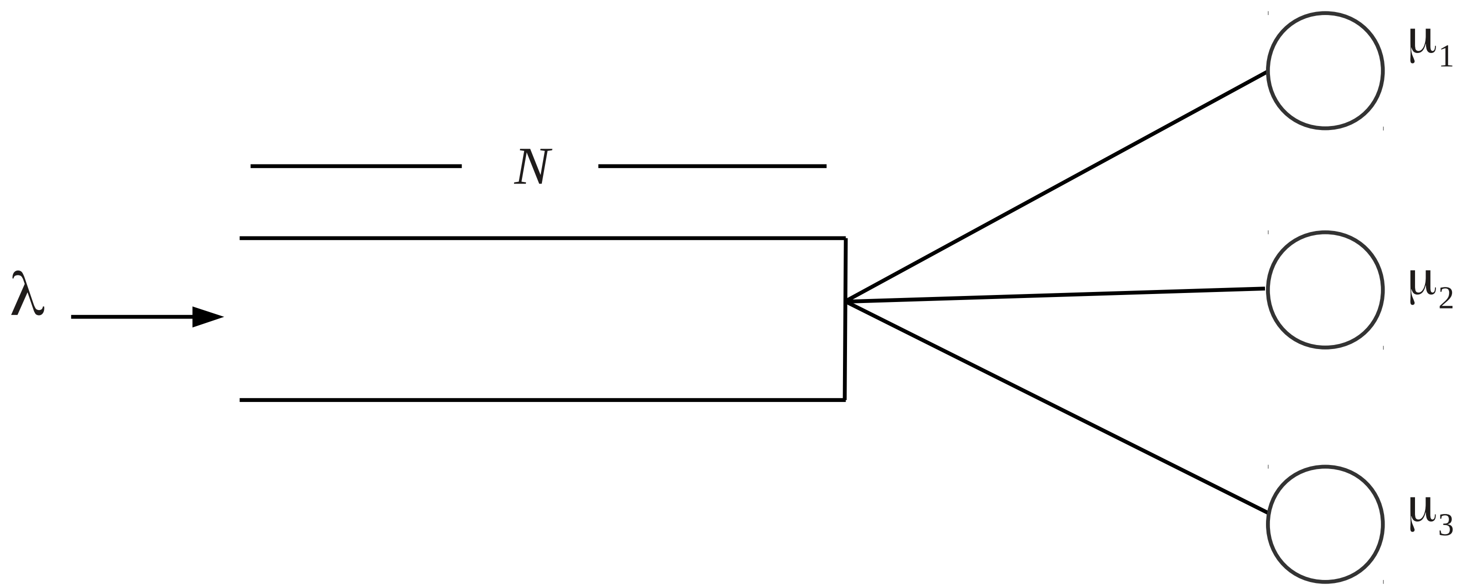


Prop. approx. better with high heterogeneity and servers

Content

- Motivation
- Contribution
- Multi Heterogeneous System First Model
  - Analysis of First Model
- Multi Heterogeneous System Second Model
  - Analysis of Second Model
- Conclusion
- References

Multi Heterogeneous System Second Model (*M/Mi/3/N*)



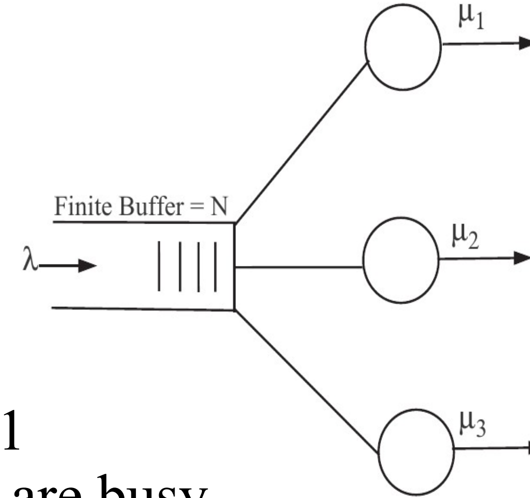
Assumption for Second Model

- $\mu_1 \geqslant \mu_2 \geqslant \mu_3$
- Poisson distribution with rate $\boldsymbol{\lambda}$,
- Exponential distribution with rate $\boldsymbol{\mu_i}$,
- ***3*** number of servers,
- Fastest Server First (FSF) allocation
- Because of FSF, lower bound performance metrics

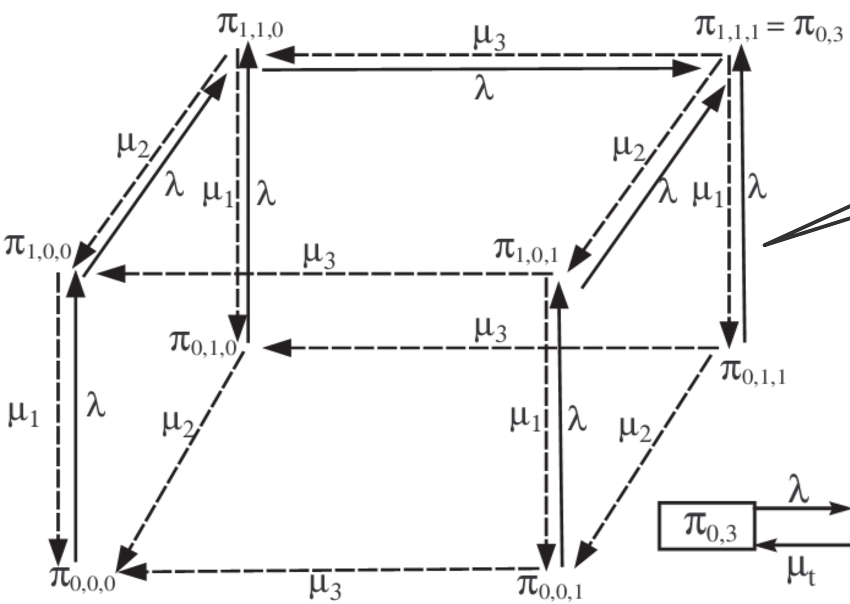
Content

- Motivation
- Contribution
- Multi Heterogeneous System First Model
  - Analysis of First Model
- Multi Heterogeneous System Second Model
  - Analysis of Second Model
- Conclusion
- References

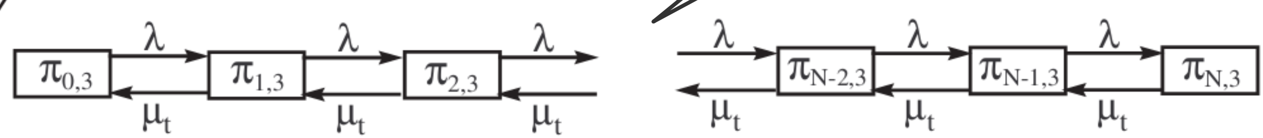
Notations



λ	Job arrival rate,
μ_1, μ_2, μ_3	Service rate of 1st, 2nd, and 3rd servers,
$\pi_{i,j,k}$	State probability of servers, idle:0, busy:1
$\pi_{n,3}$	State probability of nth state when servers are busy
μ_t	Total service rates

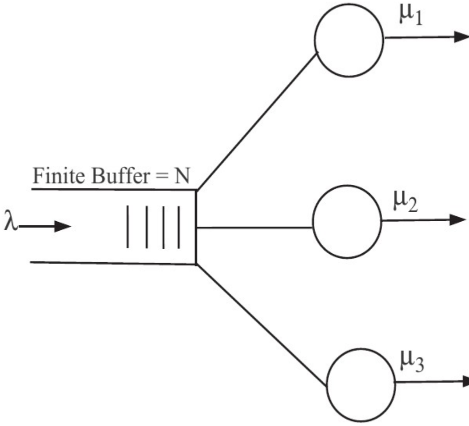


Server state transition diagram



Jobs state transition diagram

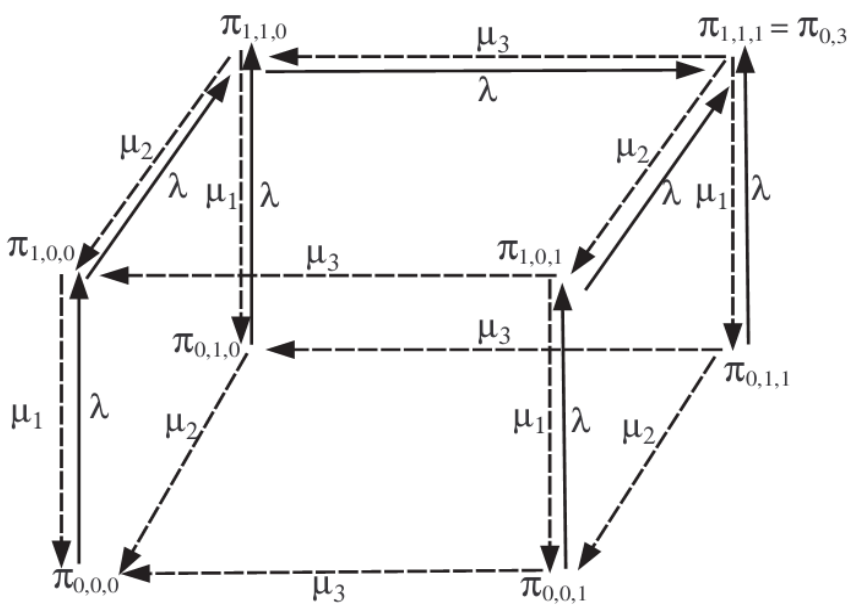
Probability of States

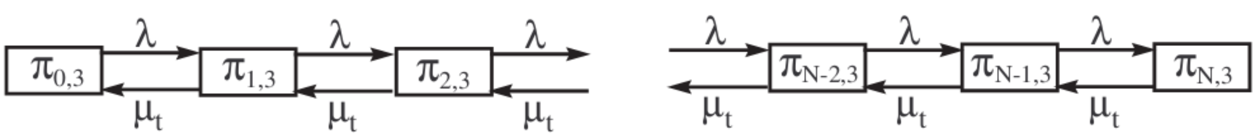


$$\lambda\,\pi_{0,3}+\mu_t\,\pi_{0,3}=\lambda\,\pi_{0,1,1}+\lambda\,\pi_{1,0,1}+\lambda\,\pi_{1,1,0}+\mu_t\,\pi_{1,3}$$

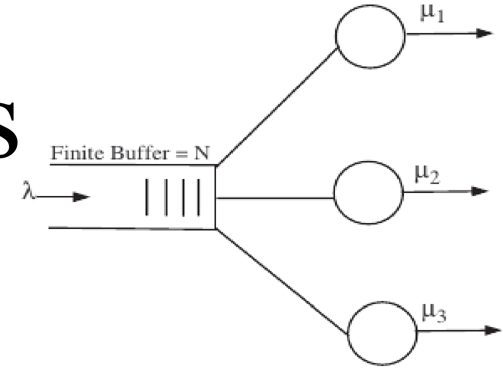
$\lambda\,\pi_{n,3}=\mu_t\,\pi_{n+1,3}$ where $0\leqslant n<N$

$\pi_{n,3}=\pi_{0,3}\left(\frac{\lambda}{\mu_t}\right)^{n}$ or $\pi_{n,3}=\pi_{N,3}\left(\frac{\lambda}{\mu_t}\right)^{n-N}$ where $0\leqslant n\leqslant N$





Probability of States



$$\lambda\pi_{1,0,0} + \mu_1\pi_{1,0,0} = \lambda\pi_{0,0,0} + \mu_2\pi_{1,1,0} + \mu_3\pi_{1,0,1}$$

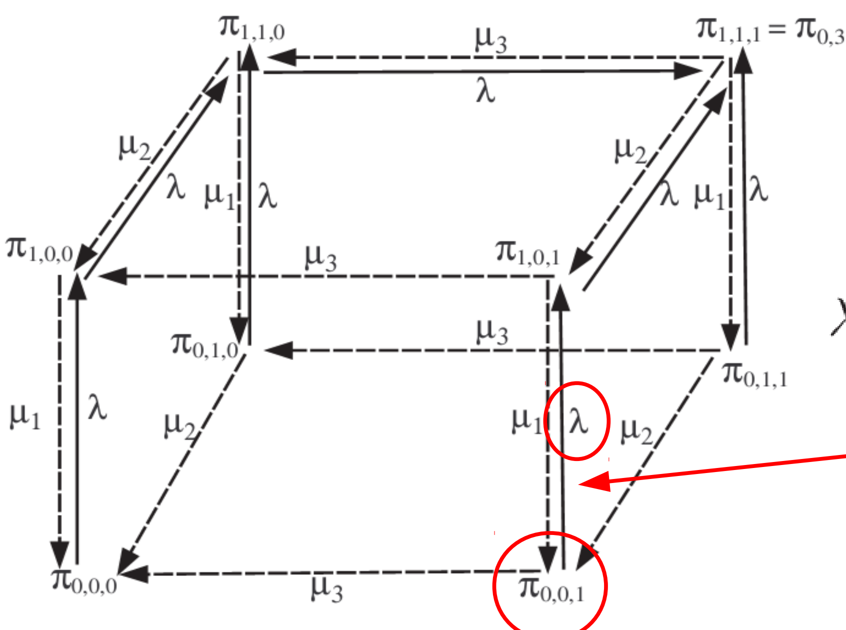
$$\lambda\pi_{0,1,0} + \mu_2\pi_{0,1,0} = \mu_1\pi_{1,1,0} + \mu_3\pi_{0,1,1}$$

$$\lambda\pi_{0,0,1} + \mu_3\pi_{0,0,1} = \mu_1\pi_{1,0,1} + \mu_2\pi_{0,1,1}$$

$$\lambda\pi_{1,1,0} + \mu_1\pi_{1,1,0} + \mu_2\pi_{1,1,0} = \lambda\pi_{1,0,0} + \lambda\pi_{0,1,0} + \mu_3\pi_{1,1,1}$$

$$\lambda\pi_{0,1,1} + \mu_2\pi_{0,1,1} + \mu_3\pi_{0,1,1} = \mu_1\pi_{1,1,1}$$

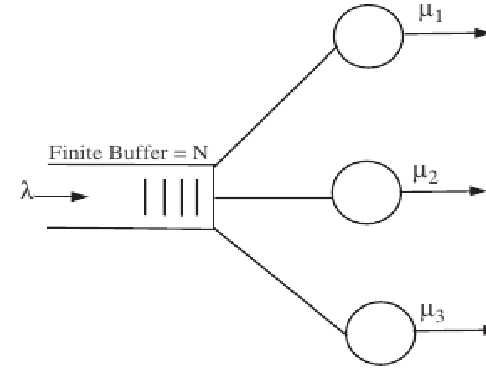
All derived formulas are invalid



$$\lambda\pi_{1,0,1} + \mu_1\pi_{1,0,1} + \mu_3\pi_{1,0,1} = \mu_2\pi_{1,1,1}$$

$$\lambda\pi_{1,0,1} + \mu_1\pi_{1,0,1} + \mu_3\pi_{1,0,1} = \mu_2\pi_{1,1,1} + \lambda\pi_{0,0,1}$$

Different Approach

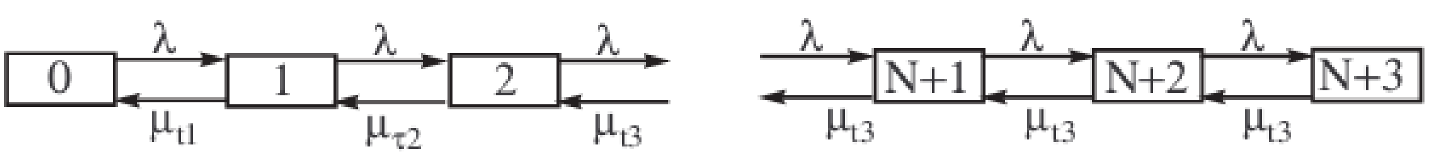


- Hard to develop approximations after correction
- Similar methodology with first model
- Only Job states (no server states)
- Can be expendable <u>*n*</u> number of servers

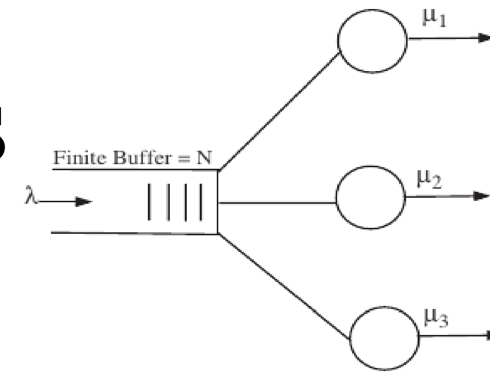
Notations

λ	Job arrival rate,
μ_1, μ_2, μ_3	Service rate of 1st, 2nd, and 3rd servers,
p_i	State probability of jobs in queue
E_n^l	Lower bound average queue length of the system,
E_T^l	Lower bound average waiting time of the system,
P_B^l	Lower bound queue drop rate
γ^u	Upper bound throughput of queue
ρ	Utilization of the system, or $\rho = \lambda / \mu_{t3}$

$\mu_{t1} = \mu_1$ $\quad \mu_{t2} = \mu_1 + \mu_2$ $\quad \mu_{t3} = \mu_1 + \mu_2 + \mu_3$



Probability of States



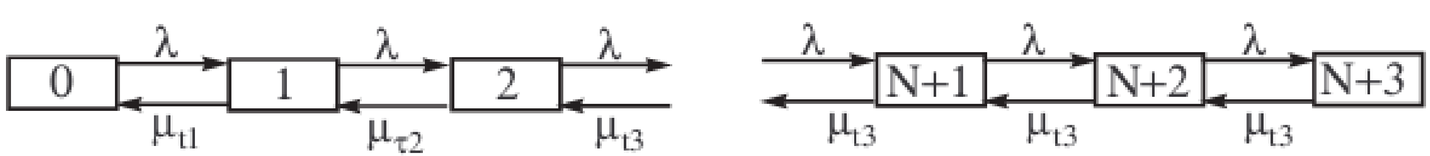
$$\lambda\, p_0 = \mu_{t1}\, p_1 \Leftrightarrow p_1 = p_0 \frac{\lambda}{\mu_{t1}}$$

$$\lambda\, p_1 = \mu_{t2}\, p_2 \Leftrightarrow p_2 = p_0 \frac{\lambda^2}{\mu_{t1}\mu_{t2}}$$

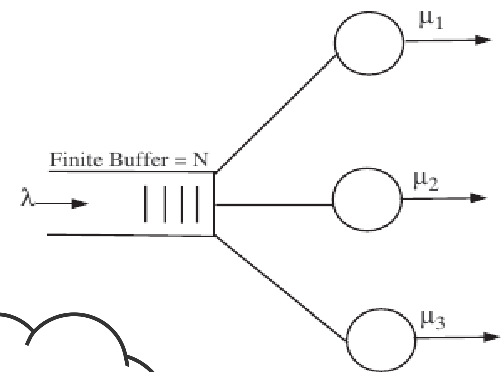
$$\lambda\, p_{i-1} = \mu_{t3}\, p_i \Leftrightarrow p_i = p_0 \frac{\lambda^i}{\mu_{t1}\mu_{t2}\mu_{t3}^{i-2}} = p_0 \frac{\mu_{t3}^2 \rho^i}{\mu_{t1}\mu_{t2}}$$ where $3 \leqslant i \leqslant N+3$ $\quad \rho = \lambda / \mu_{t3}$

$$\sum_{i=0}^{N+3} p_i = 1$$ and $\rho > 0$

$$p_0 = \begin{cases} \dfrac{1}{1 + \dfrac{\lambda}{\mu_{t1}} + \dfrac{\lambda^2}{\mu_{t1}\mu_{t2}} + \dfrac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \left(\dfrac{\rho^3 - \rho^{N+4}}{1-\rho} \right)} & \rho \neq 1 \\ \dfrac{1}{1 + \dfrac{\lambda}{\mu_{t1}} + \dfrac{\lambda^2}{\mu_{t1}\mu_{t2}} + \dfrac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} (N+1)} & \rho = 1 \end{cases}$$



Queue Drop Rate and Throughput



Lower bound because of FSF allocation

$$P_B^l = p_{N+3} = p_0 \frac{\mu_{t3}^2 \rho^{N+3}}{\mu_{t1}\mu_{t2}}$$

$$p_i = p_0 \frac{\mu_{t3}\rho^i}{\mu_{t1}\mu_{t2}}$$

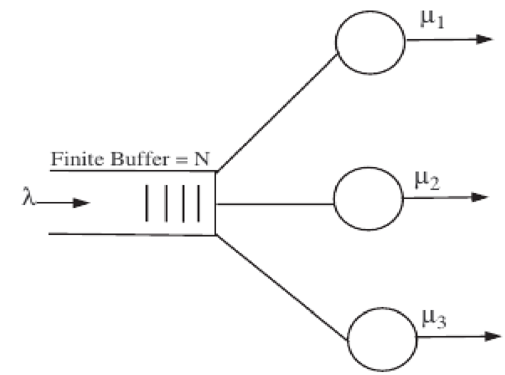
Upper bound because of FSF allocation

$$\gamma^u = \lambda(1 - P_B^l)$$

Arrival rate

Not blocking probability

Average Queue Length and Waiting Time



General formula of average queue length for single server

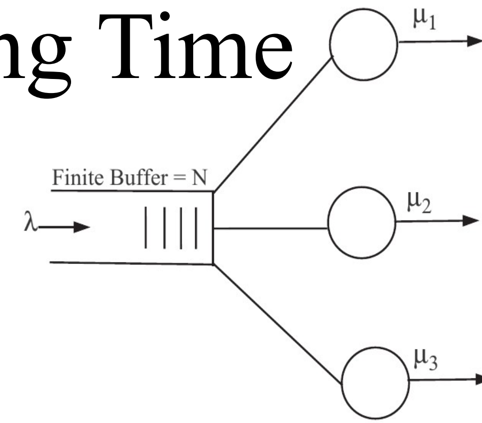
$$E_n = \sum_{i=1}^{N} i p_i$$

Modified formula of average queue length for multi servers

$$E_n^l = \sum_{i=4}^{N+3} (i-3)\, p_i$$

$$E_n^l = p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \sum_{i=4}^{N+3} (i-3)\rho^i$$ because $$p_i = p_0 \frac{\mu_{t3}\rho^i}{\mu_{t1}\mu_{t2}}$$

Average Queue Length and Waiting Time



Lower bound average queue length

$$E_n^l = \begin{cases} p_0 \dfrac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \rho^4 \left(\dfrac{1-(N+1)\rho^N + N\rho^{N+1}}{(1-\rho)^2} \right) & \rho \neq 1 \\ p_0 \dfrac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \dfrac{N(N+1)}{2} & \rho = 1 \end{cases}$$

From Little's Law

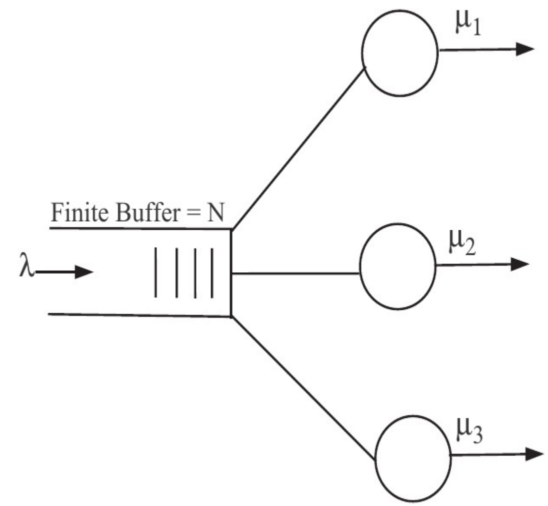
Lower bound average waiting time

$$E_T^l = \frac{E_n^l}{\gamma^u} = \frac{E_n^l}{\lambda(1-P_B^l)}$$

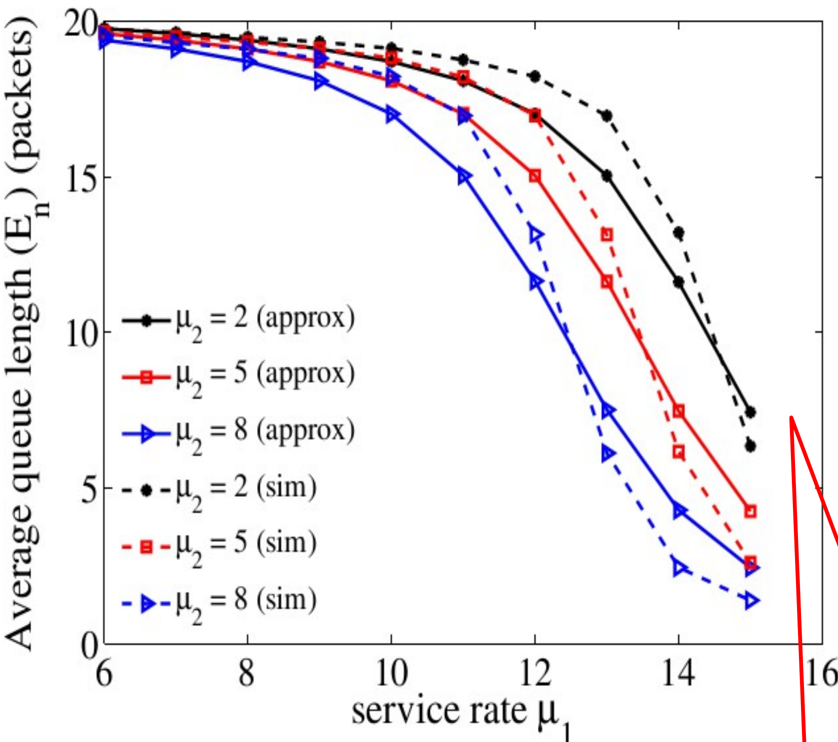
Throughput or effective arrival rate

Result for Second Model

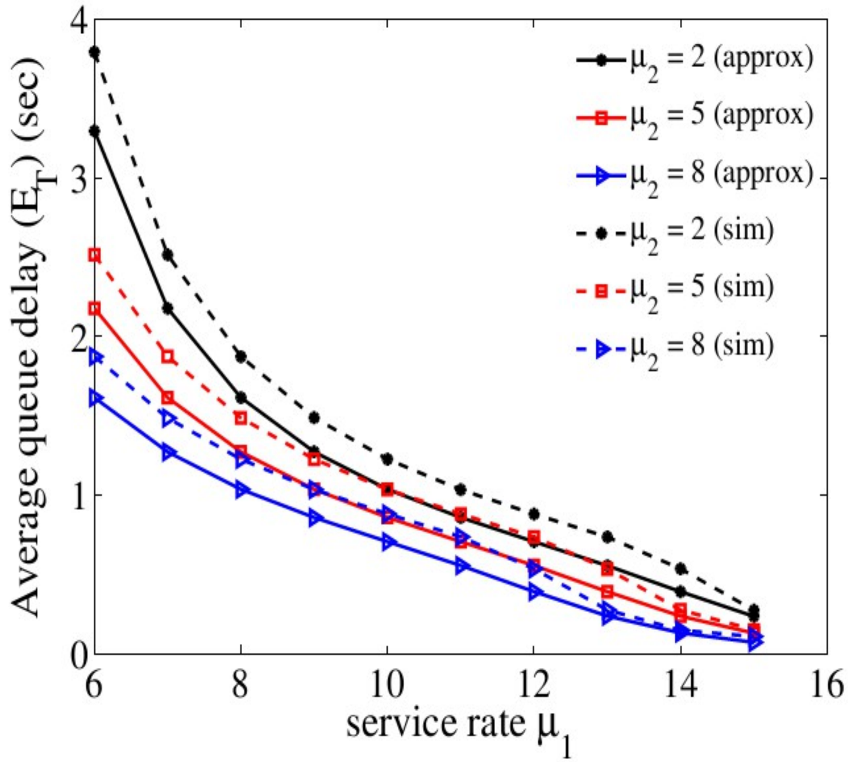
- Arrival rates
- Average queue length
- Average queue delay
- Drop Probability
- Throughput
- Buffer Size



Queue length and Waiting time (Delay)



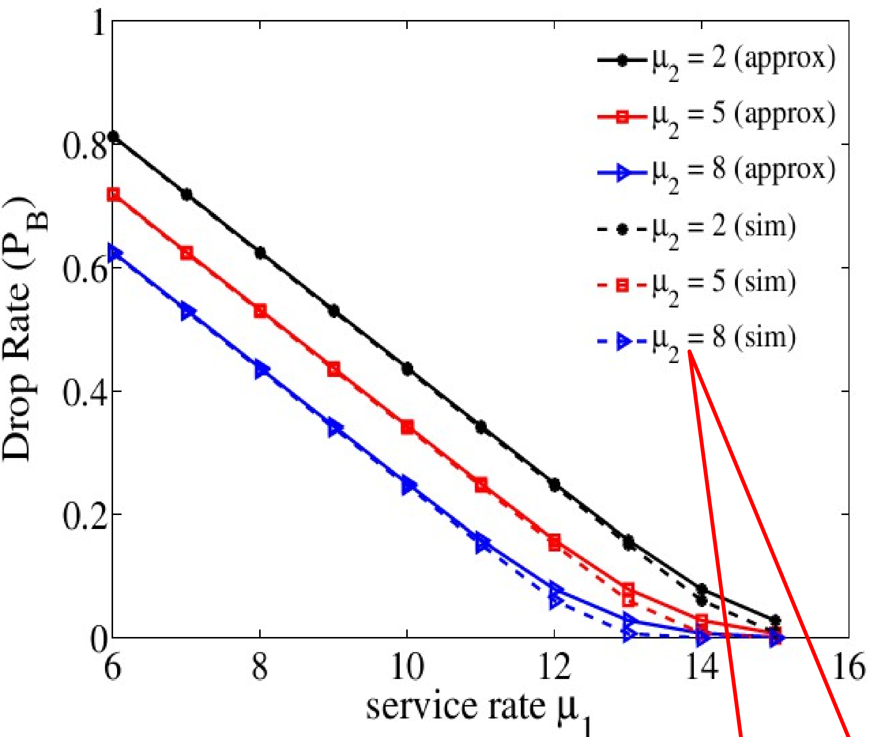
Queue length comparison between analytical and simulation



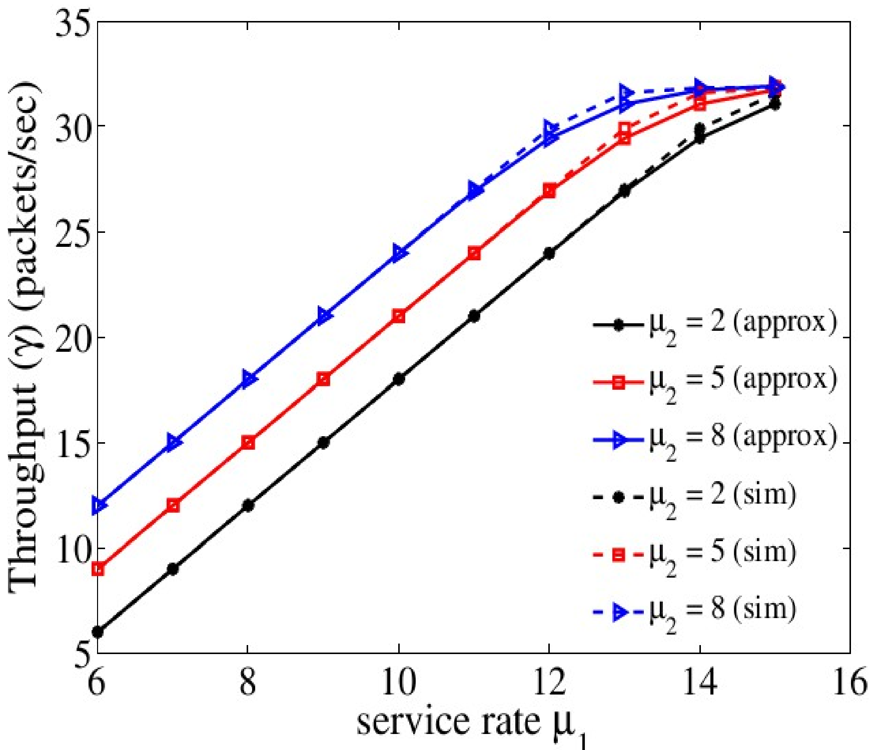
Waiting time comparison between analytical and simulation

Good approximation

Drop rate and Throughput



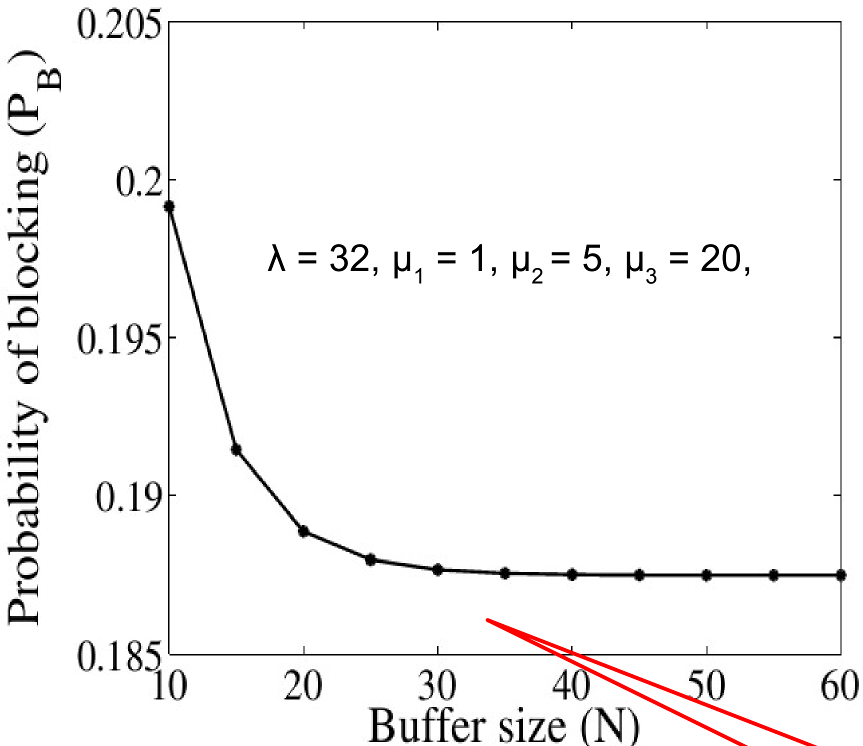
Drop rate comparison between analytical and simulation



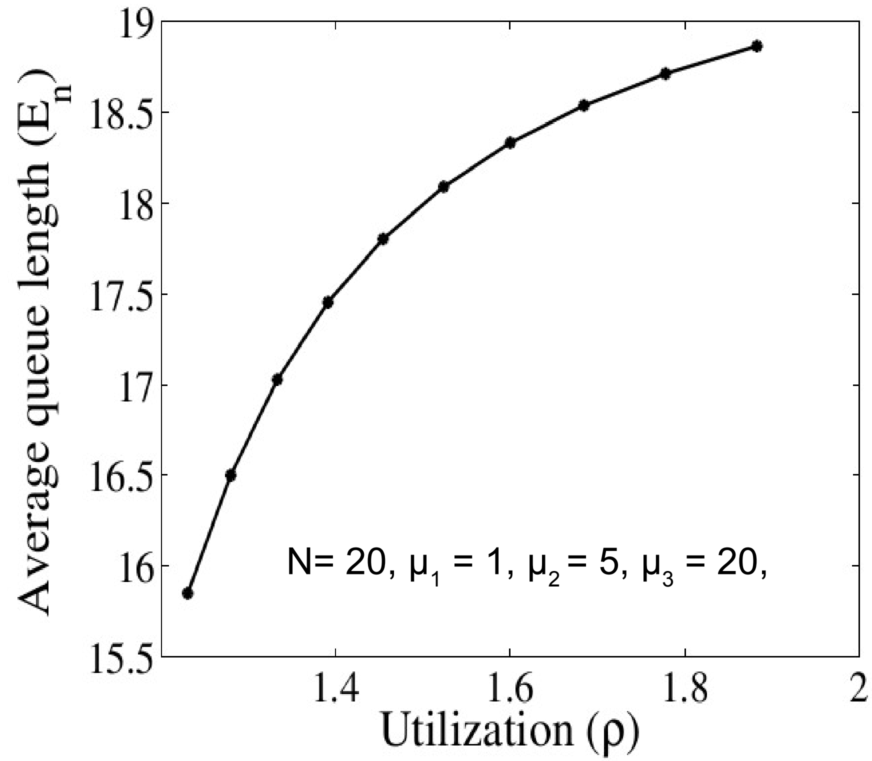
Throughput comparison between analytical and simulation

Very good approximation

Drop Rate and Queue Length



Effect of buffer size to drop rate



Effect of utilization to queue length

After N=30, buffer size has no effect

Content

- Motivation
- Contribution
- Multi Heterogeneous System First Model
    - Analysis of First Model
- Multi Heterogeneous System Second Model
    - Analysis of Second Model
- Conclusion
- References

Conclusion

- Approximation formulas developed
  - Queue length
  - Waiting time
  - Drop and Throughput for finite queue
- Verified by simulation
- Tested by different allocations,
  - FSF, SSF, RCS

Future Work

- Different allocation can be used
  - LUF : Low utilization first
- Allocation policies behaviors in in finite queue
- Developing Probabilistic allocation approximation
- Considering multi class system with different class and flexibility level

Questions

References

[1] T. Heath, B. Diniz, E. V. Carrera, W. M. Jr., and R. Bianchini, “Energy conservation in heterogeneous server clusters,” in Principles and Practice of Parallel Programming, Chicago, Illinois, June 2005, pp. 186–195.

[2] S. Gurumurthi and S. Benjaafar, “Modeling and analysis of flexible queueing systems,” Naval Research Logistics, vol. 51, pp. 755–782, June 2004.

[3] F. S. Q. Alves, H. C. Yehia, L. A. C. Pedrosa, F. R. B. Cruz, and L. Kerbache, “Upper bounds on performance measures of heterogeneous M/M/c queues,” Mathematical Problems in Engineering, vol. 2011, p. 18, May 2011.

[4] C. Misra and P. K. Swain, “Performance analysis of finite buffer queueing system with multiple heterogeneous servers,” in 6th international conference on Distributed Computing and Internet Technology, ser. ICDCIT’10, Bhubaneswar, India, Feb 2010, pp. 180–183.

[5] G. Appenzeller, I. Keslassy, and N. McKeown, “Sizing router buffers,” Computer Communication Review, vol. 34, pp. 281–292, Oct 2004.

Probability of States of First Model

$$1 = \sum_{j=0}^{\infty} p_j = \sum_{j=0}^{c} p_j + \sum_{j=c+1}^{\infty} p_j$$

$$p_0^{-1} = 1 + \sum_{j=1}^{c} \left(\frac{\lambda^j}{\prod_{i=1}^{j} \mu_{ti}} \right) + \sum_{j=c+1}^{\infty} \left(\frac{\lambda^j}{\left(\prod_{i=1}^{c} \mu_{ti} \right) \left(\mu_{tc}^{j-c} \right)} \right)$$

$$p_0^{-1} = 1 + \sum_{j=1}^{c-1} \left(\frac{\lambda^j}{\prod_{i=1}^{j} \mu_{ti}} \right) + \sum_{j=c}^{\infty} \left(\frac{\lambda^j}{\left(\prod_{i=1}^{c} \mu_{ti} \right) \left(\mu_{tc}^{j-c} \right)} \right)$$

$$p_0^{-1} = 1 + \sum_{j=1}^{c-1} \left(\frac{\lambda^j}{\prod_{i=1}^{j} \mu_{ti}} \right) + \left(\frac{\mu_{tc}^{c}}{\prod_{i=1}^{c} \mu_{ti}} \right) \sum_{j=c}^{\infty} \rho^j$$

$$\sum_{j=c}^{\infty} \rho^j = \frac{\rho^c}{1 - \rho}$$

$$p_0 = \frac{1}{1 + \sum_{j=1}^{c-1} \left(\frac{\lambda^j}{\prod_{i=1}^{j} \mu_{ti}} \right) + \left(\frac{\lambda^c}{(1-\rho) \prod_{i=1}^{c} \mu_{ti}} \right)}$$

Probability of States of Second Model

$$1 = \sum_{j=0}^{N+3} p_j = p_0 + p_0 \frac{\lambda}{\mu_{t1}} + p_0 \frac{\lambda^2}{\mu_{t1}\mu_{t2}} + \sum_{j=3}^{N+3} p_0 \frac{\mu_{t3}^2 \rho^j}{\mu_{t1}\mu_{t2}}$$

$$p_0^{-1} = 1 + \frac{\lambda}{\mu_{t1}} + \frac{\lambda^2}{\mu_{t1}\mu_{t2}} + \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \sum_{j=3}^{N+3} \rho^j$$

$$p_0 = \frac{1}{1 + \frac{\lambda}{\mu_{t1}} + \frac{\lambda^2}{\mu_{t1}\mu_{t2}} + \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \sum_{j=3}^{N+3} \rho^j}$$

$$p_0 = \frac{1}{1 + \frac{\lambda}{\mu_{t1}} + \frac{\lambda^2}{\mu_{t1}\mu_{t2}} + (N+1)\frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}}}$$ if $\rho = 1$

$$p_0 = \begin{cases} \frac{1}{1+\frac{\lambda}{\mu_{t1}}+\frac{\lambda^2}{\mu_{t1}\mu_{t2}}+\frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}}\frac{\rho^3-\rho^{N+4}}{1-\rho}} & \rho \neq 1 \\ \frac{1}{1+\frac{\lambda}{\mu_{t1}}+\frac{\lambda^2}{\mu_{t1}\mu_{t2}}+(N+1)\frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}}} & \rho = 1 \end{cases}$$

$$\sum_{j=3}^{N+3} \rho^j = \frac{\rho^3 - \rho^{N+4}}{1-\rho}$$ if $\rho \neq 1$