Κεφάλαιο 6: Προσομοίωση ενός συστήματος αναμονής

Τεχνικές Εκτίμησης Υπολογιστικών συστημάτων

Γιάννης Γαροφαλάκης

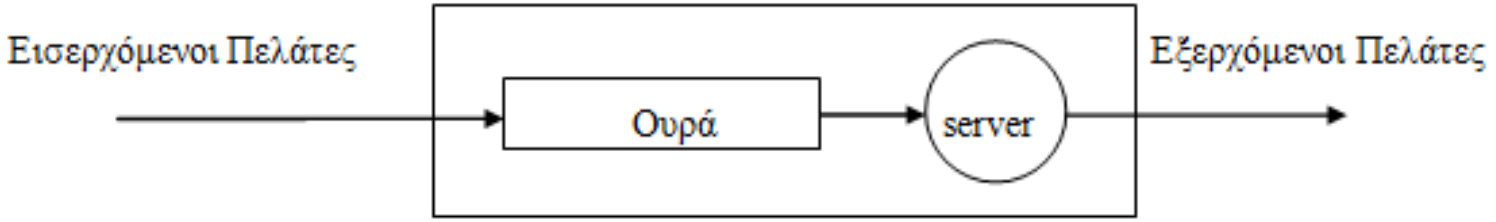
Καθηγητής

Διατύπωση του προβλήματος (1)

- Τα συστήματα αναμονής (queueing systems), βρίσκονται πίσω από τα περισσότερα μοντέλα μελέτης της απόδοσης υπολογιστικών συστημάτων.
  - Φαινόμενα καθυστερήσεων λόγω της απαίτησης χρήσης περιορισμένων πόρων από πολλούς “πελάτες”
    - σε απλούς υπολογιστές
    - σε πολύπλοκα συστήματα

Διατύπωση του προβλήματος (2)

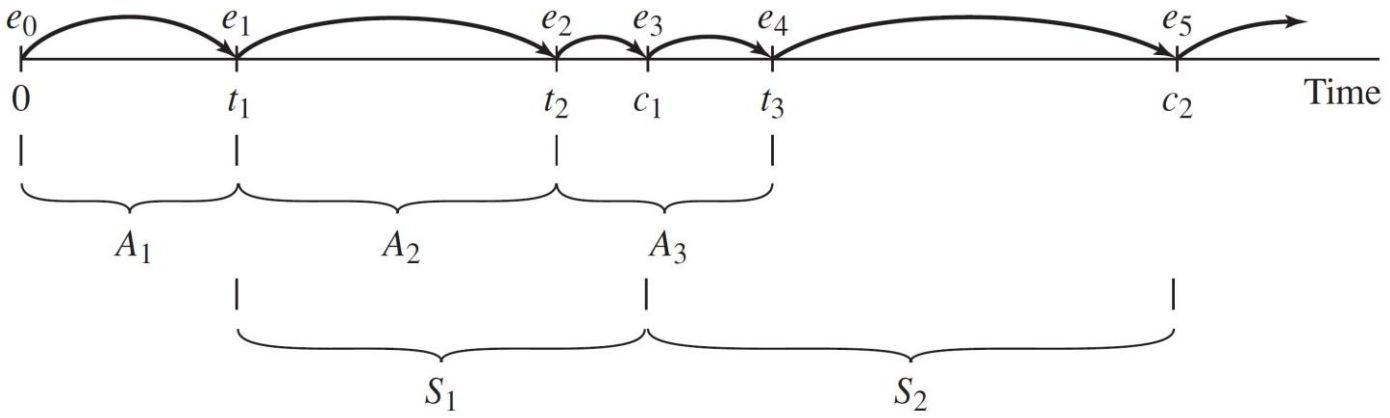
- Έστω το παρακάτω σύστημα:



- *Χρόνοι μεταξύ διαδοχικών αφίξεων* $A_1, A_2, \ldots$
  - Τυχαίες μεταβλητές
  - Ανεξάρτητες
  - Όμοια κατανεμημένες

Διατύπωση του προβλήματος (3)

- 'Ενας πελάτης που φθάνει και βρίσκει τον εξυπηρετητή (server) άδειο, αρχίζει αμέσως την εξυπηρέτησή του, ενώ εάν τον βρει κατειλλημένο μένει στο τέλος της ουράς αναμονής.
- 'Οταν ο server ελευθερωθεί, παίρνει τον πρώτο πελάτη της ουράς (αν υπάρχει κάποιος), δηλαδή έχουμε FIFO πολιτική εξυπηρέτησης.
- Οι *χρόνοι εξυπηρέτησης των πελατών* $S_1, S_2, \ldots$είναι επίσης ανεξάρτητες, όμοια κατανεμημένες τυχαίες μεταβλητές.



Διατύπωση του προβλήματος (4)

- Η προσομοίωση θα αρχίσει από τη "μηδενική" κατάσταση του συστήματος, δηλαδή με άδειο σύστημα.
  - Τη στιγμή 0, θα αρχίσουμε να περιμένουμε την άφιξη του πρώτου πελάτη, η οποία θα γίνει μετά από χρόνο A_1.
- Θέλουμε να προσομοιώσουμε το σύστημα μέχρις ότου ένας συγκεκριμένος αριθμός n πελατών περάσει από **την ουρά**.
  - Η προσομοίωση θα σταματήσει όταν ο n -στός πελάτης θα εισέλθει στον εξυπηρετητή.

Διατύπωση του προβλήματος (5)

- Μέτρηση απόδοσης συστήματος
  - Αναμενόμενη *Μέση καθυστέρηση στην ουρά* $d(n)$
    - Το $d(n)$ θα πρέπει κανονικά να βρίσκεται ως η μέση τιμή πολλών (πρακτικά άπειρων) μέσων καθυστερήσεων πελατών.
  - Από *μία* εκτέλεση του προσομοιωτή, στην οποία παρατηρούνται καθυστερήσεις στην ουρά $D_1, D_2, \ldots, D_n$ των n πελατών, μία προφανής εκτίμηση του είναι:

$$\hat{d}(n) = \frac{\sum_{i=1}^{n} D_i}{n}$$

Διατύπωση του προβλήματος (6)

- Η έννοια της καθυστέρησης, περιλαμβάνει φυσικά και την περίπτωση ένας πελάτης να μη χρειασθεί να περιμένει.
  - Ο πρώτος πελάτης θα έχει οπωσδήποτε $D_1 = 0$
- $\hat{d}(n)$
  - Δεν είναι ο μαθηματικός μέσος όρος μιας τυχαίας μεταβλητής, αφού δεν έχουμε τυχαίες παρατηρήσεις της ίδιας τυχαίας μεταβλητής.
  - Είναι μια εκτίμηση που βασίζεται σε ένα "δείγμα" μεγέθους 1, αφού βασίζεται μόνο σε μία εκτέλεση του προσομοιωτή.

Διατύπωση του προβλήματος (7)

- *Μέσος αριθμός πελατών στην ουρά* $q(n)$

  Υπολογίζεται πάνω στο (συνεχή) χρόνο.

  - $Q(t)$
    - αριθμός των πελατών στην ουρά τη χρονική στιγμή t για κάθε πραγματικό αριθμό $t \geq 0$
  - $T(n)$
    - ο χρόνος που απαιτείται για να παρατηρήσουμε τις n καθυστερήσεις στην ουρά.
  - Τότε, για κάθε χρόνο t μεταξύ 0 και $T(n)$, το $Q(t)$ είναι ένας μη-αρνητικός ακέραιος.

Διατύπωση του προβλήματος (8)

- p_i είναι το ποσοστό (με τιμές μεταξύ 0 και 1) του χρόνου που το $Q(t)$ είναι ίσο με i, τότε το $q(n)$ ορίζεται ως:

$$q(n) = \sum_{i=0}^{\infty} i p_i$$

- η εκτίμηση του $q(n)$ από μία εκτέλεση του προσομοιωτή, δίνεται από τις εκτιμήσεις των p_i δηλαδή:

$$\hat{q}(n) = \sum_{i=0}^{\infty} i \hat{p}_i \qquad (1)$$

Διατύπωση του προβλήματος (9)

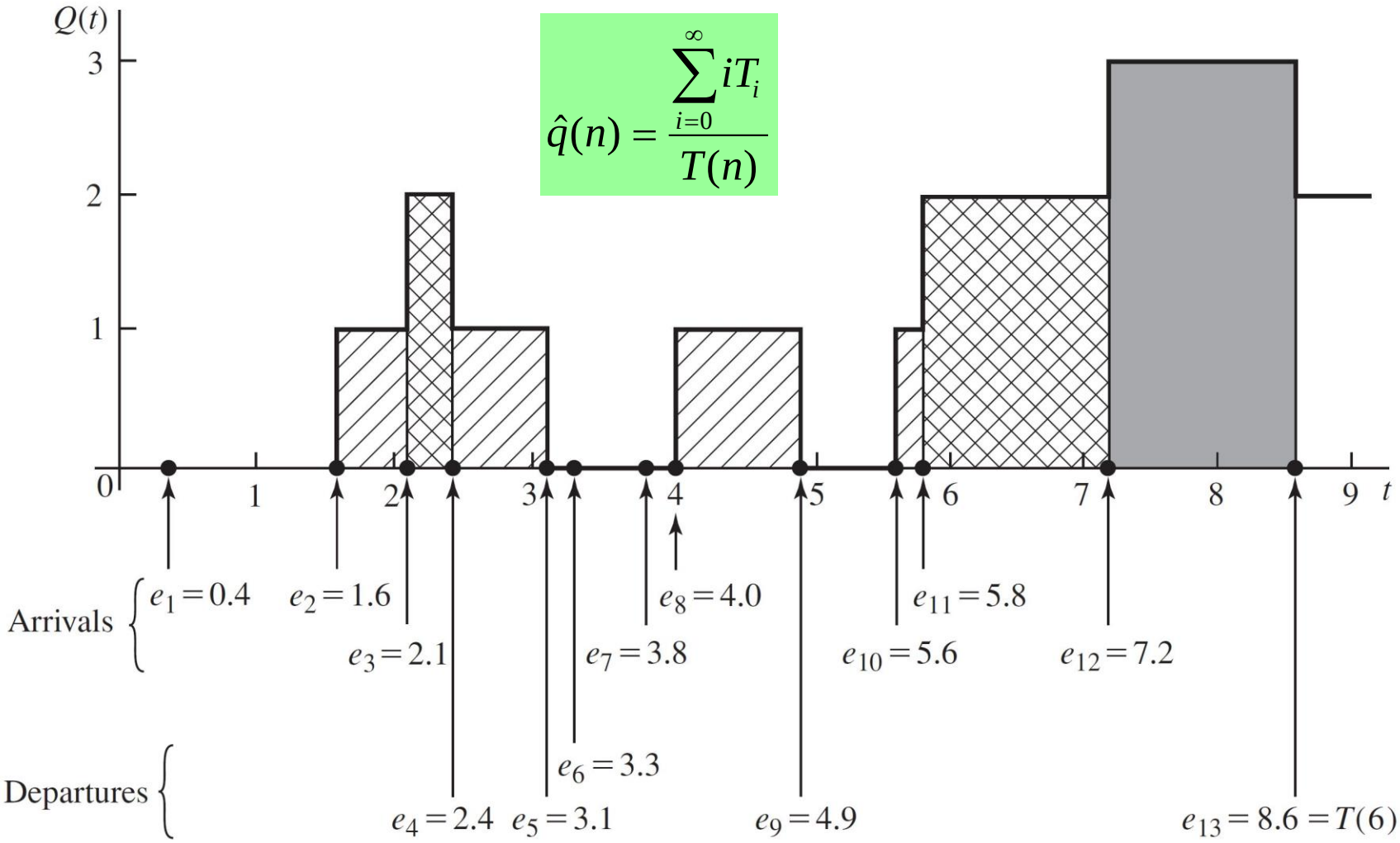
- $\hat{p}_i$ είναι τα *παρατηρούμενα* ποσοστά του χρόνου, κατά τη διάρκεια της προσομοίωσης, που υπάρχουνi πελάτες στην ουρά.
- Αν T_i είναι ο συνολικός χρόνος προσομοίωσης κατά τον οποίο η ουρά έχει μήκος i ,τότε

  $\hat{p}_i = T_i / T(n)$ όπου $T(n) = T_0 + T_1 + T_2 + \cdots$ και η (1) μπορεί να γραφεί ως:

$$\hat{q}(n) = \frac{\sum_{i=0}^{\infty} iT_i}{T(n)}$$

Διατύπωση του προβλήματος (10)

$$\hat{q}(n) = \frac{\sum_{i=0}^{\infty} iT_i}{T(n)}$$



Διατύπωση του προβλήματος (11)

- Το άθροισμα στον αριθμητή είναι η επιφάνεια κάτω από την "καμπύλη" του , μεταξύ της αρχής και του τέλους της προσομοίωσης, δηλαδή είναι το ολοκλήρωμα:

$$\sum_{i=0}^{\infty} iT_i = \int_0^{T(n)} Q(t)dt$$

οπότε η εκτίμηση του $q(n)$ δίνεται από τη σχέση:

$$\hat{q}(n) = \frac{\int_0^{T(n)} Q(t)dt}{T(n)}$$

Διατύπωση του προβλήματος (12)

- Η “αναμενόμενη” *Χρησιμοποίηση* $u(n)$ του εξυπηρετητή , είναι το μέσο ποσοστό (με τιμές μεταξύ 0 και 1) του χρόνου προσομοίωσης [από τη στιγμή 0 έως τη στιγμή $T(n)$], που ο εξυπηρετητής είναι απασχολημένος.
- Η εκτίμηση της χρησιμοποίησης , μπορεί να υπολογισθεί απ΄ ευθείας από την προσομοίωση, παρατηρώντας τις στιγμές κατά τις οποίες ο εξυπηρετητής αλλάζει κατάσταση (από άεργος σε απασχολημένος και αντίστροφα) και εκτελώντας τις κατάλληλες αφαιρέσεις και διαιρέσεις.

Διατύπωση του προβλήματος (13)

- Συνάρτηση απασχόλησης

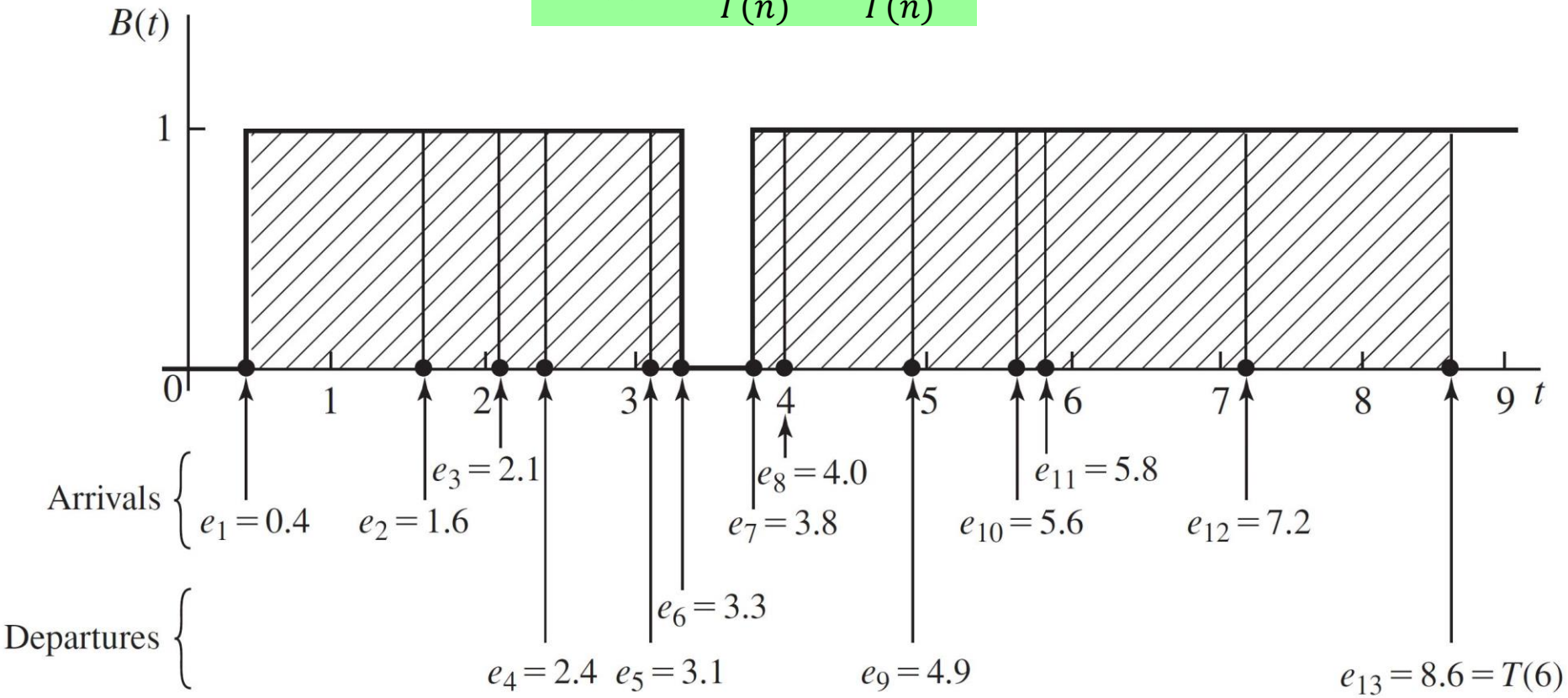
$$B(t) = \begin{cases} 1 & \text{αν ο εξυπηρετητής είναι απασχολημένος τη στιγμή } t \\ 0 & \text{αν ο εξυπηρετητής είναι αεργος τη στιγμή } t \end{cases}$$

- Το $\hat{u}(n)$ μπορεί να εκφραστεί ως το ποσοστό του χρόνου που το $B(t)$ είναι ίσο με 1.
- Αφού στο διάγραμμα του $B(t)$, το ύψος της γραφικής παράστασής του είναι πάντοτε είτε 0 ή 1, το θα μπορεί να υπολογισθεί (κατά αναλογία προς τον τρόπο υπολογισμού του $q(n)$ παραπάνω), από τη σχέση

$$\hat{u}(n) = \frac{\sum_{i=0}^{1} i\, T_i}{T(n)} = \frac{T_1}{T(n)} \quad \text{ή} \quad \hat{u}(n) = \frac{\int_0^{T(n)} B(t)dt}{T(n)}$$

Διατύπωση του προβλήματος (14)

$$\hat{u}(n) = \frac{\sum_{i=0}^{1} i\, T_i}{T(n)} = \frac{T_1}{T(n)}$$



Διατύπωση του προβλήματος (15)

- Στο σύστημα αυτό, τα γεγονότα είναι η **άφιξη** και η **αναχώρηση** ενός πελάτη.
- Οι μεταβλητές συστήματος που μας χρειάζονται για τις εκτιμήσεις των $u(n)$, $d(n)$ και $q(n)$ είναι
  - Η κατάσταση του εξυπηρετητή (0 αν είναι άεργος και 1 αν είναι απασχολημένος).
  - Ο αριθμός των πελατών στην ουρά, η χρονική στιγμή άφιξης κάθε πελάτη που βρίσκεται στην ουρά (μία λίστα) .
  - Η χρονική στιγμή εμφάνισης του πιο πρόσφατου γεγονότος.
- Η στιγμή εμφάνισης του πιο πρόσφατου γεγονότος, η οποία ορίζεται ως e_{i-1} εάν $e_{i-1} \le t < e_i$ (όπου t είναι ο παρών χρόνος της προσομοίωσης), μας χρειάζεται για τον υπολογισμό του πλάτους των ορθογωνίων παραλληλογράμμων, που χρησιμοποιούνται στις εκτιμήσεις των $q(n)$ και $u(n)$.

Η εκτέλεση του προσομοιωτή (1)

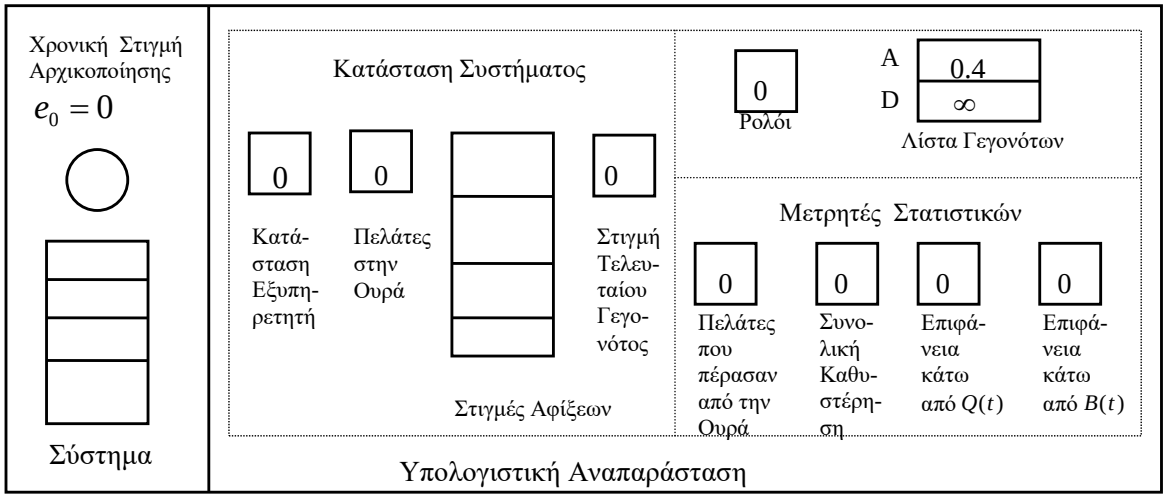
- Υπολογιστική αναπαράσταση μοντέλου προσομοίωσης
  - ΠΑΡΑΔΕΙΓΜΑ: 16 διαδοχικά γεγονότα
    - χρονική στιγμή $e_0 = 0$ και οι στιγμές $e_1, e_2, \cdots, e_{15}$
    - $n = 6$ διελεύσεις πελατών από την ουρά.
    - οι χρόνοι μεταξύ διαδοχικών αφίξεων και εξυπηρέτησης των πελατών είναι

$$A_1 = 0.4, A_2 = 1.2, A_3 = 0.5, A_4 = 1.7, A_5 = 0.2, A_6 = 1.6, A_7 = 0.2, A_8 = 1.4, A_9 = 1.9, \ldots$$

$$S_1 = 2.0, S_2 = 0.7, S_3 = 0.2, S_4 = 1.1, S_5 = 3.7, S_6 = 0.6, \ldots$$

Η εκτέλεση του προσομοιωτή (2)

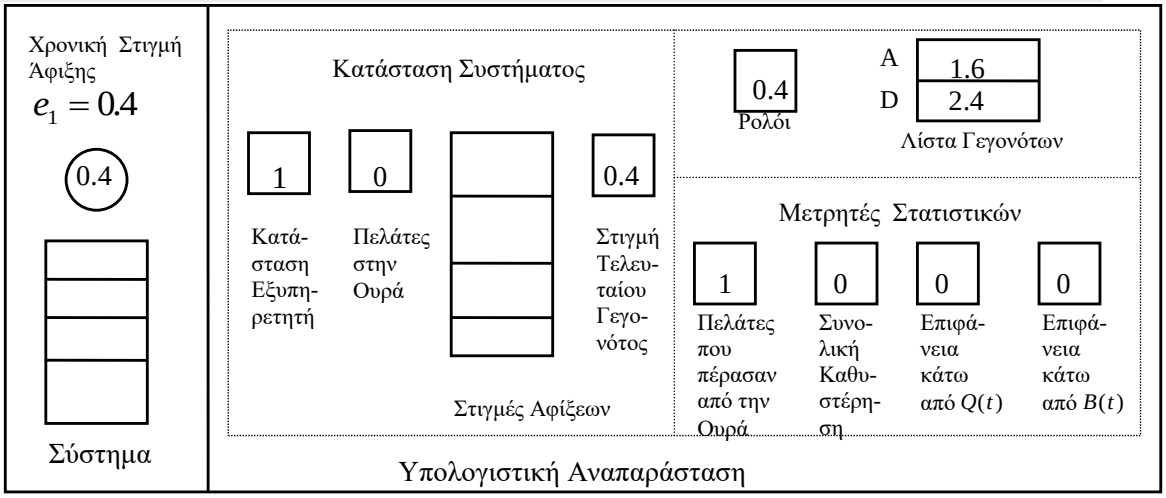
- Το βασικό στοιχείο της δυναμικής εξέλιξης του προσομοιωτή είναι η αλληλεπίδραση του ρολογιού με τη λίστα γεγονότων.
- Κατά τη διάρκεια της επεξεργασίας ενός γεγονότος, δεν εξελίσσεται ο "προσομοιωμένος" χρόνος.
  - πρέπει να ενημερώνονται οι μεταβλητές κατάστασης και οι μετρητές στατιστικών.



(a)

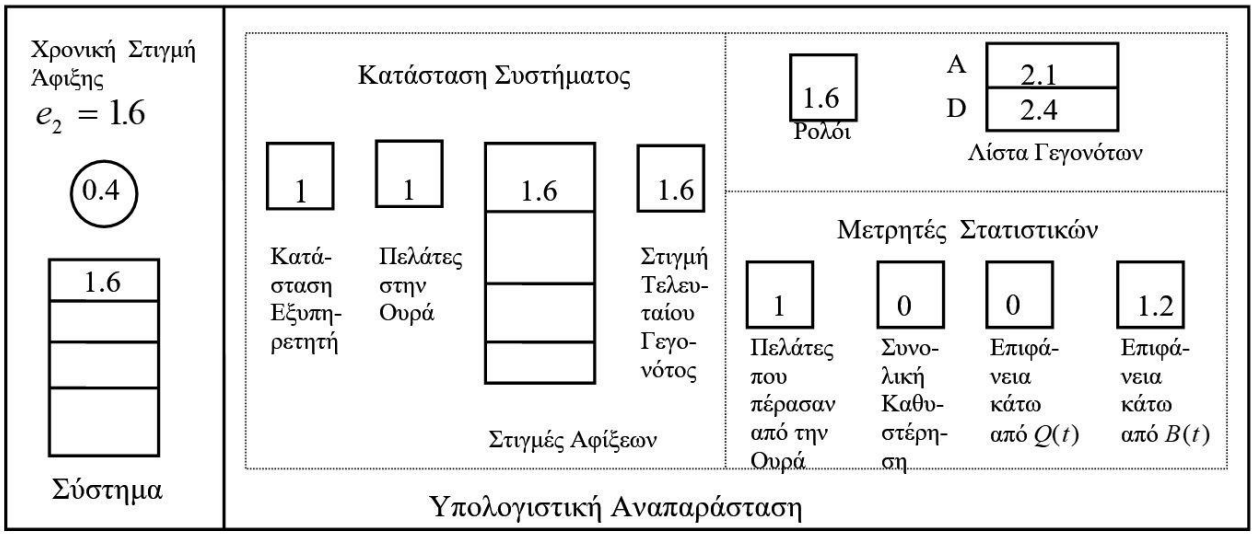
$$A_1 = 0.4, A_2 = 1.2, A_3 = 0.5, A_4 = 1.7, A_5 = 0.2, A_6 = 1.6, A_7 = 0.2, A_8 = 1.4, A_9 = 1.9, \ldots$$

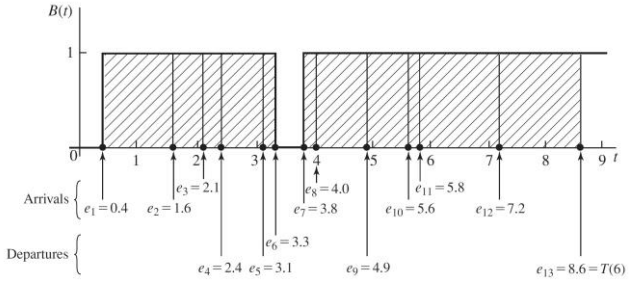
$$S_1 = 2.0, S_2 = 0.7, S_3 = 0.2, S_4 = 1.1, S_5 = 3.7, S_6 = 0.6, \ldots$$



(b)

Σχήμα 6.1. Το Σύστημα και η Υπολογιστική Αναπαράστασή του κατά τις χρονικές στιγμές εμφάνισης των γεγονότων.

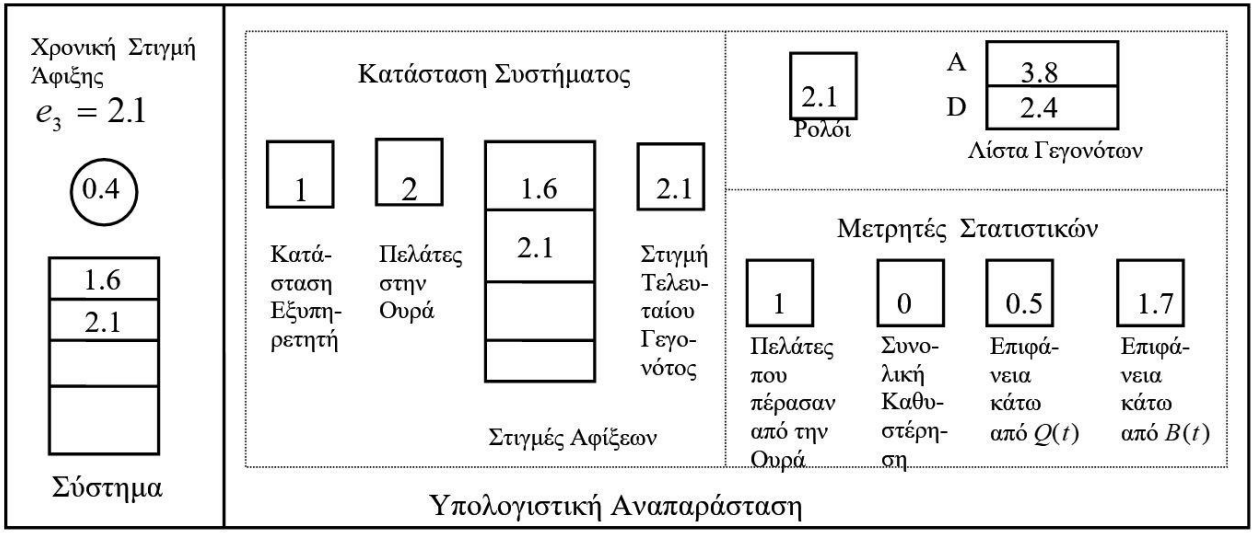


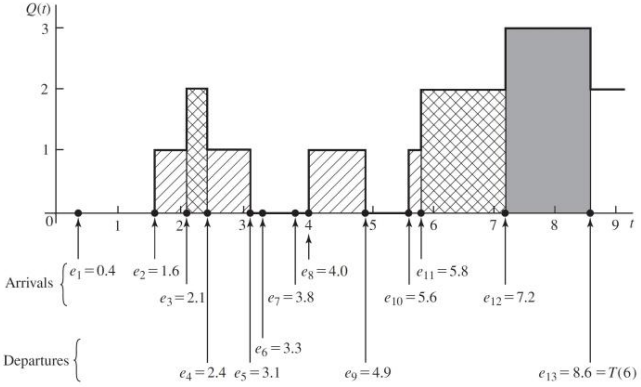


(c)

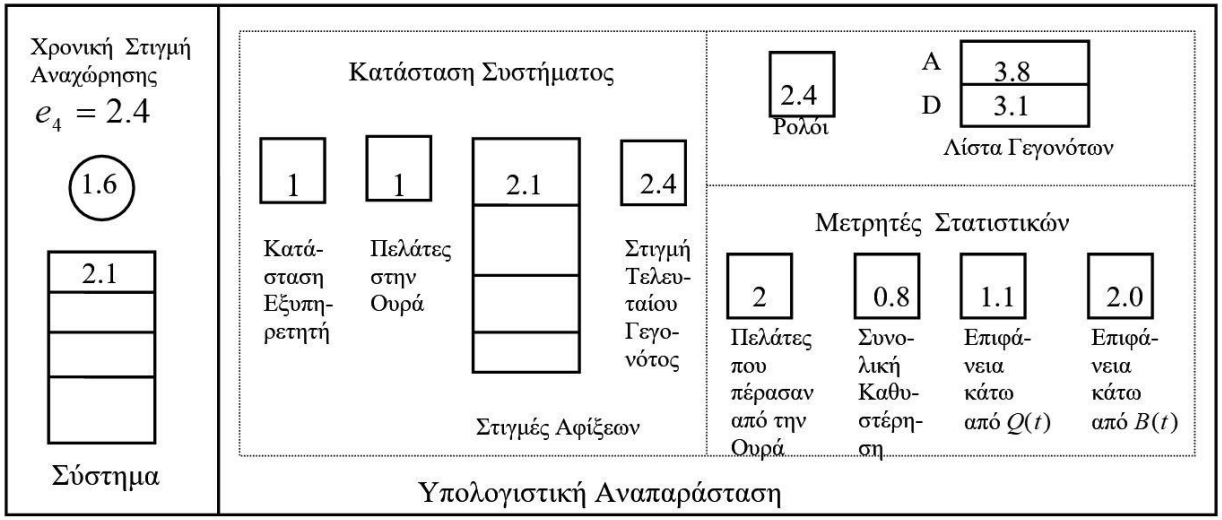
$$A_1 = 0.4, A_2 = 1.2, A_3 = 0.5, A_4 = 1.7, A_5 = 0.2, A_6 = 1.6, A_7 = 0.2, A_8 = 1.4, A_9 = 1.9, \ldots$$

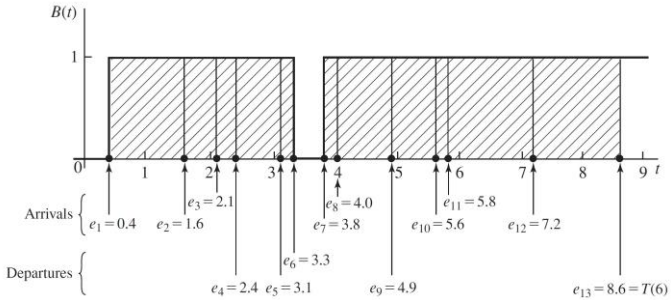
$$S_1 = 2.0, S_2 = 0.7, S_3 = 0.2, S_4 = 1.1, S_5 = 3.7, S_6 = 0.6, \ldots$$





(d)

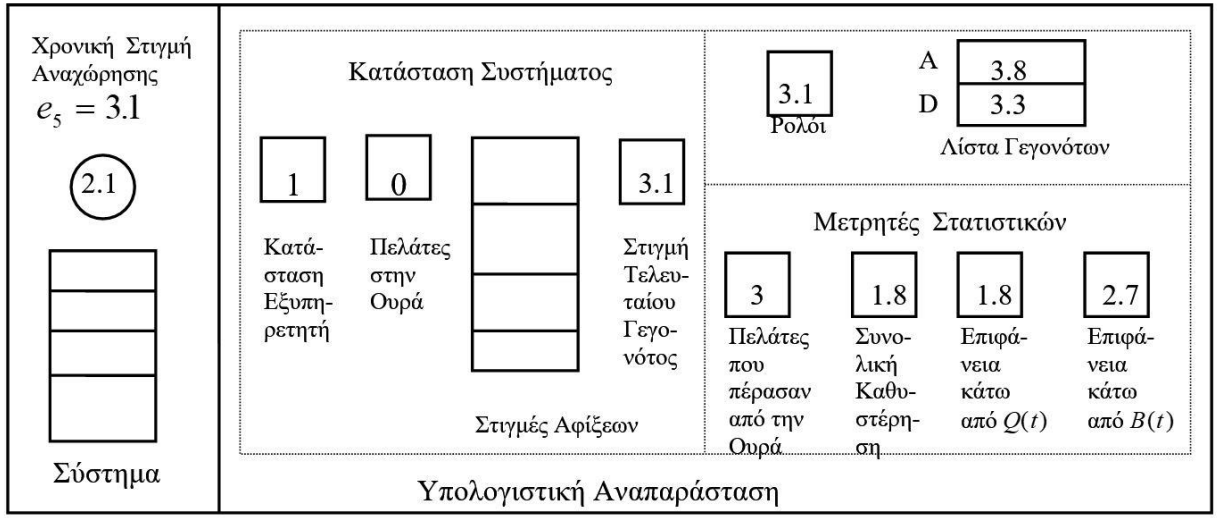


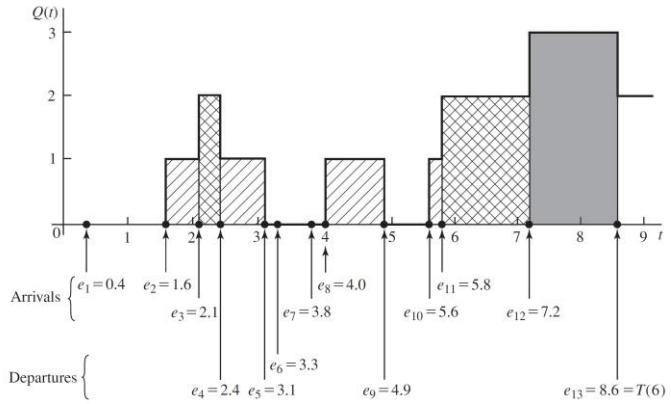


(e)

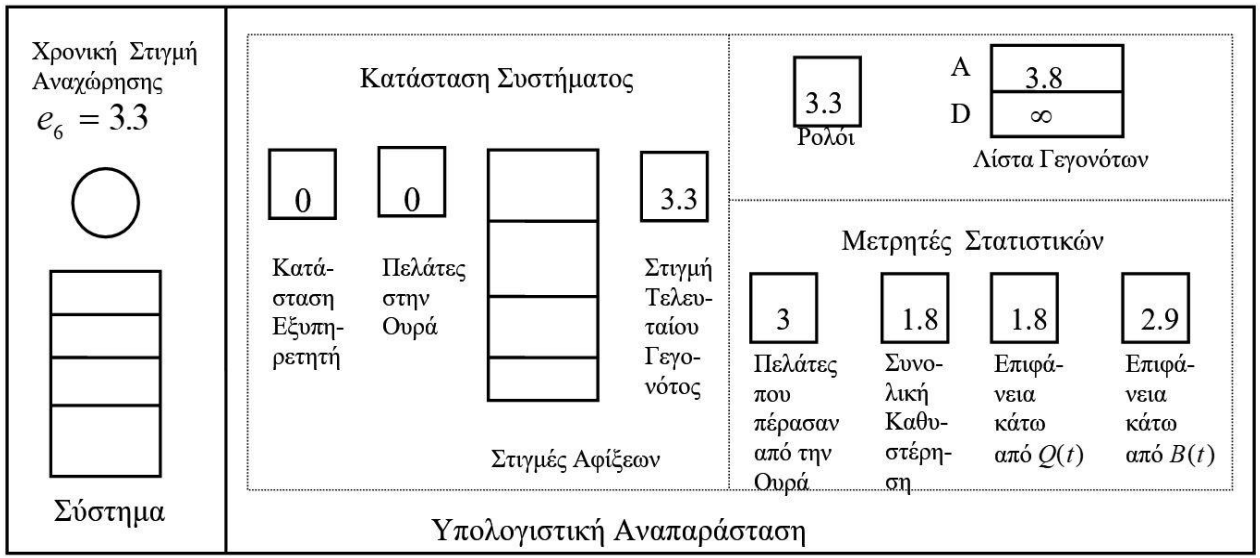
$A_1 = 0.4, A_2 = 1.2, A_3 = 0.5, A_4 = 1.7, A_5 = 0.2, A_6 = 1.6, A_7 = 0.2, A_8 = 1.4, A_9 = 1.9, \ldots$

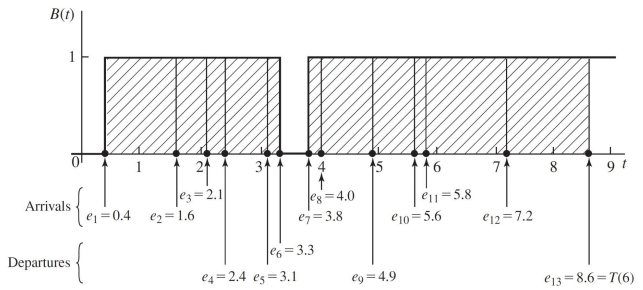
$S_1 = 2.0, S_2 = 0.7, S_3 = 0.2, S_4 = 1.1, S_5 = 3.7, S_6 = 0.6, \ldots$





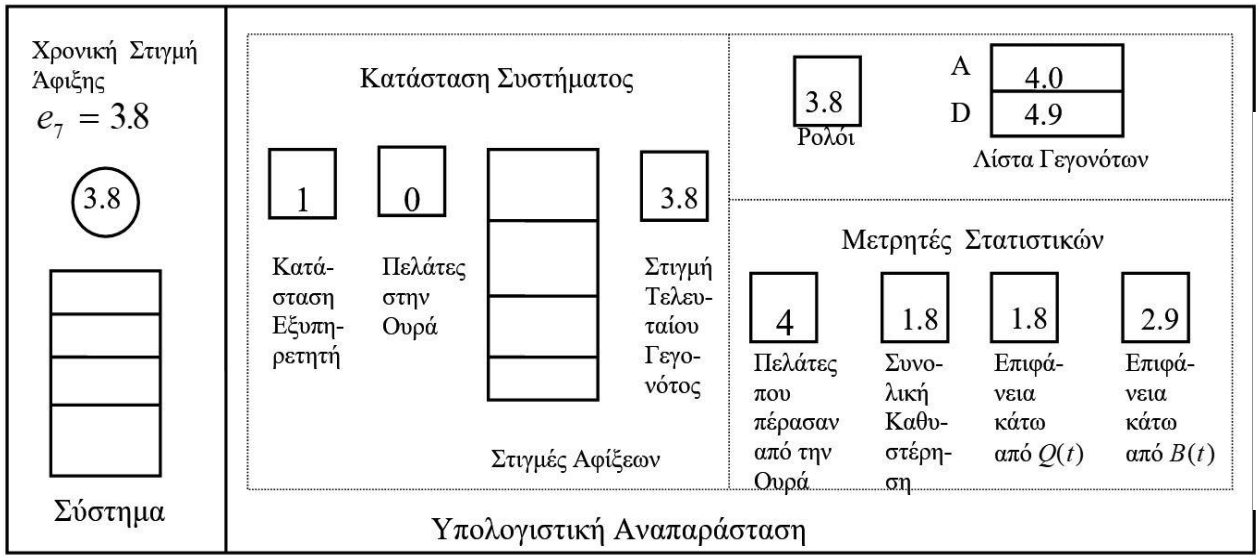
(f)

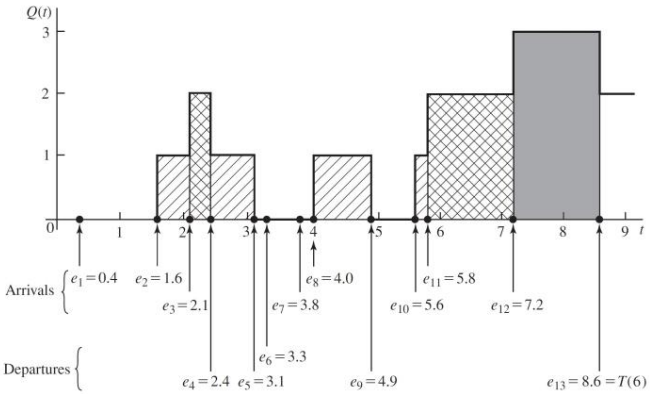




$$A_1 = 0.4, A_2 = 1.2, A_3 = 0.5, A_4 = 1.7, A_5 = 0.2, A_6 = 1.6, A_7 = 0.2, A_8 = 1.4, A_9 = 1.9, \ldots$$

$$S_1 = 2.0, S_2 = 0.7, S_3 = 0.2, S_4 = 1.1, S_5 = 3.7, S_6 = 0.6, \ldots$$





(h)

Η εκτέλεση του προσομοιωτή (3)

- Στο τέλος, από τους «μετρητές στατιστικών» υπολογίζονται οι μετρικές απόδοσης:

$$\hat{d}(n) = \frac{\sum_{i=1}^{n} D_i}{n}$$

$$\hat{d}(n) = \frac{(\text{Συνολικη καθυστερηση})}{(\text{Πελατες που περασαν απο ουρα})}$$

$$\hat{q}(n) = \frac{\sum_{i=0}^{\infty} iT_i}{T(n)}$$

$$\hat{q}(n) = \frac{(\text{Επιφανεια κάτω από } \mathrm{Q}(t))}{T(n)}$$

$$\hat{u}(n) = \frac{T_1}{T(n)}$$

$$\hat{u}(n) = \frac{(\text{Επιφανεια κάτω από } \mathrm{B}(t))}{T(n)}$$

Η εκτέλεση του προσομοιωτή (4)

- Μερικές φορές είναι εύκολο να παραβλεφθούν οι συνέπειες γεγονότων που δεν είναι συνηθισμένα κατά τη διάρκεια της προσομοίωσης, που έχουν όμως σημαντικές επιπτώσεις:
  - Είναι πιθανό να ξεχάσουμε ότι ένας πελάτης που αναχωρεί, μπορεί να αφήσει πίσω του ένα άδειο σύστημα και κατά συνέπεια πρέπει να μείνει άεργος ο εξυπηρετητής.
  - Το γεγονός της αναχώρησης πρέπει να διαγραφεί από τη λίστα γεγονότων.

Η εκτέλεση του προσομοιωτή (5)

- Μπορεί να συμβεί δύο ή περισσότερες τιμές της λίστας γεγονότων να είναι ίδιες, οπότε πρέπει να αποφασισθεί ποιό γεγονός θα ακολουθήσει.
  - Οι κανόνες που θα χρησιμοποιούνται στις περιπτώσεις αυτές
    - επηρεάζουν τα αποτελέσματα της προσομοίωσης
    - πρέπει να επιλέγονται με βάση την επιθυμητή μοντελοποίηση του συστήματος (π.χ. αναχώρηση πριν την άφιξη)
  - Όταν τα γεγονότα περιγράφονται από συνεχείς πιθανοτικές κατανομές, η πιθανότητα εμφάνισης ενός τέτοιου γεγονότος (πρέπει να) είναι 0.
  - Αν δεν υπάρχει κανόνας από το πραγματικό σύστημα, η επιλογή μπορεί να γίνει και τυχαία.

Οι κατανομές αφίξεων και εξυπηρέτησης (1)

- Οι χρόνοι μεταξύ διαδοχικών αφίξεων και οι χρόνοι εξυπηρέτησης:
  - Ανεξάρτητες τυχαίες μεταβλητές που περιγράφονται από εκθετικές κατανομές
  - Η **εκθετική κατανομή** με μέση τιμή $\beta = 1/\lambda$ είναι συνεχής, με *συνάρτηση πυκνότητας πιθανότητας (pdf)*:

$$f(x) = \frac{1}{\beta} e^{-x/\beta}$$

  - και *συνάρτηση κατανομής πιθανότητας* (PDF):

$$F(x) = \int_0^x \frac{1}{\beta} e^{-t/\beta} dt = 1 - e^{-x/\beta}$$

Οι κατανομές αφίξεων και εξυπηρέτησης (2)

- Είναι πιο συνηθισμένο οι ποσότητες εισόδου που "οδηγούν" την προσομοίωση, να δημιουργούνται από συγκεκριμένες κατανομές, παρά να θεωρούμε ότι είναι γνωστές.
- Η επιλογή της εκθετικής κατανομής είναι ουσιαστικά αυθαίρετη, αλλά με ισχυρή βάση δικαιολόγησης (Markov…).
- Το σύστημα αυτό αναμονής με έναν εξυπηρετητή και εκθετικούς χρόνους μεταξύ αφίξεων και εξυπηρέτησης, είναι το γνωστό σύστημα αναμονής

$$M/M/1$$

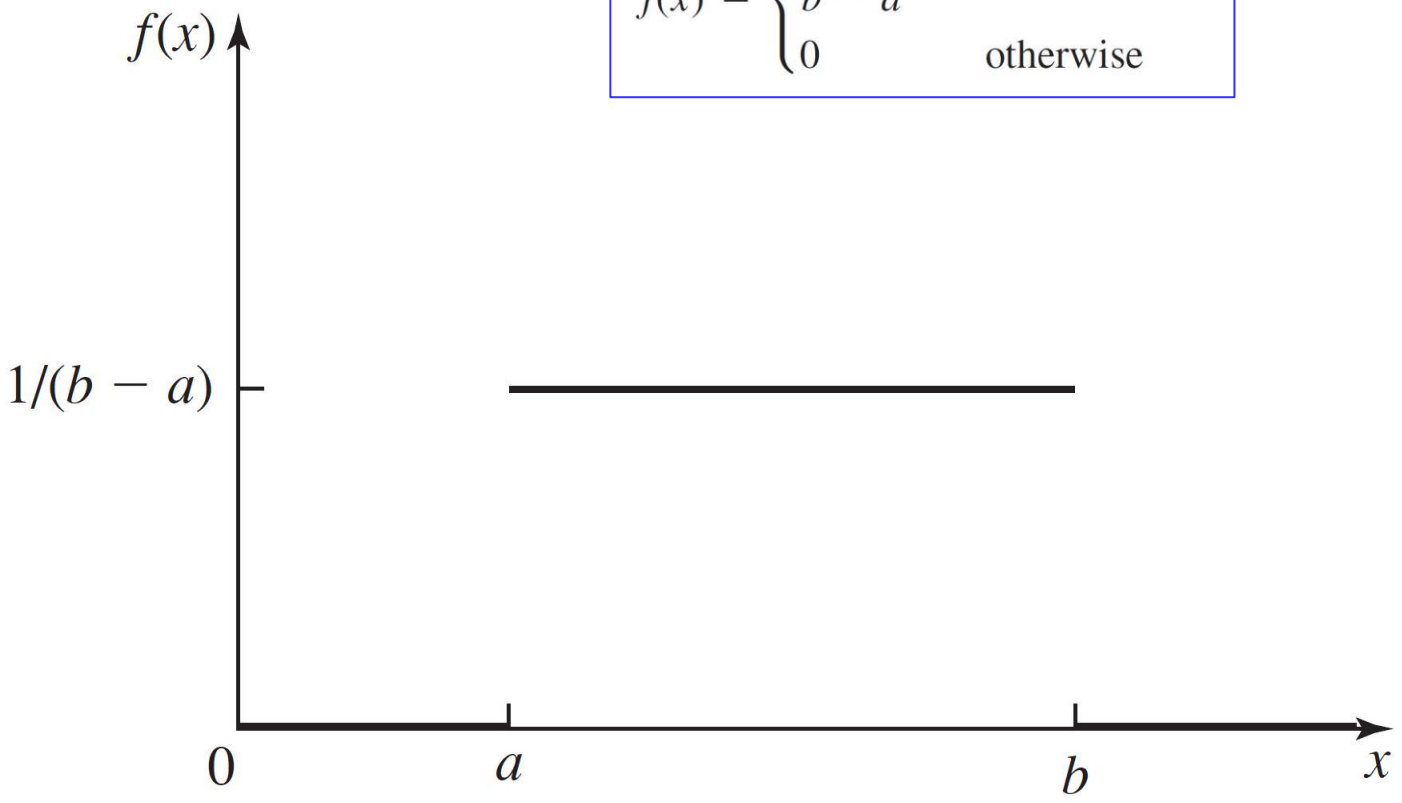
Δημιουργία τιμών κατανομής (1)

- **ΒΗΜΑ 1**
  - *Λήψη τιμής **U** από Γεννήτρια Τυχαίων Αριθμών*
    - Δημιουργία μίας «τυχαίας» τιμής *U* που είναι *ομοιόμορφα (συνεχώς) κατανεμημένη* μεταξύ 0 και 1.
    - Η κατανομή αυτή θα αναφέρεται ως $U(0,1)$ και έχει συνάρτηση πυκνότητας πιθανότητας:

$$f(x) = \begin{cases} 1 & \text{αν} \quad 0 \le x \le 1 \\ 0 & \text{αλλιως} \end{cases}$$

Η ομοιόμορφη κατανομή U(*a*, *b*)

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$



Δημιουργία τιμών κατανομής (2)

- Η πιθανότητα μία $U(0,1)$τυχαία μεταβλητή (π.χ. η X) να "πέσει" σε οποιοδήποτε υποδιάστημα

  $[x,\ x+\Delta x]$ που περιλαμβάνεται στο διάστημα [0,1], είναι (ομοιόμορφα) Δx :

  $$P(X \in [x, x+\Delta x]) = \int_x^{x+\Delta x} 1\, dy = (x+\Delta x) - x = \Delta x$$

- *Μέση τιμή*: $\frac{a+b}{2}$ για την U(a, b), οπότε:

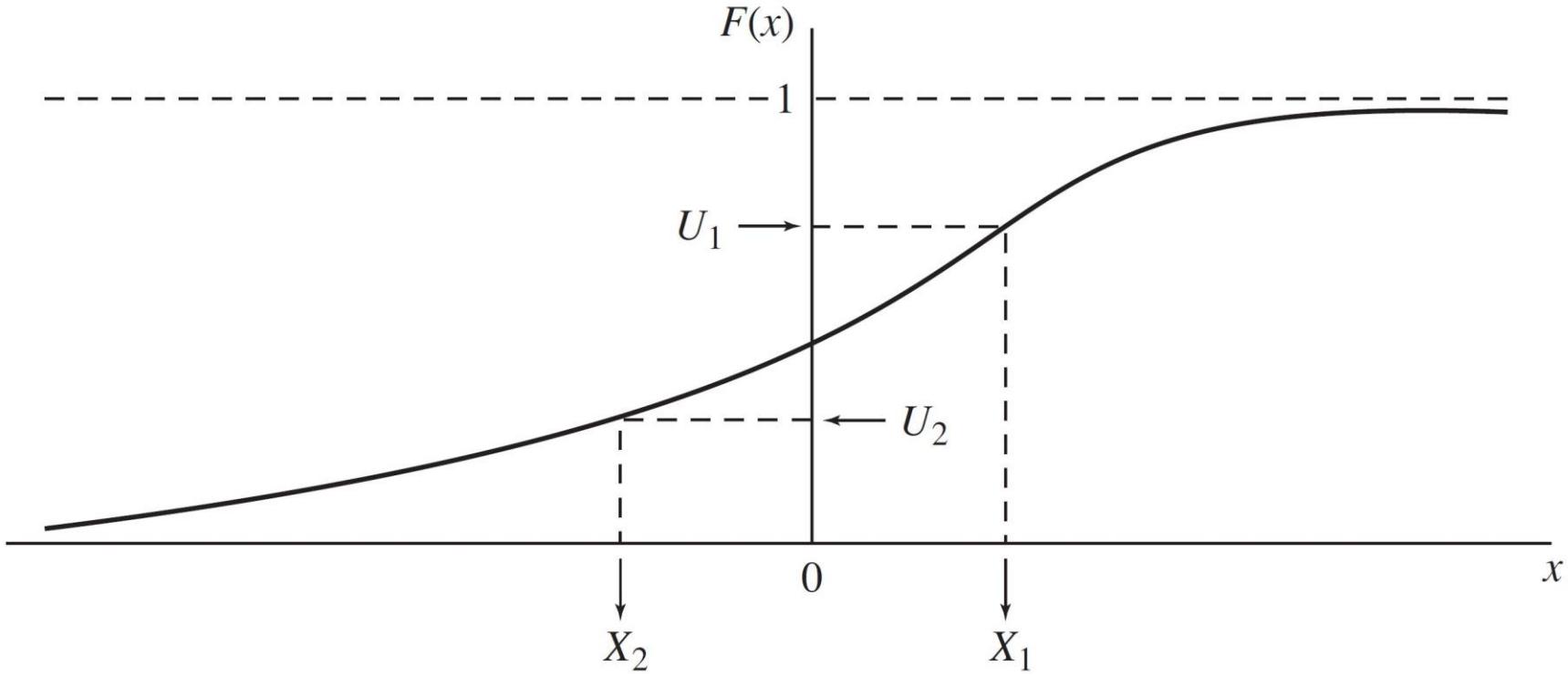
  ½ για την $U(0,1)$.

Δημιουργία τιμών κατανομής (3)

- **ΒΗΜΑ 2**
  - *Αντιστροφή της PDF της εκθετικής κατανομής με χρήση του U.*
    - Η *Συνάρτηση Κατανομής Πιθανότητας* (PDF) μιας τυχαίας μεταβλητής *X* ορίζεται για κάθε πραγματικό *x*, ως $F(x) = P(X \leq x)$ και ισχύει πάντα $0 \leq F(x) \leq 1$. Όπως και το *U*…
    - Αν μπορούμε να αντιστρέψουμε την $F(x)$, τότε λύνουμε την εξίσωση:

      $F(x) = U$ και παίρνουμε το $x = F^{-1}(U)$
    - Έτσι δημιουργούνται τιμές που προέρχονται από την $F(x)$:

Δημιουργία τιμών κατανομής (4)



Δημιουργία τιμών κατανομής (5)

- Για την *εκθετική κατανομή* με μέση τιμή β, θέτουμε:

$$U = F(x) = 1 - e^{-x/\beta} \Rightarrow$$

$$\Rightarrow e^{-x/\beta} = 1 - U \Rightarrow -x/\beta = \ln(1-U) \Rightarrow$$

$$\Rightarrow x = -\beta \ln(1-U) \quad \text{ή} \quad \boxed{x = -\beta \ln U}$$

Το τελευταίο ισχύει, διότι τόσο το ***U***, όσο και το ***1 – U***, ακολουθούν την κατανομή $U(0,1)$.

- Επιβεβαίωση ότι το $X = -\beta \ln U$ είναι μικρότερο ή ίσο με *x*, με πιθανότητα $F(x)$ όπως παραπάνω:

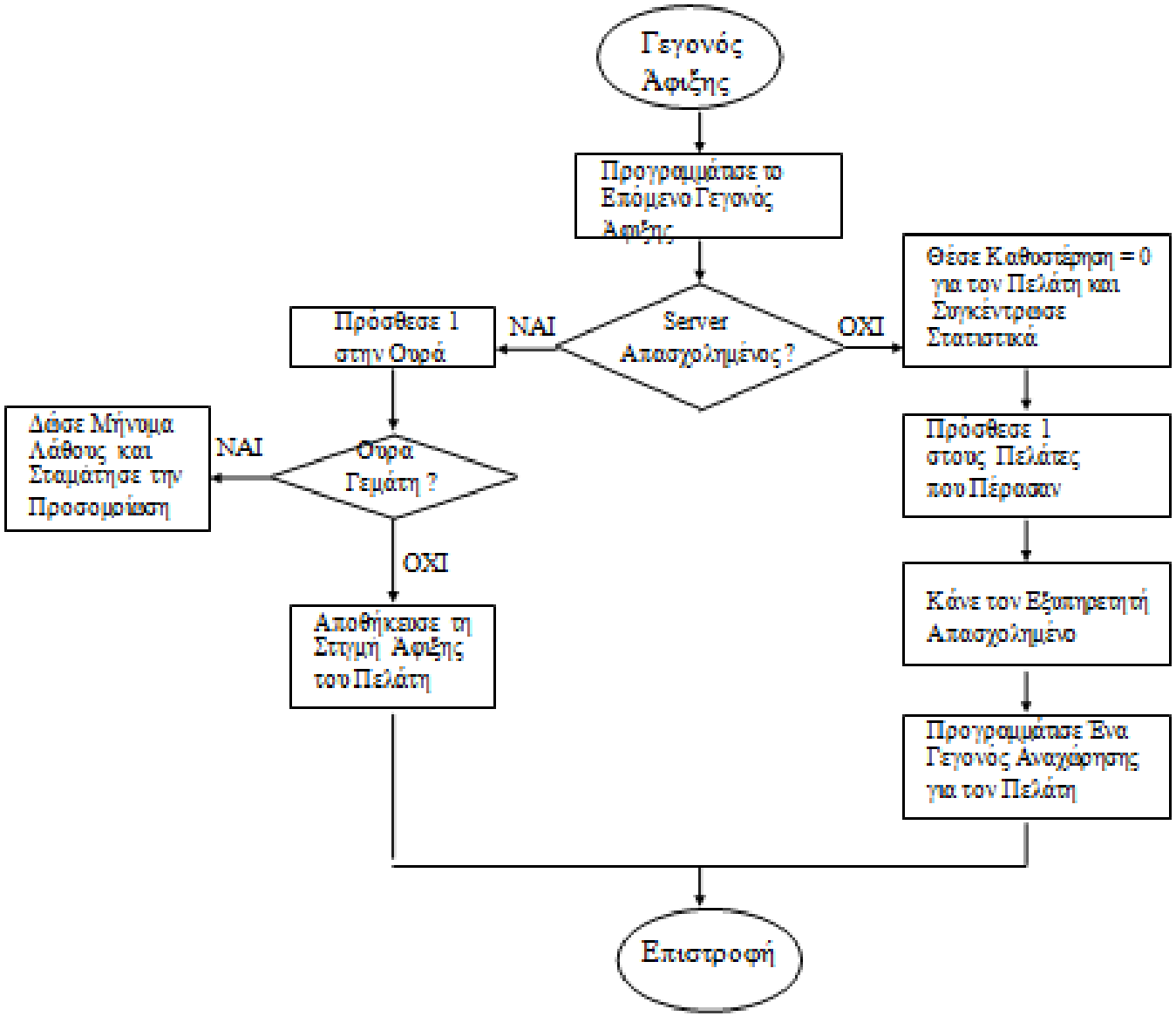
$$P(-\beta \ln U \le x) = P(\ln U \ge -\frac{x}{\beta}) = P(U \ge e^{-x/\beta}) = P(e^{-x/\beta} \le U \le 1) = 1 - e^{-x/\beta}$$

Οργάνωση και λογική του προγράμματος

- Το πρόγραμμα του προσομοιωτή πρέπει να κατασκευάζεται κατά ενότητες (modules), ώστε να ξεκαθαρίζεται η λογική και οι αλληλεπιδράσεις των τμημάτων του προγράμματος.
- Γεγονότα:

Περιγραφή Γεγονότος	*Τύπος Γεγονότος*
1. Άφιξη Πελάτη στο Σύστημα	**1**
1. Αναχώρηση Πελάτη από το Σύστημα αφού ολοκλήρωσε την εξυπηρέτησή του	**2**

Διάγραμμα ροής για το γεγονός «άφιξη»



Διάγραμμα ροής για το γεγονός «αναχώρηση»