

ΣΥΣΤΗΜΑΤΑ ΑΝΑΜΟΝΗΣ

Queuing Systems

Εισαγωγή

Βασίλης Μάγκλαρης
maglaris@netmode.ntua.gr

21/2/2018

ΠΕΡΙΕΧΟΜΕΝΑ (1/3)

http://www.netmode.ntua.gr/main/index.php?option=com_content&task=view&id=130&Itemid=48

1. Εισαγωγή

- Περιεχόμενα
- Γενική Περιγραφή Συστημάτων Αναμονής
- Τεχνικές Μελέτης & Αξιολόγησης Επίδοσης Συστημάτων Αναμονής
- Μοντέλα Τηλεπικοινωνιακών & Υπολογιστικών Συστημάτων

2. Εισαγωγή στη Θεωρία Ουρών.

- Χαρακτηριστικά & Παράμετροι Συστημάτων Αναμονής
- Μήκος Ουράς, Χρόνος Καθυστέρησης
- Νόμος Little

3. Γνώσεις από Θεωρία Πιθανοτήτων

- Εκθετική Κατανομή
- Κατανομή Poisson
- Ιδιότητα Απώλειας Μνήμης (Markov)
- Στοχαστικές Ανελίξεις

ΠΕΡΙΕΧΟΜΕΝΑ (2/3)

4. **Μοντέλο Γεννήσεων - Θανάτων (Birth - Death Processes).**
5. **Συστήματα Markov και Εξισώσεις Ισορροπίας**
 - Ανάλυση απλών ουρών M/M/1
6. **Άλλες ουρές Markov**
 - Μεταβάσεις Εξαρτώμενες από την Κατάσταση
 - Ουρές με Απώλειες (M/M/1/N)
 - Ουρές με Πολλαπλούς Εξυπηρετητές: M/M/m, M/M/m/K, M/M/m/m (Erlang – B)

ΠΕΡΙΕΧΟΜΕΝΑ (3/3)

7. Προσομοίωση Απλών Συστημάτων Αναμονής
8. Ανοικτά και Κλειστά Δίκτυα Ουρών
9. Ουρές με μη Εκθετική Εξυπηρέτηση
 - Ιδιότητα PASTA (Poisson Arrivals See Time Averages)
 - Απόδειξη Τύπου Little
 - Στοιχεία Θεωρίας Αναγεννήσεων (Renewal Theory)
 - Ενσωματωμένη Αλυσίδα Markov (Embedded Markov Chain)
 - Ανάλυση Ουράς M/G/1
10. Παραδείγματα & Εφαρμογές
 - Ανάλυση Υπολογιστικών Συστημάτων
 - Ανάλυση & Σχεδίαση Τηλεφωνικών Κέντρων
 - Ανάλυση Δικτύων Internet

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: $E(a) = 1/\lambda$ sec
- Μέσος ρυθμός εξυπηρέτησης πελατών: μ πελάτες/sec
- Χρόνος εξυπηρέτησης πελάτη – τυχαία μεταβλητή s , μέσος όρος: $E(s) = 1/\mu$ 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

ΜΕΘΟΔΟΙ ΑΞΙΟΛΟΓΗΣΗΣ ΜΟΝΤΕΛΩΝ ΑΝΑΜΟΝΗΣ

- **Αναλυτική ή/και αριθμητική αξιολόγηση** απλοποιημένων μοντέλων συστημάτων ΤΠΕ (Τεχνολογιών Πληροφορικής & Επικοινωνιών) για προσέγγιση βασικών παραμέτρων σχεδιασμού (π.χ. χωρητικότητας γραμμών, μεγέθους buffer, τοπολογία δικτύου)
- **Προσομοίωση (simulation)** για αξιολόγηση και προσδιορισμό παραμέτρων προτεινόμενων λύσεων ώστε να ικανοποιούνται προδιαγραφές QoS (Quality of Service)

ΜΕΘΟΔΟΙ ΑΞΙΟΛΟΓΗΣΗΣ ΜΟΝΤΕΛΩΝ ΑΝΑΜΟΝΗΣ

- **Αναλυτική ή/και αριθμητική αξιολόγηση** απλοποιημένων μοντέλων συστημάτων ΤΠΕ (Τεχνολογιών Πληροφορικής & Επικοινωνιών) για προσέγγιση βασικών παραμέτρων σχεδιασμού (π.χ. χωρητικότητα γραμμών, μεγέθους buffer, τοπολογία δικτύου)
- **Προσομοίωση (simulation)** για αξιολόγηση και προσδιορισμό παραμέτρων προτεινόμενων λύσεων ώστε να ικανοποιούνται προδιαγραφές QoS (Quality of Service) και QoE (Quality of Experience) με οικονομία σε CAPEX (Capital Expenses) & OPEX (Operational Expenses)
- **Εξομοίωση (emulation)** με διαμόρφωση εικονικών (virtual) υποδομών και πειραματισμό κάτω από τεχνητές ακραίες συνθήκες (επιθέσεις, καταστροφές...)
- **Δοκιμές** σε αληθινά συστήματα στο εργαστήριο και ανάλυση συμπεριφοράς κάτω από τυποποιημένα ενδεικτικά σενάρια χρήσης - benchmarking
- **Συνεχείς μετρήσεις** σε εγκατεστημένα συστήματα και αξιολόγηση εναλλακτικών σεναρίων αναβάθμισης - επανασχεδιασμού

ΠΑΡΑΜΕΤΡΟΙ (1/3)

– Ένταση φορτίου (traffic intensity)

Σε περίπτωση 1 ουράς, 1 εξυπηρετητή:

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

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

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:

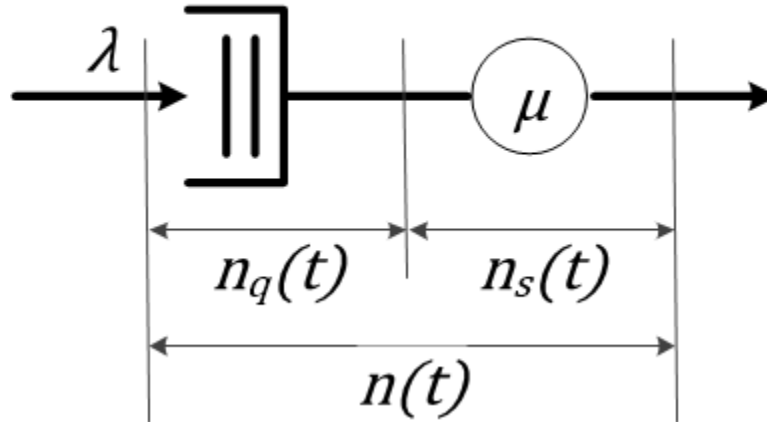
$$\gamma = \lambda(1 - P\{\text{blocking}\}) \leq \lambda, \quad \gamma < \mu$$

όπου $P\{\text{blocking}\}$ 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 (2/3)

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



– Μέσος ρυθμός απωλειών, ποσοστό απωλειών, πιθανότητα απώλειας πελάτη

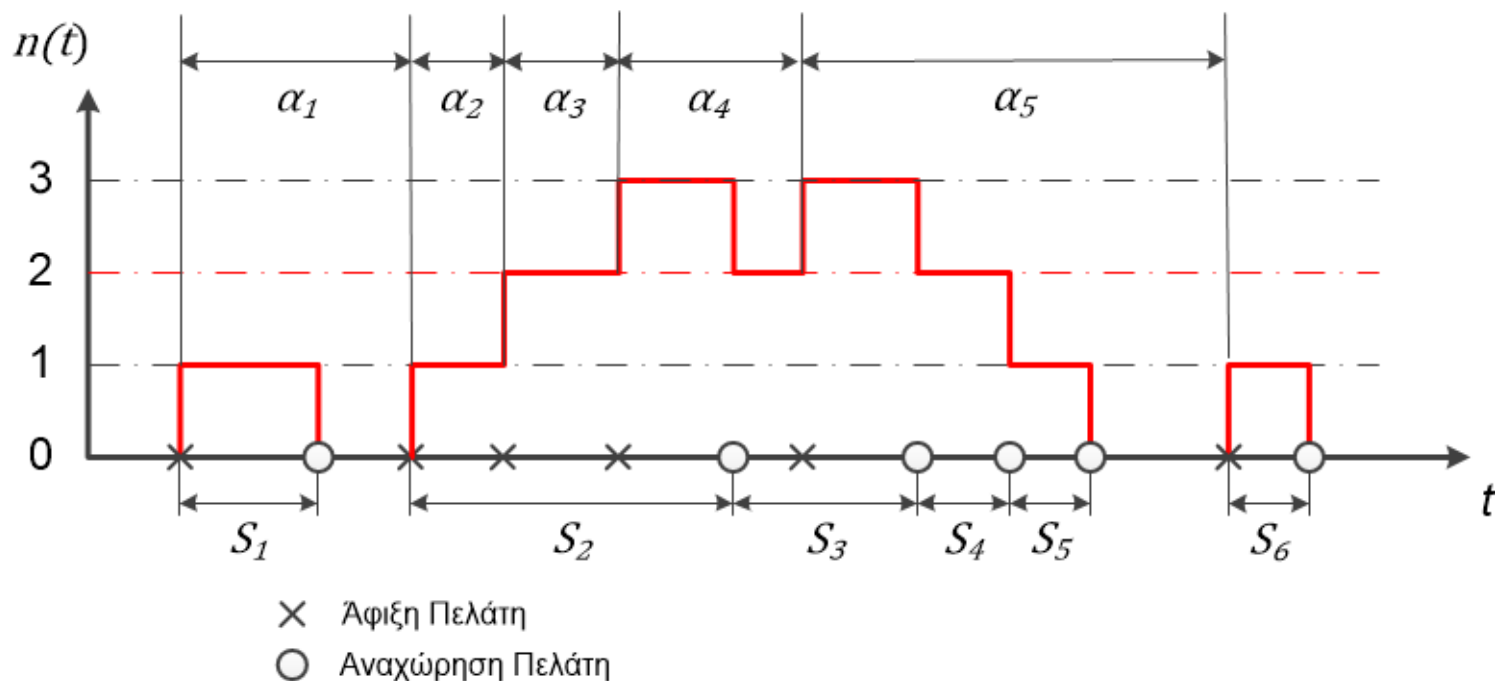
- Σε περίπτωση 1 ουράς, 1 εξυπηρετητή
Μέσος ρυθμός απωλειών: $\lambda - \gamma$
Ποσοστό απωλειών: $\frac{\lambda - \gamma}{\lambda} = P\{\text{blocking}\}$

– Βαθμός χρησιμοποίησης εξυπηρετητή (server utilization)

- Σε περίπτωση 1 ουράς, 1 εξυπηρετητή
 $u \triangleq \gamma / \mu$

ΠΑΡΑΜΕΤΡΟΙ (3/3)

Εξέλιξη Αριθμού Πελατών στο Σύστημα



– Αριθμός πελατών (κατάσταση)

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

– Μέσος αριθμός πελατών $E\{n(t)\}$

– Μέσος χρόνος καθυστέρησης (average time delay)

Μέσος χρόνος αναμονής (waiting time) + Μέσος χρόνος εξυπηρέτησης

$$E(T) = E(W) + E(s)$$