Εισαγωγή στις Ουρές Αναμονής

Αντώνης Οικονόμου
aeconom@math.uoa.gr

3 Μαρτίου 2008

1 Βασική περιγραφή

Ένα σύστημα εξυπηρέτησης ή ουρά αναμονής (queueing system, queue) είναι στην ουσία ένα σύστημα εισόδου - εξόδου στο οποίο υπεισέρχεται τυχαιότητα. Αυτός ο ορισμός είναι πολύ γενικός και περιλαμβάνει πολλές πραγματικές καταστάσεις που παρατηρούμε στην καθημερινή ζωή, καθώς και σε περίπλοκα τεχνολογικά συστήματα. Ιστορικά, η συγκεκριμένη επιστημονική περιοχή άρχισε να αναπτύσσεται στις αρχές του 20ου αιώνα, όταν ο δανός A.K. Erlang δημοσίευσε κάποιες εργασίες του για τη μαθηματική μοντελοποίηση του συνωστισμού σε τηλεφωνικά δίκτυα. Η μεγάλη επιτυχία αυτών των μεθόδων στη μελέτη πραγματικών συστημάτων έδωσε τεράστια ώθηση στη περαιτέρω ανάπτυξη της θεωρίας των ουρών αναμονής καθώς και των εφαρμογών της και σε άλλα πεδία.

Τα βασικά χαρακτηριστικά ενός συστήματος εξυπηρέτησης είναι η διαδικασία αφίξεων (arrival process), οι χρόνοι εξυπηρέτησης (service times), ο αριθμός των παροχέων εξυπηρέτησης - υπηρετών (number of servers), η χωρητικότητα του συστήματος (system capacity) και η πειθαρχία ουράς (queue discipline). Για το λόγο αυτό ο βρετανός πιθανοθεωρητικός Kendall εισήγαγε ένα σύστημα ονοματολογίας για τις πιο απλές ουρές που περιγράφει συνοπτικά αυτά τα χαρακτηριστικά. Η ονοματολογία του Kendall έχει τη μορφή $A/B/c/k(\quad)$, όπου τα A, B είναι γράμματα, τα c, k αριθμοί και μέσα στην παρένθεση γράφεται μια ακροστοιχίδα γραμμάτων. Καθεμιά από τις 5 παραμέτρους της ονοματολογίας του Kendall αναφέρεται στα 5 χαρακτηριστικά που περιγράψαμε παραπάνω.

Η διαδικασία αφίξεων περιγράφει το πώς έρχονται οι πελάτες στο σύστημα. Είναι συνήθως μια ανανεωτική διαδικασία, δηλαδή θεωρούμε ότι οι χρόνοι μεταξύ διαδοχικών αφίξεων πελατών είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με κάποια γνωστή γενική κατανομή. Η διαδικασία αφίξεων αντιστοιχεί στο γράμμα A της ονοματολογίας Kendall. Οι τιμές που παίρνει είναι GI ή G (General independent), M (Memoryless, Markovian), D (Deterministic) και E_r (Erlang-r) για τις περιπτώσεις που οι ενδιάμεσοι χρόνοι μεταξύ των αφίξεων είναι γενικοί, εκθετικοί, σταθεροί και Erlang-r αντίστοιχα. Υπάρχουν βέβαια και άλλες τιμές για το γράμμα A που αντιστοιχούν σε κατανομές που εμφανίζονται σπανιότερα στη βιβλιογραφία.

Οι χρόνοι εξυπηρέτησης θεωρούνται στα κλασικά μοντέλα επίσης ανεξάρτητοι και ισόνομοι και αντιστοιχούν στο γράμμα B της ονοματολογίας Kendall που παίρνει τις ίδιες τιμές με το γράμμα A. Οι τιμές GI και G για το A και το B σηματοδοτούν ακριβώς το ίδιο, δηλαδή ανεξάρτητους ισόνομους χρόνους με γενική κατανομή. Παρόλα αυτά συνηθίζεται εθιμικά η τιμή GI για το A και η τιμή G για το B στα περισσότερα βιβλία και σημειώσεις που κυκλοφορούν διεθνώς. Θα διατηρήσουμε αυτή τη σύμβαση και στις παρούσες σημειώσεις.

Ο αριθμός των υπηρετών αναφέρεται στο πόσοι είναι οι παράλληλοι υπηρέτες που εξυπηρετούν τη ροή των πελατών που εισέρχεται στο σύστημα. Με την έννοια "παράλληλοι" υπηρέτες εννοούμε ότι υπάρχει μια κοινή ουρά για όλους και οι πελάτες πηγαίνουν στον πρώτο υπηρέτη που θα αδειάσει, αν όλοι οι υπηρέτες είναι απασχολημένοι, ή διαλέγουν στην τύχη κάποιον από τους άδειους υπηρέτες, αν υπάρχουν ελεύθεροι υπηρέτες. Ο αριθμός των υπηρετών αντιστοιχεί στον αριθμό c της ονοματολογίας Kendall.

Η χωρητικότητα του συστήματος εκφράζει το μέγιστο πλήθος πελατών που μπορεί να χωρέσει το σύστημα, συμπεριλαμβανομένων τόσο αυτών που περιμένουν να εξυπηρετηθούν όσο και αυτών που βρίσκονται σε διαδικασία εξυπηρέτησης. Αν ένα σύστημα έχει φτάσει στο μέγιστο της χωρητικότητάς του και αφιχθεί ένας πελάτης τότε στο κλασικό μοντέλο των ουρών αναμονής ο πελάτης απορρίπτεται και θεωρείται χαμένος για πάντα. Φυσικά υπάρχουν και μοντέλα στα οποία οι πελάτες που αποχωρούν λόγω της περιορισμένης χωρητικότητας επανέρχονται αργότερα με την ελπίδα να υπάρχει διαθέσιμη θέση στο σύστημα. Στην περίπτωση αυτή μιλάμε για μοντέλα με επαναπροσπάθειες ή επαναδοκιμές (retrials). Σε τέτοια μοντέλα είναι απαραίτητο να αποσαφηνισθεί η διαδικασία με την οποία οι πελάτες επανέρχονται στο σύστημα και έτσι τα μοντέλα αυτά δεν περιγράφονται στο πλαίσιο της ονοματολογίας Kendall. Προς το παρόν, επομένως, μένουμε στο πλαίσιο των μοντέλων χωρίς επαναπροσπάθειες, όπου η χωρητικότητα του συστήματος αντιστοιχεί στον αριθμό k της ονοματολογίας του Kendall.

Η πειθαρχία ουράς είναι ο τρόπος με τον οποίο το σύστημα διαλέγει ποιόν πελάτη θα εξυπηρετήσει, μόλις βρεθεί κάποιος διαθέσιμος υπηρέτης. Η πλέον κλασική πειθαρχία ουράς είναι η FCFS (First-Come-First-Served) κατά την οποία οι πελάτες εξυπηρετούνται σύμφωνα με τη σειρά της άφιξής τους. Έτσι, μόλις αδειάσει ένας υπηρέτης, επιλέγεται για εξυπηρέτηση ο πελάτης που έχει αφιχθεί πρώτος από όλους που περιμένουν. Η πειθαρχία αυτή μοιάζει η πιο δίκαιη με πρώτη ματιά και χρησιμοποιείται περισσότερο από οποιαδήποτε άλλη στην περίπτωση που οι πελάτες είναι άνθρωποι και μάλιστα έχουν οπτική επαφή με το τί συμβαίνει στο σύστημα (και επομένως μπορούν και βλέπουν πότε φθάνουν οι άλλοι πελάτες). Σε διάφορες εφαρμογές, πάντως, χρησιμοποιούνται και άλλες πειθαρχίες ουράς, όπως η LCFS (Last-Come-First-Served) κατά την οποία οι πελάτες εξυπηρετούνται αντίστροφα από τη σειρά αφιξής τους, η SIRO (Service-In-Random-Order) όπου οι πελάτες εξυπηρετούνται τυχαία, χωρίς να λαμβάνεται υπόψιν η σειρά αφιξής τους, η SSTF (Shortest-Service-Time-First) όπου επιλέγεται προς εξυπηρέτηση ο πελάτης που έχει το μικρότερο χρόνο εξυπηρέτησης κ.α. Υπάρχουν επίσης πειθαρχίες ουράς για την περίπτωση που υπάρχουν διάφορα είδη πελατών και το σύστημα τους αντιμετωπίζει διαφορετικά. Στην

περίπτωση αυτή είναι δυνατόν κάποια είδη πελατών να έχουν προτεραιότητα έναντι κάποιων άλλων, οπότε μιλάμε για πειθαρχίες ουρών με προτεραιότητες. Γενικά, αν σκεφθούμε πρακτικές εφαρμογές των ουρών αναμονής θα συνειδητοποιήσουμε ότι υπάρχει μεγάλη ποικιλία στις πειθαρχίες ουράς που χρησιμοποιούνται (σκεφτείτε τα ταμεία για πελάτες με μέχρι 10 προϊόντα στα supermarket, τα ταμεία για επιχειρηματικές συναλλαγές στις τράπεζες, τις κρατήσεις θέσεων σε εστιατόρια κλπ.).

Η χωρητικότητα του συστήματος k και/ή πειθαρχία ουράς μπορεί να παραλείπονται στην ονοματολογία Kendall, αυτό συμβαίνει στην περίπτωση που έχουμε απεριόριστη χωρητικότητα ($k = \infty$) ή πειθαρχία FCFS αντίστοιχα.

Παράδειγμα 1: Η $M/M/1$ ουρά είναι ένα σύστημα εξυπηρέτησης με Poisson διαδικασία αφίξεων (ανεξάρτητους εκθετικούς ενδιάμεσους χρόνους αφίξεων), εκθετικούς χρόνους εξυπηρέτησης και 1 υπηρέτη που έχει άπειρη χωρητικότητα και λειτουργεί υπό την πειθαρχία ουράς FCFS.

Παράδειγμα 2: Η $GI/E_2/1/5(SIRO)$ ουρά είναι ένα σύστημα εξυπηρέτησης με ανανεωτική διαδικασία αφίξεων (ανεξάρτητους και ισόνομους ενδιάμεσους χρόνους αφίξεων), Erlang-2 χρόνους εξυπηρέτησης και 1 υπηρέτη που έχει χωρητικότητα για 5 πελάτες και λειτουργεί υπό την πειθαρχία ουράς SIRO.

Πολλές φορές η διαδικασία αφίξεων και/η κατανομή των χρόνων εξυπηρέτησης δεν είναι ακριβώς γνωστή. Στην περίπτωση αυτή μπορεί να δίνεται κάποια αδρή πληροφορία για το πώς έρχονται και πώς εξυπηρετούνται οι πελάτες. Π.χ. μπορεί να δίνεται ο μέσος ενδιάμεσος χρόνος a μεταξύ δυο διαδοχικών αφίξεων και ο μέσος χρόνος εξυπηρέτησης b. Ισοδύναμα, μπορεί να δίνεται ο ρυθμός αφίξεων $\lambda = 1/a$ και ο ρυθμός εξυπηρέτησης $\mu = 1/b$. Τα a και b έχουν τη φυσική έννοια της περιόδου των διαδικασιών των αφίξεων και των εξυπηρετήσεων αντίστοιχα, ενώ τα λ και μ αντιστοιχούν στη φυσική έννοια της συχνότητας. Με τόσο ελλειπή πληροφορία, βεβαίως, τα αποτελέσματα που θα προκύψουν από τη μαθηματική ανάλυση μπορεί να είναι τελείως αναξιόπιστα. Για την εξαγωγή ασφαλών συμπερασμάτων χρειάζεται να είναι γνωστή η κατανομή των αντίστοιχων χρόνων ή τουλάχιστον κάποιες ροπές ανώτερης τάξης. Μετά τη μέση τιμή, η διασπορά των χρόνων μεταξύ των αφίξεων και/ή των χρόνων εξυπηρέτησης επηρεάζει σημαντικά την απόδοση ενός συστήματος.

2 Μέτρα απόδοσης συστήματος

Αφού περιγραφεί ένα σύστημα, το πρόβλημα που τίθεται είναι να προβλέψουμε πώς θα συμπεριφέρεται. Τυπικά ερωτήματα που απασχολούν τη θεωρία των ουρών αναμονής είναι:

1. Πόσοι πελάτες θα βρίσκονται στο σύστημα κατά μέσο όρο μια τυχούσα χρονική στιγμή;
2. Πόσο χρόνο θα περάσει στο σύστημα κατά μέσο όρο ένας πελάτης;
3. Ποιό ποσοστό του χρόνου του θα βρίσκεται απασχολημένος ένας υπηρέτης που δουλεύει στο συγκεκριμένο σύστημα;

Όπως βλέπουμε υπάρχουν ερωτήματα που απασχολούν το διαχειριστή του συστήματος, που βλέπει το σύστημα συνολικά, σαν εξωτερικός παρατηρητής (ερώτημα 1), ερωτήματα που απασχολούν τους πελάτες, που επιδρούν στο σύστημα μόνο παροδικά και κατόπιν φεύγουν (ερώτημα 2) και ερωτήματα που απασχολούν τους υπηρέτες, που επιδρούν διαρκώς στο σύστημα (ερώτημα 3). Είναι σημαντικό να κατανοηθεί ότι αυτοί οι παράγοντες του συστήματος (διαχειριστής, πελάτες και υπηρέτες) έχουν διαφορετικές οπτικές και για το λόγο αυτό υπάρχουν μέτρα απόδοσης του συστήματος που σχετίζονται με την οπτική του καθενός.

Για το διαχειριστή του συστήματος η πιο σημαντική πληροφορία είναι ο αριθμός των πελατών που βρίσκονται στο σύστημα μια τυχαία χρονική στιγμή. Έτσι ορίζουμε

- $Q(t)$ τον αριθμό των πελατών στο σύστημα τη στιγμή t,
- $Q_q(t)$ τον αριθμό των πελατών στο χώρο αναμονής τη στιγμή t (ο δείκτης q μπαίνει για να θυμίζει ότι αναφερόμαστε στο πλήθος των πελατών στην ουρά - queue),
- $Q_s(t)$ τον αριθμό των πελατών στο χώρο εξυπηρέτησης τη στιγμή t (ο δείκτης s μπαίνει για να θυμίζει ότι αναφερόμαστε στο πλήθος των πελατών υπό εξυπηρέτηση - service).

Παρατηρούμε ότι

$$Q(t) = Q_q(t) + Q_s(t).$$

Ενδιαφερόμαστε για το τί συμβαίνει στο σύστημα σε κατάσταση ισορροπίας, δηλαδή καθώς $t \to \infty$ και για το λόγο αυτό μας ενδιαφέρουν οι οριακές πιθανότητες

$$p_j = \lim_{t\to\infty} \Pr[Q(t) = j],\ j = 0, 1, \ldots. \tag{1}$$

Η p_j είναι η οριακή πιθανότητα σε συνεχή χρόνο να βρίσκονται j πελάτες στο σύστημα. Μια πρώτη της ερμηνεία επομένως είναι ότι εκφράζει την πιθανότητα να βρίσκονται j πελάτες στο σύστημα, αν το κοιτάξουμε σε κάποια χρονική στιγμή που έχει παρέλθει πολύς χρόνος από την έναρξη λειτουργίας του συστήματος και επομένως η επίδραση της αρχικής κατάστασης του συστήματος (αριθμός πελατών σε αυτό κατά την έναρξη της λειτουργίας του) έχει χαθεί. Είναι όμως σημαντικότερο να την αντιλαμβανόμαστε ως το μακροπρόθεσμο ποσοστό του χρόνου που βρίσκονται j πελάτες στο σύστημα, δηλαδή σαν το όριο (με πιθανότητα 1)

$$p_j = \lim_{t\to\infty} \frac{\text{Χρόνος στο [0,t] που βρίσκονται } j \text{ πελάτες στο σύστημα}}{t} = \lim_{t\to\infty} \frac{\int_0^t 1\{Q(u) = j\}du}{t}, \tag{2}$$

όπου με $1\{Q(u) = j\}$ συμβολίζουμε τη δείκτρια τυχαία μεταβλητή του ενδεχομένου $\{Q(u) = j\}$ που παίρνει την τιμή 1 όταν $Q(u) = j$ και την τιμή 0 διαφορετικά.

Είναι όμως οι εκφράσεις - ορισμοί (1) και (2) ισοδύναμες; Η απάντηση είναι ότι "ναί, είναι ισοδύναμες" κάτω από πολύ γενικές συνθήκες, συγκεκριμένα όταν η στοχαστική διαδικασία $\{Q(t)\}$ είναι

αναγεννητική διαδικασία (regenerative process). Πρακτικά αυτό σημαίνει ότι πρέπει να υπάρχουν χρονικά σημεία στην εξέλιξη της $\{Q(t)\}$ στα οποία η διαδικασία συμπεριφέρεται σαν να ξεκινάει από την αρχή. Στις ουρές αναμονής κατά κανόνα αυτό ισχύει και μάλιστα τα σημεία αυτά είναι οι χρονικές στιγμές που στο σύστημα φθάνει ένας πελάτης που το βρίσκει κενό. Για μια αυστηρή συζήτηση πάνω στο θέμα των αναγεννητικών διαδικασιών μπορεί κανείς να ανατρέξει στα βιβλία των Φακίνου (2003) και Wolff (1989). Όλα τα συστήματα που θα μελετήσουμε είναι αναγεννητικά και κατά συνέπεια οι οριακές πιθανότητες ισούνται με ποσοστά, όπως είδαμε στην περίπτωση των (1) και (2) παραπάνω. Για το λόγο αυτό σε ό,τι ακολουθεί θα παραθέτουμε ενιαία τις δυο εκφράσεις, χωρίς περαιτέρω μνεία.

Συμβολίζοντας με Q την οριακή τυχαία μεταβλητή που έχει την κατανομή $(p_j : j = 0, 1, \ldots)$ ορίζουμε $\overline{Q}$ να είναι το οριακό μέσο πλήθος πελατών στο σύστημα (ισοδύναμα το μακροπρόθεσμο μέσο πλήθος πελατών στο σύστημα) που δίνεται ως

$$\overline{Q} = E[Q] = \lim_{t\to\infty} \frac{1}{t} \int_0^t Q(u)du. \tag{3}$$

Αν έχουμε προσδιορίσει την $(p_j : n = 0, 1, \ldots)$ (πράγμα που δεν είναι πάντα εύκολο) τότε μπορούμε να υπολογίσουμε το $\overline{Q}$ από τη σχέση $E[Q] = \sum_{j=0}^{\infty} jp_j$. Σε κάποιες περιπτώσεις το $\overline{Q}$ μπορεί να υπολογιστεί και απευθείας με πιθανοθεωρητικά επιχειρήματα, χρησιμοποιώντας κάποια βασικά αποτελέσματα που θα παρουσιαστούν παρακάτω. Ομοίως με το $\overline{Q}$ ορίζονται τα $\overline{Q}_q$ και $\overline{Q}_s$.

Για τους πελάτες το πιο σημαντικό μέτρο απόδοσης είναι ο χρόνος που παραμένουν στο σύστημα. Έτσι ορίζουμε

- S_n το χρόνο παραμονής στο σύστημα του n-οστού πελάτη,
- W_n το χρόνο αναμονής στην ουρά (μέχρι να αρχίσει η εξυπηρέτηση) του n-οστού πελάτη,
- X_n το χρόνο εξυπηρέτησης του n-οστού πελάτη.

Παρατηρούμε ότι

$$S_n = W_n + X_n.$$

Ενδιαφερόμαστε και πάλι για το τί συμβαίνει σε κατάσταση ισορροπίας και γι αυτό ενδιαφερόμαστε για την οριακή κατανομή του χρόνου παραμονής

$$\begin{aligned} F_S(x) &= \lim_{n\to\infty} \Pr[S_n \le x] \\ &= \lim_{n\to\infty} \frac{\text{Πλήθος πελατών που παραμένουν για χρόνο} \le x \text{ μεταξύ των πρώτων } n}{n} \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1\{S_k \le x\},\ x \ge 0, \end{aligned}$$

όπου με $1\{S_k \le x\}$ συμβολίζουμε τη δείκτρια συνάρτηση του ενδεχομένου $\{S_k \le x\}$ (ο k-οστός πελάτης να παραμείνει στο σύστημα το πολύ για χρόνο x) που παίρνει την τιμή 1 όταν $S_k \le x$ και την τιμή 0 διαφορετικά.

Συμβολίζοντας με S την οριακή τυχαία μεταβλητή που έχει την κατανομή $(F_S(x) : x \geq 0)$ ορίζουμε $\overline{S}$ να είναι ο οριακός μέσος χρόνος παραμονής ενός πελάτη στο σύστημα (ισοδύναμα ο μακροπρόθεσμος μέσος χρόνος παραμονής στο σύστημα) που δίνεται ως

$$\overline{S} = E[S] = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} S_k.$$

Ο υπολογισμός του $\overline{S}$ είναι εύκολος αν έχουμε υπολογίσει την κατανομή $(F_S(x) : x \geq 0)$ ή την αντίστοιχη συνάρτηση πυκνότητας πιθανότητας $(f_S(x) : x \geq 0)$. Πράγματι χρειάζεται να χρησιμοποιήσουμε τη σχέση $E[S] = \int_0^\infty x f_S(x)dx = \int_0^\infty (1 - F_S(x))dx$ και επομένως πρόκειται για έναν υπολογισμό ρουτίνας (ο οποίος μπορεί πάντως να έχει αρκετές πράξεις). Σε κάποιες περιπτώσεις όμως, το $\overline{S}$ μπορεί να υπολογιστεί και απευθείας ταυτόχρονα με το $\overline{Q}$ με πιθανοθεωρητικά επιχειρήματα. Ομοίως με το $\overline{S}$ ορίζονται τα $\overline{W}$ και $\overline{X}$.

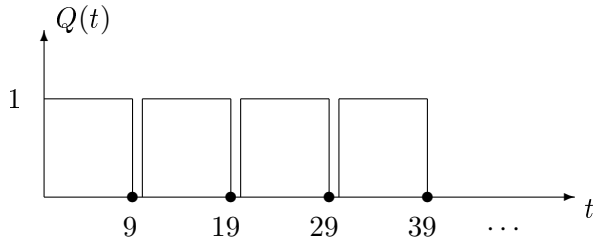
Ως κύκλος απασχόλησης του συστήματος ορίζεται το διάστημα από την αναχώρηση ενός πελάτη που αφήνει το σύστημα κενό, μέχρι την επόμενη αναχώρηση πελάτη που θα αφήσει το σύστημα κενό. Κάθε τέτοιος κύκλος αρχίζει με ένα χρονικό διάστημα που το σύστημα παραμένει κενό μέχρι να εμφανιστεί ο πρώτος πελάτης. Το διάστημα αυτό αναφέρεται ως μια περίοδος αργίας του συστήματος. Από τη στιγμή που θα αφιχθεί ο πρώτος πελάτης το σύστημα θα είναι συνεχώς απασχολημένο μέχρι να τελειώσει ο κύκλος απασχόλησης. Το διάστημα αυτό αναφέρεται ως μια περίοδος συνεχούς λειτουργίας του συστήματος. Οι διάρκειες των περιόδων αργίας I, των περιόδων συνεχούς λειτουργίας Y και των κύκλων απασχόλησης Z μας ενδιαφέρουν κυρίως από την οπτική σκοπιά των υπηρετών και του διαχειριστή του συστήματος αφού τα μέτρα αυτά είναι σημαντικά για να ληφθούν αποφάσεις σχετικά με τη συντήρηση του συστήματος ή με διακοπές-διαλείματα των υπηρετών. Ισχύει ότι $Z = I + Y$. Για το λόγο αυτό ενδιαφερόμαστε για τη μελέτη των αντίστοιχων οριακών κατανομών $F_I(x)$, $F_Y(x)$ και $F_Z(x)$, ή τουλάχιστον των αντίστοιχων μέσων τιμών τους $\overline{I}$, $\overline{Y}$ και $\overline{Z}$.

3 Εμφυτευμένες διαδικασίες σε στιγμές αφίξεων και αναχωρήσεων

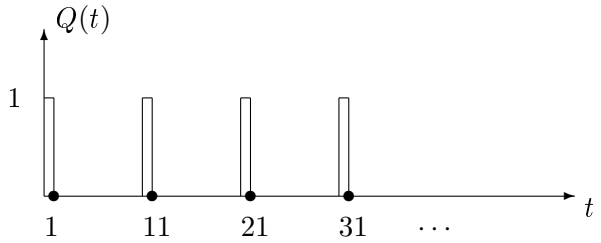
Όπως είπαμε παραπάνω ο διαχειριστής του συστήματος, που μπορεί να θεωρηθεί ως ένας εξωτερικός παρατηρητής του συστήματος, έχει μια διαφορετική αντίληψη από τους πελάτες του συστήματος. Για το λόγο αυτό ορίσαμε και τα διάφορα μέτρα απόδοσης. Για να γίνει περισσότερο κατανοητή η διαφορά των δυο οπτικών, διαχειριστή-εξωτερικού παρατηρητή και πελατών ας φανταστούμε ότι έχουμε ένα $D/D/1$ σύστημα και ας εξετάσουμε τί αντιλαμβάνεται ο διαχειριστής και τί οι πελάτες.

Σε ένα πρώτο σενάριο, ας υποθέσουμε ότι έχουμε αφίξεις σε σταθερά χρονικά διαστήματα, κάθε 10 λεπτά και ότι οι χρόνοι εξυπητέτησης είναι επίσης σταθεροί και ίσοι με 9 λεπτά. Τότε ο διαχειριστής βλέπει το σύστημα πολύ απασχολημένο, για την ακρίβεια βλέπει ότι το 90% του χρόνου στο σύστημα υπάρχει 1 πελάτης ενώ μόνο ένα 10% του χρόνου το σύστημα είναι άδειο. Από την άλλη μεριά κάθε

πελάτης βλέπει το σύστημα άδειο τη στιγμή που φθάνει σε αυτό (αφού ο προηγούμενος πελάτης έχει φύγει πριν 1 λεπτό).



Σε ένα δεύτερο σενάριο, ας υποθέσουμε ότι οι χρόνοι μεταξύ των αφίξεων είναι και πάλι σταθεροί και ίσοι με 10 λεπτά, αλλά τώρα θεωρούμε ότι οι χρόνοι εξυπηρέτησης διαρκούν 1 λεπτό για κάθε πελάτη. Τώρα ο διαχειριστής βλέπει το σύστημα πολύ λίγο απασχολημένο, αφού το 10% μόλις του χρόνου υπάρχει 1 πελάτης ενώ το 90% του χρόνου το σύστημα είναι άδειο. Από την άλλη μεριά η εντύπωση που αποκομίζει κάθε πελάτης φθάνοντας στο σύστημα δεν διαφέρει από την εντύπωση που έχουν οι πελάτες στο πρώτο σενάριο: Και πάλι κάθε πελάτης βλέπει το σύστημα άδειο τη στιγμή που εισέρχεται σε αυτό.



Το συμπέρασμα είναι ότι οι οπτικές του εξωτερικού παρατηρητή και του πελάτη μπορεί να δίνουν πολύ διαφορετικές εικόνες για το ίδιο σύστημα. Επίσης δυο συστήματα μπορεί να μοιάζουν παρόμοια υπό τη μία οπτική (όπως τα δυο σενάρια υπό την οπτική των πελατών) και να είναι πολύ διαφορετικά υπό την άλλη οπτική (όπως τα δυο σενάρια υπό την οπτική του διαχειριστή). Αν σκεφτούμε ότι οι πελάτες δεν είναι παθητικές οντότητες αλλά μπορεί να αποφασίζουν τι θα κάνουν σε σχέση με το σύστημα (π.χ. να μπουν σε αυτό ή να φύγουν) έχει μεγάλη σημασία να ποσοτικοποιήσουμε με κάποιο τρόπο το πώς αντιλαμβάνονται το συνωστισμό του συστήματος οι πελάτες κατά την άφιξή τους ή την αναχώρησή τους.

Προς το σκοπό αυτό έστω $t_1 < t_2 < t_3 < \ldots$ οι διαδοχικές στιγμές αφίξεων και $\tau_1 < \tau_2 < \tau_3 < \ldots$ οι διαδοχικές στιγμές αναχωρήσεων των πελατών. Ξεκινώντας από τη στοχαστική διαδικασία $\{Q(t)\}$ που περιγράφει τον αριθμό των πελατών στο σύστημα σε συνεχή χρόνο, ορίζουμε τις εμφυτευμένες διαδικασίες $\{Q_n^-\}$ και $\{Q_n^+\}$ σε στιγμές αφίξεων και αναχωρήσεων πελατών αντίστοιχα. Έτσι ορίζουμε

- $Q_n^- = Q(t_n^-)$ τον αριθμό των πελατών αμέσως πριν την n-οστή άφιξη πελάτη (δηλαδή των αριθμό των παρόντων πελατών που βλέπει ο n-οστός πελάτης καθώς εισέρχεται στο σύστημα),

- $Q_n^+ = Q(\tau_n^+)$ τον αριθμό των πελατών αμέσως μετά την n-οστή αναχώρηση πελάτη (δηλαδή των αριθμό των πελατών που αφήνει πίσω του κατά την έξοδο του ο n-οστός πελάτης που φεύγει από το σύστημα).

Ενδιαφερόμαστε για τις αντίστοιχες οριακές κατανομές που περιγράφουν την εντύπωση που διαμορφώνει ένας πελάτης τη στιγμή που εισέρχεται στο σύστημα και τη στιγμή που αναχωρεί από αυτό, εφόσον το σύστημα βρίσκεται σε κατάσταση ισορροπίας. Συγκεκριμένα ορίζουμε:

$$\begin{aligned} r_j &= \lim_{n\to\infty} \Pr[Q_n^- = j] \\ &= \lim_{n\to\infty} \frac{\text{Πλήθος αφίξεων που βρίσκουν } j \text{ πελάτες στο σύστημα μεταξύ των πρώτων } n}{n} \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1\{Q_k^- = j\},\ j = 0, 1, \ldots \end{aligned}$$

και

$$\begin{aligned} d_j &= \lim_{n\to\infty} \Pr[Q_n^+ = j] \\ &= \lim_{n\to\infty} \frac{\text{Πλήθος αναχωρήσεων που αφήνουν } j \text{ πελάτες στο σύστημα μεταξύ των πρώτων } n}{n} \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1\{Q_k^+ = j\},\ j = 0, 1, \ldots. \end{aligned}$$

Οι r_j και d_j είναι οι οριακές πιθανότητες να βρίσκονται j πελάτες στο σύστημα σε στιγμές αφίξεων και αναχωρήσεων αντίστοιχα, ενώ η p_j είναι η οριακή πιθανότητα σε συνεχή χρόνο. Όπως και η p_j, έτσι και οι r_j και d_j έχουν ερμηνείες ως ποσοστά. Συγκεκριμένα η r_j εκφράζει το μακροπρόθεσμο ποσοστό των αφίξεων που βρίσκουν j πελάτες στο σύστημα, ενώ η d_j εκφράζει το μακροπρόθεσμο ποσοστό των αναχωρήσεων που αφήνουν j πελάτες στο σύστημα.

Δεν υπάρχει κάποιος λόγος οι οριακές κατανομές (p_j), (r_j) και (d_j) να συμπίπτουν και γενικά αυτό δεν ισχύει. Πράγματι, παρατηρείστε ότι στα δυο σενάρια για την $D/D/1$ ουρά που περιγράψαμε παραπάνω είχαμε $r_0 = 1$ και $r_j = 0,\ j \geq 1$, ενώ στο πρώτο σενάριο είχαμε $p_0 = 0.1$, $p_1 = 0.9$, $p_j = 0,\ j \geq 2$ και στο δεύτερο είχαμε $p_0 = 0.9$, $p_1 = 0.1$, $p_j = 0,\ j \geq 2$.

4 Ρυθμός συνωστισμού - Ευστάθεια

Το μέσο ποσό εργασίας που εισέρχεται προς διεκπεραίωση στο σύστημα ανά χρονική μονάδα αναφέρεται ως ο ρυθμός συνωστισμού ρ και ισούται με το γινόμενο του ρυθμού αφίξεων λ επί το μέσο χρόνο εξυπηρέτησης b:

$$\rho = \lambda b.$$

Το ποσό της εργασίας που μπορεί να διεκπεραιώσει το σύστημα ανά χρονική μονάδα είναι ίσο με το πλήθος των υπηρετών c αφού κάθε υπηρέτης μπορεί να διεκπεραιώσει μια μονάδα εργασίας ανά χρονική

μονάδα. Έτσι για να είναι το σύστημα ευσταθές και να να μην απειρίζεται η ουρά αναμένουμε διαισθητικά ότι θα πρέπει το μέσο ποσό εργασίας που εισέρχεται ανά χρονική μονάδα να είναι μικρότερο από τη μέγιστη δυνατότητα διεκπεραίωσης του συστήματος ανά χρονική μονάδα. Πράγματι, αποδεικνύεται ότι

Θεώρημα 1 - Ευστάθεια: *Στο $GI/G/c$ σύστημα με συνεχή κατανομή για τους ενδιάμεσους χρόνους μεταξύ των αφίξεων και/ή τους χρόνους εξυπηρέτησης ισχύει ένα ακριβώς από τα παρακάτω:*

1. *$\rho < c$ οπότε το σύστημα είναι ευσταθές, δηλαδή υπάρχουν οι κατανομές (p_j), (r_j) και (d_j) και είναι $p_j > 0,\ \ j \geq 0$ και $\sum_{j=0}^{\infty} p_j = 1$ (και όμοια για τις (r_j) και (d_j)).*

2. *$\rho \geq c$ οπότε το σύστημα είναι ασταθές, δηλαδή το πλήθος των πελατών απειρίζεται καθώς $t \to \infty$ και $p_j = r_j = d_j = 0,\ \ j \geq 0$.*

Η απόδειξη του αποτελέσματος αυτού είναι ιδιαίτερα περίπλοκη. Ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο βιβλίο των Baccelli and Bremaud (1994) όπου αποδεικνύονται θεωρήματα ευστάθειας και άλλα θεωρητικά αποτελέσματα κάτω από γενικές συνθήκες.

5 Ιδιότητα μεμονωμένων μεταβάσεων και ιδιότητα PASTA

Όπως είπαμε οι οριακές κατανομές (p_j), (r_j) και (d_j) γενικά δεν συμπίπτουν. Υπάρχουν όμως δυο περιπτώσεις στις οποίες κάποιες από αυτές συμπίπτουν. Συγκεκριμένα έχουμε τα ακόλουθα αποτελέσματα.

Θεώρημα 2 - Ιδιότητα μεμονωμένων μεταβάσεων: *Σε συστήματα εξυπηρέτησης στα οποία οι πελάτες έρχονται και αναχωρούν μεμονωμένα, δηλαδή δεν υπάρχουν ομαδικές αφίξεις ούτε ομαδικές αναχωρήσεις οι οριακές κατανομές σε στιγμές αφίξεων και σε στιγμές αναχωρήσεων συμπίπτουν: $(r_j) = (d_j)$.*

Θεώρημα 3 - Ιδιότητα Poisson Arrivals See Time Averages (PASTA): *Σε συστήματα εξυπηρέτησης στα οποία οι πελάτες έρχονται σύμφωνα με μια διαδικασία Poisson (δηλαδή οι ενδιάμεσοι χρόνοι μεταξύ των αφίξεων είναι ανεξάρτητοι και ισόνομοι με εκθετική κατανομή) οι οριακές κατανομές σε στιγμές αφίξεων και σε συνεχή χρόνο συμπίπτουν: $(r_j) = (p_j)$.*

Η διαισθητική αιτιολόγηση της ιδιότητα PASTA είναι ότι η διαδικασία Poisson μοντελοποιεί την ιδέα των εντελώς τυχαίων αφίξεων στο χρόνο και επομένως η παρατήρηση του αριθμού των πελατών κατά τη στιγμή της άφιξη ενός πελάτη στο σύστημα ισοδυναμεί με την παρατήρηση του αριθμού των πελατών σε μια τυχαία χρονική στιγμή (σε συνεχή χρόνο).

Η αυστηρή απόδειξη αυτών των αποτελεσμάτων κάτω από τόσο γενικές συνθήκες είναι αρκετά απαιτητική. Ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο βιβλίο των Baccelli and Bremaud (1994), όπου υπάρχουν και άλλα συναφή αποτελέσματα. Στο βιβλίο του Φακίνου (2003) υπάρχει μια περιγραφή αυτών των αποδεικτικών ιδεών.

Φυσικά μπορούμε να συνδυάσουμε τα δυο αυτά αποτελέσματα και τότε έχουμε ότι σε συστήματα που οι πελάτες έρχονται σύμφωνα με μια διαδικασία Poisson και έχουμε μεμονωμένες μεταβάσεις (αφίξεις, αναχωρήσεις) όλες οι οριακές κατανομές συμπίπτουν: $(p_j) = (r_j) = (d_j)$.

6 Ο νόμος του Little

Ο νόμος του Little είναι ένα πολύ γενικό αποτέλεσμα που συνδέει το μέσο πλήθος πελατών στο σύστημα $E[Q]$, το ρυθμό αφίξεων λ και το μέσο χρόνο παραμονής ενός πελάτη $E[S]$ σε αυτό. Συγκεκριμένα έχουμε

Θεώρημα 4 - Νόμος του Little: *Έστω ένα σύστημα εξυπηρέτησης με μέσο πλήθος πελατών $E[Q]$, ρυθμό αφίξεων λ και μέσο χρόνο παραμονής πελάτη $E[S]$. Τότε*

$$E[Q] = \lambda E[S].$$

Διαισθητικά, το αποτέλεσμα του Little μπορεί να γίνει κατανοητό θεωρώντας ότι κάθε πελάτης πληρώνει 1 χρηματική μονάδα ανά χρονική μονάδα παραμονής του στο σύστημα. Τότε ο διαχειριστής του συστήματος λαμβάνει $E[Q]$ χρηματικές μονάδες στη μονάδα του χρόνου, αν υποθέσουμε ότι η πληρωμή γίνεται κατά τρόπο "συνεχή". Από την άλλη μεριά η μέση είσπραξη του διαχειριστή στη μονάδα του χρόνου θα πρέπει να είναι η ίδια αν οι πελάτες πληρώνουν "προκαταβολικά", δηλαδή αν με την είσοδό τους στο σύστημα δίνουν όλο το ποσό για την παραμονή τους. Αλλά τότε θα έχουμε κατά μέσο όρο λ πελάτες ανά χρονική μονάδα και ο καθένας θα πληρώνει $E[S]$ χρηματικές μονάδες, οπότε η συνολική είσπραξη του διαχειριστή θα είναι $\lambda E[S]$ χρηματικές μονάδες στη μονάδα του χρόνου. Αφού τα δυο ποσά πρέπει να είναι ίσα (συνεχής είσπαραξη - προκαταβολική είσπραξη) έχουμε τη σχέση $E[Q] = \lambda E[S]$. Αυστηρές αποδείξεις του θεωρήματος Little έχουν γίνει κάτω από πολύ γενικές συνθήκες. Ο ενδιαφερόμενος αναγνώστης παραπέμπεται στις εργασίες των Little (1961) και Stidham (1974). Στο βιβλίο του Φακίνου (2003) υπάρχει ένα σκαρίφημα αυτών των αποδεικτικών ιδεών.

Το αποτέλεσμα του Little μπορεί να εφαρμοστεί και σε υποσυστήματα ενός συστήματος, δίνοντας ενδιαφέρονται αποτελέσματα. Στην περίπτωση αυτή το $E[Q]$ θα αναφέρεται στο μέσο πλήθος πελατών στο συγκεκριμένο υποσύστημα, το λ στο ρυθμό άφιξης στο συγκεκριμένο υποσύστημα και το $E[S]$ στο χρόνο παραμονής ενός πελάτη στο συγκεκριμένο υποσύστημα.

Θεωρώντας ως υποσύστημα το χώρο αναμονής ενός συστήματος (δηλαδή την ουρά) παίρνουμε τη σχέση

$$E[Q_q] = \lambda E[W],$$

δηλαδή ο μέσος αριθμός πελατών στην ουρά ισούται με το ρυθμό αφίξεων επί το μέσο χρόνο αναμονής ενός πελάτη μέχρι να αρχίσει η εξυπηρέτησή του.

Θεωρώντας ως υποσύστημα το χώρο εξυπηρέτησης ενός συστήματος παίρνουμε τη σχέση

$$E[Q_s] = \lambda E[X] = \lambda b = \rho,$$

δηλαδή ο μέσος αριθμός πελατών στο χώρο εξυπηρέτησης που προφανώς ταυτίζεται με τον μέσο αριθμό απασχολημένων υπηρετών ισούται με το ρυθμό συνωστισμού του συστήματος. Έτσι έχουμε μια δεύτερη ερμηνεία του ρυθμού συνωστισμού. Όχι μόνο είναι το μέσο ποσό εργασίας που εισέρχεται στο σύστημα ανά χρονική μονάδα αλλά επιπλέον εκφράζει και το μέσο αριθμό απασχολημένων υπηρετών μια τυχούσα χρονική στιγμή. Επειδή ο μέσος αριθμός απασχολημένων υπηρετών ισούται με το πλήθος c των υπηρετών επί την πιθανότητα ένας υπηρέτης να είναι απασχολημένος συμπεραίνουμε ότι

$$\text{Πιθανότητα απασχολημένου υπηρέτη=Ποσοστό του χρόνου απασχόλησης υπηρέτη} = \frac{\rho}{c}.$$

Ειδικά για την $GI/G/1$ ουρά έχουμε

$$\rho = E[Q_s] = 0\Pr[Q_s = 0] + 1\Pr[Q_s = 1] = \Pr[Q_s = 1] = \Pr[Q \geq 1] = 1 - p_0,$$

επομένως στην περίπτωση αυτή έχουμε

$$p_0 = \text{Πιθανότητα κενού συστήματος} = 1 - \rho.$$

Η ιδιότητα PASTA σε συνδυασμό με το θεώρημα Little μπορούν να χρησιμοποιηθούν για να υπολογίσουμε με πιθανοθεωρητικούς συλλογισμούς και ελάχιστους υπολογισμούς τα μέτρα απόδοσης $E[Q]$ και $E[S]$ για αρκετά συστήματα, χωρίς να χρειαστεί να υπολογίσουμε τις αντίστοιχες κατανομές. Η ανάλυση αυτή αναφέρεται συχνά ως ανάλυση μέσης τιμής (Mean Value Analysis - MVA) και είναι ένα ιδιαίτερα ισχυρό εργαλείο.