Δ.Π.Μ.Σ Μαθηματικά της Αγοράς και της Παραγωγής

Μαθηματικά Υποδείγματα Παραγωγής, Εφοδιαστικής και Υπηρεσιών ΙΙ - Ουρές Αναμονής ΜΑΘΗΜΑ 5ο

Γιάννης Δημητρακόπουλος, PhD in OR
Τμήμα Μαθηματικών
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Περιεχόμενα Μαθήματος

1. Μοντέλα Συστημάτων Εξυπηρέτησης
- Τί είναι Σύστημα Εξυπηρέτησης;
- Πού υπεισέρχεται η αβεβαιότητα;
- Η έννοια της καθυστέρησης (αναμονής)
- Στοχαστική Μοντελοποίηση
- Βασικές έννοιες
- Βασικά Μέτρα Απόδοσης

2. Μαρκοβιανές Αλυσίδες Συνεχούς Χρόνου
- Ορισμός και Βασικές Έννοιες
- Ένα ισοδύναμο μοντέλο – Στασιμότητα
- Διαδικασίες Γέννησης - Θανάτου

3. Απλές Μαρκοβιανές Ουρές
- M/M/1
- Τροποποιήσεις της M/M/1
- M/M/s
- M/M/s/k (για s=1)
- Πεπερασμένος Πληθυσμός Πελατών

4. Μοντέλα με Γενικούς Χρόνους Εξυπηρέτησης
- Η M/G/1 Ουρά
- Τύπος Pollaczek- Khintchine
- Εφαρμογές

5. Μοντέλα με Προτεραιότητες
- Preemptive και non-Preemptive Μοντέλα

6. Δίκτυα Ουρών
- Ουρές σε Σειρά
- Ανοικτά Δίκτυα Jackson
- Equivalence property και product form – Εξισώσεις Κίνησης

Ουρά Αναμονής

A/B/s/k (queue disc)

A Κατανομή διαδ. αφίξεων

B Κατανομή διαδ. Εξυπηρετήσεων

$s =$ πλήθος παράλληλων υπηρετών

$k =$χωρητικότητα συστήματος (in queue $+s$)

Απλές Μαρκοβιανές Ουρές

M/M/s (M=Εκθετικούς Χρόνους)

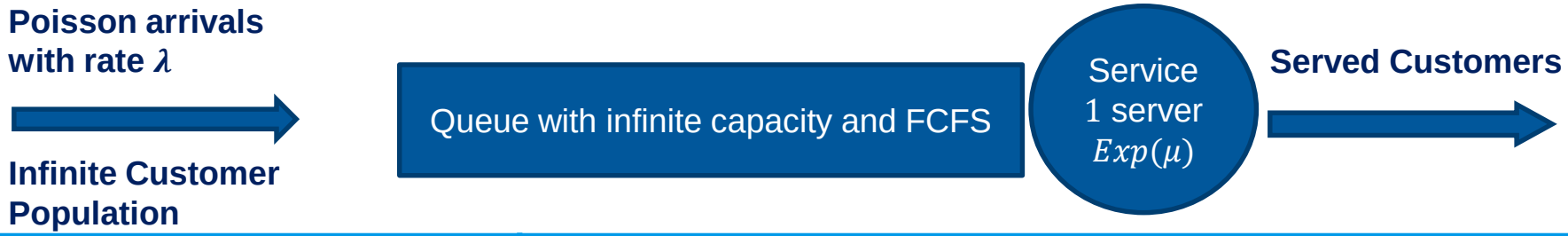
Διαδικασία Αφίξεων Poisson ρυθμού λ

Χρόνοι εξυπηρετησης i.i.d. $\sim Exp(\mu), E(\boldsymbol{W_s}) = \frac{1}{\mu}$

s παράλληλοι υπηρέτες - άπειρη χωρητικότητα

Πειθαρχία ουρά FCFS - $\rho = \frac{\lambda}{s\mu} < 1$ (Ευστάθεια)

Η M/M/1 Ουρά



Βασικά Χαρακτηριστικά

Διαδικασία Αφίξεων Poisson ρυθμού λ

Χρόνοι εξυπηρετησης i.i.d. $\sim Exp(\mu)$,

$s = 1$ υπηρέτης

Άπειρη χωρητικότητα ($k = \infty$).

Εισέρχονται όλοι στο σύστημα

Πειθαρχία ουρά FCFS

$\rho = \frac{\lambda}{\mu} < 1$ (Ευστάθεια)

Βασικά Αποτελέσματα: Αν $\lambda < \mu$

Στάσιμη Κατανομή Πελατών $N \sim Geom(\rho)$ στο N_0

$$\pi_n = P(N = n) = \rho^n(1-\rho), \rho = \frac{\lambda}{\mu}$$

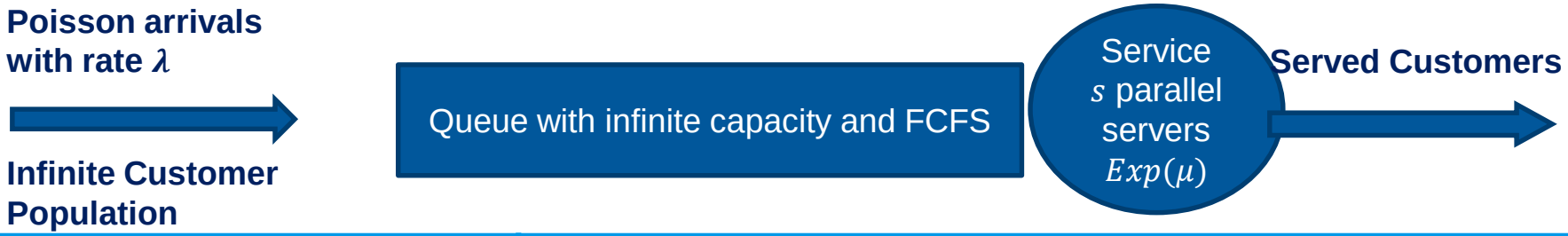
Ποσοστό Χρόνου σύστημα άδειο $\pi_0 = 1 - \rho$

Μέσο Πλήθος Πελατών $L = \frac{\lambda}{\mu-\lambda}$ (σύστημα), $L_q = \frac{\lambda\rho}{\mu-\lambda}$ (αναμονή)

Μέσος Χρόνος Παραμονής $W = \frac{1}{\mu-\lambda}$ (σύστημα), $\mathrm{W_q}\ \frac{\rho}{\mu-\lambda}$ (αναμονή)

Κατανομή Χρόνου Παραμονής $\boldsymbol{W} \sim Exp\big((1-\rho)\mu\big)$

Η M/M/s Ουρά



Βασικά Χαρακτηριστικά

Διαδικασία Αφίξεων Poisson ρυθμού λ

Χρόνοι εξυπηρετησης i.i.d. $\sim Exp(\mu)$,

s παράλληλοι υπηρέτες

Άπειρη χωρητικότητα ($k = \infty$).

Εισέρχονται όλοι στο σύστημα

Πειθαρχία ουρά FCFS

$\rho = \frac{\lambda}{s\mu} < 1$ (Ευστάθεια)

Βασικά Αποτελέσματα: Αν $\lambda < s\mu$

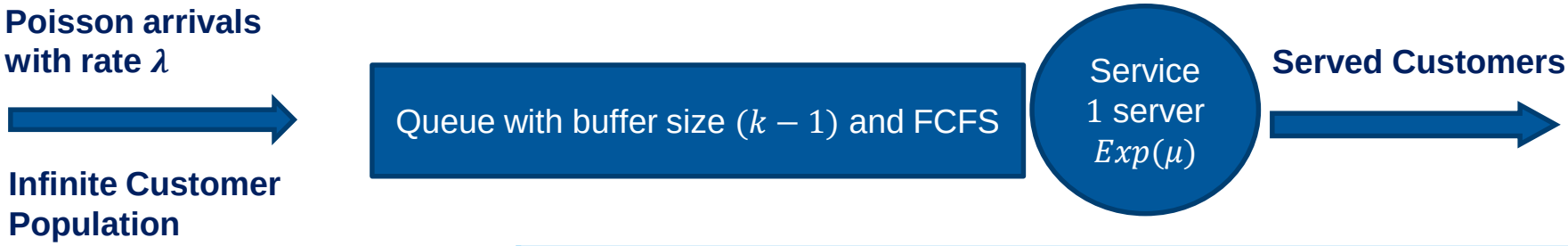
Στάσιμη Κατανομή Πελατών

$\pi_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}\pi_0, 0 \le n \le s-1$ και $\pi_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{s!s^{n-s}}\pi_0, n \ge s$

Ποσοστό Χρόνου σύστημα άδειο $\pi_0 = \left(\sum_{n=0}^{s-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!(1-p)}\right)^{-1}$

- $L_q = \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!}\pi_0\frac{\rho}{(1-\rho)^2}$, $W_q = \frac{L_q}{\lambda}$ (από Little)
- $W = W_q + \frac{1}{\mu}, \mathrm{L} = \lambda\mathrm{W} = \lambda\left(W_q + \frac{1}{\mu}\right) = \mathrm{L_q} + \frac{\lambda}{\mu}$

Η M/M/1/k Ουρά



Βασικά Χαρακτηριστικά

Διαδικασία Αφίξεων Poisson ρυθμού λ

Χρόνοι εξυπηρετησης i.i.d. $\sim Exp(\mu)$,

1 υπηρέτης. Πειθαρχία ουρά FCFS

Πεπερασμένη χωρητικότητα $(k < \infty)$.

Δεν εισέρχονται όλοι στο σύστημα

1 θέση εξυπ. $k-1$ θέσεις στην αναμ.

Ευσταθές για κάθε λ, μ.

Βασικά Αποτελέσματα: για $\rho \neq 1$

Στάσιμη Κατανομή Πελατών

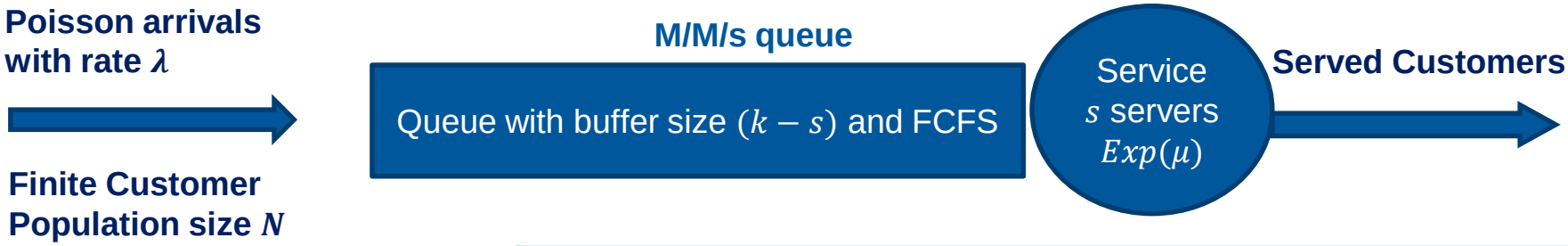
$\pi_0 = \frac{1-\rho}{1-\rho^{k+1}}$ και $\pi_n = \rho^n \frac{1-\rho}{1-\rho^{k+1}}, 1 \leq n \leq \mathrm{k}$

Ρυθμός Πραγματικών Αφίξεων $\bar{\lambda} = \lambda(1-\pi_k)$

$L = E(N) = \frac{\rho}{1-\rho} - \frac{(\mathrm{k}+1)\rho^{\mathrm{k}+1}}{1-\rho^{k+1}}$ και $W = \frac{L}{\bar{\lambda}} = \frac{\frac{\rho}{1-\rho} - \frac{(\mathrm{k}+1)\rho^{\mathrm{k}+1}}{1-\rho^{k+1}}}{\lambda(1-\pi_k)}$ (από Little)

$$L_q = \mathrm{L} + \rho = \frac{\rho}{1-\rho} - \frac{(\mathrm{k}+1)\rho^{\mathrm{k}+1}}{1-\rho^{k+1}} - 1 + \frac{1-\rho}{1-\rho^{k+1}}, W_q = \frac{L_q}{\bar{\lambda}}$$

Πεπερασμένος Πληθυσμός Πελατών



Βασικά Χαρακτηριστικά

Πηγή Εισόδου Μεγέθους N – M/M/s

Διαδικασία Αφίξεων Poisson ρυθμού λ

Χρόνοι εξυπηρετησης i.i.d. $\sim Exp(\mu)$,

1 υπηρέτης. Πειθαρχία ουρά FCFS

Άπειρη χωρητικότητα $(k = \infty)$.

Εισέρχονται όλοι στο σύστημα

Ευσταθές για κάθε λ, μ. λόγω πεπερασμένου πλήθους καταστάσεων

Βασικά Αποτελέσματα: για $s = 1$

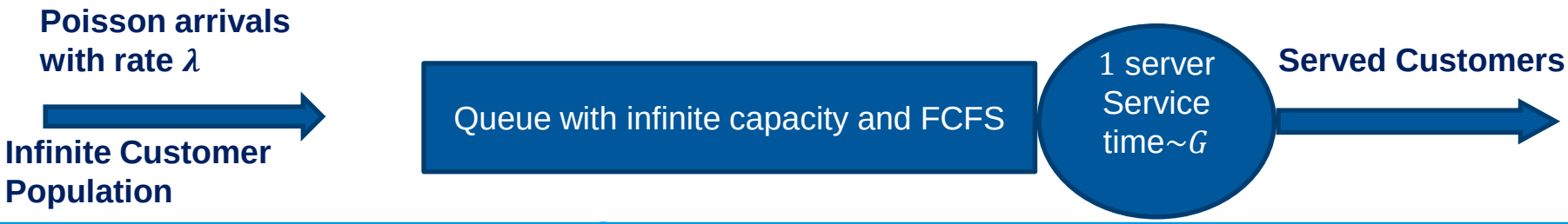
Στάσιμη Κατανομή Πελατών

$$\pi_0 = \left(\sum_{n=0}^{N} \frac{\left(\frac{\lambda}{\mu}\right)^n \mathrm{N!}}{(\mathrm{N-n})!} \right)^{-1} \text{ και } \pi_n = \frac{\left(\frac{\lambda}{\mu}\right)^n \mathrm{N!}}{(\mathrm{N-n})!}, 1 \le n \le \mathrm{N}$$

Ρυθμός Πραγματικών Αφίξεων $\bar{\lambda} = \lambda(\mathrm{N} - L)$

- $L_q = N - \frac{\mu}{\lambda}(1 - \pi_0)$, $W_q = \frac{L_q}{\lambda}$ (από Little)
- $L = L_q + \rho = N - \frac{\mu}{\lambda}(1 - \pi_0) + \frac{\mu}{\lambda}$, $W = \frac{L}{\lambda}$ (από Little)

Η M/G/1 Ουρά – Non Markovian Model



Βασικά Χαρακτηριστικά

Διαδικασία Αφίξεων Poisson ρυθμού λ

Χρόνοι εξυπηρετησης i.i.d. $\sim G, \mathrm{E}[\mathrm{W_s}] = \frac{1}{\mu}$

$s = 1$ υπηρέτης. Άπειρη χωρητικότητα.

Εισέρχονται όλοι στο σύστημα.

Πειθαρχία ουρά FCFS.

$\rho = \lambda \cdot \frac{1}{\mu} < 1$ (Ευστάθεια)

Βασικά Αποτελέσματα: Αν $\rho < 1$

Στάσιμη Κατανομή Πελατών $N \sim \pi_n$ δύσκολος υπολ.

Ποσοστό Χρόνου σύστημα άδειο $\pi_0 = 1 - \rho$

Μέσο Πλήθος Πελατών στην αναμονή (Τύπος P-K)

$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}, \sigma^2 = \mathrm{Var}(\mathrm{W_s})$$

Μέσος Χρόνος Αναμονής $W_q = \frac{L_q}{\lambda}$

Μέσος Χρόνος Παραμονής $\mathrm{W} = \mathrm{W_q} + \frac{1}{\mu}$

Μέσο Πλήθος Πελατών $\mathrm{L} = \lambda W$

Μοντέλα Προτεραιοτήτων (Queues with Priorities)

Βασικά Χαρακτηριστικά

Διαδικασία Αφίξεων Poisson ρυθμού Λ.

Δεν ισχύει η πειθαρχία FCFS για το σύνολο των πελατών. s παράλληλοι υπηρέτες.

Οι πελάτες έχουν βαθμό προτεραιότητας και εξυπηρετούνται με βάση αυτόν.

Κλάσεις Προτεραιότητας $i = 1, \dots, N$, $i = 1$ υψηλότερη, $i = N$ χαμηλότερη.

Απαιτήσεις εξυπηρέτησης του πελάτη της κλάσης i απαιτούν χρόνο $Exp(\mu_i)$.

Δύο είδη προτεραιοτήτων: Non-preemptive και preemptive priorities.

Non-preemptive: Εάν έλθει πελάτης υψηλότερης προτεραιότητας από αυτούς που ήδη εξυπηρετούνται, τότε αναμένει την πρώτη ολοκλήρωση εξυπηρέτησης για να ξεκινήσει την εξυπηρέτησή του.

Preemptive: Ο πελάτης υψηλότερης προτεραιότητας διακόπτει την εξυπηρέτηση οποιουδήποτε πελάτη της χαμηλότερης δυνατής προτεραιότητας (δεν έχει σημασία ποιού), ο οποίος μπαίνει πρώτος στον αντίστοιχό του χώρο αναμονής. (resume or repeat)

Πρακτικές Εφαρμογές: ΤΕΠ Νοσοκομείου, Private Banking, First and Coach Class in Airlines

Μοντέλα Προτεραιοτήτων (Queues with Priorities)

Βασικό Μοντέλο

Διαδικασία Αφίξεων Poisson ρυθμού Λ.

s παράλληλοι υπηρέτες.

N Κλάσεις Προτεραιότητας Πελατών $i = 1, \dots, N$, $i = 1$ υψηλότερη, $i = N$ χαμηλότερη.

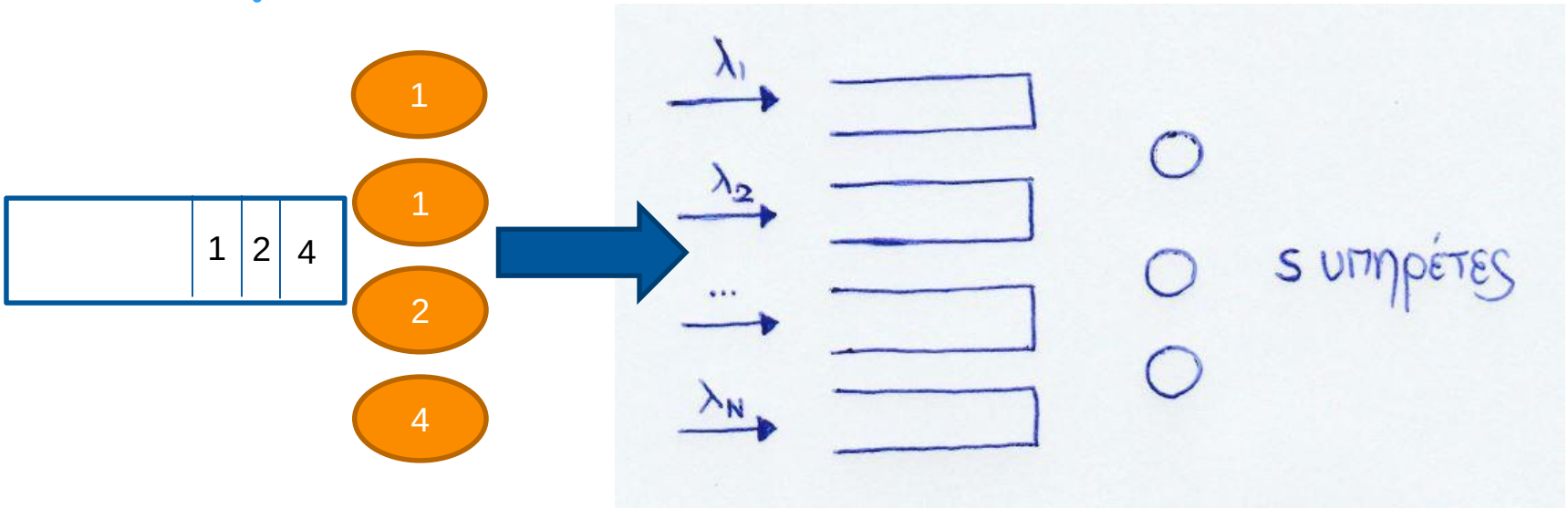
Κάθε φορά που ελευθερώνεται υπηρέτης, εξυπηρετεί τον επόμενο πελάτη με την υψηλότερη προτεραιότητα. Πελάτες της ίδιας κλάσης εξυπηρετούνται με FCFS.

Για κάθε κλάση πελατών k θεωρούμε:

- Διαδικασία αφίξεων πελατών της κλάσης$-k$ Poisson ρυθμού λ_k.
- Απαιτήσεις εξυπηρέτησης του πελάτη της κλάσης i απαιτούν χρόνο $Exp(\mu_i)$.

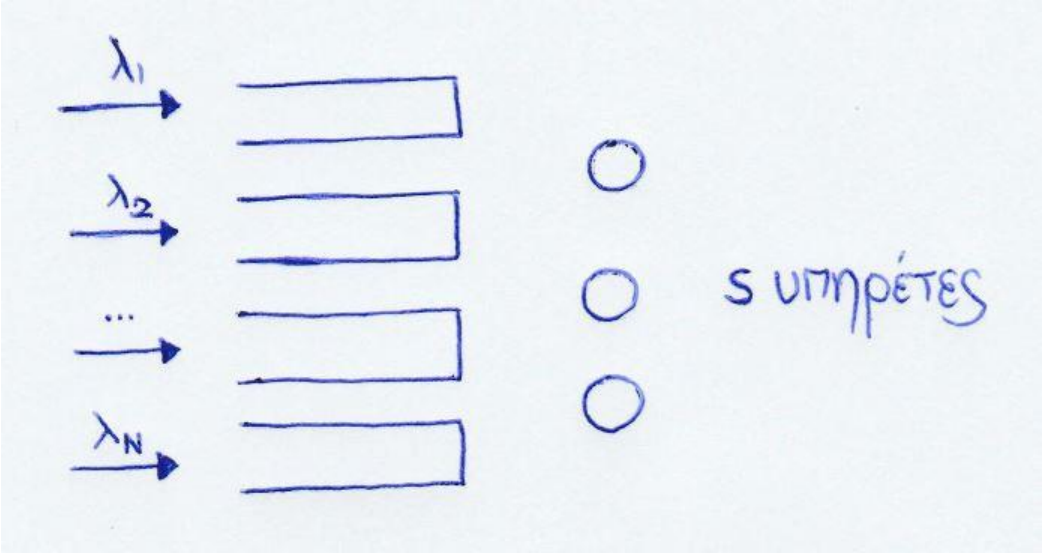
ΙΔΕΑ Αν $\mu_k = \mu$ τότε το συνολικό σύστημα είναι M/M/s (χωρίς να λάβω υπόψη την προτεραιότητα)

Ένα ισοδύναμο Μοντέλο



- Οι πελάτες διαμορφώνουν N παράλληλες ουρές σύμφωνα με τη κλάση προτεραιότητας.
- $i = 1$ υψηλότερη, $i = N$ χαμηλότερη.
- λ_i ο ρυθμός αφίξεων της κλάσης i και $\lambda_1 + \lambda_2 + \cdots + \lambda_N = \Lambda$.
- Σε κάθε ουρά πειθαρχία FCFS.
- Χρόνοι Εξυπηρέτησης στην ουρά $O_i \sim Exp(\mu_i)$ από απασχολημένο υπηρέτη.
- Οι ελεύθεροι υπηρέτες ανατίθενται στις ουρές $O_1, O_2, \ldots, O_N$.

Non-Preemptive Priorities



Άφιξη πελάτη υψηλότερης προτεραιότητας που βρίσκει τους υπηρέτες απασχολημένους περιμένει την ολοκλήρωση της εξυπηρέτησης και την απελευθέρωση υπηρέτη για να εξυπηρετηθεί.

Μελέτη δύο βασικών περιπτώσεων: (NP-A) s υπηρέτες και $\mu_i = \mu$, (NP-B) $s = 1$ και μ_i διαφ.

NP-A: s υπηρέτες και $\mu_i = \mu$ για κάθε $i = 1, \dots, N$

- Το συνολικό σύστημα (χωρίς τους βαθμούς προτεραιότητας) είναι **M|M|s**
- Συνολικός ρυθμός άφιξης $\Lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_N$
- Χρόνοι Εξυπηρέτησης i.i.d. $\sim \mathrm{Exp}(\mu)$. Οι διαφορετικές κλάσεις έχουν τις ίδιες απαιτήσεις εξυπηρέτησης
- Το συνολικό σύστημα είναι ευσταθές ανν $\frac{\Lambda}{s\mu} < 1$.
- Εάν το συνολικό σύστημα είναι ευσταθές, τότε προφανώς και τα επιμέρους υποσυστήματα εξυπηρέτησης που διαμορφώνουν οι πελάτες των διαφορετικών κλάσεων είναι επίσης ευσταθή.
- Επιμέρους Υποσυστήματα: Μελέτη σε μία ουρά οι $\{O_1\}, \{O_1, O_2\}, \dots, \{O_1, O_2, \dots, O_N\}$.

Ρυθμοί Άφιξης: $\lambda_1, \quad \lambda_1 + \lambda_2, \ \dots, \ \lambda_1 + \lambda_2 + \cdots \lambda_N = \Lambda$

NP-A: s υπηρέτες και $\mu_i = \mu$ για κάθε $i = 1, \dots, N$

Μέτρα Απόδοσης

Αν $s = 1$, τότε $A = \frac{\mu^2}{\lambda}$

Υπολογίζουμε τη σταθερά $A = s! \frac{s\mu - \Lambda}{r^s} \sum_{j=0}^{s-1} \frac{r^j}{j!} + s\mu$, με $r = \Lambda/\mu$

και την ακολουθία $\{B_k\}_{k \in \{0,1,2,\dots,N\}}$ με

$$B_0 = 1 \text{ και } B_k = 1 - \frac{1}{s\mu} \sum_{i=1}^{k} \lambda_i \,, \mathrm{k} = 1,2, \dots, \mathrm{N}$$

Μέσος Χρόνος Παραμονής ενός πελάτη της O_k: $W_k = \frac{1}{AB_{k-1}B_k} + \frac{1}{\mu}, k = 1,2,, \dots, N.$

Για τα υπόλοιπα μέτρα απόδοσης της εφαρμόζουμε τους γνωστούς σχετικούς O_k:

$$L_k = \lambda_k W_k, W_{q_k} = W_k - \frac{1}{\mu}, L_{q_k} = \lambda_k W_{q_k}, k = 1,2,, \dots, N$$

Για το συνολικό σύστημα:

$L_{\text{συνολικό}} = L_1 + L_2 + \cdots + L_N$, και $\Lambda W_{\text{συνολικό}} = \lambda_1 W_1 + \lambda_2 W_2 + \cdots + \lambda_N W_N$ (από Νόμο Little)

NP-B: $s = 1$ υπηρέτης και μ_i για $i = 1, \dots, N$

- Συνολικός ρυθμός άφιξης $\Lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_N$
- Χρόνοι Εξυπηρέτησης i.i.d. $\sim \mathrm{Exp}(\mu_i)$
- Το συνολικό σύστημα είναι ευσταθές ανν $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} + \cdots + \frac{\lambda_N}{\mu_N} < 1.$
- Εάν το συνολικό σύστημα είναι ευσταθές, τότε προφανώς και τα επιμέρους υποσυστήματα εξυπηρέτησης που διαμορφώνουν οι πελάτες των διαφορετικών κλάσεων είναι επίσης ευσταθή, καθώς η O_k ευσταθής ανν $\sum_{i=1}^{k} \frac{\lambda_i}{\mu_i} < 1$

NP-B: $s = 1$ υπηρέτες και μ_i για $i = 1, \dots, N$

Μέτρα Απόδοσης

Υπολογίζουμε τις ακολουθίες $\{\alpha_k\}, \{b_k\}$, με

$$a_k = \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i^2}, b_0 = 1 \text{ και } b_k = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i}, \text{k} = 1,2, \dots, \text{N}$$

Μέσος Χρόνος Παραμονής ενός πελάτη της O_k: $W_k = \frac{a_k}{b_{k-1} b_k} + \frac{1}{\mu_k}, k = 1,2,, \dots, N.$

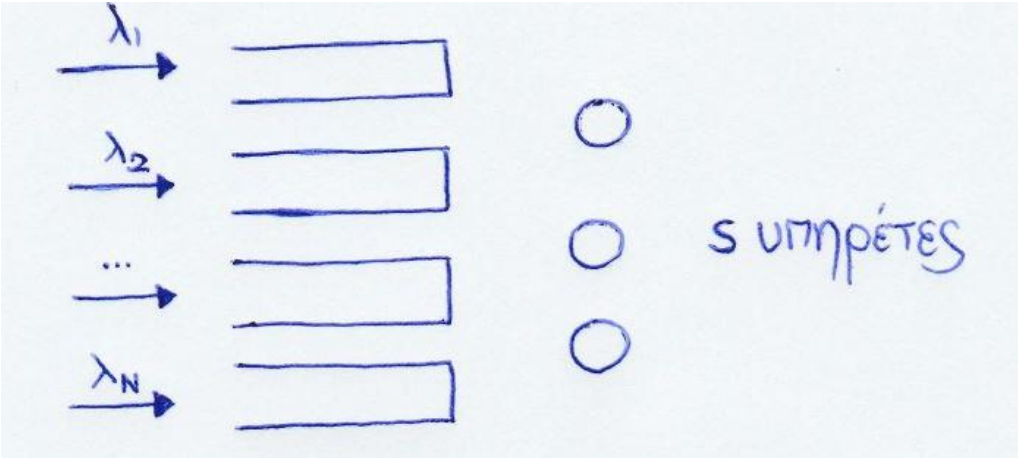
Για τα υπόλοιπα μέτρα απόδοσης της εφαρμόζουμε τους γνωστούς σχετικούς O_k:

$$L_k = \lambda_k W_k, W_{q_k} = W_k - \frac{1}{\mu}, L_{q_k} = \lambda_k W_{q_k}, k = 1,2,, \dots, N$$

Για το συνολικό σύστημα:

$L_{\text{συνολικό}} = L_1 + L_2 + \cdots + L_N$, και $\Lambda W_{\text{συνολικό}} = \lambda_1 W_1 + \lambda_2 W_2 + \cdots + \lambda_N W_N$ (από Νόμο Little)

Preemptive Priorities



Άφιξη πελάτη υψηλότερης προτεραιότητας που βρίσκει τους υπηρέτες απασχολημένους σταματάει την εξυπηρέτηση του πελάτη χαμηλότερης προτεραιότητας, τον μετακινεί πρώτο στην κλάση του, και ξεκινάει τη δική του εξυπηρέτηση.

Σημαντική Παρατήρηση: Στα preemptions δεν επηρεάζεται η διαδικασία εξυπηρέτησης πελάτη που διακόπτεται η εξυπηρέτησή του, λόγω της αμνήμονης ιδιότητας της Exp.

Μελέτη δύο βασικών περιπτώσεων: (PR-A) $s = 1$ και $\mu_i = \mu$, (PR-B) $s > 1$ και $\mu_i = \mu$ διαφ.

PR-A: $s = 1$ υπηρέτης και $\mu_i = \mu$ για κάθε $i = 1, \dots, N$

Μέτρα Απόδοσης

Υπολογίζουμε την ακολουθία $\{B_k\}_{k\in\{0,1,2,\dots,N\}}$ με

$$B_0 = 1 \text{ και } B_k = 1 - \frac{1}{\mu}\sum_{i=1}^{k}\lambda_i \text{ , k} = 1,2, \dots, \text{N}$$

Μέσος Χρόνος Παραμονής ενός πελάτη της O_k: $W_k = \frac{\frac{1}{\mu}}{B_{k-1}B_k}, k = 1,2,, \dots, N.$

Για τα υπόλοιπα μέτρα απόδοσης της εφαρμόζουμε τους γνωστούς σχετικούς O_k:

$$L_k = \lambda_k W_k, W_{q_k} = W_k - \frac{1}{\mu}, L_{q_k} = \lambda_k W_{q_k}, k = 1,2,, \dots, N$$

Για το συνολικό σύστημα:

$L_{\text{συνολικό}} = L_1 + L_2 + \cdots + L_N$, και $\lambda W_{\text{συνολικό}} = \lambda_1 W_1 + \lambda_2 W_2 + \cdots + \lambda_N W_N$ (από Νόμο Little)

PR-B: $s > 1$ υπηρέτες και $\mu_i = \mu$ για κάθε $i = 1, \dots, N$

Μέτρα Απόδοσης

Δεν υπάρχουν κλειστοί τύποι για τον υπολογισμό του μέσου χρόνου παραμονής ενός πελάτη της κλάσης $-k$ στο σύστημα. Χρησιμοποιώ ένα αναδρομικό σχήμα.

Βασική Εφαρμογή: Στα ΤΕΠ του Νοσοκομείου Αιγίου εργάζονται στη βάρδια δύο γιατροί, με χρόνο εξέτασης που ακολουθεί $Exp(3)$ για όλους τους ασθενείς. Οι ασθενείς φθάνουν σύμφωνα με διαδικασία Poisson ρυθμού $\lambda = 2$ και ταξινομούνται σύμφωνα με την κατάστασή τους σε κρίσιμα με πιθανότητα $\mathrm{q}_1 = 0.1$, σοβαρά με πιθανότητα $q_2 = 0.3$ και σταθερά περιστατικά με πιθανότητα $q_3 = 0.6$. Οι ασθενείς εξυπηρετούνται με προτεραιότητες σύμφωνα με preemptive πειθαρχία ουράς, όπου τα κρίσιμα περιστατικά έχουν την υψηλότερη προτεραιότητα.

Για την επίλυση θεωρώ τα συστήματα των ουρών $\{\mathrm{O}_1\}, \{\mathrm{O}_1, \mathrm{O}_2\}, \dots, \{\mathrm{O}_1, \mathrm{O}_2, \dots, \mathrm{O}_\mathrm{N}\}$.

Ρυθμοί Άφιξης: $\lambda_1, \quad \lambda_1 + \lambda_2, \ \dots, \ \lambda_1 + \lambda_2 + \cdots \lambda_\mathrm{N} = \Lambda$

Ουρές με Προτεραιότητες - Ενδεικτική Βιβλιογραφία

- **Hillier and Lieberman, “Introduction to Operations Research”, 8th edition** Chapter 17 Queueing Theory, par. 8, p.p. 804-809.