PhD Course: Performance & Reliability Analysis of IP-based mobile Communication Networks

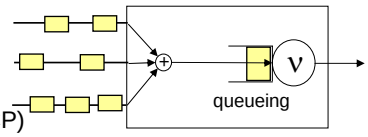
Søren Asmussen, Andrea Bondavalli, Henrik Schiøler, Hans-Peter Schwefel

- **Day 1** Basics Modeling approaches & bursty traffic models (HPS) [NJ14 4-117]
- **Day 2** Wireless Link Models & Network Models (HS) [???]
- **Day 3** Network Models cntd (HS), Stochastic Control (HPS) [A5-006]
- **Day 4** Simulation Techniques, Rare Event Simulation (SA) [A6-308]
- **Day 5** Dependability Modeling (AB) [NJ14, 3-117]

Organized by HP Schwefel & H Schiøler

Intro: Packet-Based Transport

- Advantages of Packet-Based Transport (as opposed to circuit switched)
  - Flexibility
  - Optimal Use of Link Capacities, Multiplex-Gain for bursty traffic
- Drawbacks
  - Buffering/Queueing at routers can be necessary
  - Delay / Jitter / Packet Loss can occur
  - Overhead from Headers (20 Byte IPv4, 20 Byte TCP)

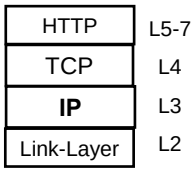


… and it makes performance modeling harder!!

Main motivation for Performance Modeling:

- Network Planning
- Evaluation/optimization of protocols/architectures/etc.

Challenges in Packet Switched Setting



Challenges in IP networks:

- Multiplexing of packets at nodes (L3)
- Burstiness of IP traffic (L3-7)
- Impact of Dynamic Routing (L3)
- Performance impact of transport layer, in particular TCP (L4)
- Wide range of applications → different traffic & QoS requirements (L5-7)
- Feedback: performance → traffic model, e.g. for TCP traffic, adaptive applications

Challenges in Wireless Networks:

- Wireless link models (channel models)
- MAC & LLC modeling
- RRM procedures
- Mobility models
- Cross layer optimization

Goal of this course

- Application of stochastic models often advantageous
  - Stochastic description of traffic
  - Stochastic mobility models
  - Use of randomization in protocols (e.g. MAC)
- Goal of the course:

  Provide an understanding of a selection of (mainly) stochastic modeling techniques in the context of network performance & reliability analysis for IP-based networks

Content

1. Intro & Review of basic stochastic concepts
   - Random Variables, Exp. Distributions, Stochastic Processes
   - Markov Processes: Discrete Time, Continuous Time, Chapman-Kolmogorov Eqs., steady-state solutions
   - Example: WLAN 802.11 DCF Modeling
2. Simple Qeueing Models
   - Kendall Notation
   - M/M/1 Queues, Birth-Death processes
   - Multiple servers, load-dependence, service disciplines
   - Transient analysis
   - Priority queues
3. Traffic Measurements and Analysis
4. Simple analytic models for bursty traffic
   - Markov modulated Poisson Processes (MMPPs)
5. MMPP/M/1 Queues and Quasi-Birth Death Processes

Note: slide-set will be complemented by some formulas and the mathematical derivations on the black-board!

Fundamental concepts I

- Probabilities
  - 'Random experiment' with set of possible results Ω
  - Axiomatic definition on event set $\wp(\Omega)$
    - $0 \le Pr(A) \le 1$; $Pr(\emptyset)=0$; $Pr(A \cup B)=Pr(A)+Pr(B)$ if $A \cap B=\emptyset$ [$A,B \in \wp(\Omega)$]
  - Conditional probabilities: $Pr(A|B)=Pr(A \cap B) / Pr(B)$
- Random Variables (RV)
  - Definition: $X: \Omega \rightarrow R^{(+)}$; $Pr(X=x)=Pr(X^{-1}(x))$
  - Probability density function f(x), cumulative distribution function $F(x)=Pr(X \le x)$, reliability function (complementary distr. Function) $R(x)=1-F(x)=Pr(X>x)$
  - Expected value, moments, coefficient of variation, covariance
  - Relevant Examples, e.g.:

Fundamental concepts II: Exponential Distributions

Important Case: Exponentially distributed RV

- Single parameter: rate λ
- Density function $f(x)= \lambda \exp(-\lambda x)$, $x>0$
- Cdf: $F(x)=1-\exp(-\lambda x)$, Reliability function: $R(x)=\exp(-\lambda x)$
- Moments: $E\{X\}=1/\lambda$; $Var\{X\}=1/\lambda^2$, $C^2=1$

Important properties:

- Memoryless: $Pr(X>x+y \mid X>x) = \exp(-\lambda y)$
- Properties of two indep. exponential RV: X with rate λ, Y with rate μ
  - Distribution of min(X,Y): exponential with rate $(\lambda+\mu)$
  - $Pr(X<Y)= \lambda/(\lambda+\mu)$

Fundamental concepts III: Stochastic Processes

- Definition of process (X_i) (discrete) or (N_t) (continuous)
- Simplest type: X_i independent and identically distributed (iid)
- Relevant Examples:
  - Inter-arrival time process: X_i
  - Counting Process:
    $N(t) = \max\{n \mid \Sigma_{i=1}^{n-1} X_i \leq t\}$, alternatively $N_i(\Delta) = N(i\Delta) - N([i-1]\Delta)$

Important Example: Poisson Process

- Assume i.i.d. exponential packet inter-arrival times (rate λ): $X_i:=T_i-T_{i-1}$
- Counting Process: Number of packets N_t until time t
  - $Pr(N_t=n)= (\lambda t)^n \exp(-\lambda t) / n!$
- Properties:
  - Merging: arrivals from two independent Poisson processes with rate λ_1 and λ_2 → Poisson process with rate $(\lambda_1+ \lambda_2)$
  - Thinning: arrivals from a Poisson process of rate λ are discarded independently with probability p → Poisson process with rate (1-p) λ
  - Central Limit Theorem: superposition of n independent processes results in the limit $n\to\infty$ in a Poisson process (under some conditions on the processes)

Markov Processes I:Discrete Time

- State-Space: finite or countable infinite, w/o.l.g. E={0,1,2,...,K} (K=∞ also allowed)
- Transition probabilities: p_{jk}=Pr(transition from state j to state k)
- X_i = RV indicating the state that the Markov process is in in step i
- 'Markov Property': State in step i only depends on state in step i-1
  - $Pr(X_i=k)=\Sigma_j [Pr(X_{i-1}=j)*p_{jk}]$
- Computation of state probabilities
  - Probability vector $\underline{\pi}_i$
  - Matrix notation for behavior of $\underline{\pi}_i$
  - Steady-state solution
- Assumption here: irreducibility (no transient states), non-periodicity, and homogenuity

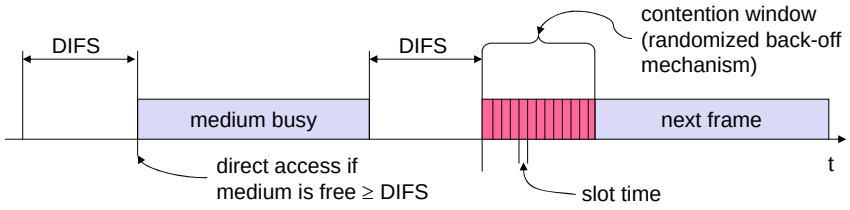
Markov Processes II: Continuous Time

- Defined by
  - State-Space: finite or countable infinite, w/o.l.g. E={0,1,2,...,K} (K=∞ also allowed)
  - Transition rates: μ_{jk}
    - Holding time in state j: exponential with rate $\Sigma_{k\neq j} \mu_{jk}$
    - Transition probability from state j to k: $\mu_{jk} / \Sigma_{l\neq j} \mu_{jl} =: \mu_{jk} / \mu_j$
- X_t = RV indicating the current state at time t; $\pi_i(t):=Pr(X_t=i)$
- 'Markov Property': transitions do not depend on history but only on current state $t_0<t_1<...<t_n$, $\forall i_0,i_1,...i_n,j \in E$

$$P\{X(t_{n+1}) = j \mid X(t_n) = i_n, \ldots, X(t_0) = i_0\} = P\{X(t_{n+1}) = j \mid X(t_n) = i_n\},$$

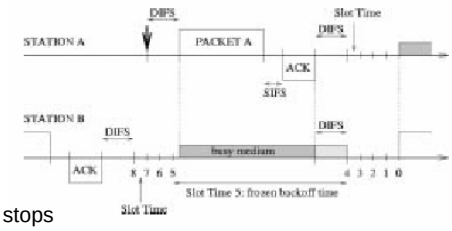
- Computation of steady-state probabilities
  - Chapman Kolmogorov Equations: $d\pi_i(t)/dt = - \mu_i \pi_i(t) + \Sigma_{j\neq i} \mu_{ji} \pi_j(t)$
  - Flow-balance equations, steady-state ($t\rightarrow\infty$): $\mu_i \pi_i = \Sigma_{j\neq i} \mu_{ji} \pi_j$
  - Matrix Notation: generator matrix $\underline{Q}$, probability vector $\underline{\pi}(t)$, limit $\underline{\pi}$
- Here: restriction to irreducible, homogeneous processes

Example Application: 802.11MAC - CSMA/CA



station ready to send starts sensing the medium (Carrier Sense)

– if the medium is free for the duration of an Inter-Frame Space (IFS, depends on service type)
  - the station starts sending
– if the medium is busy
  - the station has to wait for a free IFS
  - the station must additionally wait a random back-off time (multiple of slot-time)
  - if another station occupies the medium during the back-off time of the station, back-off timer stops



Random back-off

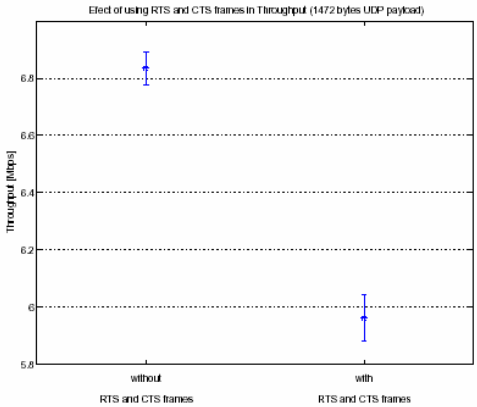
- If multiple stations are waiting for the medium to become available
  → potential for repeated collisions
- To break symmetry: randomization
  – Each station randomly choses integer counter value in [0,CW] (Contention Window)
  – when medium was idle for a slot-time → back-off counter is decreased
  – Transmission only started when counter=0 and medium idle
- Congestion Window Size
  – Initial setting: CW=7
  – When Collisions detected (missing ACKs)
    → CW is doubled
  – After successful transmission
    → CW set back to initial value

Access methods: variations

- DCF CSMA/CA (mandatory) – basic access method
  - collision avoidance via randomized „back-off“ mechanism
  - minimum distance between consecutive packets
  - ACK packet for acknowledgements (not for broadcasts)
- DCF w/ RTS/CTS (optional) – handshaking access method
  - avoids hidden terminal problem
- PCF (optional)
  - access point polls terminals according to a list

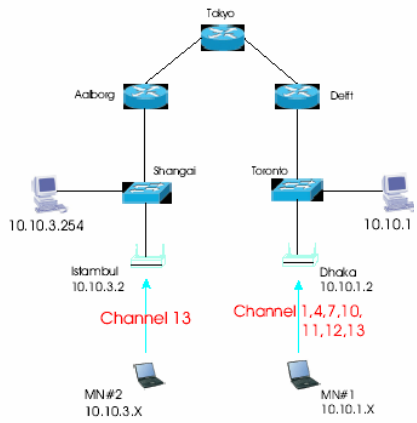
802.11b Performance Measurements

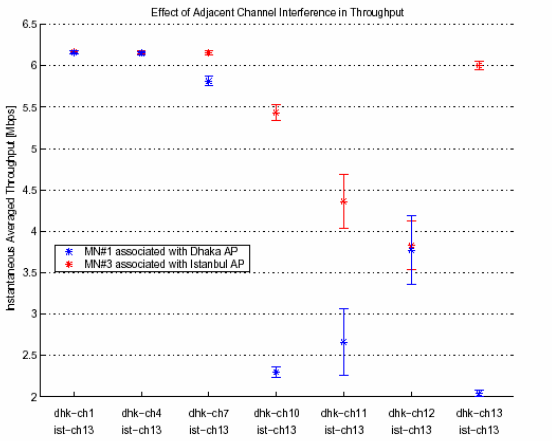
Scenario:
1 STA connected to AP
11Mb/s 802.11b WLAN



		UDP payload [bytes]		
Mean Throughput [Mbps]		700	1000	1472
RTS-CTS frames	Disable	5.0642	5.9442	6.8354
	Enable	4.2108	5.0605	5.9614
Performance Degradation [%]		16.8	14.9	12.8

802.11b Performance: Measurements (2)





802.11:Theoretical max Throughput

- based on deterministic analysis of time duration of steps during transmission
- without consideration of collisions and channel errors
- see J.Jun, P.Peddabachagari, M.Sichitiu: 'Theoretical Maximum Throughput of IEEE 802.11 and its Applications.' IEEE NCA, 2003.

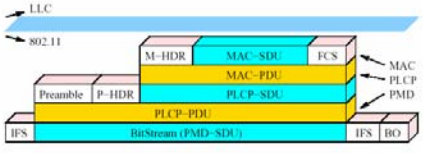


Figure 2. Overhead at different sublayers of IEEE 802.11

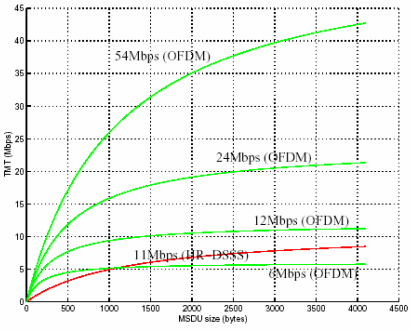


Figure 6. TMT curve for CSMA/CA - 11 Mbps HR-DSSS, OFDM

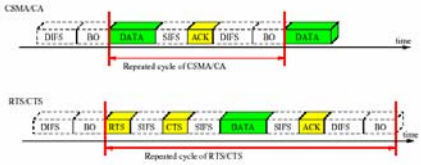


Figure 3. Timing diagram for CSMA/CA and RTS/CTS

Goal: Analytic Model for 802.11 DCF MAC

Model for throughput (+ delay + collision probabilities)

see G. Bianchi, 'Performance Analysis of the IEEE 802.11 DCF', IEEE J. sel. A. in Comm., 2000.

... including the following aspects:

- Medium-sharing by n stations
- Collisions between packet transmissions from different stations
- Random back-off
- Geometrically increasing congestion window upon collisions

... simplifying assumptions:

- All nodes always have data to transmit
- *Ideal channel (collisions only reason for corrupted transmissions)*
- *All stations can hear each other*
- ‚Time-slot synchronisation' between stations
- Fixed packet size

First Approach: Markovian Model

Discrete time model (at embedding points of state changes)

Necessary elements in state-space for each station:

- Congestion window size $W_0=CW_{min}=:W$, $W_1=2\,CW_{min}$, $W_2=4\,CW_{min}$, ..., $W_m=CWmax$
- Current value of back-off counter, $BC \in \{0,...,W_i-1\}$
- Number of states $S = W \sum_{i=0}^{m} 2^i = (2^{m+1}-1)\,W$

State space for all n stations: $\left((W^{(1)}, BC^{(1)}), (W^{(2)}, BC^{(2)}), ..., (W^{(n)}, BC^{(n)})\right)$

- Transmission states
  - Successful
  - Collision

Transition Probabilities

- Product space representation
  - If n independent processes → Kronecker product $\underline{P}_n=\underline{P}^{\otimes n}$ [dim $\underline{P}_n=S^n$]
    [in continuous time settings often Kronecker sums $\underline{A}\oplus\underline{B}=\underline{A}\otimes\underline{I} + \underline{I}\otimes\underline{B}$, see Day 3]
  - However, transitions depend on other stations (collision, successful?) !

Decoupling approximation

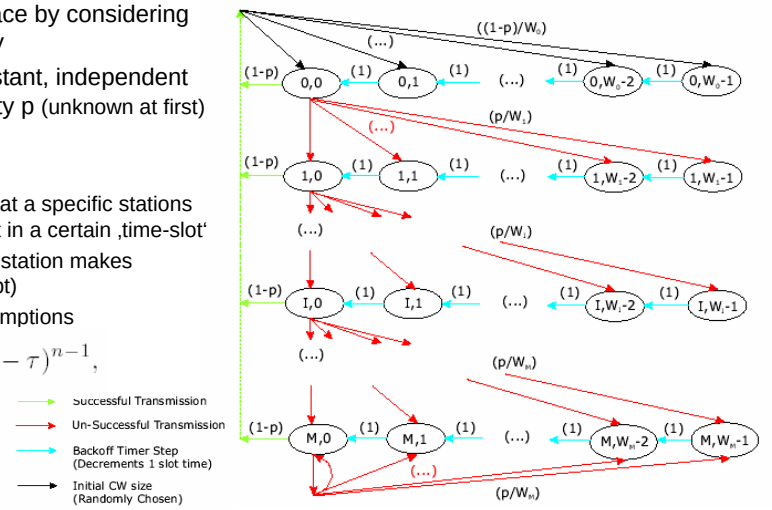
Avoid large state-space by considering single station only

→ Assumption: constant, independent collision probability p (unknown at first)

Computation:

- Let τ:=Probability that a specific stations attempts to transmit in a certain ‚time-slot'
- p = Prob.(collision | station makes transmission attempt)
- Independence assumptions

$$p = 1 - (1 - \tau)^{n-1},$$



Solution from fix-point equations

- From independence assumptions $p = 1 - (1 - \tau)^{n-1},$
- τ computed from steady-strate probabilities π(i,j) of Markov chain:

  $\tau = \sum_{i=0}^{M} \pi(i,0)$

  and π(i,j) from Markov chain balance equations (enumerate states in vector $\underline{\pi}$):

  $\underline{\pi}\,\underline{P}(p) = \underline{\pi}$
- Explicit solution from Markov chain structure

$$\tau = \frac{2(1-2p)}{(1-2p)(W+1) + pW(1-(2p)^m)}$$

→ Solution (p*, τ*) of non-linear equation system via numerical methods [existence of unique solution can be shown]

Throughput computation

- Overall aggregated throughput:

  Λ=E(payload info per time-slot) / E(duration of time-slot)

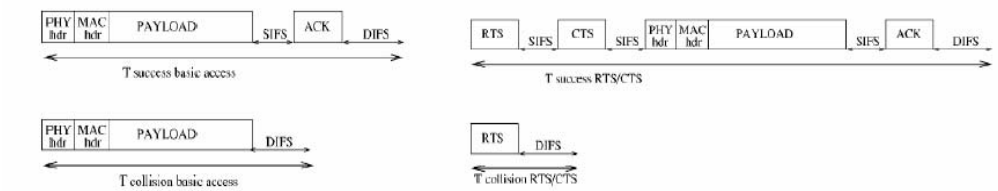
$$\Lambda = \frac{P_s P_{tr} \mathrm{DP}}{(1-P_{tr})\sigma + P_{tr} P_s T_s + P_{tr}(1-P_s) T_c} \qquad (7.7)$$

where,

- DP - Data Payload (we assume that all packets have the same size)
- T_s - Average time the channel is sensed busy because of a successful transmission
- T_c - Average time the channel is sensed busy by the non-colliding STAs during a collision
- σ - Slot-Time (specified by the standard)
- P_{tr} - Probability that there is at least one transmission in the considered slot time
- P_s - Probability that exactly one STA transmits on the channel conditioned on the fact that at least one STA transmits

Duration of transmission slots

- Depending on whether RTS/CTS activated

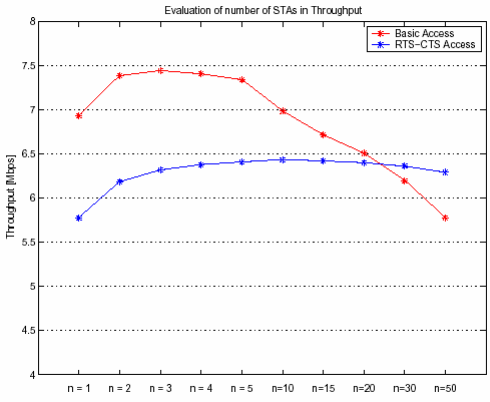


Analytic Results: RTS/CTS handshake

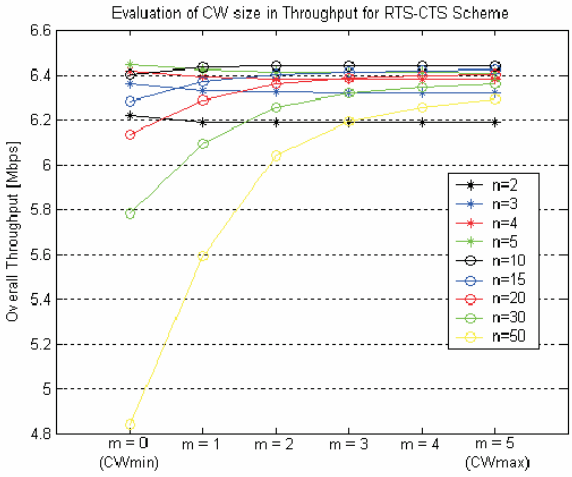
Scenario:

- 802.11b WLAN with 11Mb/s raw throughput
- Infrastructure setting: one AP, n stations

See Master Thesis of Rui Martins



Analytic Results: Impact of Maximum Congestion Window



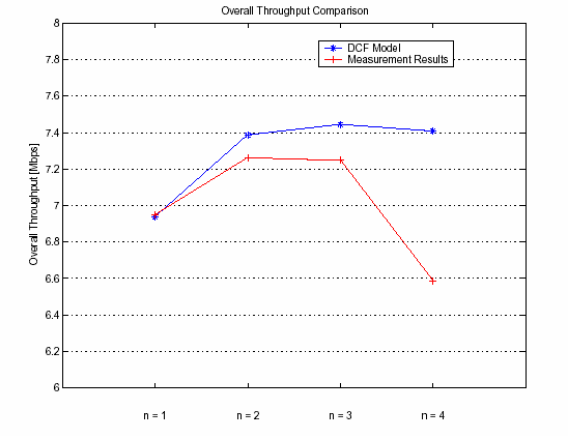
- m=0 corresponds to CW_max=32
- m=5 corresponds to CW_max=1024

Scenario:

- 802.11b WLAN with 11Mb/s raw throughput
- Infrastructure setting: one AP, n stations

See Master Thesis of Rui Martins

Comparison with Measurements



Technology	802.11b
Transmission Data Rate [Mbps]	11
Control frames transmission rate [Mbps]	2
CW_{min}	32
CW_{max}	1024 ($m = 5$)
$MSDU_{size}$ [bytes]	1514
PLCP PPDU frames	short

Modifications: non-ideal channel

- Transmission may be not successful, even if no collision, due to transmission errors
- First approach
  - Independent, constant error probability p_e (for complete frame transmission)
  - Modified fixpoint equations:

    $p=1-(1-\tau)^{n-1}(1-p_e)$ $\quad \tau = \frac{2(1-2p)}{(1-2p)(W+1)+pW(1-(2p)^m)}$

    And $P_s \rightarrow P_s(1-p_e)$ in enumerator of throughput formula
  - However, more accurate analysis requires to differentiate different slot times depending on whether
    - [RTS corrupted
    - CTS corrupted]
    - Data packet corrupted (no ACK sent → shorter slot)
    - ACK corrupted
- More complicated error models: correlated errors, see Gupta & Kumar, p. 286-290.

Challenges: multi-hop scenarios

- Even without mobility ...
- Consequence of more complex topologies
  - Number of stations in range may vary
    - Even worse: transmission range of different type of packets will vary

    - Further complications when potential asymmetric transmission ranges

  → Topological relations need to be described & careful consideration of different collision cases (RTS/CTS/DATA/ACK) is required
- For multi-hop transmissions:
  - subsequent transmission events of same packet are correlated
  - Starvation/buffer-overflow of intermediate nodes can occur (→tandem queue models)
  - End-to-end multi-hop delay and throughput analysis still research topic

References and Acknowledgments

Sources for this lecture:

- Rui Martins: 802.11b WLAN performance measurements and optimizations. Master thesis, Aalborg University, June 2004.
- Jangeun Jun, Pushkin Peddabachagari, Mihail Sichitiu: 'Theoretical Maximum Throughput of IEEE 802.11 and its Applications'. IEEE Conference NCA, Boston, 2003.
- G. Bianchi: 'Performance Analysis of the IEEE 802.11 Distributed Coordination Function', IEEE Journal on selected areas in communication, Vol 18, No 3, March 2000.
- Kumar, Gupta: A Performance Analysis of the 802.11 Wireless LAN Medium Access Control. Communications in Information and Systems, Vol. 3, No. 4, 2004.
- A. Graham: Kronecker Products and matrix calculus with applications. Ellis Horwood. 1981.

Other WLAN modeling approaches:

- Berger-Sabbatel, Duda, Gaudoin, Heusse, Rousseau: Fairness and its impact on delay in 802.11 Networks. Globecom 2004.

Exercises MAC modelling:

Implement a MATLAB program to solve the fix-point equations for the WLAN MAC model of Bianchi. Compare the resulting throughput with a model modification taking independent error probabilities into account. Investigate the impact of the channel errors on the throughput.

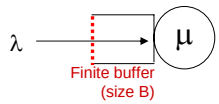
Content

1. Intro & Review of basic stochastic concepts
   - Random Variables, Exp. Distributions, Stochastic Processes
   - Markov Processes: Discrete Time, Continuous Time, Chapman-Kolmogorov Eqs., steady-state solutions
2. Simple Qeueing Models
   - Kendall Notation
   - M/M/1 Queues, Birth-Death processes
   - Multiple servers, load-dependence, service disciplines
   - Transient analysis
   - Priority queues
3. Traffic Measurements and Analysis
4. Simple analytic models for bursty traffic
   - Markov modulated Poisson Processes (MMPPs)
5. MMPP/M/1 Queues and Quasi-Birth Death Processes

Queueing Models: Kendall Notation

X/Y/C[/B] Queues (example: M/M/1, GI/M/2/10, M/M/10/10, ...)

- X: Specifies Arrival Process
  - M=Markovian →Poisson
  - GI=General Independent→ iid
- Y: Specifies Service Process (M,G(I),...)
- C: Number of Servers
- B: size of finite waiting room (buffer)
  [also counting the packet in service]
  - If not specified: B=∞
- Often also specified: service discipline
  - FIFO: First-In-First-Out (default)
  - Processor Sharing: PS
  - Last-in-first-out LIFO (preemptive or non-preemptive)
  - Earliest Deadline First (EDF), etc.

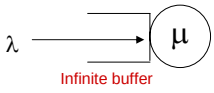


Scope here: Point-process models as opposed to fluid-flow queues

M/M/1 queue

- Poisson arrival of packets (first 'M' → Markovian) with rate λ
- Exponentially distributed service times of rate μ (second 'M')
- Single Server (1)
- FIFO service discipline

→ Q_t = Number of packets in system is continuous-time Markov Process



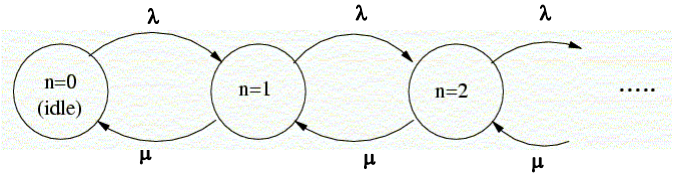
'Derived' Parameter:

- Utilization, $\rho = \lambda / \mu$: if $\rho \geq 1$, instable case (no steady-state q.l.d)

Performance Parameters

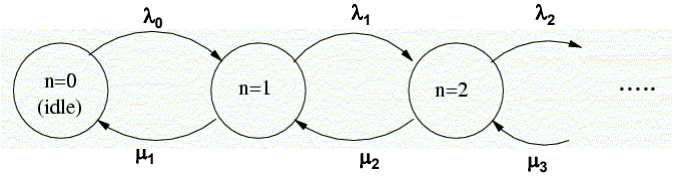
- Queue-length distribution: $\underline{\pi(t)}$, steady-state limit: $\pi = \lim \underline{\pi(t)}$ (if $\rho < 1$)
- Queue-length that an arriving customer sees
- Waiting/System time distribution
- Buffer-Overflow Probability for level B = Pr(arriving customers sees buffer occupancy B or higher)
- First Passage Times, Transient queue-length/overflow probabilities

M/M/1 queue: Performance



- Birth-Death Process
  - Probability of i packets in queue [using flow-balance equations]
    π_i := **Pr(Q=i) = (1-ρ)* ρ^i** , where $\rho = \lambda/\mu < 1$
  - Probability of idle server: **π_0 = (1-ρ)**
  - Average Queue-length: E{Q}= **ρ/(1-ρ)**
  - Average Delay (System Time): E{S}= E{Q}/ λ = 1/(μ-λ)
  - Buffer Overflow Probabilities (PASTA principle)
    $Pr(Q^{(a)} \geq B) = Pr(Q \geq B) = \rho^B$

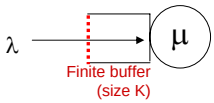
General Birth-Death Processes



- Steady-State Probabilities (from balance equations):
  $\pi_i := Pr(Q=i) = \pi_0 \prod_{k=0}^{i-1} \lambda_k \; / \; \prod_{k=1}^{i} \mu_k$

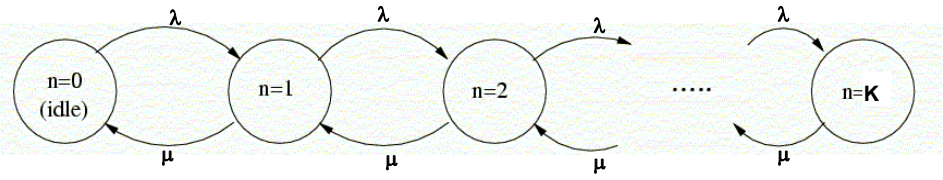
- Models in this class, e.g.
  - M/M/1/B
  - M/M/C, M/M/C/C
  - Load-dependent services, discouraged arrivals

Packet-based link model: M/M/1/K queue



- Assumptions
  - Poisson arrival of packets with rate λ
  - Exponentially distributed service times of rate μ
  - Single Server
  - Finite waiting room (buffer) for K packets
- Suitable e.g. for modeling 'bottleneck' link in packet-based wireless networks
  - [Full network models: see Day 4]

M/M/1/K queue: Performance



- From Birth-Death Process Theory:
  - Probability of i packets in queue

$$\pi_i := Pr(Q=i) = (1-\rho)/(1-\rho^{K+1}) * \rho^i \quad \text{, where } \rho = \lambda/\mu \neq 1,$$

i=0,…,K

  - Probability of packet loss:

$$p_{(loss)} = \pi_K = (1-\rho)/(1-\rho^{K+1}) * \rho^K$$

  - Average Delay:

$$\check{D} = 1/[\lambda (1-p_K)] * \rho/(1-\rho^{K+1}) * [(1-\rho^K)/(1-\rho) - K \rho^K]$$

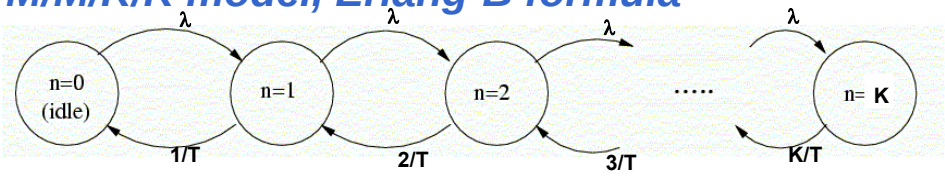
The circuit switched scenario

- K channels
- Users allocate one channel per call for certain call duration
- If all channels are allocated additional starting calls are blocked
- How many channels are necessary to achieve a call certain maximal blocking probability?

Common Model Assumptions:

- Calls are arriving according to a Poission Process (justified for large user population, limit theorems for stochastic processes) with rate λ
- Call durations are exponentially distributed with mean T (okay for voice calls)

Computation of blocking probabilities: M/M/K/K model, Erlang-B formula



- Finite Birth-Death Process:
  - Probability of i calls active

$$\pi_i := Pr(n=i) = \pi_0 (\lambda T)^i / i! \quad , i=1,\ldots,K$$

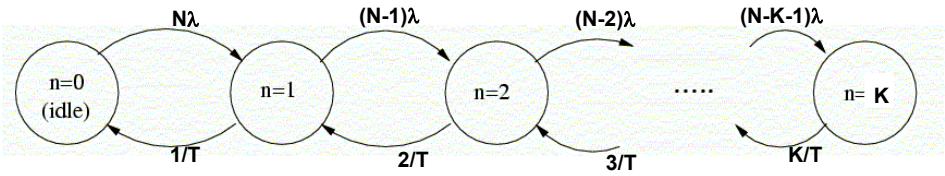
where $\pi_0 = 1/[\Sigma(\lambda T)^i / i!]$ (sum taken over i=0 to K)

  - Probability of blocked call:

$$p_{(Blocking)} = \pi_K = \pi_0 (\lambda T)^K / K!$$

[also known as Erlang-B formula]

Finite Number of Users: Engset Model



- Finite number of N independent users
  - Each indep. user with exponential time between calls (mean 1/λ)
  - Average call duration T
- Finite Birth-Death Process:
  - Probability of i calls active

$$\pi_i := Pr(n=i) = \pi_0 (\lambda T)^i \, N! \, / (N-i)! / i! \quad , i=1,\ldots,K$$

where π_0 from normalization

Exercise 1:

Use MATLAB to... (unless specified otherwise, use mu=1 for convenience)

1. (optional) Plot the Queue-length distribution for the M/M/1/K queue with rho in {0.5, 0.9, 1.0, 1.1, 2.0}. Plot a curve of the dropping probability for rho in [0;5], discuss its behavior for rho>1.
2. (optional) Implement the Q matrix for an M/M/1/K queue with rho=0.6, K=5, and plot transient probabilities for pi_0(t), given that the queue is empty at t=0.
3. One router in a network (to be modelled as M/M/1 queue) is apparently creating large delays. The admin decides to add a second routing processor to it. In principle, the admin now has the choice to upgrade to an M/M/2 system, but he could also configure separate queues for the new routing module (packets will be randomly directed to one of the two queues); alternatively, he could replace the routing processor with another one that is twice as fast. Which solution is performance-wise better? Plot the ratio of the mean queuelengths for varying arrival rate lambda.
4. DiffServ-like scenario: Assume a single server (rate mu) at which two types of packets, High priority and low-priority, arrive (according to a Poisson process with rate lambda_l and lambda_h). Two queues are used to store the packets, but low-priority packets are only serviced when there are no high-priority packets. Also, an arrival of a high-priority packet pre-empts a currently ongoing service of a low-priority packet.
   - Draw the Markov chain for such priority model for buffer-sizes 2 packets for HP, 3 packets for LP.
   - Implement a MTALB function which allows to set-up the Q matrix of the Markov chain for arbitrary lambda_l and mu, and for arbitrary buffer-size for the low-priority packets (the high priority buffer is fixed to size 2).
   - Compute the steady-state probability vector plus the loss probabilities for varying lambda_h.

Content

1. Intro & Review of basic stochastic concepts
   - Random Variables, Exp. Distributions, Stochastic Processes
   - Markov Processes: Discrete Time, Continuous Time, Chapman-Kolmogorov Eqs., steady-state solutions
2. Simple Qeueing Models
   - Kendall Notation
   - M/M/1 Queues, Birth-Death processes
   - Multiple servers, load-dependence, service disciplines
   - Transient analysis
   - Priority queues
3. Traffic Measurements and Analysis
4. Simple analytic models for bursty traffic
   - Markov modulated Poisson Processes (MMPPs)
5. MMPP/M/1 Queues and Quasi-Birth Death Processes

Daily Profiles: Stationarity, informally

- "A stationary process has the property that the mean, variance and autocorrelation structure do not change over time."
- Informally: "we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality)."

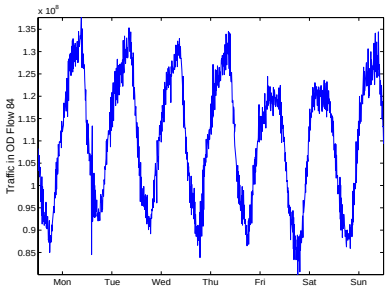
NIST/SEMATECH e-Handbook of Statistical Methods

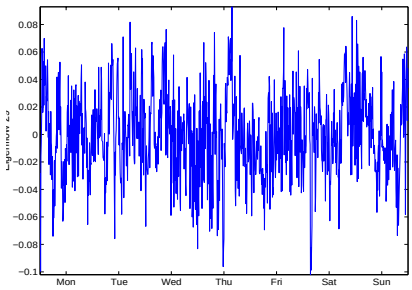
http://www.itl.nist.gov/div898/handbook/

Source: M. Crovella

The 1-Hour / Stationarity Connection

- Non-stationarity in traffic is primarily a result of varying human behavior over time
- The biggest trend is diurnal
- This trend can usually be ignored up to timescales of about an hour, especially in the “busy hour”





Source: M. Crovella

Traffic Modeling: A Reasonable Approach

- Fully characterizing a stochastic process can be impossible
  - Potentially infinite set of properties to capture
  - Some properties can be very hard to estimate
- A reasonable approach is to concentrate on two particular properties:

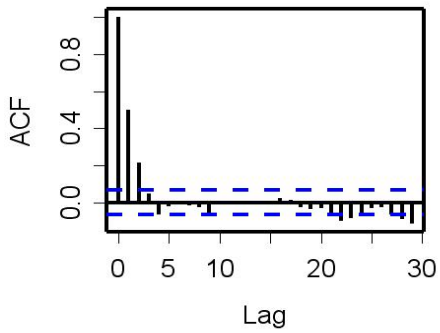
  marginal distribution and autocorrelation

Source: M. Crovella

Measuring Autocorrelation

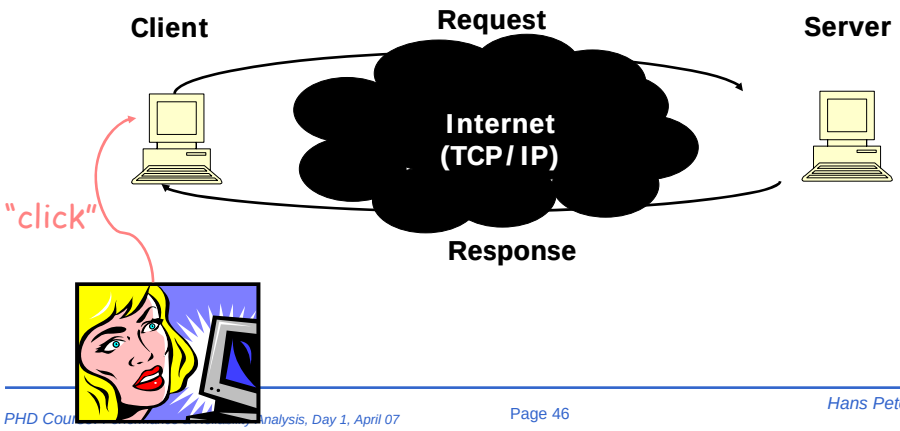
Coefficient of Autocorrelation (assumes stationarity):

$R(k) = Cov(X_n, X_{n+k})/Var[X_0] = (E[X_n X_{n+k}] - E^2[X_0]) / Var[X_0]$

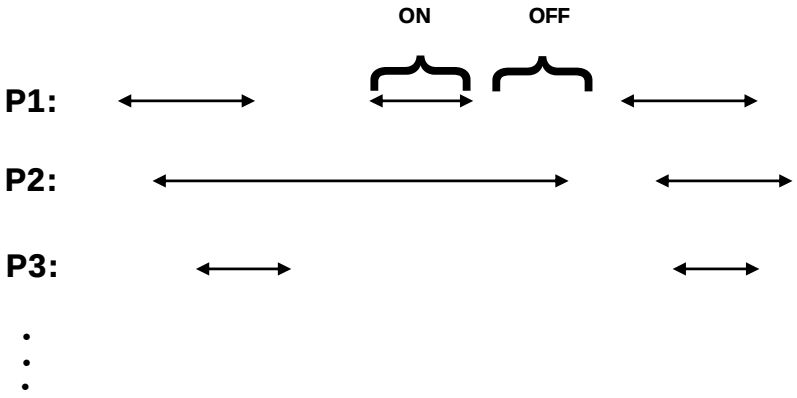


How Does Autocorrelation Arise?

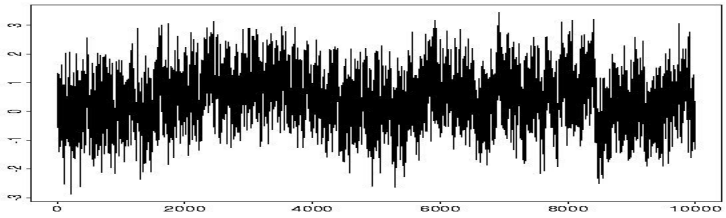
Network traffic is the superposition of *flows*

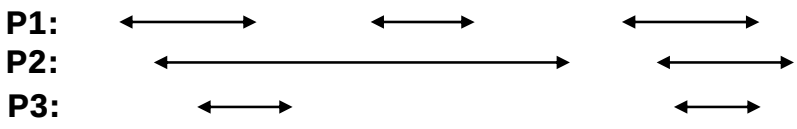


Why Flows?: Sources appear to be ON/OFF



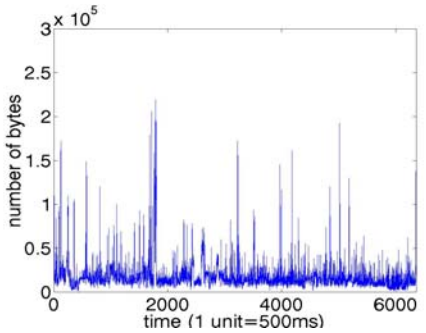
Superposition of ON/OFF sources ⇒ Autocorrelation

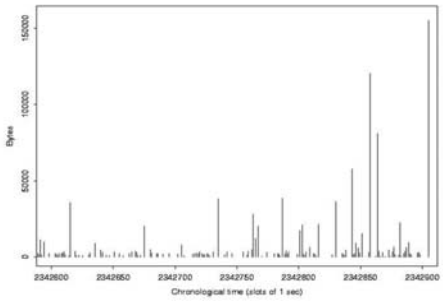


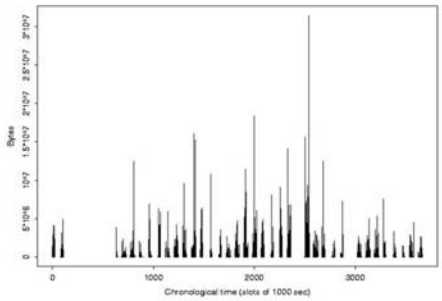


Low Multiplexed Traffic

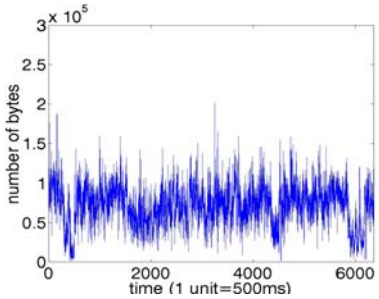
- Marginals: highly variable
- Autocorrelation: low

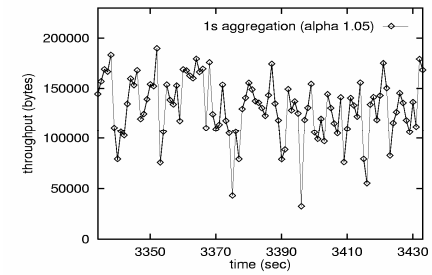


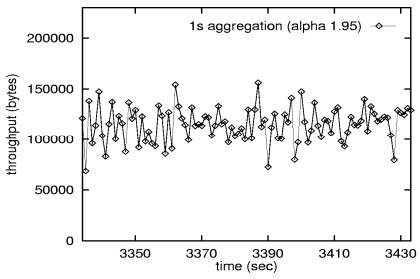




Highly Multiplexed Traffic







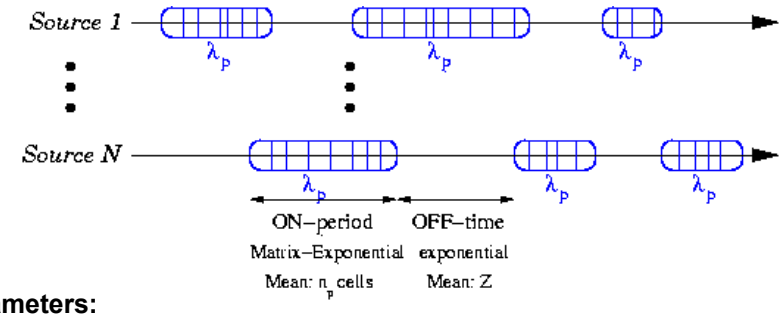
Models for bursty traffic

- Poisson assumption for packet arrivals may be applicable for highly aggregated traffic (core networks), but otherwise traffic tends to be bursty
  - High data rates in ftp download but less activity between downloads
  - http: activities after mouse-clicks
  - Video streaming: high data rates in frame transmissions
  - Interactive Voice: talk and silent periods
- Model Modifications:
  - Bulk Arrival processes
  - MMPPs

(optional) Bulk Arrival Models

- Queue-length at arrival instances increases not only by 1, but by a Random Variable B, the bulk-size
- Parameter set of model
  - Bulk arrival process, e.g. exponential with rate λ
  - Bulk-Size distribution: p_i (e.g. geometric)
  - Service rate (single packet)
- Steady-state solution for mean system time [Chaudhry & Templeton 83]:
  $E\{S\} = [E\{B\}+E\{B^2\}] / [2 E\{B\} \mu (1-\rho)]$
- E.g. $M^{(B)}/M/1$ queue with geometrically distributed B
  - Transition Diagram and $\underline{Q}$ Matrix for bulk-arrival model

Bursty Models: ON/OFF Models



Parameters:

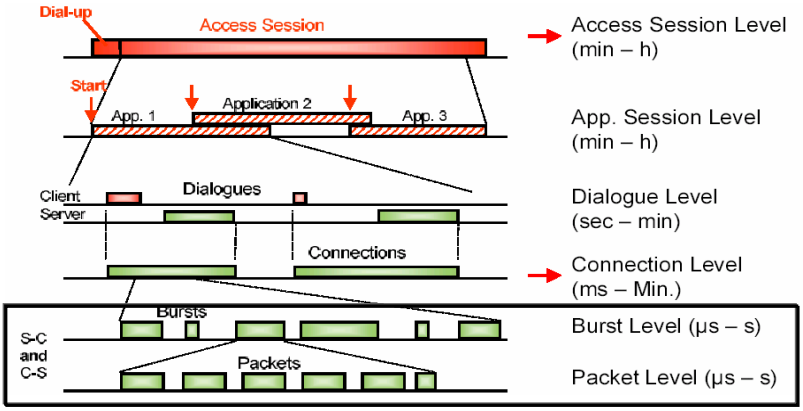
- N sources, each average rate κ
- During ON periods: peak-rate λ_p
- Mean duration of ON and OFF times

'bursty' traffic, when $\lambda_p >> \kappa$

$\kappa = \lambda_p$ ON/(ON+OFF)

Outlook: General hierarchical models

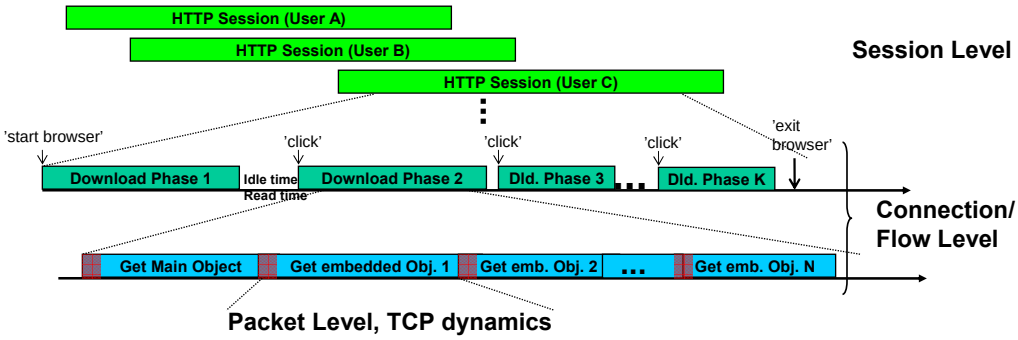
- Frequently used: Several levels with increasing granularity
  - E.g. 3 levels: sessions, connections, packets
  - Or: 5-level model:



Example: HTTP traffic model

- 'Main' objects contain zero or more embedded objects that the browser retrieves

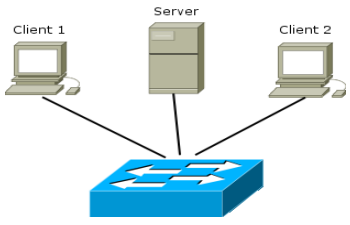
→ Correlated requests for embedded objects within retrieval of main object



- **Statistics:**
  - Session arrivals: Renewal process (Poisson)
  - Idle time: heavy-tail
  - # embedded objects: geometric (measurements e.g. mean 5)
  - Object size: heavy-tailed

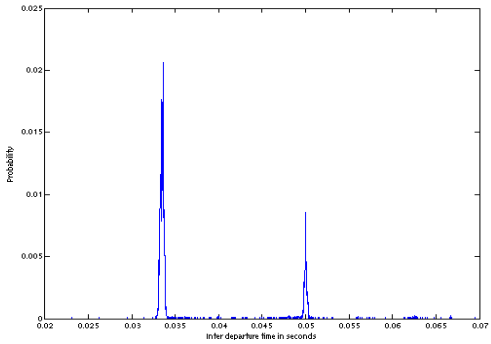
Example: Gaming Traffic

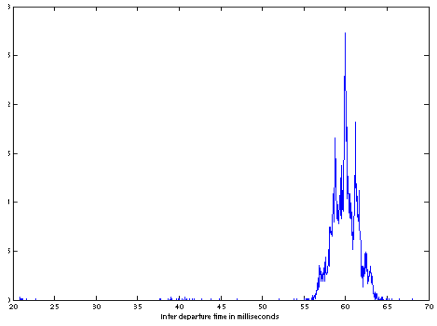
- Traffic Measurements of Counter-Strike Session
  - Architecture:
    - 2 clients, 1 server
    - Within switched 100Mb/s Ethernet
- Measurement of UDP traffic
  - Inter-packet time distribution (inter-departure time)
  - Packet Size distribution
  - Correlation properties (auto-, cross-)
- … for four different flows
  - Flow 1: Client 1 → Server
  - Flow 2: Client 2 → Server
  - Flow 3: Server → Client 1
  - Flow 4: Server → Client 2



Gaming: Inter-departure time distribution

- Upstream: two modes, around 33 ms and 50 ms
- Downstream: single mode, around 60ms

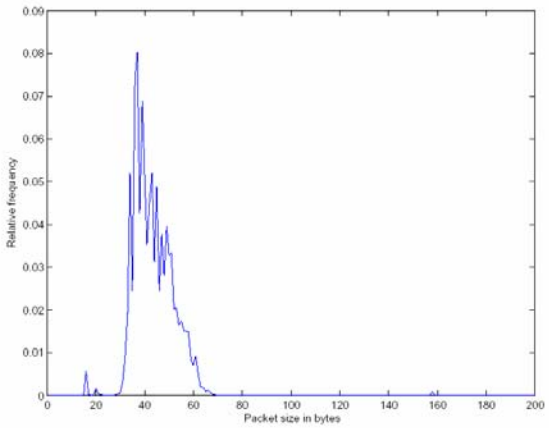




Gaming: Packet size distribution

Flow	Mean packet size	Standard deviation
Flow 1	43.47 bytes	8.81 bytes
Flow 2	40.32 bytes	8.56 bytes
Flow 3	54.59 bytes	70.39 bytes
Flow 4	56.19 bytes	81.13 bytes

Flow 1 packet size distribution



Gaming: Modeling of Gaming Traffic

- Downstream (server → client)
  - i.i.d. normally distributed
- Upstream (client → server)
  - Mix of two normal distributions
  - Normal distributions fitted with the mean and standard deviation of the traffic sample
  - Selection of 'short' or 'long' inter-departure time determined by discrete time Markov chain
    - Transition matrix derived from traffic samples

$$P_{flow1} = \begin{bmatrix} \text{slow} \rightarrow \text{slow} & \text{slow} \rightarrow \text{fast} \\ \text{fast} \rightarrow \text{slow} & \text{fast} \rightarrow \text{fast} \end{bmatrix} = \begin{bmatrix} 0.712406 & 0.28557 \\ 0.350855 & 0.649145 \end{bmatrix}$$

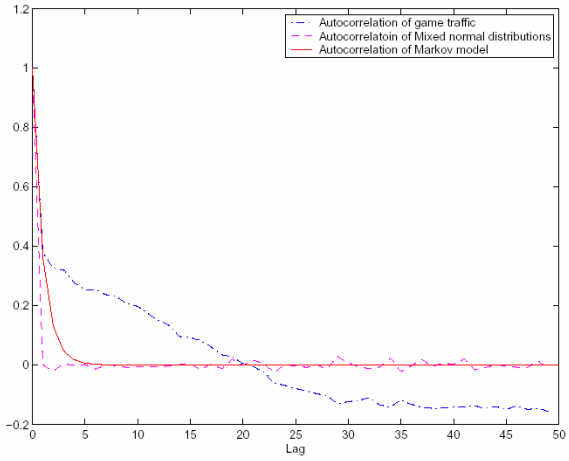
$$P_{flow2} = \begin{bmatrix} \text{slow} \rightarrow \text{slow} & \text{slow} \rightarrow \text{fast} \\ \text{fast} \rightarrow \text{slow} & \text{fast} \rightarrow \text{fast} \end{bmatrix} = \begin{bmatrix} 0.136926 & 0.863074 \\ 0.209484 & 0.788296 \end{bmatrix}$$

Gaming: Auto-correlation of inter-packet times

Coefficient of auto-correlation

→ Seemingly some periodic properties in game traffic

- Not achieved (of course) by two-state Markov model!



Content

1. Intro & Review of basic stochastic concepts
   - Random Variables, Exp. Distributions, Stochastic Processes
   - Markov Processes: Discrete Time, Continuous Time, Chapman-Kolmogorov Eqs., steady-state solutions
2. Simple Qeueing Models
   - Kendall Notation
   - M/M/1 Queues, Birth-Death processes
   - Multiple servers, load-dependence, service disciplines
   - Transient analysis
   - Priority queues
3. Traffic Measurements and Analysis
4. Simple analytic models for bursty traffic
   - Markov modulated Poisson Processes (MMPPs)
5. MMPP/M/1 Queues and Quasi-Birth Death Processes

Markov Modulated Poisson Processes

- $\underline{\underline{Q}}$: generator[1] of modulating Markov process.
- The Poisson arrival rates λ_i generated while the modulating process is in state i, summarized in the Matrix $\underline{\underline{\Lambda}}$:

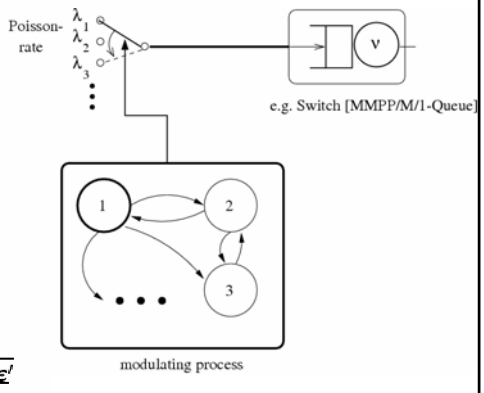
$$\underline{\underline{\Lambda}} = \begin{bmatrix} \lambda_1 & 0 & \dots \\ 0 & \lambda_2 & \ddots \\ 0 & \dots & \ddots \end{bmatrix}.$$

steady-state probability vector of modulating process,

$$\underline{\pi} \text{ with } \underline{\pi}\underline{\underline{Q}} = \underline{0} \text{ and } \underline{\pi}\underline{\varepsilon}' = 1.$$

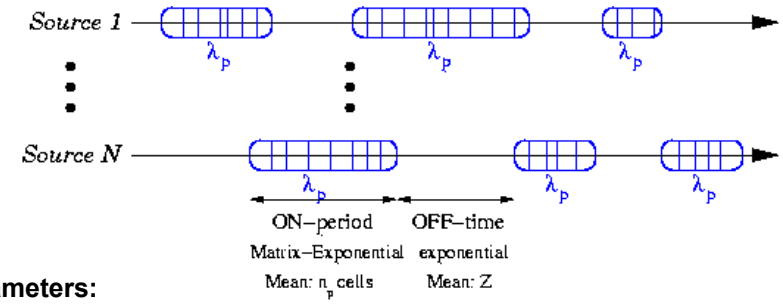
Average generated arrival rate: $\Lambda = \underline{\pi}\underline{\underline{\Lambda}}\underline{\varepsilon}'$

Average inter-arrival time: $E\{X\} = \frac{1}{\Lambda} = \frac{1}{\underline{\pi}\underline{\underline{\Lambda}}\underline{\varepsilon}'}$



Example: Single-Source ON/OFF model

Multiplexed ON/OFF Models



Parameters:

- N sources, each average rate κ
- During ON periods: peak-rate λ_p
- Mean duration of ON and OFF times

ON & OFF Times exponential → MMPP representation with N+1 states
(Exercises!)

Content

1. Intro & Review of basic stochastic concepts
   - Random Variables, Exp. Distributions, Stochastic Processes
   - Markov Processes: Discrete Time, Continuous Time, Chapman-Kolmogorov Eqs., steady-state solutions
2. Simple Qeueing Models
   - Kendall Notation
   - M/M/1 Queues, Birth-Death processes
   - Multiple servers, load-dependence, service disciplines
   - Transient analysis
   - Priority queues
3. Traffic Measurements and Analysis
4. Simple analytic models for bursty traffic
   - Markov modulated Poisson Processes (MMPPs)
5. MMPP/M/1 Queues and Quasi-Birth Death Processes

MMPP/M/1 Queues

- State space: <n,i>, n=# packets in queue, i=state of modulating process
- order states lexiographically

→ Generator matrix Block-tri-diagonal: *Ouasi-Birth-Death Process*

$$\hat{\underline{Q}} = \begin{bmatrix} \underline{\mathcal{Q}} - \underline{\mathcal{L}} & \underline{\mathcal{L}} & 0 & 0 & \cdots \\ \nu\underline{\mathrm{I}} & \underline{\mathcal{Q}} - \underline{\mathcal{L}} - \nu\underline{\mathrm{I}} & \underline{\mathcal{L}} & 0 & \cdots \\ 0 & \nu\underline{\mathrm{I}} & \underline{\mathcal{Q}} - \underline{\mathcal{L}} - \nu\underline{\mathrm{I}} & \underline{\mathcal{L}} & \cdots \\ 0 & 0 & \nu\underline{\mathrm{I}} & \underline{\mathcal{Q}} - \underline{\mathcal{L}} - \nu\underline{\mathrm{I}} & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

- Steady-state probability vector: $\hat{\underline{\pi}} = [\hat{\underline{\pi}}(0), \hat{\underline{\pi}}(1), \hat{\underline{\pi}}(2), \ldots]$. with $\hat{\underline{\pi}}\hat{\underline{Q}} = \underline{0}$
- Matrix Form of Balance Equations → Matrix Geometric Solution

$$\hat{\underline{\pi}}(k) = \hat{\underline{\pi}}(0)\underline{\mathrm{R}}^k$$

- From boundary conditions: $\hat{\underline{\pi}}(0) = \underline{\pi}(\underline{\mathrm{I}} - \underline{\mathrm{R}})$

Matrix-Geometric Factor: R

- From Balance equations. Sufficient condition for R:

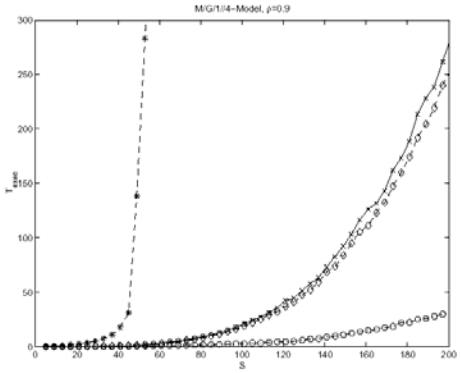
$$\underline{\mathrm{A}}_0 + \underline{\mathrm{R}}\underline{\mathrm{A}}_1 + \underline{\mathrm{R}}^2\underline{\mathrm{A}}_2 = \underline{0} \quad \text{with} \quad \underline{\mathrm{A}}_0 = \underline{\mathcal{L}},\ \underline{\mathrm{A}}_1 = \underline{\mathcal{Q}} - \underline{\mathcal{L}} - \nu\underline{\mathrm{I}},\ \underline{\mathrm{A}}_2 = \nu\underline{\mathrm{I}}.$$

- Performance Parameters:
  - Scalar queue-length probabilities $r(k) = \hat{\underline{\pi}}(k)\underline{\varepsilon}' = \underline{\pi}(\underline{\mathrm{I}} - \underline{\mathrm{R}})\underline{\mathrm{R}}^k\underline{\varepsilon}', \quad k = 0, 1, \ldots$
  - Average queue-length $E\{Q\} = \sum_{k=0}^{\infty} k\,\hat{\underline{\pi}}(k)\underline{\varepsilon}' = \underline{\pi}\underline{\mathrm{R}}(\underline{\mathrm{I}} - \underline{\mathrm{R}})^{-1}\underline{\varepsilon}'$
- How to solve quadratic matrix equation:
  - Simple iteration $\underline{\mathrm{R}}_0 = \underline{0}$

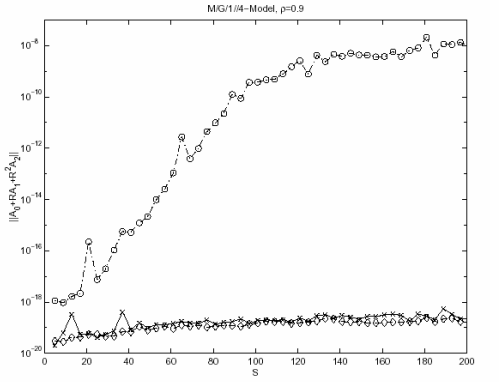
    $\underline{\mathrm{R}}_{n+1} = -(\underline{\mathrm{A}}_0 + \underline{\mathrm{R}}_n^2\underline{\mathrm{A}}_2)\underline{\mathrm{A}}_1^{-1}$
  - Spectral decomposition (not discussed here)

Computational aspects

- Simple iteration method (*) may lead to long processing times



- Numerical properties



(optional) MMPP/M/1/B Queues

- Finite buffer-space, loss model
- Generator matrix

$$\hat{\underline{Q}} = \begin{bmatrix} \overline{\underline{A}_1} & \underline{A}_0 & \underline{0} & \cdots & \\ \underline{A}_2 & \underline{A}_1 & \underline{A}_0 & \underline{0} & \cdots \\ \underline{0} & \ddots & \ddots & \ddots & \underline{0} \\ & \underline{0} & \underline{A}_2 & \underline{A}_1 & \underline{A}_0 \\ & & & \underline{A}_2 & \underline{A}_1 \end{bmatrix}.$$

$$\begin{aligned} \text{with} \quad & \underline{A}_0 = \underline{\mathcal{L}}, \quad \underline{A}_1 = \underline{\mathcal{Q}} - \underline{\mathcal{L}} - \nu\underline{I}, \quad \underline{A}_2 = \nu\underline{I}, \\ \text{and} \quad & \overline{\underline{A}_1} = \underline{A}_1 + \underline{A}_2 = \underline{\mathcal{Q}} - \underline{\mathcal{L}}, \quad \underline{A}_1 = \underline{A}_0 + \underline{A}_1 = \underline{\mathcal{Q}} - \nu\underline{I} \end{aligned}$$

- Mixed matrix geometric solution

$$\hat{\underline{\pi}}(i) = \underline{a}\mathbf{R}^i + \underline{b}\mathbf{S}^{B-i}, \quad i = 0, ..., B,$$

$$\begin{aligned} \underline{A}_0 + \mathbf{R}\underline{A}_1 + \mathbf{R}^2\underline{A}_2 &= \underline{0}, \\ \mathbf{S}^2\underline{A}_0 + \mathbf{S}\underline{A}_1 + \underline{A}_2 &= \underline{0}. \end{aligned}$$

- Approaches to obtain matrix factor S

Application: VoIP with silence supression

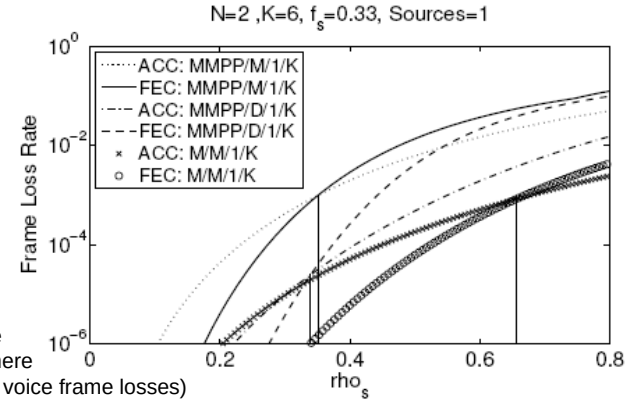
- VoIP packets: small payload, e.g. 2/3 of data volume due to header
- Usually using small buffers in the network (to avoid long delays)
- Packet loss can result from
  - unreliable (wireless links)
  - losses due to buffering
- Mechanisms to reduce impact of losses (at the price of additional delay) include
  - Forward error correction (FEC): redundancy, e.g. by transmitting payload N times
  - Frame Accumulation (ACC): multiple (N) voice frames in one IP packet → reduced volumes

What mechanism is better in congestion scenarios?

Results: VoIP with silence supression

… depends on bottleneck utilization and on traffic model!

- FEC is better for low utilization, ACC better for high load
- Utilization value at which cross-over occurs depends on traffic model, here
  - approx 65% for Poisson traffic
  - approx 35% for bursty ON/OFF traffic
- Service time distribution (M versus D) only little influence on location of cross-over point here (but large influence on resulting voice frame losses)



See S. Præstholm, HP Schwefel, S.V. Andersen, 'A Comparative Study of Forward Error Correction and Frame Accumulation for VoIP over congested networks', 20th International Teletraffic Congress (ITC 20), June 2007.

References

- Traffic Measurements
  - H. Gogl, 'Measurement and Characterization of Traffic Streams in High-Speed Wide Area Networks', VDI Verlag, 2001.
  - E. Matthiesen, J. Larsen, F. Olufsen: ‚Quality of Service for Computer Gaming – An Evaluation of DiffServ', Student Report, Aalborg University, Spring 04. www.control.auc.dk/~04gr832b/
- M. Neuts: 'Matrix geometric Solutions in Stochastic Models', John Hopkins University Press, 1981.
- M. Neuts: 'Structured stochastic matrices of M/G/1 type and their applications.' Dekker, 1989.
- G. Latouche, V. Ramaswami: 'Introduction to matrix-analytic methods in stochastic modeling'. ASA-SIAM Series on Statistics and Applied Probability 5. 1999.
- H.-P. Schwefel: 'Performance Analysis of Intermediate Systems Serving Aggregated ON/OFF Traffic with Long-Range Dependent Properties', Dissertation, TU Munich, 2000. [Appendices B,C,D,F]
- K. Meier-Hellstern, W. Fischer: 'MMPP Cookbook', Performance Evaluation 18, p.149-171. 1992.
- P. Fiorini et al.: 'Auto-correlation Lag-k for customers departing from Semi-Markov Processes', Technical Report TUM-I9506, TU München, July 1995.
- M. Crovella: 'Network Traffic Modeling', PhD lecture, Aalborg University, February 2004.

Summary

1. Intro & Review of basic stochastic concepts
   - Random Variables, Exp. Distributions, Stochastic Processes
   - Markov Processes: Discrete Time, Continuous Time, Chapman-Kolmogorov Eqs., steady-state solutions
2. Simple Qeueing Models
   - Kendall Notation
   - M/M/1 Queues, Birth-Death processes
   - Multiple servers, load-dependence, service disciplines
   - Transient analysis
   - Priority queues
3. Traffic Measurements and Analysis
4. Simple analytic models for bursty traffic
   - Markov modulated Poisson Processes (MMPPs)
5. MMPP/M/1 Queues and Quasi-Birth Death Processes

Exercises II:

- A network operator asks you to assist him in dimensioning his access router. The operator expects that during the daily busy hours, the router is used by N=10 users, each of them independently generates traffic according to an ON/OFF process with exponential ON and OFF periods. On periods have mean ON=10s with a Poisson packet rate of λ_p=6pck/sec for a single user. OFF periods show mean Z=20 seconds. The aggregated traffic stream of N=10 users can be described by an MMPP with K=11 states.
- a. Determine the Q and the L matrix of the MMPP (in MATLAB).
- b. Compute the stationary probability vector pi of the MMPP. What is the average packet rate generated by the MMPP?
- c. Determine the queue-length distribution of an MMPP/M/1 queue with service rate nu=30pck/sec via the following steps.
- i. Compute the coefficient matrices A0,A1,A2 for the quadratic matrix equation for the rate matrix R.
- ii. Solve the quadratic matrix equation via the simple iterative method from the lecture. Measure the time that Matlab needs to do so.
- iii. (Optional) Use spectral decomposition to solve for R and compare the run-time for that algorithm. Compare the obtained R matrix with the previous one.
- iv. Use R to compute and plot the queue-length probabilities and the average queue-length.