Slide supporting material

Lesson 3: Random Variables, Stochastic Processes; Traffic Engineering, QoS

Giovanni Giambene

Queuing Theory and Telecommunications: Networks and Applications
2nd edition, Springer

Random Variables and Probability Generating Functions

Random Variables

- Taxonomy:
  - **Continuous value**: domain Ω
    - Real axis (also a semi-axis)
    - Segment
  - **Discrete values**: domain {1, 2, …}
    - Finite values
    - Infinite values

Probability density function (pdf) $f_X(x)$ and Probability Distribution Function (PDF) $F_X(x)$:

$$\text{Prob}\{X \leq x\} = \int_{\Omega} f_X(x)dx$$

Probability mass function:

$$\text{Prob}\{N = k\} = p(k)$$

Summary of Discrete Distributions

Name	Probability mass function, p(k)	Mean	Variance
Discrete Uniform	$p(k)=\dfrac{1}{N}, \quad k=1,2,\ldots,N$	$\dfrac{N+1}{2}$	$\dfrac{N^2-1}{12}$
Bernoulli	$p(k)=\begin{cases} p, & k=1 \\ 1-p, & k=0 \end{cases}$	p	$p(1-p)$
Binomial	$p(k)=\binom{N}{k}p^k(1-p)^{N-k}, \quad k=0,1,\ldots,N$	Np	$Np(1-p)$
Modified Geometric	$p(k)=p(1-p)^{k-1}, \quad k=1,2,\ldots$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Poisson	$p(k)=\dfrac{\rho^k}{k!}e^{-\rho}, \quad k=1,2,\ldots$	ρ	ρ

Summary of Continuous Distributions

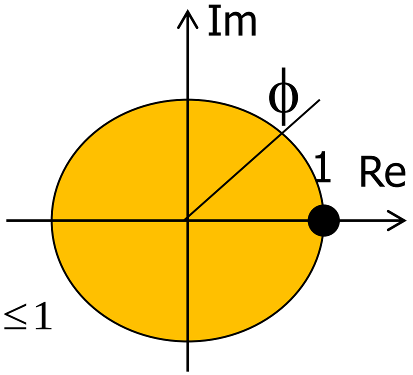
Name	Probability density function, pdf	Mean	Variance
Uniform $a \le x \le b$	$f(x)=\begin{cases}\frac{1}{b-a}, & a \le x \le b \\ 0, & otherwise\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential $x \ge 0$	$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x \ge 0 \\ 0, & otherwise\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal $-\infty \le x \le +\infty$	$f(x)=\frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$	μ	σ^2
Pareto $x \ge k$	$f(x)=\frac{\gamma k^{\gamma}}{x^{\gamma+1}} \quad x \ge k$	$\frac{\gamma k}{\gamma-1}$	$\frac{\gamma k^2}{(\gamma-1)^2(\gamma-2)}$

Discrete Random Variables

Probability Generating Function (PGF) for Discrete Random Var.

- The PGF is a transform for integer-valued discrete random variables having all the same sign (e.g., N) for which we know the probability mass function [e.g., Prob{N = k}]. The PGF is defined in the **complex domain** (variable $z \in C$) and is similar to a z-transform:

$$N(z) = E\left[z^N\right] = \sum_k z^k \mathrm{Prob}\{N = k\}, \qquad \text{for } |z| \le 1$$



- Basic properties:
  - N(z) is a power series with non-negative coefficients (probab.)

$$N(z=1) = \sum_k \mathrm{Prob}\{N = k\} = 1 \quad (normalization) \qquad N(z=0) = \mathrm{Prob}\{N = 0\} \le 1$$

$$|N(z)| \le 1 \quad \text{for} \quad |z| \le 1 \quad (bound \;\; condition)$$

  - A complex function is characterized by a **radius of convergence** ϕ: a complex function (z domain) is convergent for $|z| < \phi$ and diverges for $|z| > \phi$. On the circle $|z| = \phi$ there is at least one singularity. On the basis of the bound condition, the ϕ **value of the PGF must be at least one**: a PGF is convergent inside and on the unit disc $|z| \le 1$.
    - In z = 1 there can be a singularity that can be removed (Abel theorem).

Special Cases of PGFs for Deterministic Values

- It is interesting to note that also "1" can be seen as z^0 and therefore it is the PGF or the deterministic value "0".
- Moreover, also z and z^2 are PGFs of the deterministic values "1" and "2", respectively.

Abel Theorem

- A PGF is a power series with non-negative coefficients.
- In the **limiting case of a PGF with a radius of convergence just equal to 1**, the Abel theorem can be applied to prove that N(z) has a finite limit for $z \to 1^-$ and due to the normalization condition the value of this limit must be equal to 1:

$$\lim_{z \to 1^-} N(z) = 1$$

  - This theorem will be applied to the M/G/1 case, where the PGF P(z) of the state has a removable singularity by means of the Hôpital rule at z = 1.

Mean Value: Use of the First Derivative of PGF

- The **mean value** of a random variable X is defined as:

$$E[X] = \begin{cases} \int_{-\infty}^{+\infty} x f_X(x)\,dx & \text{for a continuous variable} \\ \sum_i x_i P\{X = x_i\} & \text{for a discrete variable} \end{cases}$$

- In the case of the discrete-value random variable X, we can obtain E[X] from the derivative of the PGF X(z) of X as:

$$X'(z) = \frac{d}{dz} \sum_k z^k \text{Prob}\{X = k\} =$$

$$= \text{under the assumption of series } \textit{uniform convergence} =$$

$$= \sum_k \frac{d}{dz} z^k \text{Prob}\{X = k\} = \sum_k k z^{k-1} \text{Prob}\{X = k\}$$

$$\Rightarrow X'(z = 1) = \sum_k k \text{Prob}\{X = k\} = E[X]$$

$$\Rightarrow \quad E[X] = X'(1)$$

PGFs provide an easy way to compute mean values.

Mean Square Value: Use of the Second Derivative of PGF

- The **mean square value** of a random variable X is defined as:

$$E\left[X^2\right]=\begin{cases}\displaystyle\int_{-\infty}^{+\infty} x^2 f_X(x)dx & \text{for a continuous variable}\\ \displaystyle\sum_i x_i^2 P\{X=x_i\} & \text{for a discrete variable}\end{cases}$$

- In the case of the discrete-value random variable X, we can obtain the mean square value of X from the second derivative of its PGF:

$$X''(z)=\frac{d}{dz}\sum_k kz^{k-1}\text{Prob}\{X=k\}=$$

$$=\text{under the assumption of series } \textit{uniform convergence}=$$

$$=\sum_k k\frac{d}{dz}z^{k-1}\text{Prob}\{X=k\}=\sum_k k(k-1)z^{k-2}\text{Prob}\{X=k\}$$

$$\Rightarrow X''(z=1)=\sum_k k^2\text{Prob}\{X=k\}-\sum_k k\text{Prob}\{X=k\}$$

$$\Rightarrow \quad E\left[X^2\right]=X''(1)+X'(1)$$

PGFs provide an easy way to compute mean square values.

- Finally, **variance** is obtained as:

$$\text{Var}[X]=E\left[(X-E[X])^2\right]=E\left[X^2\right]-\{E[X]\}^2=X''(1)+X'(1)-[X'(1)]^2$$

PGFs of Geometric and Poisson Distributions

- The discrete random variable N is **geometrically** distributed if its probability mass function can be represented as:

$$\mathrm{Prob}\{N=k\}=(1-q)q^k, \quad 0<q<1, \quad k=0,1,2,\ldots.$$

  - PGF:

Geometric series

$$N(z)=\sum_{k=0}^{+\infty} z^k \mathrm{Prob}\{N=k\}=\sum_{k=0}^{+\infty}(1-q)(zq)^k=\frac{1-q}{1-zq}$$

- The discrete random variable N is **Poisson** distributed if its probability mass function can be represented as:

$$\mathrm{Prob}\{N=k\}=\frac{\rho^k}{k!}e^{-\rho}, \quad \rho>0, \quad k=0,1,2,\ldots.$$

  - PGF:

Exponential series

$$N(z)=\sum_{k=0}^{+\infty} z^k \mathrm{Prob}\{N=k\}=\sum_{k=0}^{+\infty}\frac{(z\rho)^k}{k!}e^{-\rho}=e^{-\rho}\sum_{k=0}^{+\infty}\frac{(z\rho)^k}{k!}=$$

$$=e^{-\rho}\times e^{z\rho}=e^{\rho(z-1)}$$

PGFs of Bernoulli and Binomial Distributions

- The discrete random variable N is **Bernoulli** distributed if its probability mass function can be represented as:

$$N = \begin{cases} 1, & \text{with } probability\ p \\ 0, & \text{with } probability\ 1-p \end{cases}$$

  - PGF:

$$N(z) = 1 - p + zp$$

- The discrete random variable N is **binomially** distributed if its probability mass function can be represented as:

$$\text{Prob}\{N = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 < p < 1, \ k = 0, 1, 2, \ldots, n$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial coefficient

  - PGF:

$$N(z) = \sum_{k=0}^{n} z^k \text{Prob}\{N = k\} = \sum_{k=0}^{n} \binom{n}{k} (zp)^k (1-p)^{n-k} =$$

$$= \text{by invoking the binomial Newton formula} =$$

$$= (1 - p + zp)^n$$

N.B. Sum of iid Bernoulli random variables yields a Binomial random variable.

PGF: Sum of Independent Random Variables and Inversion

- Let us consider two discrete independent random variables: X with distribution Prob{X = k} and PGF X(z) and Y with distribution Prob{Y = h} and PGF Y(z). We need to characterize the PGF of W = X + Y
  - The PGF W(z) is related to X(z) and Y(z) as follows:

$$W(z) = X(z)Y(z)$$

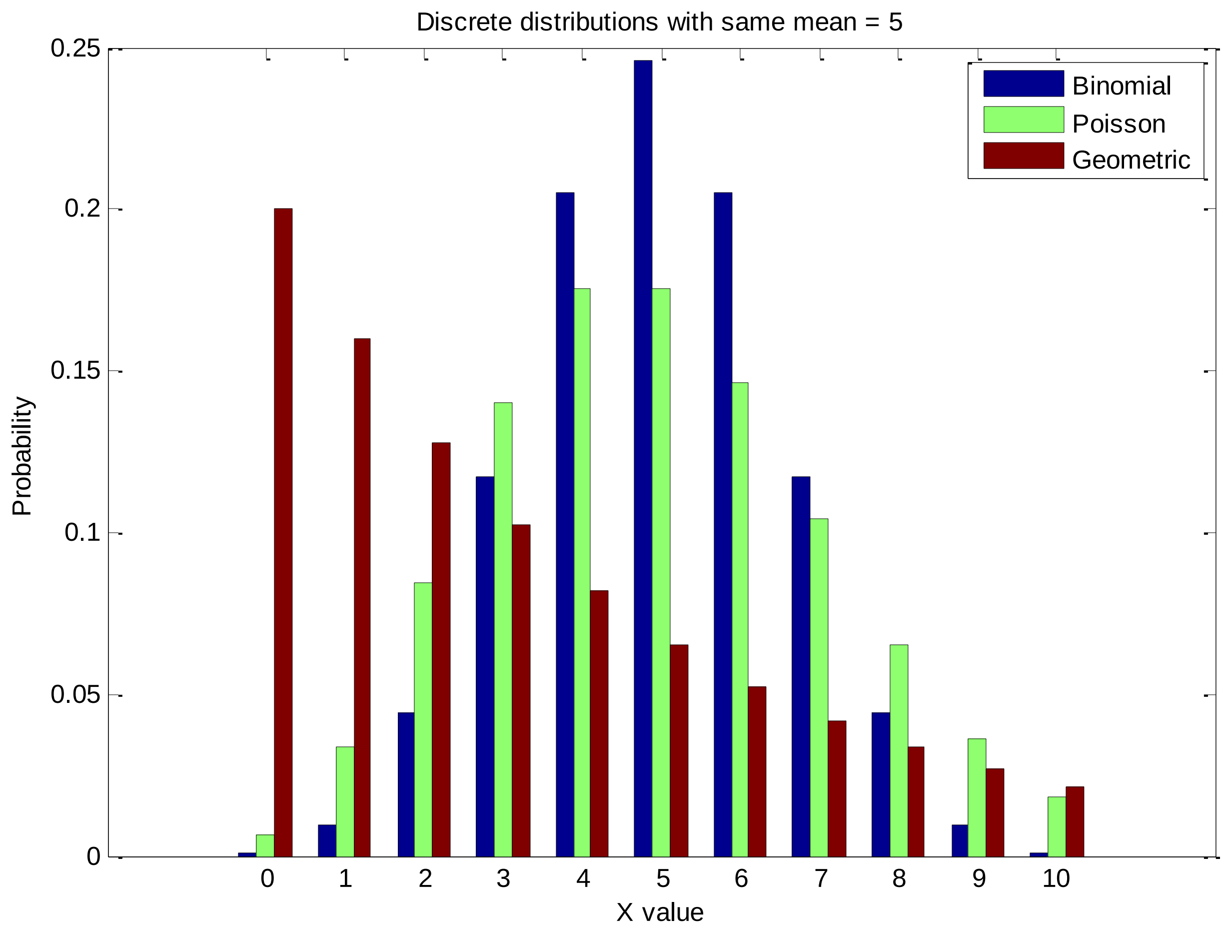
  - Special case: sum of **independent identically distributed (iid)** random variables with Bernoulli distribution, yielding a Binomial distribution.

- **Inversion**: for some derivations in the field of queuing theory a random variable N can be characterized in terms of its PGF N(z). It is therefore important to invert N(z) to derive the probability distribution Prob{N = k}. By definition, N(z) can be seen as a Taylor series expansion centered at z = 0 (i.e., MacLaurin series expansion). Hence, a simple inversion method can be based on the formulas to derive the coefficients of the **MacLaurin series expansion** as:

$$\text{Prob}\{N = k\} = \frac{1}{k!}\frac{d^k}{dz^k}N(z)\bigg|_{z=0}$$

This method can be easily implemented in Matlab® as shown in Lesson No. 19.

Comparison of Probability Mass Functions (Same Mean Value, 5)



Matlab© code:

```
X=0:1:10;
l=length(X);
A(1,X+1)=binopdf(X,10,0.5);
A(2,X+1)=poisspdf(X,5);
A(3,X+1)=geopdf(X,1/5);
bar(X,A')
legend('Binomial','Poisson','Geometric')
xlabel('X value')
ylabel('Probability')
title('Discrete distributions with same
mean = 5')
```

PGF: Compound Variables

- We consider independent discrete random variables N_i ($i = 1, 2, \ldots, M$) with probability mass functions Prob$\{N_i = k\}$ and PGFs $N_i(z)$. We are interested in characterizing the new random **compound variable** Y obtained as follows:

$$Y = \sum_{i=1}^{M} N_i$$

where M is a discrete random variable with probability mass function Prob$\{M = j\}$ and PGF $M(z)$.

If random variables N_i are iid: $N_i(z) = N(z)$ $\Rightarrow$ $$Y(z) = \sum_j [N(z)]^j \,\mathrm{Prob}\{M = j\} = M[N(z)]$$

- Special cases:
  - Sum of a geometric-distributed number of iid geometric-distributed variables yielding a geometric distribution;
  - Sum of a Poisson-distributed number of iid Bernoulli variables yielding a Poisson distribution.

Continuous Random Variables

The Exponential Distribution and the Memoryless Property

- The continuous random variable X is **exponentially distributed** if it has the following probability density function (pdf) and probability distribution function (PDF):

$$f_X(t) = \mu e^{-\mu t} \quad x \geq 0 \quad , \quad F_X(t) = 1 - e^{-\mu t} \quad t \geq 0$$

where $\mu > 0$ is the mean rate with the dimension of time^{-1}

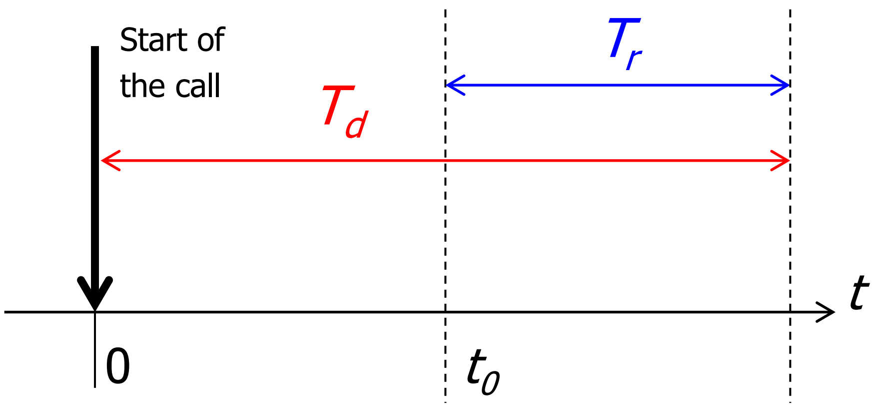
- Mean value $E[X] = 1/\mu$ and mean square value $E[X^2] = 2/\mu^2$

- Let us assume that T_d, exponentially-distributed with mean rate μ, is the duration of a phenomenon (e.g., phone call) started at time $t = 0$. We examine the same phenomenon at time $t = t_0$ and we assume that it is still active: $T_d > t_0$. We can prove that the **residual length** of the event, $T_r = T_d - t_0$, is still exponentially distributed with mean rate μ. This is the **memoryless property** of the exponential distribution. The exponential distribution is **the sole** continuous random variable for which the memoryless property is valid.

Memoryless Property of the Exp. Distribution: an Example

- Phone call started at time t = 0 and with exponentially-distributed length T_d (mean rate μ):

$$f_{T_d}(t) = \mu e^{-\mu t}, \quad t \geq 0$$



- Assuming $T_d > t_0$, the residual phone duration after time t_0, T_r, has the same distribution of T_d:

$$f_{T_r}(t) = \mu e^{-\mu t}, \quad t \geq 0$$

The min Property for the Exponential Distribution

- Let us consider the minimum of variables X_i for i = 1, 2, …, n that are independent with exponential distributions and rates μ_i. Then, the new random variable $\min_i \{X_i\}$ is still exponentially distributed with mean rate $\Sigma_i \mu_i$.
- Let us examine the case with n = 2. In general, we have random variables X and Y for which we know the joint pdf $f_{XY}(x,y)$ and, of course, the related marginal pdfs. We need to characterize the distribution $F_W(w)$ of $W = \min\{X, Y\}$:

$$
\begin{aligned}
F_W(w) &= \text{Prob}\{W \le w\} = \text{Prob}\{\{X \le w\} \overset{\text{or}}{\cup} \{Y \le w\}\} = \\
&= \text{Prob}\{X \le w\} + \text{Prob}\{Y \le w\} - \text{Prob}\{\{X \le w\} \overset{\text{and}}{\cap} \{Y \le w\}\} = \\
&= F_X(w) + F_Y(w) - \text{Prob}\{X \le w \overset{\text{and}}{,} Y \le w\} = \text{if X and Y st.indip.} = \\
&= F_X(w) + F_Y(w) - F_X(w) \times F_Y(w)
\end{aligned}
$$

Joint distribution

- If X and Y have independent exponential distributions $F_X(t) = 1 - e^{-\mu 1 t}$ and $F_Y(t) = 1 - e^{-\mu 2 t}$, then $F_W(w) = 1 - e^{-(\mu 1 + \mu 2)t}$.

The Pareto Distribution and Heavy-Tailed Distributions

- The continuous random variable X has a Pareto distribution if it has the following probability density and probability distribution functions:

$$f_X(x) = \frac{\gamma k^\gamma}{x^{\gamma+1}} \quad x \geq k \qquad F_X(x) = 1 - \left(\frac{k}{x}\right)^\gamma \quad x \geq k$$

where γ is a real positive number (*shape parameter*) and k is a positive translation term.

The mean value is finite for $\gamma > 1$: $E[X] = \frac{\gamma k}{\gamma - 1}$

The variance is finite for $\gamma > 2$: $Var[X] = \frac{\gamma k^2}{(\gamma-1)^2(\gamma-2)}$

Important note: the mean (variance) of a random variable can be **infinite depending on the γ value**!

- A random variable X is said to be **heavy-tailed** if its complementary distribution fulfills (definitely) the following condition that entails **infinite variance**:

$$\text{Prob}\{X > x\} \propto x^{-\gamma}, \text{ where } 0 < \gamma \leq 2$$

- The Pareto distribution is heavy-tailed if $0 < \gamma \leq 2$.

Examples of Use of Distributions in Telecommunications

- **Geometric distribution**: number of retransmission attempts with ARQ for the correct delivery of a packet, where each transmission attempt has an independent probability q to fail.
- **Poisson distribution**: number of sessions generated by a user for a given application in a given interval of time.
- **Bernoulli distribution**: describes the success / failure probability of a transmission attempt for a bit or a packet.
- **Binomial distribution**: describes the success / failure probability of a transmission attempt of a packet of bits with bit-to-bit memoryless error behavior (sum of independent identically-distributed Bernoulli variables).
- **Exponential distribution**: duration of a classical phone call / lifetime of an electronic equipment / lifetime of a subatomic particle.
- **Pareto distribution**: length of a file (discretized version of). It is used in the characterization of the self-similar Internet traffic.

Exercises/Homework

- We have to invert the probability generating functions in simple cases to determine the related probability mass functions:

$$A(z) = \frac{1}{2} + \frac{z^3}{4} + \frac{z^5}{4} \qquad \text{has distrib.:} \quad A = \begin{cases} 0, \text{ with probability } \frac{1}{2} \\ 3, \text{ with probability } \frac{1}{4} \\ 5, \text{ with probability } \frac{1}{4} \end{cases}$$

$$B(z) = e^{2(z-1)} \qquad \text{has distrib.:} \quad \text{Prob}\{B = k\} = \frac{2^k}{k!} e^{-2} \quad \text{for } k = 0, 1, 2, \ldots.$$

Poisson distribution

$$C(z) = z^2 e^{2(z-1)} \qquad \text{has distrib.:} \quad \text{Prob}\{C = k\} = \frac{2^{k-2}}{(k-2)!} e^{-2} \quad \text{for } k = 2, 3, 4, \ldots$$

Translated Poisson distribution

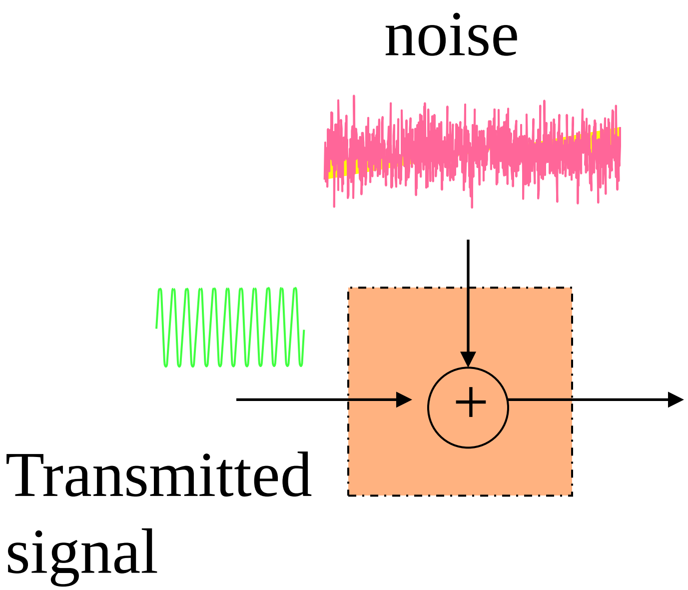
Stochastic Processes

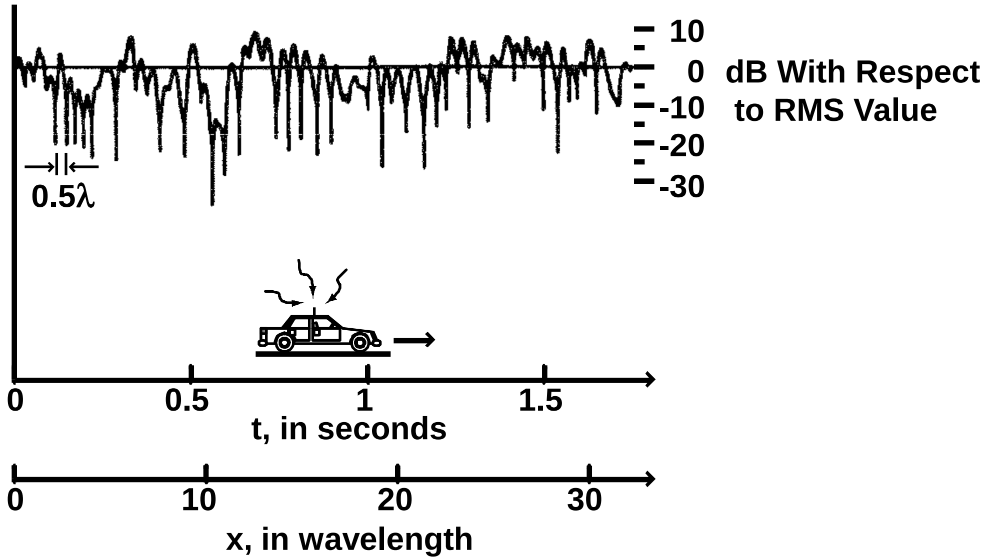
Stochastic Processes

- A stochastic process X(t) is identified by a different distribution of X at different time instants t. A stochastic process is characterized by:
  - The **state space**, that is the set of all possible values that can be assumed by X(t). Such space can be **continuous** or <u>**discrete**</u> (in such a case the stochastic process is named **chain**).
  - **Time variable**: variable t can belong to a <u>**continuous**</u> set or to a **discrete** one.
  - **Correlation characteristics** among X(t) random variables at different instants t.

Stochastic Processes (cont'd)

- There are different examples of stochastic processes in telecommunications





Stationary Processes

- Definition of **mean**: $$E[X(t)] = \int_{-\infty}^{+\infty} x \times pdf_X(x)dx = \mu_X(t)$$
- Definition of **autocorrelation**: $$R_X(t_1, t_2) = E[X(t_1)X(t_2)]$$

  where $X(t_1)$, $X(t_2)$ are random variables obtained from the process $X(t)$ at times t_1 and t_2

- The **strict-sense stationary process** entails that its joint distribution on a set of time instants does not vary for their translation.
- A random process is said to be **wide-sense stationary**, if its mean is constant and its autocorrelation only depends on the distance from instants (does not vary with a shift in the time origin):

$$E[X(t)] = \mu_X = \text{constant} \qquad R_X(t_1, t_2) = R_X(t_1 - t_2) = R_X(\tau)$$

- In this course we will only consider strict-sense stationary processes.

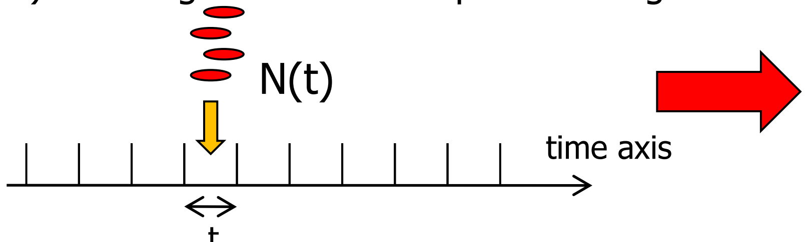
Arrival Processes

- Typical stochastic processes are related to the arrival of traffic in the networks:
  - **Number of calls** (or packets or sessions) arrived in a given time interval;
  - **Interarrival time** between two consecutive arrivals of calls (or packets or sessions).

Arrival Process / Point Process Characterization

- Let us refer for the moment to continuous-time processes. An **arrival process** is a stochastic process, where transitions are only possible between adjacent increasing states.

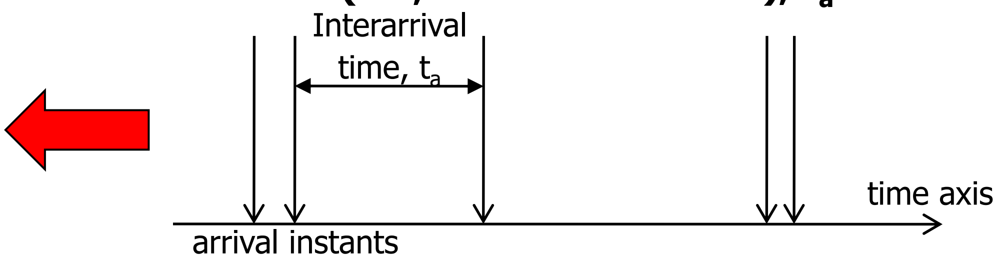
- An arrival process can be seen as a **point process** on the positive real axis, i.e., arrival of points on R^+. An arrival process can be characterized in two different ways:
  - **Number of arrivals in a generic interval t**: We can group the arrivals (i.e., points on the positive real axis) counting the number of points falling on intervals of given size t, N(t).



Prob{N(t)=k} distribution

  - **Distribution of times between arrival (i.e., interarrival times), t_a**

$f_{ta}(t)$ probability density function



Poisson Arrival Process

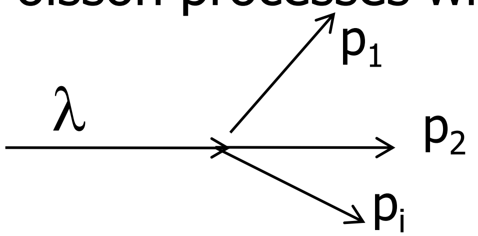
- A Poisson process is characterized by a number of arrivals in a given interval of duration t, N(t), according to the following **Poisson distribution** with mean arrival rate λ:

$$\mathrm{Prob}\{N(t)=k\}=\frac{(\lambda t)^k}{k!}e^{-\lambda t},\ \text{for any interval of duration } t$$

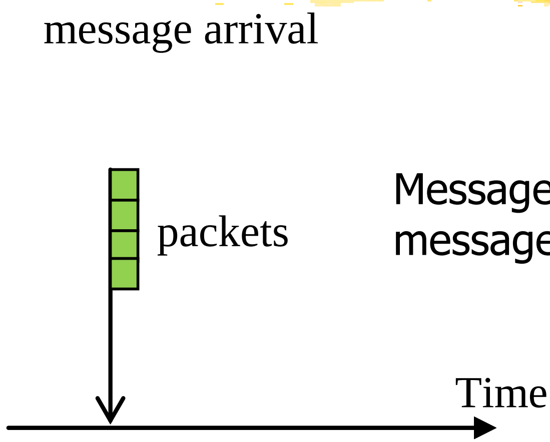
- We have a Poisson arrival process with mean rate λ if and only if the **interarrival times are exponentially distributed** with mean rate λ (i.e., mean value $1/\lambda$): $f_a(t)=\lambda e^{-\lambda t},\quad t\geq 0$
- **Poisson arrivals in disjoint intervals are independent**.
- Each Poisson arrival carries (for instance) a voice call or a packet. **The Poisson process is here used as a traffic generator**.
- A Poisson arrival process is **compound** if every Poisson arrival implies the instantaneous generation of a **group of arrivals**.
  - For instance: The arrival of IP packets segmented in a number of layer 2 packets (frames).

Poisson Arrival Process: Properties

- **Sum property**: The sum of independent Poisson processes i = 1, 2, …, n with mean rates λ_i for i = 1, 2, …, n is still a Poisson process with mean rate $\Sigma_i \lambda_i$.
  - Let us consider the case n = 2. Then on a given interval t we have to sum the number of Poisson arrivals $N_1(t)$ and $N_2(t)$ with respective PGFs as $N_1(z) = e^{\lambda_1 t(z-1)}$ and $N_2(z) = e^{\lambda_2 t(z-1)}$. The total arrival process is $N(t) = N_1(t) + N_2(t)$. Since $N_1(t)$ and $N_2(t)$ are independent processes, the PGF of N(t), N(z), is obtained as the product of the PGF $N_1(t)$, $N_1(z)$, and that of $N_2(t)$, $N_2(z)$: $N(z) = N_1(z) \times N_2(z) = e^{(\lambda_1+\lambda_2)\, t(z-1)}$. Hence, we can deduce that N(z) is related to a Poisson arrival process with mean rate $\lambda_1 + \lambda_2$.
- **Random splitting property**: The probabilistic division of a Poisson process with mean rate λ in sub-processes with related probabilities p_i for i = 1, 2, …, n generates Poisson processes with mean rates $\lambda\, p_i$, respectively.
- These two properties are used to study the traffic in the networks.

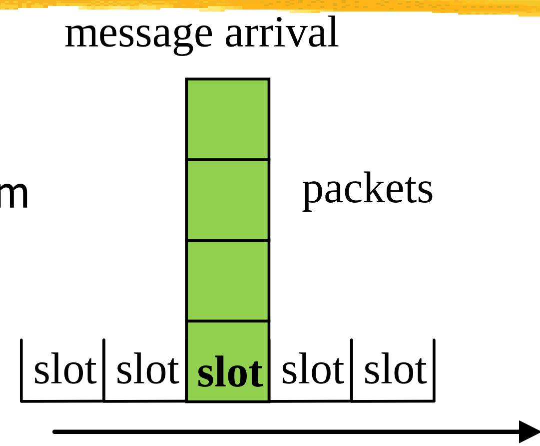


Compound Arrival Processes: Continuous and Discrete Time



Continuous-time compound arrival process

Messages have iid lengths from message to message



Discrete-time compound arrival process

- In the **continuous-time case**, all the packets of a message arrive together at the same instant; this is well suited to model the arrival of packets at a queue in a host (operating system).
- In the **discrete-time case**, the packets of a message arrive in the same slot; this is well suited to model the transmission messages to a remote node in a store-and-forward network.
- In both cases we have a **packet-based traffic model**.

Traffic Engineering and Definition of Traffic Intensity

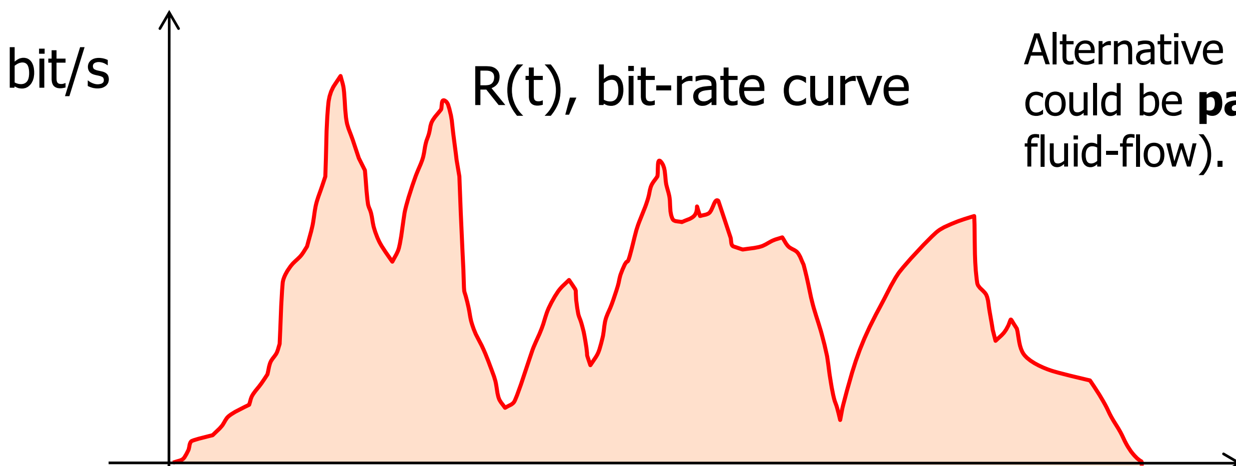
Traffic Engineering: a Definition

- Traffic engineering encompasses the application of scientific principles and technology to the **measurement, modelling, characterisation, and control** of multi-media multi-class traffic and the application of such knowledge and techniques to achieve specific performance objectives, including the planning of network capacity under QoS guarantee, and the efficient, reliable transfer of information.
  - The major objective of traffic engineering is **to improve network performance while maintaining the QoS requirements** through the optimisation of network resources.
  - The need **to allocate and balance resources among different traffic classes** to achieve the best use of network resources is a crucial traffic engineering problem.

Memorandum of Understanding, COST Action 290 “Traffic and QOS Management in Wireless Multimedia Networks: WI-QOST”, 2004.

Traffic Generated by Sources

- **Traffic is generated by a source, that is an application running on a host**.
- The traffic generated by a source can be seen as a bit-rate as a function of time according to a **stochastic process R(t)**. This is a **fluid-flow model**. We can therefore determine the mean, the variance, and, in general, the distribution of the bit-rate.



Alternative traffic models could be **packet-based** (not fluid-flow).

Time, t

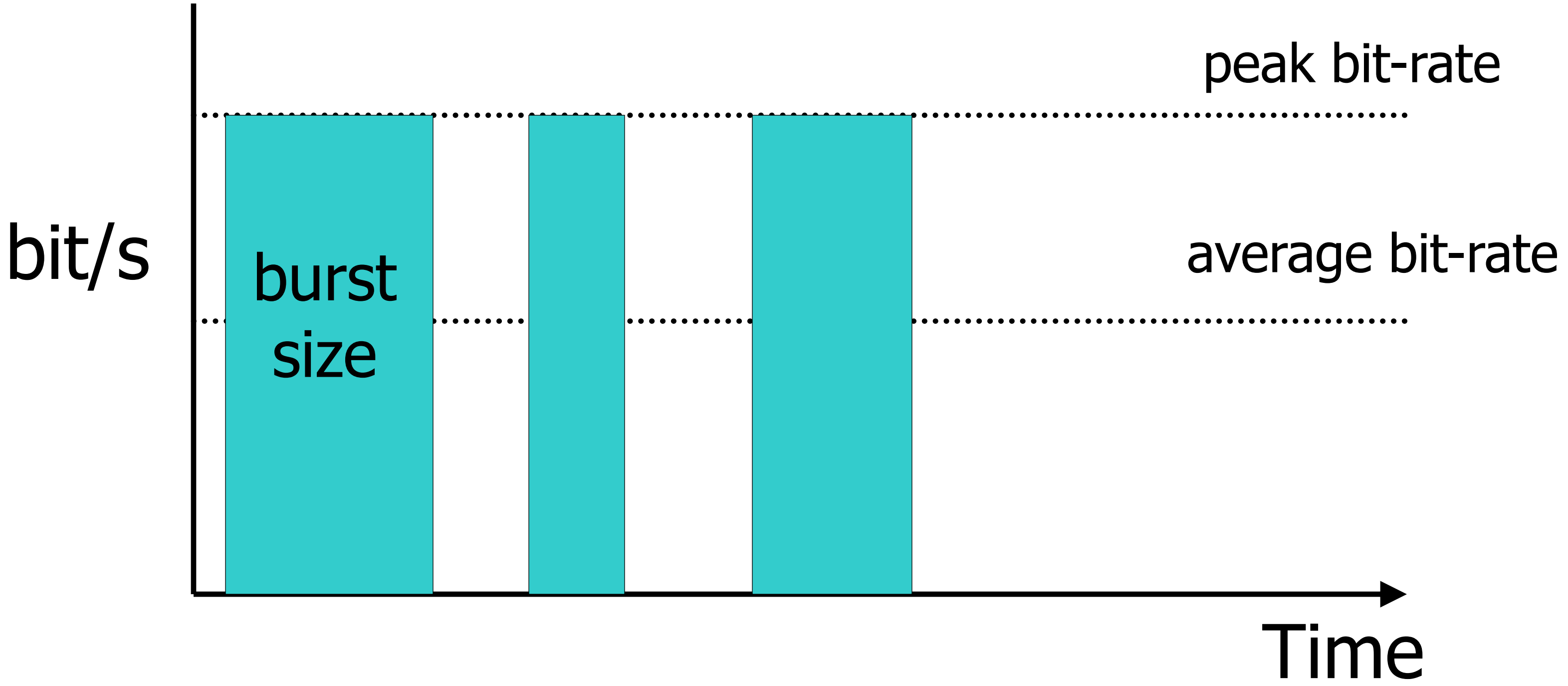
$$\text{Arrival curve}\left(\text{t}\right)=\int_0^t R(\tau)\,d\tau$$

Traffic Characterization: Parameters

- **Different flows (generated by different applications) have distinct traffic patterns.**
- A given traffic pattern can be described using several traffic parameters (the only average rate is not enough):
  - **Peak rate**: maximum rate in any time interval
  - **Average rate**: long-term average
  - **Burst size**: duration of traffic peaks.

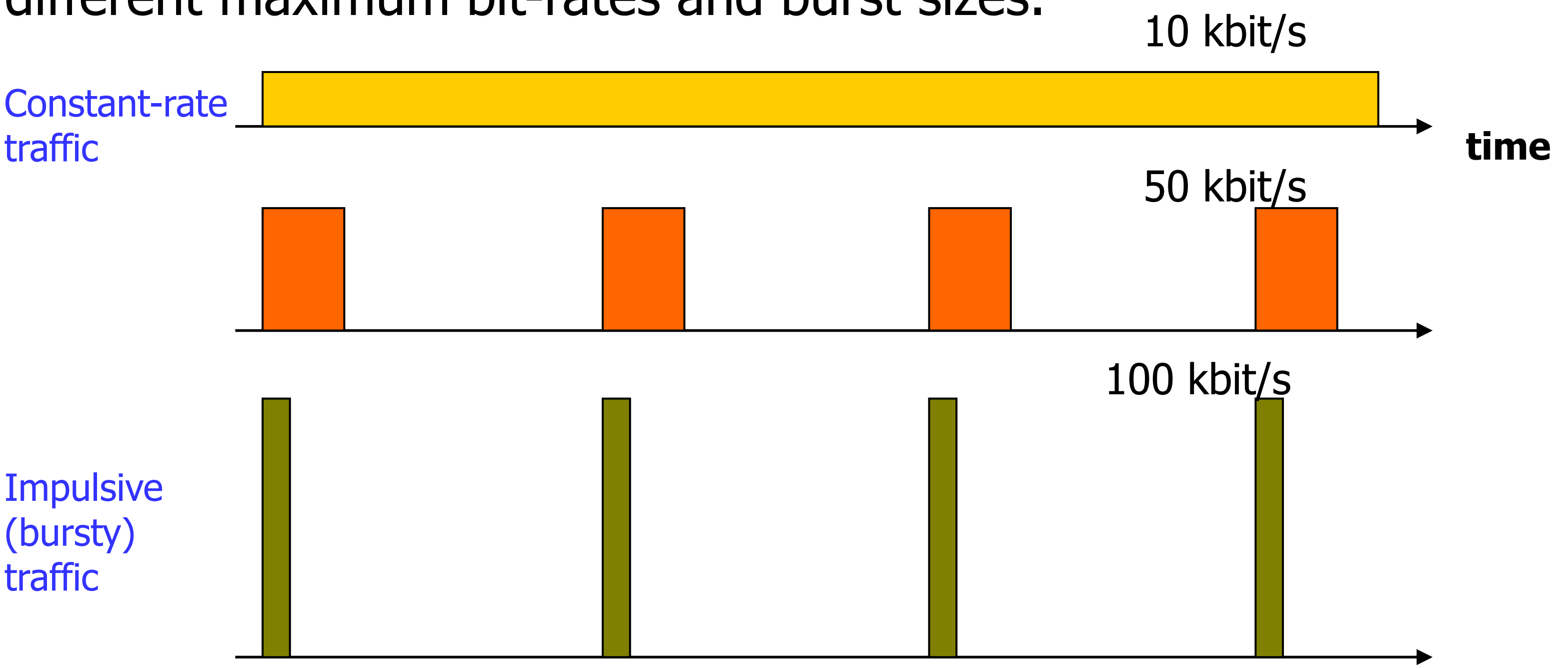
Traffic Parameters, Illustrated

- An example of the traffic bit-rate graph (fluid-flow) as a function of time:



Traffic Patterns, an Example

All traffic flows below have the same average bit-rate (10 kbit/s), but different maximum bit-rates and burst sizes.



The **burstiness** of a traffic flow β is defined as **the ratio of the peak bit-rate and the mean bit-rate.** In the previous examples, burstiness β is 1, 5 and 10.

Traffic Intensity: Erlang

- **Traffic intensity** is a basic parameter in traffic engineering. It represents a fundamental **characteristic of a traffic flow when it arrives at a suitable service facility**.
  - Let λ denote the mean arrival rate of the traffic (packets or calls per second).
  - Let E[X] denote the mean service duration (e.g., transmission time) of each service request (packet or call).
  - Then, the **traffic intensity ρ is obtained as $\rho = \lambda E[X]$ and it is measured in Erlangs (even if ρ is dimensionless).**
    - **ρ denotes the percentage of time the service facility is busy in serving this traffic.**

The Danish engineer Agner Krarup Erlang was a pioneer of the queuing theory.
A. K. Erlang, "Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges", *Post Office Electrical Engineers Journal*, Vol. 10, 1917.

Traffic Intensity: Erlang

We refer here to a packet-based traffic for the definition of the traffic intensity even if the definition could be modified to refer to fluid-flow traffic models.

- **Traffic intensity** is a ba[…] engineering. It represents […] **characteristic of a traff[…] a suitable service facility**.
  - Let λ denote the mean arrival rate of the traffic (packets or calls per second).
  - Let E[X] denote the mean service duration (e.g., transmission time) of each service request (packet or call).
  - Then, the **traffic intensity ρ is obtained as $\rho = \lambda E[X]$ and it is measured in Erlangs (even if ρ is dimensionless).**
    - **ρ denotes the percentage of time the service facility is busy in serving this traffic.**

The Danish engineer Agner Krarup Erlang was a pioneer of the queuing theory.
A. K. Erlang, "Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges", *Post Office Electrical Engineers Journal*, Vol. 10, 1917.

QoS

Types of Applications

- Classical data applications are **“elastic”** and tolerate delays and losses and can adapt to congestion.
- “Real-time” applications may be **“inelastic”**.
- **The terms “elastic” or “inelastic” have to be intended in relation to the bit-rate constraints of the application.**

Different Characteristics for Traffic Flows

- **Bursty traffic (elastic traffic)** creates difficulties for the network since it entails a low utilization of network resources for long times, but suddenly causes congestion in network buffers. This type of traffic (data traffic, **non-real-time**) is sensitive to packet losses. TCP-based traffic (e.g., HTTP, FTP) can be bursty.
  - $\beta >> 1$

- **Constant bit-rate traffic (inelastic traffic)** is typical of **real-time**, time-critical applications and needs high priority to be managed with low delays in network buffers. This type of traffic is less sensitive to packet losses depending on the robustness of the application codec. Voice/audio (MP3) and video (H.264) can be represented by constant bit-rate traffic.
  - $\beta = 1$

What is Quality of Service?

- In the field of telephony, quality of service was defined by ITU-T in **Recommendation E.800** (1994 and subsequent revisions).
  - E. 800 defines QoS as “collective effect of service performance which determines the degree of satisfaction of a user of the service”.
- **QoS has today a very broad scope from PHY layer issues to application level ones.**
- QoS entails the ability to provide **different priority** levels to different applications, users, or data flows, or to guarantee a **certain level of performance to a data flow** (e.g., a required throughput, mean delay, etc.).

ITU-T, "E.800: Definitions of terms related to quality of service", last revision on September 2008 (http://www.itu.int/rec/T-REC-E.800-200809-I/en).

Some QoS Metrics

- Main QoS metrics are:
  - Mean delay [s] to cross a node or to cross a whole network
  - Packet loss rate [%] at IP or MAC layers
  - Blocking probability [%] at PHY or MAC layer
  - Jitter (delay variation)

With streaming (video and audio) traffic, a de-jittering buffer is needed on the receiver side to compensate for delay variations due to jitter.

Further details on QoS parameters and approaches are provided in Lesson No. 14 when dealing with QoS support in the Internet (IP networks).

Traffic Source Types, Requirements, Models

	Characteristics	QoS Requirements	Traffic Model
Voice	* Alternating talk-spurts and silence intervals. * Talk-spurts produce constant packet-rate traffic	Delay < ~150 ms Jitter < ~30 ms Packet loss < ~1%	* Two-state (on-off) Markov Modulated Rate Process (MMRP) * Exponentially distributed time in each state
Video	* Highly bursty traffic (when encoded) * Long range dependencies	Delay < ~ 400 ms Jitter < ~ 30 ms Packet loss < ~1%	K-state (on-off) Markov Modulated Rate Process (MMRP)
Interactive **FTP, Telnet, Web (HTTP)**	* Poisson type * Sometimes batch-arrivals, or bursty, or sometimes on-off	Zero or near-zero packet loss Delay may be important	Poisson, Poisson with batch arrivals, Two-state MMRP

For more details, see Lesson No. 17.

QoS-Enforcing Approaches

- QoS can be achieved by:
  - **Resource reservation** (e.g., integrated services in IP networks, as shown in Lesson No. 14)
  - **Prioritization** (e.g., differentiated services in IP networks, as shown in Lesson No. 14)

- QoS can be applied:
  - **Per flow**: individual, unidirectional streams
  - **Per aggregate**: two or more flows belonging to the same traffic class have common QoS management and share resources.

Differentiated QoS Levels for the Applications

- In ITU-T Recommendation G.1010, **applications have been classified in 8 groups according to error tolerance and delay requirements**.

Error sensitivity

Error tolerant	Conversational voice and video	Voice/video messaging	Streaming audio and video	Fax
Error intolerant	Command/control (eg Telnet, interactive games)	Transactions (eg E-commerce, WWW browsing, Email access)	Messaging, Downloads (eg FTP, still image)	Background (eg Email arrival)
	Interactive (delay <<1 sec)	Responsive (delay ~2 sec)	Timely (delay ~10 sec)	Non-critical (delay >>10 sec)

Time criticality

ITU-T Y.1541

- As far as the traffic classification is concerned, we may refer to the categorization in ITU-T Y.1541 ("Network Performance Objectives for IP-Based Services"), which defines **8 QoS traffic classes at IP layer** (even if in what follows we consider only 6 QoS traffic classes).
- With Y.1541, traffic classes refer to
  - Application layer characteristics,
  - Connectivity requirements (queuing mechanisms at nodes and routing types),
  - **Mean delay, loss percentage, and delay jitter (delay variation) tolerance**.

ITU-T Y.1541 Traffic Classes

IP layer classification

Priority, urgency

QoS class	Applications	Node mechanism	Network techniques
0	Real-time, jitter sensitive, high interaction (VoIP)	Separate queue with preferential servicing, traffic grooming	Constrained routing and distance
1	Real-time, jitter sensitive, interaction (VoIP)		Less constrained routing and distances
2	Data transfer, high interaction	Separate queue, drop priority	Constrained routing and distance
3	Data transfer, interaction		Less constrained routing and distances
4	Error non sensitive (bulk data, video streaming)	Long queue, drop priority	Any route/path
5	Traditional IP-based applications	Separate queue (lowest priority)	Any route/path

ITU-T Y.1541 Requirements

Network parameter	Condition	QoS Classes					
		Class 0	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**
IPTD (IP Packet Transfer Delay)	Less than	100 ms	400 ms	100 ms	400 ms	1 s	Unspecified
IPDV (IP Packet Delay Variation)	Less than	50 ms	50 ms	Unspecified	Unspecified	Unspecified	Unspecified
IPLR (IP Packet Loss Ratio)	Less than	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	Unspecified
IPER (IP Packet Error Ratio)	Less than	1×10^{-4}					Unspecified

GEO satellite networks cannot guarantee the requested QoS levels to Class 0 and 1 services due to the high latency.

QoS Support Techniques

- **QoS support requires the adoption of coherent solutions at the different layers of the OSI protocol stack**.
  - **PHY**: Selection of appropriate modulation and coding level to guarantee a certain Bit Error-Rate (BER) at the receiver.
  - **MAC**: Call Admission Control (CAC), traffic-class-based queuing, traffic shaping/policing, scheduling, prioritization.
  - **Network**: DiffServ (or IntServ), IP traffic routing, Explicit Congestion Notification (ECN), IP buffer management techniques (e.g., RED).
  - **Transport**: Network layer buffer size selection, TCP acceleration techniques (e.g., use of proxies).
  - **Application**: Codec selection.

QoS versus QoE

- QoS is ensuring that network elements apply consistent treatments to traffic flows as they traverse the network.
- **Quality of Experience (QoE) is subjective and relates to the QoS actually perceived by a user**. This applies to voice, multimedia, and data.
  - **ITU-T Recommendation P.10/G.100**, defines QoE as "the overall acceptability of an application or service, as perceived subjectively by the end-user".
  - QoE includes complete end-to-end system effects (client, terminal, network, and service infrastructure).
  - Overall acceptability may be influenced by user expectations and context.

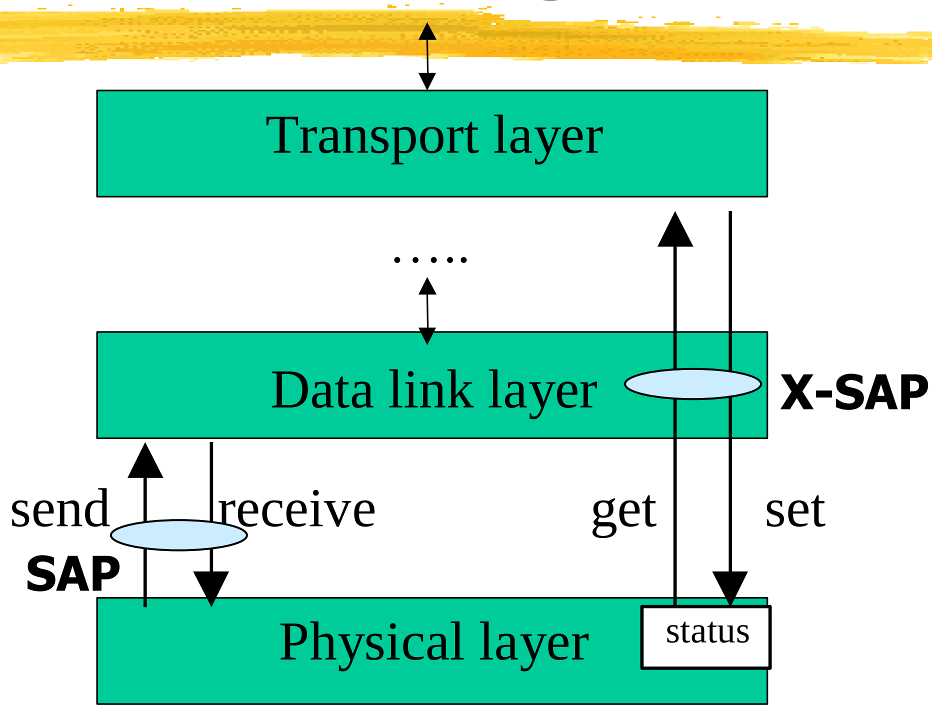
QoS versus Efficiency

- **System efficiency** and **QoS support** are essential requirements, but they can represent **conflicting needs**.
  - System efficiency is an important requirement for network operators to provide services at competitive costs.
  - QoS support is mandatory for end users who do not care about resource utilization, but expect a good service level.

QoS vs. Efficiency: Cross-Layer Air Interface Design

- Cross-layer air interface design is a novel approach that modifies the classical ISO/OSI protocol stack to achieve an efficient use of resources with QoS support.

- **Signaling and protocol coordination is achieved also between non-adjacent layers through new X-SAPs.**



G. Giambene, S. Kota, "Cross-layer Protocol Optimization for Satellite Communications Networks: A Survey", *International Journal of Satellite Communications and Networking*, Vol. 24, pp. 323-341, September-October 2006.

G. Giambene (Editor). *Resource Management in Satellite Networks: Optimization and Cross-Layer Design*. Springer, 2007, ISBN 978-0-387-36897-9, New York, NY

ETSI TR 102 676 ("Satellite Earth Stations and Systems (SES); Broadband Satellite Multimedia (BSM); Performance Enhancing Proxies (PEPs)", 2009.

ITU-R "Cross-layer QoS Provisioning in IP-based Hybrid Satellite-Terrestrial Networks", Document 4B/196, Question ITU-R 287/4, 22 September 2011.

Thank you!

giovanni.giambene@gmail.com