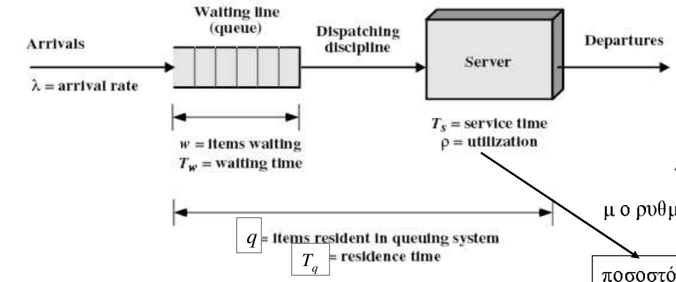
Θεωρία ουρών αναμονής

Ουρές αναμονής

- Έστω εξυπηρετητής που εξυπηρετεί ένα πακέτο ανά sec. Έστω, επίσης, ότι σε αυτόν φτάνει 1 πακέτο ανά sec. Τότε, αν η κατανομή των πακέτων είναι ομοιόμορφη δεν υπάρχει πρόβλημα.
- Αν όμως η μ.τ. των αφίξεων είναι 1 πακέτο / sec αλλά υπάρχει και μια μικρή διακύμανση, τότε μπορεί σε μια χρονική στιγμή να μην έρθει κανένα πακέτο, στην επόμενη όμως μπορεί να φτάσουν 2, 3 ή περισσότερα.
- Δημιουργείται αμέσως η ανάγκη για ύπαρξη χώρου προσωρινής ενταμίευσης (buffer) και είναι σημαντικός ο σωστός υπολογισμός του μεγέθους του χώρου αυτού.

Ουρά αναμονής με έναν εξυπηρετητή



$$\rho = \frac{\lambda}{\mu}$$

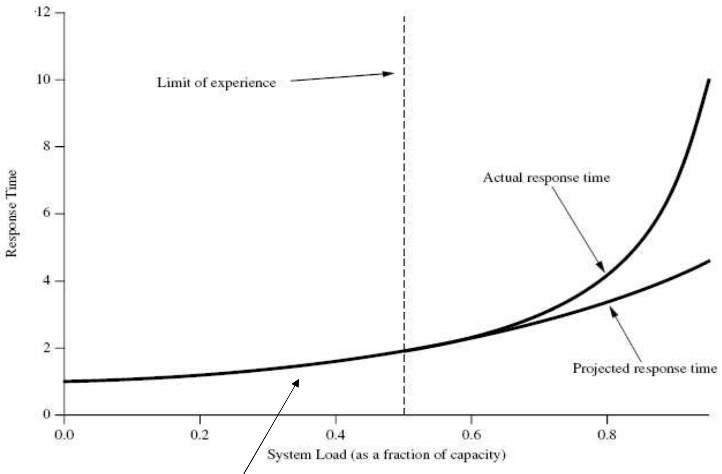
μ ο ρυθμός εξυπηρέτησης

ποσοστό χρήσης, ή η πιθανότητα ένα αντικείμενο που μόλις έφτασε να περιμένει

- Για ρ=1 το σύστημα φτάνει σε κόρο.
- Ο θεωρητικός μέγιστος ρυθμός αφίξεων που μπορούν να εξυπηρετηθούν είναι $\lambda_{\max} = \frac{1}{T_s}$
- Στην πράξη υπάρχει κόρος πολύ νωρίτερα.

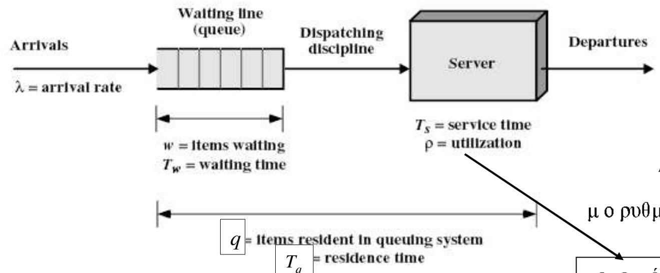
Βασικές έννοιες ουρών

- Σύστημα ουρών αναμονής
  - Ένα δίκτυο δεδομένων στο οποίο φτάνουν πακέτα, περιμένουν σε διάφορες ουρές αναμονής, εξυπηρετούνται σε διάφορα σημεία, και εξέρχονται μετά από κάποιο χρονικό διάστημα.
- Ρυθμός Αφίξεων (Arrival rate)
  - Ο αριθμός αφίξεων στη μονάδα του χρόνου (μέση τιμή)
- Βαθμός κατάληψης (Occupancy)
  - Αριθμός πακέτων μέσα στο σύστημα (μέση τιμή)
- Χρόνος παραμονής στο σύστημα (delay)
  - Ο χρόνος από την άφιξη μέχρι στην έξοδο των πακέτων (μέση τιμή για πολλά πακέτα)



Η προηγούμενη εμπειρία
δεν είναι πάντα ασφαλής μέθοδος πρόβλεψης

Ουρά αναμονής με έναν εξυπηρετητή



$$\rho = \frac{\lambda}{\mu}$$

μ ο ρυθμός εξυπηρέτησης

ποσοστό χρήσης, ή
η πιθανότητα ένα
αντικείμενο που μόλις
έφτασε να περιμένει

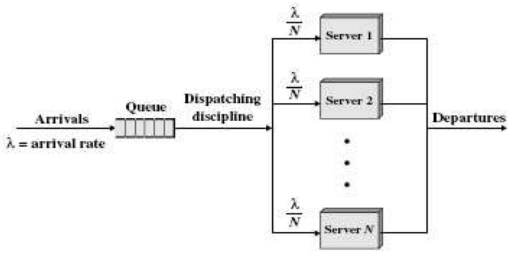
- Για ρ=1 το σύστημα φτάνει σε κόρο.
- Ο θεωρητικός μέγιστος ρυθμός αφίξεων που μπορούν να εξυπηρετηθούν είναι $\lambda_{\max} = \frac{1}{T_s}$
- Στην πράξη υπάρχει κόρος πολύ νωρίτερα.

Χαρακτηριστικά των μοντέλων

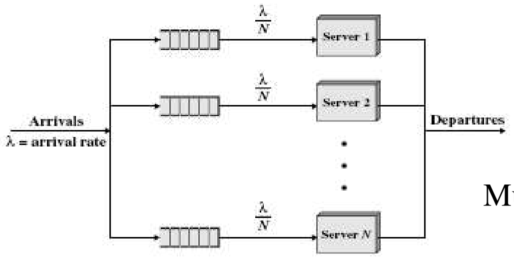
1. Πλήθος αντικειμένων προς άφιξη: θεωρείται άπειρος
2. Μέγεθος ουράς αναμονής: θεωρείται άπειρο. Αν είναι πεπερασμένο κάποια αντικείμενα μπορεί να απορριφθούν από το σύστημα
3. Σειρά επεξεργασίας των εισερχομένων αντικειμένων:
   a) FIFO (FCFS)
   b) LIFO
   c) Ανάλογα με το χρόνο επεξεργασίας (δύσκολη αναλυτική περιγραφή)
      a) πρώτα τα μικρότερα
      b) πρώτα τα μεγαλύτερα
   d) Ανάλογα με προτεραιότητες

Stability & Steady State

- Ένα απλό σύστημα ουράς είναι σταθερό αν
  `packet arrival rate < system transmission capacity`
- Για μια απλή ουρά ο λόγος
  `packet arrival rate / system transmission capacity`
  ονομάζεται utilization factor (ποσοστό χρήσης) = λ/μ
  - Περιγράφει την επιβάρυνση – χρήση της ουράς
- Σε ένα μη σταθερό σύστημα τα πακέτα συσσωρεύονται σε μνήμες προσωρινής αποθήκευσης (Buffers) και / ή απορρίπτονται.
- Σε ασταθή συστήματα με μεγάλα buffers οι καθυστερήσεις μερικών πακέτων γίνονται πολύ μεγάλες.
  - Μηχανισμοί ελέγχου ροής χρησιμοποιούνται για να ελέγξουν το ρυθμό αφίξεων
  - Η θέσπιση προτεραιοτήτων στις ροές θέτουν όρια στην καθυστέρηση πακέτων που ανήκουν σε σημαντικές εφαρμογές/υπηρεσίες.
- Σταθερά συστήματα με κατανομές αφίξεων που είναι στάσιμες (stationary) στο χρόνο, τείνουν σε μια σταθερή κατάσταση (steady state)



Multiple server queue



Multiple single-server queues

Μερικές βασικές σχέσεις

Γενικές σχέσεις		Single Server	Multiserver
$q = \lambda T_q$	Little's formula	$\rho = \lambda T_s$	$\rho = \lambda \frac{T_s}{N}$
$w = \lambda T_w$	Little's formula	$q = w + \rho$	$q = w + N\rho$
$T_q = T_s + T_w$			

Ο νόμος του Little

- Για δεδομένο ρυθμό αφίξεων ο χρόνος παραμονής στο σύστημα είναι ανάλογος του βαθμού κατάληψης (occupancy)

  $q = \lambda T$

  όπου

  q: ο μέσος αριθμός πακέτων στο σύστημα

  λ: ο ρυθμός αφίξεων (πακέτα στην μονάδα χρόνου)

  T: μέση καθυστέρηση (χρόνος στο σύστημα) ανά πακέτο
- Παραδείγματα:
  - Στις βροχερές μέρες υπάρχει περισσότερη κυκλοφορία
  - Τα «φαστφουντάδικα» χρειάζονται μικρότερο χώρο εστίασης σε σύγκριση με κανονικά εστιατόρια με τον ίδιο ρυθμό άφιξης πελατών.
  - Μεγάλες μνήμες αναμονής (buffers) και υψηλοί ρυθμοί αφίξεων έχουν σαν αποτέλεσμα μεγαλύτερες καθυστερήσεις

Ερμηνεία του Νόμου του Little

- Έστω ένα Λούνα Πάρκ: έρχονται άνθρωποι, διασκεδάζουν για κάποιο χρονικό διάστημα σε διάφορα παιχνίδια και φεύγουν
- Έστω ότι πληρώνουν 1 € ανά μονάδα χρόνου.
- Ο ρυθμός εσόδων του πάρκου είναι q € ανά μονάδα χρόνου (q: μέσος αριθμός ανθρώπων στο πάρκο)
- Ο ρυθμός με τον οποίον πληρώνουν οι πελάτες είναι λ x T ανά μονάδα χρόνου (λ: ρυθμός αφίξεων, T: χρόνος ανά άτομο)
- Μακροπρόθεσμα:

  Ο ρυθμός εσόδων = Ρυθμό πληρωμών

  ή

  $q = \lambda T$

Μερικές βασικές σχέσεις

Η πιθανότητα x αφίξεων στη μονάδα του χρόνου	$P(x)=\frac{\lambda^x e^{-\lambda}}{x!}$
Η πιθανότητα να μην υπάρχουν αντικείμενα στο σύστημα	$P_0=1-\frac{\lambda}{\mu}$
Η πιθανότητα ένα αντικείμενο που μόλις έφτασε να πρέπει να περιμένει	$\rho=\frac{\lambda}{\mu}=\lambda \mathrm{T}_s$
Ο μέσος αριθμός αντικειμένων στη ουρά (περιμένουν εξυπηρέτηση)	$w=\frac{\lambda^2}{\mu(\mu-\lambda)}=\frac{\rho^2}{1-\rho}$
Ο μέσος αριθμός αντικειμένων στο σύστημα	$q=w+\rho=\frac{\rho}{1-\rho}$
Ο μέσος χρόνος παραμονής ενός αντικειμένου στη ουρά αναμονής	$T_w=\frac{w}{\lambda}$
Ο μέσος χρόνος παραμονής στο σύστημα	$T_q=\frac{T_s}{1-\rho}=T_s+T_w$
Η πιθανότητα να υπάρχουν N αντικείμενα στο σύστημα	$\Pr[Q=N]=(1-\rho)\rho^{\mathrm{N}}$
Η πιθανότητα να υπάρχουν λιγότερα από N αντικείμενα	$\Pr[Q\le N]=\sum_{i=0}^{N}(1-\rho)\rho^i$

Παράδειγμα 1:
Εξυπηρέτηση πελατών σε κατάστημα
Ρυθμός αφίξεων

Έστω, ότι οι πελάτες εισέρχονται τυχαία και ανεξάρτητα μεταξύ τους (κατανομή Poisson). Η πιθανότητα να έχουμε x αφίξεις σε μια χρονική στιγμή δίνεται από την:

$$P(x)=\frac{\lambda^x e^{-\lambda}}{x!},\quad x=0,1,2,...$$

όπου λ η μ.τ. αφίξεων στη μονάδα του χρόνου ή ο ρυθμός αφίξεων. Έστω, ότι μετά από μέτρηση προκύπτει ότι ο αριθμός άφιξης πελατών είναι 45 / ώρα. Επομένως, για ένα λεπτό είναι λ=45/60=0,75 αφίξεις /λεπτό (e=2,7183).

$$P(0)=\frac{\lambda^0 e^{-0,75}}{0!}=0,4724,\quad P(1)=\frac{0,75^1 e^{-0,75}}{1!}=0,3543,...$$

Παράδειγμα 1:
Εξυπηρέτηση πελατών σε κατάστημα
Χρόνος εξυπηρέτησης

- Αν οι χρόνοι εξυπηρέτησης ανήκουν σε εκθετική κατανομή, τότε η πιθανότητα ο χρόνος εξυπηρέτησης να είναι μικρότερος από κάποιον χρόνο t, είναι:

$$P(service\,time \le t) = 1 - e^{-\mu t}$$

Αν εξυπηρετούνται 60 παραγγελίες την ώρα, είναι μ = 1 παραγγελία /λεπτό. Άρα η πιθανότητα να εξυπηρετηθεί σε λιγότερο από ½ λεπτό είναι:

$$P(service\,time \le 0,5) = 1 - e^{-1\cdot 0,5} = 0,3935$$

Παράδειγμα:
χαρακτηριστικά λειτουργίας

Η πιθανότητα x αφίξεων σε ένα λεπτό	$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$
Η πιθανότητα να μην υπάρχουν πελάτες στο κατάστημα	$P_0 = 1 - \frac{\lambda}{\mu} = 0,25$
Η πιθανότητα ένας πελάτης, που μόλις έφτασε, να πρέπει να περιμένει	$\rho = \frac{\lambda}{\mu} = 0,75$
Ο μέσος αριθμός πελατών στη ουρά (περιμένουν εξυπηρέτηση)	$w = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{1-\rho} = 2,25$
Ο μέσος αριθμός πελατών στο κατάστημα	$q = w + \rho = 3$
Ο μέσος χρόνος παραμονής ενός πελάτη στη ουρά αναμονής	$T_w = \frac{w}{\lambda} = 3$ λεπτά
Ο μέσος χρόνος παραμονής στο σύστημα	$T_q = \frac{T_s}{1-\rho} = T_s + T_w = 4$ λεπτά
Η πιθανότητα να υπάρχουν Ν πελάτες στο κατάστημα	$\Pr[Q = N] = (1-\rho)\rho^{N}$
Η πιθανότητα να υπάρχουν λιγότεροι από Ν πελάτες	$\Pr[Q \le N] = \sum_{i=0}^{N} (1-\rho)\rho^i$

Στόχοι

- Γνωρίζοντας (α) το ρυθμό αφίξεων, λ, και (β) το χρόνο εξυπηρέτησης, T_s, θέλουμε να βρούμε:
  - Τα αντικείμενα σε αναμονή
  - Το χρόνο αναμονής
  - Τα αντικείμενα στην ουρά
  - Το χρόνο παραμονής στην ουρά

  μ.τ. και σ

  - Τι buffer πρέπει να τοποθετήσω για να έχω πιθανότητα απόρριψης < 0,001 ;
  - Ποια είναι η τιμή του N, για την οποία: Pr[αντικείμενα σε αναμονή<N]=0,999 ;

Σημαντικές υποθέσεις

1. Ο ρυθμός αφίξεων υπακούει σε κατανομή Poisson, ή με άλλα λόγια

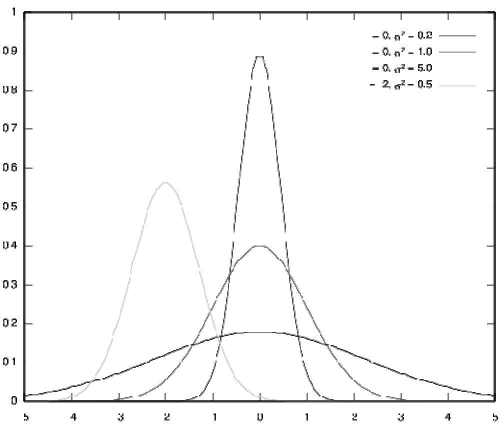
   οι χρόνοι μεταξύ των αφίξεων ανήκουν σε εκθετική κατανομή, ή αλλιώς ότι

   οι αφίξεις συμβαίνουν τυχαία και ανεξάρτητα η μια από την άλλη.
2. Οι υπολογισμοί γίνονται ακόμα πιο εύκολοι αν υποθέσουμε ότι οι χρόνοι εξυπηρέτησης ακολουθούν είτε εκθετική κατανομή είτε ομοιόμορφη κατανομή.

Αναπαράσταση - συμβολισμός

- X/Y/N/Z/K/L (Kendall, 1951)
  - X: κατανομή του χρόνου μεταξύ των αφίξεων
  - Y: κατανομή του χρόνου εξυπηρέτησης
  - N: ο αριθμός των εξυπηρετητών
  - Z: ο τύπος της ουράς, π.χ. FCFS
  - K: ο μέγιστος επιτρεπόμενος αριθμός πελατών σε αναμονή
  - L: το μέγεθος του πληθυσμού από το οποίο προέρχονται οι αφίξεις
- Οι πιο κοινές κατανομές συμβολίζονται ως:
  - G: γενικά ανεξάρτητες αφίξεις ή χρόνοι εξυπηρέτησης
  - M: αρνητική εκθετική κατανομή
  - D: ντετερμινιστικές αφίξεις ή σταθεροί χρόνοι εξυπηρέτησης
- Έτσι, για παράδειγμα, M/M/1 σημαίνει ουρά αναμονής με έναν εξυπηρετητή, ρυθμό αφίξεων Poisson, εκθετικούς χρόνους εξυπηρέτησης.
- Κανονικά θα έπρεπε να αναφέρεται ως: M/M/1/FCFS/∞/∞

Κοινές κατανομές

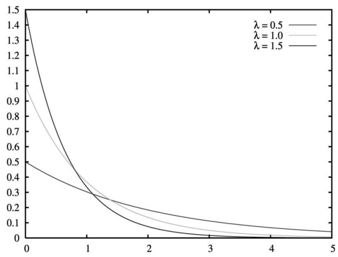


Η κανονική κατανομή

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

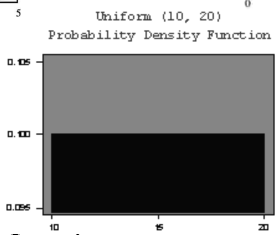
$$x \in (-\infty, \infty)$$

Κοινές κατανομές

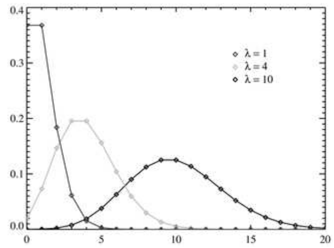


Εκθετική κατανομή:

$$P(x;\mu) = \mu e^{-\mu x}$$



Ομοιόμορφη κατανομή

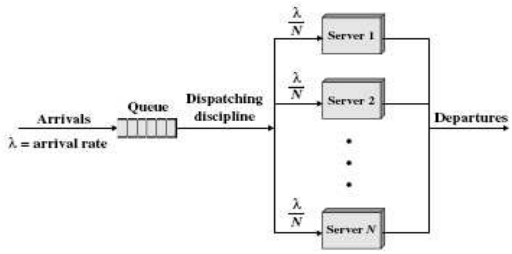


Κατανομή Poisson

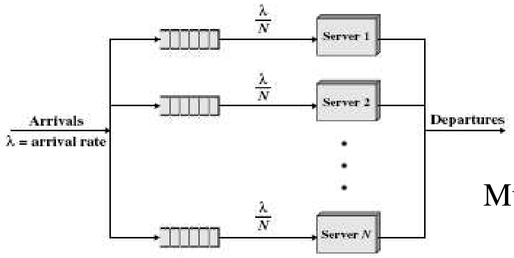
$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Βελτίωση της λειτουργίας μιας ουράς αναμονής

- Η βελτίωση της λειτουργίας συνίσταται στη μείωση του χρόνου παραμονής στο σύστημα.
- Οι γενικές προσεγγίσεις βελτίωσης συνίστανται στις:
  - Αύξηση του μέσου ρυθμού εξυπηρέτησης (μείωση του χρόνου εξυπηρέτησης) που γίνεται είτε μέσω της χρήσης νέας τεχνολογίας είτε μέσω κάποιας ευφυούς σχεδίασης
  - Προσθήκη επιπλέον καναλιών εξυπηρέτησης



Multiple server queue



Multiple single-server queues

Παράδειγμα 2:
Εξυπηρέτηση πελατών σε κατάστημα

- Στο παράδειγμα του fast-food η διοίκηση αποφάσισε την πρόσληψη ενός παραγγελιοδόχου που θα βοηθά τον υπάλληλο που ετοιμάζει την παραγγελία. Με το νέο σχεδιασμό ο ρυθμός εξυπηρέτησης ανέβηκε στους 75 πελάτες ανά ώρα. Αυτό δίνει μ=75/60=1,25 πελάτες ανά λεπτό. Έτσι τώρα:

Βρείτε τις παραμέτρους: ρ, P_0, w, q, T_w, T_q

Παράδειγμα 3

- Έστω εταιρεία που διαθέτει έναν πολύ ακριβό εξυπηρετητή για εξειδικευμένη σχεδίαση γραφικών. Κατά τη διάρκεια μιας τυπικής εργάσιμης ημέρας (8-ωρο) οι 10 μηχανικοί της εταιρείας χρησιμοποιούν τον server για 30 λεπτά κατά μέσο όρο. Παραπονιούνται, δε, ότι είναι δύσκολο να βρουν ελεύθερο τον server και μάλιστα υποστηρίζουν πως πολλές φορές χρειάζεται να περιμένουν περισσότερο από μια ώρα για να τον χρησιμοποιήσουν. Ο διευθυντής της εταιρείας θεωρεί τους ισχυρισμούς τους υπερβολικούς διότι υπολογίζει ότι ο εξυπηρετητής χρησιμοποιείται μόνο κατά τα 5/8 της ημέρας. Ποιος έχει δίκιο;

Σχέσεις που χρειάζονται

- ρ : χρήση (utilization) = 5/8
- Μέσος χρόνος αναμονής: $\mathrm{T}_w = \frac{\rho \mathrm{T}_s}{1-\rho} = 49,99\,\mathrm{min}$
- Χρόνος αναμονής στο 10% των περιπτώσεων:

$$m_{\mathrm{T}_w}(90) = \frac{\mathrm{T}_w}{\rho} \times \ln\left(\frac{100\rho}{100-r}\right) = \frac{50}{5/8} \times \ln\left(\frac{100(5/8)}{100-90}\right) = 146,6\,\mathrm{min}$$

- Ρυθμός αφίξεων $\lambda = \frac{10}{8}(hour) = \frac{10}{8 \times 60} = 0,021$ μηχ. / λεπτό
- Μέσος αριθμός μηχανικών σε αναμονή $w = \lambda \mathrm{T}_w = 1,0416$ μηχ.

Παράδειγμα 4: Δρομολογητής

- Έστω ότι τα πακέτα ενός LAN δρομολογούνται προς το διαδίκτυο μέσω ενός δρομολογητή. Έστω, επίσης, ότι τα πακέτα φτάνουν στο δρομολογητή με ρυθμό λ=5 πακέτα/sec, και το μέγεθος τους ακολουθεί εκθετική κατανομή με μέσο μήκος 144 bytes. Η ταχύτητα διασύνδεσης με τον έξω κόσμο είναι 9600 bps. Θέτουμε τις παρακάτω ερωτήσεις:

Παράδειγμα 4: Δρομολογητής

1. Ποιος είναι ο μέσος χρόνος παραμονής στο δρομολογητή;

Η ερώτηση ζητάει το μέσο χρόνο παραμονής στην ουρά, T_q.

λ= 5 packets/sec,

$$T_s = \frac{144 \times 8}{9600} = 0,12\sec$$

$\rho=\lambda T_s=0,6$

$$T_q = \frac{T_s}{1-\rho} = 0,3\sec$$

Παράδειγμα 4: Δρομολογητής

2. Ποιος είναι ο μέσος αριθμός πακέτων στο δρομολογητή, υπολογίζοντας όσα αναμένουν να εκπεμφθούν μαζί με αυτό που μόλις εκπέμπεται;

$$q = \frac{\rho}{1-\rho} = 1,5\, packets$$

Παράδειγμα 4: Δρομολογητής

3. Πόσα πακέτα βρίσκονται μέσα στο δρομολογητή το 90% του χρόνου; και (4) πόσα στο 95% του χρόνου;

Η πιθανότητα να έχω Ν αντικείμενα μέσα στην ουρά δίνεται από τη σχέση:

$$\Pr[Q = N] = (1-\rho)\rho^{N}$$

Η οποία σε μορφή αθροίσματος είναι:

$$\frac{r}{100} = \sum_{k=0}^{m_q(r)} (1-\rho)\rho^{k} = 1-\rho^{1+m_q(r)}$$

Όπου $m_q(r)$ είναι το μέγιστο πλήθος πακέτων που αναμένεται στην ουρά το r ποσοστό του χρόνου. Επομένως, πρέπει να βρούμε το $m_q(r)$ για r=90 και r=95.

$$\rho^{1+m_q(r)} = 1-\frac{r}{100} \Rightarrow m_q(r) = \frac{\ln(1-\frac{r}{100})}{\ln\rho} - 1$$

r=90 → $m_q(r)$=4

r=95 → $m_q(r)$=5

buffer size!!

Παράδειγμα 5: Εξυπηρετητής βάσης δεδομένων

- Έστω ένα τοπικό δίκτυο με 100 PC κι ένα database server. Ο μέσος χρόνος εξυπηρέτησης των ερωτημάτων είναι 0,6s με ίση τυπική απόκλιση. Στις ώρες αιχμής ο αριθμός ερωτημάτων φτάνει στα 20 ανά λεπτό. Θέλουμε να απαντήσουμε στα παρακάτω ερωτήματα:
  - Ποιος είναι ο μέσος χρόνος απόκρισης;
  - Αν ο μέγιστος αποδεκτός χρόνος απόκρισης θεωρηθεί 1,5s ποιο ποσοστό αύξησης στο φορτίο μπορεί να συμβεί πριν φτάσουμε στο μέγιστο;
  - Αν η χρήση της υπηρεσίας αυξηθεί κατά 20%, ο χρόνος απόκρισης θα αυξηθεί περισσότερο ή λιγότερο από 20%;

Παράδειγμα 5: Εξυπηρετητής βάσης δεδομένων

- Υποθέτω ουρά M/M/1 και δεν λαμβάνω υπόψη την επίδραση του LAN στην καθυστέρηση. Τότε, ο ρυθμός χρήσης της υπηρεσίας είναι:

$\rho=\lambda T_s$=(20queries/min)(0,6sec/transmission)/(60sec/min)=0,2

$$T_q = \frac{T_s}{1-\rho} = \frac{0,6}{1-0,2} = 0,75\sec$$

Παράδειγμα 5: Εξυπηρετητής βάσης δεδομένων

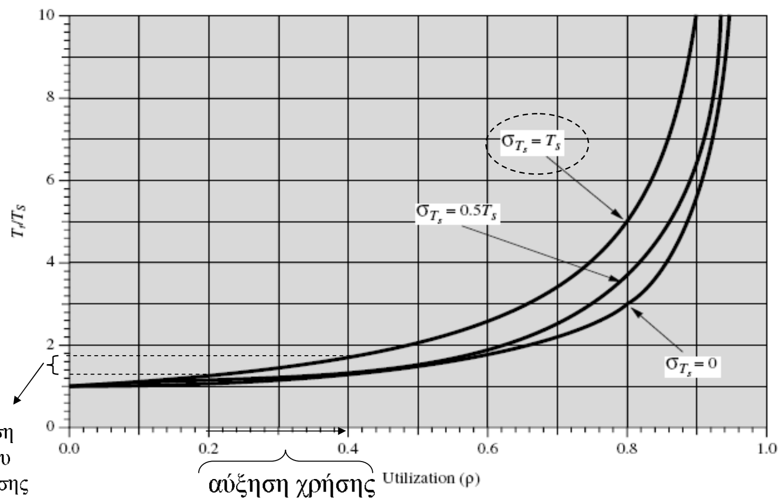
- Η δεύτερη ερώτηση περιέχει ένα παράδοξο στη διατύπωσή της. Μέγιστος χρόνος σημαίνει ότι δεν δεχόμαστε τιμές μεγαλύτερες από αυτόν. Η αλλιώς, $P(T_q > 1,5 \text{ sec})=0$. Εναλλακτικά, μπορούμε να απαιτήσουμε το 90% των χρόνων απόκρισης να είναι μικρότερο από 1,5 sec. Τότε:

$$m_{T_q}(r) = T_q \times \ln\left(\frac{100}{100-r}\right) = \frac{T_s}{1-\rho}\ln\left(\frac{100}{100-r}\right) \Rightarrow$$

$$m_{T_q}(90) = 1,5 = \frac{T_s}{1-\rho}\ln(10) \Rightarrow 1-\rho = \frac{0,6}{1,5}2,3 \Rightarrow \rho = 0,08$$

- Δηλαδή, η χρήση πρέπει να κατέβει στο 8% για να μπουν τα 1,5 sec στο 90ο ποσοστιμόριο.

Παράδειγμα 5: Εξυπηρετητής βάσης δεδομένων



Αύξηση χρόνου απόκρισης 30%

αύξηση χρήσης 100%