En español

Efecto del uso de tiempos de atención *heavy-tailed* sobre el modelo básico de líneas de espera y sus medidas de desempeño

Lina M. Rangel Martínez[1] y Jorge A. Alvarado valencia[2]

RESUMEN

La reciente aparición de modelos generatrices de líneas de espera con tiempos de atención *heavy-tailed* y su comprobación empírica implican la necesidad de conocer el comportamiento de las medidas clásicas de desempeño de una línea de espera bajo estas condiciones. El objetivo del estudio fue el de analizar el comportamiento de L_q (longitud promedio de la fila) y W_q (tiempo promedio de espera en fila) variando los parámetros capacidad del sistema, nivel de utilización promedio (ρ) y número de servidores para líneas de espera con tiempos de atención *heavy-tailed*, y contrastar dicho comportamiento con los resultados clásicos basados en procesos de Poisson, usando para ello la simulación de eventos discretos. Los resultados mostraron que la sensibilidad de los modelos con tiempos de atención *heavy-tailed* a variaciones en los parámetros es mayor que la de los modelos basados en procesos de Poisson. En particular, a partir de capacidades de sistema de 1.000 entidades ciertos procesos *heay-tailed* pueden considerarse infinitos, y la importancia del número de servidores es mayor en los procesos *heavy-tailed* analizados que en los procesos de Poisson. Por último, la utilización de L_q y W_q como medidas de desempeño es inadecuada para tiempos de atención *heavy-tailed* al generar resultados inestables y contraintuitivos.

Palabras clave: *líneas de espera, distribuciones* heavy-tailed*, tiempos de servicio, distribución de Pareto, modelos generatrices*.

Recibido: junio 26 de 2009
Aceptado: julio 10 de 2010

Introducción

La evidencia empírica reciente sugiere que algunas de las distribuciones de probabilidad de tiempos de servicio en líneas de espera no presentan colas exponenciales, y por tanto, obtener valores extremos en tiempos de servicio deja de ser algo improbable (Stidham, 2002).

Ejemplos empíricos de la aparición de este tipo de distribuciones son: el procesamiento y respuesta de correo electrónico (Barabási, 2005), el tiempo de atención en un taller automotriz (Alvarado *et al*.,2008) y el tiempo de transmisión de un archivo en telecomunicaciones (Mitzenmacher, 2004).

[1] Ingeniera Industrial. Profesora Catedrática, Facultad de Ingeniería, Departamento de Ingeniería Industrial, Pontificia Universidad Javeriana, Bogotá, Colombia. lrangel@javeriana.edu.co.
[2] Ingeniero Industrial. M. Sc., en Analytics. Profesor Asistente, Facultad de Ingeniería, Departamento de Ingeniería Industrial, Pontificia Universidad Javeriana, Bogotá, Colombia. jorge.alvarado@javeriana.edu.co.

In English

The consequences of heavy-tailed service time distribution on a basic queuing model and its performance indicators

Lina Rangel Martínez[3] and Jorge A. Alvarado valencia[4]

ABSTRACT

Recent research showing theoretical generative models for heavy-tailed service time queues and its empirical validation implies the need for a better knowledge of the key performance indicators' behaviour under such assumption. The behaviour of the average length of the queue (L_q) and the average waiting-time (W_q) were analysed through simulation, varying system capacity, average service utilisation factor (ρ) and the number of servers in the systems as parameters. Comparisons were also made with service times based on Poisson processes. The results showed more sensitive variations of L_q and W_q for heavy-tailed service times than for Poisson-based service times. Systems having a capacity of over 1,000 entities might be considered as being systems having infinity capacity and the number of servers has a greater importance in heavy-tailed ruled processes than in Poisson processes. There was a lack of adequacy of L_q and W_q as key performance indicators for heavy-tailed service times, leading to unexpected and unstable results.

Keywords: *queuing system, heavy-tailed distribution, service time, Pareto distribution, generative model*.

Received: jun 26th 2009
Accepted: jun 10th 2010

Introduction

Recent empirical evidence has suggested that some of the probability distributions in service time queues do not exhibit exponential tails; therefore, attaining extreme values for service times ceases to be unlikely (Stidham, 2002).

Some empirical examples of the appearance of this type of distribution are processing of and replying to e-mails (Barabási, 2005), waiting time at an automotive repair shop (Alvarado *et al*., 2008) and the transmittal time of a file in telecommunications (Mitzenmacher, 2004)

[3] Industrial Engineer. Professor, School of engineering, Departament of industrial engineering, Pontificia Universidad Javeriana, Bogotá, Colombia. lrangel@javeriana.edu.co.
[4] Industrial Engineer. M. Sc., en Analytics. Assistant Professor, Departament of industrial engineering, Departament of industrial engineering, Pontificia Universidad Javeriana, Bogotá, Colombia. jorge.alvarado@javeriana.edu.co.

En español

Dos modelos teóricos han sido propuestos para explicar esta situación: el primero sugiere que si el tamaño del trabajo a procesar que llega a la fila es una variable aleatoria con colas no exponenciales (tal es el caso de los archivos en telecomunicaciones), el tiempo de procesamiento de ese trabajo tendrá como consecuencia colas no exponenciales (Willinger, 1995); el segundo plantea que cuando el servidor es un ser humano que debe utilizar su propio criterio para decidir cuál trabajo procesar primero, y no hay seres humanos físicamente en la fila, se genera un fenómeno denominado prioridades percibidas, que lleva a postergar durante largos tiempos el procesamiento de un trabajo que ya ha ingresado al servidor y por consiguiente los tiempos de atención tendrán colas no exponenciales (Barabási, 2005; Alvarado *et al*., 2008). De lo anterior se deduce la existencia de un importante conjunto de fenómenos que podría ser modelado con más precisión utilizando distribuciones con colas no exponenciales.

Las distribuciones con colas no exponenciales suelen ser denominadas distribuciones *heavy-tailed*, si bien el uso de esta expresión ha sido algo confuso en la literatura (Embrechts *et al*., 1997; Mitzenmacher, 2004). Este artículo se ceñirá a la siguiente definición, que expresa de manera clara la naturaleza no exponencial de la cola: dada una variable aleatoria no negativa X, su función de distribución acumulada $F(x)=P(X\le x)$ y su función acumulada complementaria, o cola, $\overline{F}(x)=1-F(x)=P(X>x)$, una función de distribución para la variable aleatoria *X*, es denominada *heavy-tailed* si $\overline{F}(x)>0$,y $\underset{x\to\infty}{Lim}\ P(X>x+y\mid X>x)=\underset{x\to\infty}{Lim}\ \frac{\overline{F}(x+y)}{\overline{F}(x)}=1$ para $y\ge 0$ (Sigman, 1999, pp. 261-262). En contraste, una cola exponencial presenta el comportamiento $\lim_{x\to\infty}\frac{\bar{F}(x+y)}{\bar{F}(x)}=c*e^{-tx}$, siendo c y t constantes.

Entre el conjunto de las distribuciones *heavy-tailed* son ampliamente conocidas las distribuciones de Weibull (en análisis de confiabilidad) y de Pareto (en análisis financiero).

Uno de los mayores desafíos que presentan la mayoría de las distribuciones *heavy-tailed* para el análisis de líneas de espera es la no convergencia de la media y la varianza. Sin el cumplimiento del supuesto de media y varianza finitas, los resultados clásicos de la teoría de colas para modelos M/G/1/FIFO/ ∞ / ∞ de Cohen (Cohen, 1973, pp. 343-353) y Pakes (Pakes, 1975, pp. 555-564) no son válidos. Dada la dificultad matemática de las distribuciones *heavy-tailed*, no se conocen soluciones cerradas para líneas de espera con tiempos de atención *heavy-tailed*, y sólo hasta años recientes Whitt (Whitt, 2000, pp. 71-87) probó que si el tiempo de servicio tiene una distribución de probabilidad *heavy-tailed*, la longitud de la fila y el tiempo de espera tendrán también una distribución *heavy-tailed,* con cotas superior e inferior para estos valores bajo ciertos parámetros, resultado válido sólo para los sistemas M/G/1/FIFO/ ∞ / ∞ .Basados en estos antecedentes, la investigación tuvo como objetivo simular numéricamente el comportamiento de un modelo de líneas de espera donde el tiempo de atención esté regido por una distribución *heavy-tailed* bajo diversas condiciones de capacidad del sistema, utilización y número de servidores para analizar las variaciones de sus medidas de desempeño y la viabilidad misma de estas medidas (longitud promedio de la fila L_q y tiempo promedio de espera W_q).

In English

Two theoretical models have been proposed for explaining this case. The first has suggested that the size of the job in the queue to be processed is a random variable with no exponential tail (such is the case of telecommunication files), the processing time of this job will consequently have no exponential tails (Willinger,1995). The second proposes that a phenomenon called perceived priorities is created when the server is a human being using his/her personal criteria to select the order of job completion (and there are no human beings physically in queue). This leads to the lengthy postponement of the processing time of a job which has already reached the server and, consequently, services times will have no exponential tails (Barabási, 2005; Alvarado *et al*., 2008). Deduced from the foregoing is the existence of an important set of phenomena which could be more accurately modelled by distributions having non-exponential tails.

Distributions having non-exponential tails are often called heavy-tailed, though the term has been somewhat confused in the literature (Embrechts *et al*., 1997; Mitzenmacher, 2004). This article will adhere to the following definition which clearly expresses the non-exponential nature of the tail. Given a non-negative, random variable X and its cumulative distribution function $F(x)=P(X\le x)$ and its cumulative complimentary function, or tail, $\overline{F}(x)=1-F(x)=P(X>x)$, a distribution function for random variable *X* is called *heavy-tailed* if $\overline{F}(x)>0$, and $\underset{x\to\infty}{Lim}\ P(X>x+y\mid X>x)=\underset{x\to\infty}{Lim}\ \frac{\overline{F}(x+y)}{\overline{F}(x)}=1$ for $y\ge 0$ (Sigman, 1999, pp. 261-262). In dissimilarity, the exponential tail's behaviour is $\lim_{x\to\infty}\frac{\bar{F}(x+y)}{\bar{F}(x)}=c*e^{-tx}$, c and t being constants.

Weibull distributions (in reliability engineering) and Pareto (in financial analysis) are extensively recognised in heavy-tailed distribution sets.

One of the major challenges for most heavy-tailed distributions in queuing analysis is the non convergence of the median–variance framework. Without the fulfilling the median values and finite variance assumptions, the standard tail theory results for Cohen's M/G/1/FIFO/ ∞ / ∞ model (Cohen, 1973, pp. 343-353) and Pakes (Pakes, 1975, pp. 555-564) are not valid. Given the mathematical difficulty of heavy-tailed distributions, there are no known closed-form solutions for queuing models having heavy-tailed service times. Not until recent years did Whitt (Whitt, 2000, pp. 71-87) demonstrate that if a service time has a heavy-tailed distribution probability, then the length of the queue and waiting time will also have a heavy-tailed distribution, with higher or lower datum for these values under certain parameters. The result is only valid for M/G/1/FIFO/ ∞ / ∞ systems. Based on this case history, this research was aimed at numerically simulating the behaviour of a queuing model where service time is controlled by heavy-tailed distribution in diverse system capacity conditions, use and number of servers to examine the variation of the values of its performance indicators and the viability of these measurements (average length of the queue L_q and average waiting-time W_q).

En español

Para facilitar el logro de este objetivo los resultados se compararon con los obtenidos para el modelo clásico de líneas de espera con tiempos de atención basados en procesos de Poisson.

Desarrollo experimental

El experimento utilizó simulación de eventos discretos, que es una buena alternativa cuando no se conocen aproximaciones o resultados cerrados para una línea de espera y se desea realizar un análisis exploratorio del modelo bajo rangos significativos de sus parámetros (Neuts, 1998; Stewart, 1994; Gross, 2009), como es el caso aquí analizado. La evaluación sistemática de variaciones en los factores que influyen en un fenómeno evaluados en un experimento controlado —como es la simulación— se suele llevar a cabo de manera eficiente mediante un diseño experimental estadístico (Kuehl, 2001).

Diseño experimental

Entre los factores que la teoría clásica de líneas de espera considera fundamentales en el análisis se escogieron para este trabajo la distribución de los tiempos de servicio, la capacidad del sistema, el nivel de utilización y la cantidad de servidores. La disciplina de la fila no se tuvo en cuenta por considerarse que estaba implícita en el modelo de prioridades percibidas, explicado en la introducción para tiempos de atención *heavy-tailed*. El número de clientes potenciales del sistema se mantuvo infinito, pues hacerlo finito tan sólo lograría limitar y enmascarar el efecto que la distribución *heavy-tailed* tendría sobre la longitud de la fila y el tiempo de espera. En cuanto a la distribución de probabilidad de los tiempos de llegada, se mantuvo la utilización de un proceso de Poisson para hacer más comparables los resultados con las ecuaciones cerradas de la teoría clásica de líneas de espera. Los factores y niveles definidos generaron un modelo mixto con tres factores fijos o constantes y uno aleatorio (los parámetros de las distribuciones), cuya combinación daba lugar a 1.575 posibles tratamientos. Se decidió realizar un experimento factorial completo dado que la capacidad computacional no era una limitación y ello permitía el análisis exploratorio más amplio posible.

Parámetros de las distribuciones de probabilidad de los tiempos de servicio

Se seleccionó como paradigmática de las distribuciones *heavy-tailed* la distribución de Pareto por ser una de las distribuciones de esta clase más conocidas y utilizadas, y por tener la importante propiedad de que dependiendo de los valores de sus parámetros, su media y su varianza pueden o no converger, lo cual permite una exploración más amplia de los efectos de las distribuciones *heavy-tailed* (Andriani y McKelvey, 2005, pp. 219-223). La distribución Pareto tiene la siguiente distribución de probabilidad:

$$f(x) = k * \frac{x_{min}^{k}}{x^{k+1}} \text{ para } x \geq x_{min}; x_{min}, k > 0$$

Siendo $x_{\min}$ un parámetro de posición (mínimo de la función) y k el parámetro de forma, cuyo valor determina si tanto la media como la varianza convergen: en el rango $(0,1]$ tanto la media como la varianza divergen, en $(1,2]$ la media diverge pero la varianza no, y en $(2,\infty)$ ambos valores convergen (Mitzenmacher, 2004, pp. 228).

Así mismo, cuanto mayor sea el valor de k se tendrán densidades de Pareto más concentradas en las proximidades del mínimo, es decir, densidades menos dispersas (Newman, 2005, pp. 325-327).

In English

To facilitate this, the results were evaluated against those obtained for the classic queuing model with service times based on Poisson processes.

Experimental development

This experiment simulated discreet events, which is a viable alternative when an exploratory analysis of the model is desired in the absence of approximations or closed results for a queuing system (Neuts, 1998; Stewart, 1994; Gross, 2009). The systematic evaluation of factor variations influencing a pheno-menon, evaluated in a controlled experiment (i.e. simulation), is efficiently completed through a statistical experimental design (Kuehl, 2001).

Experimental design

Among the analysis factors which traditional queuing theories consider fundamental, the distribution of service-times, system capabilities, utilisation level and server quantity were selected for this study. The queuing discipline was not considered since it was deemed implicit in the model of perceived priorities explained in the introduction regarding heavy-tailed service times. The number of potential clients the system might have was left as an infinite value since making it finite would only limit and mask the effect that heavy-tailed distribution would have on the length of the queue and waiting time. Regarding arrival time probability distribution, a Poisson process was used to increase the comparability of the results with the closed equations of traditional queuing theories. The factors and levels defined generated a mixed model containing three permanent, constant factors and one random one (the distribution parameters); this combination allowed for 1,575 possible treatments. Given that computational capabilities were not a limitation, a complete factorial experiment was performed; this allowed for the broadest possible exploratory analysis.

Service-time probability distribution parameters

Because it is most often used and is the most popular of distributions, and because its mean and its variance may or may not converge depending on the values of its parameters, the Pareto distribution was selected as exemplary of heavy-tailed distributions. This very important property allowed for a broader exploration of the effects of heavy-tailed distributions (Andriani and McKelvey, 2005, pp. 219-223). The Pareto distribution has the following probability distribution:

$$f(x) = k * \frac{x_{min}^{k}}{x^{k+1}} \text{ being } x \geq x_{min}; x_{min}, k > 0$$

$x_{\min}$ being a location parameter (minimum to the function) and k the shape parameter. Its value determines whether both the median and the variance converge; within the $(0,1]$ range both the median and the variance diverge, in the $(1,2]$ the median diverges but not the variance and in the $(2,\infty)$ both values converge (Mitzenmacher, 2004, pp. 228).

Equally, the higher the value of k, the higher the Pareto densities will be around the minimum, that is, less disperse densities (Newman, 2005, pp. 325-327)

En español

La función de probabilidad cumple con la condición de colas no exponenciales, puesto que su cola se rige por la expresión

$$\Pr(X > x) = \left(\frac{x}{x_m}\right)^k$$

El interés se centró en determinar si existían variaciones importantes de L_q y W_q asociadas a la convergencia de la media y la varianza de una distribución *heavy-tailed*, por lo que se generaban tres intervalos de interés en la distribución Pareto, en los cuales se seleccionó aleatoriamente un punto que representara cada uno de los niveles de interés de la investigación, como se explica a continuación.

Primero, se estableció una relación entre la distribución exponencial (proceso de Poisson) y la distribución Pareto de manera tal que alguna de sus medidas de tendencia central —diferente de la media— fuese equivalente. Se decidió utilizar la mediana por ser una medida de tendencia central robusta. Siendo la mediana de la distribución exponencial $\beta * \ln(2)$ (Ross, 2006, p. 419) y la mediana de la distribución Pareto $x_{\min}\sqrt[k]{2}$ (Janicki y Simpson, 2005, p. 294), se realizó la siguiente equivalencia:

$$\beta * \ln(2) = x_{\min}\sqrt[k]{2}$$

De donde se obtiene la ecuación de equivalencia:

$$k = \ln\left(\frac{2 * x_{min}}{\beta * \ln(2)}\right)$$

En segundo lugar, se estableció una tasa media de servicio fija para todo el sistema μ con valor unitario, por lo que $\beta = \frac{1}{\mu} = 1$

En tercer lugar, y a partir de los rangos de convergencia de la distribución Pareto y la ecuación de equivalencia, se formaron rangos de convergencia basados en el parámetro x_{min}.

Por último, se generó un número aleatorio dentro de cada rango de convergencia de x_{min} y se establecieron estos tres valores como niveles aleatorios para el factor parámetro de la distribución del tiempo de servicio, obteniendo los valores para *k* a partir de la ecuación de equivalencia.

Número de servidores

Se realizaron variaciones en el rango de 1 a 15 servidores, tomando como referencia la propuesta mostrada en el texto bibliográfico de Hillier y Lieberman para modelos sin entradas Poisson (Hillier y Lieberman, 2005, pp. 802). De acuerdo con ello, los niveles seleccionados fueron: 1, 2, 3, 5, 7, 10 y 15 servidores.

Capacidad del sistema

La capacidad del sistema puede ser considerada como finita o infinita. Para definir los niveles del factor se hizo una transferencia de metodologías utilizadas en otras ciencias, como la neurociencia y la física, cuyos ámbitos de dominio y escalas de observación varían de acuerdo con potencias de 10. De esta manera, los niveles para la capacidad finita variaron desde 10 entidades (10^1) para un sistema pequeño, hasta 10.000 entidades (10^4). Se consideró que potencias superiores a cinco entrarían, para efectos prácticos, dentro de la categoría de servicios sin capacidad limitada.

Factor promedio de utilización del servicio (ρ)

Los valores de este factor se variaron en el rango (0, 1) para abrir la posibilidad de alcanzar el estado estable.

In English

The probability function fulfils the non-exponential tail condition, since its tail is regimented by the following expression

$$\Pr(X > x) = \left(\frac{x}{x_m}\right)^k$$

The focus was centred on determining whether important L_q and W_q variations existed which were associated with median and variance conversions of a heavy-tailed distribution. Three interested intervals were generated in the Pareto distribution; from each, a point representative of each of the investigation's levels of interest was randomly selected, as explained below.

A relationship was established between exponential distribution (Poisson process) and the Pareto distribution so that one of its main trend measurements (other than the median) was equal. The mean was used as it is a strong main trend measurement.

Exponential distribution mean being $\beta * \ln(2)$ (Ross, 2006, pp. 419) and Pareto distribution mean $x_{\min}\sqrt[k]{2}$ (Janicki and Simpson, 2005, pp. 294), the following comparison was made:

$$\beta * \ln(2) = x_{\min}\sqrt[k]{2}$$

From where the following comparison equation was obtained:

$$k = \ln\left(\frac{2 * x_{min}}{\beta * \ln(2)}\right)$$

Secondly, a fixed average service rate having unitary value μ was established for the entire system, therefore $\beta = \frac{1}{\mu} = 1$

Thirdly, parameters were established from Pareto distribution convergence ranges and the comparison equation convergence ranges based on x_{min}.

Lastly, a random number was created within each x_{min} convergence range and these three values were set as random levels for the parameter factor of service time distribution, obtaining values for k based on the comparison equation.

Number of servers

Variations ranging from 1 to 15 servers were performed, using the assumption presented in Hillier and Lieberman regarding models having no Poisson input (Hillier and Lieberman, 2005. pp. 802). Based on these, the levels selected were 1,2,3,5,7,10 and 15 servers.

System capacity

The system's capacity may be considered as finite or infinite. Methodologies were transferred from other sciences like neuroscience and physics to determine factor levels, their range domain and observation scales varying by powers of 10. Based on this, the levels of finite capacity varied from 10 entities (10^1) for a small system, up to 10,000 entities (10^4). All powers over five were considered, for practical purposes, included within the category of services without limited capacity.

Average service use (ρ factor)

The values of this factor were altered within a range of (0, 1) to unlock the possibility of reaching the rate stability state.

En español

En el rango que se definió, se estableció un salto constante de 0,2, obteniéndose 5 niveles para este factor.

Simulación

Todo modelo de línea de espera se representa por medio de tres instancias básicas: una fuente de generación de entidades que representa la llegada de éstas al sistema, un conjunto de servidores que atienden a las entidades, sobre los cuales se genera la cola, y una salida por donde se despachan las entidades una vez han sido atendidas. Considerando esta estructura, se realizaron los modelos para ser simulados (Figura 1). Adicional a esta estructura, se agregó una validación relacionada con la capacidad del sistema para restringir la entrada de entidades una vez éste se ha copado (Figura 2).

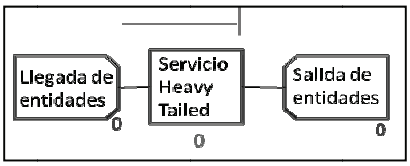

Figura 1. Modelo de simulación en ARENA para sistemas con capacidad infinita

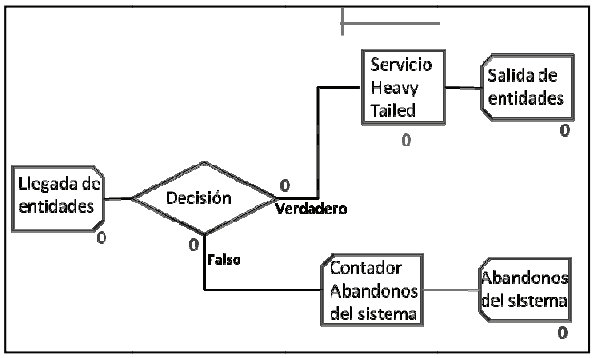

Figura 2. Modelo de simulación en ARENA para sistemas con capacidad finita

No se encontró literatura disponible que permitiese estimar un tamaño de muestra mínimo para un experimento con tres factores fijos con diferente cantidad de niveles y un factor aleatorio. La complejidad misma de definir tamaños de muestra adecuados cuando se presentan factores aleatorios es bien conocida (Montgomery, 2008). Se decidió realizar el mayor número de réplicas que el tiempo y la capacidad computacional permitieran. Puesto que los tamaños de muestra para diseños con cuatro factores fijos no superan las 1.000 unidades en los casos asintóticos (Kuehl, 2001), se consideró que la realización de 10.000 réplicas para cada uno de los 1.575 tratamientos daría un margen de seguridad suficiente para los resultados.

Con respecto a la duración de cada una de las réplicas, se buscó representar un mes de trabajo continuo en una empresa, es decir, 30 días de 24 horas, para un total de 720 horas/mes. Buscando una mayor precisión para el experimento, se aproximó este valor a 1.000 horas/réplica

Resultados

El análisis de los resultados arrojados para la prueba de análisis de varianza realizada en el experimento diseñado permite establecer la existencia de diferencias entre las medias de los factores.

In English

A steady jump of 0.2 in the determined range was recognised, obtaining five levels for this factor.

Simulation

Every queuing model is represented through three basic scenarios: an entity generating source which represents input into the system, a group of servers which service the entities forming the queues and an output to dismiss the entities once they have been serviced. The simulation models were created bearing in mind this structure (Figure 1). A validation mechanism was added to this structure which was related to the system's capacity to restrict the entrance to additional entities once it had filled (Figure 2).

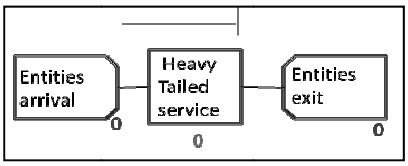

Figure 1. ARENA simulation model for infinite capacity systems

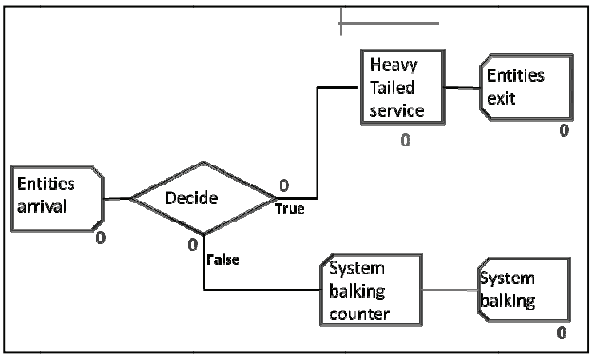

Figure 2. ARENA simulation model for finite capacity systems

No literature was available for estimating the minimum sample size required for an experiment having three constant factors, each having a different number of levels and one random factor. The complexity of determining a suitable sample size when random factors are present is well known (Montgomery, 2008). It was determined that the highest number of replications allowed by time and computational capacity would be carried out. It was assumed that ten thousand replications for each of the 1,575 treatments would provide results having an adequate safety margin because sample sizes for designs with four constant factors do not exceed a thousand units in asymptotical cases (Kuehl, 2001).

The duration of each replication represented a month of continuous work in a factory setting, meaning 30, 24 hour days, a total of 720 hours/month. This value was rounded up to 1,000 hours/replication seeking more precision in the experiment.

Results

Evaluating the results produced by variance analysis ascertained the existence of differences between factor medians.

En español

Se encontraron interacciones significativas (*valor p<0,001*) entre los factores listados a continuación. Interacciones de segundo orden se encontraron entre: capacidad y número de servidores; capacidad y nivel de utilización; número de servidores y nivel de utilización. Se encontró interacción de tercer orden entre capacidad, servidores y nivel de utilización; e interacción de cuarto orden entre capacidad, servidores, nivel de utilización y parámetros de la distribución Pareto.
El análisis hecho para identificar el sentido de las interacciones es descriptivo, siempre y cuando la interacción como tal sea significativa desde un punto de vista estadístico.

Variaciones generales en las medidas de desempeño

Suponer distribuciones exponenciales de los tiempos de servicio lleva a una subestimación del verdadero valor de L_q y W_q si en realidad el proceso presenta tiempos de atención *heavy-tailed*. Por ejemplo, para una línea de espera con capacidad finita, el máximo valor promedio alcanzado de L_q fue 98 entidades, mientras que con una distribución *heavy-tailed* bajo las mismas condiciones del sistema anterior, este valor ascendió a 437 entidades: un incremento del 345%.
En el caso W_q el valor promedio máximo alcanzado con capacidad finita es de 99 horas, mientras que con una distribución *heavy-tailed* bajo las mismas condiciones del sistema, este valor asciende a 273 horas: un incremento del 175%.

Efectos debidos a la capacidad del sistema

Al realizar modificaciones en la cantidad de servidores disponibles en un sistema o en su nivel de utilización, se encuentra que los valores de L_q no varían en sistemas con capacidades superiores a 100 entidades cuando se tienen tiempos de servicio *heavy-tailed*, es decir, que los sistemas con capacidades superiores a este valor, bajo esta distribución, pueden ser considerados como de capacidad infinita (Figura 3).

Esta situación ocurre igualmente para W_q pero solo cuando el factor de utilización es inferior a 0,6. Por encima de este valor, los resultados pierden estabilidad.

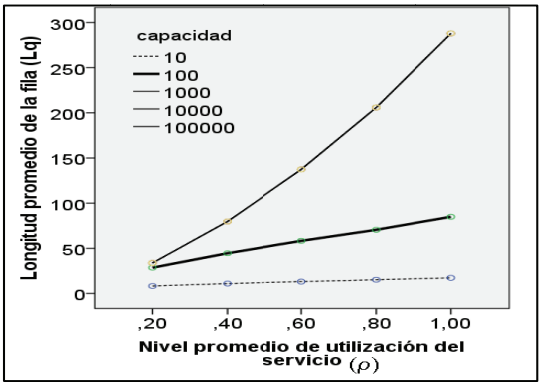

Figura 3. Valores de L_q según capacidad del sistema y □ con tiempos de servicio heavy-tailed.

En la figura 4 se observa una tendencia de crecimiento prácticamente lineal para W_q.

In English

Significant interactions were established (p<0.001) between the factors listed below. Secondary interactions were found between capacity and number of servers, capacity and utilisation level and the number of servers and utilisation level. The interaction of a third level was found between capacity, servers, utilisation level and Pareto distribution parameters.

The investigation for identifying the sense of the interactions was descriptive, provided the interaction was significant from a statistical point of view.

General variations in performance indicator

The assumption of service-time exponential distributions led to underestimating the true value of L_q and W_q if in fact the process presented heavy-tailed service-times. For example, for a finite capacity queuing system, the maximum average value reached for L_q was 98 entities, while this value ascended to 437 entities with a heavy-tailed distribution in the same conditions for the previous system (a 345% increase).

The maximum average value reached with finite capacity was 99 hours in the case of W_q whilst this value ascended to 273 hours with a heavy-tailed distribution in the same system conditions (an175% increase).

Effects due to system capacity

When altering the available server quantities in a system and/or in its utilisation level, it was evident that the values of L_q did not vary in systems having capacities exceeding 100 entities when heavy-tailed service-times existed (i.e. systems having capabilities exceeding this value, with this distribution, may be considered to have infinite capacity, Figure 3).

This state also occurred with W_q but only when the utilisation factor was under 0.6. A lack of stability resulted above this value.

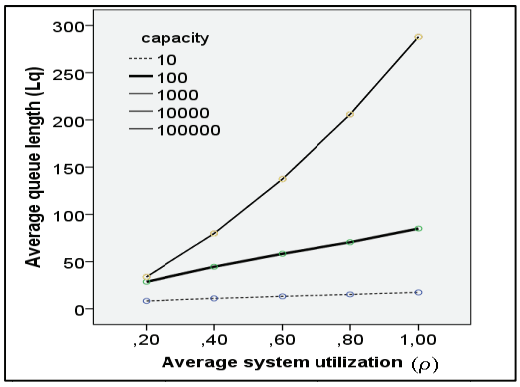

Figure 3. Values of L_q per system capacity and □□with heavy-tailed service-times

Figure 4 illustrates an almost linear growth tendency for W_q.

En español

Para una capacidad de 10 entidades, el cambio de W_q al pasar de un sistema con $\rho = 0.2$ a uno con un nivel de utilización muy cercano a 1 es del 22%, mientras que para un sistema con capacidad infinita es de 194%.

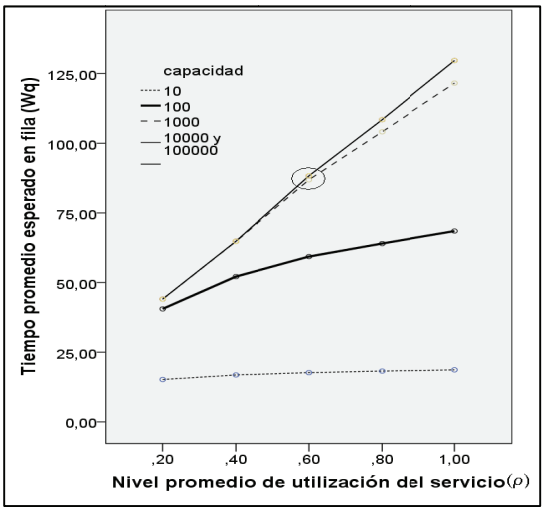

Figura 4. Valores de W_q según la capacidad del sistema y su nivel de utilización con tiempos de servicio heavy-tailed

Caso contrario sucede en los sistemas con tiempos de servicio exponenciales, donde un sistema con capacidad para 10.000 entidades no puede ser considerado como infinito, pues L_q y W_q toman valores diferentes frente al sistema con 100.000 entidades (Figura 5). La tasa de crecimiento de W_q es mayor que en los sistemas *heavy-tailed* (compárese la forma de las figuras 4 y 5). Sin embargo, los valores de L_q y W_q en sistemas exponenciales es menor que en sistemas *heavy-tailed*.

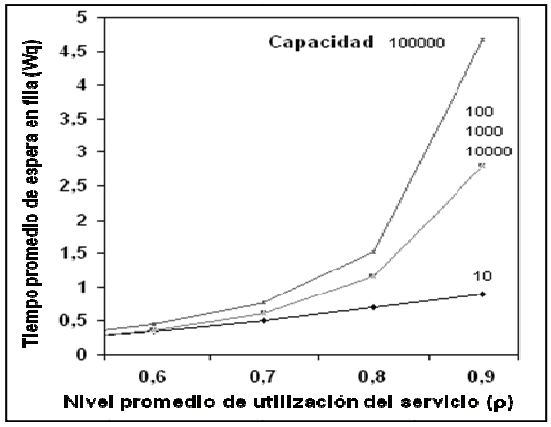

Figura 5. Valores de W_q según la capacidad del sistema y su nivel de utilización con tiempos de servicio exponenciales

Efectos debidos al número de servidores

Al aumentar la cantidad de servidores del sistema utilizando la distribución exponencial en los tiempos de servicio, los valores de L_q se reducen excepto en niveles de utilización altos, donde se pierde el estado estable. Estos resultados son los esperados de acuerdo a la teoría.

In English

The change in W_q going from a $\rho = 0.2$ system to one with a utilisation level very close to 1 was 22% for a capacity of 10 entities, while for an infinite capacity system this percentage was 194%.

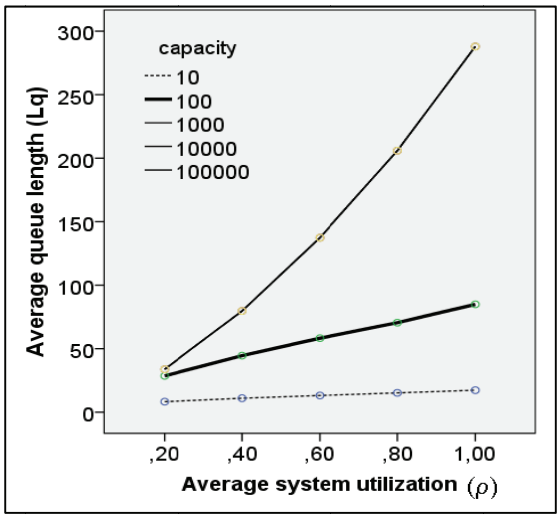

Figure 4. Values of W_q per system capacity and its utilisation level with heavy-tailed service-times

A contrary case occurred in systems having exponential service-times where a system with capacity for 10,000 entities could not be considered infinite, since L_q and W_q took on different values when faced with a system with 100,000 entities (Figure 5). The growth rate of W_q was greater than in heavy-tailed systems (compare Figures 4 and 5). However, the values for L_q and W_q in exponential systems were lower than in heavy-tailed systems.

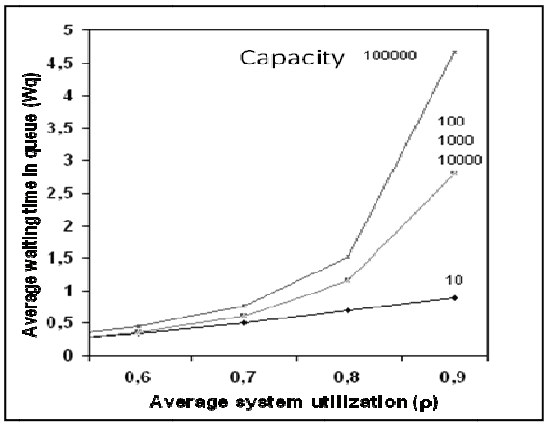

Figure 5. Values of W_q per system capacity and the level of utilisation with exponential service-times

Effects due to number of servers

When the number of servers in the system was increased using exponential distribution in service-times, the values of L_q were reduced except during high utilisation levels, where stability was lost. According to the theory, these were the expected results.

En español

Para el caso de las distribuciones *heavy-tailed* se observa un comportamiento atípico cuando se tiene un único servidor en el sistema (Figura 6), pues sólo con aumentar a dos servidores, el valor de L_q se reduce en gran proporción y no varía de allí en adelante al aumentar el nivel de utilización. Esto implicaría que si se tienen distribuciones *heavy-tailed* importa menos el nivel de utilización del sistema, ganando relevancia la cantidad de servidores disponibles en él.

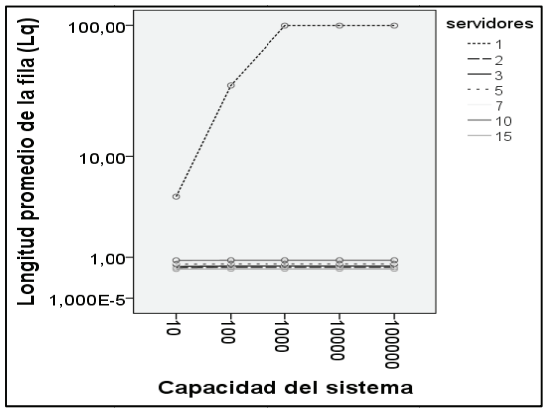

Figura 6. Valores de L_q según la cantidad de servidores en el sistema con tiempos de servicio heavy-tailed

Cuando se tienen capacidades superiores a 1.000 entidades en sistemas con tiempos de servicio *heavy-tailed* , W_q no disminuye consistentemente como sería de esperarse al aumentar el número de servidores, sino que presenta un comportamiento contraintuitivo. Como puede apreciarse en la figura 7 el valor de W_q se incrementa entre 1 y 5 servidores. Para una capacidad de 1.000 entidades, a partir de 5 servidores W_q comienza a disminuir; para más de 1.000 entidades, el aumento atípico de W_q se mantiene.
Este comportamiento atípico en sistemas con distribuciones *heavy-tailed* puede ser consecuencia de la no convergencia de su media o varianza para ciertos valores.

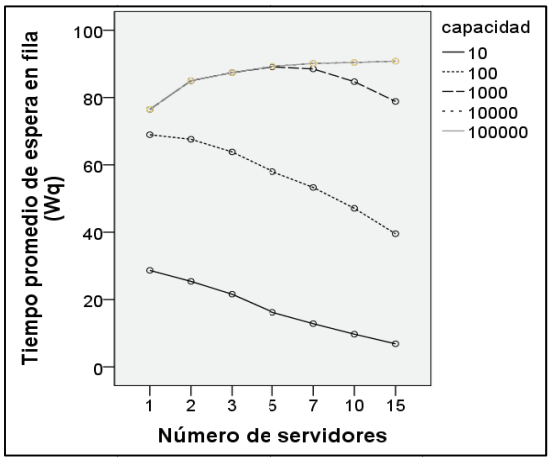

Figura 7. Valores de W_q según la capacidad del sistema y la cantidad de servidores con tiempos de servicio heavy-tailed

In English

Atypical behaviour was observed in heavy-tailed distributions when there was only one server in the system (Figure 6); by simply increasing the number of servers to two, the value of L_q was greatly reduced and it did not vary when the utilisation level was increased. This would imply that the level of utilisation of the system is not as important as the number of servers available on it regarding heavy-tailed distributions.

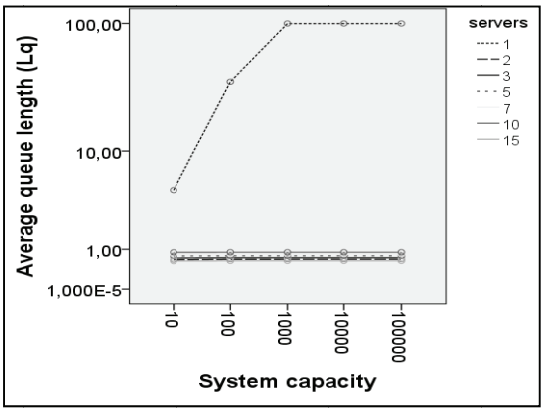

Figure 6. Values per the number of servers in the system having heavy-tailed service-times

When capacities exceeded 1,000 entities in systems having heavy-tailed service-times, contrary to expectations, W_q did not consistently decrease with the increase of server quantity; instead it presented a counterintuitive pattern. As illustrated by Figure 7 the value of W_q became increased between one and five servers. W_q began to decrease with a capacity for 1,000 entities, and from five servers; the atypical increase of W_q continued for more than 1,000 entities.
This atypical behaviour in heavy-tailed distributions may have been the consequence of the non-convergence of its median and/or variance for certain values.

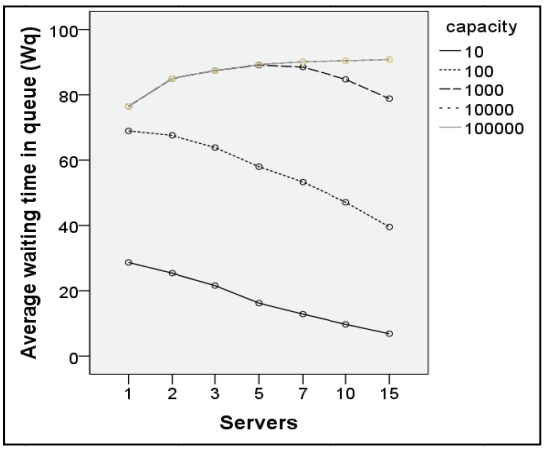

Figure 7. Values of W_q per system capacity and the number of servers having heavy-tailed service-times

En español

Efectos debidos al factor de utilización ρ

En la medida en que aumenta el nivel de utilización del sistema, se incrementan de forma acelerada L_q y W_q, con independencia de la distribución de los tiempos de servicio, como era de esperar. Para capacidades mayores, se hacen más evidentes las variaciones de L_q y W_q conforme crece ρ.
En sistemas con tiempos de servicio exponenciales los valores de L_q se hacen realmente representativos, es decir, mayores a una entidad, para niveles superiores a 0,7, independientemente de la cantidad de servidores en el sistema y su capacidad. Para los sistemas con distribuciones *heavy-tailed* las variaciones del indicador son representativas, y mayores a las obtenidas con tiempos de servicio exponenciales, desde un nivel de utilización de 0,2 únicamente cuando el sistema tiene 1 servidor (Figura 8), de manera similar a lo mostrado en la sección anterior, lo cual representa un comportamiento atípico. De 2 servidores en adelante los valores y variaciones de L_q son mínimas y no significativas.

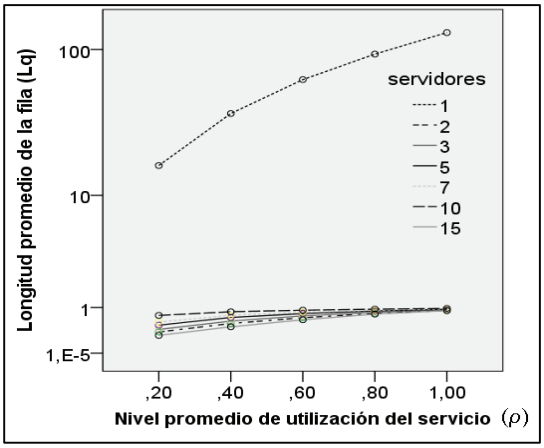

Figura 8. Valores de L_q según la cantidad de servidores en el sistema y ρ con tiempos de servicio heavy-tailed

Se encontró que en la medida en que se incrementa la capacidad del sistema junto con el nivel de utilización, no se desestabiliza el sistema en cercanías a un nivel $\rho = 1$ para distribuciones de los tiempos de servicio exponenciales, mientras que con tiempos de servicio *heavy-tailed* el comportamiento de desestabilización se sigue presentando. Esto evidencia cómo los sistemas con distribuciones exponenciales son más predecibles y arrojan resultados más confiables con niveles de utilización y capacidades elevados.

Las variaciones de W_q al incrementar el nivel de utilización del sistema de 0,2 a 1 se hacen cada vez más pequeñas en la medida en que se aumenta la cantidad de servidores disponibles en el sistema, hasta llegar a ser nulas para 15 servidores (Figura 9).
Este comportamiento evidencia que en sistemas con pequeñas capacidades el tiempo promedio que una entidad espera en la fila antes de ser atendida parece no depender de la tasa de llegada de entidades al sistema ni de la tasa de servicio.

Efectos debidos a los parámetros de la distribución

Las mayores variaciones de L_q y W_q dependen del nivel de utilización de los sistemas con tiempos de servicio exponenciales, más que de la capacidad del sistema y el número de servidores;

In English

Effects due to utilisation factor ρ

As the level of system utilisation increased, the increase of L_q and W_q accelerated, independently from service-time distribution, as anticipated. Variations in L_q and W_q became more apparent As ρ grew for greater capacities.

The values of L_q became truly representative in systems having exponential service-times (i.e. greater than one entity, for levels exceeding 0.7, independently of the number of servers in the system and their capacity). Indicator variations were representative for heavy-tailed distribution systems and were greater than those obtained with exponential service-times, from a 0.2 utilisation level only when the system had only one server (Figure 8), similar to that presented in the previous section, which represented an atypical behaviour. The values and the variations of L_q were minimal from two servers and on, and were not significant.

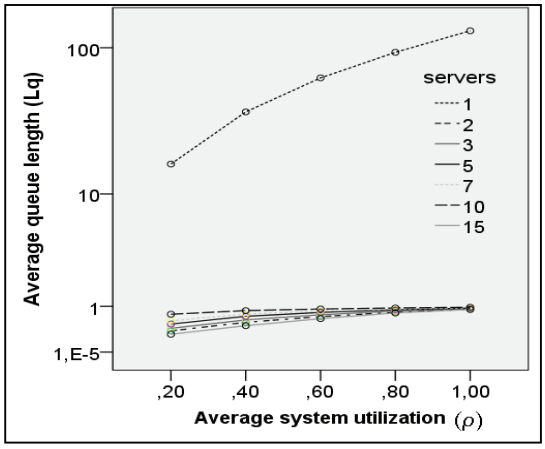

Figure 8. Values of L_q per the number of servers in the system and ρ with heavy-tailed service-times

It was established that as system capacity and utilisation levels increased, the system was not destabilised around a $\rho = 1$ level for exponential service-time distributions, while this destabilization was still perceptible with heavy-tailed service-times. This confirmed that systems having exponential distributions were more predictable and yielded more reliable results with high levels of utilisation and capacities.

Variations of W_q as the level of system utilisation was increased from 0.2 to 1 decreased as the number of available servers in the system was increased, before becoming cancelled out at 15 servers (Figure 9).
This behaviour provided evidence that in systems having small capacities the entity's average waiting-time in queue before being served did not seem to depend on either the rate of entity arrivals into the system or service rate.

Effects due to distribution parameters

The greatest variation of L_q and W_q depended on the level of system utilisation of exponential service-times, more than the system capacity and the number of servers;

En español

mientras que para sistemas con tiempos de servicio *heavy-tailed*, es el parámetro de forma (*k*) el que determina la magnitud de la variación, es decir, la convergencia de la media o de la varianza (Tablas 1 y 2).

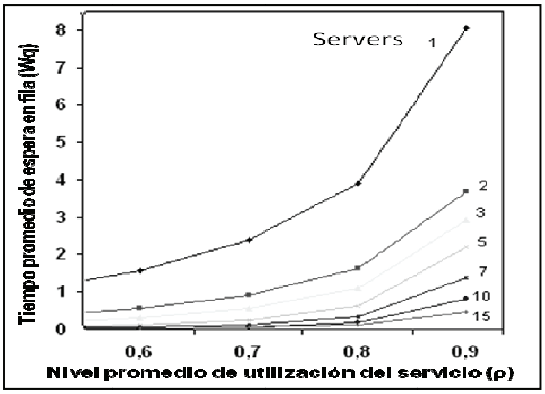

Figura 9. Valores de W_q según la capacidad del sistema y ρ con tiempos de servicio exponenciales

Tabla 1. Impacto del parámetro de forma en sistema heavy-tailed sobre L_q

Situación	*K*	L_q
Media y varianza convergen	$k > 2$	Prácticamente cero
Varianza converge	$1<k<2$	Puede llegar a ser igual a 40 entidades
Nada converge	$k<1$	Puede llegar a ser igual a 70 entidades

Tabla 2. Impacto del parámetro de forma en sistema heavy-tailed sobre W_q

Situación	*K*	W_q
Media y varianza convergen	$k > 2$	Prácticamente cero. Sin incrementos
Varianza converge	$1<k<2$	Puede llegar a ser igual a 90 horas. Incremento acelerado al incrementar ρ
Nada converge	$k<1$	Puede llegar a ser igual a 190 horas. Incremento desacelerado al incrementar ρ

Conclusiones

La investigación mostró que las medidas de desempeño clásicas de las líneas de espera presentan comportamientos atípicos e inestables cuando la distribución de los tiempos de atención tiene un comportamiento *heavy-tailed*. Esto puede deberse tanto a la tendencia de las distribuciones *heavy-tailed* a generar valores extremos con mayor frecuencia como a la no convergencia de la media o la varianza de la distribución *heavy-tailed*. De este modo se concluye que los indicadores basados en promedios tales como L_q y W_q no son los más adecuados para medir el desempeño de una línea de espera cuando la distribución de los tiempos de servicio es *heavy-tailed*. Puesto que los modelos generatrices teóricos de distribuciones *heavy-tailed* presentados en la introducción cubren una amplia gama de situaciones reales, es necesario para investigadores y profesionales considerar la posibilidad de la presencia de modelos *heavy-tailed* cuando se presenten valores extremos en las muestras, en vez de simplemente hallar una explicación para ellos y posteriormente, descartarlos.

In English

meanwhile, for systems having heavy-tailed service-times, it was the (*k*) parameter which determined the magnitude of the variation (i.e. convergence of the median and/or variance) (Tables 1 and 2).

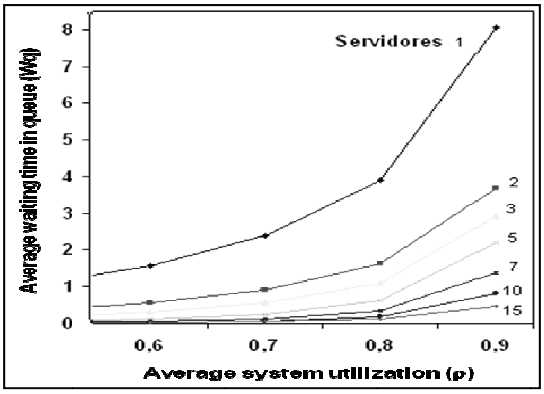

Figure 9. Values of W_q per system capacity and ρ , with exponential service-times

Table 1. Impact of the shape parameter in heavy-tailed systems over L_q

Scenario	*K*	L_q
Mean and variance Convergence	$k > 2$	Approximately zero
Variance convergence	$1<k<2$	Might reach 40 entities
No convergence	$k<1$	Might reach 70 entities

Table 2. Impact of the shape parameter in heavy-tailed systems over W_q

Scenario	*K*	W_q
Mean and variance convergence	$k > 2$	Approximately zero Without increases
Variance converge	$1<k<2$	Might have reached 90 hours Steep increases as □□increased
No convergence	$k<1$	Might have reached 190 hours Slow increases as □□increased

Conclusions

The investigation demonstrated that traditional queuing system performance indicators presented atypical and unstable behaviour when the distributions of service-times had a heavy-tailed behaviour. This may have been due to both the tendency of heavy-tailed distributions towards more frequently generate extreme values and the non-convergence of the median and/or the variance of the heavy-tailed distribution. It was thus concluded that the indicators based on averages, such as L_q and W_q were not the most appropriate for measuring the performance of a queuing system when the distribution of service-times was heavy-tailed. Because theoretic generative models of heavy-tailed distributions in the introduction cover a broad range of real life situations, it is necessary for researchers to consider the possibility of the presence of heavy-tailed models when extreme values appear in samples, instead of simply finding an explanation for them and subsequently discarding them.

En español

Trabajos futuros pueden explorar nuevos y más robustos indicadores para líneas de espera cuyos tiempos de servicio puedan ser modelados con variables *heavy-tailed*.

Otro resultado clave de la investigación es la mayor sensibilidad de las distribuciones *heavy-tailed* a cambios en los parámetros tales como la capacidad del sistema y el número de servidores, lo cual debe tenerse en cuenta al momento de tomar decisiones en líneas de espera con tiempos de atención *heavy-tailed*. La realización de un análisis de riesgo en la toma de decisiones bajo estas condiciones es otra vía de investigación que se abre.

Apéndice: nomenclatura

L_q: Longitud promedio de la fila
W_q: Tiempo promedio de espera en fila
ρ: Nivel promedio de utilización
x_{min} : Parámetro de posición de la distribución Pareto
k: Parámetro de forma de la distribución Pareto
β: Parámetro de la distribución exponencial

In English

Future studies may explore new and stronger indicators for queuing systems and their service-times may be modelled with heavy-tailed variables.

Another key result of the investigation was the greater susceptibility of heavy-tailed distributions to parameter changes such as system capacity and the number of servers; this must be considered when making decisions regarding queuing systems having heavy-tailed service-times. The completion of a risk analysis during the decision-making process in these conditions is another means of investigation.

Appendix: nomenclature

L_q: Average length of the queue
W_q: Average in queue waiting time
ρ: Average utilisation level
x_{min} : Pareto distribution position parameter
k: Pareto distribution shape parameter
β: Exponential distribution parameter

Bibliografía / References

Alvarado, J. A., Montoya, J. R., Rangel, L. M., Analyse par simulation de l'impact de la modélisation du temps de service avec une distribution heavy-tailed: étude de Cas d'un atelier de maintenance automobile., Mosim 2008, Proceedings of the 7eme conference international de modelisation et Simulation, Paris, 2008.

Andriani, P., McKelvey, B., Why Gaussian statistics are mostly wrong for strategic organization., Strategic *Organization,* Vol. 3, 2005, pp. 219–223.

Barabási, A. L., The origin of bursts and heavy-tailed in human dynamics., Nature, Vol. 435, 2005, pp. 435–439.

Cohen, J. W., Some results on regular variation for distributions in queuing and fluctuations theory., Journal of Applied Probability, Vol. 10, 1973, pp. 343–353.

Embrechts, P., Kluppelberg, C.,Mikosch, T., Modeling extremal events for Insurance and finance., New York, Springer-Verlag, 1997.

Gross, D., Fundamentals of queuing theory., 4th ed., New York, John Wiley & Sons, 2009.

Hillier, F., Lieberman, G. J., Operations Research., 8th ed., México, McGraw- Hill. 2005.

Janicki, H. P., Simpson, E., Changes in the size distribution of US Banks: 1960–2005., Economic Quarterly - Federal Reserve Bank of Richmond, Vol. 92, No. 4, 2005, pp. 291-316.

Kuehl, R., Diseño de experimentos: Principios estadísticos de diseño y análisis de investigación., México, International Thomson Editores, 2001.

Mitzenmacher, M., A brief history of generative models for power law and lognormal distributions., Internet Algorithms, Vol. 1, No. 2, 2004, pp. 226–251.

Montgomery, D., Design and analysis of experiments., 7th edition, 2008, Wiley.

Neuts, M. F., Computer experimentation in applied probability., Journal of applied probability, Vol. 25A,1988, pp. 31-43

Newman, M., Power laws, Pareto distributions and Zipf's law., Contemporary Physics, Vol. 46, 2005, pp. 323-351.

Pakes, A. G., On the tails of waiting-time distributions., Journal of Applied Probability, Vol. 12, 1975, pp. 555–564.

Ross, S., A first course in probability., 7th ed., New Jersey, Pearson Prentice Hall, 2006.

Sigman, K., Apendix: A primer on heavy-tailed distributions., Queuing Systems, Vol. 33, No. 1-3, Dic., 1999, pp. 261–275.

Stidham, S., Analysis, Design, and Control of Queuing Systems., Operation research, Vol. 50, 2002, pp. 197-216.

Stewart, W. J., Introduction to the numerical solution of Markov chains, New Jersey, Princeton University Press, 1994.

Whitt, W., The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting time distribution., Queuing Systems, Vol. 36, 2000, pp. 71-87.

Willinger, W., Traffic modelling for high-speed networks: theory versus practice., Stochastic Networks, IMA Volumes in Mathematics and its applications 71, Springer-Verlag, New York, 1995, pp.169-181.