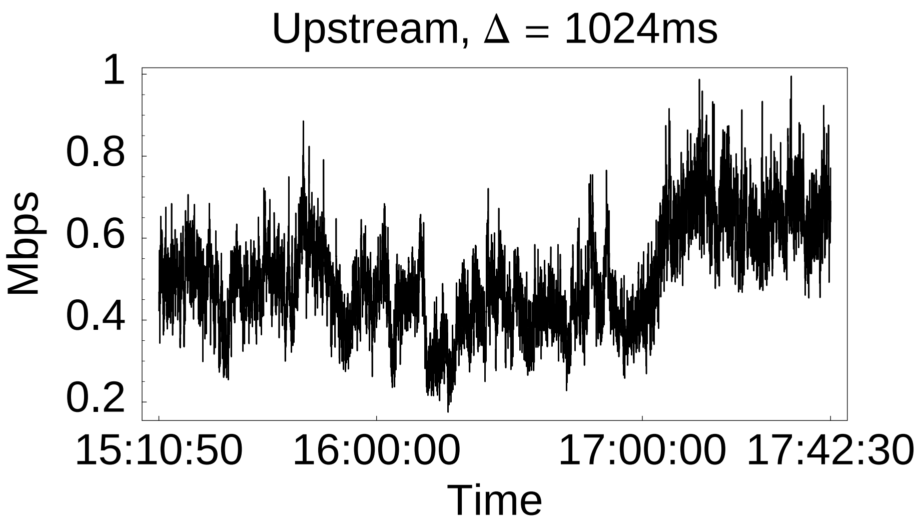
Testing the Gaussian character of access network traffic

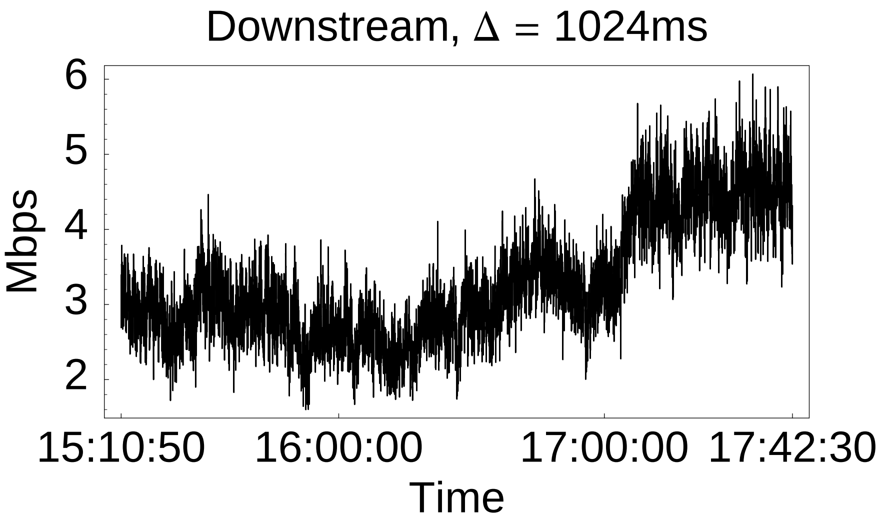
Jorma Kilpi, Ilkka Norros

VTT Information Technology

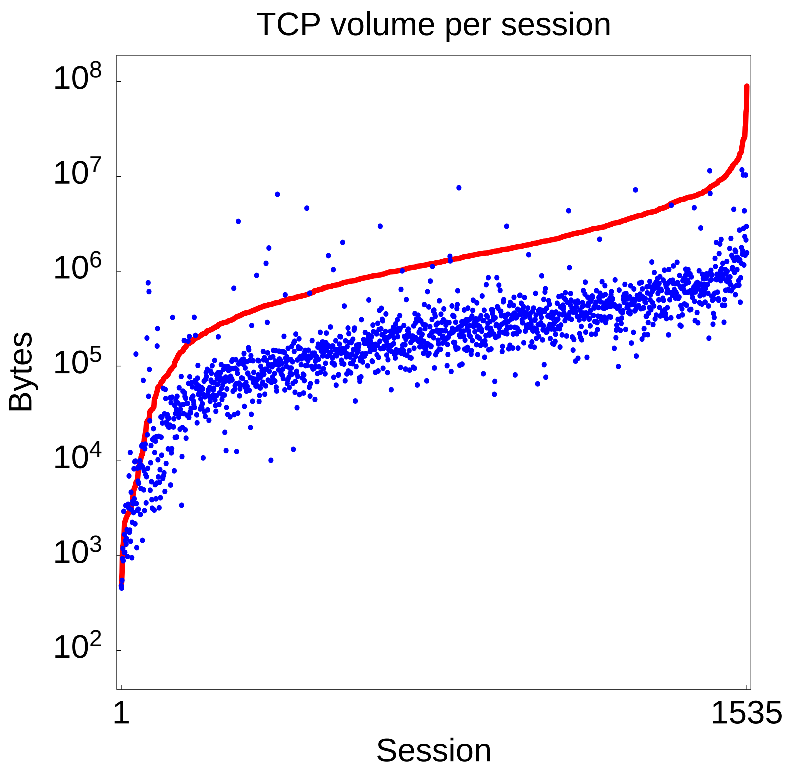
October 15, 2001

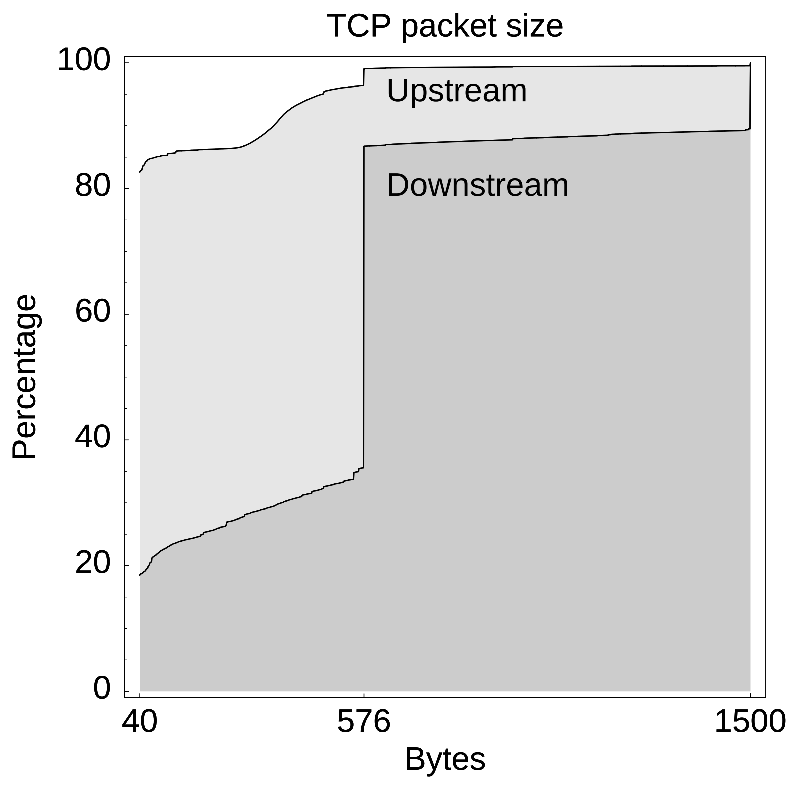
Time series of TCP traffic:





Description of the data:

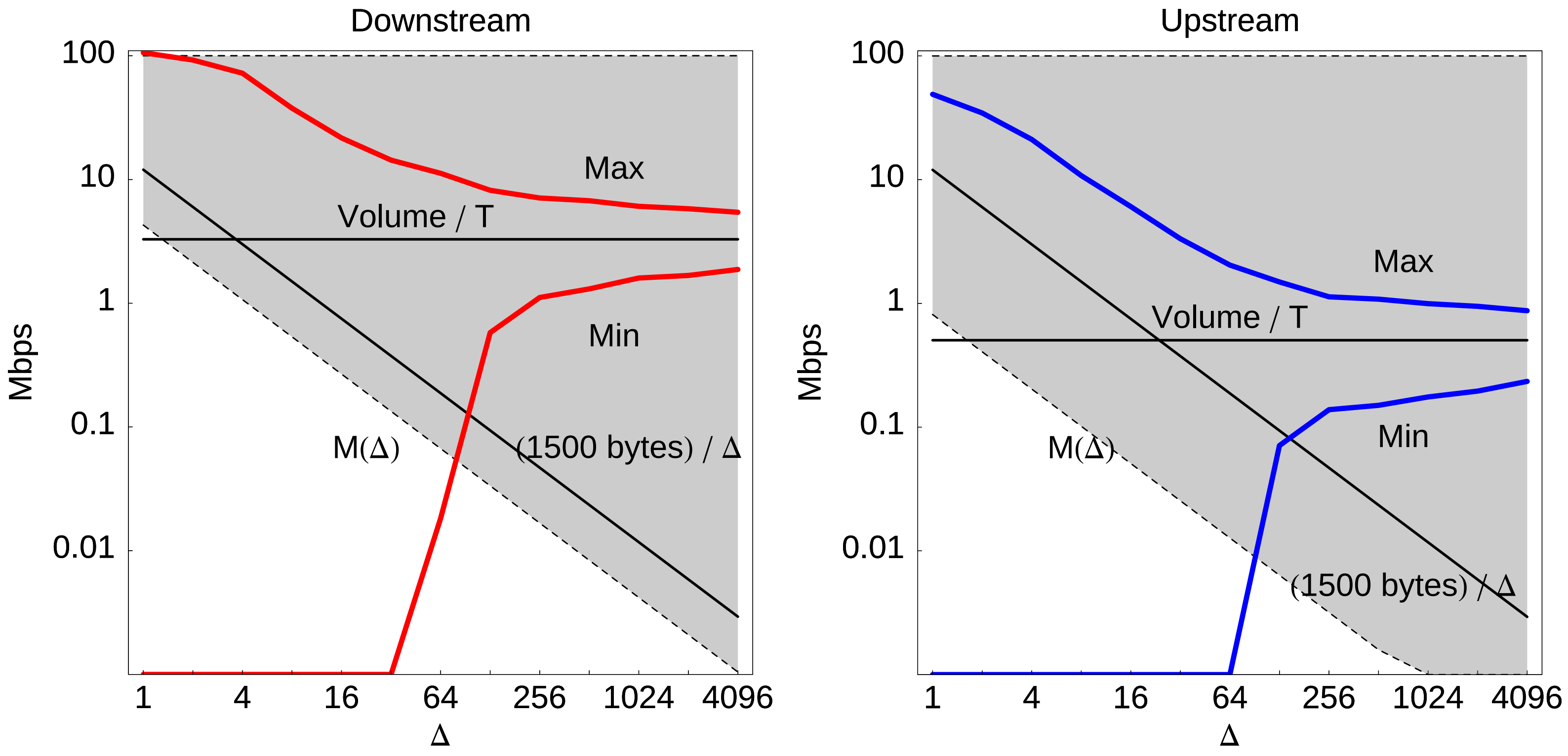




Why should traffic be Gaussian?

- *Horizontal* aggregation: working at a sufficiently large resolution.
- *Vertical* aggregation: presence of a large number of contributors.
- When capacity is not a restriction the contributors can be assumed to be independent, hence *Central Limit Theorem* (CLT) should work for vertical aggregation.
- $M(\Delta) = (\text{mean packet size})/\Delta$. Mean packet sizes were 102 bytes upstream and 536 bytes downstream.

Horizontal aggregation:



Vertical aggregation:

- For fixed $0 < p < 1$ find the smallest number n such that

$$P\left\{\sum_{j=1}^{n} X_j > M(\Delta)\right\} \geq p.$$

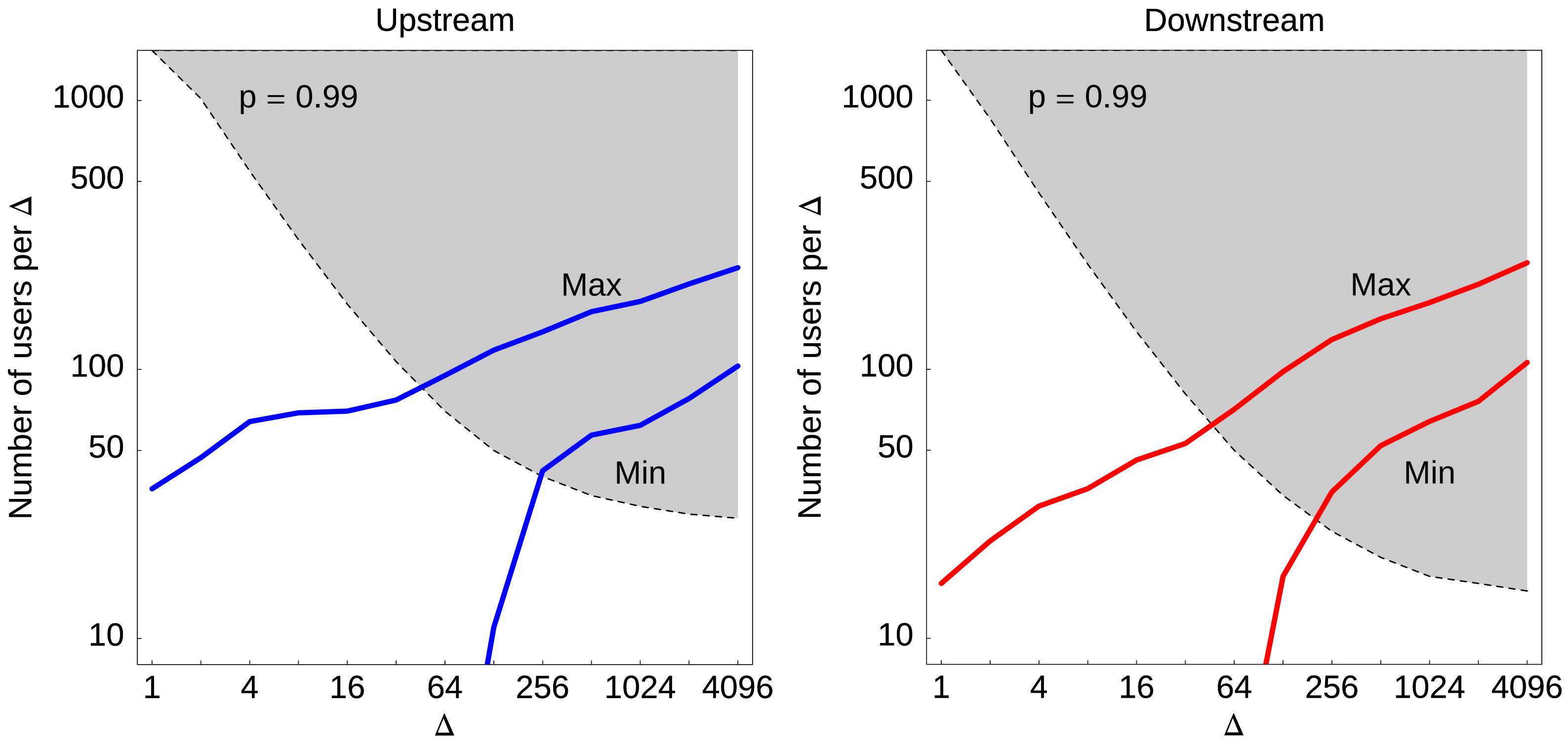
- Normal approximation: If

$$P\left\{\sum_{j=1}^{n} X_j > M(\Delta)\right\} \approx 1 - \Phi\left(\frac{M(\Delta) - \mu n}{\sigma\sqrt{n}}\right),$$

  then n is the minimal solution of

$$\frac{M(\Delta) - \mu n}{\sigma\sqrt{n}} \leq \Phi^{-1}(1-p).$$

Vertical aggregation:



Description of the methods:

- *Quantile-Quantile plots:*

$$\left(\Phi^{-1}\left(\frac{i}{N+1}\right), x_{(i)}\right), \qquad i = 1, \ldots, N,$$

  N is the sample size, $x_{(1)}, \ldots, x_{(N)}$ are order statistics and Φ^{-1} is inverse of standard normal cdf.

- $\Phi^{-1}_{\mu,\sigma}(p) = \mu + \sigma\Phi^{-1}(p)$, $0 < p < 1$.
- Plots are presented together with least square error sense fitted line.

Description of the methods:

- *Goodness of fit:*

$$r_Q = \frac{\sum_{i=1}^{N}(x_{(i)} - \overline{x})(q_i - \overline{q})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(q_i - \overline{q})^2}},$$

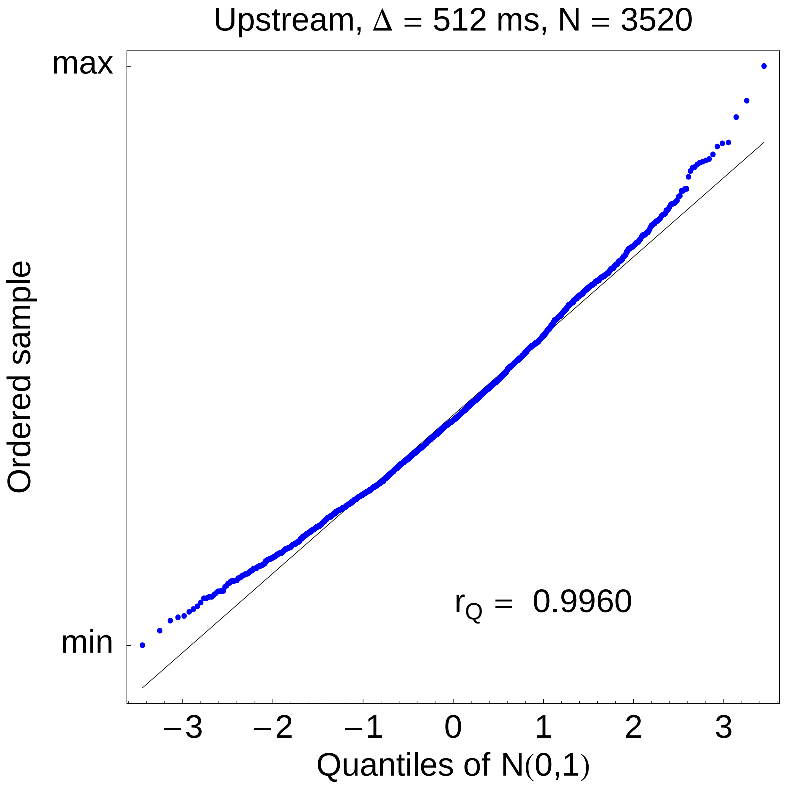
where $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and q_i are quantiles of the model distribution with $\overline{q} = \frac{1}{N}\sum_{i=1}^{N} q_i$. Clearly $|r_Q| \leq 1$ and $r_Q = 1$ if and only if all points lie perfectly on a straight line.

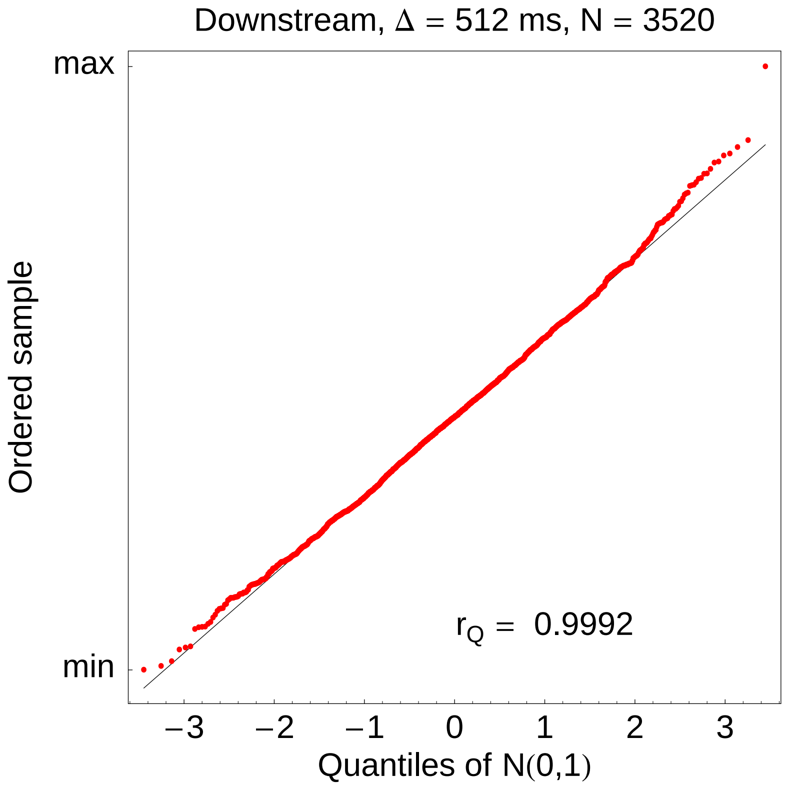
- *Shapiro-Wilk test of normality:*

$$\left(Ez_{(i)}, x_{(i)}\right), \qquad i = 1, \ldots, N,$$

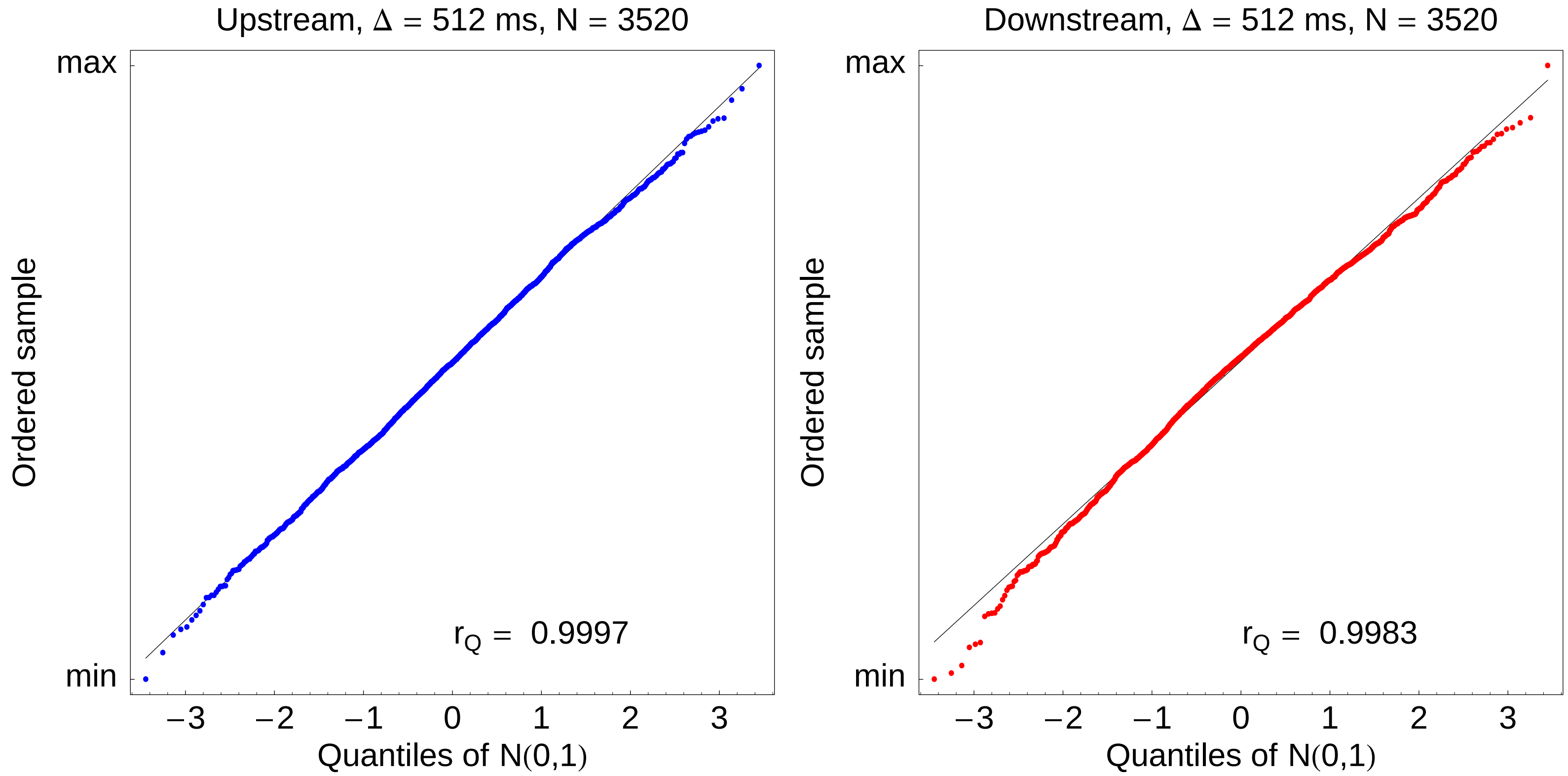
where z_i, $i = 1, \ldots, N$ is a sample of size N from standard normal distribution, test statistic is r_Q^2 where $q_i = Ez_{(i)}$ and $\overline{q} = 0$ by symmetry.

Examples of normal probability plots:

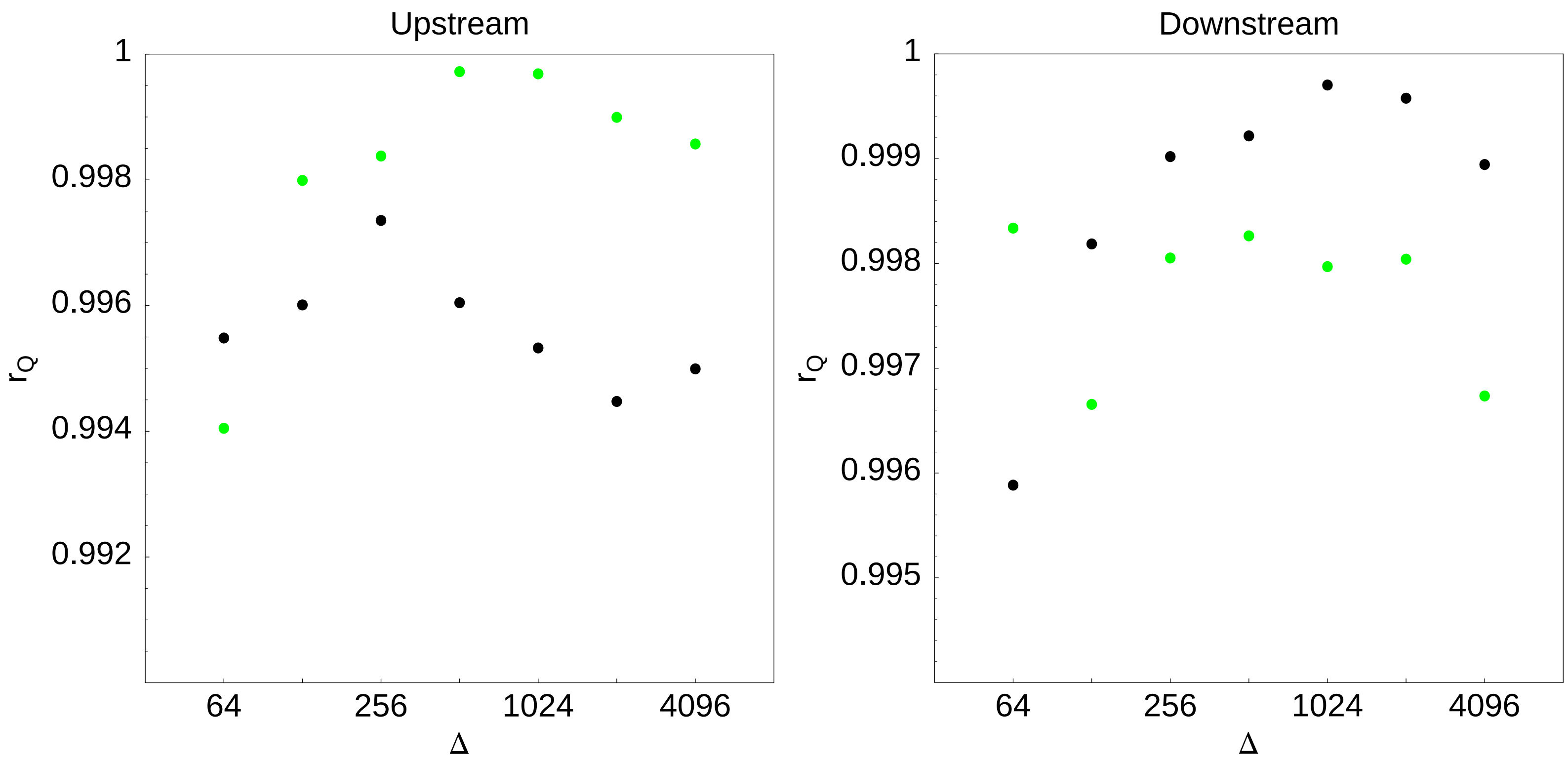




Examples of log-normal probability plots:



First results:



Desription of the methods:

- *Coefficient of skewness*

$$\frac{E(X-\mu)^3}{\sigma^3}$$

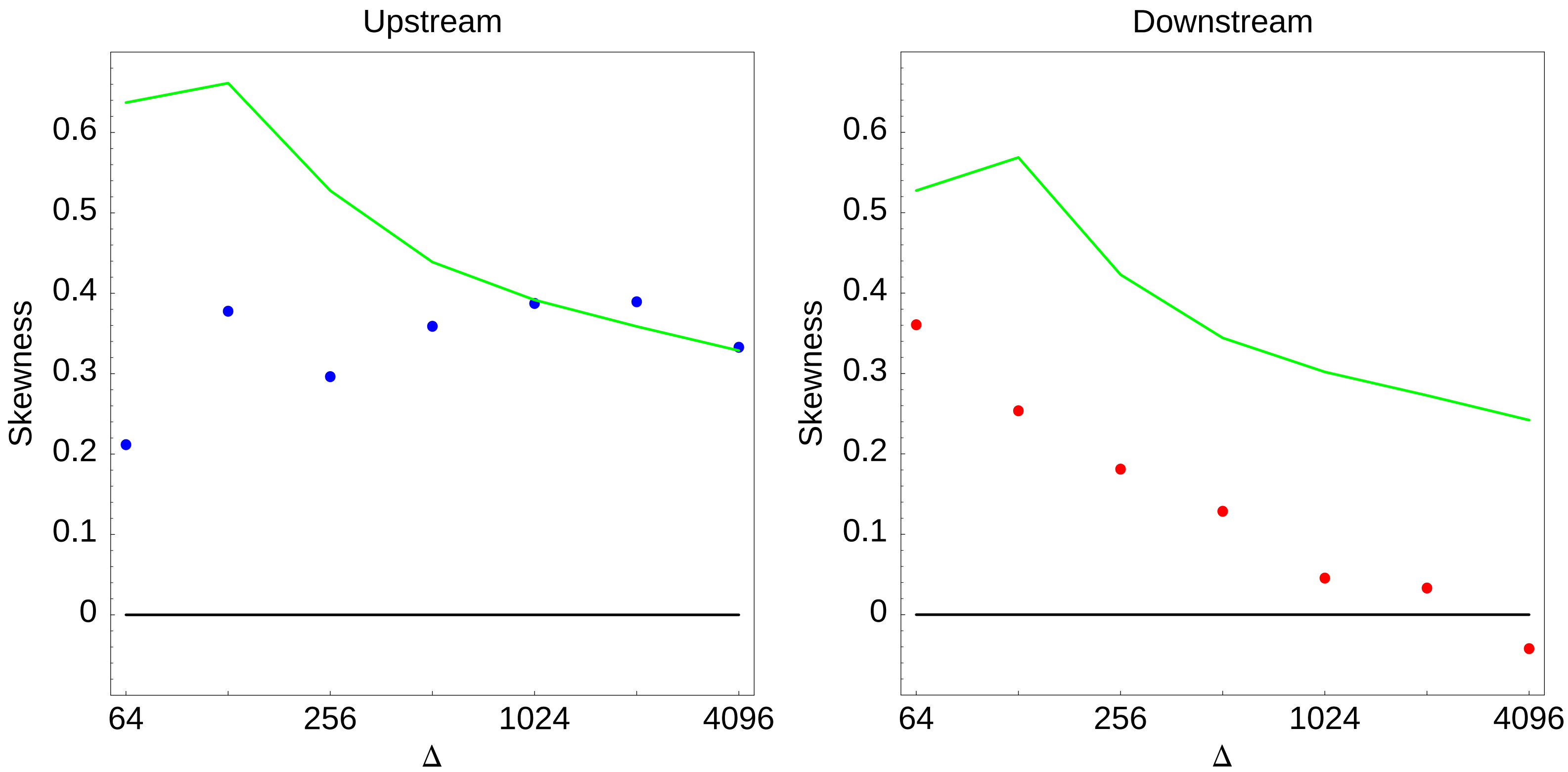
  measure the lack of symmetry.

- Skewness of normal distribution is always 0, skewness of log-normal distribution is

$$(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$$

- Other: Pearson's first coefficient of skewness, Pearson's second moment of skewness, Quartile coefficient of skewness.

Skewness:


Upstream
Downstream
Skewness
0.6
0.5
0.4
0.3
0.2
0.1
0
64
256
1024
4096
Δ

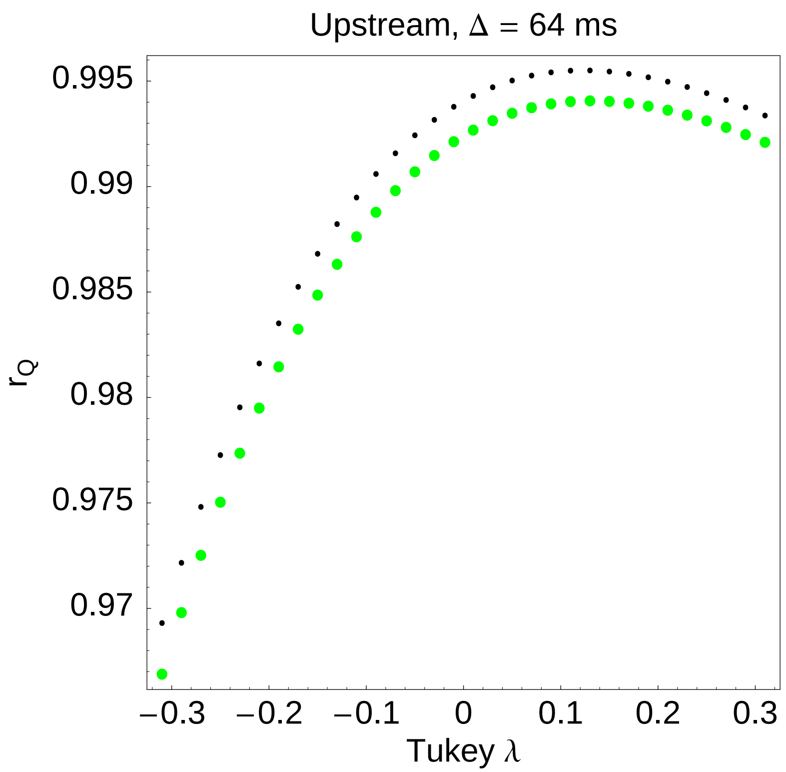
Description of the methods:

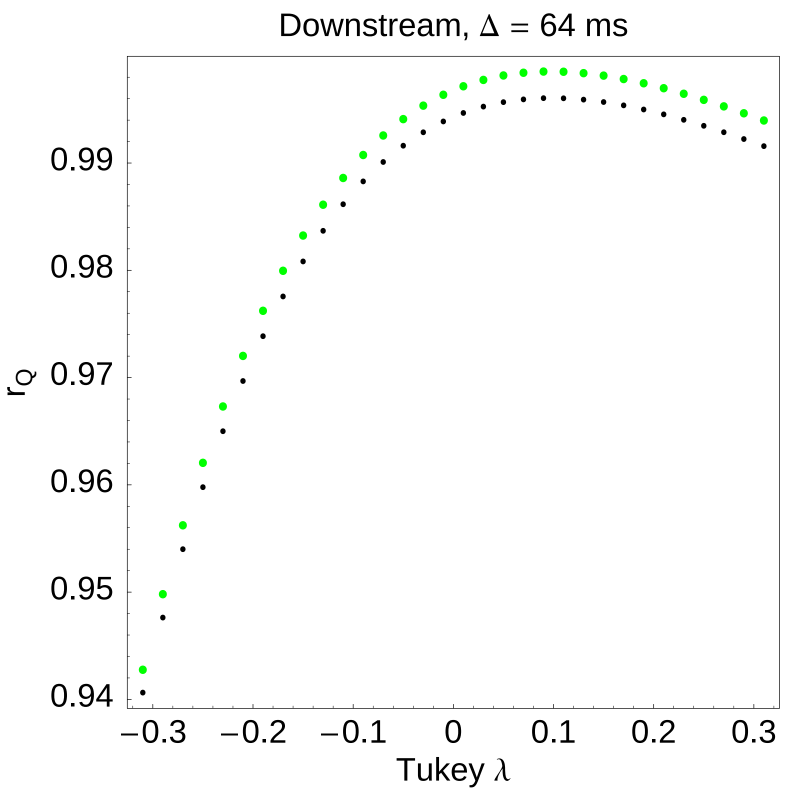
- *Probability plot correlation coefficient plot* (PPCC plot) is a plot of correlation coefficients r_Q against the corresponding shape parameters for given distributional family. The maximum value of r_Q corresponds to the optimal value of the shape parameter.

- Tukey λ-distribution:
  - Quantile function is $F^{-1}(p) = \frac{1}{\lambda}(p^\lambda - (1-p)^\lambda)$
  - The Tukey λ is a family of distributions that can approximate a number of common **symmetric** distributions, for example

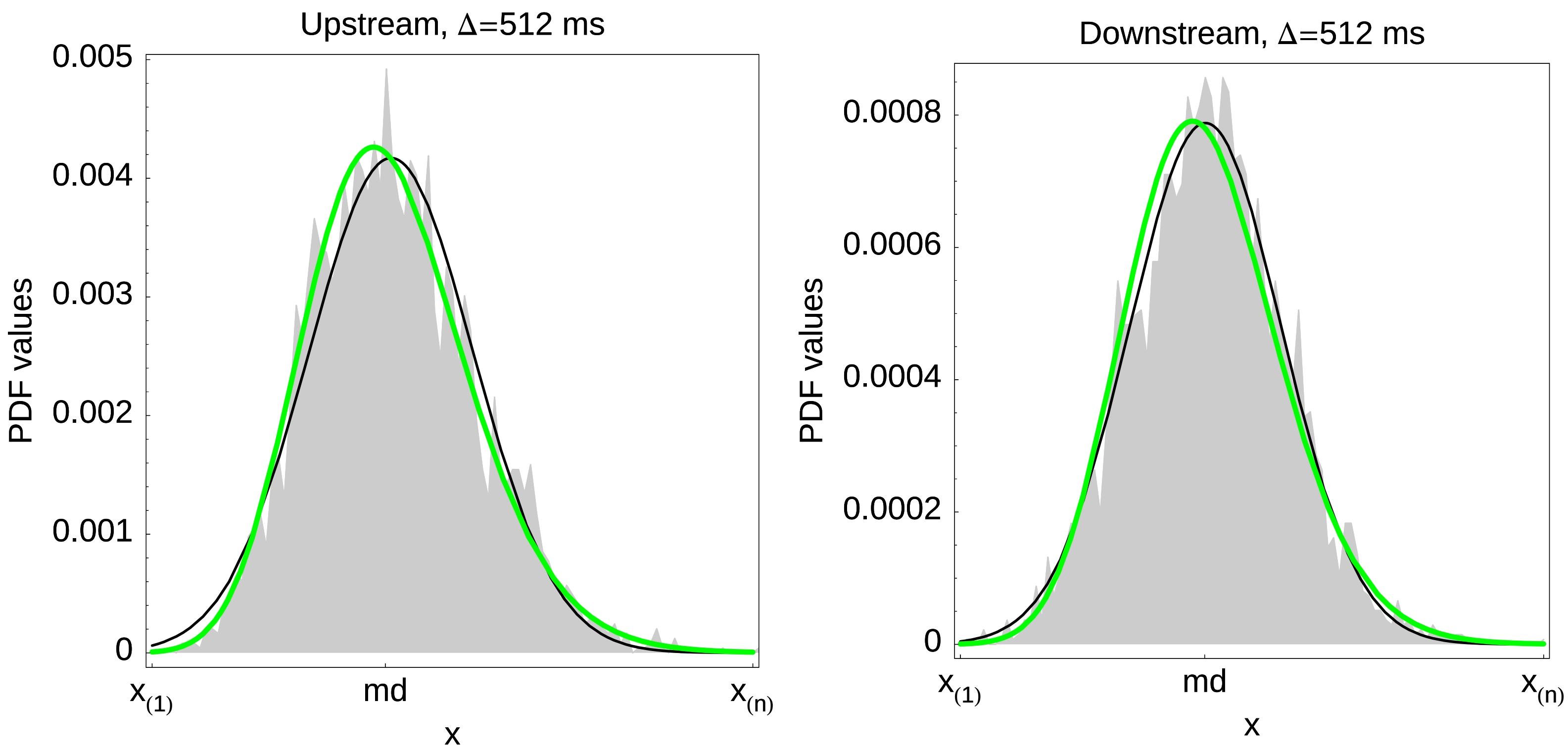
λ	Distribution
-1.00	approximately Cauchy
0.00	exactly logistic
0.14	approximately normal
0.50	U-shaped
1.00	exactly uniform

Tukey λ PPCC plot:

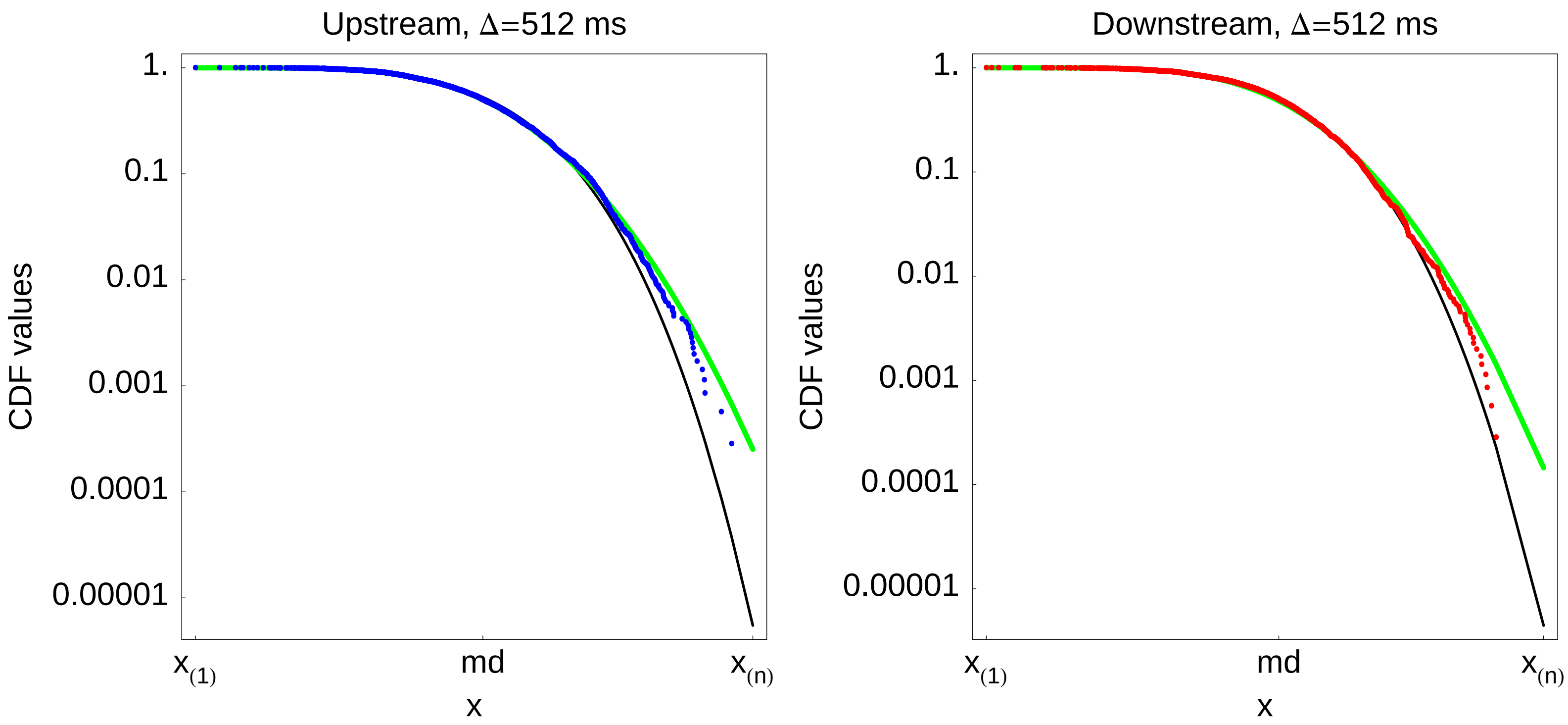




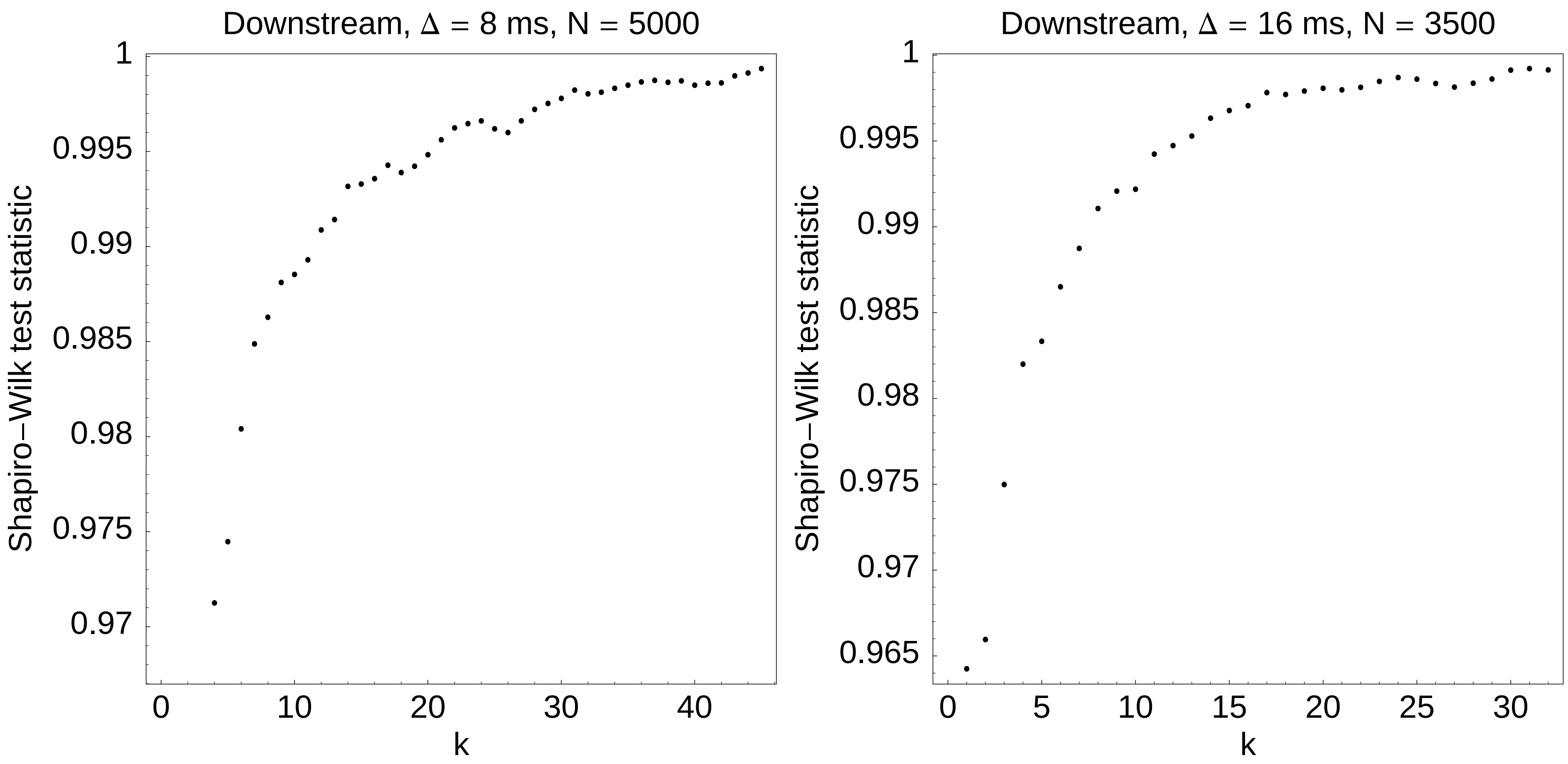
Empirical histogram *vs.* model pdf:



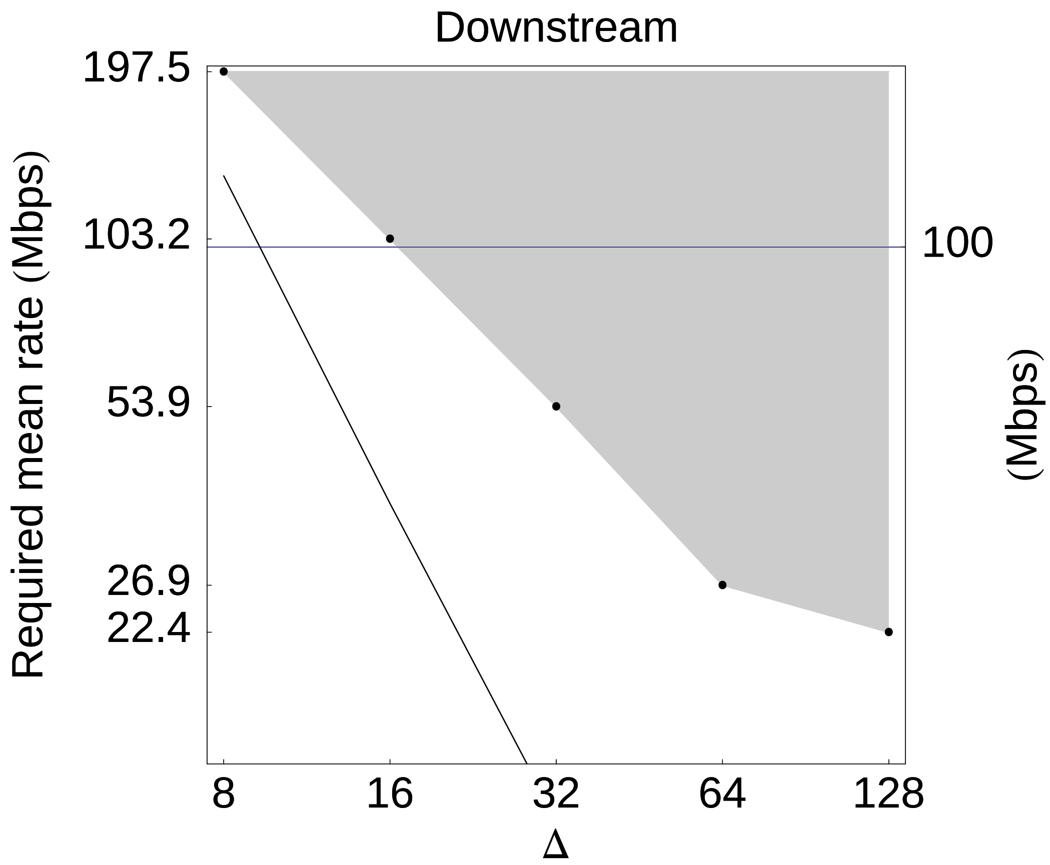
Empirical tail *vs.* model tails:



Overlapping increases the value of r_Q:



Vertical aggregation required for a good fit:



The problem of choosing a suitable significance level

- *Type I error*: normality hypothesis is rejected when it should be accepted
- *Type II error*: normality hypothesis is accepted when it should be rejected
- Which type of errors of decision we want to minimize?
- *P-value*: maximum probability to obtain a given sample given that normality (the null hypothesis) is true:

P-value	Standard conclusion
$\ldots > 0.05$	"not significant"
$0.05 \geq \ldots \geq 0.01$	"probably significant"
$0.01 > \ldots$	"highly significant"

- For example: How good fit is required for the approximate performance formulaes for queues with Gaussian input to hold?