4. Traffic modelling and measurements

Contents

- Traditional modelling of telephone traffic
- Traffic variations
- Traffic measurements
- Traditional modelling of data traffic
- Novel models for data traffic

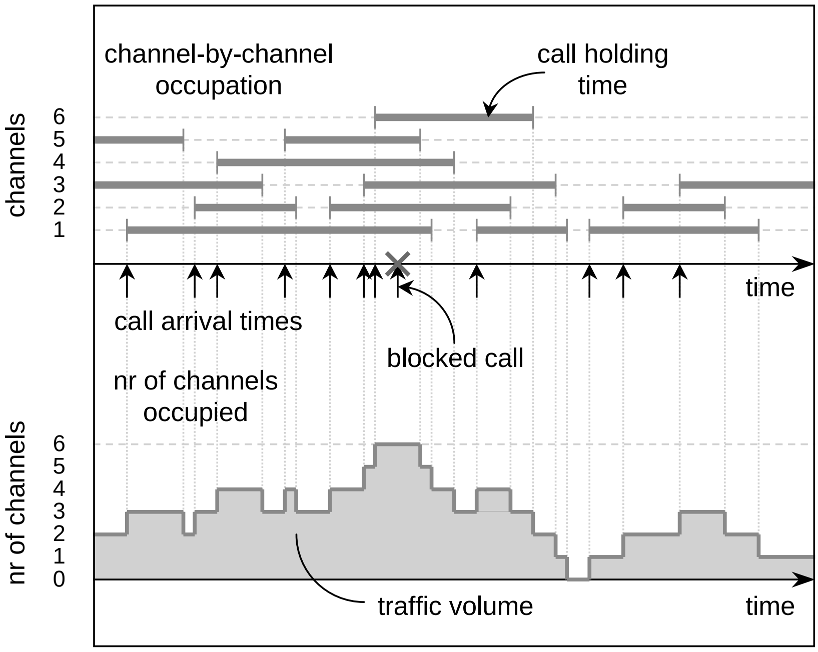
Modelling of telephone traffic

- In telephone networks:

Traffic ↔ Calls

- Traffic model (for a single link) should specify
  - the type of the call arrival process
  - the distribution of call holding times
- These together specify
  - the **traffic process** that tells the number of ongoing calls
    = number of occupied channels
    = instantaneous intensity of the traffic carried (in erlangs)
- Note:
  - **Traffic volume** refers to
    the amount of carried traffic during some time interval
    = integral of the instantaneous traffic intensity over this interval

Traffic process

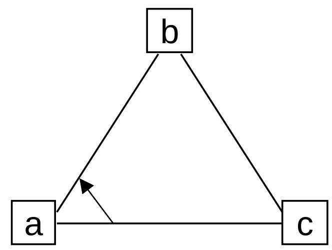


Call arrival process (1)

- **Aggregated traffic** in trunk network
  - Traditional model: **Poisson process** (with some intensity $\lambda > 0$)
    - In a short time interval of length Δ, there are two possibilities: either a new call arrives (with probability $\lambda\Delta$) or nothing happens (with probability $1 - \lambda\Delta$)
    - Disjoint intervals are independent of each other
    - As a result: call interarrival times are independently and exponentially distributed with mean $1/\lambda$
  - This is found to be a good model when user population is large ("infinite") and users make independent decisions (which is the case for the links in the trunk network)
  - Corresponding teletraffic models are loss models:
    - **Erlang model** (finite link capacity)
    - **Poisson model** (infinite link capapcity)

Call arrival process (2)

- **Overflow traffic** in trunk network
  - due to alternative routing
  - Model: **Interrupted Poisson process** (IPP)
  - In addition to the traffic process itself, there is a **modulating process** that tells whether the arrivals of an ordinary Poisson process will be realized or not
  - In the overflow model, the modulating process is the traffic process of the original (direct) link (how?)
  - Traffic stream consists of the calls blocked in the direct link



direct route: a - c
alternative route: a - b - c

Call arrival process (3)

- Traffic generated by an **individual user** (subscriber)
  - Traditional model: exponential on-off process
    - The user alternates between on and off states
      - When on, a call is going on
      - When off, the user is “idle”
    - The times spent in different states are assumed to be independent and exponentially distributed (with state-dependent mean)
- Traffic generated by a **superposition of users** in access network
  - Finite number of individual users
    - modelled separately as above
    - making independent decision
  - Corresponding teletraffic models are loss models:
    - Engset model (insufficient link capacity)
    - Binomial model (sufficient link capacity)

Call holding time (1)

- Basic assumption:
  - call holding times are **independent** and **identically distributed**
- Distribution of call holding times
  - Traditional model: **exponential** distribution
    - one parameter $\Rightarrow$ simple!
    - **memoryless property**: given that the holding time is at least (any) t, the probability that the call will end in a short time interval $(t, t+\Delta)$ depends just on Δ (but not on t)
    - exponential tail
  - More complicated models:
    - normal distribution (two parameters: mean and variance)
    - log-normal distribution (two parameters)
    - hyper-exponential (with two parameters)
    - Weibull distribution (with two/three parameters)

Call holding time (2)

- Call holding time distribution is typically different for
  - business and residential calls
  - daytime and evening calls
  - ordinary and “data” calls (fax, Internet access, etc.)

Contents

- Traditional modelling of telephone traffic
- Traffic variations
- Traffic measurements
- Traditional modelling of data traffic
- Novel models for data traffic

Traffic variations in different time scales (1)

- **Predictive variations**
  - **Trend** (years)
    - traffic growth: due to
      - existing services (new users, new ways to use, new tariffs)
      - new services
  - **Regular year profile** (months)
  - **Regular week profile** (days)
  - **Regular day profile** (hours)
    - including "busy hour"
  - Variations caused by predictive (regular and irregular) **external events**
    - regular: e.g. Christmas day
    - irregular: e.g. World Championships, televoting
  - Note: different profiles for different types of user groups
    - e.g. business vs. residential users

Traffic variations in different time scales (2)

- **Non-predictive variations**
  - **Short term stochastic variations** (seconds - minutes)
    - random call arrivals
    - random call holding times
  - **Long term stochastic variations** (hours - ...)
    - random deviations around the profiles
    - each day, week, month, etc. is different
  - Variations caused by non-predictive **external events**
    - e.g. earthquakes, hurricanes

Busy hour (1)

- For dimensioning,
  - an estimate of the traffic load is needed
- In telephone networks,
  - standard way is to use so called **busy hour** traffic for dimensioning

Busy hour ≈ the continuous 1-hour period for which the traffic volume is greatest

  - This is unambiguous only for a single day (let’s call it **daily peak hour**)
  - For dimensioning, however,
    - we have to look at not only a single day but many more (why?)
  - At least three different definitions for busy hour (covering several days) have been proposed:
    - Average Daily Peak Hour (ADPH)
    - Time Consistent Busy Hour (TCBH)
    - Fixed Daily Measurement Hour (FDMH)

Busy hour (2)

- Let
  - N = number of days during which measurements are done (e.g. N = 10)
  - $V_n(\Delta)$ = measured traffic volume during 1-hour interval Δ of day n
  - $\max_\Delta V_n(\Delta)$ = daily peak hour traffic volume of day n
- Average Daily Peak Hour (ADPH) traffic volume:

$$V_{\text{ADPH}} = \frac{1}{N}\sum_{n=1}^{N} \max_\Delta V_n(\Delta)$$

- Time Consistent Busy Hour (TCBH) traffic volume:

$$V_{\text{TCBH}} = \max_\Delta \frac{1}{N}\sum_{n=1}^{N} V_n(\Delta)$$

- Fixed Daily Measurement Hour (FDMH) traffic volume:

$$V_{\text{FDMH}} = \frac{1}{N}\sum_{n=1}^{N} V_n(\Delta_{\text{fixed}})$$

Busy hour (3)

- Average Daily Peak Hour (ADPH) traffic:

$$a_{\mathrm{ADPH}} = \frac{1}{\Delta} \cdot V_{\mathrm{ADPH}}$$

- Time Consistent Busy Hour (TCBH) traffic:

$$a_{\mathrm{TCBH}} = \frac{1}{\Delta} \cdot V_{\mathrm{TCBH}}$$

- Fixed Daily Measurement Hour (FDMH) traffic:

$$a_{\mathrm{FDMH}} = \frac{1}{\Delta} \cdot V_{\mathrm{FDMH}}$$

- It can be shown (how?) that

$$a_{\mathrm{FDMH}} \le a_{\mathrm{TCBH}} \le a_{\mathrm{ADPH}}$$

Contents

- Traditional modelling of telephone traffic
- Traffic variations
- Traffic measurements
- Traditional modelling of data traffic
- Novel models for data traffic

Traffic measurements (1)

- Traffic measurements are needed for
  - network design and traffic management
    - a basis for dimensioning
    - traffic modelling
    - traffic predictions
    - traffic control (e.g. connection admission control, dynamic routing)
    - congestion control (e.g. congestion detection)
  - but also for
    - getting accounting information
- More and more information about traffic is needed because of
  - new users, new ways to use, new tariffs (as for existing services and networks)
  - new services and networks
  - increasingly tough competition

Traffic measurements (2)

- Traffic measurements in telephone networks
  - traffic on different links
    - traffic process (carried traffic intensity = number of occupied channels)
    - call arrival process (interarrival times)
    - call holding times
  - traffic on different trunk network nodes
    - distribution of incoming traffic from different directions
    - distribution of outgoing traffic in different directions
  - traffic on different access network nodes
    - distribution according to the type of traffic source
      - e.g. residential vs. business subscribers
    - use of different services

Traffic measurements (3)

- Traffic measurements in Internet/LAN
  - traffic on different links
    - traffic process (carried traffic intensity in bits per second)
  - traffic on different network nodes
  - traffic at different protocol levels
    - packet level (IP)
      - packet arrival process (interarrival times)
      - packet lengths
    - connection level (TCP)
      - connection arrival process
      - connection holding times
        (per applications: ftp / http / email / telnet etc.)
      - total amount of information transferred
        (per applications: ftp / http / email / telnet etc.)

Analysis of traffic measurements

- Traditional statistical methods:
  - parameter estimation
    - traffic intensity
    - traffic variability (short term variance, coefficient of variation)
    - traffic peakedness
  - estimation of probability density function
  - auto-correlation
- New approach:
  - scalability analysis
    - self-similarity
    - multifractal characterization

Estimation of the traffic intensity based on measurements

- Consider the traffic process (in a link of a telephone network)
  - Traffic is measured during some interval $[0,T]$ (e.g. busy hour)
  - Let $V(T)$ denote the traffic volume during this interval (random variable!)
- Purpose is to estimate the (carried) traffic intensity
  - assuming that it is constant
  - based on these measurements
- A natural **estimate** for a is

$$\hat{a} = \frac{V(T)}{T}$$

- It is **unbiased**, that is: its expectation is a,

$$E[\hat{a}] = \frac{E[V(T)]}{T} = a$$

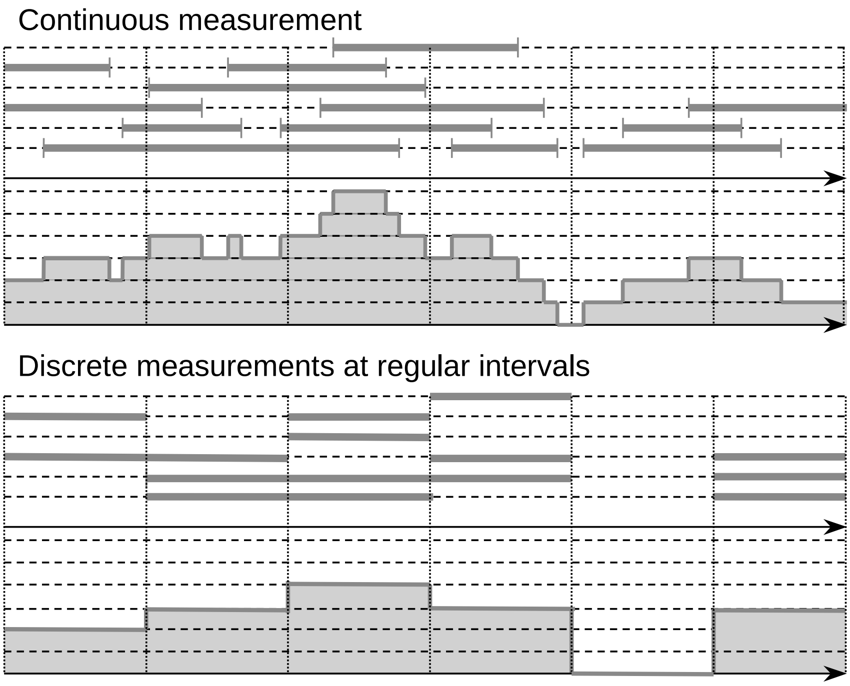
Different measurement modes (1)

- **Continuous measurement**
  - Register
    - the number of occupied channels at time 0
    - the starting times of all connections during interval $[0,T]$
    - the stopping times of all connections during interval $[0,T]$
  - Thus, it is possible to reconstruct the actual traffic process
    - giving an **exact** value for the traffic volume $V(T)$
- **Discrete measurements** at regular intervals (of length Δ)
  - Register
    - the number of occupied channels $X(t)$ at times $t = 0, \Delta, 2\Delta, ..., T\text{-}\Delta$
  - Traffic volume during interval $[0,T]$ is **estimated** by

$$\hat{V}_\Delta(T) = \sum_{n=0}^{(T/\Delta)-1} X(n\Delta) \cdot \Delta$$

  - Note: the estimate approaches to $V(T)$ as $\Delta \downarrow 0$

Different measurement modes (2)



On the accuracy of the estimate (1)

- Continuous measurement
  - Estimate $\hat{a}$ itself is a random variable with relative error

$$\frac{D[\hat{a}]}{E[\hat{a}]} = \frac{D[V(T)]}{E[V(T)]} = \frac{D[V(T)]}{aT}$$

  - Assume that
    - calls arrive according to a Poisson process
    - call holding times are exponentially distributed with mean $h = 1$
    - link capacity is infinite
  - Then the relative error is approximately

$$\frac{1}{\sqrt{a}} \quad \text{(when } T \text{ is very small)}$$

$$\frac{\sqrt{2}}{\sqrt{a}\sqrt{T}} \quad \text{(when } T \text{ is large enough)}$$

On the accuracy of the estimate (2)

- Discrete measurements at regular intervals of length Δ
  - Estimate $\hat{a}_\Delta$ itself is again a random variable with relative error

$$\frac{D[\hat{a}_\Delta]}{E[\hat{a}_\Delta]} = \frac{D[\hat{V}_\Delta(T)]}{E[\hat{V}_\Delta(T)]} = \frac{D[\hat{V}_\Delta(T)]}{aT}$$

  - Note that in this case
    - in addition to the random deviation between a and $\hat{a} = V(T)/T$, estimate $\hat{a}_\Delta$ includes the measurement error (deviation between $V(T)$ and its estimate)
  - Under the same assumptions as above, the relative error is approximately

$$\frac{\sqrt{\Delta}}{\sqrt{a}\sqrt{T}} \cdot \frac{\sqrt{1+\exp(-\Delta)}}{\sqrt{1-\exp(-\Delta)}}$$ (when T is large enough)

Example

- Accuracy requirement:
  - an estimate of a with max. relative error p = 5%
- Assume:
  - traffic intensity a = 100 erlangs
- Continuous measurement:
  - measurement interval T should be at least

$$T \geq \frac{2}{a \cdot p^2} = \frac{2}{100} \cdot \left(\frac{100}{5}\right)^2 = 8.0 \text{ (mean holding times)}$$

- Discrete measurements at regular intervals of length Δ = 1 (hold. time):
  - measurement interval T should be at least

$$T \geq \frac{\Delta}{a} \cdot \frac{1+\exp(-\Delta)}{1-\exp(-\Delta)} \cdot \frac{1}{p^2} \cong \frac{2.164}{100} \cdot \left(\frac{100}{5}\right)^2 \cong 8.7 \text{ (mean holding times)}$$

Contents

Traditional modelling of data traffic

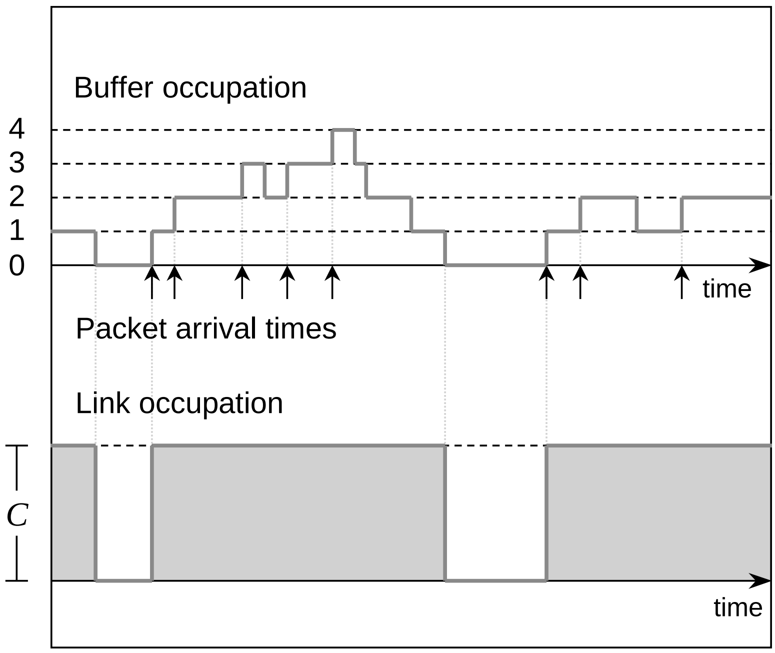
- Connection level
  - new connections arrive according to a Poisson process
    $\Rightarrow$ connection interarrival times independent and exponentially distributed
  - connection holding times are independent and exponentially distributed
  - infinite system model (since no connection admission control)
- Packet level
  - new packets arrive according to a Poisson process
    $\Rightarrow$ packet interarrival times independent and exponentially distributed
  - packet lengths are independent and exponentially distributed
    $\Rightarrow$ packet transmission times (in links) independent and exponentially distributed
  - queueing model

Traffic process at the packet level (1)

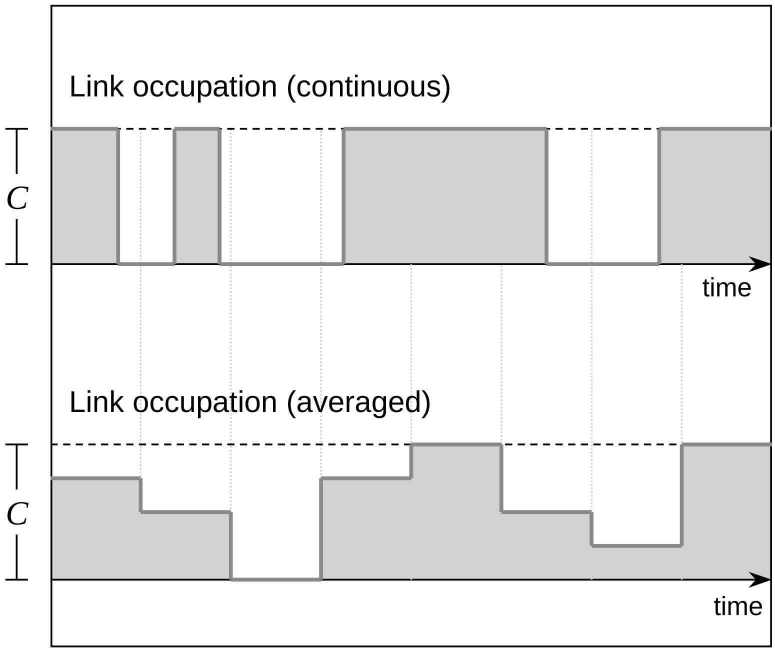
- Consider the traffic process at the packet level
- In continuous time,
  - there are just two possibilities: a link is either
    - **busy** (with the whole link capacity C in use) or
    - **idle**

    depending on
    whether there are packets to be transmitted in the buffer or not
  - thus, link occupancy can take just two different values: 0 or C
  - note: when a packet is being transmitted, it takes the whole link capacity
- However, by averaging this process (over time intervals),
  - link occupancy can have any value between 0 or C

Traffic process at the packet level (2)

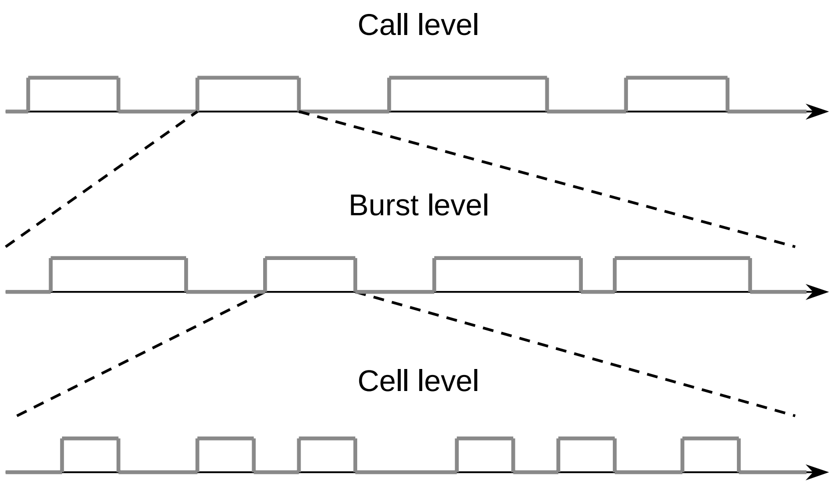


Traffic process at the packet level (3)



Modelling of ATM traffic (1)

- Three different time scales:



Modelling of ATM traffic (2)

- Call level
  - "traffic unit" = connection
  - loss model (for CBR and VBR connections)
- Burst level
  - "traffic unit" = burst of varying length (and possibly of varying rate)
  - (traditional) fluid buffer models:
    - superposition of exponential ON-OFF sources (A-M-S model)
    - burst arrivals according to Poisson process (Kosten model)
- Cell level
  - "traffic unit" = fixed length cell
  - queueing models:
    - superposition of periodic sources (N*D/D/1)
    - cell arrivals according to Poisson process (M/D/1)
    - discrete time Markov arrival processes (MAP/D/1)

Contents

- Traditional modelling of telephone traffic
- Traffic variations
- Traffic measurements
- Traditional modelling of data traffic
- Novel models for data traffic

Bellcore measurements

- Ethernet (LAN) measurements by Leland, Willinger, … (‘89-’92)
  - high-accuracy recording of hundreds of millions Ethernet packets
    - including both the arrival time and the length
  - see: IEEE/ACM Trans. Networking, vol. 2, nr. 1, pp. 1-15, February 1994
- Conclusions:
  - Ethernet traffic seems to be **extremely varying**
    - presence of “burstiness” across an extremely wide range of time scales (from microseconds to milliseconds, seconds, minutes, hours, …)
    - bad from the performance point of view
  - Ethernet traffic is statistically **self-similar** (fractal-like)
    - it looks the same in all time scales
    - a single parameter (the Hurst parameter) describes the fractal nature
    - good from the modelling point of view (parsimony!)
  - Traditional data traffic models do not capture these properties!

Internet measurements

- Internet (WAN) measurements by Paxson and Floyd (‘93-’95)
  - both the connection and the packet level concerned
  - see: IEEE/ACM Trans. Networking vol. 3, nr. 3, pp. 226-244, June 1995
- Connection level conclusions:
  - For interactive TELNET traffic (and other user-initiated sessions),
    - connection arrivals are well-modelled by a Poisson process (with hourly fixed rates)
  - But for connections within user-initiated sessions (FTP data, HTTP) and machine-generated connections
    - connection arrivals are more **bursty** than in a Poisson process (and even correlated)
- Packet level conclusions
  - empirical distribution of TELNET packet interarrival times is
    - **heavy-tailed** (not exponential as traditionally modelled)

New models for data traffic

- Subexponential distributions (“worse than exponential tail”)
  - e.g. log-normal, Weibull and Pareto distributions
- Heavy-tailed distributions (“power-law tail”)
  - e.g. Pareto distribution (with location parameter a and shape parameter β)

$$P\{X > x\} = (a/x)^{\beta}, \quad x \geq a > 0, \ \beta > 0$$

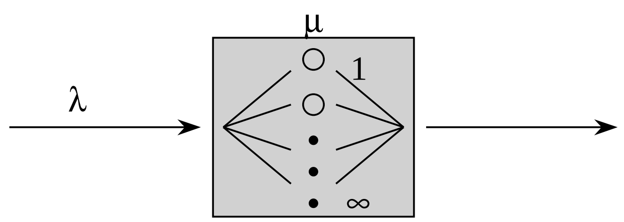
- Processes exhibiting long range dependence (LRD)
  - e.g. self-similar and asymptotically self-similar processes
- Self-similar processes
  - e.g. **fractional Brownian motion** (FBM)
    - suitable for describing aggregated traffic (in trunk network)
    - just three parameters (thus, parsimonious!)
    - one of them, so called **Hurst parameter** H, describes the grade of long range dependence (when in the interval (½,1))

Self-similarity, long range dependence and heavy tails

- If a stochastic process is **self-similar** (or asymptotically self-similar) with positive correlations,
  - then it exhibits **long range dependence** (LRD)
- Self-similarity and long range dependence are related to
  - **heavy tailed distributions**
    - tail of the distribution decreases as a power function (which is much slower than exponentially)
- In teletraffic models, this refers e.g. to distributions of
  - packet lengths and packet interarrival times,
  - connection holding times and connection interarrival times

Example on heavy tails, self-similarity and long range dependence

- Consider an infinite system (M/G/∞)
  - new customers arrive according to a Poisson process
  - service times independent and identically distributed
  - service time distribution heavy-tailed with an infinite variance
    - e.g. Pareto distribution with shape parameter $\beta < 2$
- Then the traffic process (number of customers in the system) is
  - asymptotically self-similar (and, thus, long range dependent)



THE END