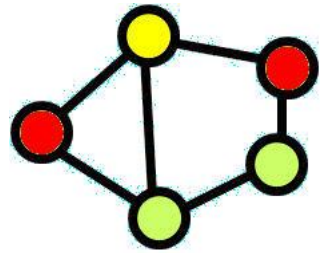
Slide supporting material



Lesson 7: M/G/1 Queuing Systems Analysis

Giovanni Giambene

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

Motivations for the Use of the M/G/1 Theory

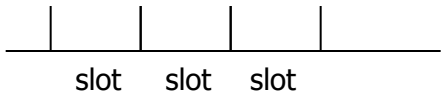
- **The assumption of Poisson arrivals may be reasonable** since the Poisson model is a limiting condition of the binomial distribution.
  - Many potential customers decide independently about arriving.
  - Each of them has a small probability of arriving in any particular time interval.
    - Probability of one arrival in a small interval is approximately proportional to the length of the interval itself.
- **The exponential distribution for the service time is no longer a good approximation in current packet-switched networks:** layer 2 packets may have a fixed length; files may have a length better modeled by a heavy-tailed distribution, e.g., Pareto distribution. Then, a general service time has to be considered.
- M/G/1 theory can be used for modeling different aspects of the networks.

M/G/1 Queues

- In the M/G/1 theory, the arrival process is Poisson with mean arrival rate λ, but, the service time is not exponentially distributed.
  - The service process has some memory: if there is a request in service at a given instant, the residual service time of the request has a distribution that depends on the **elapsed service time**.
  - A similar theoretical method to that of M/G/1 queues can be applied to solve G/M/1 ones.

Imbedded Markov Chains

- **2-D system state** for M/G/1 queues: $S(t) = \{n(t) , \tau(t)\}$.
  - $n(t)$: Number of requests in the system at instant t;
  - $\tau(t)$: Elapsed time from the beginning of the service of the currently-served request.
- To simplify the study, the M/G/1 queue is analyzed at **imbedding instants** ζ_i, this is as if we take snapshots of the system state at instants ζ_i when we obtain a mono-dimensional Markovian system (**imbedded Markov chain**), as detailed below.
- **Different alternatives are available to select imbedding instants ζ_i** (especially #1 and #3 below for M/G/1 cases):
  1. Service completion instants;
  2. Customer arrival instants (used in the G/M/1 case for the study of the waiting part);
  3. Regularly-spaced instants, for special cases with time-slotted service as TDM systems (e.g., ATM):



Imbedding to Service Completion Instants

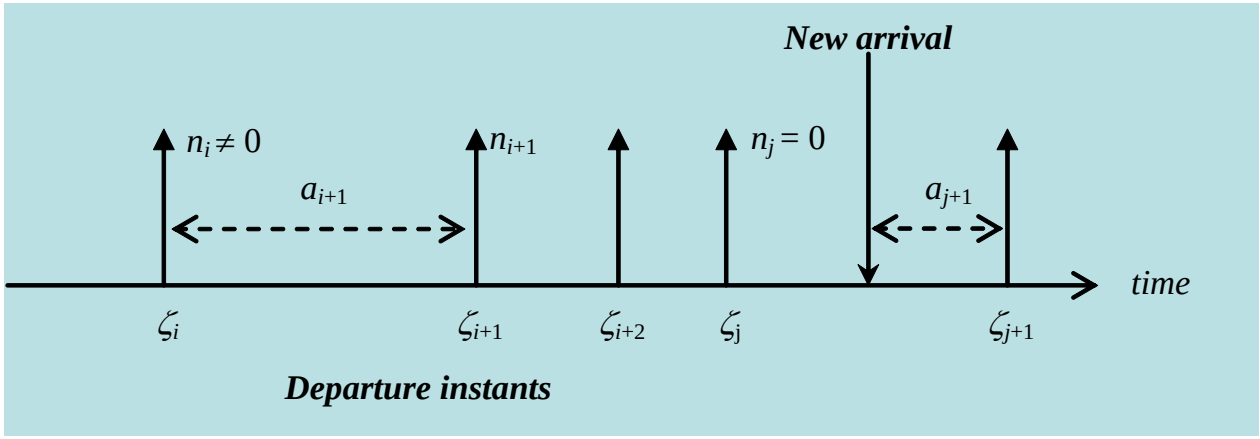
- Imbedding at service completion instants: $\tau(\zeta_i) \equiv 0$, $\forall i$ since at instant ζ_i a request has completed its service and no new request has yet started its service.
  - n_i denotes the number of requests in the queue **soon after** the service completion of the i-th request (instant ζ_i^+).
  - a_i denotes the number of requests arrived at the queue **during** the service time of the i-th request (ending at instant ζ_i^-).
  - At instants ζ_i, the state becomes **mono-dimensional**:

$$S(\zeta_i) \equiv n(\zeta_i) = n_i$$

Imbedding to Service Completion Instants (cont'd)

- **If $n_i \neq 0$**, at the subsequent instant of service completion the following balance is valid: $\mathbf{n_{i+1} = n_i - 1 + a_{i+1}}$.
  - Note that **among all requests in the queue, we do not pose special conditions on the request that has been served.**

- **If $n_i = 0$**, we have to wait for the next arrival that is immediately served, so that at the next completion instant ζ_{i+1}^+ the system just contains the arrivals occurred during the service time of the last request; we have: $\mathbf{n_{i+1} = a_{i+1}}$.

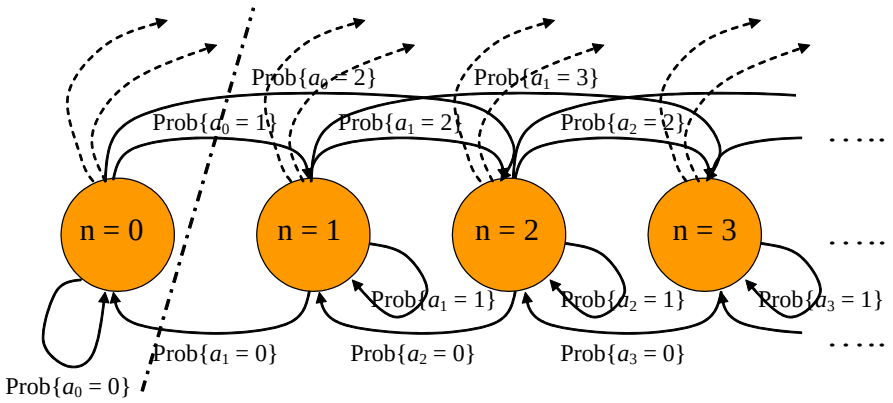


System Description

$$n_{i+1} = n_i - I(n_i) + a_{i+1}$$

where I(x) = 1, x > 0; I(x) = 0, x = 0 (Heaviside function).

- The above **difference equation** describes the behavior of the M/G/1 queue at imbedding instants.
- Since the variables at the instant ζ_{i+1} only depend on the variables at instant ζ_i, **the M/G/1 system is characterized by a discrete-time Markov chain at imbedding instants ('semi-Markov chain')**, as shown below.



In general, the arrival process is state-dependent.

The definitions/characteristics of both n_i and a_i depend on the selection of imbedding instants.

In general, the solution of the discrete-time Markov chain (i.e., determining the state probability distribution) requires a **matrix-geometric approach** or writing **cut equilibriums and an iterative solution approach**.

We will use an **approach in the z domain** by adding some assumptions.

Solution in the z-Domain with Additional Assumptions

- Let us assume that the **M/G/1 queue admits a steady state**.
  - P_n denotes the probability (at regime) to have n requests in the queue
- We focus on the difference equation that is solved in the z-domain (i.e., PGF) and we use the following **assumptions**:
  - Memoryless arrival process (a_i is memoryless: i.e., a_i independent of a_{i-1}, independent of a_{i-2}, etc.). This is a more general condition than a Poisson process: we use the '**M'/G/1 notation, where 'M' stands for a general memoryless arrival process** (e.g., a Bernoulli arrival process of packets on a slot basis).
  - Arrival process independent of the number of requests in the queue (n_i and a_i are independent). This assumption is not needed using the cut equilibrium or matrix-geometric approach.

$$\sum_h z^{n_{i+1}} P_{n_{i+1}} = \sum_k z^{n_i - I(n_i)} P_{n_i} \sum_j z^{a_{i+1}} P_{a_{i+1}} \quad \Rightarrow \quad P(z)[z - A(z)] = P_0(z-1)A(z) \quad (*)$$

where P(z) is the PGF of the state probability distribution, n_i, and A(z) is the PGF of the number of arrivals in the service time of a request, a_i.

Solution in the z-Domain with Additional Assumptions

- Let us assume that the **M/G/1 queue admits a steady state**.
  - P_n denotes the probability (at regime) to have n requests in the queue

On both sides we take triple sum on n_{i+1},n_i, a_{i+1} by using the joint probability P(n_{i+1},n_i, a_{i+1}). The result shown here is obtained after manipulations based on independence assumptions and marginal distributions.

is solved in the z-domain
mptions:

ess: i.e., a_i independent of a_{i-1},
eneral condition than a Poisson
where 'M' stands for a
e.g., a Bernoulli arrival process

of requests in the queue (n_i
s not needed using the cut
or matrix-geometric approach.

$$\sum_h z^{n_{i+1}} P_{n_{i+1}} = \sum_k z^{n_i - I(n_i)} P_{n_i} \sum_j z^{a_{i+1}} P_{a_{i+1}} \quad \Rightarrow \quad P(z)[z - A(z)] = P_0(z-1)A(z) \quad (*)$$

where P(z) is the PGF of the state probability distribution, n_i, and A(z) is the PGF of the number of arrivals in the service time of a request, a_i.

Solution in the z-Domain with Additional Assumptions

- Let us assume that the **M/G/1 queue admits a steady state**.
  - P_n denotes the probability (at regime) to have n requests in the queue
- We focus on the difference equat… (i.e., PGF) and we use the followi…
  - Memoryless arrival process (a_i is … independent of a_{i-2}, etc.). This is … process: we use the '**M'/G/1 no… <u>general memoryless</u> arrival p…** of packets on a slot basis).
  - Arrival process independent of th… and a_i are independent). This ass… equilibrium or matrix-geometric approach.

To obtain this result we do not pose special conditions on the service discipline apart the **conditions for the applicability of the insensitivity property**.

$$\sum_h z^{n_{i+1}} P_{n_{i+1}} = \sum_k z^{n_i - I(n_i)} P_{n_i} \sum_j z^{a_{i+1}} P_{a_{i+1}} \quad \Rightarrow \quad P(z)[z - A(z)] = P_0(z-1)A(z) \quad (*)$$

where P(z) is the PGF of the state probability distribution, n_i, and A(z) is the PGF of the number of arrivals in the service time of a request, a_i.

Solution in the z-Domain with Additional Assumptions

- Let us assume that the **M/G/1 queue admits a steady state**.
  - P_n denotes the probability (at regime) to have n requests in the queue
- We focus on the difference equation th… (i.e., PGF) and we use the following **as…**
  - Memoryless arrival process (a_i is memor… independent of a_{i-2}, etc.). This is a more… process: we use the '**M'/G/1 notation**… **general memoryless arrival process**… of packets on a slot basis).
  - Arrival process independent of the num… and a_i are independent). This assumptio… equilibrium or matrix-geometric approach.

Subscripts are here omitted because we assume to study the probability distribution at regime, that is for $i \to \infty$.

$$\sum_h z^{n_{i+1}} P_{n_{i+1}} = \sum_k z^{n_i - I(n_i)} P_{n_i} \sum_j z^{a_{i+1}} P_{a_{i+1}} \quad \Rightarrow \quad P(z)[z - A(z)] = P_0(z-1)A(z) \quad (*)$$

where P(z) is the PGF of the state probability distribution, n_i, and A(z) is the PGF of the number of arrivals in the service time of a request, a_i.

Solution in the z-Domain (cont'd)

- We can derive P(z) as:

$$P(z) = P_0 \frac{z-1}{z - A(z)} A(z)$$

- In this P(z) formula we have an **apparent singularity at z = 1**, but we can apply the **Abel theorem** to state that it exists the lim of P(z) for z → 1^- -pole-zero cancellation- and should be necessarily equal to 1 for the **normalization condition**. Therefore, we can solve this limit by means of the Hôpital rule:

$$\lim_{z \to 1^-} P_0 \frac{z-1}{z - A(z)} A(z) = 1 \Leftrightarrow P_0 \times \lim_{z \to 1^-} \frac{1}{1 - A'(z)} = 1 \Leftrightarrow P_0 = 1 - A'(1)$$

Abel theorem + normalization Hôpital rule

Solution in the z-Domain (cont'd)

- Deriving with respect to z both sides of the z-equation (*) and computing the result at z = 1, at the different orders of the derivative we obtain first the **empty queue probability** P_0 and then the **mean number of requests** in the queue N:
  - First derivative: $P_0 = 1 - A'(1)$ (normalization condition);
  - Second derivative: $$N = P'(1) = A'(1) + \frac{A''(1)}{2[1 - A'(1)]}$$
- **The PGF of the state probability distribution P(z) only depends on the PGF A(z) that, in turn, depends on the characteristics of the arrival process, the imbedding instants, and the distribution of the service time.**
  - These results are **insensitive to the service discipline** adopted for the queue.
  - This solution is for a generalized queue (not only Poisson arrivals).
  - **Stability condition** is $P_0 > 0 \Leftrightarrow$ **A'(1) < 1 Erl**; **A'(1) is the traffic intensity**.

Solution of the M/G/1 Queue for Poisson Arrivals

- **Assumptions: Poisson arrival process and system imbedded at the service completion instants.**
- A(z) can be computed considering the PGF of the number of arrivals in a given interval t, $A(z \mid t) = e^{\lambda t(z-1)}$ and then removing the conditioning by means of the probability density function of the service time, g(t) [with corresponding Laplace transform $\Gamma(s)$]:

$$A(z) = \int_0^{+\infty} e^{\lambda t(z-1)} g(t) dt = \Gamma\left(s = -\lambda(z-1)\right)$$

s to z domain transform:
$s = -\lambda(z-1)$

or equivalently

$$A\left(z = 1 - \frac{s}{\lambda}\right) = \Gamma(s)$$

z to s domain inverse transform:
$z = 1 - s/\lambda$

Solution of the M/G/1 Queue for Poisson Arrivals (cont'd)

- We obtain: **$A'(1)=\lambda E[X]$ = traffic intensity** ρ and $A''(1)=\lambda^2 E[X^2]$.

- Then, we can determine the mean number of requests in the system N as:

$$N = A'(1) + \frac{A''(1)}{2[1-A'(1)]} = \lambda E[X] + \frac{\lambda^2 E[X^2]}{2(1-\lambda E[X])}$$

Queuing term

- Then, the mean delay T is obtained dividing N by λ according to the Little theorem:

$$T = \frac{N}{\lambda} = E[X] + \frac{\lambda E[X^2]}{2[1-\lambda E[X]]}$$

Service part

Pollaczek-Khinchin formula

M/D/1 Queue

- In this system, arrivals are according to a Poisson process with mean rate λ and have a **fixed, constant service time**, x. This is for instance the case of the transmission of packets of a given size on a link with constant capacity.
- **Imbedding points** are at the end of the service of a request.
- We can directly apply the **Pollaczek-Khinchin formula** to determine the mean delay as:

$$T = x + \frac{\lambda x^2}{2[1-\lambda x]}$$

- For completeness, we have also $A(z) = e^{\lambda x(z-1)}$ and

$$P(z) = (1-\lambda x)\frac{(z-1)e^{\lambda x(z-1)}}{z - e^{\lambda x(z-1)}}$$

$M^{[L(z)]}/D/1$ Queue

- This is a case with a **bulk (or compound) Poisson arrival process** with PGF of the message length L(z) in packets. The lengths of messages are iid.
- Each packet transmission time is here denoted by T.
- We are interested in determining the PGF of the number of packets in the buffer, P(z), and the mean **packet delay.**
- **We imbed the system at the end of a packet transmission.** We can apply the M/G/1 theory with some approximation. **We derive A(z)**, the PGF of the number of packets arrived in the service time of a packet:

$$A(z \mid n) = L^n(z)$$

$$A(z) = \sum_n L^n(z) \frac{(\lambda T)^n}{n!} e^{-\lambda T} = e^{\lambda T(L(z)-1)}$$

$$A'(1) = \lambda T L'(1)$$

$$A''(1) = \left[\lambda T L'(1)\right]^2 + \lambda T L''(1)$$

- We can write the classical M/G/1 difference equation with some **approximation in the case $n_i = 0$**. The mean number of packets in the system N_p and the mean delay for the transmission of a packet T_p are:

$$N_p = A'(1) + \frac{A''(1)}{2\left[1 - A'(1)\right]} \qquad T_p = \frac{N_p}{\lambda L'(1)} [\mathrm{s}]$$

The Little theorem is here applied to a compound process

$M^{[L(z)]}/D/1$ Queue

The classical M/G/1 difference equation can be used as a first **approximation**: we consider $n_{i+1} \approx a_{i+1}$ for $n_i = 0$ (i.e., we neglect the existence of the packets after the first one in a message arriving at an empty buffer). We can remove this approximation by using the M/G/1 theory with 'different service times', as shown in Lesson No. 9.

arrival process with
ns of messages are iid.

ber of packets in the

nsmission. We can
e derive A(z), the PGF
f a packet:

$$A(z) = \sum_n L^n(z) \frac{(\lambda T)^n}{n!} e^{-\lambda T} = e^{\lambda T(L(z)-1)}$$

$$A'(1) = \lambda T L'(1)$$

$$A''(1) = \left[\lambda T L'(1)\right]^2 + \lambda T L''(1)$$

- We can write the classical M/G/1 difference equation with some **approximation in the case $n_i = 0$**. The mean number of packets in the system N_p and the mean delay for the transmission of a packet T_p are:

$$N_p = A'(1) + \frac{A''(1)}{2\left[1 - A'(1)\right]} \qquad T_p = \frac{N_p}{\lambda L'(1)} [\text{s}]$$

The Little theorem is here applied to a compound process

$M^{[L(z)]}/D/1$ Queue

- This is a case with a
  PGF of the message
- Each packet transmi
- We are interested in
  buffer, P(z), and the
- **We imbed the sys**
  to apply the M/G/1 t
  number of packets a

$$A(z \mid n) = L^n(z)$$

$$A(z) = \sum_n L^n(z) \frac{(\lambda T)^n}{n!} e^{-\lambda}$$

- We can write the cla
  approximation in
  system N_p and the n

$$N_p = A$$

$$\frac{\lambda^2 L^{(2)}(1)}{2[1-A'(1)]} \qquad \lambda L'(1)$$

applied to a compound process

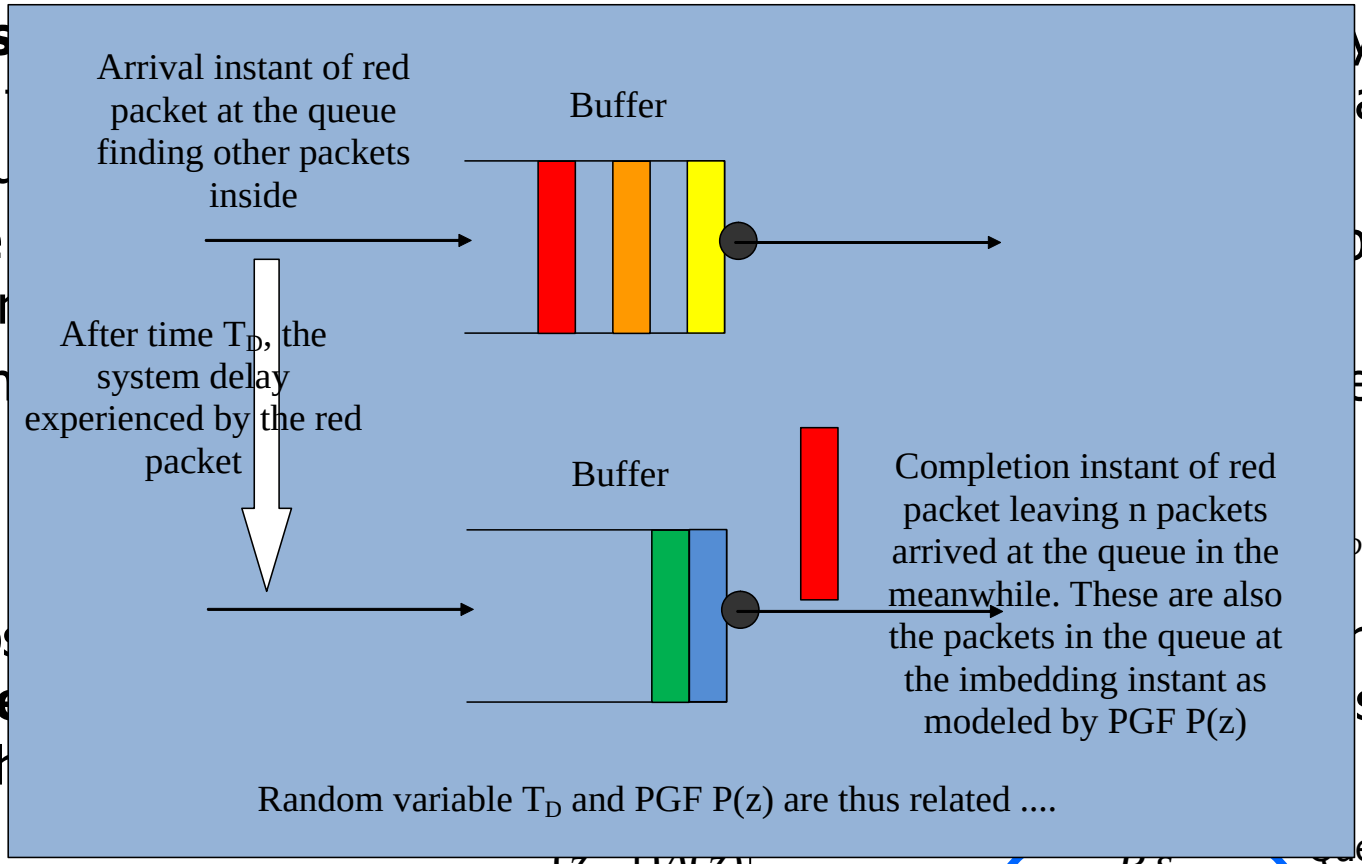
The same system admits another M/G/1 model **working at the level of messages**; imbedding points are now at the end of message service times (transmissions). This is a trivial application of the Pollaczek-Khinchin formula: $A(z) = e^{\lambda T(z-1)}$.

These **two models** for the same system are both interesting: the $M^{[L(z)]}/D/1$ model characterizes the system **at the level of packets** (number, delay); instead, the M/G/1 model characterizes the system **at the level of messages** (number, delay).

M/G/1 Delay Distribution in the FIFO Case with Poisson Arrivals

- **At s** ystem are a requ
- The bability distr
- Bein eing the

Subs ng the **inve** sform of th



$$T_D(s) = P(z)\Big|_{z=1-s/\lambda} = P_0 \frac{(z-1)A(z)}{z-A(z)}\Bigg|_{z=1-s/\lambda} = \Gamma(s) \times \frac{P_0 s}{s-\lambda+\lambda\Gamma(s)}$$

$[s=-\lambda(z-1)]$

euing delay

M/G/1 Delay Distribution in the FIFO Case with Poisson Arrivals

- **At service completion instant, the n requests left in the system are those arrived during the system delay T_D experienced by a request from input to output.**
- The probability distribution for n coincides with the state probability distribution with PGF P(z).
- Being $f_{TD}(t)$ the density function of the system delay [$T_D(s)$ being the Laplace transform], we can write in the z-domain:

The actual unknown term is $f_{TD}(t)$.

$$P(z) = \int_0^{+\infty} e^{\lambda t(z-1)} f_{T_D}(t)dt = T_D[s = -\lambda(z-1)]$$

Substituting the P(z) expression for the M/G/1 queue and using the **inverse transform z = 1 – s/λ**, we obtain the Laplace transform of the delay distribution:

$$T_D(s) = P(z)\Big|_{z=1-s/\lambda} = P_0 \frac{(z-1)A(z)}{z - A(z)}\Bigg|_{z=1-s/\lambda} = \Gamma(s) \times \frac{P_0 s}{s - \lambda + \lambda\Gamma(s)}$$

Queuing delay

Delay Distribution Analysis for the $M^{[L(z)]}/D/1$ Case with FIFO

- In the FIFO case with a **bulk (compound) Poisson arrival process** with PGF of the message length in packets L(z), the PGF of the number of packets in the buffer, P(z), and the Laplace transform of the probability density function of the **packet system delay**, $T_{Dp}(s)$, are related by means of the condition **$s = \lambda[1 - L(z)]$**.

- If L(z) is the PGF of a **modified geometric distribution** with mean value L we have [where $L^{-1}(.)$ is the inverse function of L(z)]:

$$s = \lambda\left[1 - \frac{z/L}{1 - z\left(1 - \frac{1}{L}\right)}\right]$$

inversion $\Rightarrow$

$$z = L^{-1}\left(1 - \frac{s}{\lambda}\right) = \frac{s - \lambda}{s\left(1 - \frac{1}{L}\right) - \lambda}$$

- This expression z = z(s) can be substituted in P(z) of the M/G/1 solution to obtain $T_{Dp}(s)$ as:

$$T_{D_p}(s) = [1 - \lambda T L'(1)]\frac{s \times e^{-\lambda T s/\lambda}}{(s - \lambda)L - [s(L-1) - \lambda L] \times e^{-\lambda T s/\lambda}}$$

M/G/1 Theory Generalization

- **Kleinrock principle** (also by P. J. Burke): for queuing systems where the state changes at most by +1 or −1 (we refer here to **the actual variations in the number of requests in the queue, not to what are the state changes between imbedding points**), the system distribution as seen by an arriving customer will be the same as that seen by a departing customer.
  - Hence, the state probability distribution by imbedding the queue at the departure instants is equal to the state probability distributions at arrival instants.
- Due to the **PASTA** property, the state probability distribution at arrival instants is valid at generic instants (random observer).
  - **The state probability distribution at the service completion instants coincides with the distribution of the continuous-time system (random observer).**

L. Kleinrock. *Queueing Systems*. New York: Wiley, 1975

M/G/1 Theory Generalization

- **Kleinrock principle** (also by P. J. Burke): for queuing systems where the state changes at most by +1 or −1 (we refer here to **the actual variations in the number of requests in the queue, not to what are the state change**
  distribut
  as that
  - Hence
    the de
    arriva

For a compound Poisson process the generalization considered here is not applicable. The Kleinrock principle is not applicable.

- Due to the **PASTA** property, the state probability distribution at arrival instants is valid at generic instants (random observer).
  - **The state probability distribution at the service completion instants coincides with the distribution of the continuous-time system (random observer).**

L. Kleinrock. *Queueing Systems*. New York: Wiley, 1975

M/G/1 Theory Generalization

- **Kl** … queuing sy… 1 or −1 (we re… **mber of re… e state ch…** ystem dis… be the same as…
  - … the queue at … distributions at arrival inst…

In the case of a **Bernoulli arrival process** on a slot basis (for which we can apply the 'M'/G/1 theory), the **BASTA** analogous property holds, so that we can reapply the generalization result below.

- Due to the **PASTA** property, the state probability distribution at arrival instants is valid at generic instants (random observer).
  - **The state probability distribution at the service completion instants coincides with the distribution of the continuous-time system (random observer).**

L. Kleinrock. *Queueing Systems*. New York: Wiley, 1975

M/G/1 Theory Generalization (cont'd)

- As a further proof of the generalization of the state probability distribution of M/G/1 at generic instants, we could use the following **heuristic considerations**.

- The Pollaczek-Khinchin formula can also be applied to the **M/M/1 queue** (imbedding points at the service completion instants), where mean and mean square values of the service time X are so related (exponential distribution case): $E[X^2] = 2E[X]^2$.

$$T = E[X] + \frac{\lambda E[X^2]}{2[1-\lambda E[X]]}$$

$$\Rightarrow \quad T = E[X] + \frac{2\lambda E[X]^2}{2[1-\lambda E[X]]} = \frac{E[X] - \lambda E[X]^2 + \lambda E[X]^2}{[1-\lambda E[X]]} = \frac{E[X]}{1-\lambda E[X]}$$

exponential service time

classical M/M/1 result

- We note that **we obtain again the classical M/M/1 result that is valid at any instant, not only at imbedding points**.

Numerical Inversion Method for P(z)

- The PGF P(z) of an M/G/1 queue has typically an expression that cannot be inverted to obtain the state probability distribution. A **numerical inversion method** is needed.

- As explained in Lesson No. 3, P(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 obtained looking at the definition of P(z) :

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

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

M/G/1 Theory and Heavy-Tailed-Distributed Service Times

- Heavy-tailed (Pareto) distributions for the service time are frequent in modern traffic. One disadvantage of using these distributions is that their Laplace transforms often have no closed-form expressions and are thus not easy to manipulate.

- The M/G/1 state probability distribution depends on A(z), the PGF of the number of arrivals in a service time. Moreover, the mean delay is given by the Pollaczek-Khinchin formula, which requires to use mean and mean square values of the service time. **With heavy-tailed distributions, we can have infinite mean and/or variance, which may entail some paradoxical situations for the queues**, as discussed below referring to the Pareto distribution case with shape parameter γ.

- In the M/Pareto/1 case, we need to have a **finite mean value** of the Pareto service time (thus entailing $\gamma > 1$) in order to have a **stable queue.**

M/Pareto/1 Queue

- If $1 < \gamma \leq 2$, the Pareto service time has finite mean and infinite variance (i.e., **heavy tails**). This entails that the queue is stable (there exists the state probability distribution as well as the distribution of the delay), but the mean delay is infinite. Hence, **this is a very special (degenerate) case, where the infinite mean delay does not imply the instability of the queue**!

- The PGF of the state probability distribution, P(z), depends on A(z) computed as follows:

$$A(z) = \sum_{n=0}^{\infty} z^n \int_k^{+\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \frac{\gamma k^\gamma}{t^{\gamma+1}} dt = \gamma k^\gamma \int_k^{+\infty} e^{-\lambda t(1-z)} t^{-\gamma-1} dt \overset{s=\lambda(1-z)}{=} \gamma k^\gamma \int_k^{+\infty} e^{-st} t^{-\gamma-1} dt$$

The integral in A(z) cannot be expressed in a closed form. It can be represented by means of the *incomplete Gamma function*, Γ(a, y):

$$A(z) = \gamma k^\gamma \int_k^{+\infty} e^{-st} t^{-\gamma-1} dt \Big|_{s=\lambda(1-z)} = \gamma (sk)^\gamma \Gamma(-\gamma, sk)\Big|_{s=\lambda(1-z)}, \quad where \;\; \Gamma(a, y) = \int_y^{+\infty} e^{-t} t^{a-1} dt$$

- If $\gamma > 2$, the Pareto distribution has finite mean and finite variance so that the mean delay is finite. In this case, the Pareto distribution is not heavy-tailed.

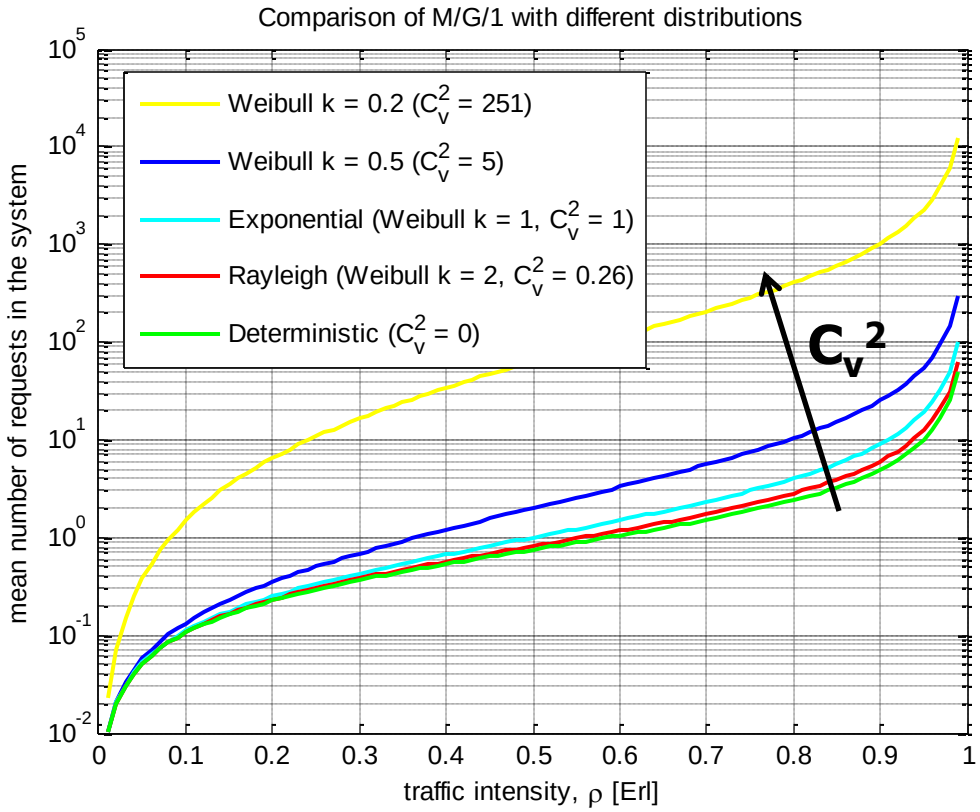
M/G/1 Mean Number of Requests for Different Serv. Time Distrib.

- Let us compare the mean delay of an M/G/1 queue for different service distributions **with the same mean arrival rate λ and mean service time E[x]**. Let $\rho = \lambda E[x] < 1$ Erl denote the traffic intensity.
- The different service time distributions are characterized by the coefficient of variation C_v: $C_v^2 = \frac{Var[X]}{E[X]^2}$. The exponential distribution has $C_v = 1$.
- **The coefficient of variation C_v is 0 for a *deterministic* random variable, is 1 for an exponential distribution, is greater than 1 for the hyper-exponential distribution, and tends to ∞ for heavy-tailed distributions.**
- Let us compare the mean number of requests in the system for exponential and general service times (i.e., M/M/1 vs. M/G/1):

$$N_{M/M/1} = \frac{\lambda E[X]}{1-\lambda E[X]} \quad vs. \quad N_{M/G/1} = \lambda E[X] + \frac{\lambda^2 E[X^2]}{2(1-\lambda E[X])} = \lambda E[X] + \frac{(\lambda E[X])^2 (C_v^2+1)}{2(1-\lambda E[X])} = N_{M/M/1}\left[1+\lambda E[X]\left(\frac{C_v^2-1}{2}\right)\right]$$

We have: $N_{M/M/1} < N_{M/G/1} \Leftrightarrow C_v^2 = \frac{Var[X]}{E[X]^2} > 1$ $\quad N_{M/M/1} > N_{M/G/1} \Leftrightarrow C_v^2 = \frac{Var[X]}{E[X]^2} < 1$

Comparison .... (cont'd)



Weibull distribution:

$$f_{\beta,k}(t) = \frac{k}{\beta}\left(\frac{t}{\beta}\right)^{k-1} e^{-\left(\frac{t}{\beta}\right)^k}, \quad t \geq 0$$

The Weibull distribution is used since varying parameter k, we can obtain distributions with different C_v^2 values from low values (< 1) to high values (> 1).

- **At a parity of ρ, the mean waiting time of the M/G/1 queue increases with C_v^2, the square coefficient of variation of the service time.**

First Exercises on M/G/1 Theory

Exercise #1

- We have a buffer of a transmission line that receives messages coming from two independent processes:
  - *First traffic*: Poisson message arrival process with mean rate λ_1 and exponentially-distributed service time with mean rate μ_1;
  - *Second traffic*: Poisson message arrival process with mean rate λ_2 and exponentially-distributed service time with mean rate μ_2.
- Assuming $\mu_1 \neq \mu_2$, we have to determine the mean delay from the message arrival (total arrival process sum of both processes) to the buffer to its transmission completion.

Solution of Exercise #1

- The first and the second arrival processes are at the input of the buffer. Since they are independent Poisson processes, their sum is still Poisson with mean rate $\lambda_1 + \lambda_2$.
- **The service time probability density function, f(t), is not exponential**; it can be derived as weighted sum of the probability density functions related to the two different input flows:

$$f(t) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \mu_1 e^{-\mu_1 t} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \mu_2 e^{-\mu_2 t}$$

Hyper-exponential service time distribution ($C_v > 1$)

- We model this buffer by means of an **M/G/1 queue**: we imbed the chain at the instants of message transmission completion and we use the **Pollaczek-Khinchin formula**.

$$T = E[X] + \frac{(\lambda_1 + \lambda_2) E[X^2]}{2[1 - (\lambda_1 + \lambda_2) E[X]]}$$

where

$$E[X] = \frac{\lambda_1}{\lambda_1 + \lambda_2} \times \frac{1}{\mu_1} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \times \frac{1}{\mu_2}$$

$$E[X^2] = \frac{\lambda_1}{\lambda_1 + \lambda_2} \times \frac{2}{(\mu_1)^2} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \times \frac{2}{(\mu_2)^2}$$

Stability: $(\lambda_1 + \lambda_2)E[X] = \lambda_1/\mu_1 + \lambda_2/\mu_2 < 1$ Erl

The intensities of the two traffic flows sum.

Exercise #2

- We consider a link with a transmission buffer where messages arrive according to a Poisson process with mean arrival rate λ.
- Each message is formed of a random number of packets, each requiring a time T to be transmitted (**compound Poisson process**). L(z) denotes the PGF of the message length in packets that also corresponds to the PGF of the message transmission time in T units.
- Note:
  - All the packets of the same message arrive simultaneously.
  - The arrival process and the transmission one are continuous-time (non-time-slotted).
- It is requested to determine the mean message delay for a generic L(z) by selecting **suitable imbedding instants**.

Exercise #2

The arrival process at the packet level is compound Poisson; instead, the same arrival process is simply Poisson at the message level.

- ion buffer where messages
  ss with mean arrival rate λ.
- m number of packets, each requiring a time T to be transmitted (**compound Poisson process**). L(z) denotes the PGF of the message length in packets that also corresponds to the PGF of the message transmission time in T units.
- Note:
  - All the packets of the same message arrive simultaneously.
  - The arrival process and the transmission one are continuous-time (non-time-slotted).
- It is requested to determine the mean message delay for a generic L(z) by selecting **suitable imbedding instants**.

Solution of Exercise #2

- Let us **imbed the system at the instants of message transmission completion: this is the best option to measure the performance at the message level (imbedding at the end of packet transmission is not suitable to determine the mean message delay)**.
- Let n_i represent the number of messages in the buffer at the end of the transmission of the i-th message; let a_i denote the number of messages arrived at the buffer during the service time of the i-th message.
- We have a classical **M/G/1 queue** with Poisson arrival process. Then, we directly apply the **Pollaczek-Khinchin** formula to derive the mean message delay:

$$T_m = L'(1)T + \frac{\lambda[T]^2[L''(1)+L'(1)]}{2[1-L'(1)\lambda T]} \quad [\text{seconds}]$$

Mean square value of the message transmission time

Mean value of the message transmission time

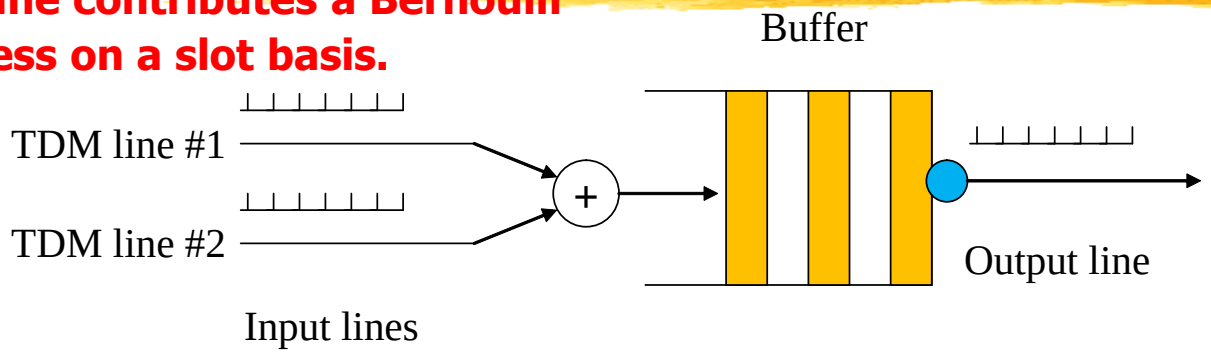
The stability condition is $\lambda TL'(1) < 1$ Erl

Exercise #3 (ATM-like case, 'M'/D/1 queue)

- Let us consider that fixed-size packets arrive at a transmission buffer from two TDM input lines: line #1 and line #2. The transmission of packets from the buffer is according to a TDM output line.
  - Input and output slots have the same duration. Input TDM lines are synchronous each other and synchronous with the output line as well.
  - A slot of the input line #1 carries a packet with probability *p*; a slot of the input line #2 carries a packet with probability *q*. A packet needs a slot to arrive and to be stored in the buffer before it can be sent (store-and-forward case).
  - The arrival processes on the two lines are memoryless and independent.
- It is requested to determine the mean delay that a packet experiences from the arrival at the buffer to the end of its transmission. **This is a first example of discrete-time system that we solve by means of an 'M'/D/1 queue.**

Solution of Exercise #3

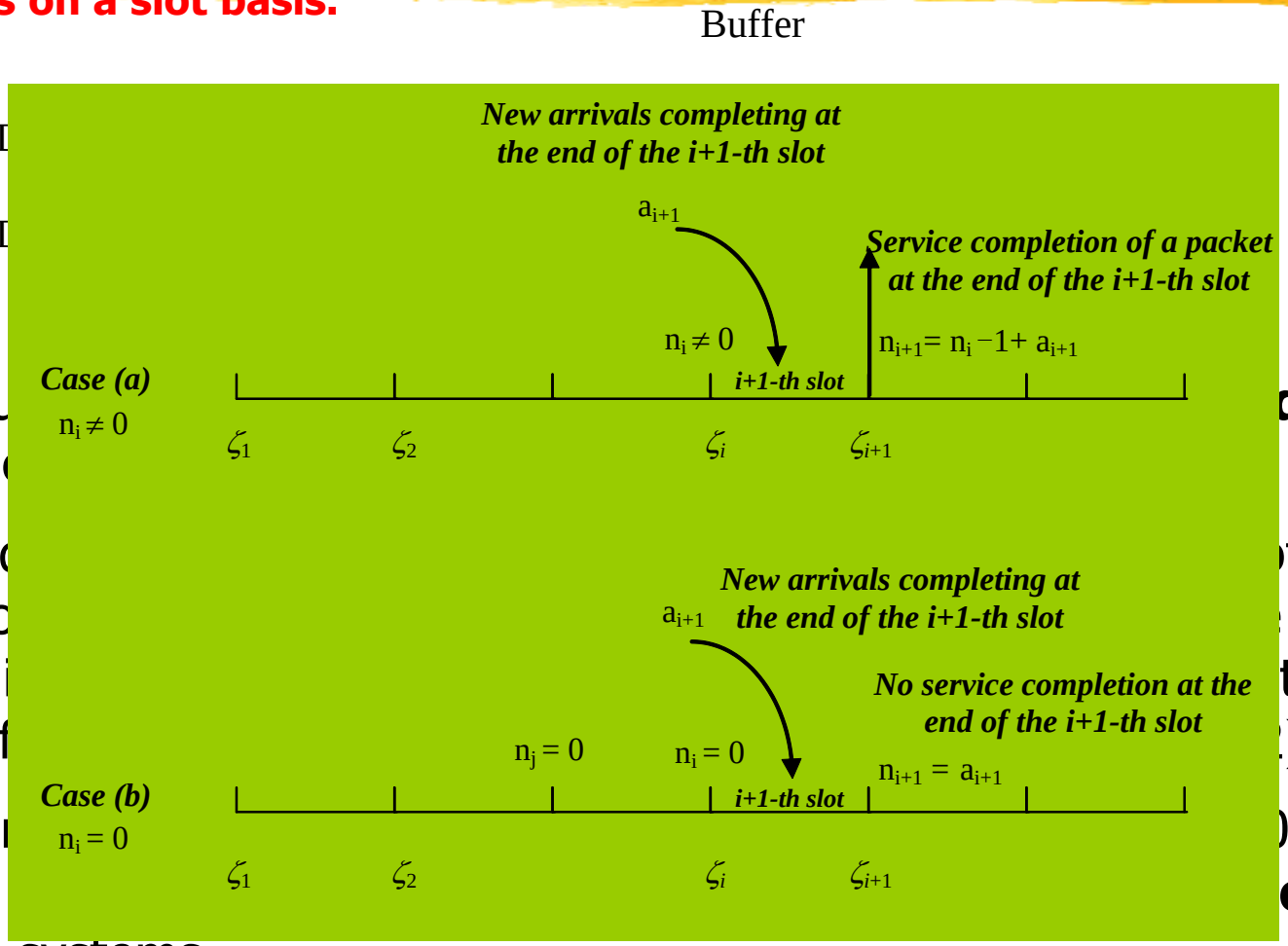
Each input line contributes a Bernoulli arrival process on a slot basis.



- We study this discrete-time system by **imbedding at the end of the slots of the output TDM line**.
- Let n_i denote the number of packets in the buffer at the end of the i-th slot. Let a_i denote the number of packets arrived from the two input lines in the buffer during the i-th slot (we consider here the sum of the independent input processes from lines #1 and #2).
- We can write the following balance: $n_{i+1} = n_i - 1 + a_{i+1}$ for $n_i > 0$ and $n_{i+1} = a_{i+1}$ for $n_i = 0$. This is the **classical difference equation** of M/G/1 systems.

Solution of Exercise #3

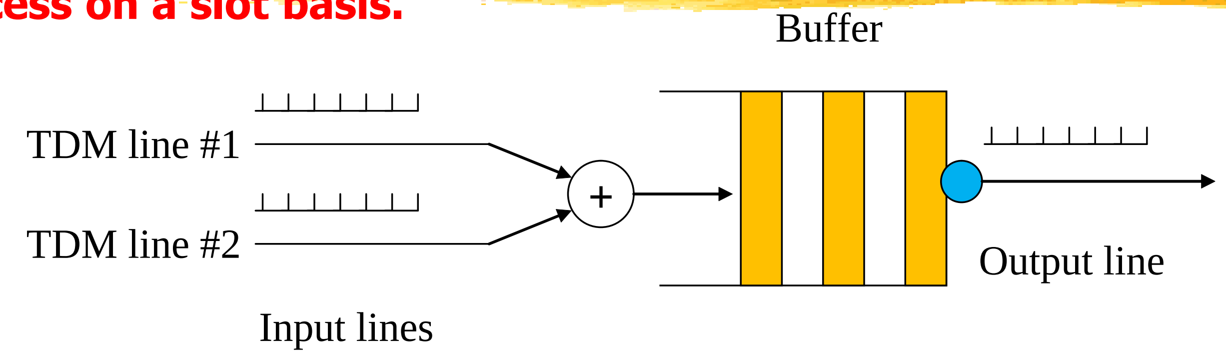
Each input line contributes a Bernoulli arrival process on a slot basis.



- We stu… **d of the sl…**
- Let n_i … of the i-th slo… two input li… the sum of …).
- We ca…) and n_{i+1} = … **on** of M/G/1 systems.

Solution of Exercise #3

Each input line contributes a Bernoulli arrival process on a slot basis.



- We study this discrete-time system by **imbedding at the end of**

We consider a classical assumption for this type of systems: a packet must have completely arrived (1 slot) before its transmission can start, according to the store-and-forward approach.

- ...e buffer at the end of the ...kets arrived from the two ...ot (we consider here the ...from lines #1 and #2).
- We ... ite the following balance: $n_{i+1} = n_i - 1 + a_{i+1}$ for $n_i > 0$ and $n_{i+1} = a_{i+1}$ for $n_i = 0$. This is the **classical difference equation** of M/G/1 systems.

Solution of Exercise #3 (cont'd)

- The mean number of packets in the buffer is:

$$N_c = A'(1) + \frac{A''(1)}{2[1 - A'(1)]} \quad [\text{pkts}]$$

where A(z) is related to the sum of two independent processes (product in the z-domain of the PGFs):

$$A(z) = (1 - p + zp)(1 - q + zq) = (1 - p)(1 - q) + z[p(1 - q) + q(1 - p)] + z^2 pq =$$
$$= A_0 + zA_1 + z^2 A_2$$
$$A'(1) = A_1 + 2A_2 = p + q \qquad A''(1) = 2A_2 = 2pq$$

- The stability condition is A'(1) = $p + q < 1$ Erlang
- The mean packet delay is derived from N by using the **Little theorem**: we divide N by is A'(1), the mean number of packets arrived per slot.
  - For **time-slotted systems**, we consider the PGF of the number of arrivals in a slot A(z). Then, A'(1) represents the mean number of arrivals per slot. Therefore, we can apply the Little theorem dividing the mean number of requests by A'(1): T = N/A'(1) is expressed in slot units.

Solution of Exercise #3 (cont'd)

- The mean packet delay is:

$$T = \frac{N}{A'(1)} = 1 + \frac{A''(1)/A'(1)}{2[1 - A'(1)]} = 1 + \frac{pq}{(p+q)(1-p-q)} \quad [\text{slots}]$$

Further Application Examples of the M/G/1 Theory to Telecommunications

ARQ Scheme for Reliable Transmissions on a Link

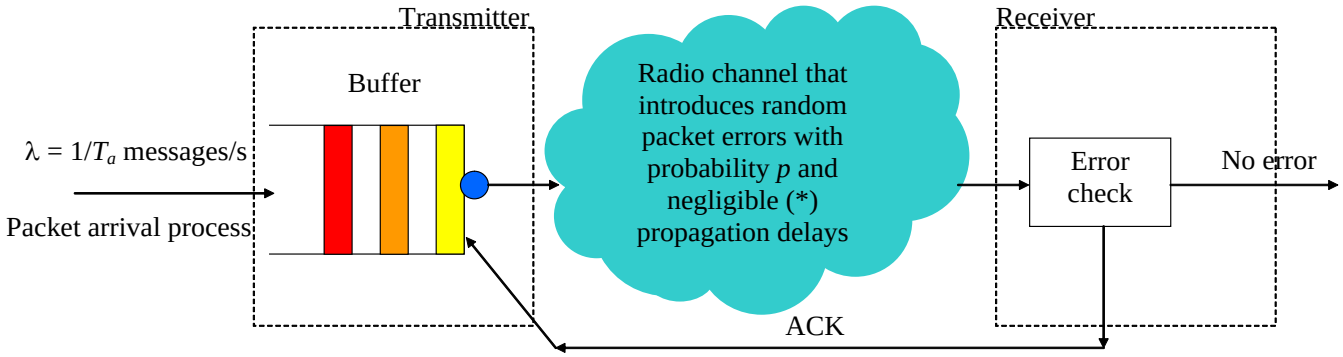
Analysis of an ARQ Scheme

- We consider a transmission system with a buffer. The transmitter is used to send packets on a radio channel. We know that:
  - Packets arrive in groups of messages (bulk arrival process)
  - Messages arrive according to exponentially-distributed intervals with mean value equal to T_a in seconds.
  - The length l_m of a message in packets is according to the following distribution (uncorrelated from message to message): $\text{Prob}\{l_m = n \text{ pkts}\} = q(1-q)^{n-1}, \quad n \in \{1, 2, \ldots\}$
  - The buffer has infinite capacity.
  - The radio channel causes a packet to be erroneously received with probability p; packet errors are uncorrelated from packet to packet.
  - An ARQ scheme is adopted.
  - Round trip propagation delays to receive ACKs are negligible with respect to the deterministic packet transmission time, T (note *): if the transmission of a packet is unsuccessful its retransmission is soon reattempted.
  - A packet sojourns in the buffer until its ACK is received.

- We have to determine the mean number of packets in the buffer and the mean delay that a packet experiences from its arrival at the buffer to its last and successful transmission.

(*) The extension of this study to a case with high propagation delays is straightforward in the ARQ stop-and-wait case.

Analysis of an ARQ Scheme: A Model with Feedback



- **Bulk arrival process**:
  - Messages arrive according to a Poisson process with mean rate $\lambda = 1/T_a$ [messages/s].
  - Each message contains a number of packets with modified geometric distribution with parameter q; 1/q = mean length of a message in packets.

- **Service process:**
  - Due to the errors introduced by the channel, each packet requires a modified geometrically distributed number of slots (with parameter 1–p) to be transmitted; 1/(1–p) = mean time in slot units to successfully transmit a packet.
  - Each slot has duration T.

(*) The extension of this study to a case with high propagation delays is straightforward in the ARQ stop-and-wait case.

Solution

The queue notation is also related to the type of imbedding instants selected.

- The arrival process is compound Poisson, but we can still use the 'M'/G/1 theory. In this case, we have an $M^{[Geom]}$/Geom/1 system.
- We imbed the chain to the **instants of successful packet transmission (i.e., without error)**; a packet could be transmitted many times to achieve a successful delivery. We can write as a first approximation the classical M/G/1 difference equation with n_i and a_i. The details of this approximation (related to the bulk arrival process) will be clarified in Lesson No. 9.
- A(z) denotes the PGF of the number of packets arrived in the time required to successfully transmit a packet, T_s. In the derivation of A(z) **three random variables need to be taken into account**:
  - Number of messages arrived in T;
  - Number of packets conveyed by each message;
  - Time necessary in slots to successfully transmit a packet by means of ARQ (neglecting the round trip propagation delay, all the ARQ schemes are almost equivalent), T_s.
- PGF of the message length in packets $L(z) = \dfrac{zq}{1 - z(1-q)}$
- PGF of the time in slot to successfully transmit a packet $T_s(z) = \dfrac{z(1-p)}{1 - zp}$
- The input traffic intensity is: $\dfrac{1}{1-p} \times \lambda T \times \dfrac{1}{q}$

Solution

- The arrival process is compound Poisson, but we can still use the 'M'/G/1 theory. In this case, we have an $M^{[Geom]}$/Geom/1 system.
- We imbed the chain to the **instants of successful packet transmission (i.e., without error)**; a packet could be transmitted many times to achieve a successful delivery. We can wr
  difference equation with n_i and
  the bulk arrival process) will b
- A(z) denotes the PGF of the nu
  successfully transmit a packet,
  variables need to be taken
  - Number of messages arrived in T
  - Number of packets conveyed by e
  - Time necessary in slots to success
    round trip propagation delay, all t
- PGF of the message length in
- PGF of the time in slot to succ
- The input traffic intensity is:

Note that we could even **imbed the queue at the end of successful message transmissions** thus obtaining a queue of the M/Geom/1, where the Geom service time is the result of the geometric number of packets per message composed with the geometric distribution of the packet service time in slots. The mean message delay is thus given by the Pollaczek-Khinchin formula. N.B. The composition of two random variables with modified geometric distributions has still a modified geometric distribution with mean value given by the product of the mean values of the composing geometric variables.

Solution (cont'd)

The derivatives of this compound function A(z) can be obtained by leaving $T_s(z)$ and L(z) in 'implicit forms', because this allows easier derivatives by using $T_s(1) = 1$, $T_s'(1) = 1/(1-p)$, $L(z) = 1$, $L'(1) = 1/q$.

- The PGF of the number of packets arrived in the time to serve one packet, A(z), is obtained by considering the twofold composition of PGFs:

$$A(z) = T_s\left[e^{\lambda T(L(z)-1)}\right]$$

This is the PGF of the number of packets arrived in T.

$$A'(1) = \frac{1}{1-p} \times \lambda T \times \frac{1}{q}$$

$$A''(1) = \left[\frac{\lambda T}{q}\right]^2 \left[\frac{2p}{(1-p)^2} + \frac{1}{1-p}\right] + \frac{\lambda T}{1-p} \times \frac{2(1-q)}{q^2}$$

- The buffer stability is assured if $A'(z = 1) < 1$ Erlang $\Rightarrow \lambda T/[q(1-p)] < 1$ Erlang.

- The mean number of packets in the ARQ sender buffer N_p and the mean delay for the correct transmission of a packet T_p are:

by means of the Little theorem

$$N_p = A'(1) + \frac{A''(1)}{2[1 - A'(1)]} \qquad T_p = \frac{qN_p}{\lambda} \quad [\text{s}]$$

Thank you!

giovanni.giambene@gmail.com