Design of IP networks with Quality of Service

Design of IP networks with QoS

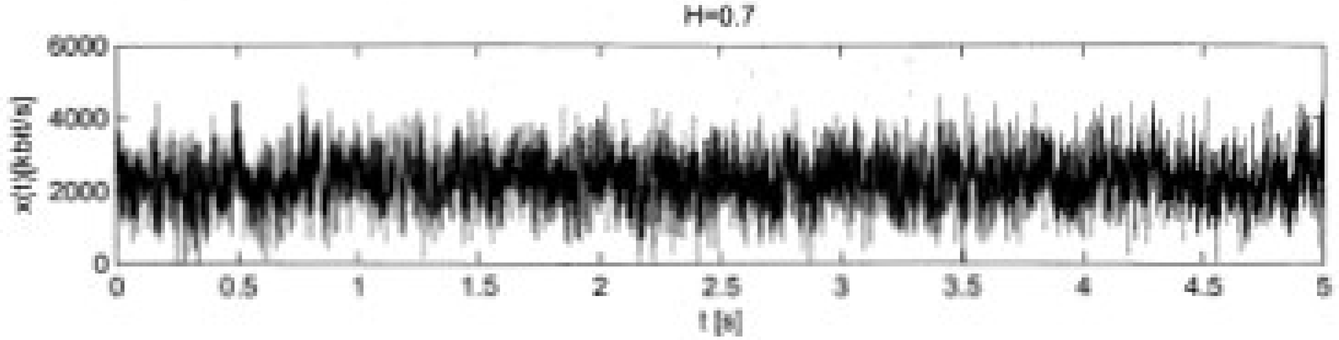
- The design of IP networks with QoS is difficult, especially with statistical service level agreements
- A significant problem is traffic modeling: usually, in classical queueing theory abstract traffic models are used, for example based on Markov chains
- These models can provide a tool for dimensioning, but with significant drawbacks
  - Real traffic sources are difficult to model with markov chains: real traffic is complex
  - Moreover, Markov chain analysis is complex, because the number of states that it is necessary to analyze grows exponentially as the number of sources increases
  - Computer analysis is needed, and models and algorithms are complex

Design of IP networks with QoS

- Network calculus is a recent discipline which is able to simplify calculations and can account for more realistic traffic models
- However, also network calculus has some drawbacks
- The classical deterministic network calculus seeks for guaranteed delay bounds, and this leads to a very conservative design in case of statistical service level agreements
- Statistical network calculus can be a remedy to this problem, but the statistical delay bounds offered by the classical statistical network calculus are still quite conservative
- The bounded variance network calculus is a new discipline which aims at obtaining tight statistical approximations of network delay
- Therefore, we will develop this flawor of network calculus and we will show through examples how it is possible to design an IP network with QoS

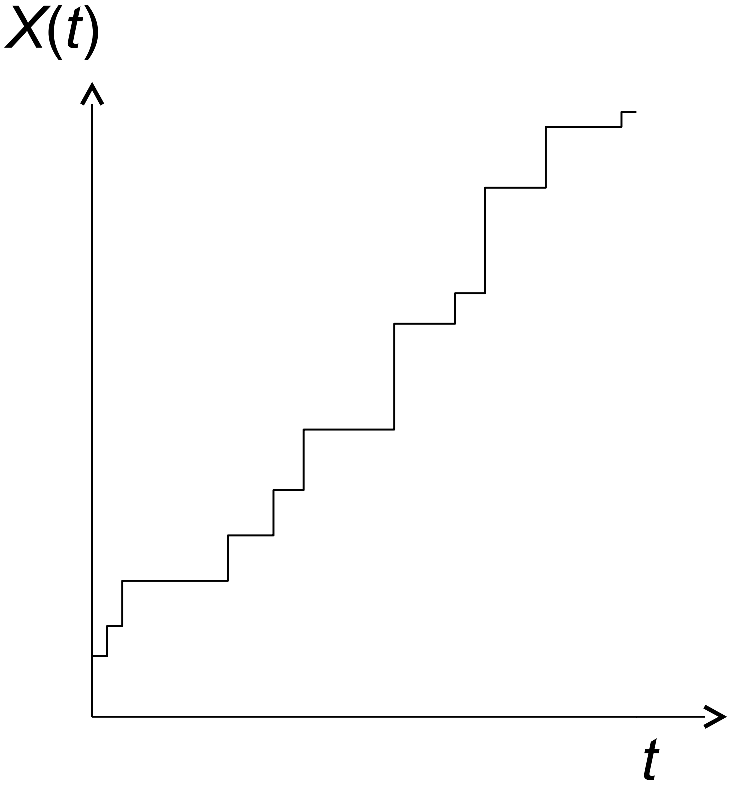
Traffic models

- The bounded variance network calculus uses a simple two-moment model of traffic
- That is, only the average value and variance of traffic is used
- This leads to an approximate analysis, but the quality of the approximation is in general very good
- Traffic is a random process which can be characterized by its instantaneous rate process
- An example of the instantaneous rate process of a variable bit rate source is shown in the figure



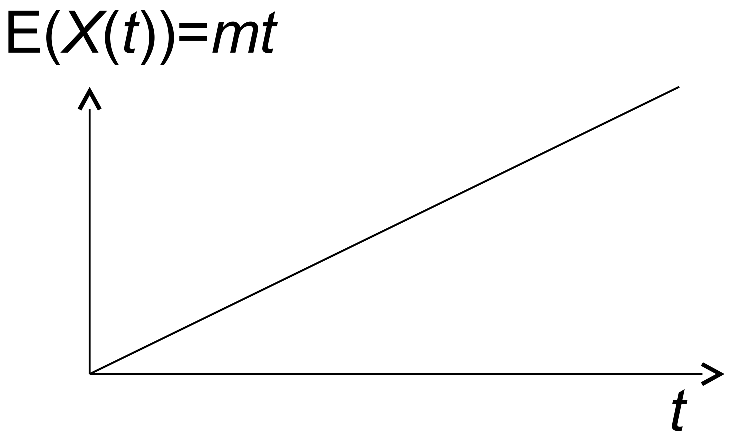
Traffic models

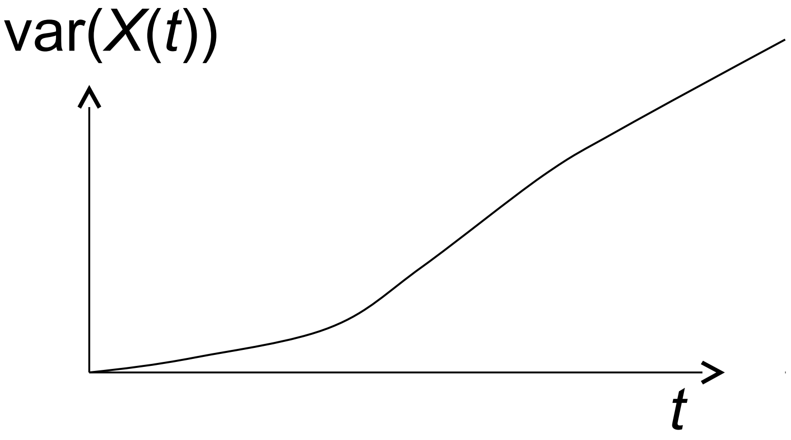
- The instantaneous rate process is referred to as r(t), and it is measured in bit/s
- Its average value is equal to m, and it is measured in bit/s
- Cumulative traffic is the total number of bits generated by a traffic source in a time interval with duration t
- It is referred to as X(t)
- An example of cumulative traffic for a variable bit rate source is shown in the figure



Traffic models

- We characterize traffic through it cumulative representation, X(t)
- In particular, we consider the two first moments of cumulative traffic, that is, mean and variance
- The mean value of X(t) is equal to mt
- The variance of X(t) is reterred to as var(X(t)) and its specific value depends on the type of traffic source
- An example of mean and variance of a cumulative traffic is shown in the figure (for a linearly-bounded variance traffic)



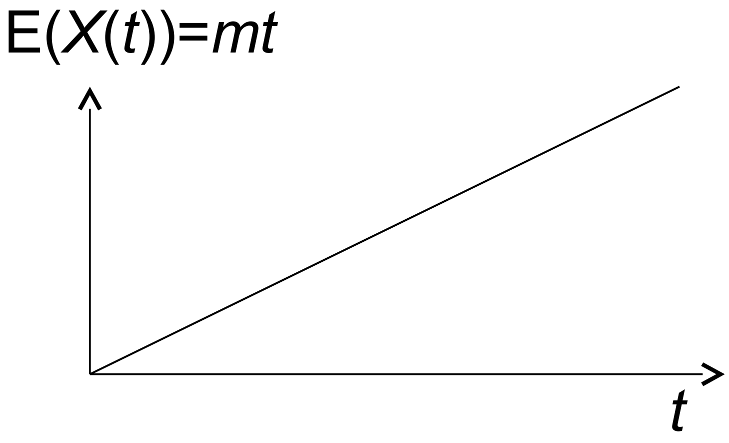


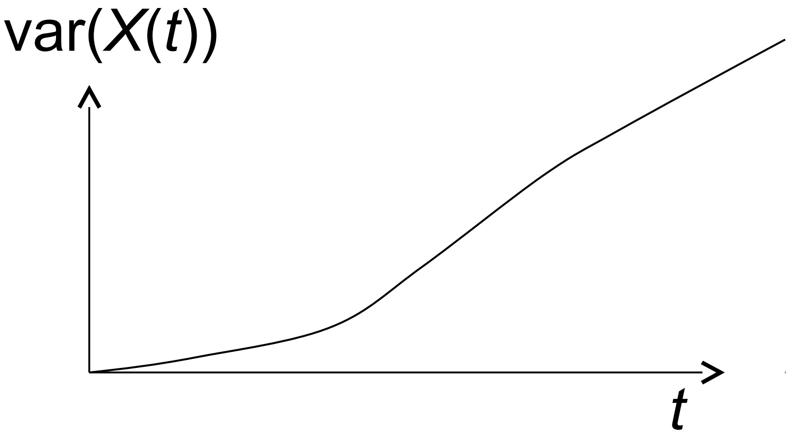
Traffic models

- We distinguish three main types of traffic sources
  - Deterministic
  - Variable bit rate, short range dependent
  - Variable bit rate, long range dependent
- Deterministic sources have a constant instantaneous rate
- The correct procedure to dimension links for deterministic traffic is that already shown for the EF PHB
- Deterministic traffic is out of the scope of statistical network calculus

Traffic models

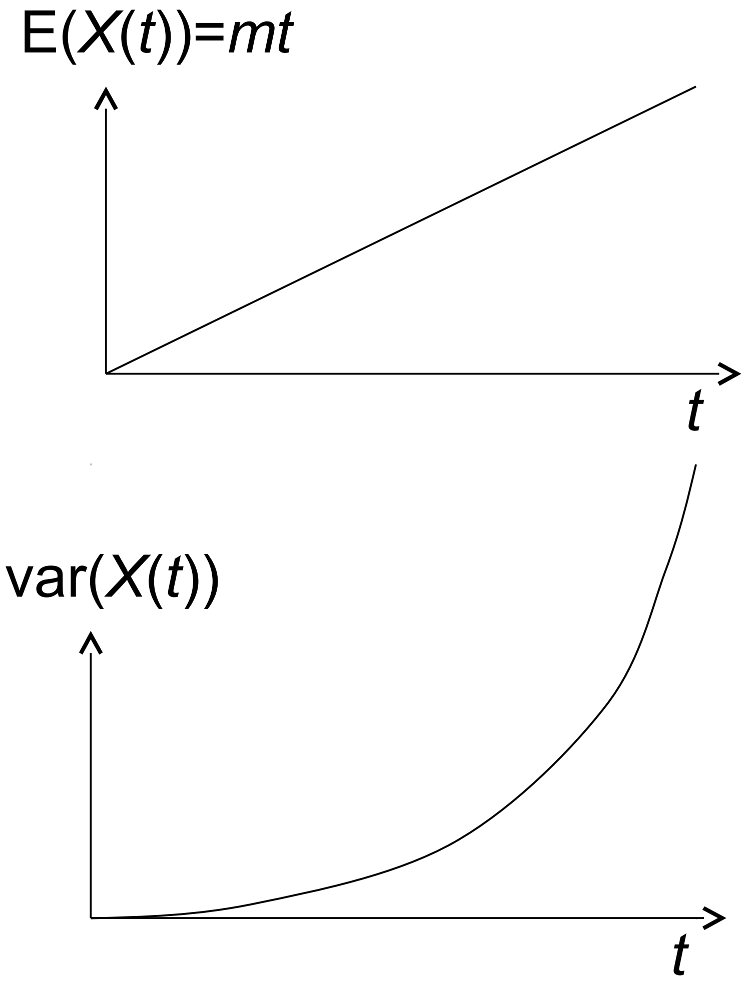
- Variable bit rate, short range dependent traffic can model sources such as VoIP and some types of data sources
- Short range dependent traffic exhibits a variance profile asymptotically linear as time grows
- In the figure an example of mean and variance of a short range dependent traffic is shown
- Frequently, there exists a linear upper bound of variance, so that we can use conservatively a completely linear model, that is, both mean and variance of short range dependent traffic can be assumed to be linear





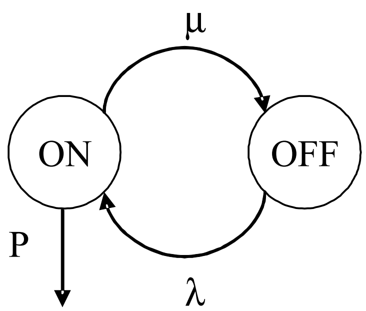
Traffic models

- Long range dependent (LRD) traffic exhibits a variance profile growing faster than a straight line as time increases
- Long range dependent traffic is widely present in the Internet
- For example, MP4 video traces are long range dependent
- Also data traffic can be long range dependent
- The variance profile of LRD traffic can be represented as $var(X(t))=kt^{\beta}$, with $\beta > 1$
- An example is shown in the figure



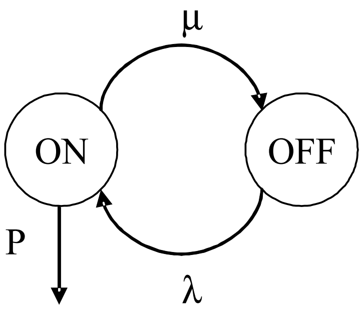
Traffic models: VoIP sources

- Many VoIP codecs can be reporesented as a two-state source, switching between active and inactive states
- In the active state, the source transmits λ packets per second and the length of each packet is L bits
- In the silent state, the source does not transmit
- The rate of transitions from active to inactive state is β
- The rate of transitions from inactive to active state is α



Traffic models: VoIP sources

- The average value and variance of the cumulative traffic generated by a VoIP source are shown in the figure
- There exists a linear envelope for the variance of the cumulative traffic
- VoIP sources are short-range dependent
- This facilitates significantly analysis and design



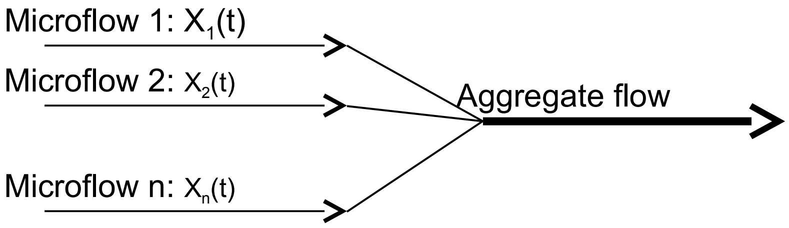
$$E\left(X(t)\right) = \lambda L \frac{\alpha}{\alpha+\beta} t$$

$$\mathrm{var}\left(X(t)\right) = \frac{\beta}{\alpha+\beta}\lambda\left(L^2+\sigma^2\right)t - 2\frac{\alpha\beta}{\left(\alpha+\beta\right)^3}\lambda^2 L^2 e^{-(\alpha+\beta)t}$$

$$\mathrm{var}\left(X(t)\right) \le \frac{\beta}{\alpha+\beta}\lambda\left(L^2+\sigma^2\right)t$$

Traffic models: aggregate flows

- The average value and variance of traffic will be used to determine performance
- In order to deal with traffic aggregates, it is fundamental to have a method for calculating the average value and variance of an aggregate flow, given the average value and variance of each individual microflow
- The assumption is that microflows are statistically independent
- In this case, the average value of the aggregate is the sum of the average values of the microflows
- In the same way, the variance of the aggregate is the sum of the variances of the microflows



$$m = \sum_{i=1}^{n} E\left(X_i\left(t\right)\right) = \sum_{i=1}^{n} m_i t$$

$$\mathrm{var}\left(X\left(t\right)\right) = \sum_{i=1}^{n} \mathrm{var}\left(X_i\left(t\right)\right)$$

Planning QoS

- An evolution of deterministic network calculus is the statistical, or stochastic, network calculus, designed to include statistical SLAs in its perimeter
- One of the most advanced current frameworks of the stochastic network calculus employs the min-plus algebra, by which it is possible to calculate statistical upper bounds on end-to-end delay
- The min-plus algebra-based stochastic network calculus is an effective method
- However, the statistical delay bounds of this framework are still very loose
- Thus, also with this type of stochastic network calculus, very conservative results are obtained, even if tighter than those provided by deterministic network calculus
- Moreover, results must always be obtained with rather complex numerical optimizations

Statistical network calculus

- Network calculus is a recent discipline which is able to simplify calculations and can account for more realistic traffic models
- However, also network calculus has some drawbacks
- The classical deterministic network calculus seeks for guaranteed delay bounds, and this leads to a very conservative design in case of statistical service level agreements
- Statistical network calculus can be a remedy to this problem, but the statistical delay bounds offered by the classical statistical network calculus are still quite conservative
- The bounded variance network calculus is a new discipline which aims at obtaining tight statistical approximations of network delay
- Therefore, we will develop this flawor of network calculus and we will show through examples how it is possible to design an IP network with QoS

Statistical network calculus

- Recently (2008) I with my research group at the Politecnico di Milano have introduced a new framework of stochastic network calculus: the *bounded variance network calculus*
- In this framework, we release the objective of obtaining true statistical upper bounds of delay
- We rather seek for tight approximations
- This means that the end-to-end delay forecasted by the bounded variance network calculus can be both higher and lower than the actual network performance
- However, the forecasted delay is usually very close to the real delay
- Moreover, it is possible to calculate rather easily end-to-end delay bounds, often analytically
- Thus, the bounded variance network calculus could be a valid means to obtain an evaluation of end-to-end delay

Representing traffic streams

- A traffic stream can be represented through its instantaneous rate $r(t)$, measured in bit/s, defined as the number of bits transmitted in a time interval with infinitesimal duration
- Stationary traffic streams will be considered, thus, $r(t)$ has a stationary average value m=E($r(t)$)
- The cumulative traffic transmitted in a time interval with duration t is an important characteristic of a traffic stream; it is referred to as $X(t_0, t_0+t)$, where t_0 is a reference time instant
- Since traffic is assumed to be stationary, the distribution of $X(t_0, t_0+t)$ is identical to the distribution of $X(t_1, t_1+t)$, for all t_1
- Therefore, the initial reference time instant can be neglected and, statistically, $X(t)$ fully represents the traffic process

Traffic streams: instantaneous rate and cumulative traffic

- The instantaneous rate and cumulative traffic are related by the relation

$$r(t) = \frac{dX(t)}{dt}$$

- The inverse relation is

$$X(t) = \int_0^t r(\tau)\,d\tau$$

Traffic streams: cumulative traffic

- The first and second moment of cumulative traffic are particularly important
- The first moment of cumulative traffic is

$$E\left(X\left(t\right)\right)=E\left(\int_{0}^{t}r\left(\tau\right)d\tau\right)=\int_{0}^{t}E\left(r\left(\tau\right)\right)d\tau=mt$$

- The variance of cumulative traffic is

$$\operatorname{var}\left(X\left(t\right)\right)=2\int_{0}^{t}\left(R_{r}\left(\tau\right)-m^{2}\right)\left(t-\tau\right)d\tau$$

- Where $R_r(\tau)$ is the autocorrelation function of the instantaneous rate

Traffic streams: cumulative traffic

- The autocovariance of the instantaneous rate is defined as $C_r(\tau)= R_r(\tau)-m^2$
- The shape of the rate’s autocovariance function determines the queueing behavior of traffic
- Short-range dependent traffic has an autocovariance function decreasing exponentially as time grows, that is

$$C_r(t) : e^{-t} \text{ for } t \to \infty$$

- Long-range dependent traffic has a sub-exponential autocovariance function

$$C_r(t) : t^{-\gamma} \text{ for } t \to \infty, \gamma > 0$$

Traffic streams: cumulative traffic

- The variance profile of the cumulative traffic of a short-range dependent stream is asymptotically linear for growing time

$$\operatorname{var}\left(X\left(t\right)\right) : kt \text{ for } t \to \infty, k > 0$$

- The variance profile of the cumulative traffic of a long-range dependent stream grows with a power law

$$\operatorname{var}\left(X\left(t\right)\right) : kt^{2-\gamma} \text{ for } t \to \infty, k > 0$$

- It will be shown that the queueing behavior of a traffic stream depends on its variance profile
- The queueing behavior with a power-law variance profile is quite different from that of a stream with a linear variance profile

Self-similar traffic

- Very often long-range dependence and self-similarity are closely related features of traffic, that is, very frequently self-similar traffic is long-range dependendt and vice-versa
- Self-similarity is a property of traffic such that a scaled version of the traffic stream, both in time and amplitude, has the same statistical properties of the original traffic stream
- In particular, the following property holds

$$X(t) \overset{d}{=} a^{-H} X(at), \quad \forall a > 0, t \geq 0$$

Self-similar traffic

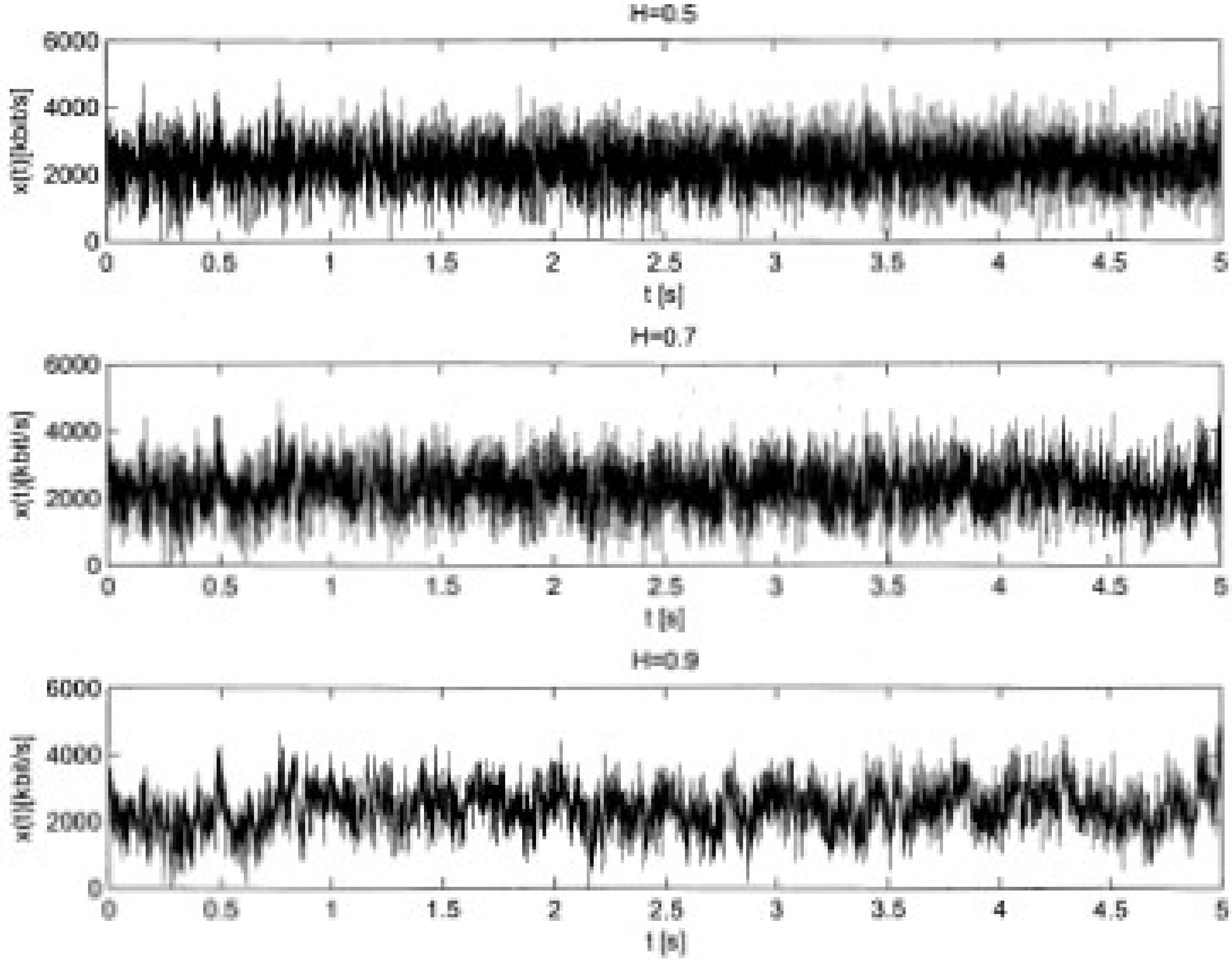
- An example of both self-similar and long-range dependent traffic is the fractional Gaussian traffic (fGt)
- The cumulative traffic of a fGt stream exhibits the following properties

$$E\left(X\left(t\right)\right)=mt$$

$$\mathrm{var}\left(X\left(t\right)\right)=amt^{2H}$$

- *m* is the average rate of the stream, *a* is a constant measured in (bit . s), and *H* is the Hurst parameter of traffic, usually ranging from ½ to 1
- The larger *H*, the longer the correlation of the traffic process

Fractional Gaussian traffic



- Three traces of fGt traffic are represented in the figures
- The parameters are
- *m*=2.279 Mbit/s
- *a*= 773950 bit.s
- *H*=0.5, 0.7, 0.9
- With H=0.5 the traffic process is memoryless (i.e., it is a white noise)
- With *H*=0.7 it exhibits a stronger correlation
- The correlation is higher with *H*=0.9

Long-range dependent traffic

- It is well known that in IP networks in general and in the Internet in particular long-range dependent traffic is a significant fraction of the transported volume of traffic
- Knowing how to deal with long-range dependent traffic is thus important, but long-range dependent traffic falls out of the perimeter of classic markovian analysis
- Therefore, new methods to calculate the network performance with long-range dependent traffic are needed
- Stochastic network calculus is able to treat short-range dependent and long-range dependent traffic in a unique framework

Two-moment analysis

- The bounded-variance network calculus works with the first two moments of traffic, that is, with the average value and variance of cumulative traffic
- This corresponds to account for the average value and autocovariance function of the instantaneous rate
- By accounting for just two moments of traffic leads to an approximated analysis, if traffic is not Gaussian
- The positive aspects of the two-moment analysis are
  - It is simple
  - It is general
  - Analytical results are possible
  - Multihop analysis is possible
  - Approximations can be very good, especially when aggregates of many micro-flows are studied

Aggregate flows

- When a stream *X*(*t*) is an aggregate of *N* independent micro-flows *x*$_i$(*t*), its average value and variance are calculated as

$$E\left(X\left(t\right)\right)=\sum_{i=1}^{m}E\left(x_i\left(t\right)\right)=\sum_{i=1}^{m}m_i t$$

$$\mathrm{var}\left(X\left(t\right)\right)=\sum_{i=1}^{m}\mathrm{var}\left(x_i\left(t\right)\right)$$

- If the micro-flows are statistically identical, that is, E(*x*$_i$(*t*)) = *m* and var(*x*$_i$(*t*)) = var(*x*$_j$(*t*)), then

$$E\left(X\left(t\right)\right)=Nmt$$

$$\mathrm{var}\left(X\left(t\right)\right)=N\,\mathrm{var}\left(x_i\left(t\right)\right)$$

Two-moment analysis

- Let us consider a scheduler where in service class *i* the cumulative traffic $X_i(t_1, t_2)$ is offered in the time interval $[t_1, t_2]$
- In the same time interval, a cumulative traffic $Y_i(t_1, t_2)$ is served for class *i*
- At time *t*, $Q_i(t)$ is defined as the amount of traffic backlogged in the queue of service class
- The time interval $[t_1, t_2]$ is defined as a *backlog interval* for the queue of service class *i* if for all *t* in $[t_1, t_2]$, $Q_i(t) > 0$
- The *minimal backlogged input traffic* $X_i^*(t_1, t_2)$ is the minimal amount of traffic that must be offered to service class i to keep continually its queue in a backlog state in $[t_1, t_2]$
- The *available service* $Y_i^*(t_1, t_2)$ for class *i* in $[t_1, t_2]$ is the traffic served for class *i* in $[t_1, t_2]$ when the class is offered its minimal backlogged input traffic

Two-moment analysis

- In general, in the time interval [t_1, t_2] the amount of served traffic for class *i* is referred to as $Y_i(t_1, t_2)$
- The *virtual delay* of the traffic of class *i* is defined as

$$D_i(t) = \min\left(\Delta t \geq 0 : X_i(t_0, t) \geq Y_i(t_0, t + \Delta t)\right)$$

- A delay-oriented SLA is formalized as

$$\Pr\left(D_i(t) > d_i\right) \leq p_i$$

- For the event ($D_i(t)>d_i$) the following identity holds

$$\left(D_i(t) \geq d_i\right) \equiv \left(X_i(t_1, t) - Y_i(t_1, t + d_i) > 0\right)$$

Two-moment analysis

- Thus,

$$\Pr\left(D_i\left(t\right)\geq d_i\right)\equiv\Pr\left(X_i\left(t_1,t\right)-Y_i\left(t_1,t+d_i\right)>0\right)$$

- and the following inequality holds

$$\Pr\left(D_i\left(t\right)\geq d_i\right)\leq\Pr\left(\max_{t\geq t_1}\left(X_i\left(t_1,t\right)-Y_i\left(t_1,t+d_i\right)\right)>0\right)$$

- Moreover, by definition of available service

$$\Pr\left(D_i\left(t\right)\geq d_i\right)\leq\Pr\left(\max_{t\geq t_1}\left(X_i\left(t_1,t\right)-Y_i^*\left(t_1,t+d_i\right)\right)>0\right)$$

Two-moment analysis

- The previous analysis is general and holds for non-stationary traffic
- By exploiting the assumption that traffic is stationary, it is possible to simplify the formalism:

$$\Pr\left(D_i \geq d_i\right) \leq \Pr\left(\max_{t\geq 0}\left(X_i\left(t\right) - Y_i^*\left(t+d_i\right)\right) > 0\right)$$

- The last step is to introduce the notion of service envelope
- For class *i*, the service envelope *S*$_i$(*t*) is a stochastic process such that $\forall t \geq 0, \forall z \geq 0 : \Pr\left(Y_i^*\left(t\right) \geq z\right) \geq \Pr\left(S_i\left(t\right) \geq z\right)$
- That is, the service envelope is a statistical lower bound of the available service

Two-moment analysis

- Thus, the following inequality can be established

$$\Pr\left(D_i \geq d_i\right) \leq \Pr\left(\max_{t\geq 0}\left(X_i\left(t\right) - S_i\left(t + d_i\right)\right) > 0\right)$$

- The problem of calculating the probability of violating a statistical delay SLA is reduced to the calculus of the previous probability
- However, two problems must still be faced
  - First, the service envelope of the relevant schedulers must be known
  - Second, the calculus of the probability shown in this slide is very complex and a simplified path must be found
- The first problem is simple, as the service envelope of the most important schedulers are already known
- The second problem requires more analysis

Two-moment analysis

- Let us consider for example the simple case of a FIFO scheduler with a link capacity equal to *C*
- The service envelope is elementary: *S*(*t*)=*Ct*
- In fact, the amount of service available for *X*(*t*) in a time interval of duration *t* is equal to the line capacity times the duration of the service interval
- The calculus of the SLA becomes

$$\Pr\left(D \geq d\right) \leq \Pr\left(\max_{t \geq 0}\left(X(t) - C(t+d)\right) > 0\right)$$

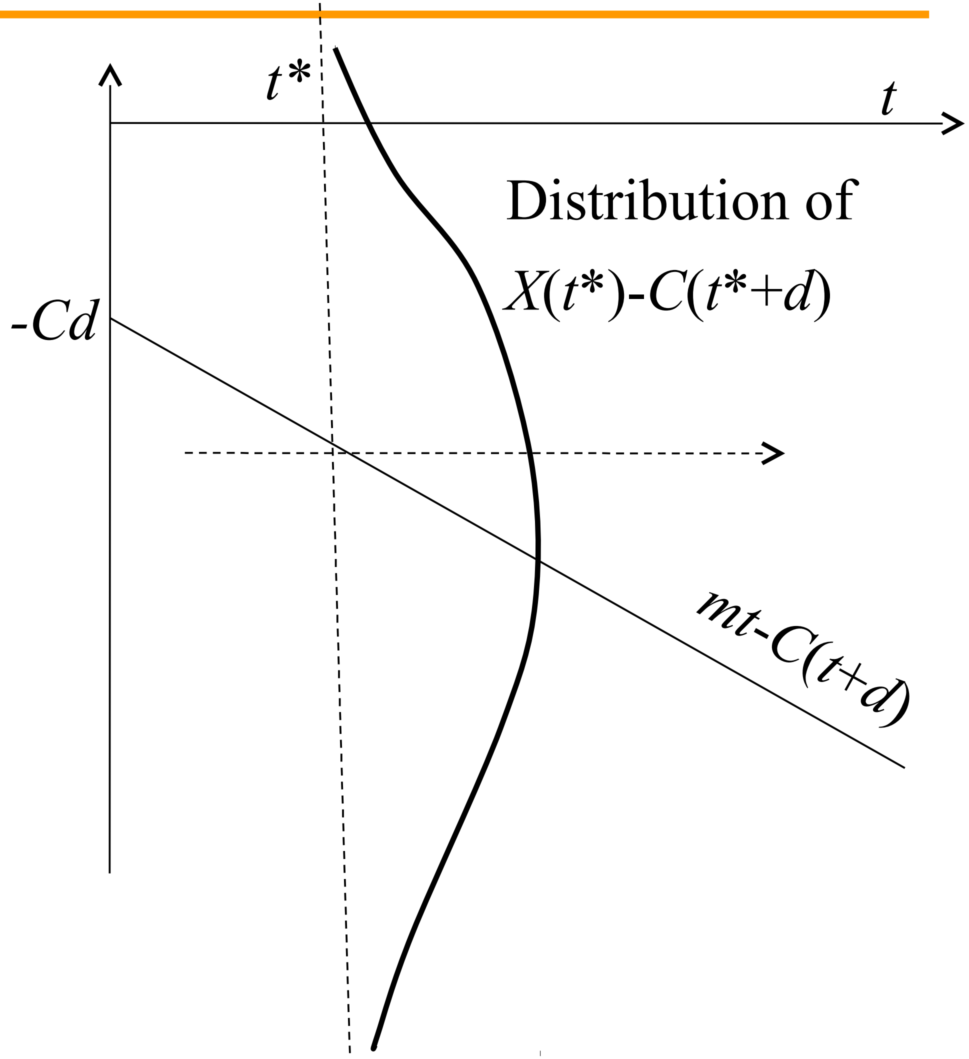
- As reported before, this calculus is very complex; in practice, we must calculate the probability that the absolute maximum of the process *X*(*t*)-*C*(*t*+*d*) is greater than 0

Two-moment analysis

- Since the calculus is complex, simplified methods have been introduced
- Choe and Shroff propose the Maximum Variance Asymptotic (MVA) upper bound, providing an upper bound for the probability of violating the statistical delay SLA, when traffic is Gaussian
- The derivation of the MVA bound is quite complex, it is based aon extreme values theories
- Only the main result will be presented

Two-moment analysis

- If the average instantaneous rate of the input traffic process is m, the average value of $X(t)$ is equal to mt
- The process $X(t)$-$C(t+d)$ has a negative drift, as its average rate decreases as mt-$C(t+d)$
- The variance of $X(t)$-$C(t+d)$ increases as time grows and it is equal to var($X(t)$)
- The time instant t^* in which the process $X(t)$-$C(t+d)$ is more likely to become positive is when the ratio of its average value to its standard deviation is minimal

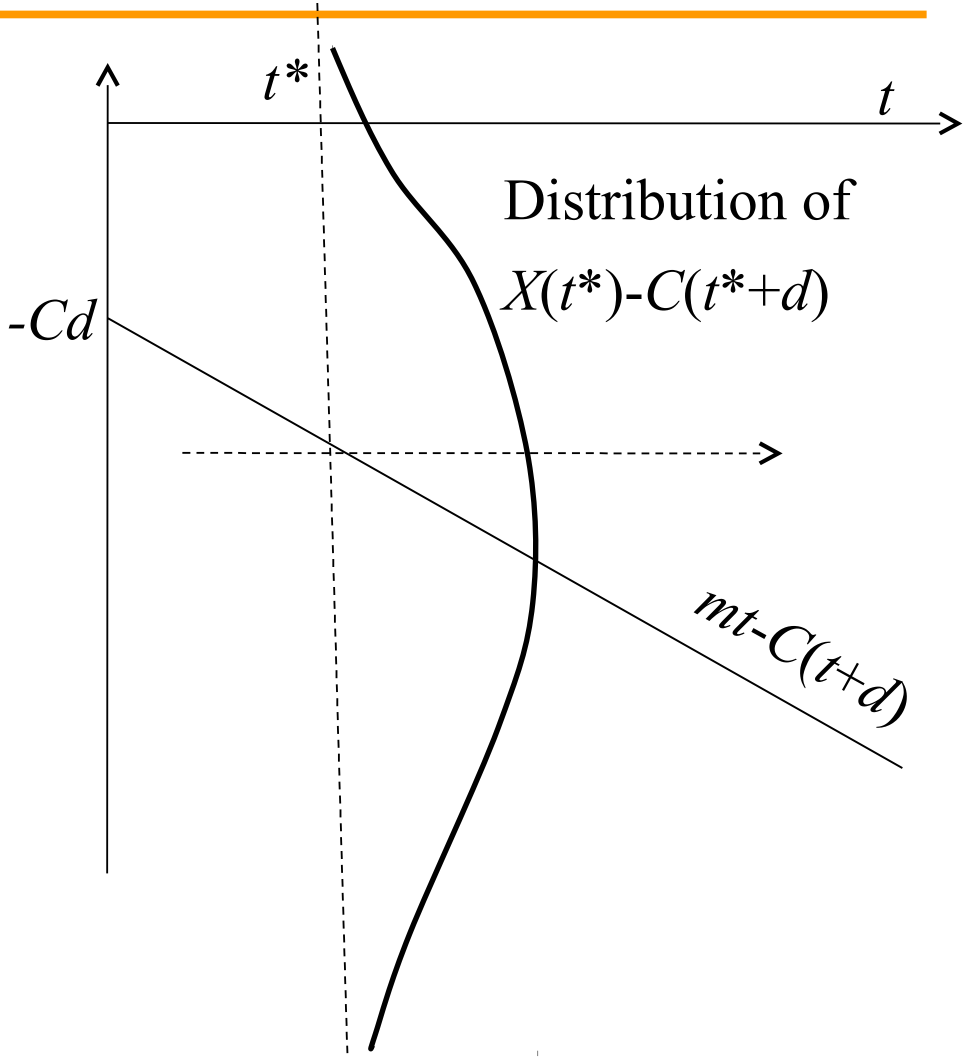


Two-moment analysis

- Choe and Shroff, through extreme values analysis, have proved that if traffic is Gaussian

$$\Pr\left(X\left(t^* - C\left(t^* + d\right)\right) \geq 0\right) \geq \Pr\left(D > d\right)$$

- Thus, the problem is reduced to the calculus of *t**, the time instant in which it is more likely that *X*(*t*)-*C*(*t*+*d*) crosses the time axis



Two-moment analysis

- The procedure (Maximum Variance Asymptotic upper bound) is carried out as follows
- First, define

$$\alpha_i(t) = -\frac{E\left(X_i(t) - S(t+d_i)\right)}{\sqrt{\operatorname{var}\left(X(t) - S(t+d_i)\right)}}$$

- Second, calculate

$$\alpha_{i,\min} = \min_{t \geq 0} \alpha_i(t)$$

- Third, calculate

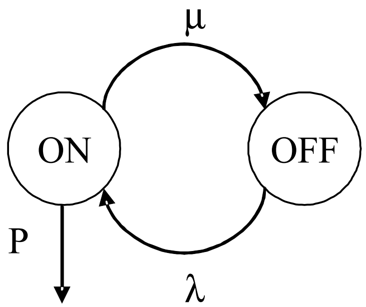
$$\Pr\left(D_i \geq d\right) \leq e^{-\frac{\alpha_{i,\min}^2}{2}}$$

Two-moment analysis

- The illustrated procedures produces statistical upper bounds on delay SLAs when traffic is Gaussian
- When traffic is not Gaussian, by applying the same procedure (i.e., by neglecting the third and higher-order moments of traffic) we obtain an approximation rather than an upper bound
- The quality of the approximation grows as the number of micro-flows composing the traffic aggregates increases
- This is due to the Central-Limit theorem, which states that the distribution of the sum of independent random variables converges to a Gaussian for an increasing number of summed variables

Two-moment analysis: voice codecs

- In order to apply this analysis to practical cases, it is necessary to characterize the first and second moments of relevant traffic streams
- The first type of considered stream is a variable bit rate voice codec that can be represented as a two-state Markov chain, for example, the G.726 codec



- The parameters of this type of source have been described previously

The G.726 voice codec

- For this codec, the average value of the cumulative traffic *x*(*t*) is equal to

$$E\left(x\left(t\right)\right)=\frac{\lambda}{\lambda+\mu}Pt$$

$$\operatorname{var}\left(x\left(t\right)\right)=2\frac{\lambda\mu}{\left(\lambda+\mu\right)^{3}}P^{2}t-2\frac{\lambda\mu}{\left(\lambda+\mu\right)^{4}}P^{2}\left(1-e^{-\left(\lambda+\mu\right)t}\right)$$

- The variance can be linearly upper bounded

$$\operatorname{var}\left(x\left(t\right)\right)\leq 2\frac{\lambda\mu}{\left(\lambda+\mu\right)^{3}}P^{2}t$$

Markov sources in general

- Given a traffic source modeled with a generic markov chain, with an arbitrary number of states, it can be difficult to calculate its variance
- A recent theorem by Giacomazzi and Verticale proves that all Markov sources admit a linear upper-bound for the variance of the cumulative generated traffic
- This simplifies significantly the analysis as, if a scheduler is subject to traffic streams which can be modelled as Markov sources, the variance of the incoming cumulative traffic will be linear
- That is, for the traffic offered to service class *i*:

$$E\left(X_i\left(t\right)\right)=r_i t$$

$$\operatorname{var}\left(X_i\left(t\right)\right)\leq r_i b_i t$$

Markov sources in general and the FIFO scheduler

- Therefore, by analyzing a scheduler which is offered traffic with a linear variance envelope, that scheduler will be studied automatically, in one shot, for all possible combinations of any set of Markov sources, with arbitrary transitions rate and per-state transmission speeds
- The analysis will start from the FIFO scheduler
- We consider a FIFO scheduler with line capacity equal to *C*
- We asume that *N* statistically identical traffic streams with linear variance envelope are offered
- The average rate of an individual traffic stream is *r*
- The variance of the cumulative traffic of each individual traffic stream is equal to *rbt*

Markov sources in general and the FIFO scheduler

- The MVA is applied

$$\alpha(t) = -\frac{Nrt - C(t+d)}{\sqrt{Nrbt}}$$

$$\alpha_{\min} = 2\sqrt{\frac{Cd(C-Nr)}{Nrbt}}$$

$$\Pr(D>d) \approx e^{-\frac{\alpha_{\min}^2}{2}} = \exp\left(-2\frac{C(C-Nr)}{Nrb}d \div\right)$$

$$E(D) = \frac{Nrb}{2C(C-Nr)}$$

G.726 codec and FIFO scheduler

- For a G.726 codec, the parameters are

$$r = \frac{\lambda}{\lambda+\mu} P$$

$$rb = 2\frac{\lambda\mu}{\left(\lambda+\mu\right)^3} P^2$$

- Thus

$$\Pr\left(D > d\right) \approx \exp\left(-2\frac{C\left(C - N\frac{\lambda}{\lambda+\mu}P\right)}{2N\frac{\lambda\mu}{\left(\lambda+\mu\right)^3}P^2} d \right)$$

FIFO scheduler with linear-variance sources : capacity allocation

- Given a FIFO scheduler multiplexing *N* streams, each with average rate *r* and variance of cumulative traffic *rbt*, the capacity allocation procedure consists in finding the line capacity *C* needed to fulfill the statistical delay SLA (*d*, *p*)
- This is done by calculating the probability of violating the delay SLA and by solving the resulting equation for *C*
- The result is

$$C \geq \frac{Nr}{2} + \sqrt{\left(\frac{Nr}{2}\right)^2 + \frac{Nb(-\ln p)}{2d}}$$

FIFO scheduler with linear-variance sources: admission control

- Given a scheduler with a line capacity equal to *C*, offered traffic from statistically identical streams with average rate *r* and variance of cumulative traffic *rbt*, the admission control calculus finds the maximum number of sources that can be accepted in order to fulfill the statistical delay SLA (*d*, *p*)
- This is done by calculating the probability of exceeding the delay bound and by solving it for *N*
- The result is

$$N \le \frac{2C^2 d}{2Cdr - b\ln p}$$

FIFO scheduler with fractional Gaussian traffic

- Fractional Gaussian traffic has average value of cumulative traffic equal to *mt* and variance equal to amt^{2H}
- Given a FIFO scheduler with line capacity equal to *C*, we calculate the probability of violating the statistical delay SLA (*d, p*)
- By applying the MVA:

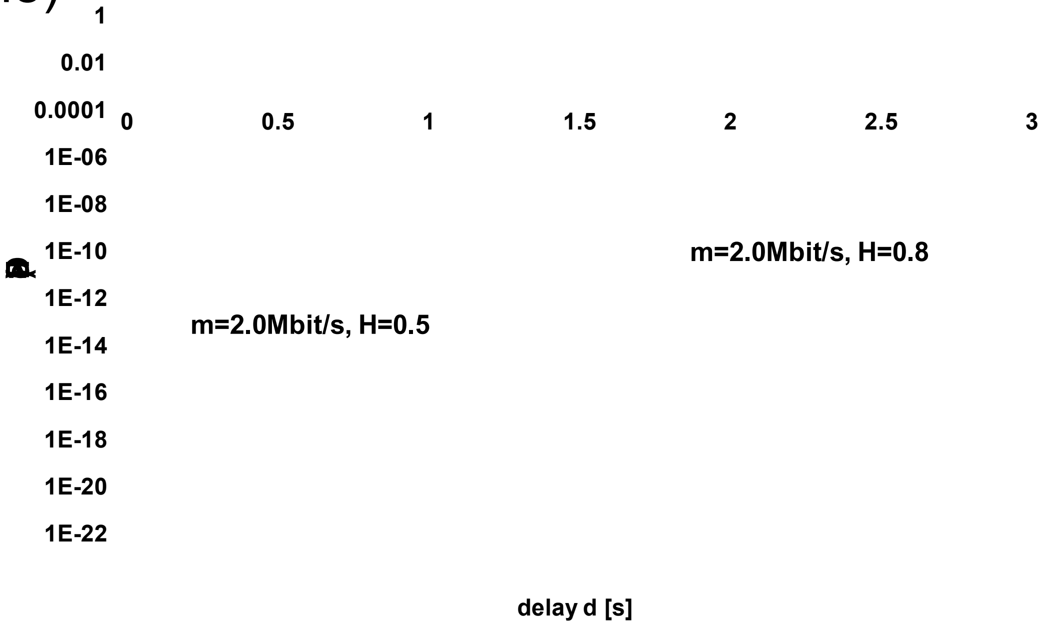
$$\alpha(t) = -\frac{mt - C(t+d)}{\sqrt{amt^{2H}}}$$

$$\alpha_{\min} = \frac{1}{\sqrt{am}} \frac{1}{H^H (1-H)^{1-H}} C^{1-H} (C-m)^H d^{1-H}$$

$$\Pr(D > d) \approx e^{-\frac{\alpha_{\min}^2}{2}} = \exp\left(-\frac{1}{2am} \frac{1}{H^{2H} (1-H)^{2-2H}} C^{2-2H} (C-m)^{2H} d^{2-2H}\right)$$

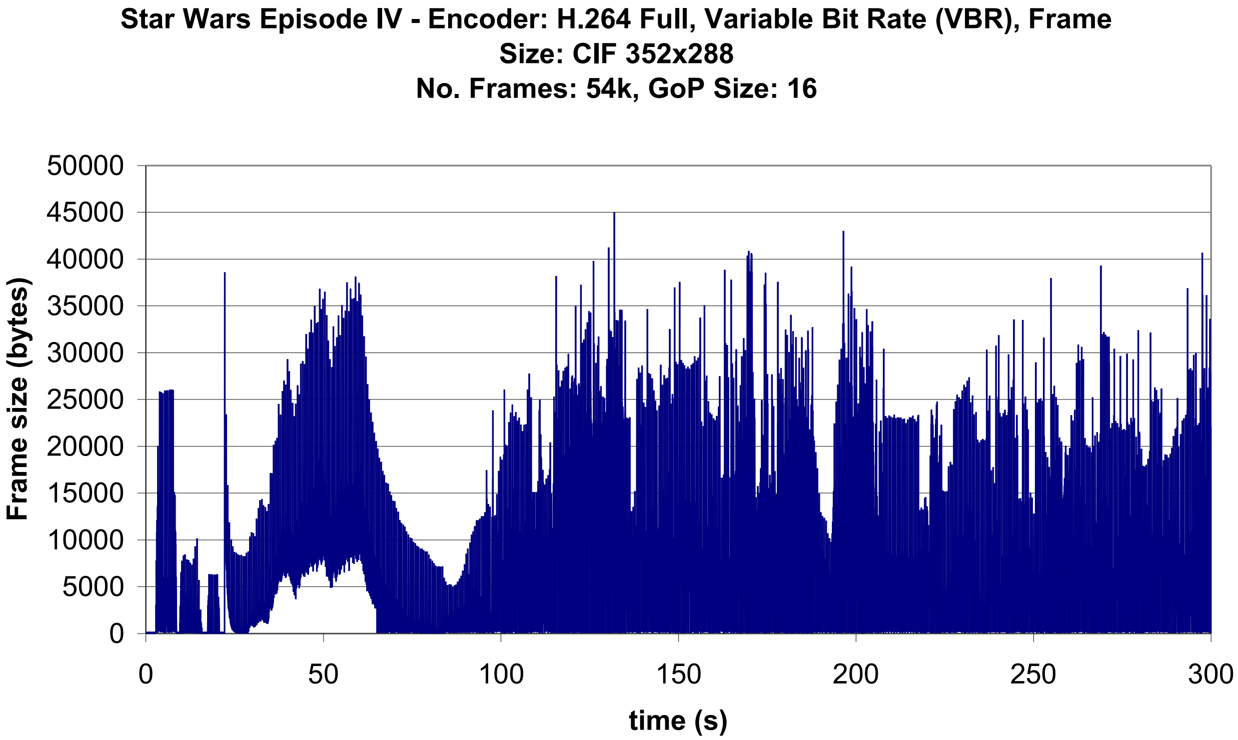
FIFO scheduler with fGt

- The delay distribution for a FIFO scheduler with fGt traffic is Weibull and not exponential
- The Weibull distribution is fat-tailed, that is, it decreases very slowly
- This means that extremely high delay peaks are possible
- This does not hold for short-range dependent traffic with a linear variance envelope (red line)



fGt equivalents of video traffic

- Recently (Giacomazzi-Saddemi) a method to obtain a fGt equivalent of video traces has been proposed in the literature
- With this method, it is possible to treat video traffic traces as fGt traffic, obtaining delay curves similar to those of the real traces

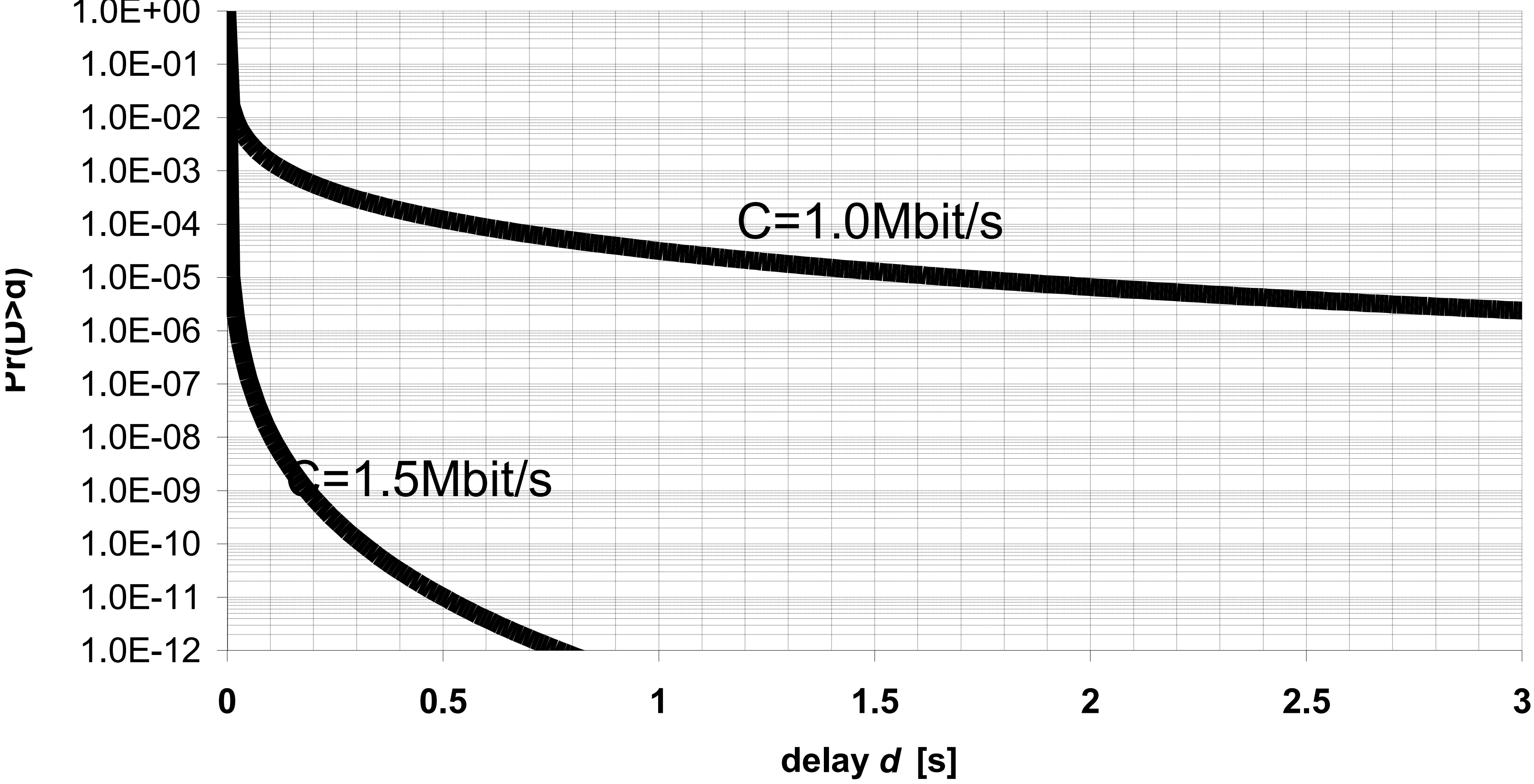


fGt equivalents of video traffic

- For the Star Wars video reported in the previous slide, transported over an Ethernet link with IEEE 802.1Q LAN feature, with maximum payload length of the IP packet equal to 500 bytes, the features of the equivalent fGt traffic are
  - *m* = 283,000 bit/s
  - *a*=180,500 bit s
  - *H*=0.8988
- The resulting delay distribution is that of the equivalent fGt traffic, which approximates very well the real delay curve
- The plot of the delay distribution is shown in the next slide

fGt equivalents of video traffic

Star Wars Episode IV - Encoder: H.264 Full, Variable Bit Rate (VBR), Frame Size: CIF 352x288 No. Frames: 54k, GoP Size: 16, Ethernet transport with IEEE 802.1Q VLAN, IP video payload size = 500 bytes


1.0E+00
1.0E-01
1.0E-02
1.0E-03
1.0E-04
1.0E-05
1.0E-06
1.0E-07
1.0E-08
1.0E-09
1.0E-10
1.0E-11
1.0E-12
Pr(D>d)
C=1.0Mbit/s
C=1.5Mbit/s
delay *d* [s]