SMMSO 2015

Equilibrium abandonment strategies in a cloud management system (CMS): A queueing approach

Gopinath Panda Veena Goswami† A D Banik

School of Basic Sciences
Indian Institute of Technology Bhubaneswar, India.

† KIIT University Bhubaneswar, India.

10th Conference on Stochastic Models of Manufacturing and Service Operations- SMMSO 2015

Volos, Greece

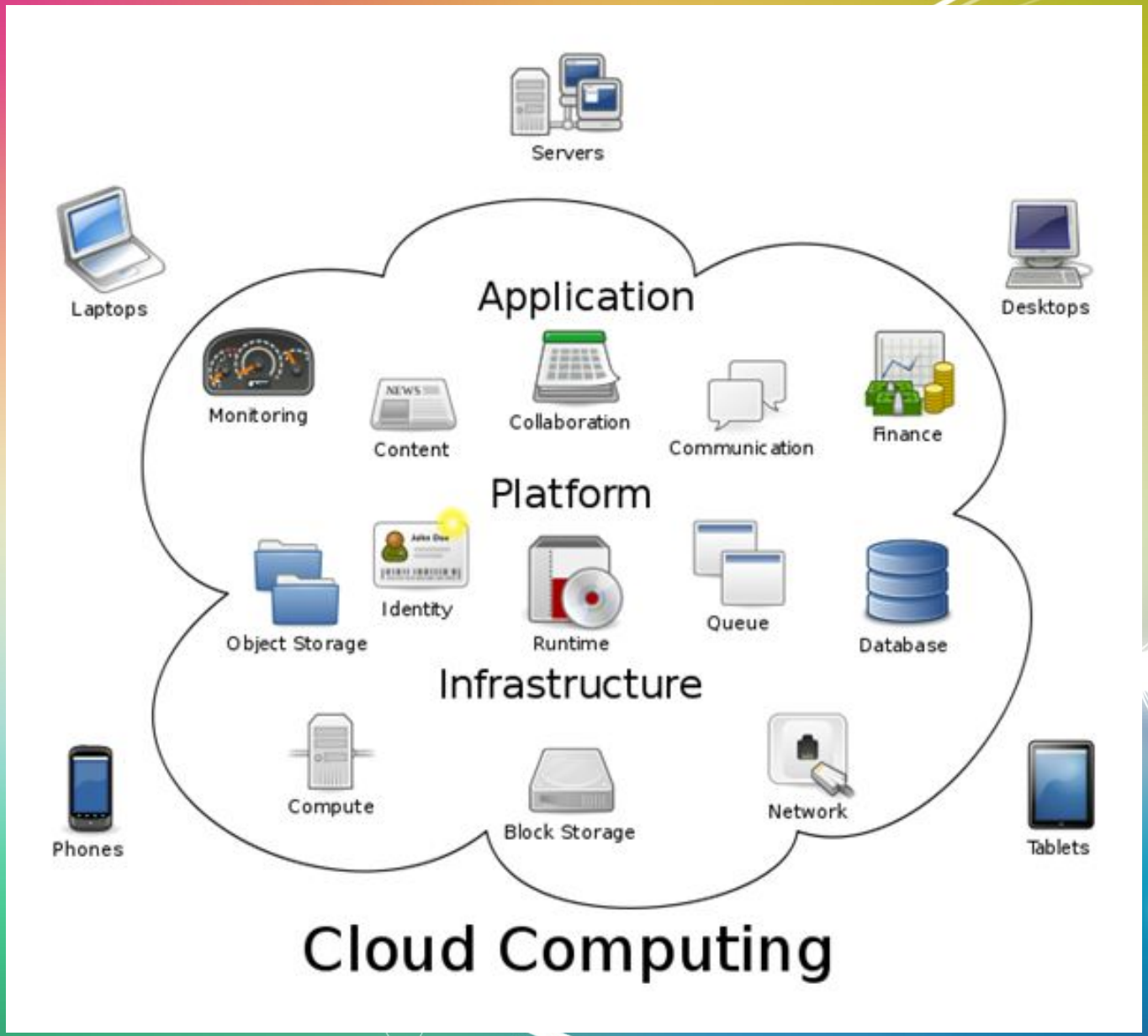
3 June 2015

Outline

Introduction

- Cloud computing (CC) service model
  - Infrastructure-as-a-Service (IaaS)
  - Platform-as-a-Service (PaaS)
  - Software-as-a-Service (SaaS)
- CC deployment models:
  Public, Private, Hybrid.
- CC environment: Cloud User (CU);
  Cloud Service Provider (CSP).

SMMSO 2015


Servers
Laptops
Desktops
Application
Monitoring
Content
Collaboration
Communication
Finance
Platform
Identity
Object Storage
Runtime
Queue
Database
Infrastructure
Compute
Block Storage
Network
Phones
Tablets
Cloud Computing

IaaS

- CSPs $---> $ virtual servers (CPUs running several choices of operating systems and a customized software stack with storage and networking).
- CUs $---> $ manage applications, data, runtime, middleware, and OSes.
- Amazon Web Services offers: S3 (storage), EC2 (virtual servers), Cloudfront (content delivery), Cloudfront Streaming (video streaming),SimpleDB (structured datastore), RDS (Relational Database), SQS (reliable messaging), and Elastic MapReduce (data processing).
- Microsoft Azure

PaaS

- CSPs $---$>platforms to develop, run and manage Web applications.
- $---$>programming languages (Python, Java, .NET, Ruby, Apex etc.), frameworks, libraries, services (Web applications) and tools (hardware, software).
- Third party manages and controls the infrastructure, including network, servers, operating systems and storage.
- CU have control over the deployed applications and configuration settings for the application hosting environment.
- Salesforce.com$--$ > enterprise CRM platform. Amazon Web Services, Google App Engine and Heroku are platforms for software development and management.

SaaS

- CSPs $---$>applications to customers over the web.
- $---$> office and messaging software, payroll processing, DBMS,CAD software, accounting, collaboration, CRM, ERP, HRM, content management, email and health care-related applications.
- Microsoft Azure, Amazon Web services, Google Apps, Salesforce, Adobe, Oracle ,SAP, Workday etc.

Benefits of cloud users

- Rapid Service
- Secure Service
- Lower Costs
- Multi-User Access
- Development Platform
- Infinite Storage
- High Scalability
- Pay per use
- Location independence

Challenges of CSP

- Cost
- Power consumption
- Management
- Profit optimization

Literature survey

Sl.	Researchers	Works
1	Zhang et al. (2010)	Cloud computing challenges
2	Ortiz et al. (2014)	Economic study of cloud applications
3	Keskin et al. (2014)	Benefit from impatient users
4	Fan et al.(2013)	Game theoretic study of cloud queue
5	Mitrani (2011)	Customers impatience and power consumption
6	Chiang (2014)	Performance analysis of cloud queue with balking and reneging

Abandonments

- **Independent:** Abandon the system independently if patience timer expires before starting of service.
- **Synchronized:** Customers wait for certain transport facility (Piosson process) to abandon the system.
  N_s decreases according to a binomial distribution.
- **Geometric:** The transport facility has limited capacity.
  N_s decreases according to a geometric distribution.

Objectives

In this model, we will study the

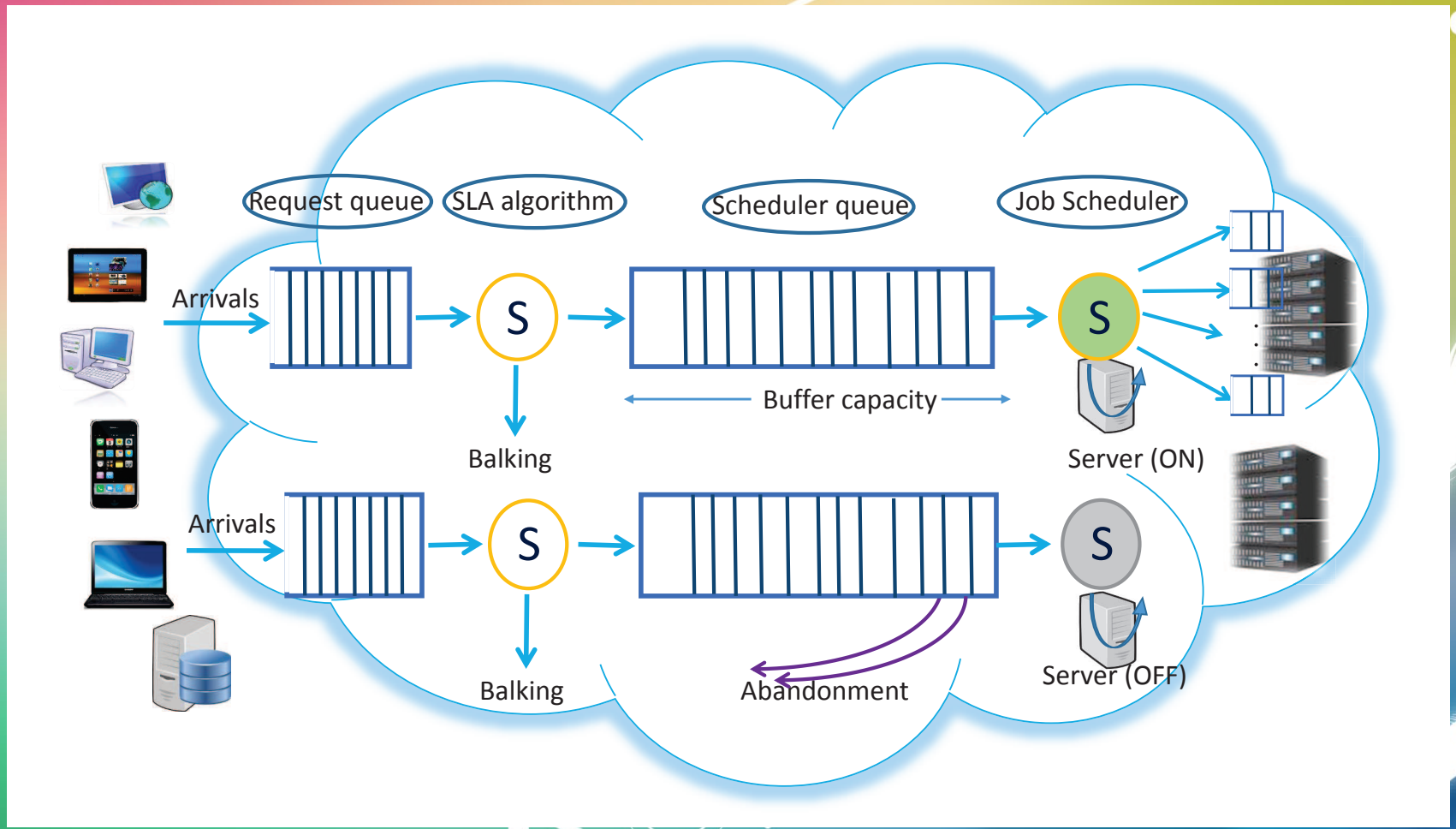
1. Equilibrium balking strategies for individual customers
2. Net benefit of customers
3. Mean sojourn time of customers
4. Profit optimization of CSP

Cloud architecture

- CU $--> $ CSP $--> $Request Queue(RQ).
- CMS selects the best sequence of jobs using a scheduling algorithm and stores them into Buffer Queue.
- Then Job Scheduler (JS) assigns the requests to suitable servers.
- After service completion, requests are stored in a transmission queue to be sent to the users.

SMMSO 2015

Cloud queueing model



Methodology

- Embedded Markov chain approach
- Supplementary variable method
- Roots method
- Approximation method
- Combinatorial method
- Simulation method

Model Parameter

- Arrival $--> $ Poisson process with rate λ.
- Service times $--> \exp(\mu)$.
- Single server with finite buffer capacity.
- First-come first-served (FCFS) service discipline.
- Traffic intensity: $\rho = \lambda/\mu$.
- Server vacation times: $\exp(\gamma)$.
- Abandonment epochs: $\exp(\phi)$.
- A request remains in the system with probability q or abandons with probability $1 - q$.

Notation

- $N_s(t)$ = number of customers in the system.
- $\zeta(t) = \begin{cases} 0, & \text{server idle,} \\ 1, & \text{server active.} \end{cases}$
- Balking: refusing to join the queue.
- Abandonment: joins the queue but leaves without service.
- $T(n,i)$- mean sojourn time of n-th user when the server is on state i.
- Net benefit of a user after service completion, $\Delta = R - C\ T(n,i)$.
- $L_e(0)$ $(L_e(1))$ buffer capacity when server is inactive (active) with $L_e(0) < L_e(1)$.

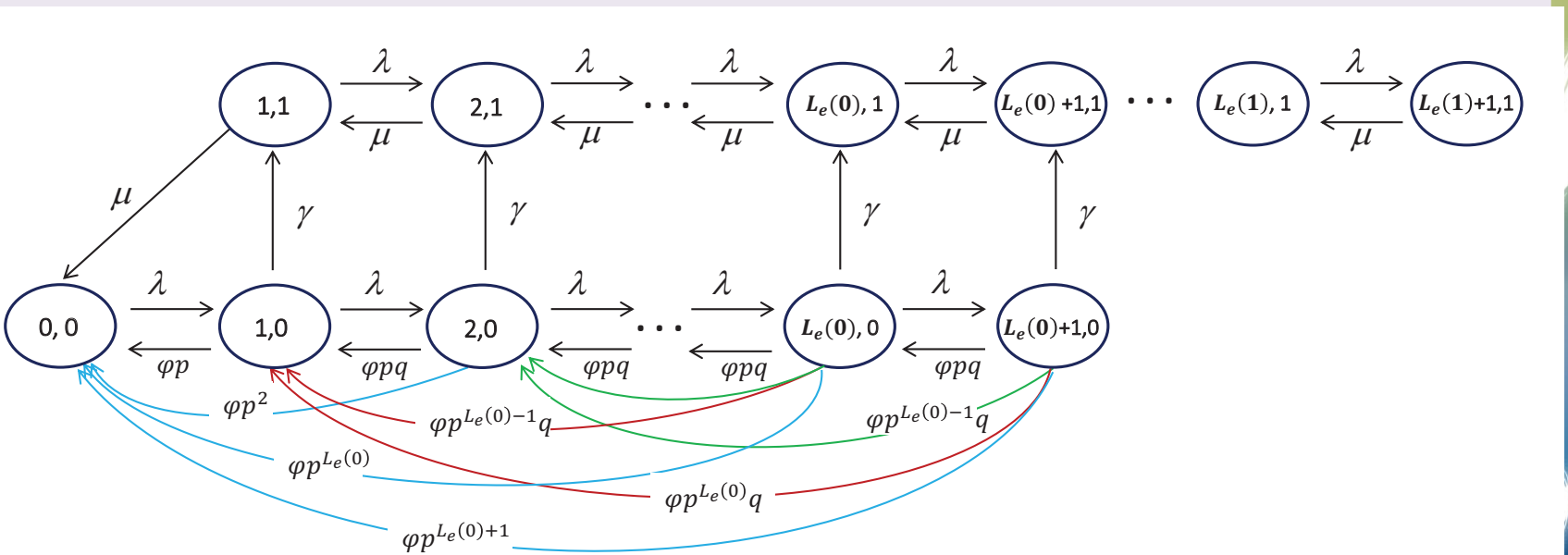
Markov chain and state transitions

- $\{(N_s(t), \zeta(t)) : t \geq 0\}$ is a continuous time Markov chain (CTMC) with state space $\Omega = \{(n, i) : \ i = 0, 1, \ n \geq i\}$.
- $\pi_{n,i}$ - stationary distribution of the CTMC.
- The non-zero transition rates between two states are

$$q_{(n,i),(n+1,i)} = \lambda, \quad n \geq i, \ i = 0, 1,$$
$$q_{(n+1,1),(n,1)} = \mu, \quad n \geq 1,$$
$$q_{(1,1),(0,0)} = \mu,$$
$$q_{(n,0),(n,1)} = \gamma, \quad n \geq 1,$$
$$q_{(n+k,0),(n,0)} = \phi p^k q, \quad 0 \leq k \leq n-1, \ n \geq 1,$$
$$q_{(n,0),(0,0)} = \phi p^n, \quad n \geq 1.$$

- Fully observable case- arriving customers observe both $N_s(t), \zeta(t)$

Transition diagram (Fully observable case)



Balance equations

$$(\lambda+\phi)\pi_{0,0} = \mu\pi_{1,1} + \phi \sum_{j=0}^{L_e(0)+1} p^j \pi_{j,0}$$

$$(\lambda+\gamma+\phi)\pi_{n,0} = \lambda\pi_{n-1,0} + \phi \sum_{j=n}^{L_e(0)+1} qp^{j-n}\pi_{j,0}, 1 \le n \le L_e(0),$$

$$(\gamma+\phi)\pi_{L_e(0)+1,0} = \lambda\pi_{L_e(0),0} + \phi q\pi_{L_e(0)+1,0},$$

$$(\lambda+\mu)\pi_{1,1} = \gamma\pi_{1,0} + \mu\pi_{2,1},$$

$$(\lambda+\mu)\pi_{n,1} = \gamma\pi_{n,0} + \mu\pi_{n+1,1} + \lambda\pi_{n-1,1}, 2 \le n \le L_e(0)+1,$$

$$(\lambda+\mu)\pi_{n,1} = \lambda\pi_{n-1,1} + \mu\pi_{n+1,1}, L_e(0)+2 \le n \le L_e(1),$$

$$\lambda\pi_{L_e(1),1} = \mu\pi_{L_e(1)+1,1}.$$

Mean sojourn time

Let us define $a = \dfrac{1}{\gamma+\phi}$, $b = \dfrac{\gamma}{\gamma+\phi}\dfrac{1}{\mu}$, $c = \dfrac{\phi}{\gamma+\phi}$

$$T(0,0) = a + b,$$

$$T(n,0) = T(0,0) + n\ b + c\ \sum_{j=1}^{n} p^j q^{n-j}\, T(n-j,0), n \geq 1,$$

$$T(n,1) = \frac{n+1}{\mu}, \quad n \geq 1.$$

On successive iteration for $n \geq 2$,

$$T(n,0) = T(1,0) \prod_{i=1}^{n-1}(1+cq^i)p + \sum_{i=2}^{n}(a+b+ibq) \prod_{j=i}^{n-1}(1+cq^j)p,$$

Threshold values

- We have $\Delta_{fo}(n,0) = R - C\ T(n,0)$.
- $R - CT(0,0) > 0$, i.e., $R > C\left(\frac{1}{\gamma+\phi} + \frac{\gamma}{\gamma+\phi}\frac{1}{\mu}\right)$.
- The threshold $(L_e(0), L_e(1))$, where $L_e(0)$ is the unique root of the equation

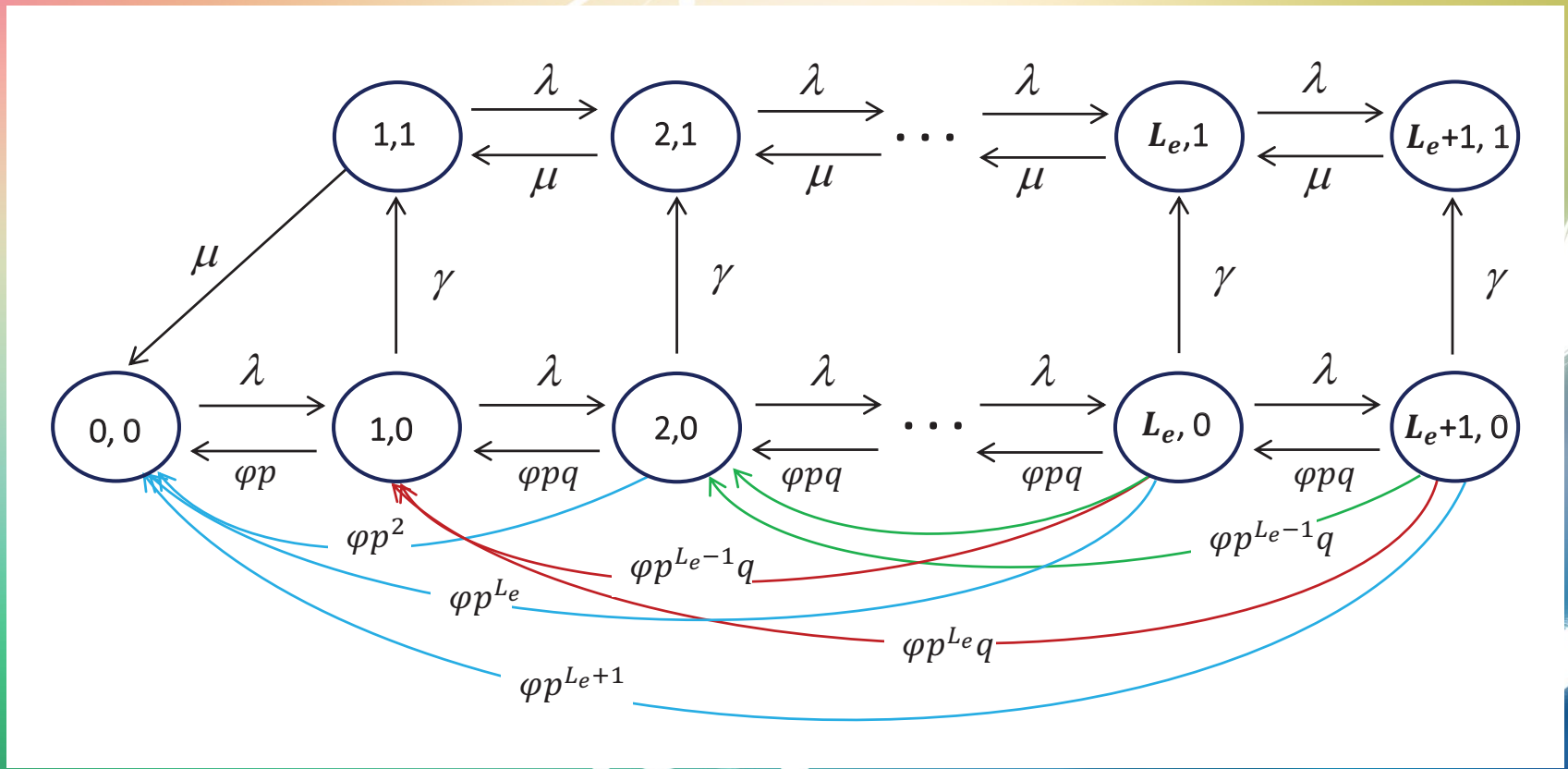
$$\frac{R}{C} = \left((1+cp)\frac{a}{q} + (2+cp)b\right)\prod_{i=1}^{n-1}(1+cq^i)p + \sum_{i=2}^{n}(a+b+ibq)\prod_{j=i}^{n-1}(1+cq^j)p,$$

and $$L_e(1) = \lfloor\frac{\mu R}{C}\rfloor - 1.$$

- Blocking probability $P_{block} = \pi_{L_e(0)+1,0} + \pi_{L_e(1)+1,1}$.

Transition rate diagram (almost observable case)

- Arriving customers observe only $N_s(t)$



Balance equations

$$(\lambda+\phi)\pi_{0,0} = \mu\pi_{1,1} + \phi\sum_{j=0}^{L_e+1} p^j\pi_{j,0},$$

$$(\lambda+\gamma+\phi)\pi_{n,0} = \lambda\pi_{n-1,0} + \phi\sum_{j=n}^{L_e+1} qp^{j-n}\pi_{j,0}, 1 \le n \le L_e,$$

$$(\gamma+\phi)\pi_{L_e+1,0} = \lambda\pi_{L_e,0} + \phi q\pi_{L_e+1,0},$$

$$(\lambda+\mu)\pi_{1,1} = \gamma\pi_{1,0} + \mu\pi_{2,1},$$

$$(\lambda+\mu)\pi_{n,1} = \gamma\pi_{n,0} + \mu\pi_{n+1,1} + \lambda\pi_{n-1,1}, \quad 2 \le n \le L_e,$$

$$\mu\pi_{L_e+1,1} = \gamma\pi_{L_e+1,0} + \lambda\pi_{L_e,1}.$$

Numerical parameters

We consider an $M/M/1/N$ queue with the following model parameters

Fully observable model

$\lambda = 2.4, \mu = 4.7, p = 0.3, \phi = 2.0, \gamma = 1.5, L_e(0) = 30, L_e(1) = 40,$ and $\rho = 0.510638.$

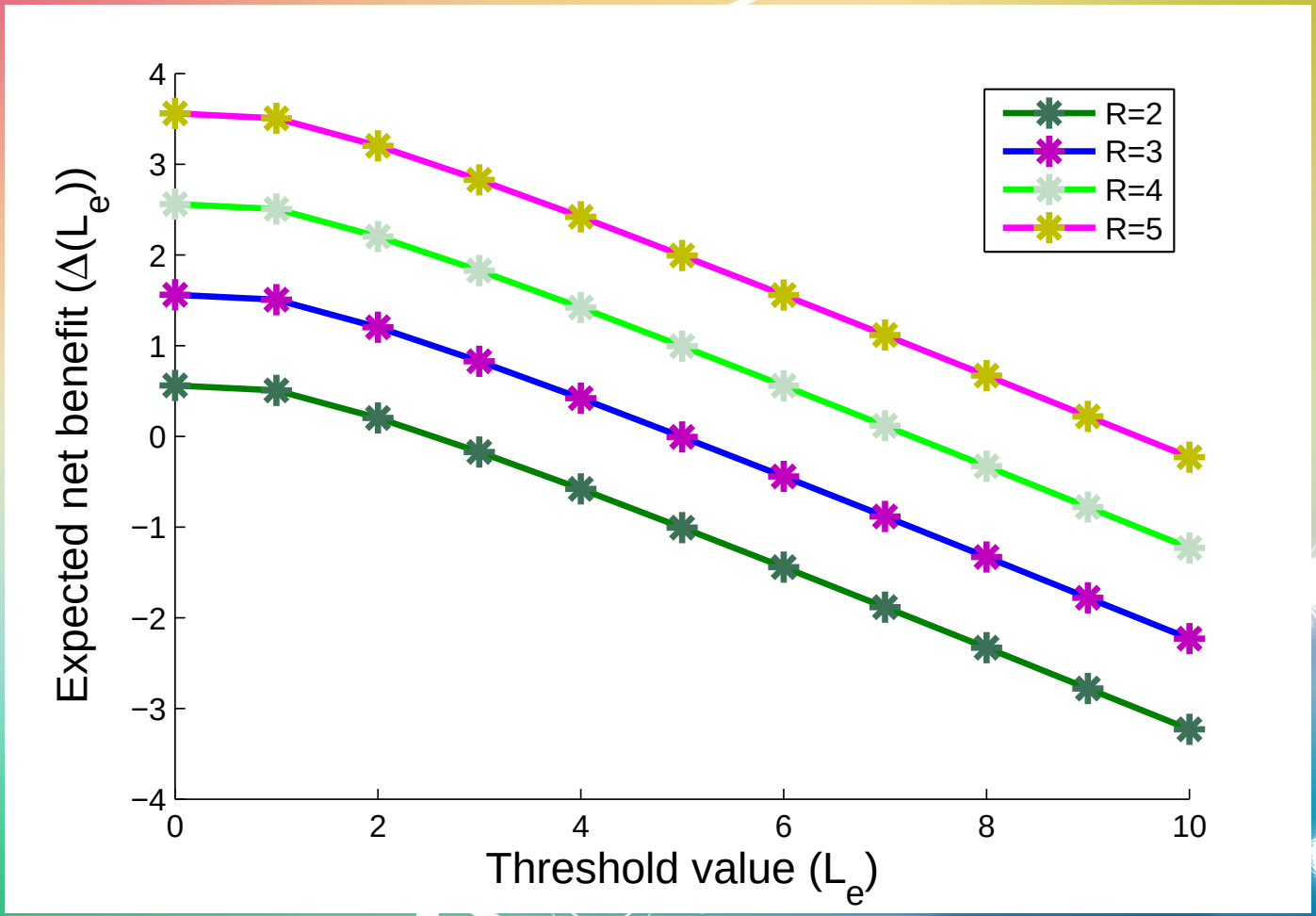
Almost observable model

$\lambda = 4.5, \mu = 10.5, \phi = 1.0, \gamma = 3.7, p = 0.3, L_e = 30$ and $\rho = 0.428571.$

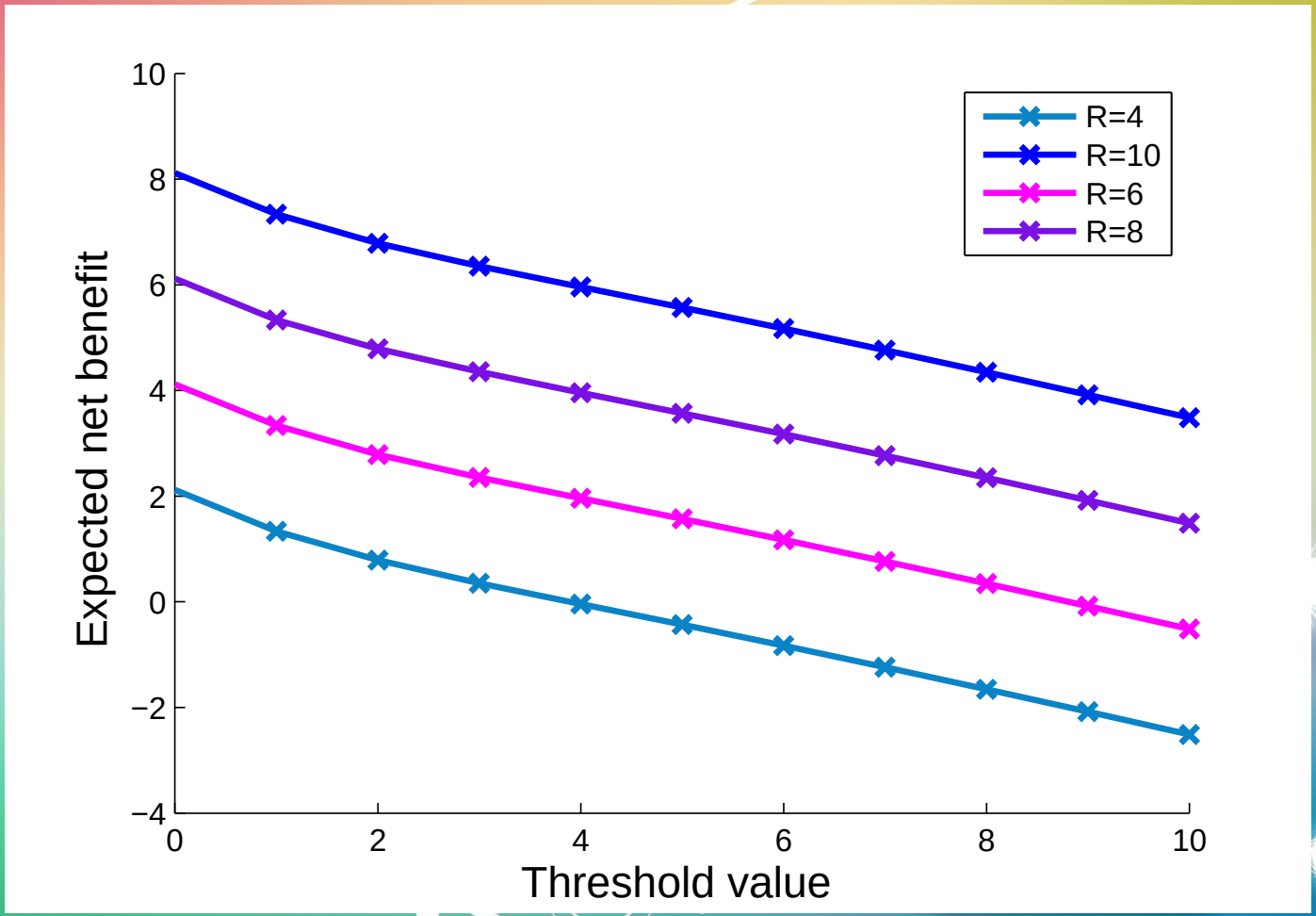
Steady state system-length distribution

Fully observable case			Almost observable case	
n	$\pi_{n,0}$	$\pi_{n,1}$	$\pi_{n,0}$	$\pi_{n,1}$
0	0.230723	0.000000	0.268966	0.000000
1	0.131489	0.097569	0.144686	0.110342
2	0.074935	0.105426	0.077832	0.106646
3	0.042705	0.085523	0.041869	0.077636
4	0.024338	0.061731	0.022523	0.050449
5	0.013870	0.041814	0.012116	0.030861
10	0.000834	0.003965	0.000546	0.001836
15	0.000050	0.000289	0.000025	0.000089
20	0.000003	0.000019	0.000001	0.000004
25	0.000000	0.000001	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮
Sum	0.536438	0.463562	0.582097	0.417903

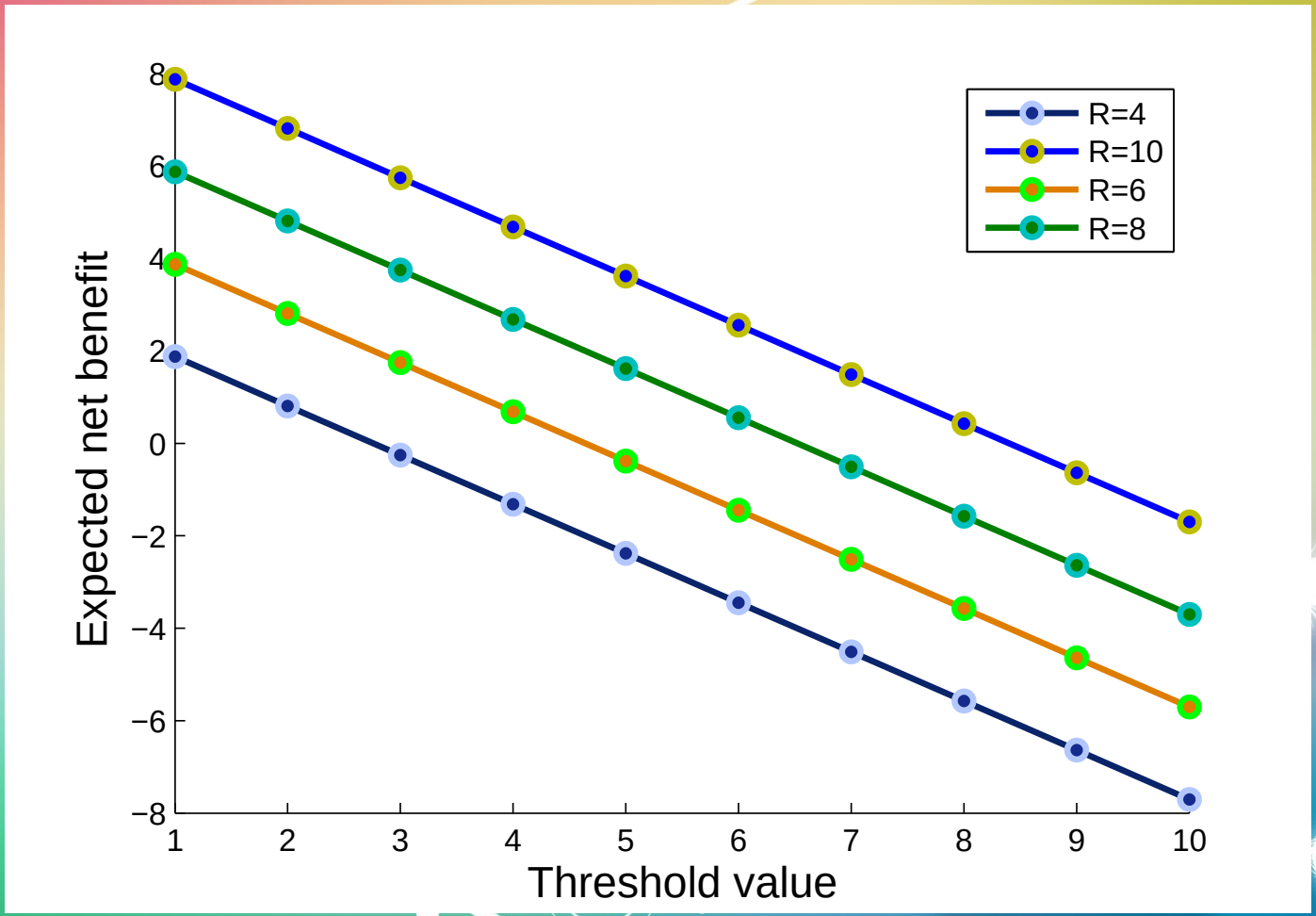
$\Delta(L_e)$ vs L_e for almost observable case



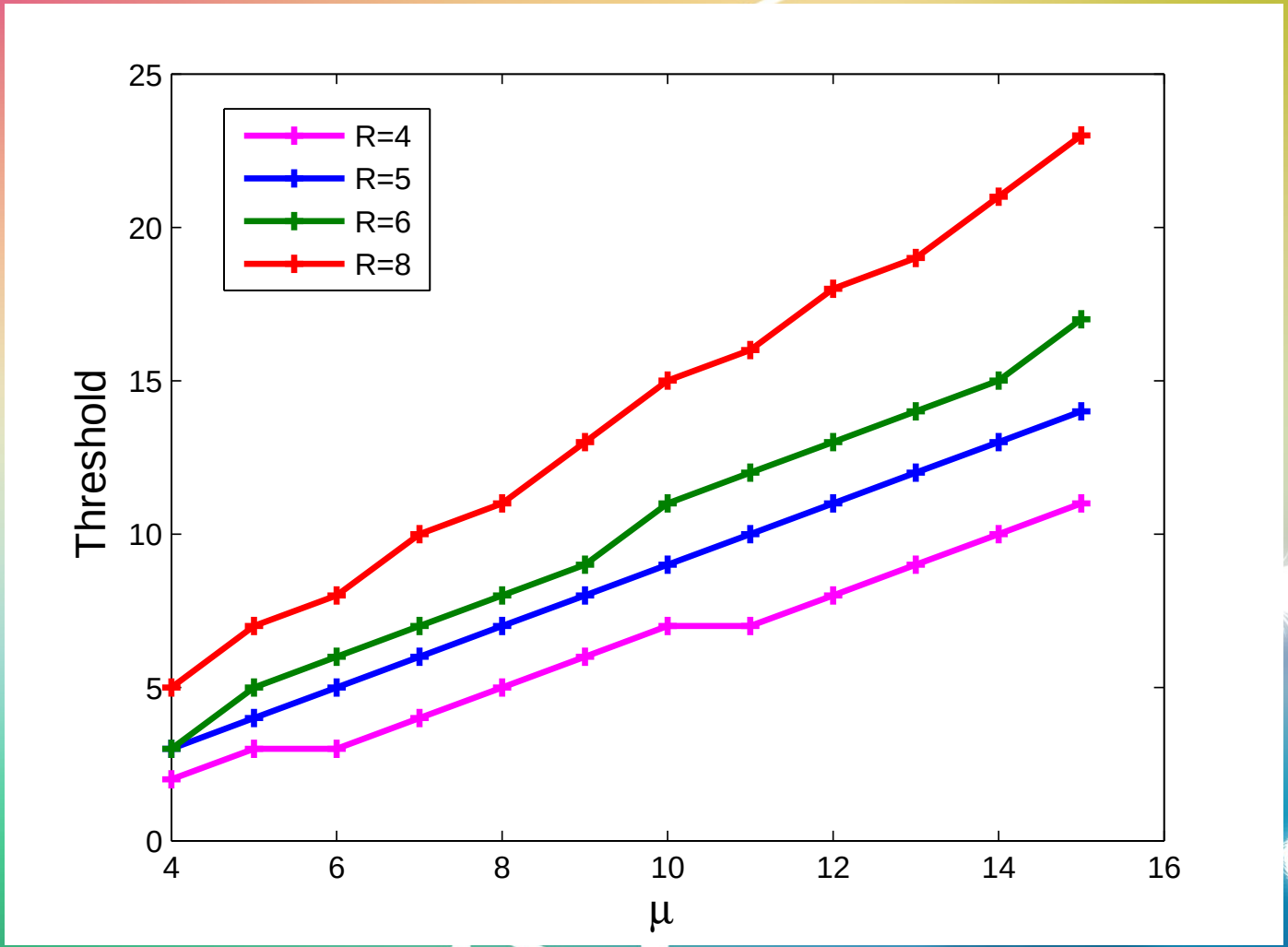
$\Delta(L_e(0))$ vs $L_e(0)$ for fully observable case



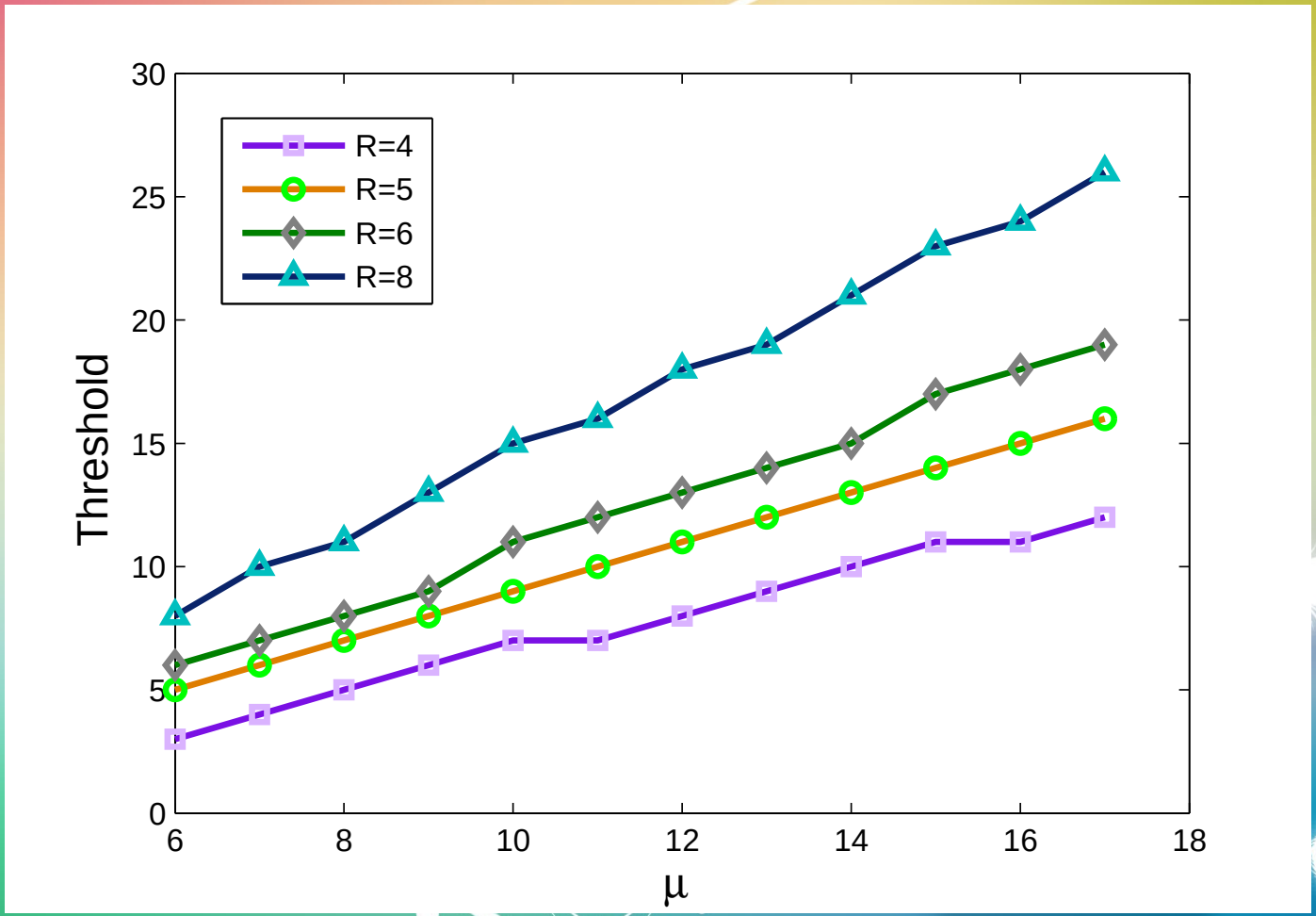
$\Delta(L_e(1))$ vs $L_e(1)$ for fully observable case

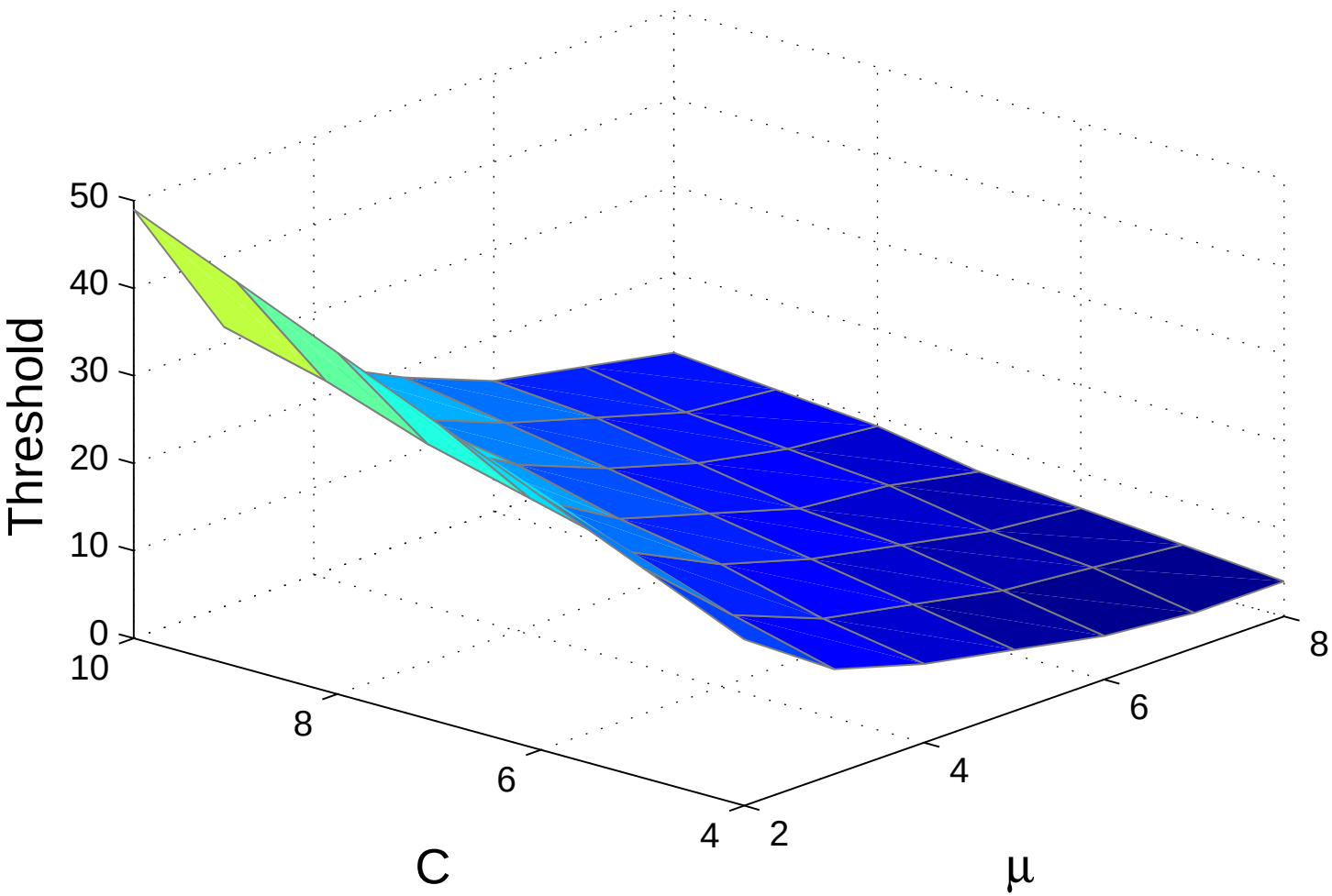


$L_e(1)$ vs μ for fully observable case

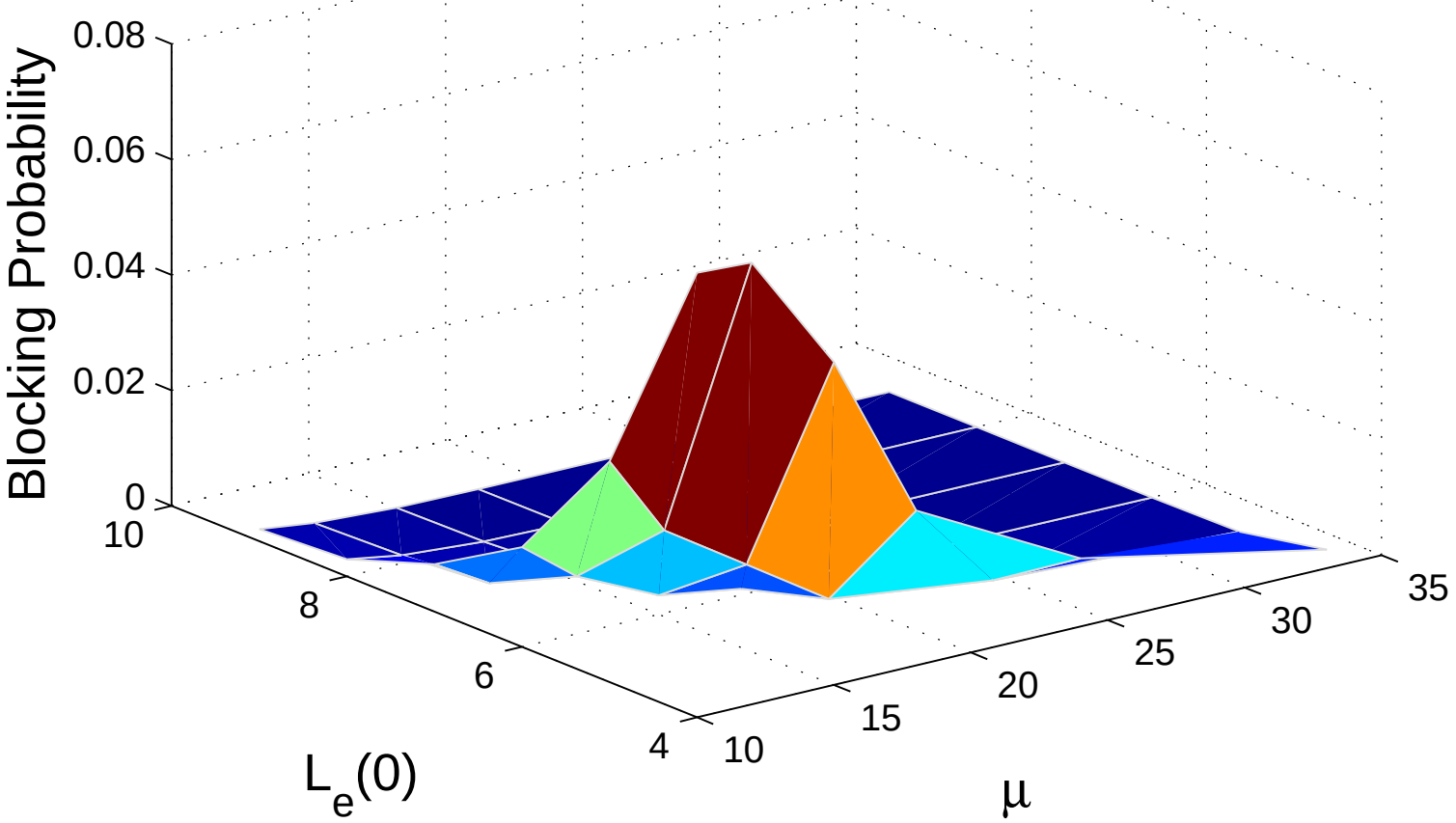


L_e vs μ for almost observable case





Threshold control in fully observable case



Blocking probability for fully observable case

Conclusion

This model will help the CSP in decision making regarding

- Vacation time of the server (power consumption).
- Waiting cost and service completion reward (profit).
- Threshold values (smooth management).
- Service rate control (cost).

Future Scope

- Multi server models
- Mobile cloud computing (Retrial queueing model, unreliable server)
- Specific cloud applications like ERP, CRM, Emergency services etc.
- A group of arrivals with general service requirement

References I

S. Dimou, A. Economou, and D. Fakinos.
The single server vacation queueing model with geometric abandonments.
Journal of Statistical Planning and Inference, 141(8): 2863–2877, 2011.

F. Zhang, J. Wang, and B. Liu.
Equilibrium balking strategies in Markovian queues with working vacations.
Applied Mathematical Modelling, 37(16):8264–8282, 2013.

Y. J. Chiang and Y. C. Ouyang.
Profit optimization in SLA-aware cloud services with a finite capacity queuing model.
Mathematical Problems in Engineering, 2014, 2014a.

References II


X. Nan, Y. He, and L. Guan.
Optimal resource allocation for multimedia cloud based on queuing model.
In *Multimedia Signal Processing (MMSP), 13th International Workshop*, pages 1–6. IEEE, 2011.


Y. J. Chiang and Y. C. Ouyang.
An Optimal Buffer Control Algorithm to Maximize Profit in Cloud Computing.
In *Computer, Consumer and Control (IS3C), 2014 International Symposium on*, pages 252–255. IEEE, 2014b.


Qi Zhang, Lu Cheng, and Raouf Boutaba.
Cloud computing: state-of-the-art and research challenges.
Journal of internet services and applications, 1(1):7–18, 2010.

References III

Raul Pena-Ortiz, Josep Domenech, Jose A Gil, and Ana Pont.
An approach for economic evaluation of cloud-based applications.
In *Cloud Networking (CloudNet), 2014 IEEE 3rd International Conference on*, pages 281–287. IEEE, 2014.

Tayfun Keskin and Nazim Taskin.
A pricing model for cloud computing service.
In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 699–707. IEEE, 2014.

References IV


Guisheng Fan, Huiqun Yu, Liqiong Chen, and Dongmei Liu.
A game theoretic method to model and evaluate attack-defense strategy in cloud computing.
In *Services Computing (SCC), 2013 IEEE International Conference on*, pages 659–666. IEEE, 2013.


Isi Mitrani.
Service center trade-offs between customer impatience and power consumption.
Performance Evaluation, 68(11):1222–1231, 2011.

Acknowledgement

- Prof. George Liberopoulos
- UNIVERSITY OF THESSALY SCHOOL OF ENGINEERING | DEPARTMENT OF MECHANICAL ENGINEERING
- INDIAN INSTITUTE OF TECHNOLOGY BHUBANESWAR
- A.D. Banik supported by DST, New Delhi, India research grant SR/FTP/MS-003/2012.

Thank You