Performance Modelling of Computer Systems

Mirco Tribastone

Institut für Informatik
Ludwig-Maximilians-Universität München

Fundamentals of Queueing Theory

Some Notable Infinite Series

For any real $|x| < 1$,

$$\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x, \tag{1}$$

$$\sum_{k=1}^{\infty} x^k = \frac{x}{1-x}, \tag{2}$$

$$\sum_{k=1}^{\infty} kx^k = \frac{x}{(1-x)^2}. \tag{3}$$

Birth-Death Processes

- Continuous-time Markov chain with states labelled $0, 1, \dots, k, \dots$
- Jumps are only allowed between neighbouring states:
  - State 0 may only make a transition to 1.
  - A state $k > 0$ may make transitions to $k - 1$ and $k + 1$.
- Population increases (births) happen at rate $\lambda_k > 0$.
- Population decreases (deaths) happen at rate $\mu_k > 0$.
- Births and deaths are independent.

Model assumptions

$$
\begin{aligned}
\mathbb{P}(\text{exactly one birth in } (t, t+\Delta t) \mid X(t) = k) &= \lambda_k \Delta t + o(\Delta t), \\
\mathbb{P}(\text{exactly zero births in } (t, t+\Delta t) \mid X(t) = k) &= 1 - \lambda_k \Delta t + o(\Delta t), \\
\mathbb{P}(\text{exactly one death in } (t, t+\Delta t) \mid X(t) = k) &= \mu_k \Delta t + o(\Delta t), \\
\mathbb{P}(\text{exactly zero deaths in } (t, t+\Delta t) \mid X(t) = k) &= 1 - \mu_k \Delta t + o(\Delta t).
\end{aligned}
$$

Chapman-Kolmogorov Equations of Birth-Death Processes

Denote

$$p_k(t) := \mathbb{P}(X(t) = k), \qquad k \geq 0.$$

By the law of total probability:

$$\begin{aligned}
p_0(t + \Delta t) &= p_0(t)(1 - \lambda_0 \Delta t) + \mu_1 p_1(t)\Delta t + o(\Delta t), \\
p_k(t + \Delta t) &= p_{k-1}(t)\lambda_{k-1}\Delta t + p_k(t)(1 - \lambda_k \Delta t)(1 - \mu_k \Delta t) \\
&\qquad + p_{k+1}(t)\mu_k \Delta t + o(\Delta t) \\
&= p_{k-1}(t)\lambda_{k-1}\Delta t + p_k(t)\left[1 - \mu_k \Delta t - \lambda_k \Delta t + \lambda_k \mu_k \Delta t^2\right] \\
&\qquad + p_{k+1}(t)\mu_{k+1}\Delta t + o(\Delta t), \qquad \text{for } k > 0.
\end{aligned}$$

Rearranging yields

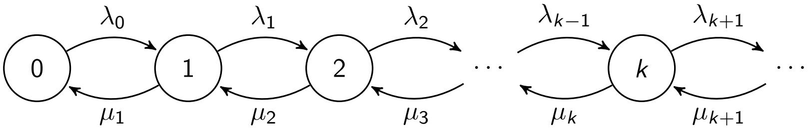
$$\begin{aligned}
\frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} &= -\lambda_0 p_0(t) + \mu_1 p_1(t) + \frac{o(\Delta t)}{\Delta t}, \\
\frac{p_k(t + \Delta t) - p_k(t)}{\Delta t} &= \lambda_{k-1} p_{k-1}(t) - (\lambda_k + \mu_k) p_k(t) \\
&\qquad + \mu_{k+1} p_{k+1}(t) + \frac{o(\Delta t)}{\Delta t}, \quad \text{for } k > 0.
\end{aligned}$$

Chapman-Kolmogorov Equations of Birth-Death Processes

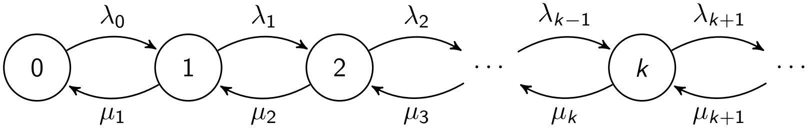
$$
\begin{aligned}
\frac{p_0(t+\Delta t) - p_0(t)}{\Delta t} &= -\lambda_0 p_0(t) + \mu_1 p_1(t) + \frac{o(\Delta t)}{\Delta t}, \\
\frac{p_k(t+\Delta t) - p_k(t)}{\Delta t} &= \lambda_{k-1} p_{k-1}(t) - (\lambda_k + \mu_k) p_k(t) \\
&\qquad + \mu_{k+1} p_{k+1}(t) + \frac{o(\Delta t)}{\Delta t}, \quad \text{for } k > 0.
\end{aligned}
$$

Taking the limit $\Delta t \to 0$ yields

$$
\begin{aligned}
\frac{dp_0(t)}{dt} &= -\lambda_0 p_0(t) + \mu_1 p_1(t), \\
\frac{dp_k(t)}{dt} &= \lambda_{k-1} p_{k-1}(t) - (\lambda_k + \mu_k) p_k(t) + \mu_{k+1} p_{k+1}(t), \qquad k > 0.
\end{aligned}
$$



A Generic Recursive Stationary Solution (1/2)



$$\frac{dp_0(t)}{dt} = -\lambda_0 p_0(t) + \mu_1 p_1(t),$$
$$\frac{dp_k(t)}{dt} = \lambda_{k-1} p_{k-1}(t) - (\lambda_k + \mu_k) p_k(t) + \mu_{k+1} p_{k+1}(t), \qquad k > 0.$$

Setting $dp_k(t)/dt = 0$ for all $k \geq 0$ yields

$$\lambda_{k-1}\pi_{k-1} - (\lambda_k + \mu_k)\pi_k + \mu_{k+1}\pi_{k+1} = 0, \quad \text{with } \lambda_i = \mu_i = 0 \text{ for all } i < 0.$$

A Generic Recursive Stationary Solution (2/2)

Rearranging $\lambda_{k-1}\pi_{k-1} - (\lambda_k + \mu_k)\pi_k + \mu_{k+1}\pi_{k+1} = 0$ gives

$$\lambda_{k-1}\pi_{k-1} - \mu_k\pi_k = \underbrace{\lambda_k\pi_k - \mu_{k+1}\pi_{k+1}}_{g(k)}.$$

Observe that

$$g(k-1) = g(k), \qquad \text{for all } k,$$

therefore $g(k)$ must be constant with k. From $dp_0(t)/dt = 0$ we get that $g(k) = 0$. Therefore, we obtain the recursive solution

$$\pi_{k+1} = \frac{\lambda_k}{\mu_{k+1}}\pi_k \Longrightarrow \underbrace{\pi_k = \pi_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}_{\text{product form type}}, \qquad k = 0, 1, 2, \ldots$$

Question: How to compute π_0?

$$\pi_0 = 1 - \sum_{k=1}^{\infty} \pi_k = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i-0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}.$$

The Poisson Process

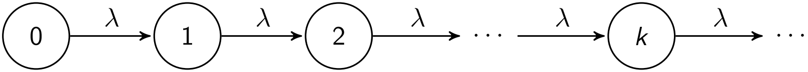
Consider a pure-birth process with $\lambda_k = \lambda$ for $k \geq 0$ and $\mu_k = 0$ for $k > 0$, and assume that $p_0(0) = 1$ and $p_k(0) = 0$ for all $k > 0$. The equations simplify to

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t)$$
$$\frac{dp_k(t)}{dt} = \lambda p_{k-1}(t) - \lambda p_k(t), \qquad k > 0.$$

Solving the first equation $p_0(t) = e^{-\lambda t}$, by induction it is proven that

$$p_k(t) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \qquad k \geq 0.$$

This is the Poisson process, a counting process with exponentially distributed increments with mean $1/\lambda$.



Properties of Poisson Process

- Mean:

$$\mathbb{E}[N(t)] = \sum_{k=0}^{\infty} k p_k(t) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$
$$= e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{(k-1)!} = e^{-\lambda t} \sum_{k=1}^{\infty} \frac{(\lambda t)^k}{(k-1)!}$$
$$= e^{-\lambda t}(\lambda t) \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} = e^{-\lambda t}(\lambda t) e^{\lambda t} = \lambda t.$$

- Variance: $\mathrm{Var}[N(t)] = \lambda t$.
- Memoryless property:

$$\mathbb{P}(N(s, s+t) = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!},$$

where $N(s, s+t)$ is defined as the number of arrivals between s and $s+t$.

Why Are Poisson Processes Relevant?

In addition to mathematical tractability, they model many phenomena:

- Arrivals of calls in a telephone network;
- Decay of radioactive elements (gamma-ray emissions);
- Army soldiers killed due to being kicked by their horses;[1]
- A large number of independent renewal processes will tend to a Poisson process.

- The birth-death process may be interpreted as a queueing system where increments (with rate λ) denote arrivals of request and decrements (with rate μ) are related to services.
- It is denoted in Kendall notation as the $M/M/1$ system:
  - Exponentially distributed interarrival times;
  - Exponentially distributed services;
  - Single-server system.

[1] Kleinrock, *Queueing Systems: Volume I — Theory*, Wiley Interscience, NY, 1975.

Kendall's Notation

A queueing system is often denoted by $A/B/C/X/Y/Z$, where:

- A gives the distribution of interarrival times (e.g., M, E, G, D, $\ldots$);
- B gives the service time distribution;
- C gives the service multiplicity $(1, 2, \ldots, \infty)$;
- X gives the system capacity;
- Y gives the customer population;
- Z gives the queue discipline (i.e., FIFO, LIFO, RANDOM, etc.).

Stationary Distribution of the $M/M/1$ System (1/2)

Consider again the equations of motion:

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) + \mu p_1(t),$$
$$\frac{dp_k(t)}{dt} = \lambda p_{k-1}(t) - (\lambda + \mu) p_k(t) + \mu p_{k+1}(t), \qquad k > 0.$$

We look for a probability vector $\pi = [\pi_0, \pi_1, \ldots, \pi_k, \ldots]$ such that

$$-\lambda \pi_0 + \mu \pi_1 = 0,$$
$$\lambda \pi_{k-1} - (\lambda + \mu)\pi_k + \mu \pi_{k+1} = 0, \qquad k > 0,$$
$$\sum_{i=0}^{\infty} \pi_i = 1.$$

- From the first equation, $\pi_1 = (\lambda/\mu)\pi_0$.
- For $k = 1$, $\lambda \pi_0 - (\lambda + \mu)\pi_1 + \mu \pi_2 = 0$, from which

$$\pi_2 = (\lambda/\mu + 1)\pi_1 - (\lambda/\mu)\pi_0 = (\lambda/\mu)^2 \pi_0.$$

Stationary Distribution of the $M/M/1$ System (2/2)

By induction one can prove that

$$\pi_k = (\lambda/\mu)^k \pi_0, \qquad \text{for } k > 0.$$

$$\pi_0 = 1 - \sum_{k=1}^{\infty} \pi_k = 1 - \sum_{k=1}^{\infty} (\lambda/\mu)^k \pi_0 \Longrightarrow \pi_0 \left(1 + \sum_{k=1}^{\infty} (\lambda/\mu)^k \right) = 1.$$

If $\lambda < \mu$ then the series converges. Therefore

$$\pi_0 = \frac{1}{1 + \frac{\lambda/\mu}{1 - \lambda/\mu}} = 1 - \lambda/\mu.$$

Setting $\rho = \lambda/\mu$ we obtain

$$\pi_k = \rho^k (1 - \rho), \qquad \text{for all } k > 0.$$

Performance Metrics for the $M/M/1$ Queue (1/2)

- Mean queue length: average number of customers in the system.

$$
\begin{aligned}
L = \sum_{k=0}^{\infty} k\pi_k = \sum_{k=0}^{\infty} k\rho^k(1-\rho) &= (1-\rho)\rho \sum_{k=1}^{\infty} k\rho^{k-1} \\
&= (1-\rho)\rho \sum_{k=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}\rho}\rho^k = (1-\rho)\rho\frac{\mathrm{d}}{\mathrm{d}\rho}\sum_{k=0}^{\infty}\rho^k \\
&= (1-\rho)\rho\frac{\mathrm{d}}{\mathrm{d}\rho}\left[\frac{\rho}{1-\rho}\right] = (1-\rho)\rho\frac{1-\rho+\rho}{(1-\rho)^2} = \frac{\rho}{(1-\rho)}.
\end{aligned}
$$

- Utilisation: probability that the server is busy.
  - Formally, it may be defined as the expected value of a function of the random variable that underlies the stationary distribution.

$$
u(X) = \begin{cases} 0 & \text{if } X = 0, \\ 1 & \text{otherwise.} \end{cases}
$$

$$
U := \mathbb{E}[u(X)] = \sum_{k=1}^{\infty} \rho^k(1-\rho) = 1 - \pi_0 = \rho.
$$

Performance Metrics for the $M/M/1$ Queue (2/2)

- Average response time. We invoke Little's law, which states that for a system in steady state

$$L = \lambda W,$$

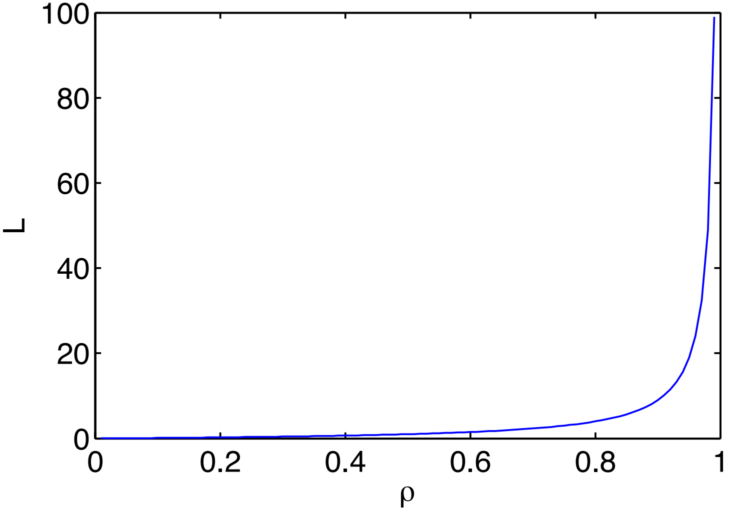
  where:

  L is the average number of users in the system;
  λ is the steady-state rate of arrivals into the system (which is equal to the throughput, i.e., the steady-state rate of departures from the system);
  W is the average response time.

- In the $M/M/1$ queue,

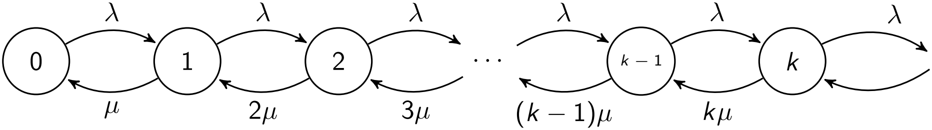
$$W = L/\lambda = \frac{\rho}{(1-\rho)\lambda} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda}.$$

$M/M/1$ Performance, Pictorially



The $M/M/\infty$ System

Service capacity is proportional to the number of customers in the system.



$$\lambda\pi_0 = \mu\pi_1 \Longrightarrow \pi_1 = \frac{\lambda}{\mu}\pi_0$$

$$\lambda\pi_1 + \mu\pi_1 = \lambda\pi_0 + 2\mu\pi_2 \Longrightarrow \pi_2 = \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 \pi_0$$

By induction, $\pi_k = \frac{1}{k!}(\lambda/\mu)^k\pi_0$, $k > 0$. From the normalisation condition,

$$1 = \pi_0 + \sum_{k=1}^{\infty} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \pi_0 = \pi_0 \left[1 + \sum_{k=1}^{\infty} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k\right] = \pi_0 \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k = \pi_0 e^{\lambda/\mu}$$

$$\Longrightarrow \quad \pi_k = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k e^{-\lambda/\mu}, \qquad k \geq 0.$$

Performance of the $M/M/\infty$ System

- Average queue length

$$L = \sum_{k=0}^{\infty} k\pi_k = \sum_{k=0}^{\infty} k \frac{1}{k!} \left(\frac{\lambda}{\mu}\right)^k e^{-\lambda/\mu} = \sum_{k=1}^{\infty} k \frac{1}{k!} \left(\frac{\lambda}{\mu}\right)^k e^{-\lambda/\mu}$$

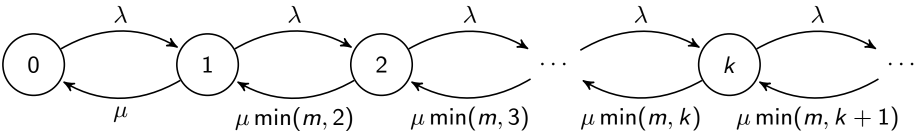
$$= e^{-\lambda/\mu} \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \left(\frac{\lambda}{\mu}\right)^k = e^{-\lambda/\mu} \left(\frac{\lambda}{\mu}\right) \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \left(\frac{\lambda}{\mu}\right)^{k-1}$$

$$= e^{-\lambda/\mu} \left(\frac{\lambda}{\mu}\right) \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n = e^{-\lambda/\mu} \left(\frac{\lambda}{\mu}\right) e^{\lambda/\mu} = \frac{\lambda}{\mu}.$$

- Average response time: $W = 1/\mu$.
- Question: What is the utilisation?

The $M/M/m$ System

A multi-server system with finite capacity m.



If in the k-th state there are fewer clients than servers then the balance equations are as in the $M/M/\infty$ system, thus yielding

$$\pi_k = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \pi_0, \qquad \text{for all } 1 \leq k \leq m.$$

In state m, $\lambda\pi_{m-1} + m\mu\pi_{m+1} = m\mu\pi_m + \lambda\pi_m$ leads to:

$$\begin{aligned}
\pi_{m+1} &= \left(1 + \frac{\lambda}{m\mu}\right)\pi_m - \frac{\lambda}{m\mu}\pi_{m-1} \\
&= \left(1 + \frac{\lambda}{m\mu}\right)\frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \pi_0 - \frac{\lambda}{m\mu}\frac{1}{(m-1)!}\left(\frac{\lambda}{\mu}\right)^{m-1}\pi_0 \\
&= \left(1 + \frac{\lambda}{m\mu}\right)\frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \pi_0 - \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \pi_0 = \frac{\lambda}{m\mu}\frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \pi_0 = \frac{1}{m}\left(\frac{\lambda}{\mu}\right)^{m+1}\pi_0.
\end{aligned}$$

The $M/M/m$ System

In general, it holds that

$$\pi_k = \begin{cases} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \pi_0, & 1 \leq k \leq m, \\ \left(\frac{\lambda}{\mu}\right)^k \frac{1}{m!}\left(\frac{1}{m}\right)^{n-m} \pi_0, & k > m, \end{cases}$$

from which one obtains

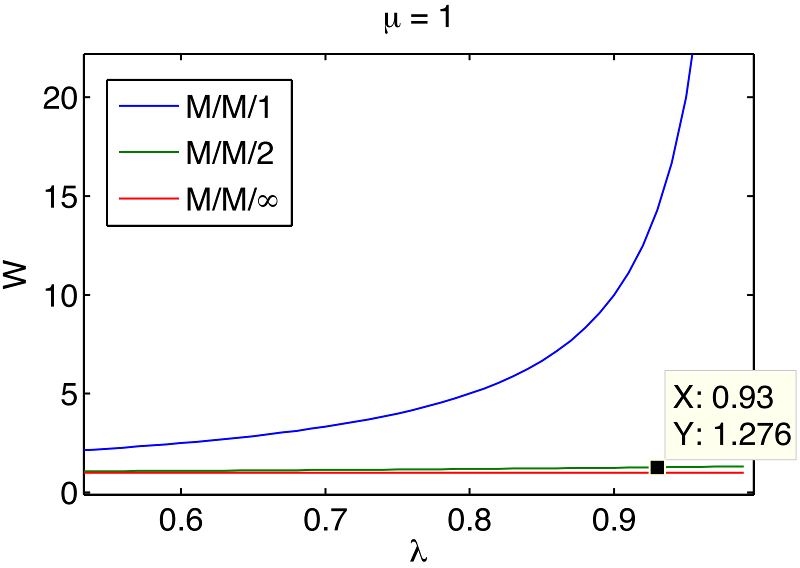
$$\pi_0 = \left[1 + \sum_{k=1}^{m-1} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k + \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \frac{1}{1-\lambda/(m\mu)}\right]^{-1}.$$

Performance Measures

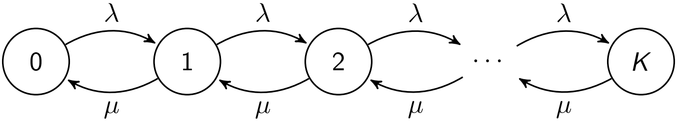
$$W = \left[\frac{(\lambda/\mu)^m \mu}{(m-1)!(m\mu-\lambda)^2}\right]\pi_0 + \frac{1}{\mu},$$

$$L = \left[\frac{(\lambda/\mu)^m \lambda\mu}{(m-1)!(m\mu-\lambda)^2}\right]\pi_0 + \frac{\lambda}{\mu}.$$

Comparison



Finite Capacity: The $M/M/1/K$ System



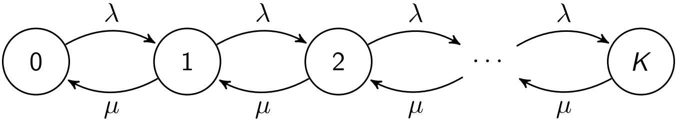
$$\lambda\pi_0 = \mu\pi_1,$$
$$\lambda\pi_k + \mu\pi_k = \lambda\pi_{k-1} + \mu\pi_{k+1}, \qquad 0 < k < K,$$
$$\lambda\pi_{K-1} = \mu\pi_K.$$

$$\pi_k = \left(\frac{\lambda}{\mu}\right)^k \pi_0, \qquad k \geq 0.$$

$$\pi_0 + \sum_{k=1}^{K} \pi_k = \pi_0 + \sum_{k=1}^{K} \left(\frac{\lambda}{\mu}\right)^k \pi_0 = \pi_0 \sum_{k=0}^{K} \left(\frac{\lambda}{\mu}\right)^k = 1$$

$$\Longrightarrow \quad \pi_0 = \left[\sum_{k=0}^{K} \left(\frac{\lambda}{\mu}\right)^k\right]^{-1} = \begin{cases} (1-\lambda/\mu)/(1-(\lambda/\mu)^{K+1}) & \text{if } \lambda \neq \mu, \\ 1/(K+1) & \text{if } \lambda = \mu. \end{cases}$$

Performance Measures for the $M/M/1/K$ System



From

$$\pi_0 = (1 - \lambda/\mu)/(1 - (\lambda/\mu)^{K+1}) \qquad \text{and} \quad \pi_k \ = \left(\frac{\lambda}{\mu}\right)^k \pi_0, \quad k \geq 0,$$

and setting $\rho = \lambda/\mu$,

$$\frac{1 - \pi_0}{1 - \pi_K} = \frac{1 - (1 - \rho)/(1 - \rho^{K+1})}{1 - (1 - \rho)\rho^K/(1 - \rho^{K+1})} = \frac{(1 - \rho^{K+1}) - (1 - \rho)}{(1 - \rho^{K+1}) - (1 - \rho)\rho^K}$$

$$= \frac{\rho - \rho^{K+1}}{1 - \rho^K} = \rho,$$

which yields the relationship

$\lambda(1 - \pi_K) = \mu(1 - \pi_0)$ [effective arrival rate = effective service rate].

Measures as Rewards

Consider the following function of a r.v. over the state space of the $M/M/1/K$ system:

$$X_a(k) = \begin{cases} \lambda & \text{, if } k \neq K, \\ 0 & \text{, if } k = K. \end{cases}$$

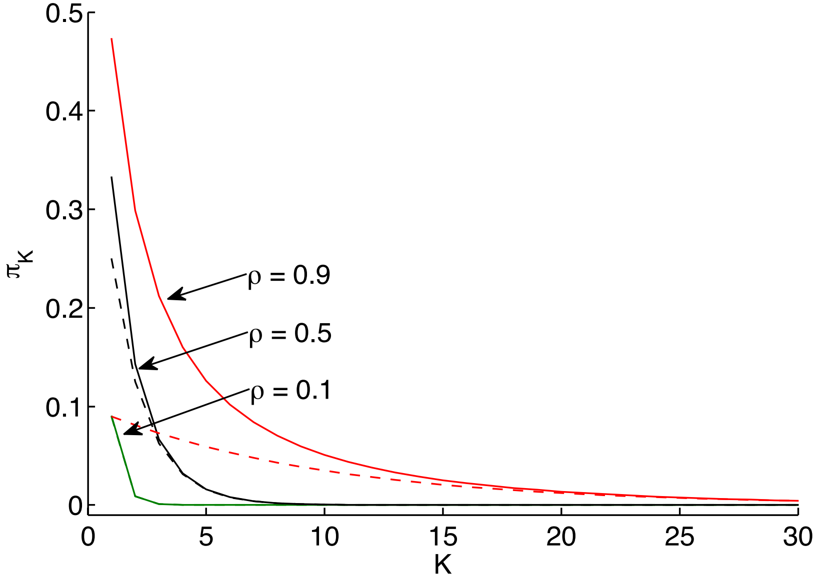
$$\mathbb{E}[X_a] = \sum_{k=0}^{K} X_a(k)\pi_k = \lambda \sum_{k=0}^{K-1} \pi_k = \lambda(1 - \pi_K).$$

Similarly, define

$$X_s(k) = \begin{cases} \mu & \text{, if } k \neq 0, \\ 0 & \text{, if } k = 0. \end{cases}$$

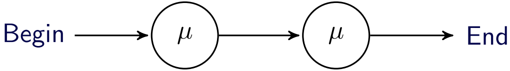
$$\mathbb{E}[X_s] = \sum_{k=0}^{S} X_s(k)\pi_k = \mu \sum_{k=1}^{K} \pi_k = \mu(1 - \pi_0)$$

Comparison



The Erlang Distribution

- Exponential service phases in tandem



- Each phase is defined by a r.v. Y with pdf

$$f_Y(y) = \mu e^{-\mu y}, \qquad y \geq 0.$$

- The total service time is given by $X = Y + Y$, which has pdf

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f_Y(y) f_Y(x-y) dy \\ &= \int_0^x \mu e^{-\mu y} \mu e^{-\mu(x-y)} dy \\ &= \mu^2 e^{-\mu x} \int_0^x dy = \mu^2 x e^{-\mu x}, \qquad x \geq 0. \end{aligned}$$

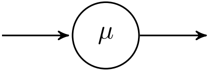
- X is called the Erlang-2 distribution (E_2).

Properties of the Erlang Distribution

- Mean and Variance

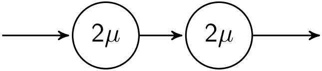
$$\mathbb{E}[X] = \int_0^{+\infty} x f_X(x) dx = \mu^2 \int_0^{+\infty} x^2 e^{-\mu x} dx = 2/\mu,$$
$$\mathrm{Var}[X] = 2/\mu^2.$$

- Compare now an exponentially distributed r.v. with rate μ and an Erlang distribution with phase 2μ.



Mean : $1/\mu$

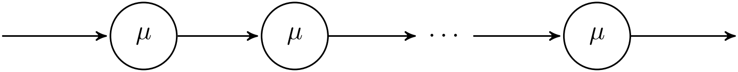
Variance : $1/\mu^2$



Mean : $1/\mu$

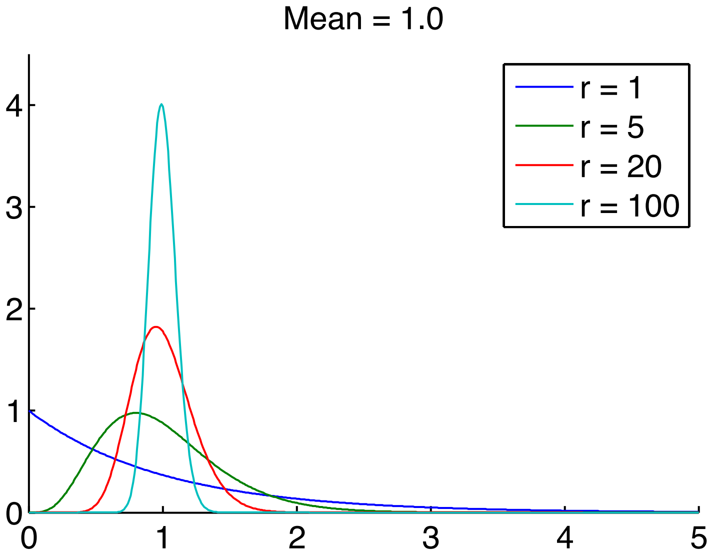
Variance : $1/2\mu^2$

- Generalisation: the Erlang-r distribution



Mean : r/μ Variance : r/μ^2

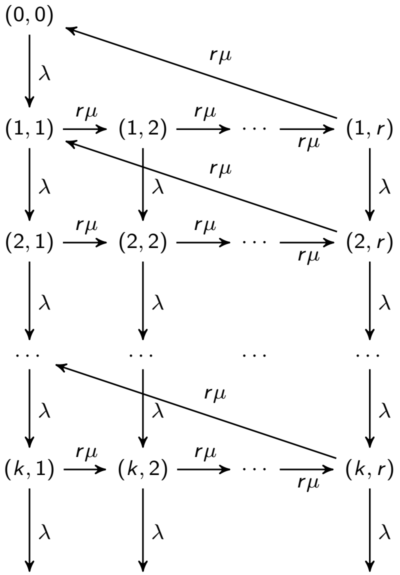
Pdfs of Some Erlang Distributions



The $M/E_r/1$ System

- Features:
  - Poisson arrivals;
  - Erlang-distributed service with r phases: if a customer is currently served at phase i, $1 \leq i \leq r$ then no other customer may be served in any of the other phases;
  - Single-server model.
- Model description: a state is described by the pair (k, i) where
  - k denotes the number of customer in the system, including the one in service;
  - i denotes the phase of service being received by the customer, $0 \leq i \leq r$; $i = 0$ when the queue is empty.

State Transition Diagram for the $M/E_r/1$ System



Analysis techniques:

1. Quasi-birth-death (QBD) form for the transition rate matrix[a]

$$Q = \begin{bmatrix} B_{00} & B_{01} & 0 & 0 & 0 & \cdots \\ B_{10} & A_1 & A_2 & 0 & 0 & \cdots \\ 0 & A_0 & A_1 & A_2 & 0 & \cdots \\ 0 & 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

2. Using z-transforms[b]

---

[a] W.J. Stewart. *Probability, Markov Chains, Queues, and Simulation*, 2009, Princeton University Press.

[b] Kleinrock, *Queueing Systems: Volume I — Theory*, Wiley Interscience, NY, 1975.