Συστήματα Αναμονής
2η Ομάδα Ασκήσεων

Ονοματεπώνυμο: Κουστένης Χρίστος
Α.Μ : el20227
Ακαδημαϊκό έτος: 2022-2023

Θεωρητική μελέτη της ουράς M/M/1

Στην άσκηση αυτή, θα μελετηθεί θεωρητικά η ουρά M/M/1. Οι αφίξεις στην ουρά ακολουθούν κατανομή Poisson με παράμετρο λ πελάτες/sec και οι εξυπηρετήσεις ακολουθούν εκθετική κατανομή με παράμετρο μ πελάτες/sec

(α)

Η συνθήκη που είναι απαραίτητη ώστε να εξασφαλίζεται η εργοδικότητα και η ισορροπία της ουράς Markov M/M/1 είναι $\rho = \frac{\lambda}{\mu} < 1$. Η ποσότητα $\rho = \frac{\lambda}{\mu}$ ονομάζεται ένταση κυκλοφορίας (traffic intensity) του συστήματος. Η ένταση κυκλοφορίας εκφράζει την πιθανότητα να μην είναι άδειο το σύστημα ή ισοδύναμα να είναι απασχολημένη η μονάδα εξυπηρέτησης, γι' αυτό ονομάζεται και βαθμός χρησιμοποίησης (utilization) της μονάδας εξυπηρέτησης. Η παραπάνω συνθήκη είναι γνωστή και ως συνθήκη Erlang.

Οι εξισώσεις

$$-(\lambda_n + \mu_n)p_n + \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} = 0,\ n \geq 1$$
$$-\lambda_0 p_0 + \mu_1 p_1 = 0$$

απευθύνονται στη γενική περίπτωση συστήματος αναμονής γεννήσεων-θανάτων όπου οι παράμετροι λ_n και μ_n αντιπροσωπεύουν τον ρυθμό γεννήσεων και τον ρυθμό θανάτων, αντίστοιχα,όταν η διαδικασία βρίσκεται στην κατάσταση n. Τέλος, $p_n(t)$ ονομάζουμε την πιθανότητα να βρίσκεται η διαδικασία στην κατάσταση αυτή τη χρονική στιγμή t. Ενδιαφερόμαστε για την ύπαρξη των οριακών πιθανοτήτων: $p_n = \lim_{t\to\infty} p_n(t)$. Η γενική λύση του συστήματος εξισώσεων είναι :

$$p_n = p_0 \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}},\ n = 1, 2, \ldots$$

όπου η πιθανότητα p0 προσδιορίζεται με βάση τη συνθήκη:

$$p_0 = \left[1 + \sum_{n=1}^{\infty} \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}}\right]^{-1}$$

Το M/M/1 είναι το απλούστερο σύστημα αναμονής και χαρακτηρίζεται από αφίξεις Poisson (εκθετικά κατανεμημένους χρόνους μεταξύ αφίξεων) και εκθετικά κατανεμημένους χρόνους εξυπηρέτησης. Οι ρυθμοί γεννήσεων–θανάτων είναι σταθεροί και ανεξάρτητοι από την κατάσταση του συστήματος:

$\lambda_n = \lambda$, n = 0, 1, 2, . . .

$\mu_n = \mu$, μ = 1, 2, . . .

Έτσι απλοποιείται η γενική λύση ως εξής :

$$
\begin{aligned}
p_n &= p_0 \left(\frac{\lambda}{\mu}\right)^n, \; n \geq 0 \\
p_0 &= \left[\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n\right]^{-1}
\end{aligned}
$$

με την προϋπόθεση ότι ισχύει $S = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n < \infty$

Δεδομένου ότι ισχύει η συνθήκη Erlang $\frac{\lambda}{\mu} < 1$ η δυναμοσειρά $\sum_{n=0}^{\infty}(\frac{\lambda}{\mu})^n$ υπολογίζεται ως

$\sum_{n=0}^{\infty}(\frac{\lambda}{\mu})^n = \frac{1}{1-\frac{\lambda}{\mu}}$.

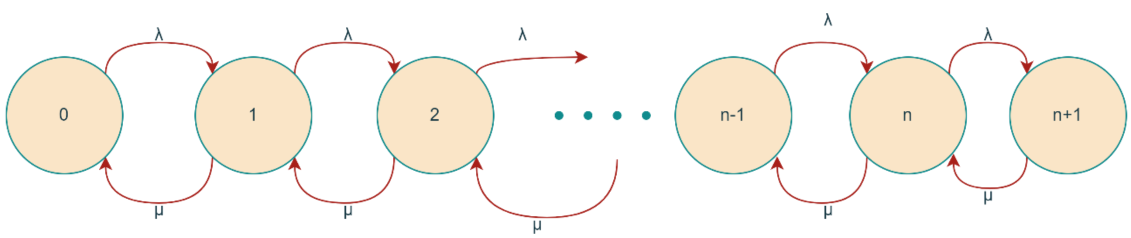
Επομένως, p_0 = 1-ρ και

έχουμε :

$$p_n = (1-\rho)\rho^n, \; n = 0, 1, 2, \ldots$$

όπου $\rho = \frac{\lambda}{\mu}$ η ένταση της κυκλοφορίας στο σύστημα.

Ακολουθεί το διάγραμμα ρυθμού μεταβάσεων της ουράς Μ/Μ/1 :



Εικόνα 1

(β)

Προκύπτει ότι ο μέσος όρος πελατών στο σύστημα που ισορροπεί είναι ίσος με:

$$\mathrm{E}[n] = \frac{\rho}{1-\rho}$$

Χρησιμοποιώντας τον τύπο του Little **(L = λW)** ο μέσος χρόνος παραμονής στο σύστημα ή αλλιώς ο μέσος χρόνος απόκρισης θα είναι :

$$T = \mathrm{E}[n]/\lambda = \frac{1}{\mu} \cdot \frac{1}{1-\rho}$$

Ο χρόνος παραμονής ενός πελάτη στο σύστημα είναι το άθροισμα του χρόνου αναμονής και του χρόνου εξυπηρέτησης. Άρα ο μέσος χρόνος αναμονής(καθυστέρησης) όταν το σύστημα βρίσκεται σε ισορροπία θα είναι θα είναι:

$$W = T - \frac{1}{\mu} = \frac{1}{\mu} \cdot \frac{\rho}{1-\rho}$$

(γ)

Θεωρητικά το σύστημα θα μπορούσε να βρίσκεται σε κατάσταση με 57 πελάτες, όμως, η πιθανότητα είναι πολύ μικρή για ρ < 1 γεγονός που καθιστά πρακτικά απίθανο αυτό το ενδεχόμενο. Ενδεικτικά για ρ = 0.5 ισχύει:

$p_{57} = (1 - 0.5) \cdot (0.5)^{57} = 3.469447 \times 10^{-18} << 1$

Παρατηρούμε ωστόσο ότι όταν το συστήμα παρουσιάζει ένταση κυκλοφορίας κοντά στη μονάδα η πιθανότητα αυτή αυξάνεται σημαντικά. Ενδεικτικά για ρ = 0.95 ισχύει :

$p_{57} = (1 - 0.95) \cdot (0.95)^{57} = 0.0026866773$

Ανάλυση ουράς M/M/1 με Octave

Στην άσκηση αυτή θα συγκρίνουμε την επίδοση διαφορετικών συστημάτων αναμονής με τη βοήθεια του πακέτου queueing του Octave.

(α)

Για να είναι το σύστημα εργοδικό πρέπει να ισχύει ρ = λ/μ < 1. Επειδή λ = 5 πελάτες/min προκύπτει ότι η εργοδικότητα εξασφαλίζεται για μ > 5 πελάτες/min.

Με βάση τη μέγιστη δυνατότητα ρυθμού εξυπηρέτησης του συστήματος μας που είναι 10 πελάτες/min προκύπτει $5 < \mu \leq 10$.

(β)

Αρχικά χρησιμοποιούμε τη συνάρτηση qsmm1() του Octave από του queueing πακέτο του Octave :

```
% system M/M/1
clc;
clear all;
close all;

lambda = 5;
mu = 5.001:0.001:10; % The only possible values so the system be ergodic

[U, R, Q, X, p0] = qsmm1 (lambda, mu)
```

Για παραμέτρους της συνάρτησης ορίζουμε την τιμη λ του συστήματος που μας δίνεται και το διάνυσμα mu που αντιστοιχεί στη μεταβλητή μ του ρυθμού εξυπηρέτησης και παίρνει τιμές στο διάστημα (5,10] όπως εξηγήσαμε στο (α) ερώτημα. Η συνάρτηση qsmm() θα επιστρέψει πέντε μεγέθη όπως φαίνεται από τον τρόπο που ορίστηκε:

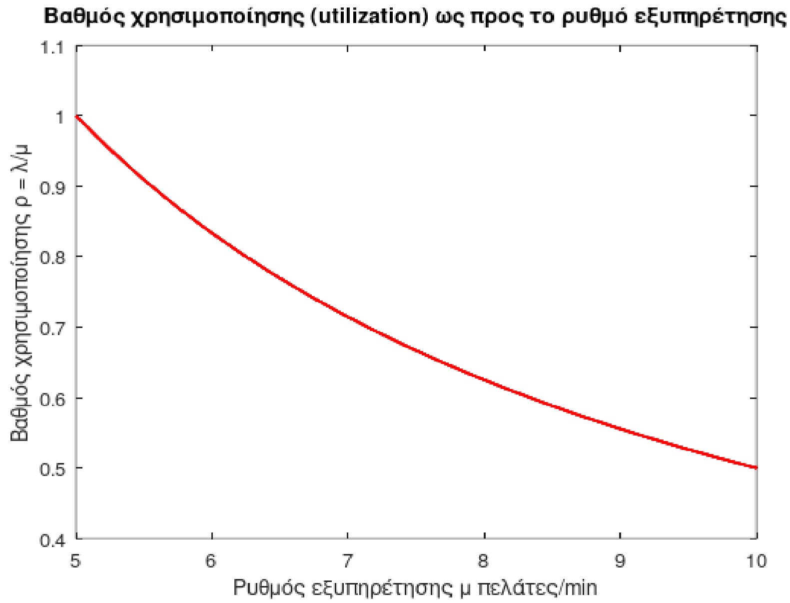
- U = Βαθμός χρησιμοποίησης του συστήματος
- R = Μέσος χρόνος απόκρισης εξυπηρετητή
- Q = Μέσος αριθμός πελατών
- X = Ρυθμαπόδοση(Throughtput)

Θα χρησιμοποιήσουμε καθένα από αυτά για τα ακόλουθα ερωτήματα

- Για το διάγραμμα του βαθμού χρησιμοποίησης (utilization) ως προς το ρυθμό εξυπηρέτησης χρησιμοποιούμε τον παρακάτω κώδικα στο Octave που μας δίνει την Γραφική 1

```
% Server Utilization
figure(1);
plot(mu,U,"r","linewidth",1.4);
title("Βαθμός χρησιμοποίησης (utilization) ως προς το ρυθμό εξυπηρέτησης");
xlabel("Ρυθμός εξυπηρέτησης μ πελάτες/min");
ylabel("Βαθμός χρησιμοποίησης ρ = λ/μ");
```

Κώδικας 1



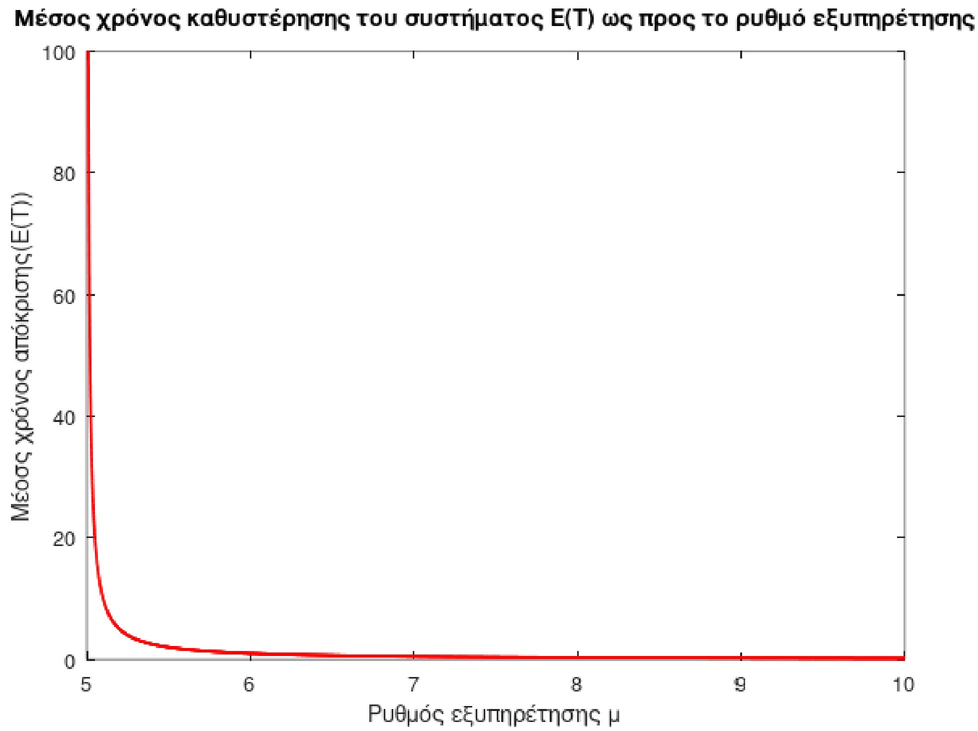
Γραφική 1

- Για το διάγραμμα του μέσου χρόνου καθυστέρησης του συστήματος E(T) ως προς το ρυθμό εξυπηρέτησης χρησιμοποιούμε τον Κώδικα 2 στο Octave που μας δίνει την Γραφική 2.

```
% Average server response time
figure(2);
plot(mu,R,"r","linewidth",1.4);
axis ([5 10 0 100]);
title("Μέσος χρόνος καθυστέρησης του συστήματος E(T) ως προς το ρυθμό εξυπηρέτησης");
xlabel("Ρυθμός εξυπηρέτησης μ");
ylabel("Μέσος χρόνος απόκρισης(E(T))");

```

Κώδικας 2



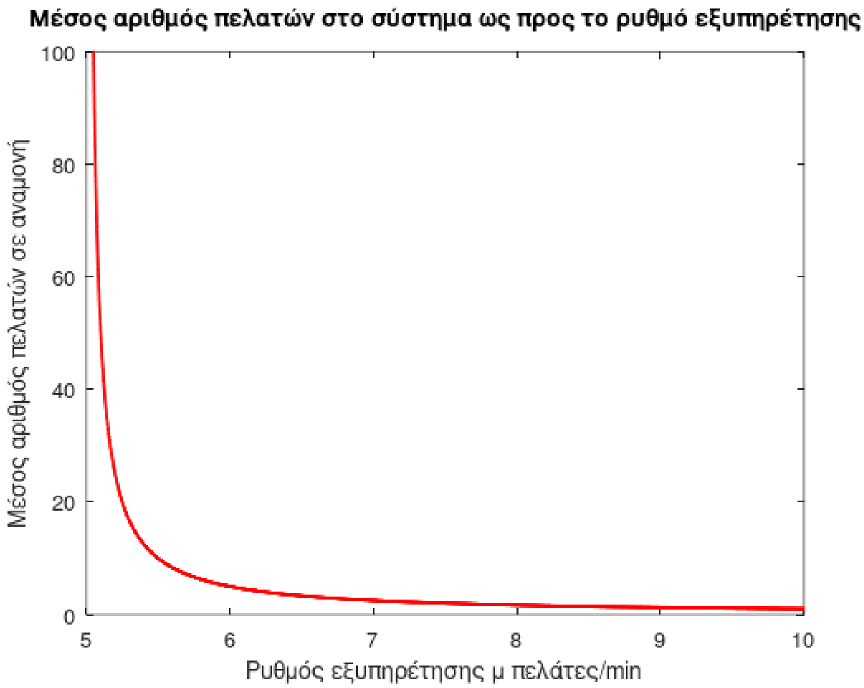
Γραφική 2

- Για το διάγραμμα του μέσου αριθμού πελατών στο σύστημα ως προς το ρυθμό εξυπηρέτησης χρησιμοποιούμε τον Κώδικα 3 στο Octave που μας δίνει την Γραφική 3.

```
% Average number of request in the system
figure(3);
plot(mu,Q,"r","linewidth",1.4);
axis ([5 10 0 100]);
title("Μέσος αριθμός πελατών στο σύστημα ως προς το ρυθμό εξυπηρέτησης");
xlabel("Ρυθμός εξυπηρέτησης μ πελάτες/min");
ylabel("Μέσος αριθμός πελατών σε αναμονή");

```

Κώδικας 3

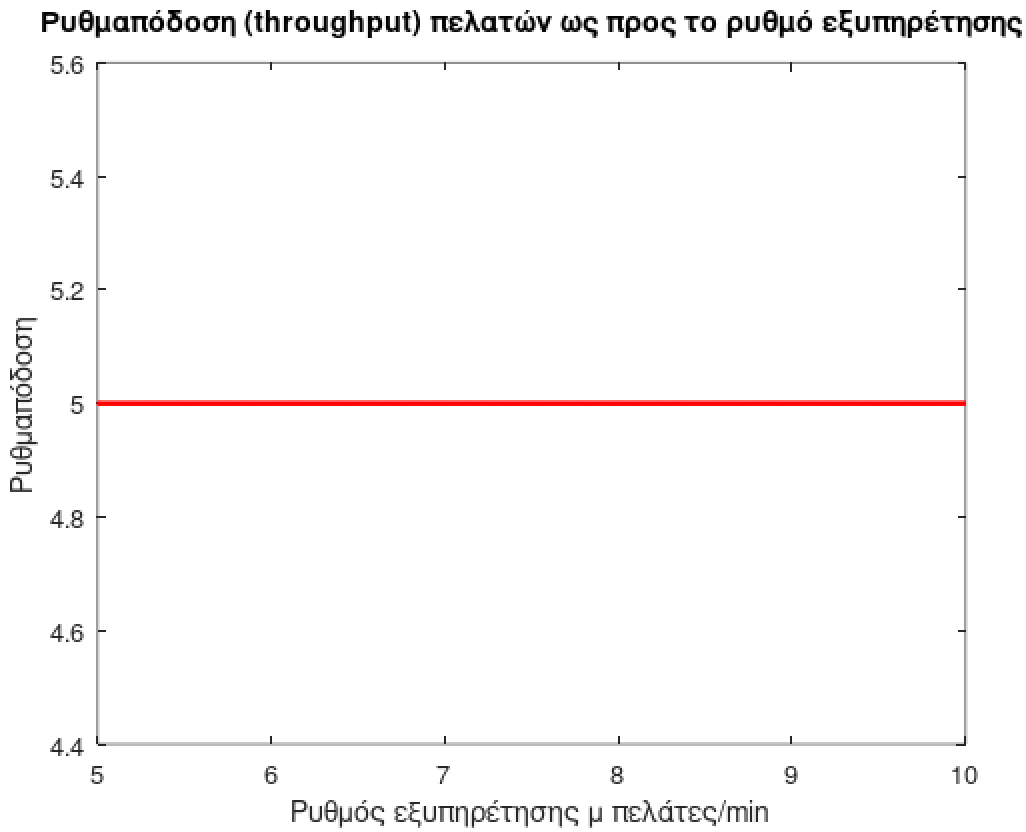


Γραφική 3

- Για το διάγραμμα του ρυθμαπόδοσης (throughput) πελατών ως προς το ρυθμό εξυπηρέτησης χρησιμοποιούμε τον Κώδικα 4 στο Octave που μας δίνει την Γραφική 4.

```
% Throughtput of the system
figure(4);
plot(mu,X,"r","linewidth",1.4);
title("Ρυθμαπόδοση (throughput) πελατών ως προς το ρυθμό εξυπηρέτησης");
xlabel("Ρυθμός εξυπηρέτησης μ πελάτες/min");
ylabel("Ρυθμαπόδοση");
```

Κώδικας 4



Γραφική 4

(γ)

Με βάση τη Γραφική 2 είναι φανερό ότι ο μέσος χρόνος καθυστέρησης του συστήματος και ο ρυθμός εξυπηρέτησης είναι αντιστρόφως ανάλογα ποσά Επομένως, μεγαλύτεροι ρυθμοί εξυπηρέτησης ευνοούν λειτουργικά το σύστημα αφού ο μέσος χρόνος καθυστέρησης ελαχιστοποιείται. Είναι λογικό λοιπόν λαμβάνοντας μόνο τον παράγοντα αυτόν υπόψιν να επιλέξουμε τον μέγιστο δυνατό ρυθμό εξυπηρέτησης δηλαδή μ = 10 πελάτες/min. Ωστόσο, σε πραγματικές εφαρμογές η συγκεκριμένη επιλογή θα ήταν πολύ κοστοβόρα και άρα θα ήταν πιο ρεαλιστικό να συμβιβαστούμε με έναν μέσο ρυθμό εξυπηρέτησης 7 ή 8 που θα μετρίαζε το κόστος εξασφαλίζοντας συγχρόνως ικανοποιητικά μεγάλη ταχύτητα.

(δ)

Είναι γνωστό ότι η ρυθμαπόδοση(throughtput) σε μια εργοδική ουρά M/M/1 χωρίς απορρίψεις πελατών ταυτίζεται με την παράμετρο λ που εκφράζει τον ρυθμο άφιξης πελατών στο σύστημα όπως προκύπτει απο τον ορισμό της ρυθμαπόδοσης :

$$\gamma = \lambda(1 - \mathrm{P}\{\mathrm{blocking}\}) \leq \lambda, \qquad \gamma < \mu$$

Αυτό επιβεβαιώνεται από τη Γραφική 4 η οποία είναι μια συνάρτηση σταθερή στην τιμή 5 που είναι και η δοθείσα για το λ.

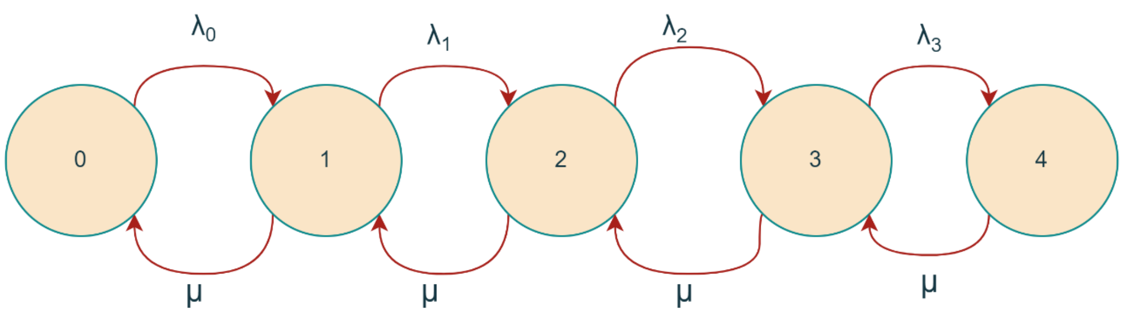
Διαδικασία γεννήσεων θανάτων (birth-death process): εφαρμογή σε σύστημα M/M/1/K

Δίνεται σύστημα αναμονής M/M/1/4, δηλαδή σύστημα με 1 εξυπηρετητή και μέγιστη χωρητικότητα 4 πελάτες. Οι αφίξεις στο σύστημα ακολουθούν κατανομή Poisson με μέσο ρυθμό $\boldsymbol{\lambda_i = \lambda/(i+1)}$, δηλαδή οι αφίξεις πελατών εξαρτώνται από την κατάσταση στην οποία βρίσκεται το σύστημα (μειώνονται ανάλογα με τον παρόντα αριθμό πελατών στο σύστημα) και ο χρόνος εξυπηρέτησης του εξυπηρετητή είναι εκθετικός με μέσο ρυθμό $\mu_i = \mu$, ανεξάρτητος από τον αριθμό πελατών στο σύστημα.

Θεωρούμε ότι **λ=5 πελάτες/sec** και **μ=10 πελάτες/sec**. Επίσης, θεωρούμε ότι αρχικά (στο χρόνο 0) στο σύστημα δεν υπάρχει κανένας πελάτης.

(α)

Το διάγραμμα ρυθμού μεταβάσεων του συστήματος αναμονής M/M/1/4 φαίνεται στην Εικόνα 2.



Εικόνα 2

Για το συγκεκριμένο σύστημα με **$\lambda = \lambda_n = \lambda/(n+1)$** και **$\mu_i = \mu$** για **$0 \leq i \leq 4$** ισχύουν οι παρακάτω εξισώσεις :

1. **$\lambda_{n-1}\, p_{n-1} = \mu_n p_n$, $1 \leq n \leq 4$** (Εξίσωση ισορροπίας)
2. **$\sum_{n=0}^{4}$ $p_n = 1$** (Συνθήκη κανονικοποίησης)

Επομένως λύνοντας βρίσκουμε τις εργοδικές πιθανότητες :

$$\boldsymbol{P_0} = \frac{\mathbf{1}}{\mathbf{1}+\frac{\lambda_0}{\mu}+\frac{\lambda_0\lambda_1}{\mu^2}+\frac{\lambda_0\lambda_1\lambda_2}{\mu^3}+\frac{\lambda_0\lambda_1\lambda_2\lambda_3}{\mu^4}} = 0.606635$$

$$\boldsymbol{P_1} = \frac{\lambda_0}{\mu}\boldsymbol{P_0} = 0.303318$$

$$\boldsymbol{P_2} = \frac{\lambda_1}{\mu}\boldsymbol{P_1} = 0.0758294$$

$$\boldsymbol{P_3} = \frac{\lambda_2}{m}\boldsymbol{P_2} = 0.0126382$$

$$\boldsymbol{P_4} = \frac{\lambda_3}{\mu}\boldsymbol{P_3} = 0.0015798$$

Η πιθανότητα της τελευταίας κατάστασης προφανώς εκφράζει

(P{blocking} = P_4 = 0.0015798) την πιθανότητα απόρριψης πελάτη λόγω πληρότητας του συστήματος.

(β)

(i)

Η μήτρα ρυθμού μεταβάσεων όπως εμφανίζεται στο Octave στην Εικόνα 3 αφού τρέξαμε τον Κωδικα 5.

```
% system M/M/1/4
% when there are 3 clients in the system, the capability of the server doubles.

clc;
clear all;
close all;

lambda = 5;

lambda_i = [lambda lambda/2 lambda/3 lambda/4];

mu = 10;
states = [0, 1, 2, 3, 4]; % system with capacity 4 states
% the initial state of the system. The system is initially empty.
initial_state = [1, 0, 0, 0, 0];


% define the birth and death rates between the states of the system.
births_B = [lambda_i(1), lambda_i(2), lambda_i(3), lambda_i(4)];
deaths_D = [mu, mu, mu, mu];
%(i)
% get the transition matrix of the birth-death process
transition_matrix = ctmcbd(births_B, deaths_D);
```

Κώδικας 5

```
>> transition_matrix
transition_matrix =

   -5.0000    5.0000         0         0         0
   10.0000  -12.5000    2.5000         0         0
         0   10.0000  -11.6667    1.6667         0
         0         0   10.0000  -11.2500    1.2500
         0         0         0   10.0000  -10.0000
```

Εικόνα 3

(ii)

Έπειτα με την εντολή ctmc του πακέτου queueing υπολογίζουμε τις εργοδικές πιθανότητες των καταστάσεων του συστήματος που αποθηκεύονται στις αντίστοιχες θέσεις του διανύσματος P.

```
P =

   6.0664e-01   3.0332e-01   7.5829e-02   1.2638e-02   1.5798e-03
```

Εικόνα 4

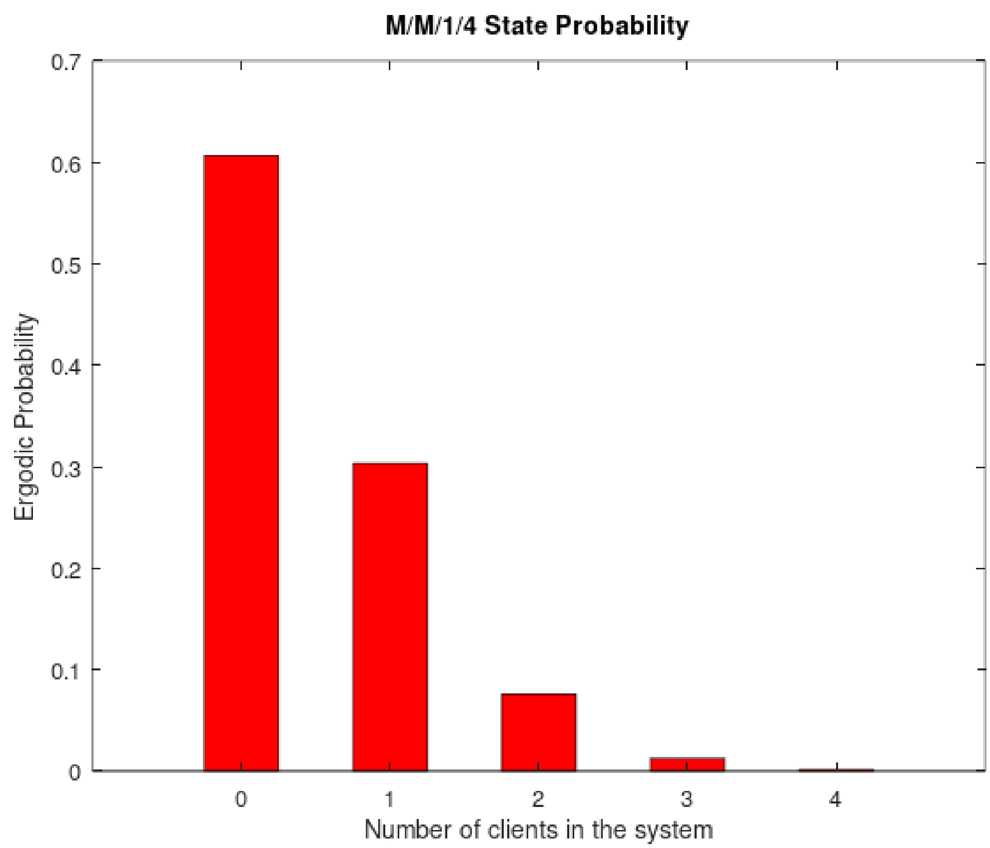
Για την εύρεση των πιθανοτήτων της Εικόνας 4 αλλά και την εκτύπωση της Γραφικής 5 χρησιμοποιήσαμε τον Κώδικα 6.

```
%(ii)
% get the ergodic probabilities of the system
P = ctmc(transition_matrix);

% plot the ergodic probabilities (bar for bar chart)
figure(1);
bar(states, P, "r", 0.5);
title("M/M/1/4 State Probability");
xlabel("Number of clients in the system");
ylabel("Ergodic Probability");

```

Κώδικας 6



Γραφική 5

Παρατηρούμε ότι τα παραπάνω αποτελέσματά μας συμβαδίζουν με εκείνα του (α) ερωτήματος.

(iii)

Με τον Κώδικα 7 εμφανίζεται στην έξοδο του τερματικού στο Octave ο μέσος αριθμός πελατών στο σύστημα, όταν εκείνο βρίσκεται σε κατάσταση ισορροπίας.

```
%(iii)
% mean number of clients in the system
mean_clients = 0;
for i = 1 : 1 : 5
  mean_clients = mean_clients + P(i)*(i-1);
endfor
display(mean_clients);
```

Κώδικας 7

```
mean_clients = 0.4992
```

Εικόνα 5

(iv)

Η πιθανότητα απόρριψης πελάτη (blocking probability) από το σύστημα, όταν εκείνο βρίσκεται σε κατάσταση ισορροπίας προκύπτει από τον Κώδικα 8 ο οποίος πρακτικά εμφανίζει στην οθόνη την τιμή της 5ης στήλης του διανύσματος P που αντιστοιχεί στην πιθανότητα άφιξης στην κατάσταση 4 του συστήματος.

```
%(iv)
display("Blocking probability = ");
display(P(5));
```

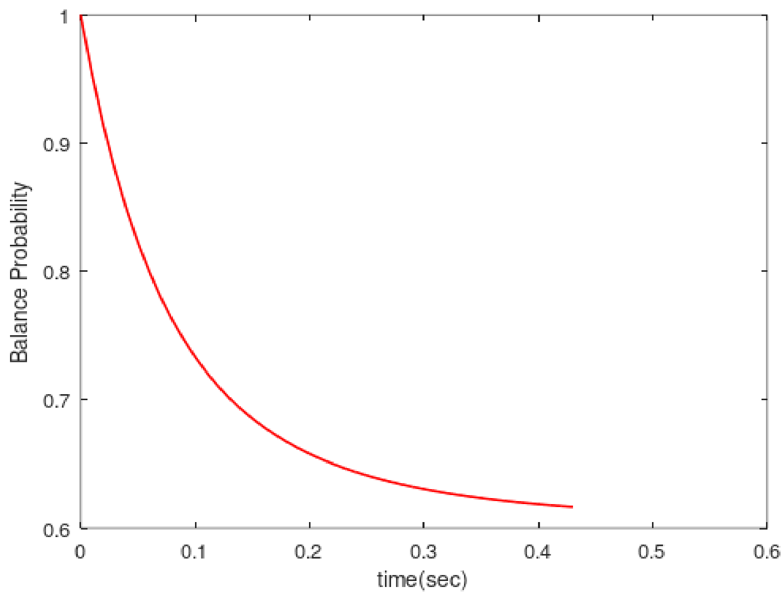
Κώδικας 8

```
Blocking probability =
1.5798e-03
```

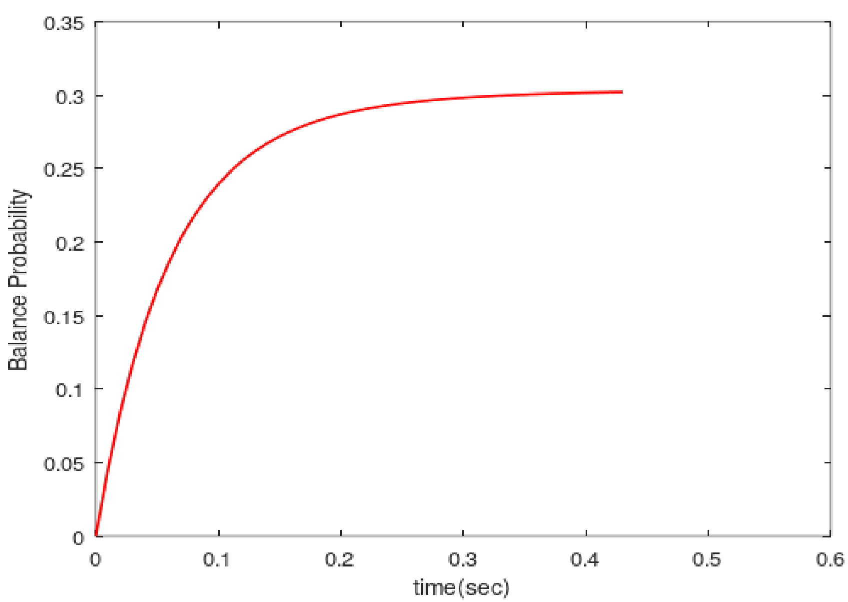
Εικόνα 6

(v)

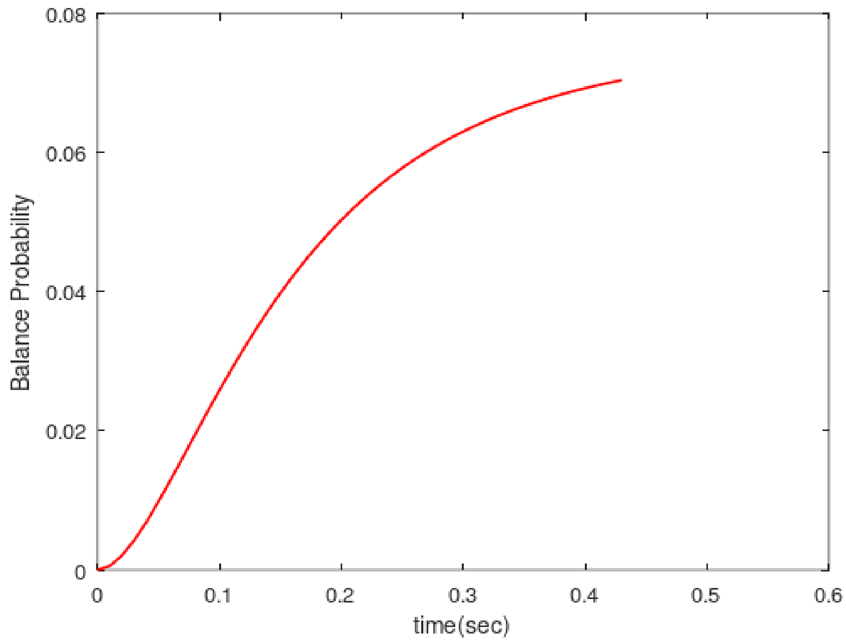
Ακολουθούν οι γραφικές, για κάθε μία από τις καταστάσεις του συστήματος, των πιθανοτήτων των καταστάσεων του συστήματος σαν συναρτήσεις του χρόνου από την αρχική κατάσταση μέχρι οι πιθανότητες να έχουν απόσταση μικρότερη του 1% από τις εργοδικές πιθανότητες του ερωτήματος (ii).



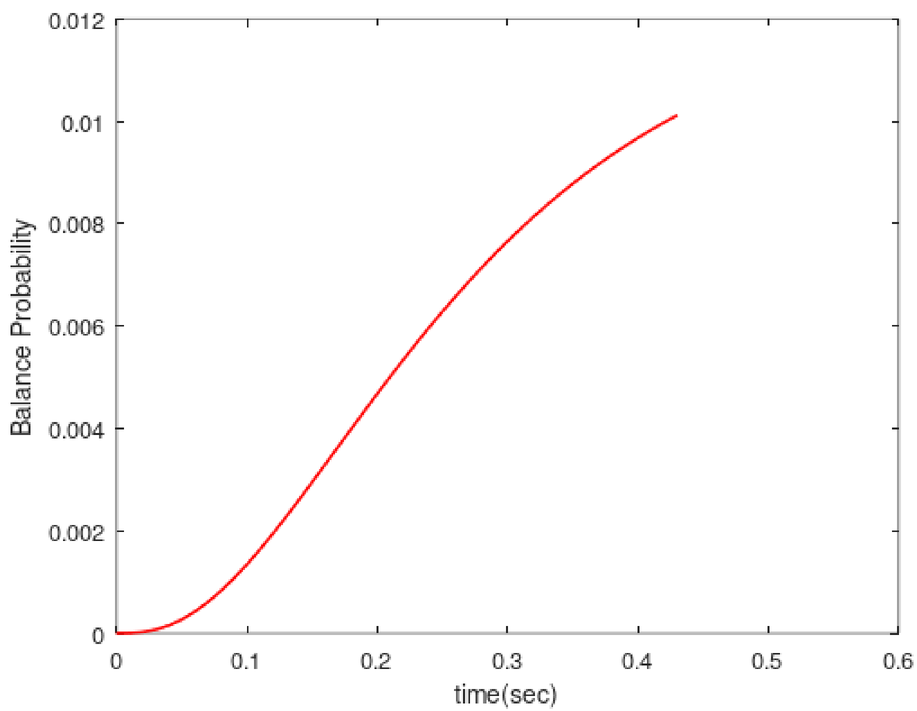
Γραφική 6 : Πιθανότητα της κατάστασης 0



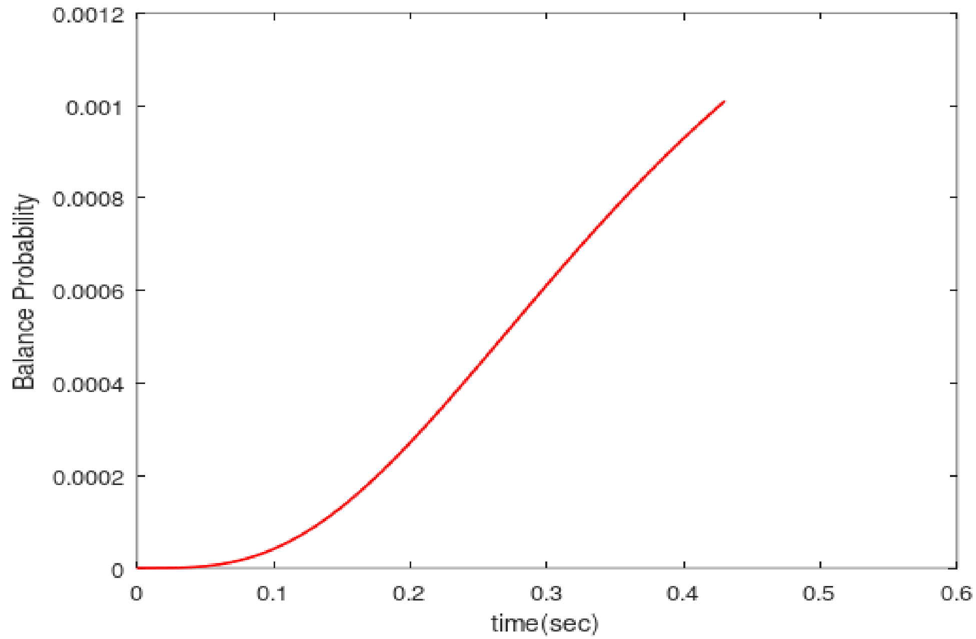
Γραφική 7 : Πιθανότητα της κατάστασης 1



Γραφική 8 : Πιθανότητα της κατάστασης 2



Γραφική 9 : Πιθανότητα της κατάστασης 3



Γραφική 10 : Πιθανότητα της κατάστασης 4

Για την εκτύπωση των γραφικών αυτών χρησιμοποιήσαμε τον παρακάτω Κώδικα

```
%(v)
% transient probability of state 0 until convergence to ergodi
for j = 1:1:5
  index = 0;
  for T = 0 : 0.01 : 50
    index = index + 1;
    P0 = ctmc(transition_matrix, T, initial_state);
    Prob0(index) = P0(j);
    if P0 - P < 0.01
     break;
    endif
  endfor

  t = 0 : 0.01 : T;
  figure(j+1);
  plot(t, Prob0, "r", "linewidth", 1.3);
  xlabel("time(sec)");
  ylabel("Balance Probability");
endfor
```

Κώδικας 9