Detailed State Probability Distribution of Infinite Servers Queues with Phase-Type Distributed Service Times

Filas de servidor infinito con distribución de tiempos de servicio tipo fase

File de serveur infini avec distribution des temps de service de type phase

Filas de servidor infinito com distribuição de tempos de serviço tipo fase

Ivan Mura*

Fecha de recepción: 15 de abril de 2015
Fecha de aprobación: 10 de junio de 2015
Pp. 29-54.

* Ph.D. en Ingeniería Electrónica, Informática y de Telecomunicaciones, Universidad de Pisa. Magister en Ciencias de la Información, Universidad de Pisa y en Information Technology Project Management, George Washington University School of Business.

Abstract

We present a detailed analysis of the M/PH/ queue, which allows determining an analytic form for both the transient and the steady-state probability distribution of customers at the various phases of the service. The analysis is based on the correspondence that can be found between the stochastic process representing the number of customers in service at the various phases of the PH distribution, and a stochastic process that represents the evolution of the number of customers in the nodes of a Jackson's network where all service centers are M/M/ queues.

Keywords

Phase-Type distribution, queuing network, steady-state, transient, analytical solution, Petri Nets.

Resumen

Este documento presenta un análisis detallado de la cola M/PH/∞, la cual permite determinar de una forma analítica, tanto para un estado transitorio como para uno estacionario, la distribución de probabilidad de los clientes en las distintas fases del servicio. El análisis se basa en la correspondencia que se puede encontrar entre el proceso estocástico que representa el número de clientes en servicio en las diferentes fases de la distribución de PH, y en un proceso estocástico que representa la evolución del número de clientes en los nodos de una red Jackson en la que todos los centros de servicio son colas M/M/∞.

Palabras clave

Distribución de tipo fase, red de colas, estado estacionario, transitorio, solución analítica, Redes Petri.

Résumé

Ce document présente une analyse détallée de la formule M/PH/∞ permettant de déterminer la distribution des probabilités des clients de différentes phases du service de façon analytique, provisoire ou stationnaire. L'analyse se base sur la correspondance pouvant être démontrée entre le procesus estocastique révélant le nombre de clients en service lors des différentes phases de distribution de PH mais également lors des processus estocastiques représentant l'évolution du nombre de clients dans les noeuds d'un réseaun Jackson dans lequel tous les centres de service sont de type M/M/∞.

Mots clefs

Distribution de type phase, état stationnaire, transitoire, solution analytique, Réseaux Petri.

Resumo

Este documento apresenta uma análise detalhada de la fila M/PH/∞, a qual permite determinar de uma forma analítica, tanto para um estado transitório como para um estacionário, a distribuição de probabilidade dos clientes nas distintas fases do serviço. A análise se baseia na correspondência que se pode encontrar entre o processo estocástico que representa o número de clientes no serviço das diferentes fases de distribuição de PH, e em um processo estocástico que representa a evolução do número de clientes nos nós de uma red Jackson na que todos os centros de serviço são filas M/M/∞.

Palavras-chave:

Distribuição do tipo fase, rede de filas, estado estacionário, transitório, solução analítica, Redes Petri.

1. Introductión

Queuing network models are extremely useful modeling tools to represent parallelism, concurrency and competition, typical features of nowadays computing and telecommunication systems. At the same time, this model family is endowed with a set of analytic solution results (Baskett *et al.,* 1975), which makes them very attractive at the moment of identifying performance bottlenecks, dimensioning systems under a design that is based on estimates of the offered workload, and comparing different architectural options in terms of their achievable performances.

In this paper, we consider queuing networks where each service center is equipped with an infinite number of servers. This type of models represents systems where the amount of time that each customer spends in the system is random and independent of the number of other customers present in the system (i.e., there is no waiting time).

This kind of models is interesting for various application purposes. For instance, it represents pure delay systems, such as transportation systems (Mandayam & Prabhakar, 2014) or telecommunications links, or it may be used to approximate the expected performance figures of multi-server systems. A more recent area of application of this model type is computational biology (Schwabe *et al.*, 2012), where the reaction processing speed is very often proportional to the number of reactants (the so-called mass-action law).

In its simplest version, the infinite server queue has independent exponentially distributed inter-arrival times and exponential services times, and following Kendall's notation it is denoted as M/M/. This queue has an especially simple analytical solution. Explicit formulas for the transient distribution of the number of customers in the systems are available (Ellis, 2010)

and the steady-state distribution is known to be of Poisson type, of parameter, where is the interarrival rate and the service rate at each server.

Important extensions of this simple infinite server queue deal with the generalization of the service time distribution. When no assumptions are made about the form of the service time distribution, the model is denoted as an M/G/∞ queue, where G stands for general distribution. While the transient solution becomes in this general scenario much more complex, the steady-state distribution of the customers in the queue is still Poisson of parameter λ/μ, insensitive to the specific form of the service time distribution and only dependent on its average μ^{-1}, (Whitt, 2002).

We study here a generalization of the service time distribution to the class of phase-type distributions (O'Cinneade, 1990; 1999), commonly denoted as PH distributions (Neuts, 1975). A PH distributed random variable can be informally described as the time until absorption of a Markov process with exactly one absorbing state. Each of the states of the Markov process is one of the phases of the distribution. The relevance of PH distributions stems from the fact that they have been shown to be dense in the field of all positive-valued distributions having a continuous (apart from the single point 0) density function, which entails that for any positive-valued distribution of this kind there is a PH distribution that can approximate it within any given accuracy level (Nelson, 1995).

The wide generality of the result for the steady-state distribution of the number of customers in an M/G/∞ queue still ensures that the M/PH/∞ queue stationary distribution is Poisson of parameter λ/μ, where μ^{-1} is the average of the PH distributed service times. However, this result only applies to the total number of customers in the queue, disregarding the stage of the service. In this study, we are interested in knowing the detailed distribution of customers, (i.e., the probability

of finding i, i ≥ 0 customers in stage j of the PH service). Whereas an equivalence between M/PH/∞ and networks of M/M/∞ queues has been derived for the moments of queue length (Nelson and & Taaffe, 2004), the analytical study of the detailed transient and steady-state distribution is a task that, to the best of our knowledge, has not been approached yet.

To tackle this problem, we will resort to a decomposition of the M/PH/∞ queue into a network of M/M/∞ queues. By a simple inspection of the state-transition probabilities, we are able to prove that the M/PH/∞ queue is equivalent, in terms of the state probability distribution, to a Jackson's queuing network (Jackson, 1957). This result allows obtaining in an analytic form the detailed transient and steady-state distribution of the customers at the various phases of service in the queue.

The rest of the paper is organized as follows: we precisely define the queuing model we will be analyzing in this paper in section 2, and we provide our equivalence result in section 3. We then detail the steady-state solution of the queue in section 4, and in section 5 its transient solution for the detailed probability distribution. section 6 is devoted to the presentation of an application of the results. Finally, our concluding remarks are provided in section 7.

2. The model of interest

In this section, we precisely define the queuing model we will be dealing with. As for the notation, we shall be using the boldface letters to denote vectors and matrices.

We consider an infinite-server queue, fed by a Markovian arrival process of intensity λ. The service times are i.i.d. random variables, which we assume to be drawn from a PH

distribution. PH distributions, proposed by Neuts (1975), result from the convolution of negative exponential distributions. They generalize the exponential, Erlang, hypo and hyper-exponential as well as Coxian distributions. PH distributions can be characterized as the time spent in the states of an absorbing Markov process until absorption occurs. If the Markov process has states *n+1* states $s_0, s_1, \ldots, s_n$, with s_0 being the absorbing state, we say that the corresponding PH distribution has *n+1* phases. Notice that, in case multiple absorbing states are present in the Markov process, they can all be collapsed into a single one, without changing the PH distribution. Obviously, a PH distribution can be expressed in terms of the infinitesimal generator matrix *Q* of its defining Markov process. Matrix $Q = ||q_{i,j}||$ can be partitioned as follows:

$$Q = \begin{bmatrix} 0 & 0 \\ v & S \end{bmatrix} \quad (1)$$

where S is the n×n sub-generator matrix, which contains the transition rates between the non-absorbing states. Because the n-dimensional vector vv, which contains the rates of absorption into state S_0, must be equal to $-S \cdot e$, where *e* is a vector whose entries are all equal to 1, the service time distribution can be characterized by *S*, plus an *1×n* probability vector α, whose entry α_i assigns the probability that phase *i, i=1,2,…,n*, be the first phase. Notice that the $\alpha \cdot e$ not necessarily need to be 1, as in the general case there may be a non-null probability of completing the service in zero time[1].

The expected value of the PH distribution is given by $-\alpha \cdot S^{-1} \cdot e$ (Buchholz et al., Chapter 2), and therefore, by Little's law (Little, 1961), the steady state distribution of customers in the queue is Poisson of parameter $-\lambda\alpha \cdot S^{-1} \cdot e$. This steady-state distribution result is valid for the total number of customers in the queue, irrespective of the service phase. However, for any given number of customers, there exist is a combinatorial

number of states of the PH service that contribute to its mass probability. Therefore, it is of interest to determine the detailed distribution of the customers in the queue. This steady-state distribution result is valid for the total number of customers in the queue, irrespective of the service phase. However, for any given number of customers, there is a combinatorial number of states of the PH service that contribute to its mass probability. Therefore, it is of interest to determine the detailed distribution of the customers in the queue.

We shall now formally define the stochastic process representing the evolution of the number of customers in the queue, enclosing enough detail in the state definition so to account for the phase of the service. Since there is no waiting time in the queue, the state of a customer is easily described by the current phase of the service, and the global state of the queue by the collection of all the customer states. Therefore, a very intuitive representation of the system state is an -dimensional vector, whose *i*-th entry is the number of customer in the queue that are experiencing the *i*-th phase of the PH service, $i = 1,2,...,n$.

The state space V of the process is therefore defined as $V = \{ y \in N^n \}$, where N^n denotes the set of natural numbers (including 0), and the transitions outgoing a given state $y=(y_1,y_2,...,y_n)$ in V are as follows:

- At the arrival of a new customer, with rate $\lambda\alpha_i$, the process passes from y to $y+e_i$, *for* $i=1,2,\ldots,n$.
- Any customer in service at phase i will move with rate $s_{i,j}$ to phase *j* of the service, therefore, the process passes from y to $y - e_i + e_j$ with rate $y_i s_{i,j,}$ for $i,j=1,2,\ldots,n$, $i \neq j$.
- Any customer in service at phase *i* will leave the queue with rate v_i, therefore, the process passes from *y* to $y - e_i$, with rate $y_i v_i$, for $i = 1,2,\ldots,n,$. Where e_i is the *n*-dimensional

vector whose entries are all 0 apart for from the *i*-th, whose value is 1, $i = 1,2,\ldots,n$. This definition of the state space is sufficiently detailed enough as to allow tracking the phases of the service for all customers in the queue.

3. The equivalence result

In this section, we will define a Jackson's network whose state transition diagram is the same as that of the M/PH/∞ queue. This equivalence result will allow determining the performance indexes of the queue via the product-form solution of the equivalent network.

Consider a Jackson's network that consists of n M/M/∞ nodes. At node , there is a Poisson arrival process of parameter $\lambda\alpha_i$, $i = 1, 2, ..., n$. The service rate at node i is $\mu_i = - s_{i,i}$, $i=1,2, ..., n,$ the routing probability from node i to node j is $\delta_{i,j} = s_{i,j} / \mu_i$, $i , j = 1, 2, ..., n, i \neq j,$ while the probability of leaving the network from node i is ν_i/μ_i, $i = 1, 2, ..., n.$

We define the state space of the stochastic process underlying the Jackson's network as the set of non-negative n-dimensional vectors, where the i-th component of the state is the number of customers at the i-th infinite server queue, $i = 1, 2, ..., n.$ Therefore, the state space of the Jackson's network is the same as , the state space of the M/PH/∞ queue.

Let us now consider the transitions that lead to state changes in the Jackson's network.

- Arrivals: the arrival of a new customer to the network happens with a total rate of $\lambda\alpha - e$, and with rate $\lambda\alpha_i$ the model passes from state $\boldsymbol{y}$ to state $\boldsymbol{y} - \boldsymbol{e}_i$, when a new customer joins the i-th queue upon arrival, for $i = 1 ,2 , ..., n.$

- Change of node: any customer at node i completes service at rate μ_i and then may join node j, with probability $s_{i,j} = s_{i,j} / \mu_i$, which means that with rate $\boldsymbol{y}_i \mu_i (s_{i,j} / \mu_i) = \boldsymbol{y}_i s_{i,j}$, the model passes from state $\boldsymbol{y}$ to state $\boldsymbol{y} - \boldsymbol{e}_i + \boldsymbol{e}_i$, for $i = 1, 2, ..., n, i \neq j$.

- Departure from the network: any customer at node i that completes its service may leave the network with probability ν_i/μ_i, which means that with rate $\boldsymbol{y}_i \mu_i (\nu_i/\mu_i) = \boldsymbol{y}_i \nu_i$, the model passes from state $\boldsymbol{y}$ to state $\boldsymbol{y} - \boldsymbol{e}_i$, for $i = 1, 2, ..., n$.

Hence, the state transition diagram of the M/M/∞ and that of the Jackson's network are the same. The performance indexes of the M/M/∞ queue can therefore be computed via the solution of the product-form Jackson's network.

Let us notice that the equivalence we just demonstrated is also valid in the opposite direction. Given a Jackson's network of n M/M/∞ nodes, with arrival rate specified by:

- a vector $\lambda = (\lambda_1 , \lambda_2 , ..., \lambda_n)$;

- a vector of service rates $\mu ; (\mu_1, \mu_2 , ..., \mu_n)$

- an $n \times (n+1)$ matrix of routing probabilities $\Theta = || \theta_{i,j} ||$, where $\Theta_{i,0}$ is the probability that a customer would depart from te network after completing service at node $i, i = 1, 2, ..., n,$ and $\Theta_{i,j}$ is the probability of joining node j after completing service at node $i, i, j, = 1, 2, ..., n$.

Then, the Jackson's network is equivalent to an $M/PH/\infty$ with arrival rate $\lambda = \lambda - e$, and whose PH service distribution is characterized by a vector $\alpha = \lambda^{-1} (\lambda_1 , \lambda_2 , ..., \lambda_n)$, a vector $\nu = (\mu_1\theta_{1,0}, \mu_2\theta_{2,0}, ..., \mu_n\theta_{n,0})$ and finally a sub-generator matrix $s = || s_{i,j} / \mu_i || = \mu_i \theta_{i,j}$, $i, j = 1, 2, ..., n,$ with the diagonal term $s_{i,i} = - \nu_i - \Sigma_{j \neq i} s_{i,j}$, $i = 1, 2, ..., n$.

Notice that, in a Jackson's network, the routing probabilities $\theta_{i,i}$ may be in general greater than 0, $i = 1, 2, ..., n$,

when we are modeling multiple consecutive visits to the same node. When we apply our transformation of the network into the equivalent M/PH/∞ queue, the feedback is implicitly modeled via a reduction of the total outgoing transition rate from the corresponding phase of the PH service.

Moreover, it is important to remark that the equivalence can be extended to networks of M/PH/∞ queues, by mapping the transitions to the absorbing state of each PH service distribution into the corresponding transition rate between queues. Here, we do not treat this extension as if it would require additional notation to demonstrate a result that is quite obvious.

4. The steady-state probability distribution

The steady-state probability distribution of the M/PH/∞ queue is computed as the product of the marginal probability distributions of the nodes of the equivalent Jackson's network.

To determine the load of each M/M/∞ node in the network, we need to solve the network traffic balance equations (Jackson, 1957). For each node , we determine the total incoming rate of customers Λ_i, $i = 1, 2, ..., n$ by solving the following system of linear equations:

$$\Lambda = \lambda + \Lambda \bullet \Delta \Rightarrow \Lambda = \lambda \bullet (I - \Delta)^{-1} \qquad (2)$$

Notice that $I - \Delta$ in equation 2 is always non-singular, because it is an irreducibly diagonally dominant matrix.

For any state $y = (y_1, y_2, ..., y_n)$, let π_y denote the steady-state probability of finding the M/PH/∞ queue in state y. To simplify the notation, let us denote by ρ_i the expected steady-state number of customers in the i-th queue of the network, $\rho_i = \Lambda_i / \mu_i$, $i = 1, 2, ..., n$. Then, owing to the product-form of the state probability of the Jackson's network, π_y is given by the expression in the equation 3:

$$\pi_y = \prod_{i=1}^{n} e^{-\rho_i} \frac{(\rho_i)^{y_i}}{y_i!} \qquad (3)$$

Now, let us denote by μ the vector of service rates, and by $\bar{\mu}$ the vector of the reciprocals of service rates. Let us notice that $\rho = (\rho_1, \rho_2, ..., \rho_n)$ can be written as $\rho = \Lambda \bullet \Delta(\bar{\mu})$, where $D(x)$ is the diagonal matrix whose i-th diagonal element is equal to x_i, $i = 1, 2, ..., n$. Then, because matrix $I - \Delta$ can be written as $-S \bullet D(\bar{\mu})$, we can get the following expression for ρ in terms of the parameters that characterize the PH distribution:

$$\rho = \Lambda \cdot D(\bar{\mu}) = \lambda \cdot (I - \Delta)^{-1} \cdot D(\bar{\mu}) = \lambda(-S \cdot D(\bar{\mu}))^{-1} \cdot D(\bar{\mu}) = \\ = -\lambda \cdot S^{-1} \cdot D(\mu) \cdot D(\bar{\mu}) = -\lambda \cdot S^{-1} = -\lambda\alpha \cdot S^{-1} \qquad (4)$$

We can use the equality stated in equation 4 to simplify the expression of the steady-state probability in equation 3, as follows:

$$\pi_y = \prod_{i=1}^{n} e^{-\rho_i} \frac{(\rho_i)^{y_i}}{y_i!} = e^{-\rho e} \prod_{i=1}^{n} \frac{(\rho_i)^{y_i}}{y_i!} = e^{-\lambda\alpha S^{-1} e} \prod_{i=1}^{n} \frac{(\rho_i)^{y_i}}{y_i!} \qquad (5)$$

Since each ρ_i is the steady-state average number of customers receiving service at phase i, $i = 1, 2, ..., n$, $-\lambda\alpha \bullet S^{-1} \bullet e$ is the average number of customers in the M/PH/∞ queue. This is

the parameter of the steady-state Poisson distribution of the total number of customers in the queue, irrespective of the phase of the service, as already known from the general result for M/G/∞ queues.

To check the correctness of the mathematical treatment, let us consider the probability of finding exactly $m \geq 0$ users in the queue at steady-state, independently of the service state. According to the Poisson distribution of parameter $-\lambda\alpha \bullet S^{-1} \bullet e$, this probability is equal to:

$$e^{-\lambda \boldsymbol{\alpha} \boldsymbol{S}^{-1} \boldsymbol{e}} \frac{(-\lambda\alpha \cdot S^{-1} \cdot e)^m}{m!} \qquad (6)$$

From our equivalence result, we get that:

$$Prob[y \cdot e = m] = \sum_{i_1=0}^{m} \sum_{i_2=0}^{m-i_1} \cdots \sum_{i_{m-2}=0}^{m-\sum_{j=1}^{m-3} i_j} \sum_{i_{m-1}=0}^{m-\sum_{j=1}^{m-2} i_j} \pi_{(i_1,i_2,\ldots,i_{n-1},m-\sum_{j=1}^{m-1} i_j)} =$$

$$= e^{-\lambda \boldsymbol{\alpha} \boldsymbol{S}^{-1} \boldsymbol{e}} \sum_{i_1=0}^{m} \sum_{i_2=0}^{m-i_1} \cdots \sum_{i_{m-2}=0}^{m-\sum_{j=1}^{m-3} i_j} \sum_{i_{m-1}=0}^{m-\sum_{j=1}^{m-2} i_j} \frac{\rho_1^{i_1}}{i_1!} \frac{\rho_1^{i_2}}{i_2!} \cdots \frac{\rho_n^{m-\sum_{j=1}^{m-1} i_j}}{(m-\sum_{j=1}^{m-1} i_j)!} = \qquad (7)$$

$$= e^{-\lambda \boldsymbol{\alpha} \boldsymbol{S}^{-1} \boldsymbol{e}} \frac{(\rho_1 + \rho_2 + \cdots \rho_n)^m}{m!} = e^{-\lambda \boldsymbol{\alpha} \boldsymbol{S}^{-1} \boldsymbol{e}} \frac{(-\lambda\alpha \cdot S^{-1} \cdot e)^m}{m!},$$

Where the equivalence before the last comes from the multinomial expansion formula. Thus, the aggregation of the detailed probability distribution allows recovering the already known solution for the distribution of the total number of customers in equation 6, which is directly obtained from the Poisson distribution.

5. The transient state probability distribution

Consider an M/M/∞ queue, with λ being the intensity of the Poisson arrival process and μ the rate of the exponential service. It is known from the literature that ρ (t), the transient mean of the queue for $t \geq 0$, can be determined by solving the following linear differential equation:

$$\frac{d}{dt}\rho(t) = \lambda - \mu\rho(t) \quad (8)$$

Plus the initial condition of the queue at time $t = 0$, which we shall assume to be given by $\rho\ (0) = 0$. This result is for instance demonstrated in Eick *et al.* (1993) and can be easily re-obtained by writing down the Chapman-Kolmogorov forward equations for the transient probability of the states of the queue, multiplying by i, for each $i \geq 0$ the equation of state i, and summing up all the resulting equations.

In Harrison and Lemoine (1981), the authors show that there is a product-form solution for queuing networks composed by $n > 1 M/M/\infty$ queues, and that at any point in time $t \geq 0$, the marginal probability distribution of queue i is Poisson of parameter $\rho_i\ (t),\ i = 1,2, \ldots, n$ where $\rho_i\ (t)$ is the time dependent expected number of customers in the queue.

For the Jackson's network equivalent to the M/PH/∞ queue, let $\rho\ (t)$ be the vector of the time dependent average number of customers $\rho_i\ (t)$. As detailed in Boucherie and Taylor (1993), $\rho_i\ (t)$, $i = 1, 2, \ldots, n$ is the solution to the following linear differential equation:

$$\frac{d}{dt}\rho_i(t) = \lambda_i + \sum_{j=1}^{n} \rho_i(t)\delta_{j,i} - \rho_i(t)\mu_i \qquad (9)$$

The first positive term on the right hand side of equation 9 is the positive contribution to the expected number of customers that originates from the Poisson incoming flow, the second one is the positive contribution of the flows joining the queue after departure from the other nodes of the network, and the third one is the negative contribution from outgoing customers.

We rewrite the differential equations in equation 9 in matrix form, as follows:

$$\frac{d}{dt}\rho(t) = \lambda + \rho(t) \cdot \Delta \cdot D(\mu) - \rho(t) \cdot D(\mu) \qquad (10)$$

Where $\Delta \; || \, s_{i,j} \, ||$, subject to the initial condition ρ (0) = 0. We can further simplify equation 10 by expressing it in terms of the sub-generator matrix S as follows:

$$\frac{d}{dt}\rho(t) = \lambda + \rho(t) \cdot (\Delta - \mathrm{I}) \cdot D(\mu) = \lambda - \rho(t) \cdot S \qquad (11)$$

For $t \geq 0$, the solution to equation 11 is given by:

$$\rho(t) = \lambda \cdot S^{-1} \cdot (I - e^{-St}) \qquad (12)$$

The $n \times n$ matrix exponential e^{st} can be computed directly when the number of phases n is limited, or the solution of the differential equations obtained via numerical integration when n is large. For any state $y \in \mathcal{V}$, the transient probability distribution $\pi_y(t)$ at time $t \geq 0$ is given by the product of marginal transient distributions, as follows:

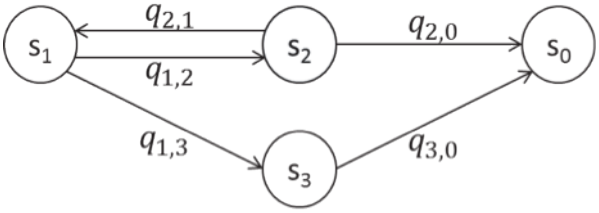
$$\pi_{\boldsymbol{y}}(t) = \prod_{i=1}^{n} e^{-\rho_i(t)} \frac{\rho_i(t)^{y_i}}{y_i!} = e^{-\boldsymbol{\rho}(t)\boldsymbol{e}} \prod_{i=1}^{n} \frac{\rho_i(t)^{y_i}}{y_i!} \quad (13)$$

The distribution $\pi_y(t)$ converges to π_y for $t \to \infty$ (Harrison & Lemoine, 1981).

6. Application example

In this section, we apply the results that have been proven in previous sections to the analysis of an example of an M/PH/∞ queue, fed by a Poisson process of parameter λ. We consider the service time distribution to be a 4-phases PH defined by the Markov process whose state transition diagram is shown in, where S_0 is the absorbing state, and we assume that $\alpha_1 = 1$, that is the PH service times always start in the phase represented by state S_1. The rate of transition from state S_i to state S_j is denoted by $q_{i,j}$, $i, j = 1, 2, 3, 4, i \neq j$.

Figure 1. State transition diagram of the example PH distribution.



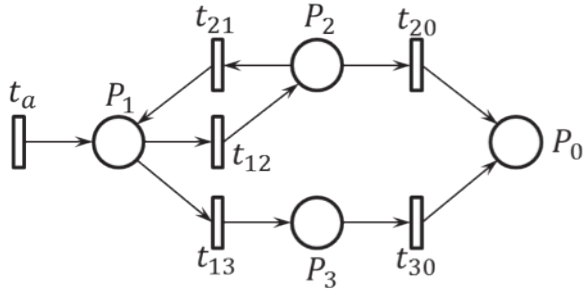
Source. By the author.

By ordering the states as S_0, S_1, S_2, S_3 the infinitesimal generator matrix can be partitioned as shown in equation14, and the sub-generator matrix S and vector ν are in this case as follows:

$$\boldsymbol{S} = \begin{bmatrix} -(q_{1,2}+q_{1,3}) & q_{1,2} & q_{1,3} \\ q_{2,1} & -(q_{2,1}+q_{2,0}) & 0 \\ 0 & 0 & -q_{3,0} \end{bmatrix}, \qquad \boldsymbol{v} = \begin{pmatrix} 0 \\ q_{2,0} \\ q_{3,0} \end{pmatrix} \quad (14)$$

To better exemplify the rationale and application of the proposed analysis approach, we build built a model of the M/PH/∞ queue as a Stochastic Petri Net (Ajmone-Marsan, 1990). The Stochastic Petri Net (SPN) shown in Figure 2 has exactly one place for each phase of the distribution and one transition for each of the possible phase transition events, plus transition t_α to model the arrival process. The correspondence between elements of the net and the M/PH/∞ is obvious and requires no further explanation. As for the rates assigned to transitions, the rate of t_a is equal to λ, and the rate of transition $t_{i,j}$ is given by the product of $q_{i,j}$ and the marking of the input place, to account for the infinite server-semantics of the service.

Figure 2. SPN model of the example M/PH/∞ queue.

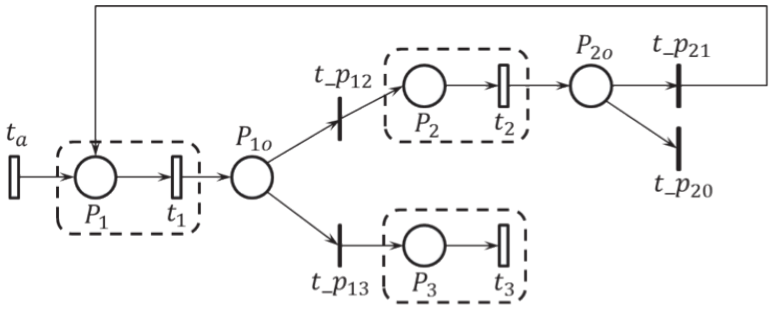


Source. By the author.

The marking of each place of the model in Figure 2 (i.e., the number of tokens contained therein) represents, at any time, the number of customers of the M/PH/∞ that are experiencing phase j of the service, j = 1, 2, 3. The number of tokens accumulated into place P_0 would represent the total number of customers served by the queue. Since this place would prevent the model from having a steady-state distribution, we simply remove it, a change that does not affect the service distribution time.

We make a simple transformation of the SPN in Figure 2 to convert it into an equivalent Generalized Stochastic Petri Net (Balbo, 2007), shown in Figure 3. In the Generalized Stochastic Petri Net (GSPN) model, the competitions among timed transitions are eliminated, and the routing among phases is now modeled by probabilistic choices of instantaneous transitions (thin bars).

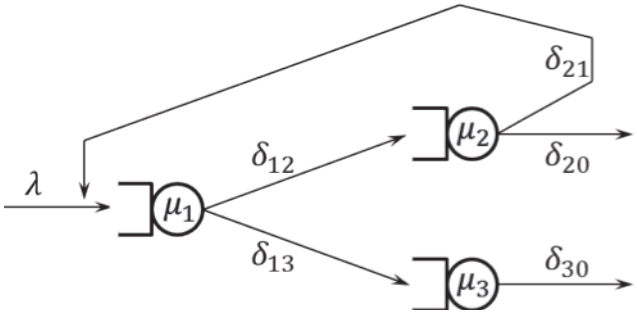
Figure 3. GSPN model of the example M/PH/∞ queue.



Source. By the author.

The conversion from the SPN to the GSPN version of the model nicely explains the rationale of the equivalence. It is easy to recognize, in the GSPN model shown in Source:, three queues, enclosed into the dashed rectangles. Therefore, we can convert the GSPN in the Jackson's network shown in Source:, where all service stations are simple M/M/∞ queues.

Figure 4. Jackson's network equivalent to the example M/PH/∞ queue.



Source. By the author.

The parameters that define the Jackson's network in Source: are obtained according to the steps detailed in section 3. The service rates μ_1 , μ_2 and μ_3 of the nodes, as per the equivalence result and the structure of the transition rates detailed in equation 14, are as follows:

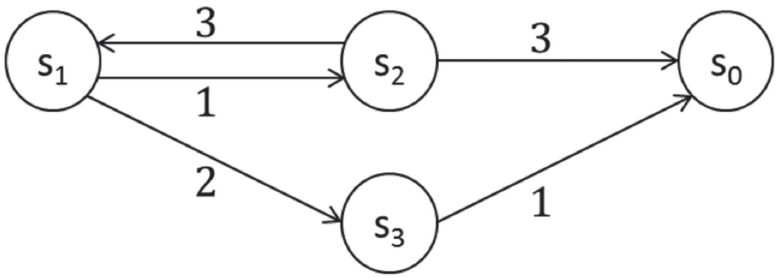
$$\mu_1 = -s_{1,1} = q_{1,2} + q_{1,3}, \mu_2 = -s_{2,2} = q_{2,1} + q_{2,0}, \mu_3 = -s_{3,3} = q_{3,0}, \quad (15)$$

While the routing probabilities are as follows:

$$\delta_{12} = q_{1,2}/\mu_1, \delta_{13} = q_{1,3}/\mu_1, \delta_{21} = q_{2,1}/\mu_2, \delta_{20} = v_2/\mu_2, \delta_{30} = v_3/\mu_3 = 1 \quad (16)$$

The Jackson's network in Figure 4 is straightforwardly solved in terms of the steady-state customer distribution. To make a concrete example, let us assume that the rates of transition in the Markov process that define the PH distributions are as shown in Figure 5, which corresponds to the pair *S* and ν shown in equation 15.

Figure 5. Absorbing Markov process defining an instance of the PH distribution.



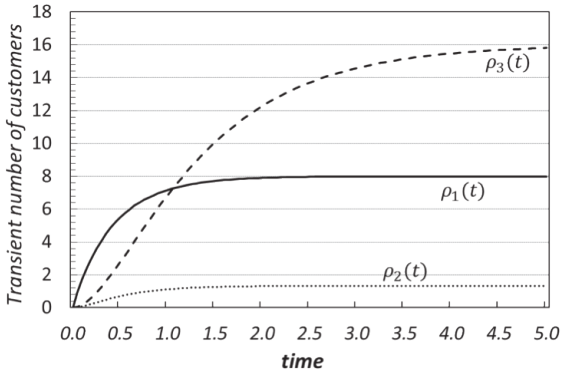
Source. By the author.

$$\boldsymbol{S} = \begin{bmatrix} -3 & 1 & 2 \\ 3 & -6 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \qquad \boldsymbol{v} = \begin{pmatrix} 0 \\ 3 \\ 1 \end{pmatrix} \quad (15)$$

Then, setting $\lambda = 20$, the average number of users in each of the stages of the PH distribution is computed by equation 4, and is given by vector $\rho = - \lambda\alpha \bullet S^{-1}$, where α (1,0,0) because we are assuming that all services start at phase 1. The expected number of customers at each stage of the service is ρ = (8.0,1.3, 16.o). By summing the three entries of vector ρ, we get the total number of users in the queue at steady-state, 25.3, which is the same result that would be obtained from the M/G/∞ equilibrium formula. From our detailed result, we can however determine the joint probability of having any number of customers in the different stages of the service, as the product of the marginal distributions of the queues, which are all Poisson, with parameters given by the entries of vector ρ.

The transient solution of the queue requires the solution of the set of linear differential equations defined in equation 10. By computing the matrix exponential with Maxima (http://maxima.sourceforge.net/) and by using the formula in equation 12, we get the solution for $\rho\ (t)$, the vector of the transient average number of users, for which we show a chart over time in., assuming the queue is empty at time $t = 0$.

Figure 6. Plot of the transient average number of customers at the PH service phases.



Source. By author.

At any point in time, vector provides the parameters of the Poisson distributions that characterize the marginal transient state probability distributions of the queues in the Jackson's network.

7. Conclusions

In this paper, we presented an approach to the detailed analysis of the transient and steady-state probability distribution for the M/PH/∞ queue. We provided a simple equivalence result assuring that the detailed state probability distribution of the number of customers at the various phases of the service has product form, and the terms of the product are computed from the state occupancy probability of a Jackson's network that only consists of M/M/∞ nodes.

We provided the explicit expression of the steady-state and transient detailed probability distribution in terms of the initial phase probability assignment at the beginning of the PH distributed service and of the PH transition matrix. The computational cost of the solution method we proposed is determined by the one required for the analysis of the equivalent Jackson's network, specifically by the solution of the traffic balance equations of the network for the steady-state, and the computation of a matrix exponential for the transient probability distribution. Finally, we remark that the equivalence result we presented in this paper is directly applicable to open or closed networks of M/PH/ queues.

8. References

Ajmone-Marsan, M. (1990). Stochastic Petri nets: An elementary introduction. *Lecture Notes in Computer Science*, 424, 1-29.

Balbo, G. (2007). Introduction to Generalized Stochastic Petri Nets. *Lecture Notes in Computer Science*, 4486, 83-131.

Baskett, F., Chandy, K., Muntz, R., & Palacios, F. (1975). Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. *Journal of the ACM*, 22(2), 248-260.

Boucherie, R., & Taylor, P. (1993). Transient product from distributions in queueing networks. *Discrete Event Dynamic Systems: Theory and Applications*, 3, 375-396.

Buchholz, P., Kriege, J., & Felko, I. (2014). Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications. *SpringerBriefs in Mathematics.* doi: 10.1007/978-3-319-06674-5__2.

Eick, S., Massey, W., & Whitt, W. (1993). The Physics of the $M_t/G/\infty$ Queue. *Operations Research*, 41(4), 731-742.

Ellis, P. (2010). The Time-Dependent Mean and Variance of the Non-Stationary Markovian Infinite Server Systems. *Journal of Mathematics and Statistics*, 6(1), 68-71.

Harrison, J., & Lemoine, A. (1981). A Note on Networks of Infinite-Server Queues. *Journal of Applied Probability,* 18(2), 561-567.

Jackson, J.R. (1957). Networks of Waiting Lines. *Operations Research*, 5(4), 518–521. doi:10.1287/opre.5.4.518.

Little, J.D.C. (1961). A Proof for the Queuing Formula: L = W. *Operations Research*, 9(3), 383–387. doi:10.1287/opre.9.3.383.

Mandayam, C., & Prabhakar, B. (2014). Traffic Congestion: Models, Cost and Optimal Transport. *In The 2014 ACM international conference on Measurement and modeling of computer systems.* SIGMETRICS '14, ACM Press, New York, USA, 553-554.

Nelson, B. L., & Taaffe, M. R. (2004). The $PH_t/PH_t/$ queueing systems: part I - the single node. *INFORMS Journal on Computing*. 16(3), 266-274.

Nelson, R. (1995). *Probability, Stochastic Processes, and Queueing Theory.* New York, NY: Springer-Verlag. doi: 10.1007/978-1-4757-2426-4

Neuts, M.F. (1975). Probability distributions of phase type. *In Liber amicorum Professor Emeritus H. Florin, Department of Mathematics.* Belgium: University of Louvain, 173-206.

O'Cinneide, C.A. (1990). Characterization of phase-type distributions. *Communications in Statistics: Stochastic Models,* 6(1), 1-57.

O'Cinneide, C.A. (1999). Phase-type distributions: open problems and a few properties. *Communications in Statistics: Stochastic Models*, 15(4), 731-757.

Schwabe, A., Rybakova, K., & Bruggeman, F. (2012). Transcription stochasticity of complex gene regulation models. *Biophysical Journal,* 103(6), 1152-1161.

Whitt, W. (2002). Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues. *Springer Series in Operations Research and Financial Engineering*. New York, NY: Springer-Verlag.