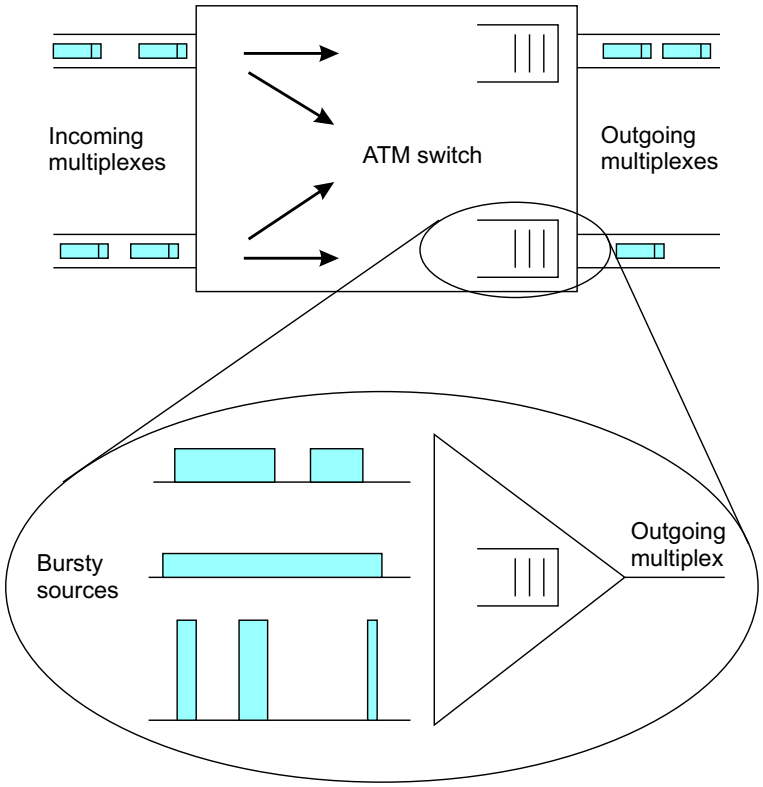
CELL LEVEL QUEUES

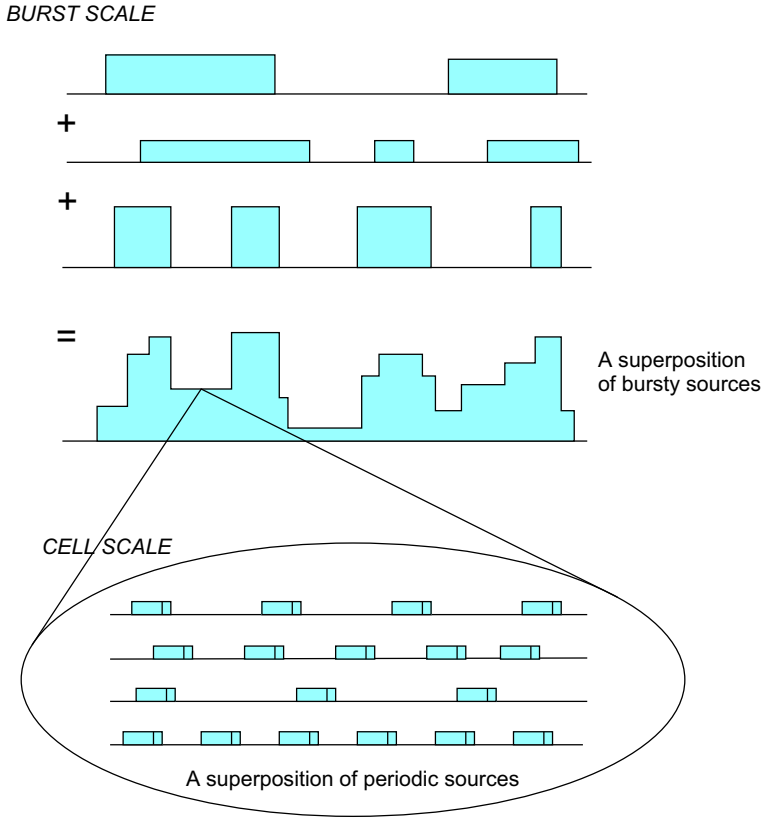
- At the cell level the time scale is short
  - traffic variations in a slower time scale (burst level) 'are not seen'
  - the burst composition can be considered fixed
  - each traffic source transmits cells approximately at (locally) constant intervals
  - different intervals for different sources; the transmission rate can also be zero when the source is silent (burst is not on)
- At cell level our interest is in the short time behaviour of the queues
  - queueing models $M/D/1$ and $N*D/D/1$
- The task is to determine the queue length distribution
  - to solve the problem we will use the so called Beneš method
  - the results can be used for dimensioning buffers in the switches
  - when the burst composition changes on a slower time scale the short time queue length distribution varies parametrically
- Another task at the cell level is to determine so called HOL (Head of Line) blocking

Cell level traffic[1]

ATM switch



Burst and cell time scales



[1]Figures adapted from J. Roberts' slides

The Beneš method for the $G/D/1$ system

- $G/D/1$ model is appropriate for ATM networks
  - a cell is of fixed length
  - the service time (transmission time to the line) is constant
  - the arrival process may be anything (general)
  - for instance $M/D/1$ or $N*D/D/1$
- For the Beneš method, we first derive so called Reich's result

Reich's result

Consider a single server system

$$\begin{cases} X_t & = \text{unfinished work (buffer content) in the queue at time } t \\ A(s,t) & = \text{work that arrives at the system in the interval } (s,t) \\ c & = \text{the service rate of the queue (draining rate, rate of discharge)} \end{cases}$$

$$\boxed{X_t = \sup_{s<t}(A(s,t) - c(t-s))}$$ Reich's result

- $A(s,t)$ is the work arrived in (s,t)
- $c(t-s)$ is the greatest amount of work that can be discharged in that interval

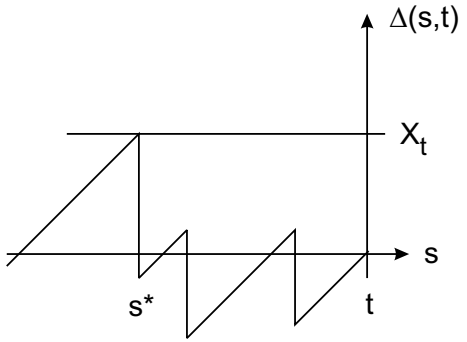
It is convenient to denote the difference of these two by

$$\boxed{\Delta(s,t) = A(s,t) - c(t-s)}$$ the excess work arrived in (s,t);
the amount of work that necessarily is backlocked

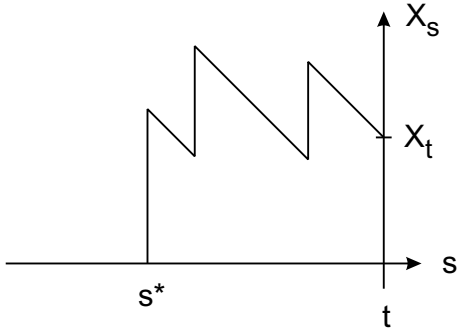
In terms of this quantity Reich's result reads

$$\boxed{X_t = \sup_{s<t} \Delta(s,t)}$$ Queue length at time t is the maximum of the excess work over all time windows preceding time t.

Reich's result (continued)



Excess work $\Delta(s,t)$ as a function of the initial time s



Queue length process X_s

Proof of Reich's result: It holds for all $s < t$ that

$$X_t \geq A(s,t) - c(t-s) = \Delta(s,t)$$

at time t the queue necessarily contains at least the excess work of the interval (s,t)

Denote by s^* the instant of time before t when the queue was empty for the last time (every stable queue is sometimes empty). Thus we have

- The server is busy throughout (s^*, t) and the amount of work discharged is $c(t-s^*)$.
- $X_t = \underbrace{X_{s^*}}_{0} + \underbrace{A(s^*,t)}_{\text{arrived work}} - \underbrace{c(t-s^*)}_{\text{discharged work}} = \Delta(s^*,t)$

Therefore, at s^* the half-inequality is satisfied as equality.
$\Delta(s,t)$ attains its maximum (as a function of s) at s^*: $X_t = \sup_{s<t} \Delta(s,t) = \Delta(s^*,t)$.

Reich's result (continued)

Corollary

An immediate corollary from the inequality

$$X_t \geq A(s,t) - c(t-s) = \Delta(s,t) \quad \forall s < t$$

is

$$\mathrm{P}\{X_t > x\} \geq \mathrm{P}\{\Delta(s,t) > x\} \quad \forall s < t$$

since whenever $\Delta(s,t) > x$ the inequality implies $X_t > x$ and consequently the probability of the latter event is greater than or equal to that of the first one.

The probability $\mathrm{P}\{X_t > x\}$ defines the tail distribution of the queue length (unfinished work). We have obtained a lower bound for this probability. As the bound applies for all $s < t$, then also the tightest lower bound is valid

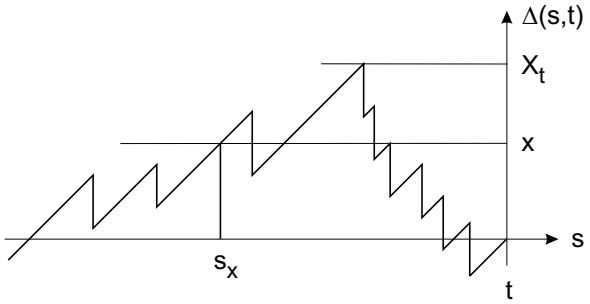
$$\mathrm{P}\{X_t > x\} \geq \sup_{s<t} \mathrm{P}\{\Delta(s,t) > x\}$$

This simple lower bound often gives useful information about the queue length distribution.

The Beneš method

We wish to determine the complementary distribution of the queue length X_t

$$Q(x) = \mathrm{P}\{X_t > x\}$$



- By Reich's result the condition $\{X_t > x\}$ implies that the maximum of $\Delta(s,t)$ (as a function of s) is greater than x.
- On the other hand, in a stable queue the arriving work cannot on the long run exceed the amount of work that can be discharged, $\lim_{s\to-\infty} \Delta(s,t) = -\infty$ (with probability 1).

Thus in each realization (path), for which $X_t > x$, there is a unique earliest time s_x, where the excess work curve crosses level x,

$$s_x = \inf\{s \le t : \Delta(s,t) = x\}$$

The event $\{X_t > x\}$ can now be partitioned according to where s_x is located. One subevent in the partitioning is $s_x \in (s, s+ds)$ or, briefly, $s_x \in ds$. The law of total probability gives

$$\boxed{\mathrm{P}\{X_t > x\} = \int_{s<t} \mathrm{P}\{s_x \in ds\}}$$

The Beneš method (continued)

The first passage time s_x of level x is characterized by two properties

a) $\Delta(s_x, t) = x$ (level x is crossed at time s_x)

b) $\Delta(s, t) < x, \quad \forall s < s_x$ (level x is not reached before s_x)

It is easy to see that the excess work is additive over consecutive intervals. Thus

$$\Delta(s, t) = \Delta(s, s_x) + \Delta(s_x, t) \qquad \text{when } s \le s_x \le t$$

From the above conditions it follows

$$\Delta(s, s_x) < 0, \quad \forall s < s_x$$

By Reich's result this means that the queue at time s_x is empty.
Property b) can therefore be replaced with the property

b') $X_{s_x} = 0$

The Beneš method (continued)

Assuming furthermore that $\Delta(s,t)$ is piecewise continuous function of s

- e.g. for discrete arrivals we have a piecewise continuous 'sawtooth curve'

we finally get the Beneš result

$$\boxed{\mathrm{P}\{X_t > x\} = \int_{s<t} \mathrm{P}\{\Delta(s,t) < x \le \Delta(s+ds,t),\ X_s = 0\}}$$

- The first condition says that level x is crossed in the interval $(s, s+ds)$.
- The second condition, $X_s = 0$, guarantees that the level x has not been reached before s.

In this form, the Beneš method is not yet an explicit solution for the problem but rather a reformulation of the problem.

- The probability still contains a condition related to the queue, $X_s = 0$.

The Beneš method, however, is very useful.

- The result is very general; no assumptions have been made about the arrival process.
- In some cases, it leads to an exact result in closed form.
- Often it can be used to derive useful approximations or bounds.

The Beneš method (continued)

The Beneš result can be written in many different ways.

- In particular, when the system is more specified, we get more concrete forms for the Beneš result.

Often we can without loss of generality denote the considered time by time 0 (instead of t). Time 0 can

- represent an arbitrary instant of time
- but it may also be a specifically selected point of time with respect to the arrival process; the Beneš result is valid even when the considered instant does not represent equilibrium

When the queue length is considered at time 0, it is convenient to use the notation $\Delta(s) = \Delta(-s, 0)$

- when the excess work function has only one argument s, it refers to the excess work arrived in a time window of length s preceding time 0
- s is now a positive number, and the initial time of the interval is $-s$

Sometimes, we also suppress the process index altogether, $X_0 = X$.

With these conventions, the Beneš result reads

$$\boxed{\mathrm{P}\{X_0 > x\} = \int_0^\infty \mathrm{P}\{\Delta(s) > x \geq \Delta(s + ds),\ X_{-s} = 0\} ds}$$

Upper bound for the queue length distribution

The 'difficulty' in the Beneš method is that the probability expression still contains a condition relating the queue length itself, $X_{-s} = 0$.

Since this is a restricting additional condition, then simply omitting the condition will increase (certainly not decrease) the probability.

Thus we obtain an upper bound for the queue length distribution

$$\boxed{\mathrm{P}\{X_0 > x\} \le \int_0^\infty \mathrm{P}\{\Delta(s) > x \ge \Delta(s+ds)\}ds}$$

This upper bound is explicit in the sense that as soon as the arrival process is specified, the probability in the integrand can, in principle, be determined and the integral calculated.

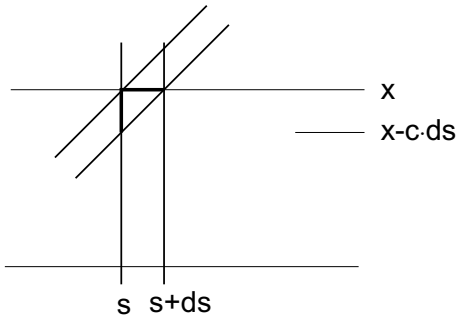
The Beneš result in case of discrete arrivals

The Beneš result given before applies for any arrival process, e.g. also for such a process where work arrives as a continuous fluid;

- fluid description is a good model in describing queueing phenomena at the burst level

In most queueing systems the arrival process is a point process;

- the arrivals occur at some discrete instants of time
- each arrival brings a finite amount of work to the queue



As the graph of the excess traffic is almost everywhere a straight line, having the slope c, then as shown in the figure, the condition of crossing level x in the interval $(s, s+ds)$ is equivalent to that the excess work $\Delta(s,0)$ at time s is in the interval $(x - c\cdot ds, x)$.

$$\begin{aligned}
\mathrm{P}\{X_t > x\} &= \int_{s<t} \mathrm{P}\{\Delta(s,t) \in (x - c\cdot ds, x), X_s = 0\} \\
&= \int_{s<t} \mathrm{P}\{\Delta(s,t) \in (x - c\cdot ds, x) \,|\, X_s = 0\}\mathrm{P}\{X_s = 0\} \\
&= \int_{s<t} \mathrm{P}\{A(s,t) \in (x + c(t-s)) - c\cdot ds, x + c(t-s) \,|\, X_s = 0\}\mathrm{P}\{X_s = 0\}
\end{aligned}$$

$$\boxed{\mathrm{P}\{X_0 > x\} = c\int_0^\infty f_{A(s)}(x + c\cdot s \,|\, X_{-s} = 0)\,\mathrm{P}\{X_{-s} = 0\}ds}$$

$f_{A(s)}(x \,|\, X_{-s} = 0)$ is the conditional pdf of the work arriving in $(-s, 0)$ given that $X_{-s} = 0$

The Beneš result for a $G/D/1$ queue

Now we restrict our system more.

- The arrivals are discrete, but the arrival process is still general (unspecified)
- The service time is assumed deterministic; each customer brings the same amount of work
  - e.g. an ATM cell; the model is well suited to describe an ATM network at cell level

We take the work brought by a customer (cell) as the unit of work and the constant service time as the unit of time (whence $c = 1$; the server accomplishes one unit of work in one unit of time). Then

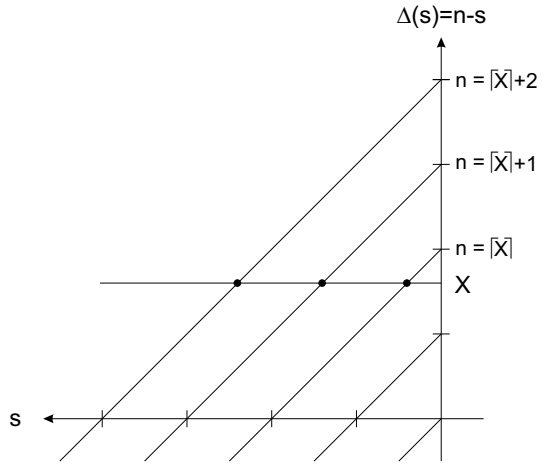
$$\begin{cases} A(s) &= \nu(s) \qquad \text{(number of arrivals in the interval } (-s, 0)) \\ \Delta(s) &= \nu(s) - s \end{cases}$$

The random variable $\nu(s)$ is an integer variable.

The level crossing condition $\Delta(s) = x$ can be satisfied only for such s which are of the form $n - x$, where n is an integer (of course, we must have $n > x$).

The possible lengths s_x of the intervals thus constitute a regular grid with spacing 1

$$s_x = n - x, \qquad n > x$$



The Beneš result for a $G/D/1$ queue (continued)

At a grid point $n - x$ the level crossing condition reads

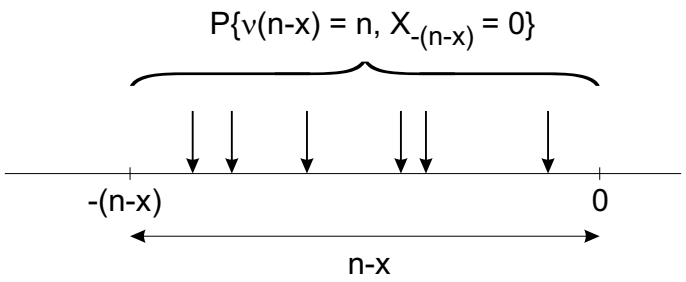
$$\nu(n-x) = n$$

in an interval of length $n - x$ there must be n arrivals;
the excess work in the interval is then $n - (n - x) = x$

The partitioning of the event $X_0 > x$ according to the fist passage time of level x reduces to a numerable set of subevents.

Correspondingly, the total probability integral reduces to a sum over the grid points:

$$\boxed{\mathrm{P}\{X_0 > x\} = \sum_{n>x} \mathrm{P}\{\nu(n-x) = n,\ X_{-(n-x)} = 0\}}$$

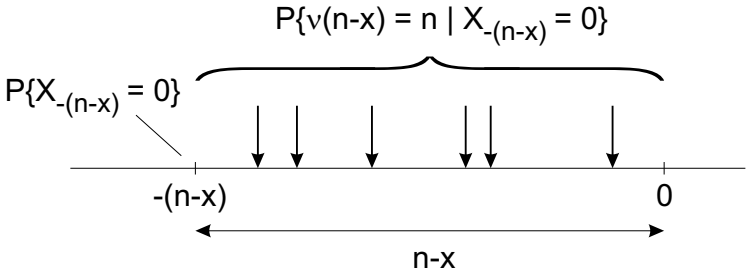


Sum over grid points of the form $n-x, (n > x)$
$\lceil x \rceil - x, \lceil x \rceil - x + 1, \lceil x \rceil - x + 2, \ldots,$
where $\lceil x \rceil$ is the smallest integer $\geq x$.

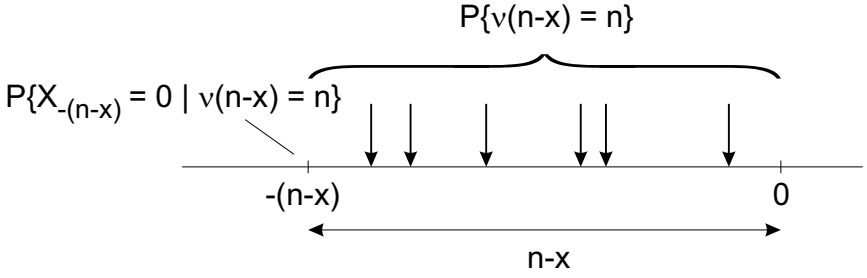
The Beneš result for a $G/D/1$ queue (continued)

By conditioning, the result can be written in two different forms:

$$\boxed{\mathrm{P}\{X_0 > x\} = \sum_{n>x} \mathrm{P}\{\nu(n-x) = n \,|\, X_{-(n-x)} = 0\}\mathrm{P}\{X_{-(n-x)} = 0\}}$$



$$\boxed{\mathrm{P}\{X_0 > x\} = \sum_{n>x} \mathrm{P}\{X_{-(n-x)} = 0 \,|\, \nu(n-x) = n\}\mathrm{P}\{\nu(n-x) = n\}}$$

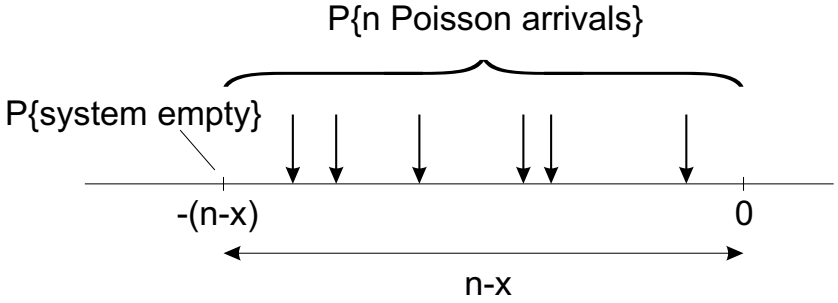


$M/D/1$ queue

- Arrivals occur according to a Poisson process with rate λ.
- The load $\rho = \lambda \cdot d$, where d $(= 1)$ is the service time.
- The number of arrivals $\nu(n-x)$ in the interval $n-x$ (service times) is Poisson distributed with mean $\rho(n-x)$,

$$\mathrm{P}\{\nu(n-x) = n\} = \frac{(\rho \cdot (n-x))^n}{n!}\, e^{-\rho\cdot(n-x)}$$

- At an arbitrary instant the queue is empty with the probability $1-\rho$.
- Poisson arrivals in the interval $(-s, 0)$ are independent of the arrivals before $-s$ and, consequently, of the queue length X_{-s}: the joint probability is broken down into a product.



$M/D/1$ queue (continued)

The complementary queue length distribution $Q(x) = \mathrm{P}\{X_0 > x\}$ is obtained as a special case of the formula for the $G/D/1$ system

$$\begin{aligned} Q(x) &= \sum_{n>x} \mathrm{P}\{\nu(n-x) = n,\ X_{-s} = 0\} \\ &= \sum_{n>x} \mathrm{P}\{\nu(n-x) = n\} \cdot \mathrm{P}\{X_{-s} = 0\} \\ &= \sum_{n>x} \frac{(\rho \cdot (n-x))^n}{n!}\, e^{-\rho\cdot(n-x)} \cdot (1-\rho) \end{aligned}$$

By using a known algebraic identity and by choosing $a = -\rho x$ and $b = \rho$

$$(1-b)\sum_0^\infty \frac{(a+n\cdot b)^n}{n!}\, e^{-(a+n\cdot b)} = 1 \quad \Rightarrow \quad (1-\rho)\sum_0^\infty \frac{(\rho\cdot(n-x))^n}{n!}\, e^{-\rho\cdot(n-x)} = 1$$

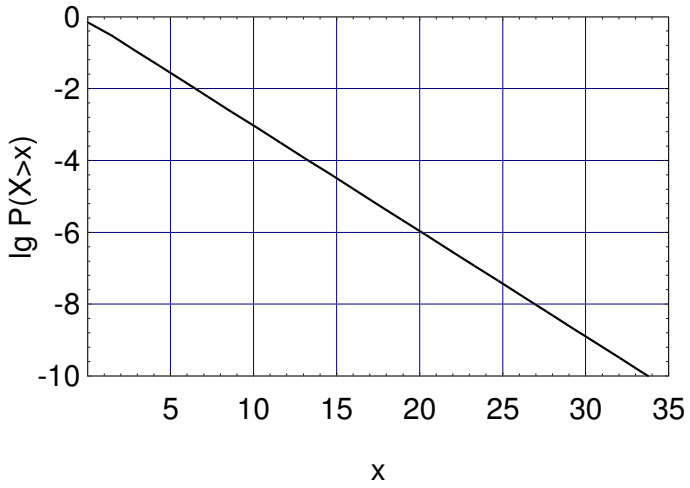
the infinite sum is transformed to a finite one (complement of one) and we arrive at

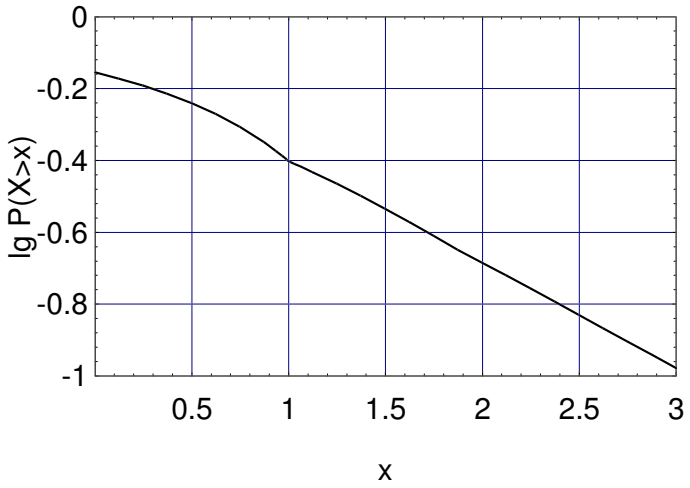
$$\boxed{Q(x) = 1 - (1-\rho)\sum_{n=0}^{\lfloor x\rfloor} \frac{(\rho\cdot(n-x))^n}{n!}\, e^{-\rho\cdot(n-x)}}$$

$M/D/1$ queue (continued)

The complementary distribution $Q(x) = \mathrm{P}\{X > x\}$ of the $M/D/1$ queue depends solely on a single parameter, viz. the load ρ.

- The distribution is very close to an exponential distribution (a straight line on logarithmic scale), as is seen from the graphs below in the case $\rho = 0.7$.





- Asymptotically (x large) the distribution indeed tends to exponential function $\mathrm{const} \cdot e^{s \cdot x}$
  - the deviation from an exponential behaviour can only be observed for small x
  - the factor of s in the exponent (negative) is the pole of the Laplace transform to be presented later, and it satisfies the following transcendental equation: $\rho(e^{-s} - 1) + s = 0$
  - for heavy load, when $\rho \approx 1$, we have approximately $s = -2(1 - \rho)/\rho$

Number of customers in an $M/D/1$ queue

We have obtained a formula for the distribution of unfinished work X in the system.

We would like to know also the distribution of the number of customers N in the system. This follows simply from the previous result by observing that

$N = \lceil X \rceil$

If for instance $X = 4.7$, then 4 customers wait and one is being served; thus there are $5 = \lceil 4.7 \rceil$ customers in the system

We see that for n integer it holds

$$\{N > n\} \equiv \{X > n\}$$

Thus we obtain

$$\boxed{\mathrm{P}\{N > n\} = \mathrm{P}\{X > n\} = Q(n)}$$

Tail distribution of the number in system is directly obtained from the tail distribution of the unfinished work at integer points.

$M/D/1$ queue according to the Pollaczek-Khinchinin transform formula

The Pollaczek-Khinchinin formula for the Laplace transform $X^*(s)$ of the unfinished work X (for its pdf) is

$$X^*(s) = \frac{1-\rho}{1-\rho\dfrac{1-S^*(s)}{s\cdot \mathrm{E}[S]}}$$

where S is the service time and $S^*(s)$ its Laplace transform (transform of the pdf)

In an $M/D/1$ system $S = \text{constant} = d\ (=1)$. Thus

$$\boxed{X^*(s) = \frac{1-\rho}{1-\rho\dfrac{1-e^{-s}}{s}}}$$

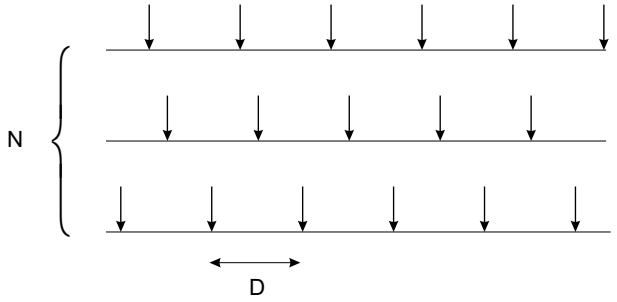
the expression derived before by the Beneš method can be derived also from this Pollaczek-Khinchinin formula by making the inverse transform.

The moments of the distribution can easily be derived from the Laplace transform

$$\boxed{\mathrm{E}[X] = \frac{\rho}{2\,(1-\rho)}} \qquad \boxed{\mathrm{V}[X] = \frac{\rho(1-(\frac{\rho}{2})^2)}{3\,(1-\rho)^2}}$$

$N*D/D/1$ queue

$N*D$ arrival process



- N independent sources
- Each source is periodic with period D
  - one arrival at intervals of D
  - random phases within one period D

Each realization of the arrival process (phases drawn in random) is periodic:

- N arrivals in each period of lengths D
- In different periods, the arrival patterns are exact copies of each other.

In the $N*D$ arrival process, the arrivals are over short time scale negatively correlated:

- If an arrival has occured, then within time D there cannot be another arrival from the same source.

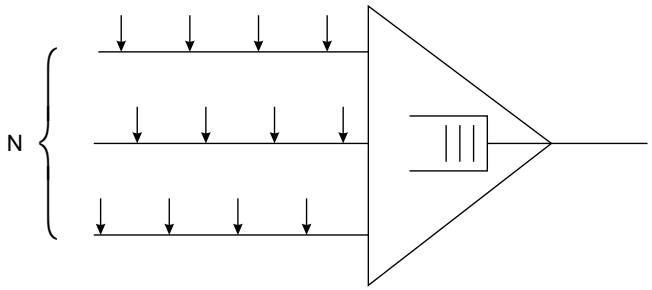
On longer time scales the arrivals are positively correlated:

- Given an arrival, then with certainty another arrival will occur after times D, $2D$ etc.

$N*D/D/1$ queue (continued)

When the service time is constant:

- Each source sends one unit of work in each period.
- We use the constant service time as the unit of time (then D is a pure (real) number).
- The load of the queue is $\rho = N/D$.



The $N*D/D/1$ queue is an appropriate model for e.g. the output buffer of an ATM switch, when the buffer is traversed by N virtual channel (VC) connections, each of them having approximately the same bit (cell) rate.

In speaking about the queue length distribution, we are not interested in the behaviour of the queue in a single realization, but on the ensemble distribution including all realizations with all possible phases of the sources.

It is convenient to look at a period starting at time 0 and consider the unfinished work, X_D, at the end of the period (which still is an arbitrary instant).

Our task is to derive the tail distribution $Q(x) = \mathrm{P}\{X_D > x\}$ of X_D.

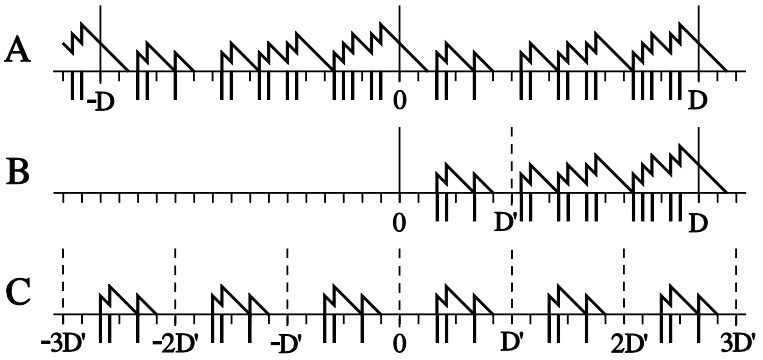
$N*D/D/1$ queue: modification of the system

The tail distribution $Q(x) = \mathrm{P}\{X_D > x\}$ of the queue length (unfinished work) at an arbitrary instant D can be solved exactly by means of the Beneš method.

To this end, we modify the system twice.

The modification is based on the following key observation:

- In a stable queue, $N < D$, the system is sometimes empty.
- By the periodicity the system is empty at some instant in each period, e.g. in the interval $(0, D)$.
- The queue at time D arises solely from the arrivals in $(0, D)$
- X_D is not changed if the arrivals before time 0 are 'switched off'



Instead of the system A we can calculate the distribution of X_D: in system B.
Now we apply the Beneš method to find the distribution of X_D in system B.

$N*D/D/1$ queue: queue length in system B

We apply the general Beneš result for the $G/D/1$ queue in the conditional form

$$Q(x) = \mathrm{P}\{X_D > x\} = \sum_{x<n\le N} \mathrm{P}\{\nu(D', D) = n\} \cdot \mathrm{P}\{X_{D'} = 0 \,|\, \nu(D', D) = n\}$$

where

- $\nu(D', D)$ = number of arrivals in (D', D)
- $n - (D - D') = x \quad$ or $\quad D' = D - n + x; \qquad$ excess work in (D, D') is x

In the sum, the grid points have been denoted by D'.

- Each of the grid points D' is identified by the number of arrivals in (D', D).
- The length of the interval $D - D' = n - x$ is fixed by the requirements that the excess work in this interval must be x.
- In order for the excess work to equal x, the number of arrivals n in the interval must be at least x (gives the lower limit of the sum).
- On the other hand, in system B there are only N arrivals before D. Thus the sum can be stopped at N (for larger values of n the first probability is 0).

$N*D/D/1$ queue: queue length in system B (continued)

Examine the probability factors in the sum

$$Q(x) = \sum_{x<n\le N} \mathrm{P}\{\nu(D', D) = n\} \cdot \mathrm{P}\{X_{D'} = 0 \,|\, \nu(D', D) = n\}$$

The probability that of the N arrivals uniformly distributed in a period of length D precisely n occur in the subinterval (D', D) with length $n - x$ is given by the binomial distribution:

$$\mathrm{P}\{\nu(D', D) = n\} = \binom{N}{n} \Big(\frac{n-x}{D}\Big)^n \Big(1 - \frac{n-x}{D}\Big)^{N-n}$$

$N*D/D/1$ queue: queue length in system B (continued)

In order to calculate the conditional probability $\mathrm{P}\{X_{D'} = 0 \,|\, \nu(D', D) = n\}$ we reverse the reasoning which led to the truncation of the arrival process:

- Given that in the interval (D', D) there are n arrivals, the rest $N - n$ arrivals must occur in the interval $(0, D')$.
- These $N - n$ arrivals are uniformly distributed in the interval $(0, D')$.
- The interval $(0, D')$ is underloaded, since
$$N - n = N - (x + (D - D')) = -(D - N) - x + D' \le D'$$
- The arrivals in $(0, D')$ can be extended to a periodic arrival process (with period D')
  - in an underloaded (stable) periodic system the queue is empty within each period, therefore, also somewhere in $(0, D')$
  - the periodic extension does not change the queue $X_{D'}$
- The empty queue probability of system B at D' is the same as that for system C.
- The load of system C is $\rho' = (N - n)/(D - n + x)$.
- The empty queue probability of system C at time D' is the same as at an arbitrary instant
$$\mathrm{P}\{X_{D'} = 0 \,|\, \nu(D', D) = n\} = 1 - \frac{N - n}{D - n + x} = \frac{D - N + x}{D - n + x}$$

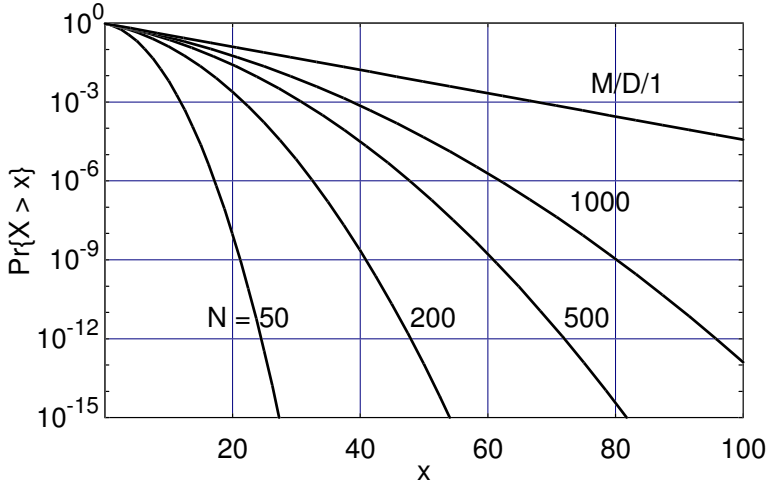
$N*D/D/1$ queue: queue length distribution

Collecting the results together we have the final result

$$Q_D^N(x) = \sum_{x<n\le N} \binom{N}{n} \Big(\frac{n-x}{D}\Big)^n \Big(1-\frac{n-x}{D}\Big)^{N-n} \frac{D-N+x}{D-n+x}$$

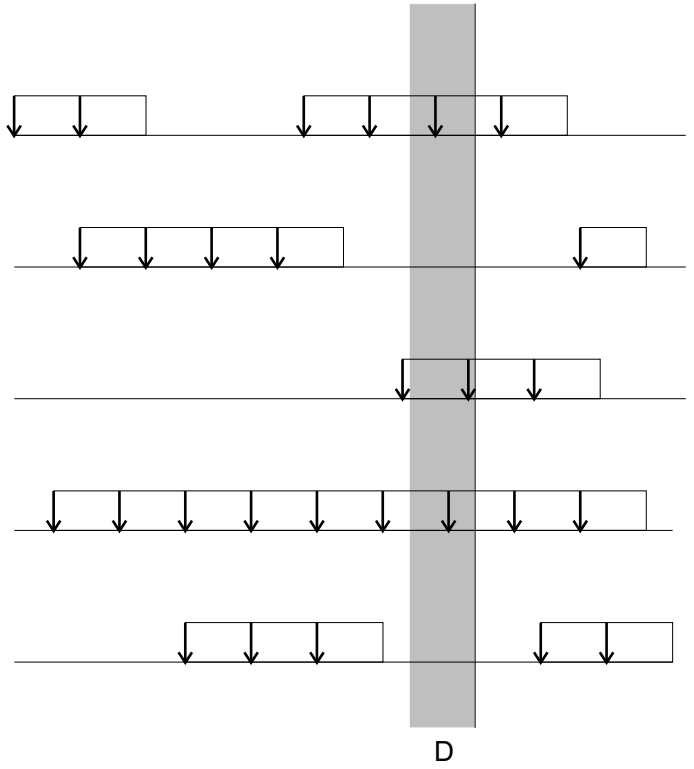
where the dependence of the queue length distribution $Q(x)$ on parameters N and D has been made explicit.

- We have obtained the distribution of the virtual waiting time.
- For integer x it gives directly the complementary distribution of the number of cells in the queues.
- The distribution of the *real* waiting time of a cell in an $N*D/D/1$ queue is the same as the virtual waiting time distribution in an $(N-1)*D/D/1$ queue, where the number of sources is one less, i.e. the distribution is $Q_D^{N-1}(x)$.



Queue length distribution for different values of N with a constant load $\rho = N/D = 0.95$.

Modulated $N*D/D/1$ queue (case $D \geq N$)



Again we have N sources.

Each source is either on or off as determined by some modulating process. The modulating process can be different for each source.

When a source is on, it sends cells periodically at constant intervals of D.

An uninterrupted on period of source is called a burst. Here we use the following interpretation. A burst

- begins with an arrival of a cell
- continues time D past the arrival of the last cell in the burst
- then the interarrival time between cells is $\geq D$

In the sequel, it suffices to know the on probability p_i of the modulating process of source i.

The restriction $D \geq N$ implies that, even if all the sources were continuously in the on state, the queue would still be stable. Thus we exclude the possibility of the so called burst scale queues in our modulated system.

Modulated $N*D/D/1$ queue (continued)

- Consider the queue at time D (arbitrary).
- The system is certainly empty at some instant within $(0, D)$.
  - this holds without modulation
  - the modulation 'rarefies' the arrivals
- The arrivals before time 0 can be discarded.
- Apply now the Beneš result to the truncated system.
- Condition the calculation on the number of sources that have the burst on at time D
  - from each source that is on, and only from these, there is one arrival in the interval $(0, D)$
  - uniformly distributed in the said interval

$$\mathrm{P}\{X > x\} = \sum_{n=0}^{N} \mathrm{P}\{n \text{ sources on}\} \cdot \sum_{x<k\le n} \mathrm{P}\{\nu(D', D) = k \text{ and } X_{D'} = 0 \,|\, n \text{ sources on}\}$$

- The latter sum equals $Q_D^n(x)$
  - the tail distribution of the $N*D/D/1$ queue: n sources, period D

Modulated $N*D/D/1$ queue (continued)

- Denote $P_n = \mathrm{P}\{n \text{ sources on}\}$
  - convolution of Bernoulli(p_i) distributions
- In terms of these, the modulated arrival process leads to a 'modulated queue length distribution'

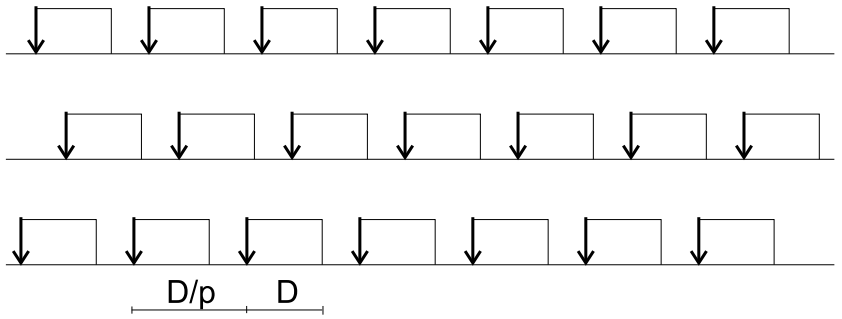
$$\boxed{\mathrm{P}\{X > x\} = \sum_{n=0}^{N} P_n Q_D^n(x)}$$

- The result depends only on the on-probabilities p_i (P_n depends on these)
  - not on other properties of the modulating process
- The result is exact
  - does not require large difference in time scales or quasi stationarity
- However, for the validity of the result the assumption $D \geq N$ (no burst scale congestion) is essential (and restrictive).

Modulated $N*D/D/1$ queue (continued)

- Assume that $p_i = p$ for all i
- P_n has binomial distribution: $P_n = \binom{N}{n} p^n (1-p)^{N-n}$
- The sum $\sum_{n=0}^{N} P_n Q_D^n(x)$ can be calculated;
  - leads to a simple result which can also be directly inferred as follows

- One can freely choose any modulating process
  - as long as the on probability is p
- One realization is the periodic process with period D/p



- Interpretation:
  - duration of the burst D
  - period D/p
  - on probability $D/(D/p) = p$

$$\boxed{\mathrm{P}\{X > x\} = Q_{D/p}^N(x)}$$

Modulated $N*D/D/1$ queue (continued)

- Through this reasoning we have, in fact, proven the mathematical identity

$$\sum_{n=0}^{N} \binom{N}{n} p^n (1-p)^{N-n} Q_D^n(x) = Q_{D/p}^N(x)$$

where

$$Q_D^n(x) = \sum_{x<k\le n} \binom{n}{k} \left(\frac{k-x}{D}\right)^k \left(1-\frac{k-x}{D}\right)^{n-k} \frac{D-n+x}{D-k+x}$$