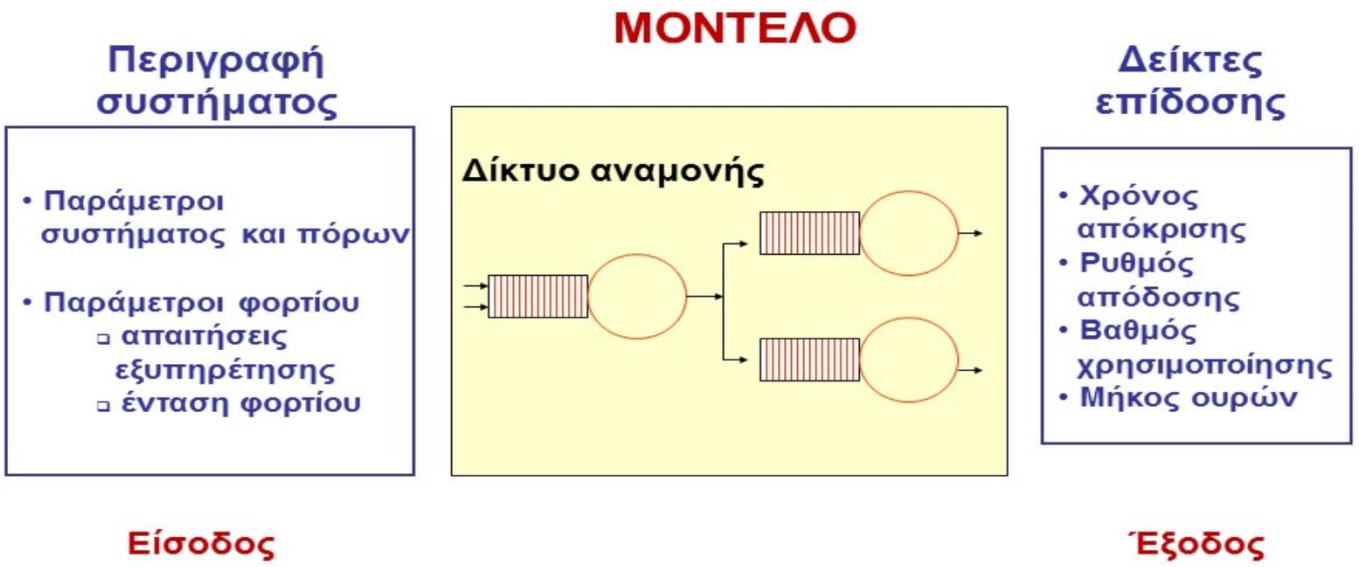
ΣΥΣΤΗΜΑΤΑ ΑΝΑΜΟΝΗΣ
(Queuing Systems)

Παράμετροι Ουρών και Συστημάτων Αναμονής

ΜΟΝΤΕΛΟ ΣΥΣΤΗΜΑΤΟΣ – ΔΙΚΤΥΟ ΑΝΑΜΟΝΗΣ (Επανάληψη)



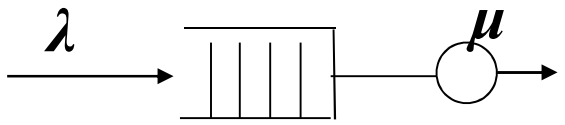
- Κυκλοφοριακή κίνηση
- Ουρές σε δρομολογητές, καταστήματα, ταχυδρομεία, τράπεζες
  - Πολλαπλοί εξυπηρετητές (servers)
  - Κοινή ουρά ή παράλληλες ουρές, προτεραιότητες
- Τηλεφωνικά κέντρα (πολλαπλοί εξυπηρετητές)
- Κόμβοι δικτύων τύπου Internet
- Πόροι υπολογιστικών συστημάτων (CPU, Μνήμη, Δίσκοι)

ΜΟΝΤΕΛΑ ΣΥΜΦΟΡΗΣΗΣ (Congestion)

- Κυκλοφοριακή κίνηση
- Ουρές σε καταστήματα, ταχυδρομεία, τράπεζες
  - Πολλαπλοί εξυπηρετητές (servers)
  - Κοινή ουρά ή παράλληλες ουρές, προτεραιότητες
- Τηλεφωνικά κέντρα (πολλαπλοί εξυπηρετητές)
- Κόμβοι δικτύων τύπου Internet
- Πόροι υπολογιστικών συστημάτων (CPU, Μνήμη, Δίσκοι)

ΚΟΙΝΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ (1/2)

- **Πελάτης:** Πελάτης τράπεζας, τηλεφωνική κλήση, πακέτο δεδομένων Internet…
- Εξυπηρετητής (**server**): Ταμίας, τηλεπικοινωνιακός πόρος (γραμμή) αφιερωμένος σε τηλεφωνική κλήση ή προώθηση πακέτου…
- Τυχαία είσοδος πελατών – «γεννήσεις», μέσος ρυθμός αφίξεων: λ πελάτες/sec
- Χρόνος μεταξύ δύο διαδοχικών αφίξεων - τυχαία μεταβλητή a, μέσος όρος: $\mathrm{E}(a) = 1/\lambda$ sec
- Μέσος ρυθμός εξυπηρέτησης πελατών: μ πελάτες/sec
- Χρόνος εξυπηρέτησης πελάτη – τυχαία μεταβλητή s, μέσος όρος: $\mathrm{E}(s) = 1/\mu$ sec/πελάτη



ΚΟΙΝΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ (2/2)

- Ουρά αναμονής (**queue**) για εξομάλυνση στατιστικών μεταβολών και απομόνωση (**buffering**) διακυμάνσεων εισόδου – εξυπηρέτησης
- Χωρητικότητα συστήματος αποθήκευσης (**queue size**) συμπεριλαμβανομένων των πελατών υπό εξυπηρέτηση
- Αριθμός εξυπηρετητών
- Πρωτόκολλο εξυπηρέτησης: First Come First Served - **FCFS** ή First In First Out - **FIFO**, Last In First Out - **LIFO**, Processor Sharing, προτεραιότητες
- **Κατάσταση συστήματος** $n(t)$: Αριθμός πελατών στο σύστημα αναμονής (ουρά + εξυπηρέτηση) σε μια χρονική στιγμή. Χρονοσειρά - time series - ή στοχαστική ανέλιξη - stochastic process - διακριτής κατάστασης & συνεχούς χρόνου
- Δρομολόγηση από ουρά σε ουρά σε περιπτώσεις δικτύων ουρών αναμονής

ΠΑΡΑΔΕΙΓΜΑΤΑ ΠΑΡΑΜΕΤΡΩΝ ΣΥΣΤΗΜΑΤΩΝ ΑΝΑΜΟΝΗΣ

- **Στοιχεία καθυστέρησης σε ένα σύστημα:** χρόνος επεξεργασίας, χρόνος αναμονής, χρόνος διάδοσης, χρόνος μετάδοσης
- **Δίκτυο μεταγωγής κυκλωμάτων (circuit switching):** ρυθμός αφίξεων κλήσεων, διάρκεια κλήσεων, ποσοστό απόρριψης κλήσεων
- **Δίκτυο μεταγωγής πακέτων (packet switching):** ρυθμός αφίξεων πακέτων, μέγεθος πακέτων, ποσοστό απόρριψης πακέτων, καθυστέρηση σε κόμβους του Internet
- **Υπολογιστικό σύστημα πολυεπεξεργασίας (windows):** αριθμός παράλληλων εντολών/προγραμμάτων υπό επεξεργασία, χρόνος ύπνωσης (sleeping time) ανά ενεργό παράθυρο, χρόνος αναζήτησης/ανταλλαγής δεδομένων στη μνήμη (I/O time), μέσος ρυθμός διεκπεραίωσης εντολών (ρυθμαπόδοση - throughput), χρόνος απόκρισης

Στοιχεία καθυστέρησης σε ένα σύστημα

- Processing Delay (χρόνος επεξεργασίας) is the time associated with the system analyzing a packet header and determining where the packet must be sent. This depends heavily on the entries in the routing table, the execution of data structures in the system, and the hardware implementation.
- Queueing Delay (χρόνος αναμονής) is the time between a packet being queued and it being sent. This varies depending on the amount of traffic, the type of traffic, and what router queue algorithms are implemented.
- Transmission Delay (χρόνος μετάδοσης) is the time needed to push a packet's data bits into the wire. This changes based on the size of the packet and the bandwidth. This does not depend on the distance of the wire, as it is solely the time to push a packet's bits into the wire, not to travel down the wire to the receiving endpoint.
- Propagation Delay (χρόνος διάδοσης) is the time associated with the first bit of the packet traveling from the sending endpoint to the receiving endpoint. This is often referred to as a delay by distance, and as such is influenced by the distance the bit must travel and the propagation speed.

ΠΑΡΑΜΕΤΡΟΙ (1/4)

– **Ένταση φορτίου (traffic intensity)**

Σε περίπτωση 1 ουράς, 1 εξυπηρετητή:

{Μέσος ***Χρόνος εξυπηρέτησης***}/ {Μέσος ***Χρόνος μεταξύ διαδοχικών αφίξεων***}

$$\rho \triangleq \frac{\left(\frac{1}{\mu}\right)}{\left(\frac{1}{\lambda}\right)} = \lambda E(s) = \lambda/\mu \quad \text{(Erlangs)}$$

Ένα ***Erlang*** αντιπροσωπεύει το φόρτο κυκλοφορίας που εξυπηρετείται από έναν εξυπηρετητή που ασχολείται το 100% του χρόνου (π.χ. 1 call-minute per minute). Ένας εξυπηρετητής ασχολείται για 30 λεπτά σε μια περίοδο μιας ώρας → μεταφέρει 0.5 Erlangs κυκλοφοριακή ένταση

– **Διεκπεραίωση πελατών – Ρυθμαπόδοση (Throughput)**

γ πελάτες/sec

Σε περίπτωση 1 ουράς, 1 εξυπηρετητή:

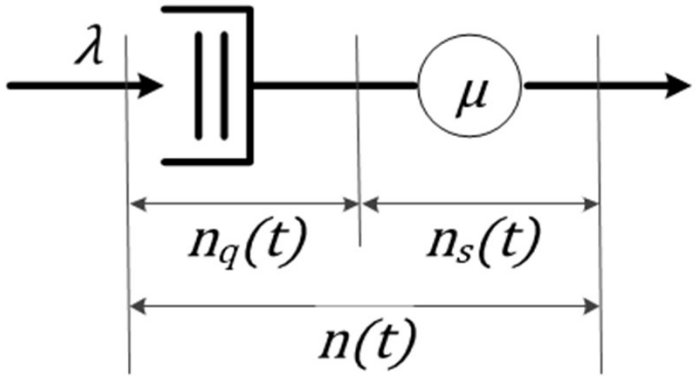
$$\gamma = \lambda(1 - \mathrm{P}\{\text{blocking}\}) \leq \lambda, \qquad \gamma < \mu$$

όπου $\mathrm{P}\{\text{blocking}\}$ είναι η πιθανότητα να χαθεί ένας πελάτης επειδή βρήκε το σύστημα πλήρες

– σε τηλεφωνικά δίκτυα: βαθμός ποιότητας, **Grade of Service - GoS**
– σε δίκτυα δεδομένων: μία παράμετρος ποιότητας υπηρεσίας, **Quality of Service - QoS**

ΠΑΡΑΜΕΤΡΟΙ (2/4)

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



- **Μέσος ρυθμός απωλειών, ποσοστό απωλειών, πιθανότητα απώλειας πελάτη**
  - Σε περίπτωση 1 ουράς, 1 εξυπηρετητή

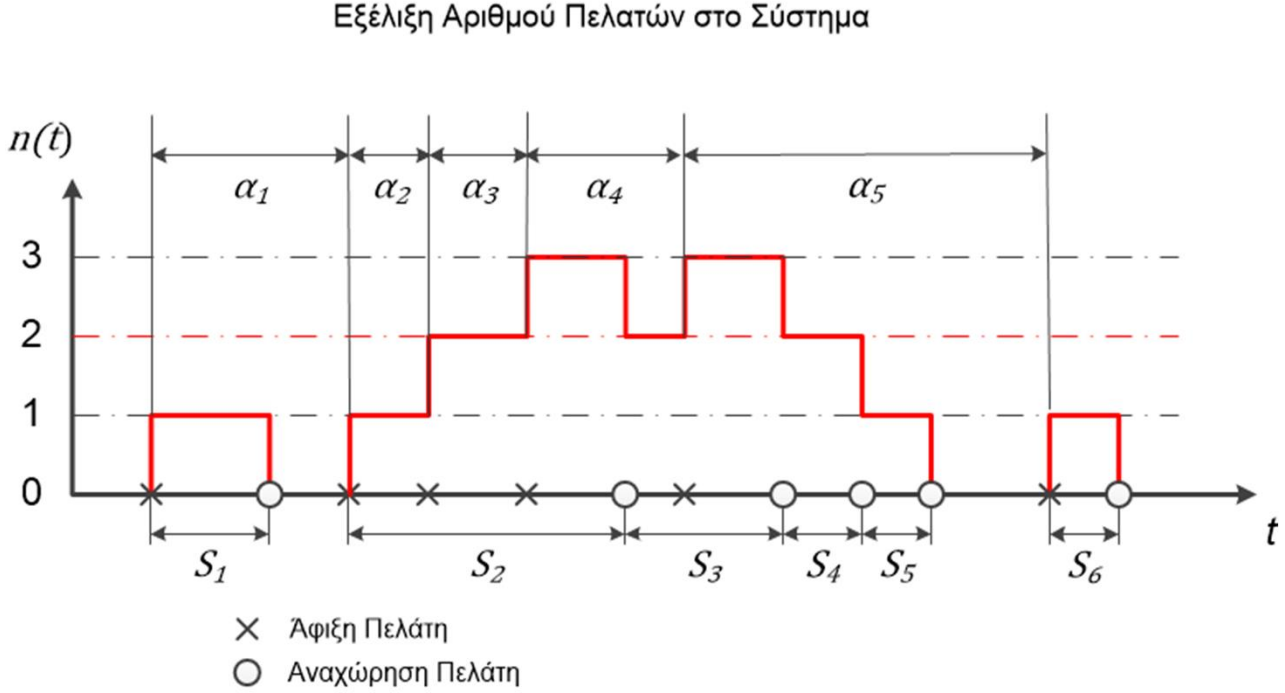
    Μέσος ρυθμός απωλειών: $\lambda - \gamma$

    Ποσοστό απωλειών: $\frac{\lambda - \gamma}{\lambda} = \mathrm{P\{blocking\}}$

- **Βαθμός χρησιμοποίησης εξυπηρετητή (server utilization)**
  - Σε περίπτωση 1 ουράς, 1 εξυπηρετητή

    $$u \triangleq \gamma / \mu$$

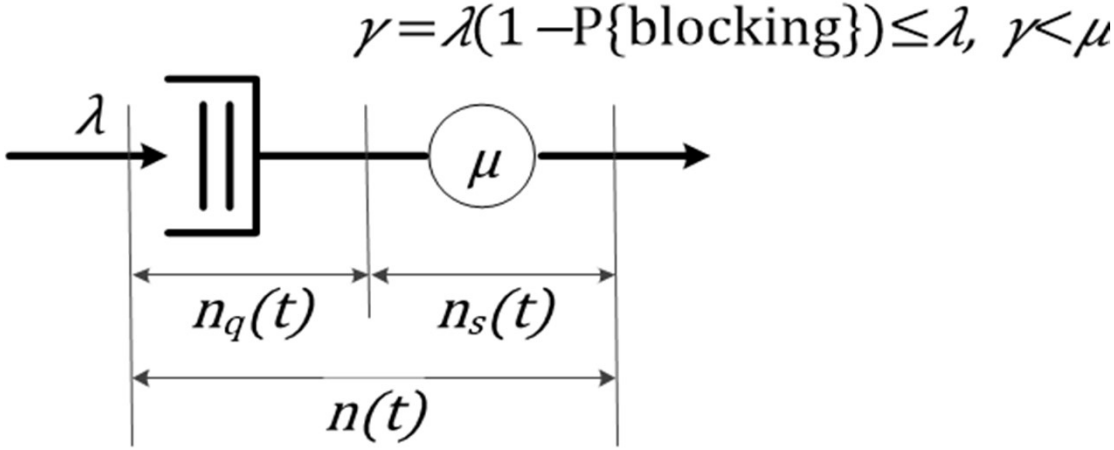
ΠΑΡΑΜΕΤΡΟΙ (3/4)



- **Αριθμός πελατών (κατάσταση)**
  $n(t)$, στοχαστική ανέλιξη – χρονοσειρά (stochastic process, time series)
- **Μέσος αριθμός πελατών** $\mathrm{E}\{n(t)\}$
- **Μέσος χρόνος καθυστέρησης (average time delay)**
  Μέσος χρόνος αναμονής (waiting time) + Μέσος χρόνος εξυπηρέτησης

$$\mathrm{E}(T) = \mathrm{E}(W) + \mathrm{E}(s)$$

ΠΑΡΑΜΕΤΡΟΙ (4/4)



- $n(t)$: **Κατάσταση συστήματος αναμονής**
- $n_q(t)$: **Αριθμός πελατών στην αναμονή**
- $n_s(t)$: **Αριθμός πελατών στην εξυπηρέτηση**
- $n(t) = n_q(t) + n_s(t)$
- $E\{n(t)\} = E\{n_q(t)\} + E\{n_s(t)\}$
- **Χρόνος καθυστέρησης**: $T = W + s$
- $E(T) = E(W) + E(s)$

ΚΑΤΑΣΤΑΣΗ ΣΥΣΤΗΜΑΤΟΣ

- $n(t) = 0,1,2,\dots,K$: Τυχαία μεταβλητή που ορίζει την **κατάσταση** του Συστήματος Αναμονής την χρονική στιγμή t . Η τυχαία συνάρτηση $n(t)$ αποτελεί **στοχαστική ανέλιξη** (διαδικασία) διακριτής κατάστασης με μεταβάσεις καταστάσεων σε συνεχή χρόνο (**discrete state, continuous time stochastic process**)

  $n(t) = n_q(t) + n_s(t) \leq K$ όπου:

  K η μέγιστη χωρητικότητα συστήματος

  $n_q(t) = 0,1,2,\dots,K-1$ ο αριθμός πελατών σε αναμονή

  $n_s(t) = 0,1$ ο αριθμός πελατών στην εξυπηρέτηση (αν έχω έναν εξυπηρετητή)

- $\mathrm{P}_k(t) \triangleq \mathrm{P}\{n(t) = k\}$: Η πιθανότητα παρουσίας k πελατών (σε αναμονή και εξυπηρέτηση) τη χρονική στιγμή t

ΙΣΟΡΡΟΠΙΑ, ΕΡΓΟΔΙΚΕΣ ΚΑΤΑΣΤΑΣΕΙΣ ΣΥΣΤΗΜΑΤΟΣ

- **ΟΡΙΣΜΟΣ ΙΣΟΡΡΟΠΙΑΣ**: Αν μια στοχαστική ανέλιξη $n(t)$ **ισορροπήσει** μετά από παρέλευση μεγάλου χρονικού διαστήματος t, το μεταβατικό φαινόμενο παρέρχεται και το σύστημα παλινδρομεί τυχαία ανάμεσα σε απείρως επισκέψιμες (**γνησίως επαναληπτικές**, **positive recurrent**) καταστάσεις $n(t) = k.$ Οι $\mathrm{P}_k(t)$ συγκλίνουν σε σταθερές τιμές $P_k > 0$ ανεξάρτητες της αρχικής κατάστασης $n(0)$

- **ΠΡΟΣΟΧΗ:** *Οι στοχαστικές ανελίξεις δεν ισορροπούν υποχρεωτικά, μόνο κάτω από ειδικές συνθήκες όπως αυτές των καλοσχεδιασμένων συστημάτων αναμονής*

- Οι απείρως επισκέψιμες καταστάσεις $n(t) = k$ συστήματος σε **ισορροπία** αποκαλούνται **εργοδικές καταστάσεις**

- Σύστημα σε ισορροπία: $\lim_{t\to\infty} \mathrm{P}_k(t) = P_k > 0$ (**εργοδικές οριακές πιθανότητες**)

  $P_k = \lim_{\mathrm{T}\to\infty} \frac{\mathrm{T}_k}{\mathrm{T}} > 0$ όπου T_k ο συνολικός χρόνος στη κατάσταση $n(t) = k$ στη διάρκεια Τ **μιας** παρατήρησης (εξέλιξης) της ανέλιξης $n(t)$

- Εργοδικοί Μέσοι Όροι των $n(t), n_q(t), n_s(t)$ συστήματος σε ισορροπία:

$$\mathrm{E}\{n(t)\} = \mathrm{E}\{n_q(t)\} + \mathrm{E}\{n_s(t)\}, \forall t$$

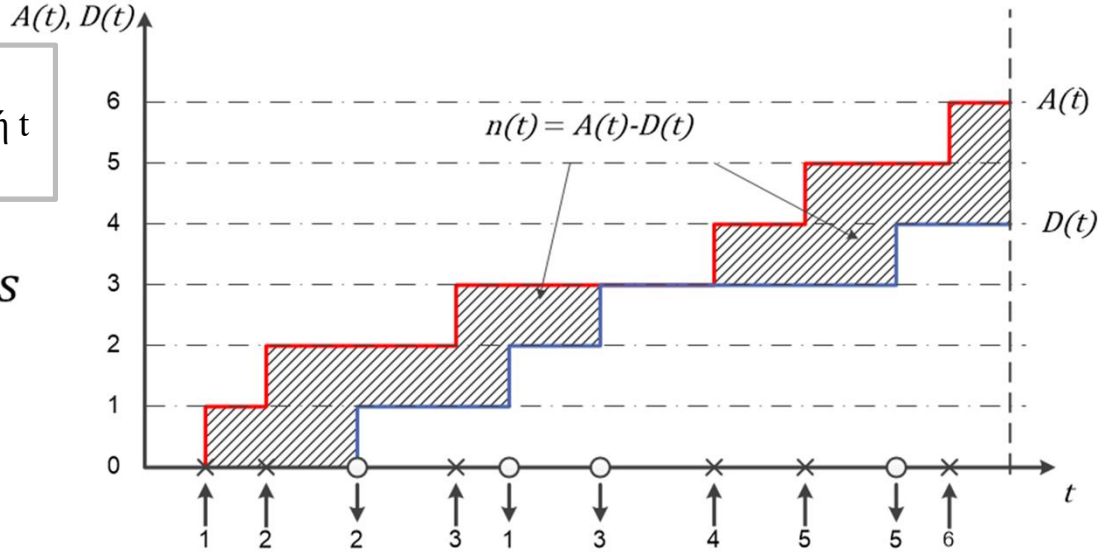
- Εργοδικοί Μέσοι Χρόνοι Καθυστέρησης (αναμονή + εξυπηρέτηση) συστήματος σε ισορροπία:

$$T = W + s, \mathrm{E}(T) = \mathrm{E}(W) + \mathrm{E}(s)$$

ΤΥΠΟΣ Little

(Σύστημα σε Ισορροπία)

A(t): συνολικός # αφίξεων μέχρι τη στιγμή t
D(t): συνολικός # αναχωρήσεων μέχρι τη στιγμή t
n(t): συνολικός # πελατών τη στιγμή t



Χρόνος καθυστέρησης: $T = W + s$

Τύπος Little:

$$\mathrm{E}(T) = \frac{\mathrm{E}\{n(t)\}}{\gamma} = \mathrm{E}(W) + \mathrm{E}(s)$$

$$= \frac{\mathrm{E}\{n_q(t)\}}{\gamma} + \frac{\mathrm{E}\{n_s(t)\}}{\gamma}$$

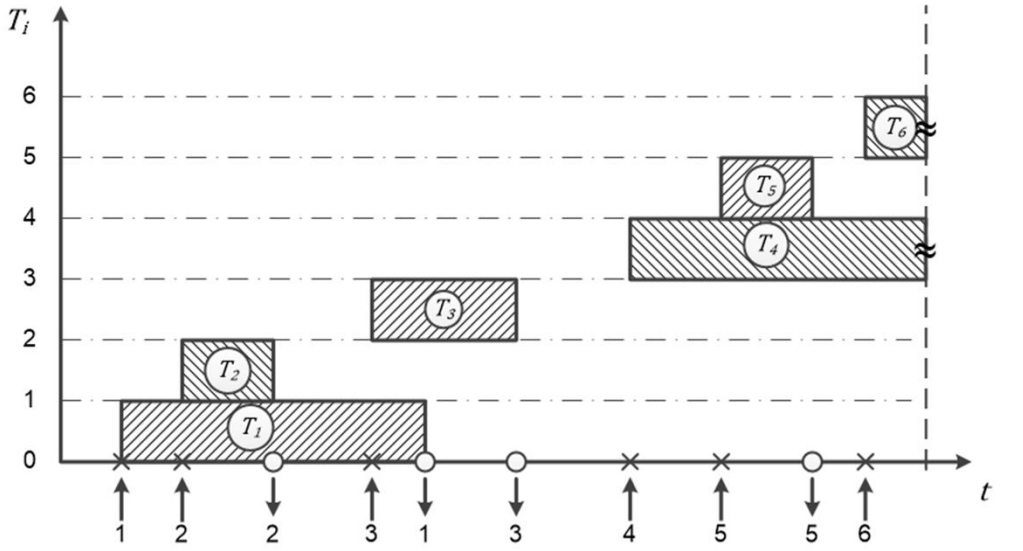
Για ουρά με **ένα** εξυπηρετητή:

$$\mathrm{E}\{n_s(t)\} = \gamma \mathrm{E}(s) = \frac{\gamma}{\mu}$$

$$= 0 \cdot \mathrm{P}\{n(t) = 0\} + \mathrm{P}\{n(t) > 0\}$$

$$= \mathrm{P}\{n(t) > 0\} =$$

(ο βαθμός χρησιμοποίησης του εξυπηρετητή $u = \frac{\gamma}{\mu} = \mathrm{P}\{n(t) > 0\}$)



ΚΑΤΑΤΑΞΗ ΟΥΡΩΝ ΑΝΑΜΟΝΗΣ

- ***A/S/N/K***
  - *A* : Τύπος διαδικασίας εισόδου πελατών
  - *S* : Τύπος τυχαίας μεταβλητής χρόνου εξυπηρέτησης
  - *N*: Αριθμός εξυπηρετητών
  - *K* : Χωρητικότητα συστήματος (μέγιστος αριθμός πελατών στην αναμονή + εξυπηρέτηση)
- *Παραδείγματα*
  - ***M/M/1:*** Αφίξεις Poisson (*Markov, Memoryless*), ανεξάρτητοι χρόνοι εξυπηρέτησης εκθετικοί (*Markov*), 1 εξυπηρετητής, άπειρη χωρητικότητα συστήματος (*μηδενικές απώλειες ή αστάθεια*)
  - ***M/D/1:*** Αφίξεις Poisson (*Markov, Memoryless*), ανεξάρτητοι χρόνοι εξυπηρέτησης σταθεροί (*Deterministic*), 1 εξυπηρετητής, άπειρη χωρητικότητα συστήματος
  - ***M/G/1/4:*** Αφίξεις Poisson (*Markov, Memoryless*), ανεξάρτητοι χρόνοι εξυπηρέτησης γενικής κατανομής (*General*), 1 εξυπηρετητής, χωρητικότητα συστήματος 4 πελάτες
  - ***M/M/4/8:*** Αφίξεις Poisson (*Markov, Memoryless*), ανεξάρτητοι χρόνοι εξυπηρέτησης εκθετικοί (*Markov*), 4 εξυπηρετητές, χωρητικότητα συστήματος 8 πελάτες: *Μοντέλο κέντρου κλήσεων (call center) με 4 χειριστές – τηλεφωνητές & μέχρι 4 κλήσεις στην αναμονή*

Η εκθετική κατανομή-(exponential distribution)

- Μια τυχαία μεταβλητή (τ.μ.) - random variable - ***X*** ακολουθεί ***Εκθετική Κατανομή*** *(Exponential Distribution)* με παράμετρο ***λ*** όταν:
- **CDF**: $F_X(t) = \mathrm{P}[X \le t] = \begin{cases} 1 - e^{-\lambda t}, & t \ge 0 \\ 0, & t < 0 \end{cases}$ και **PDF**: $f_X(t) = \frac{dF_X(t)}{dt} = \begin{cases} \lambda e^{-\lambda t}, & t \ge 0 \\ 0, & t < 0 \end{cases}$
- $\mathrm{E}[X] = \int_{t=0}^{\infty} \lambda t e^{-\lambda t} dt = 1/\lambda$
- $\mathrm{E}[X^2] = \int_{t=0}^{\infty} \lambda t^2 e^{-\lambda t} dt = 2/\lambda^2, \;\; \sigma_X^2 = \mathrm{E}[X] - (\mathrm{E}[X])^2 = 1/\lambda^2$
- Ιδιότητα έλλειψης μνήμης:
  - $\mathrm{P}[X > t + s \,|X > s] = \frac{\mathrm{P}[X>t+s,\, X>s]}{\mathrm{P}[X>s]} = \frac{\mathrm{P}[X>t+s]}{\mathrm{P}[X>s]} = e^{-\lambda t} = \mathrm{P}[X > t] = 1 - F_X(t)$

  Η εκθετική κατανομή είναι η **μόνη κατανομή συνεχούς μεταβλητής** με την ιδιότητα αυτή *(Memoryless, Markov Property)*.
- Κατανομή ελαχίστου μεταξύ ανεξάρτητων τ.μ. εκθετικά κατανεμημένων

  X1: με παράμετρο λ1

  X2: με παράμετρο λ2

  $X = \min(X1, X2)\,, F_x(\tau) = \mathrm{P}\{X \le \tau\} = 1 - \mathrm{P}\{X > \tau\} = 1 - e^{-(\lambda_1+\lambda_2)\tau}$ διότι

  $\mathrm{P}\{X > \tau\} = \mathrm{P}\{X1 > \tau, X2 > \tau\} = \mathrm{P}\{X1 > \tau\}\mathrm{P}\{X2 > \tau\} = e^{-\lambda_1\tau} e^{-\lambda_2\tau} = e^{-(\lambda_1+\lambda_2)\tau}$

  X = min{X1,X2} είναι εκθετικά κατανεμημένη με παράμετρο: λ = λ1+λ2

Στοχαστικές διαδικασίες
(Stochastic Processes – Time Series)

- Στάσιμες διαδικασίες (stationary stochastic processes) - οι από κοινού συναρτήσεις κατανομής πιθανότητας είναι αμετάβλητες σε μετατοπίσεις στο χρόνο
- Διαδικασίες Markov, ιδιότητα έλλειψης μνήμης

  $P[\boldsymbol{X}(t_{n+1})=x_{n+1}/\boldsymbol{X}(t_n)=x_n,\boldsymbol{X}(t_{n-1})=x_{n-1},\ldots,\boldsymbol{X}(t_1)=x_1] = P[\boldsymbol{X}(t_{n+1})=x_{n+1}/\boldsymbol{X}(t_n)=x_n]$
- Εργοδικότητα (ergodicity) ως προς τον μέσο όρο – μέση τιμή στο χρόνο συνάρτησης δείγματος είναι ίση με στατιστική μέση τιμή
- Διαδικασίες Γεννήσεων-Θανάτων (birth – death processes): αποτελούν μια κλάση των διαδικασιών Markov, με την επιπλέον συνθήκη ότι μεταβάσεις επιτρέπονται μόνο ανάμεσα σε γειτονικές καταστάσεις
- Διαδικασία απαρίθμησης γεγονότων (counter processes)

  $P[\boldsymbol{N}(t) = k]$: Πιθανότητα k γεγονότων στο διάστημα $(0, t)$
- Ανεξάρτητες αυξήσεις: αν οι αριθμοί των γεγονότων που λαμβάνουν χώρα σε μη επικαλυπτόμενα διαστήματα είναι ανεξάρτητοι μεταξύ τους
- Στάσιμες αυξήσεις (stationary increments): Ανεξάρτητα του χρόνου αναφοράς t *(εξάρτηση μόνο από το μήκος του διαστήματος)*

  $P[\boldsymbol{N}(t + \Delta t) - \boldsymbol{N}(t) = k] = P[\boldsymbol{N}(\tau + \Delta t) - \boldsymbol{N}(\tau) = k] = P[\boldsymbol{N}(\Delta t) = k]$