Δίκτυα Υπολογιστών

Πρόγραμμα Μεταπτυχιακών Σπουδών:
Τεχνο – Οικονομικά Συστήματα

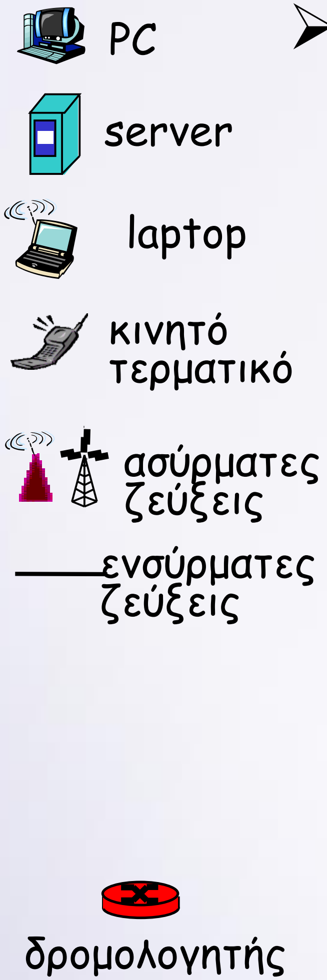
Καθηγητής Συμεών Παπαβασιλείου

Εθνικό Μετσόβιο Πολυτεχνείο
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Επικοινωνιών, Ηλεκτρονικής & Συστημάτων Πληροφορικής

(E-mail: papavass@mail.ntua.gr   Τηλ: 210 772-2550
Γραφείο: Β.3.15 Νέο Κτίριο Ηλεκτρολόγων)

Βιβλίο Αναφοράς: TCP/IP Illustrated, Vol. I, by W.R. Stevens (Addison Wesley)

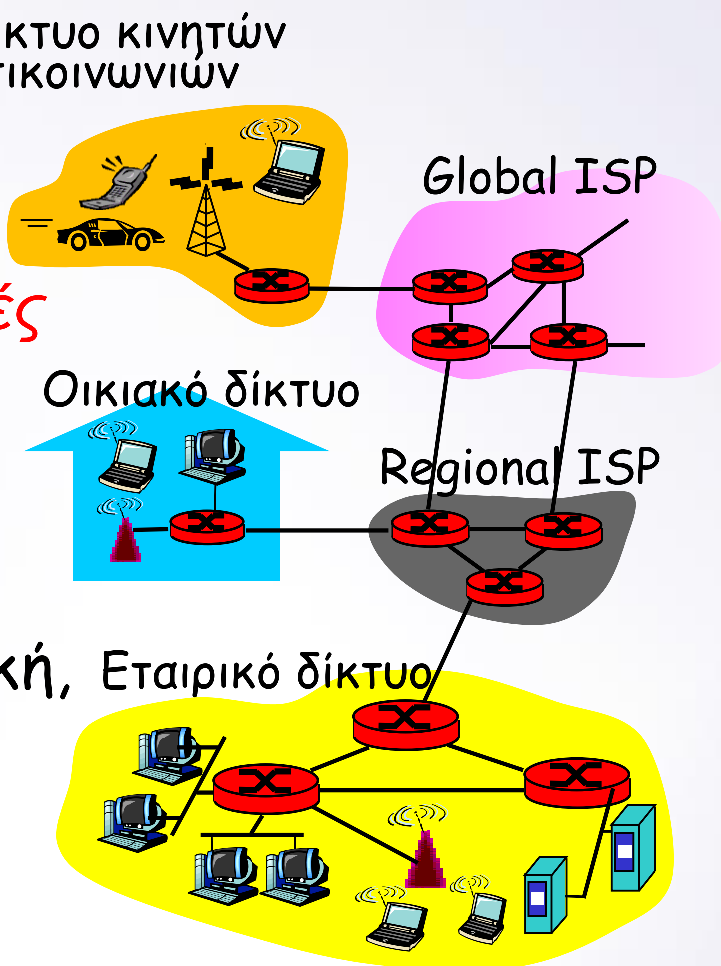
Internet



- εκατομμύρια διασυνδεδεμένων υπολογιστών: *hosts*

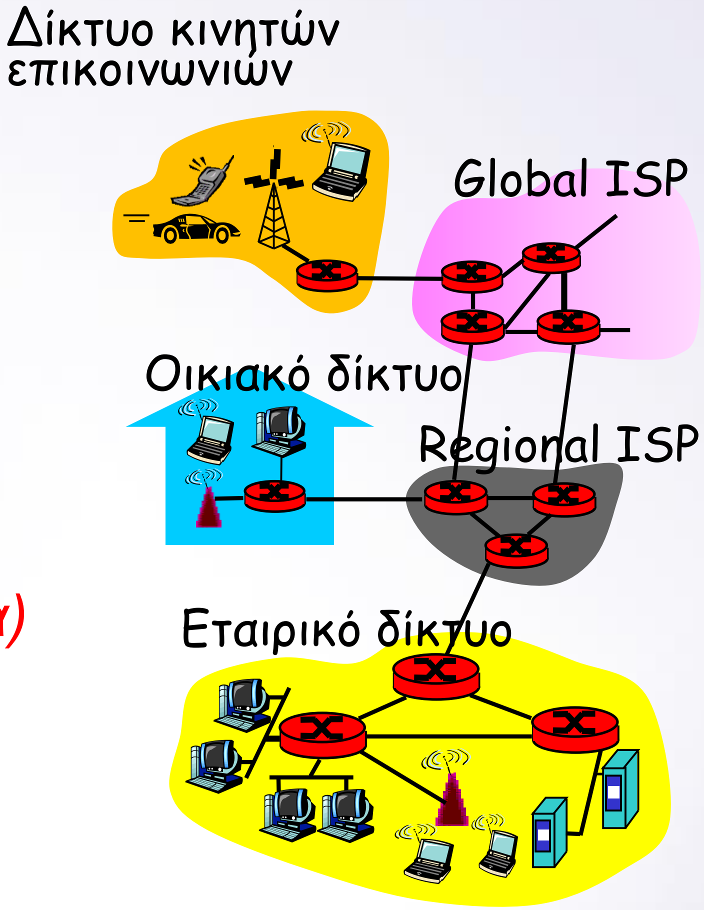
  όπου τρέχουν *δικτυακές εφαρμογές*

- *τηλεπ. ζεύξεις*
- οπτική ίνα, χάλκινος αγωγός, μικροκυματική, ζεύξη, δορυφορική
- ρυθμός μετάδοσης = *bandwidth*
- *δρομολογητές:* προωθούν πακέτα



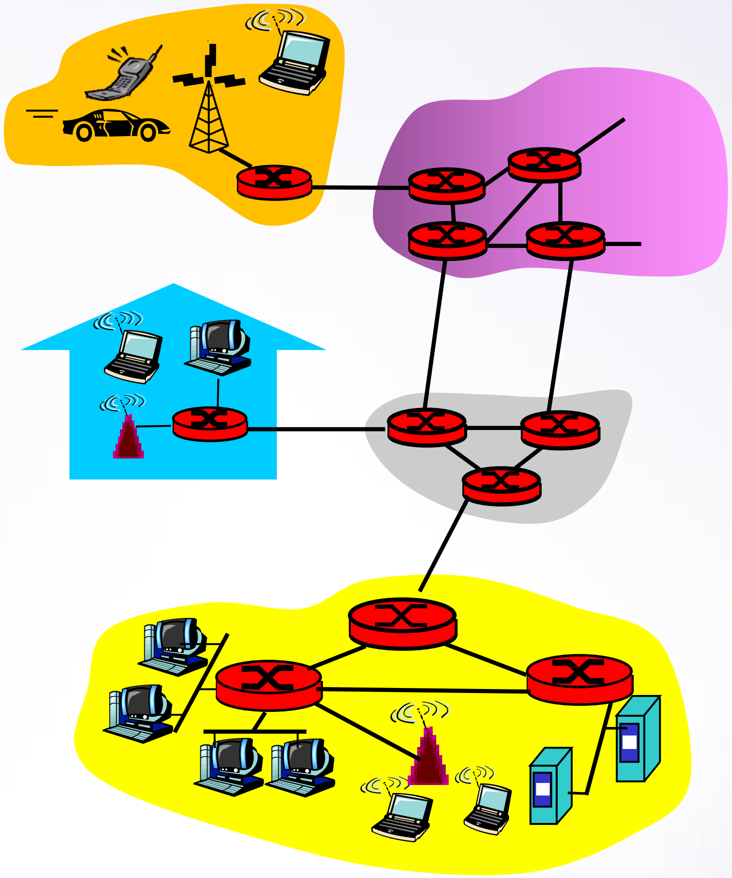
Internet

- *Πρωτόκολλα* ρυθμίζουν την αποστολή και λήψη μηνυμάτων
  - π.χ., Ethernet, IP, TCP, HTTP, Skype,
- *Internet: "δίκτυο από δίκτυα"*
  - όχι αυστηρά ιεραρχικό
  - public Internet, private intranet
- *Ανάλογο δικτύου αυτοκινητοδρόμων και διασταυρώσεων (πακέτα=οχήματα)*

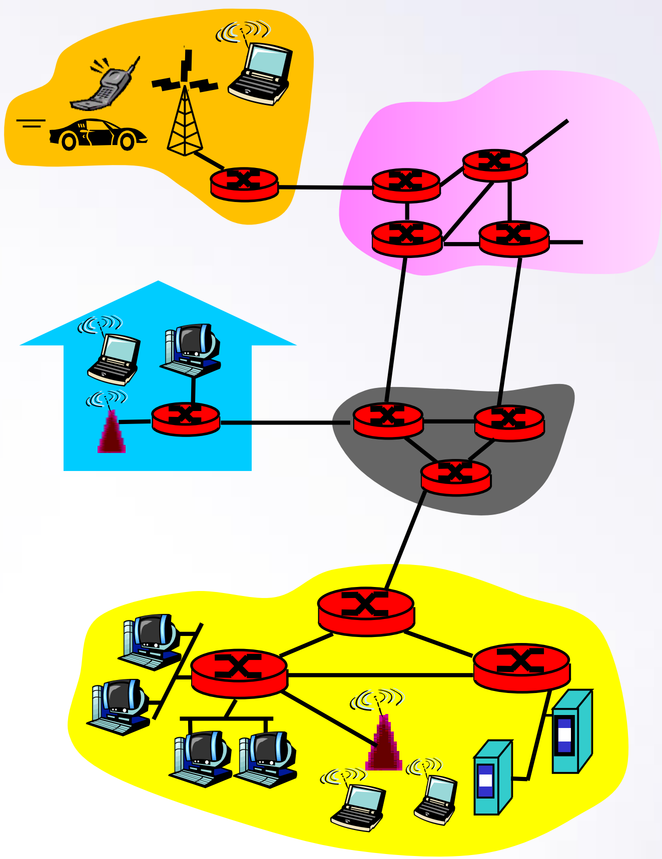


Internet: δομή δικτύου

- Άκρα του δικτύου: hosts και εφαρμογές
- Δίκτυα πρόσβασης, φυσικά μέσα: ενσύρματες, ασύρματες τηλεπ. ζεύξεις
- Δίκτυο κορμού:
  - διασυνδεδεμένοι δρομολογητές
  - δίκτυο από δίκτυα



Internet: άποψη υπηρεσίας



- Τηλεπικοινωνιακή *υποδομή* που υποστηρίζει κατανεμημένες εφαρμογές:
  - Web, VoIP, email, games, e-commerce, file sharing
- Τηλεπικοινωνιακές υπηρεσίες που παρέχονται στις εφαρμογές:
  - αξιόπιστη μεταφορά δεδομένων από την πηγή στον προορισμό
  - “best effort” (αναξιόπιστη) μεταφορά δεδομένων
- *Ανάλογο ταχυδρομικής υπηρεσίας*

Τα *πρωτόκολλα* καθορίζουν τη *μορφή* και τη *σειρά* των μηνυμάτων που *ανταλλάσσονται* μεταξύ των οντοτήτων του δικτύου και τις *ενέργειες που γίνονται* κατά την αποστολή ή λήψη των μηνυμάτων.

Economies of Scale

- Average cost per user of the network declines as network increases in size (number of users or subscribers)
- Cost of a transmission link grows at a slower rate than does its capacity or speed
- A network has certain fixed cost of operations, administration, and maintenance. Because these costs are not sensitive to network size (or at least they grow at much slower rate than the network size) the per user share of these fixed costs declines with the number of users

Network Externalities

- A network service is said to have positive externalities if its value to a user increases with the number of users (e.g. telephone service)
- Externalities provide a powerful incentive for internetworking. Additional resources needed to implement the interconnection are few if networks follow compatible standards
- Combination of network scale economies and network externalities may lead to an exponential growth in demand and supply of network services (e.g. Internet access)

Service Integration

- Economies of scope or service integration, refers to the fact that a network that currently provides one set of services may be expanded to provide new services at an additional cost that is much less than if a separate network were build to provide these new services
  - Services include: telephone, data, video, etc.

Basic Networking Mechanisms

- Multiplexing
  - Time Division Multiplexing
  - Frequency Division Multiplexing
  - Statistical Multiplexing
- Routing/Switching
- Error Control
  - Error Detection (e.g. parity bit, cyclic redundancy check)
  - Error Correction
- Flow Control (e.g. window mechanism)
- Resource Allocation (routing, bandwidth management, buffer allocation, admission control)

Switched Networks

- **Circuit Switched**
- **Packet Switched**

Virtual Circuit

(Connection oriented)

- Analogous to Telephone System
- Call setup procedure
- Route established one time for all packets
- Guaranteed delivery
- Guaranteed sequence
- Error and Flow Control handled by Network

Datagram

(Connectionless)

- Analogous to Postal Service
- No call setup
- Every packet individually routed
- No
- No
- Error and Flow Control by higher layers (end user)

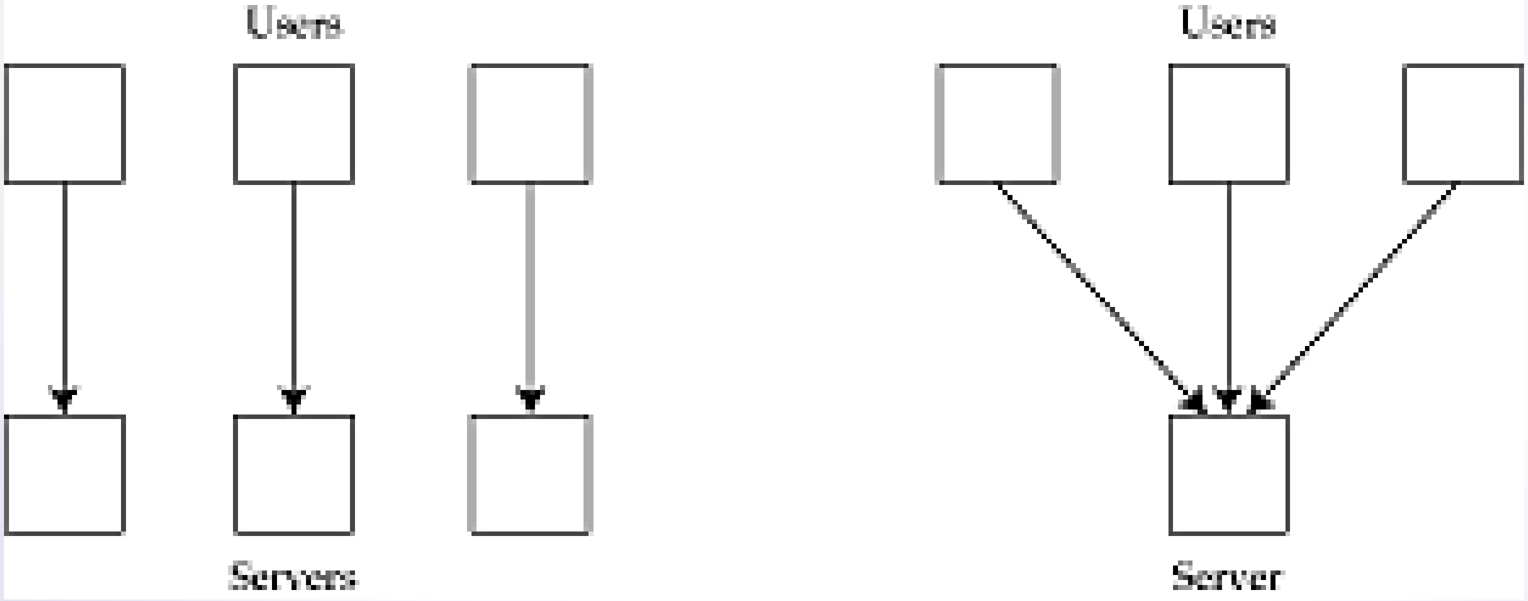
Πολυπλεξία Πόρων Δικτύου

Common design techniques

- Key concept: *bottleneck*
  - the most constrained element in a system
- System performance improves by removing bottleneck
  - but creates new bottlenecks
- In a *balanced* system, all resources are simultaneously bottlenecked
  - this is optimal
  - but nearly impossible to achieve
  - in practice, bottlenecks move from one part of the system to another

Multiplexing

- Another word for sharing
- Trades time and space for money
- Users see an increased response time, and take up space when waiting, but the system costs less



Multiplexing (contd.)

- Examples
  - Bank teller
  - multiplexed links
  - shared memory
- *Server* controls access to the shared resource
  - uses a *schedule* to resolve contention
  - choice of scheduling critical in providing quality of service guarantees

Statistical multiplexing

- Suppose resource has capacity C
- Shared by N identical tasks
- Each task requires capacity c
- If Nc <= C, then the resource is underloaded
- If at most 10% of tasks active at a given time, then C >= Nc/10 is enough
  - we have used statistical knowledge of users to reduce system cost
  - this is *statistical multiplexing gain*

Statistical multiplexing (contd.)

- **Two types: spatial and temporal**
  - **Spatial**
    - we expect only a fraction of tasks to be simultaneously active
  - **Temporal**
    - we expect a task to be active only part of the time
      - e.g silence periods during a voice call
    - Lets us assume that a a resource has capacity C. We expect the average rate of task-i, denoted by A(i) to be less than the peak rate P(i). Then:
      - Conservative sizing approach: $C_{max}=\Sigma_i P(i)$
      - Minimum sizing: $C_{min}=\Sigma_i A(i)$
      - Choose C in the range of $[C_{min}, C_{max}]$

Example of statistical multiplexing gain (SMG)

- Consider a 100 room hotel
- How many external phone lines does it need?
  - each line costs money to install and rent
  - Tradeoff (spatial multiplexing): Using a conservative approach you may use 100 lines! Using a less conservative approach you may expect that only 10% of guests will be simultaneously active. Then use 10 lines!

Example of statistical multiplexing gain (SMG)

- What if a voice call is active only 40% of the time?
  - can get both spatial and temporal statistical multiplexing gain
  - Then you may need capacity equivalent to 4 lines!
  - but only in a packet-switched network (packetized voice samples)
- Remember
  - to get SMG, we need good statistics!
  - if statistics are incorrect or change over time, we're in trouble

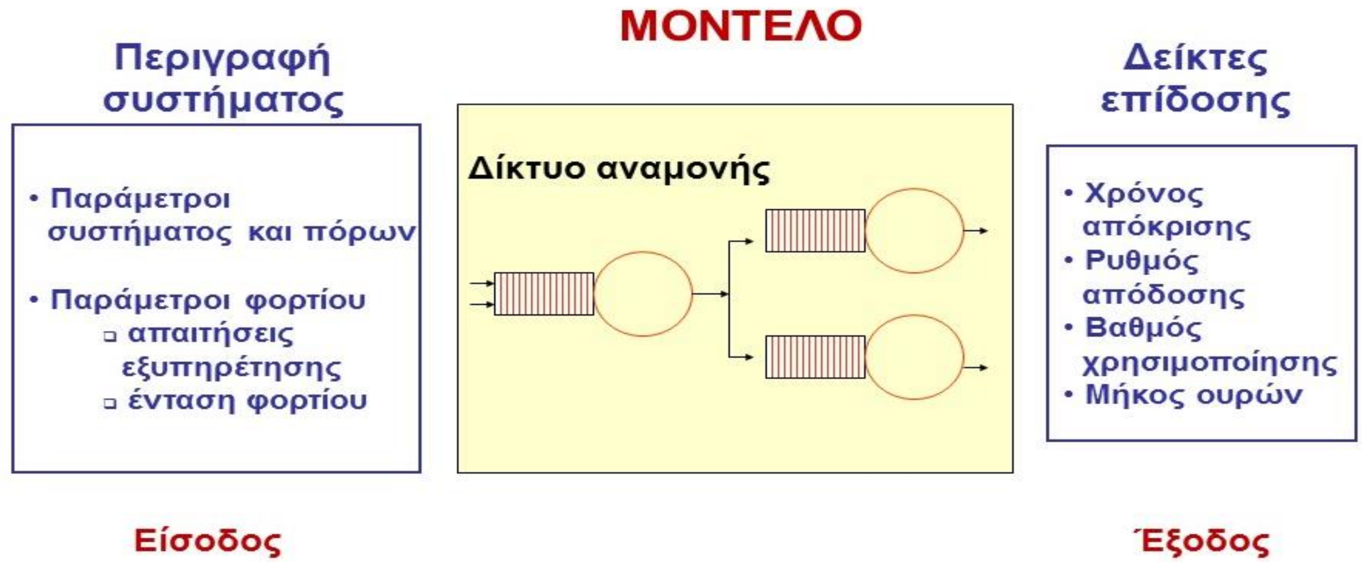
ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΑΞΙΟΛΟΓΗΣΗΣ ΕΠΙΔΟΣΗΣ & ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Τεχνικές:

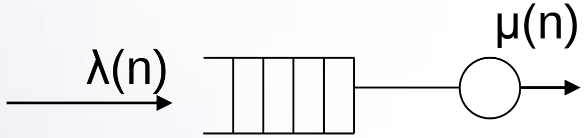
- 1η: Μετρήσεις με πραγματικές τιμές και ανάλυση αποτελεσμάτων
  - Προέκταση σε μεγαλύτερης κλίμακας συστήματα: συνήθως η συμεπτριφορά δεν είναι αναμενόμενη (π.χ. Γραμμική) σε αλλαγή φόρτου εργασίας
- 2η: Χρήση μοντέλων – μοντελοποίηση

  Γενικευμένη αναπαράσταση του συστήματος (αφαιρετική): περιλαμβάνει τα κύρια χαρακτηριστικά και αφαιρεί λεπτομέρειες που εκτιμάται ότι δεν επηρεάζουν σημαντικά την απόδοση του συστήματος (υποθέσεις). Κυκλοφορία (απαιτήσεις) χρηστών είναι στοχαστική (τυχαιότητα).
  - Αναλυτικά μοντέλα: χρήση μαθηματικής περιγραφής του συστήματος (βασισμένα κυρίως σε θεωρία αναμονής – queueing theory) και αλγορίθμων .
  - Προσομοίωση: Ανάπτυξη προγράμματος που ακολουθεί και αναπαριστά τη δυναμική εξέλιξη του συστήματος στο χρόνο.

ΜΟΝΤΕΛΟ ΣΥΣΤΗΜΑΤΟΣ – ΔΙΚΤΥΟ ΑΝΑΜΟΝΗΣ - ΜΟΝΤΕΛΑ ΣΥΜΦΟΡΗΣΗΣ



- 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)



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

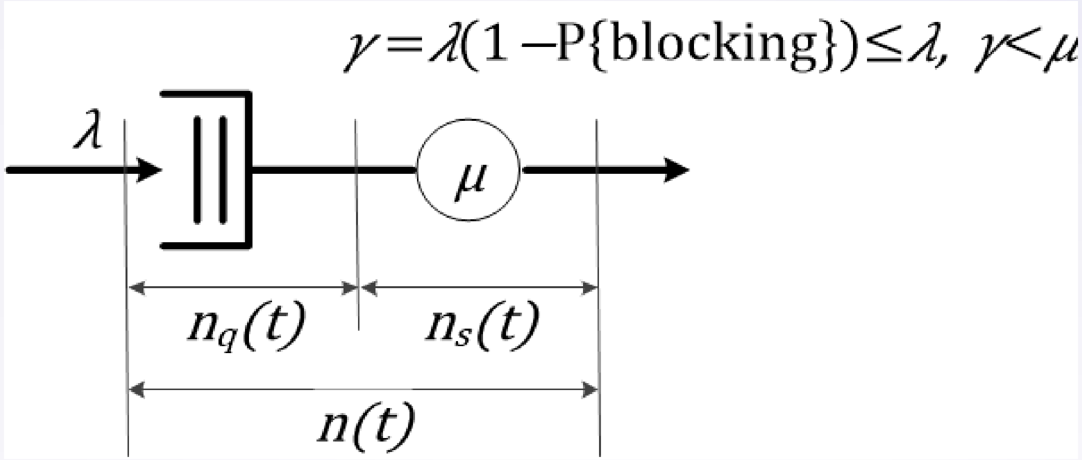
Σε περίπτωση 1 ουράς, 1 εξυπηρετητή:

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

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)



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

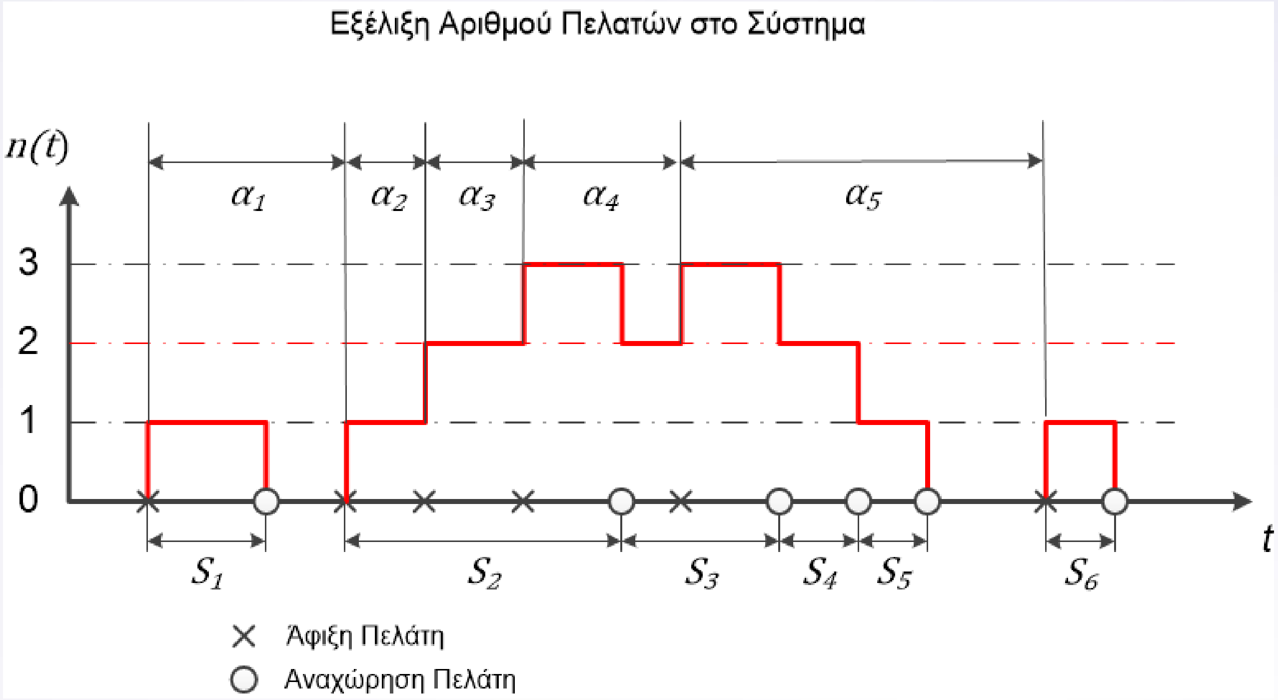
    Μέσος ρυθμός απωλειών: $\lambda - \gamma$

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

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

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

ΠΑΡΑΜΕΤΡΟΙ (3/4)



- **Αριθμός πελατών (κατάσταση)**

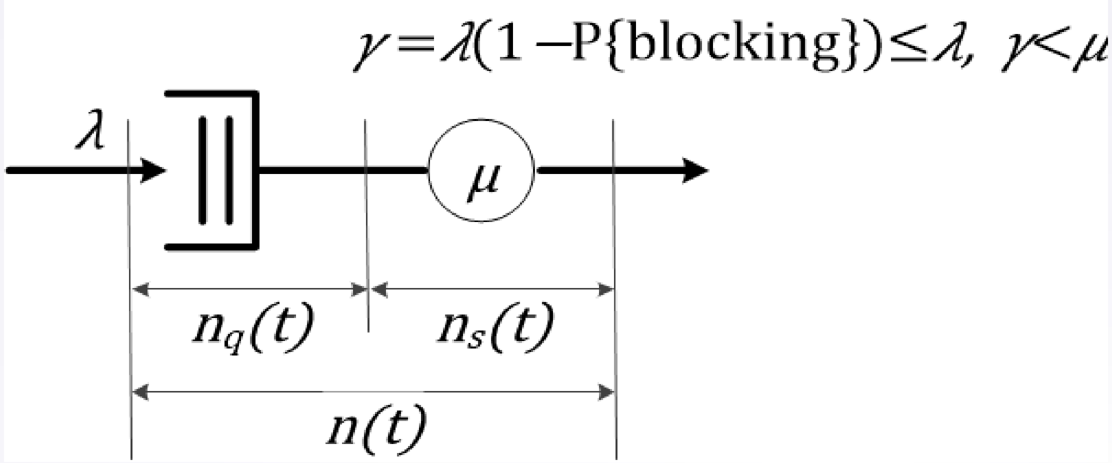
  $n(t)$, στοχαστική ανέλιξη – χρονοσειρά (stochastic process, time series)

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

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

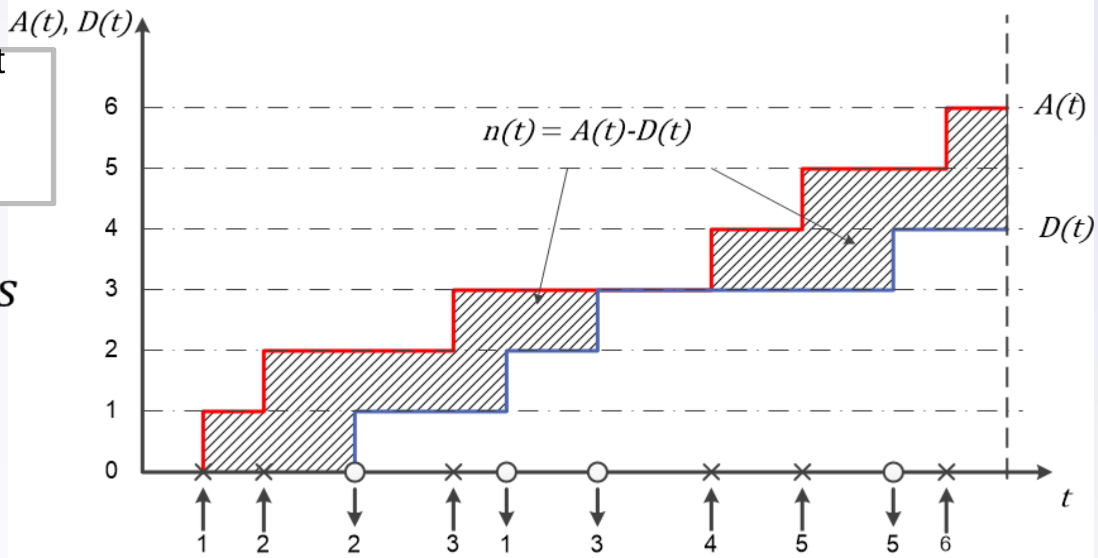
  $n_s(t) = 0,1$ ο αριθμός πελατών στην εξυπηρέτηση (αν έχω έναν εξυπηρετητή)

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

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: $T = W + s$

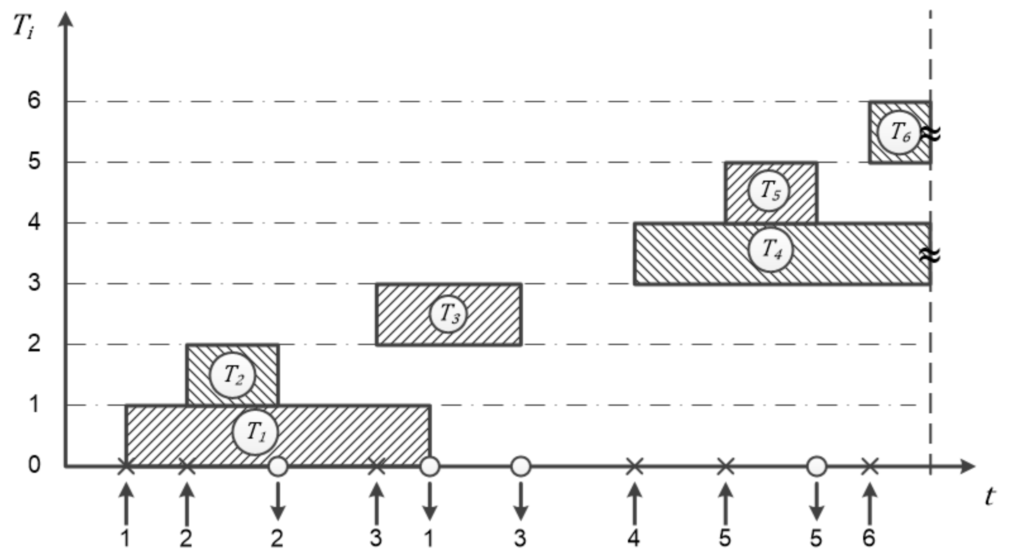
Τύπος Little:

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

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**: $F_X(t) = \mathrm{P}[X \le t] = \begin{cases} 1 - e^{-\lambda t}, t \ge 0 \\ 0, \quad t < 0 \end{cases}$ και **PDF**: $f_X(t) = \frac{dF_X(t)}{dt} = \begin{cases} \lambda e^{-\lambda t}, t \ge 0 \\ 0, \quad t < 0 \end{cases}$
- $\mathrm{E}[X] = \int_{t=0}^{\infty} \lambda t e^{-\lambda t} dt = 1/\lambda$
- $\mathrm{E}[X^2] = \int_{t=0}^{\infty} \lambda t^2 e^{-\lambda t} dt = 2/\lambda^2, \ \sigma_X^2 = \mathrm{E}[X] - (\mathrm{E}[X])^2 = 1/\lambda^2$
- 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} 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)

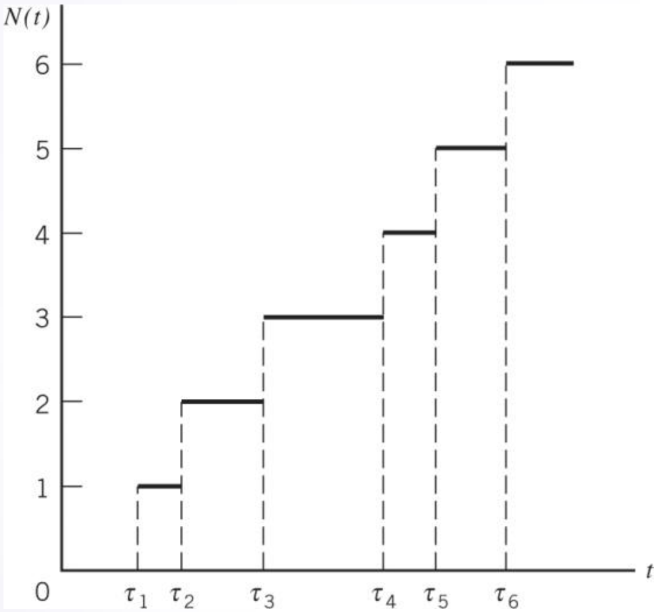
  $P[\boldsymbol{N}(t) = k]$: Πιθανότητα k γεγονότων στο διάστημα $(0, t)$
- Ανεξάρτητες αυξήσεις: αν οι αριθμοί των γεγονότων που λαμβάνουν χώρα σε μη επικαλυπτόμενα διαστήματα είναι ανεξάρτητοι μεταξύ τους
- Στάσιμες αυξήσεις (stationary increments): Ανεξάρτητα του χρόνου αναφοράς *t (εξάρτηση μόνο από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

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

Η ΔΙΑΔΙΚΑΣΙΑ ΚΑΤΑΜΕΤΡΗΣΗΣ ΓΕΓΟΝΟΤΩΝ POISSON

Η τυχαία εμφάνιση παλμών περιγράφεται σαν μια Στοχαστική Ανέλιξη Καταμέτρησης *(Counting Process)* $N(t)$ που καταμετρά τυχαία γεγονότα (αφίξεις πελατών) στο διάστημα $(0, t)$.

Ο αριθμός εμφανίσεων στο διάστημα $(t, t+T)$ είναι διακριτή τυχαία μεταβλητή $\nu = N(t+T) - N(t)$. Κάτω από συνθήκες απρόβλεπτης εξέλιξης της ανέλιξης *(τα γεγονότα εμφανίζονται ανεξάρτητα από το παρελθόν και χωρίς να επηρεάζουν το μέλλον)*, η ν ακολουθεί την **κατανομή *Poisson* με μέσο αριθμό εμφανίσεων ανάλογο του διαστήματος** T: $\mathrm{E}_T[\nu] = \lambda T$. Η σταθερά λ ορίζει τον μέσο ρυθμό *(rate)* εμφανίσεων (γεγονότα ανά μονάδα χρόνου)



Η κατανομή Poisson:

- ***n*** αφίξεις σε διάστημα ***T*** με πιθανότητα
- $P_{\boldsymbol{n}}(T) = e^{-\lambda T} (\lambda T)^{\boldsymbol{n}} / n\,!$
- $E_T(\boldsymbol{n}) = \lambda T$
- $Var_T(\boldsymbol{n}) = \lambda T$

Μέσος ρυθμός αφίξεων : *λ* πελάτες/sec

Ιδιότητες διαδικασίας Poisson:

- Οι χρόνοι μεταξύ διαδοχικών αφίξεων μιας διαδικασίας Poisson με ρυθμό λ, είναι τ.μ εκθετικά κατανεμημένες με μέση τιμή 1/λ
- Υπέρθεση ανεξάρτητων διαδικασιών Poisson
- Διάσπαση διαδικασίας Poisson με πείραμα Bernoulli

- **Διαδικασία Γεννήσεων – Θανάτων**:
  - Την χρονική στιγμή t όταν το σύστημα καταλήγει σε πληθυσμό $n > 0$ μπορεί να έχουν προηγηθεί οι ακόλουθες μεταβάσεις από την χρονική στιγμή $t\text{-}\Delta T$, $\Delta T \rightarrow 0$:
    - Μία άφιξη στο διάστημα ΔT, με πιθανότητα $\boldsymbol{\lambda_{n-1}\Delta T}$
    - Μια αναχώρηση, με πιθανότητα $\boldsymbol{\mu_{n+1}\Delta T}$
    - Τίποτα από τα δύο, με πιθανότητα $\mathbf{1- (\lambda_n+\mu_n)\Delta T}$
  - Η εξίσωση μετάβασης προκύπτει από τον τύπο συνολικής πιθανότητας:

$$P_n(t) = \lambda_{n-1}\Delta T\, P_{n-1}(t-\Delta T) + \mu_{n+1}\Delta T\, P_{n+1}(t-\Delta T) + [1- (\lambda_n+\mu_n)\Delta T]\, P_n(t-\Delta T)$$

Στο όριο, $\Delta T \rightarrow dt$:

$$[P_n(t) - P_n(t-dt)]/dt = \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t) - (\lambda_n+\mu_n)P_n(t)$$

ή

$$d\,P_n(t)/dt = \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t) - (\lambda_n+\mu_n)P_n(t)$$

και σε σταθερή κατάσταση $t \rightarrow oo$ (αν υπάρχει) :

$P_n(t) = P_n$ **:** Εργοδικές Πιθανότητες

$(\lambda_n+\mu_n)P_n = \lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1}$ (εξισώσεις ισορροπίας)

Η ουρά M/M/1 (άπειρου μεγέθους)

- Σταθεροί μέσοι ρυθμοί αφίξεων (γεννήσεων) $\boldsymbol{\lambda_n = \lambda}$**,** Poisson
- Σταθεροί μέσοι ρυθμοί εξυπηρέτησης (θανάτων) $\boldsymbol{\mu_n = \mu}$
- Εκθετικοί χρόνοι εξυπηρέτησης $\boldsymbol{s}$, $E(\boldsymbol{s}) = 1/\boldsymbol{\mu}$
- Εργοδικές πιθανότητες καταστάσεων $\boldsymbol{P_n}$
- Μέσος όρος πληθυσμού - κατάστασης $E(\boldsymbol{n})$

Η ουρά M/M/1

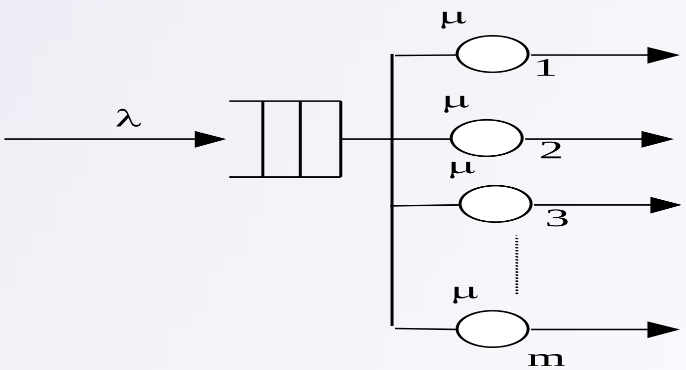
- $P_n = (1-\rho)\,\rho^n$, n = 0,1,2,…, $\rho = \lambda/\mu < 1$
- E(***n***) = $\rho/(1-\rho)$
- Νόμος του Little: E(***T***) = E(***n***)/*γ* = E(***n***)/*λ*
- E(***T***) = $(1/\mu) / (1-\rho)$

- Παράδειγμα Ανάλυσης Ουρών Markov: ***M/M/1/K*** (ουρά με μέγιστη χωρητικότητα ***K***, συμπεριλαμβανομένου του εξυπηρετουμένου)

  - Πιθανότητα απώλειας, P{blocking}

    $P_{bl} = P_K = P_o \rho^K$, $P_0 = (1-\rho)/(1-\rho^{K+1})$
  - Ρυθμαπόδοση ***(Throughput)*** $\gamma = \lambda (1- P_K)$
  - Μέση Καθυστέρηση $E(T) = E(n)/\gamma$

Παράδειγμα ανάλυσης ουράς Markov με m εξυπηρετητές M/M/m [Erlang –C]



Infinite buffer
Finite # of servers (m)

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

$$\rho_Q = \frac{P_0(\rho m)^m}{m!(1-\rho)}$$

Prob. All servers are busy

$$N_Q = P_Q \frac{\rho}{1-\rho}$$

$$W = \frac{\rho \cdot P_Q}{\lambda(1-\rho)}$$

$$T = \frac{p_Q}{m\mu - \lambda} + \frac{1}{\mu}$$

$$N = \frac{\rho p_Q}{1-\rho} + m\rho$$

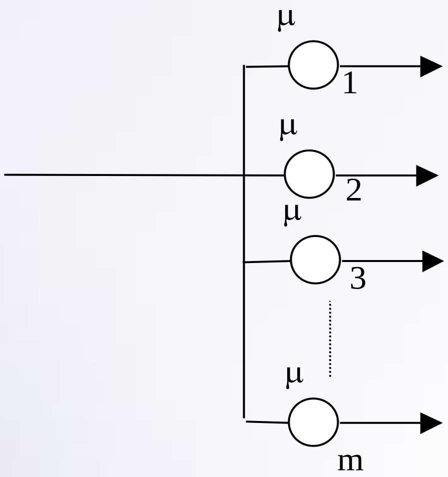
$$p_0 = \frac{1}{[\sum_{n=0}^{m-1} \frac{1}{n!}(\frac{\lambda}{\mu})^n] + \frac{1}{m!}(\frac{\lambda}{\mu})^m \frac{m\mu}{m\mu - \lambda}}$$

- Άλλα Παραδείγματα Ουρών Markov

  - ***M/M/N/K*** (***N*** εξυπηρετητές, χωρητικότητα ***K***, ***N* ≤ *K***)
    - $P_n = [\lambda/(n\mu)]\, P_{n-1}$, $n=1, 2, \ldots, N-1$
    - $P_n = [\lambda/(N\mu)]\, P_{n-1}$, $n=N, N+1, \ldots, K$
    - $P_0 + P_1 + \ldots + P_{K-1} + P_K = 1$

– ***M/M/m/m*** (m εξυπηρετητές, χωρητικότητα m) ***Erlang – B***

- Μοντέλο τηλεφωνικού κέντρου με μέσο ρυθμό κλήσεων ***λ*** (Poisson), εκθετική διάρκεια τηλεφωνήματος, μέσος χρόνος ***1/μ***, m γραμμές και απώλειες χωρίς επανάκληση (redial)



$$p_{block} = \frac{\frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m}{\sum_{n=0}^{m}\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n}$$

$$\lambda_{Block} = \lambda \cdot p_{Block}$$

ρ = λ/μ (Erlangs)

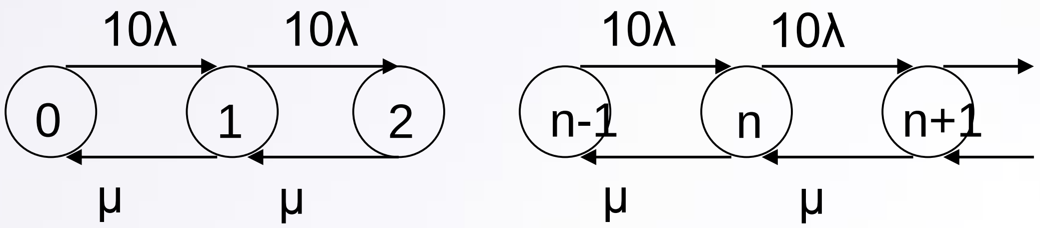
$$\boldsymbol{P_{bl} = P_m = (\rho^m/m!) / (1 + \rho + \rho^2/2 + \rho^3/3! + \ldots + \rho^m/m!)}$$

Παράδειγμα : Πιθανότητες και εξισώσεις καταστάσεων ισορροπίας

- 10 τερματικά τροφοδοτούν κοινό στατιστικό πολυπλέκτη πακέτου (μεταγωγέα – switch ή δρομολογητή – router) που εξυπηρετεί δεδομένα σε πακέτα των 1000 bits κατά μέσο όρο. Η έξοδος του πολυπλέκτη είναι γραμμή των 10 Mbps (Megabits per sec). Τα τερματικά θεωρούνται ανεξάρτητα και ισότιμα.
- Α) Προσεγγίστε τον πολυπλέκτη σαν ουρά Μ/Μ/1. Βρείτε το μέσο όρο ροής των δεδομένων ανά τερματικό ώστε η γραμμή να έχει χρησιμοποίηση 50%.
- Β) Αν ο πολυπλέκτης δεν δύναται να αποθηκεύει πάνω από 3 πακέτα (μαζί με το πακέτο υπό εξυπηρέτηση) και ο μέσος ρυθμός ροής πακέτων ανά τερματικό είναι 500 packets/sec, βρείτε τα χαρακτηριστικά της ουράς. Υποθέστε Poisson διαδικασία άφιξης πακέτων και εκθετικά κατανεμημένους χρόνους εξυπηρέτησης πακέτων.

Λύση – Τμήμα Α

Χρησιμοποιείται μοντέλο M/M/1
Η ροή πακέτων ανά τερματικό είναι λ.
Ζητούμενο: λ=? pak/sec,
Ο ρυθμός εξυπηρέτησης είναι: μ=(10000 kbits/sec)/(1 kbits/pkt) = 10000 pkts/sec

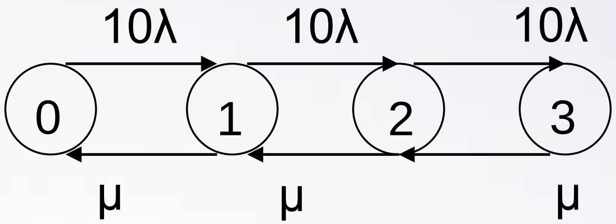


Η αθροιστική ροή πακέτων (από όλα τα τερματικά) στον πολυπλέκτη είναι: 10λ.

Ο βαθμός χρησιμοποίησης είναι: u=(10λ)/μ=0.5 →(10λ)/10000=0.5
→ λ=500 pkts/sec

Λύση – Τμήμα Β

Χρησιμοποιείται μοντέλο M/M/1/3
λ=500 pkts/sec, μ=10000 pkts/sec



$10\lambda P_0=\mu P_1 \rightarrow P_1=(10\lambda/\mu)P_0 \rightarrow P_1=0.5P0$
$10\lambda P_1=\mu P_2 \rightarrow P_2=(10\lambda/\mu)P_0^2 \rightarrow P_2=0.25P0$
$10\lambda P_2=\mu P_3 \rightarrow P_3=(10\lambda/\mu)P_0^3 \rightarrow P_3=0.125P0$

$P_0+P_1+P_2+P_3=1 \rightarrow P_0=8/15,\ P_1=4/15,\ P_2=2/15,\ P_3=1/15$

Μέσο μήκος ουράς: $E(n)=0*(8/15)+1*(4/15)+2*(2/15)+3*(1/15)=11/15$ pkts

Πιθανότητα απωλειών: $P_{bl}=P_3=1/15$

Ρυθμαπόδοση: $\gamma=10\lambda(1-P_{bl})=500*(14/15)=1400/3$ pkts/sec
$(\gamma=\mu(1-P_0))$

Μέση Καθυστέρηση: $E(\tau)=E(n)/\gamma=(11/15)/(1400/3)=11/7000$ sec

ΣΥΣΤΗΜΑΤΑ ΑΝΑΜΟΝΗΣ

Άσκηση : Μοντελοποίηση Τηλεφωνικής Ζεύξης

Μια τηλεφωνική εταιρεία εγκαθιστά σύνδεση μεταξύ δύο πόλεων όπου η αναμενόμενη κίνηση ακολουθεί κατανομή Poisson με ρυθμό 30 calls/min. Η διάρκεια των κλήσεων είναι ανεξάρτητες εκθετικά κατανεμημένες τυχαίες μεταβλητές με μέση τιμή 3 min. Πόσα κυκλώματα θα πρέπει να παρέχει η εταιρεία ώστε να εγγυηθεί ότι η πιθανότητα απόρριψης κλήσης (blocking) (επειδή όλα τα κυκλώματα είναι κατειλημμένα) είναι μικρότερη του 1%

Λύση

Θεωρούμε ένα σύστημα M/M/m/m όπου m είναι ο αριθμός κυκλωμάτων (γραμμών) που παρέχει η εταιρεία. Πρέπει να βρούμε τον μικρότερο αριθμό m για τον οποίο $p_m<0.01$ όπου p_m

$$p_m = \frac{(\lambda/\mu)^m / m!}{\sum_{n=0}^{m} (\lambda/\mu)^n / n!}$$

Έχουμε λ = 30 calls/min, 1/μ = 3 min, άρα

$\rho = \lambda/\mu = 30 \cdot 3 = 90$ Erlangs.

Με αντικατάσταση στην παραπάνω σχέση υπολογίζουμε την τιμή του m.

ΣΥΣΤΗΜΑΤΑ ΑΝΑΜΟΝΗΣ

Άσκηση II.a: Αξιολόγηση Δρομολογητή

Ένας δρομολογητής πακέτων μπορεί να επεξεργάζεται πακέτα με μέσο ρυθμό 300 pkts/sec (packets per second). Τα πακέτα καταφθάνουν στον δρομολογητή κατά μέσο όρο με ρυθμό 200 pkts/sec. Αν μοντελοποιήσουμε το σύστημα σαν M/M/1

- Μέσος # πακέτων στο σύστημα:

  $\boldsymbol{N = \lambda/(\mu-\lambda)}$ = 200/(300-200) = 2 πακέτα

- Μέσος # πακέτων σε αναμονή:

  $\boldsymbol{N_Q = \lambda^2/[\mu(\mu-\lambda)]}$ = 200^2/([300(300-200)] = 1.33 πακέτα

- Μέσος χρόνος πακέτου στο σύστημα:

  $\boldsymbol{T = 1/(\mu-\lambda)}$ = 1/(300-200) = 0.01 sec = N/λ (Νόμος Little)

- Μέσος χρόνος αναμονής (στην ουρά):

  $\boldsymbol{W = \lambda/[\mu(\mu-\lambda)]}$ = 2/300 = 0.00666 sec = N_Q/λ (ΝόμοςLittle)

- Πιθανότητα να είναι ο εξυπηρετητής απασχολημένος

  $\boldsymbol{\rho = \lambda/\mu}$ = 2/3

- Πιθανότητα το σύστημα να είναι άδειο

  $\boldsymbol{P_0=1- \rho}$ = 0.333

ΣΥΣΤΗΜΑΤΑ ΑΝΑΜΟΝΗΣ
Άσκηση II.b: Βελτίωση Επίδοσης

Για να βελτιώσουμε την απόδοση του συστήματος έχουμε δύο επιλογές:

- Να εγκαταστήσουμε ένα γρηγορότερο επεξεργαστή (αντικαθιστώντας τον παλαιό)
- Να εγκαταστήσουμε και έναν δεύτερο εξυπηρετητή

Λύση

1η Επιλογή
Αντικατάσταση του υπάρχοντος εξυπηρετητή με άλλον γρηγορότερο με ρυθμό μ = 400 pkts/second
Επαναπροσδιορισμός της απόδοσης του συστήματος M/M/1 με λ = 200 pkst/sec, μ = 400 pkts/sec και άπειρο μήκος ουράς

2η Επιλογή
Έχουμε ένα σύστημα με δύο εξυπηρετητές, ο κάθε ένας με ρυθμό μ = 200 pkts/sec. Η άφιξη των πακέτων γίνεται σε μία ουρά (buffer) με ρυθμό λ = 200 pkts/second και μετά δρομολογούνται στον πρώτο εξυπηρετητή που είναι ελεύθερος. Το σύστημα αυτό το μοντελοποιούμε σαν M/M/2 με άπειρο μήκος ουράς

ΣΥΣΤΗΜΑΤΑ ΑΝΑΜΟΝΗΣ

Άσκηση III: Μοντελοποίηση - Σύγκριση

Δύο υπολογιστές επικοινωνούν με μια γραμμή 64 kbps (kbits/sec) και υποστηρίζει 8 συνόδους (sessions). Αν το μέσο μήκος πακέτου είναι 150 bytes, ο ρυθμός άφιξης ανά σύνοδο (arrival rate/session) είναι 4 packets/second και ακολουθεί Poisson κατανομή, και ο χρόνος εξυπηρέτησης πακέτου είναι εκθετικά κατανεμημένος:

Είναι καλύτερα το δίκτυο να παρέχει σε κάθε σύνοδο το δικό της αφιερωμένο (dedicated, αποκλειστική πρόσβαση) 8 kbps κανάλι, ή είναι προτιμότερο όλες οι σύνοδοι να μοιράζονται όλη τη χωρητικότητα της γραμμής?

Θεωρείστε ότι ο χρόνος καθυστέρησης του πακέτου είναι το πιο σημαντικό κριτήριο.

Λύση

Ας θεωρήσουμε πρώτα το δίκτυο να παρέχει σε κάθε σύνοδο το δικό της (αποκλειστική πρόσβαση) dedicated 8 kbits/sec κανάλι. Τότε κάθε υποσύστημα μπορεί να μοντελοποιηθεί σαν ένα ξεχωριστό M/M/1 σύστημα με λ=4 packets/sec και ρυθμό εξυπηρέτησης 8 kbits/sec ή ισοδύναμα

$$\mu = \frac{8\times10^3}{150\times8}\frac{packets}{\sec} \Rightarrow \mu = 6.67\, packets/\sec \qquad T = \frac{1}{\mu-\lambda} = 0.375\sec onds = 375m\sec$$

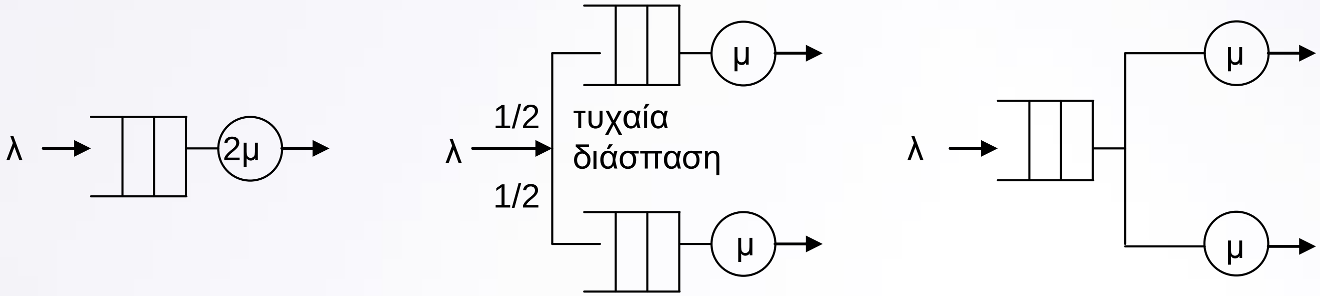
Θεωρώντας την περίπτωση όπου οι σύνοδοι μοιράζονται όλη τη χωρητικότητα της γραμμής τότε συγχωνεύουμε όλες τις συνόδους και μοντελοποιούμε το σύστημα σαν M/M/1 σύστημα: με λ=8*4=32 packets/second και ρυθμό εξυπηρέτησης 64 kbits/sec ή ισοδύναμα

$$\mu = \frac{64\times10^3}{150\times8} = 53.33\, packets/\sec \qquad T = \frac{1}{\mu-\lambda} = 0.0468\sec onds = 46.8m\sec$$

Προτιμότερη είναι η δεύτερη λύση αφού μειώνει την καθυστέρηση σημαντικά

ΑΣΚΗΣΗ IV

- Για κάθε ένα από τα συστήματα του σχήματος, υπολογίστε το μέσο χρόνο πακέτου στο σύστημα. Συγκρίνετε και επιλέξτε το καλύτερο και το χειρότερο.



Α) Μ/Μ/1 Β) Γ) Μ/Μ/2