Artículo de Investigación. Revista Ciencia e Ingeniería. Vol. 35, No. 1, pp. 47-56, diciembre-marzo, 2014.
ISSN 1316-7081. ISSN Elect. 2244-8780 Universidad de los Andes (ULA)

Diagnóstico y caracterización de series autosimilares en trazas de un dispositivo de almacenamiento con aplicaciones financieras

Diagnosis and characterization of self-similar series of storage-device traces with financial applications

Borrero Molina, Armando*, Quintero Gull, Carlos
Escuela Básica de Ingeniería, Facultad de Ingeniería. ULA.
Mérida 5101. Venezuela
borrero@ula.ve*

Resumen

En el presente trabajo se analizaron las trazas de dispositivos periféricos de almacenamiento, con la finalidad de verificar si las mismas poseían características de series autosimilares. Para ello, se dispuso de los registros de funcionamiento de un dispositivo destinado al uso de aplicaciones financieras, trabajando durante doce horas de actividad del sistema. Cada medición o registro representa una solicitud al dispositivo, compuesta por las variables: tiempo de llegada de la solicitud, identificación del dispositivo en el que se realiza la operación, dirección a la cual se dirige la solicitud, tamaño de la misma y tipo de operación a realizar (lectura/escritura). Se construyó la serie de tiempo $\boldsymbol{x\ t}$ *= Número de solicitudes a disco en un interval* $(\boldsymbol{t-1}, \boldsymbol{t})$*, y se pudo observar que la misma presenta características de procesos autosimilares, con parámetro de autosimilaridad de Hurst* $\boldsymbol{H = 0.9}$ *. Posteriormente con el análisis de la función de autocorrelación de la serie, se pudo determinar que la misma posee la característica de autosimilaridad con Dependencia de Largo Rango.*

Palabras Claves: Autosimilaridad, dispositivos de almacenamiento, series de tiempo, cargas de trabajo, trazas.

Abstract

In this paper storage-peripheral-device traces were analyzed in order to verify if they have characteristics of self-similar series. The analyzed data were the working operation records obtained from 12 hours of a system activity, used for financial applications. Each measurement or record represents a request, made by the application to the device, which is conformed by: the request arrival time, identification of the device on which a particular operation is performed, address of a particular request, size of the request and type of the operation to be performed (read /write). Time series $\boldsymbol{x\ t}$ *were made out of the number of requests to disk in a time interval* $(\boldsymbol{t-1}, \boldsymbol{t})$ *showing that the series has features which resembles a self-similar processes, with self-similarity parameter value of Hurst, of* $\boldsymbol{H = 0.9}$*. From the series autocorrelation function, it was determined that the series analyzed here has properties of self-similarity, with a long-range dependence.*

Key words: Selfsimilarity, storage devices, time series, workloads, traces.

1 Introducción

Cuando se desea desarrollar y/o mantener sistemas informáticos, es particularmente importante realizar análisis de sus trazas, es decir, estudiar los registros de la secuencia de acciones realizadas durante la ejecución o funcionamiento del sistema, pues ellas definen, representan con detalle, cualidades y características propias del sistema. Permiten conocer el comportamiento de los mismos, proporcionan mucha información acerca de su rendimiento, desempeño, ventajas y desventajas de diseño e implementación.

Es conveniente resaltar que en la literatura especializada, autores como (Ganger. 1995), (Kavalanekar. 2008), entre otros, han reseñado que existe dificultad para obtener tales registros o trazas de las cargas de trabajo ("workloads") de los sistemas informáticos.

Por tal motivo, algunos investigadores prefieren utilizar registros de actividades de sistemas de almacenamiento reales, basados en los registros de las secuencias de solici-

tudes realizadas a un disco o sistema de almacenamiento. Sin embargo, es importante señalar que aun cuando dichos registros representan un inventario bastante preciso de lo ocurrido en dicho sistema durante un intervalo de tiempo determinado, debe tenerse en cuenta que tales registros de solicitudes padecen ciertas limitaciones.

(Ganger. 1995), hace referencia a esas restricciones, cuando afirma que (**1**) Por razones no técnicas, puede ser extremadamente difícil convencer a los administradores de los sistemas de que permitan el seguimiento, registro y almacenamiento de las trazas. (**2**) Muchas veces las trazas tienden a ser grandes, a ocupar considerable espacio de disco. (**3**) Cada traza representa una medida única del comportamiento, por lo que es difícil confiar estadísticamente en los resultados. Algunas veces puede ser difícil distinguir entre las características del rendimiento real del sistema bajo prueba y el comportamiento anómalo de la traza. (**4**) Es muy difícil aislar y/o modificar características específicas de las trazas de una carga de trabajo, (por ejemplo, tasa de llegada o capacidad total del almacenamiento accedido). (**5**) No es posible estudiar trazas de cargas de trabajo futuras esperadas, pues no se puede registrar algo que todavía no existe.

Muchos investigadores han preferido realizar sus estudios con base en las trazas sintéticas, tomando en cuenta estas dificultades y además debido a que esta tarea pudiera ser un poco menos compleja. El proceso consiste en tomar un conjunto de registros de las actividades de los sistemas de almacenamiento, luego calcular algunas estadísticas que los caractericen y por último, elaborar un modelo estadístico que identifique su comportamiento. El problema estriba en que tampoco es fácil generar el modelo; y sobre todo, que existe la posibilidad de que éste no represente fielmente al sistema original, pues pudiera contener muchos supuestos simplificadores, que comprometan la validez de los resultados.

De acuerdo a (Kavalanekar, 2008), las trazas de los sistemas de almacenamiento de datos se utilizan generalmente con tres amplios fines:

Analizar características y comportamiento de sistemas específicos a partir de las cargas de trabajo registradas;
Constituir conjuntos de datos para modelizar o simular el sistema de acuerdo a esas trazas;
Extraer parámetros y heurísticas que permitan construir modelos de cargas de trabajo sintéticas.

Este trabajo estaría comprendido en el ámbito del primero de los fines mencionados, pues en él se pretende analizar las trazas de un conjunto de discos o periféricos de almacenamiento, con la finalidad de conocer el comportamiento de los mismos. Posteriormente, a mediano plazo se desea construir un modelo del sistema a partir de esas trazas y finalmente realizar simulaciones para generar trazas sintéticas con las mismas características de las originales.

2 Metodología Utilizada

Para el desarrollo de la presente investigación, se dispone de un conjunto de registros que identifican a las solicitudes que se presentan ante un grupo de veintitrés (**23**) dispositivos de almacenamiento, durante doce (**12**) horas continuas de funcionamiento de un sistema financiero. Este conjunto de datos fueron proporcionados por el Laboratorio de Ciencias de Computación de la Universidad de Bretaña Occidental, Francia, obtenidas del repositorio de trazas UMass Trace Repository (UMass 2011). En este trabajo particular, se ha utilizado solo uno de los dispositivos disponibles. La primera tarea consiste básicamente en construir una serie temporal, tomando en cuenta el tiempo de llegada de las solicitudes ante el dispositivo seleccionado. La finalidad primordial de este trabajo de investigación es determinar si dicha serie de tiempo tiene características que la identifiquen como un proceso autosimilar. Para poder establecer la autosimilaridad, es necesario precisar el grupo de parámetros o estadísticas que definan de manera concreta a esta clase de modelos.

Como es bien sabido, durante un intervalo de tiempo, se presenta una cierta cantidad de solicitudes ante los dispositivos, lo cual puede ser estudiado a través de un modelo de tráfico. Puede modelizarse a través de un proceso estocástico que representa la cantidad de solicitudes que se presentan a los dispositivos de almacenamiento durante un tiempo determinado.

Es necesario determinar cuál es el modelo de tráfico que mejor se ajusta a las características de la serie en estudio.

Es conveniente acotar que los modelos de tráfico tradicionales han trabajado bajo el supuesto de que el número de solicitudes que se realizan a los diversos dispositivos así como los tiempos entre llegadas de dichas solicitudes, son independientes entre sí. Se han desarrollado modelos, como el de autosimilaridad, (también conocido como autosimilitud o fractalidad), que se ajustan mejor al modelo de tráfico de las redes actuales, pues se ha determinado que existe cierto grado de dependencia entre estos eventos.

Desafortunadamente, su aplicación puede conducir al desarrollo de estructuras complejas de correlación, en diferentes escalas, que puede poner en duda la validez de los modelos tradicionales.

En este trabajo se estudia su aplicación al modelo de tráfico de las solicitudes a disco, con la finalidad de hacer un aporte al desarrollo de modelos autosimilares aplicados a este tipo de sistemas, y de esta forma ayudar a mejorar la comprensión del funcionamiento de los sistemas de almacenamiento, así como el desarrollo, mantenimiento y rendimiento de los mismos. Se partirá del supuesto de que el comportamiento de los accesos a disco sigue un modelo autosimilar. En caso de cumplirse este supuesto, se intentará determinar el tipo de dicha autosimilitud.

Ahora bien, es conveniente recordar que una solicitud a disco contiene información muy específica de lo que su-

cede en un subsistema de almacenamiento particular, durante un período de tiempo determinado. Una solicitud al disco está definida por cinco valores: Tiempo de llegada de la solicitud, Identificador del dispositivo, Dirección, Tamaño y Tipo de la solicitud. La primera variable identifica el momento en que la solicitud se presenta ante el dispositivo para ser tratada; el resto de las variables identifican el acceso de la solicitud al disco. La variable Identificador del dispositivo, contiene un número que indica a cuál de los dispositivos se dirige la acción. La variable Dirección se refiere a la localidad del disco en la cual comienza la solicitud. La variable Tipo especifica si la solicitud es de lectura o de escritura. La secuencia de tiempos de llegadas, se conoce como patrón de llegadas y al conjunto de variables que identifican los accesos a disco, se le denomina patrón de accesos.

3 Análisis de Resultados

Para el presente trabajo, la matriz de datos está conformada por aproximadamente $\mathbf{450.000}$ registros de solicitudes en uno de los dispositivos de almacenamiento.

La atención se centrará en la variable Tiempo de llegada, que como se ha dicho, representa el momento en el cual se origina la solicitud. Esto se realiza con el fin de caracterizar dicha variable, para definir qué tipo de distribución sigue esta colección de datos. Para ello se realizó un conjunto de operaciones en los datos, con la finalidad de transformarlos en una serie de tiempo, que se ha definido de la siguiente manera:

$\boldsymbol{x\ t}$ Número de solicitudes que llegan en el intervalo de tiempo $\boldsymbol{t-1, t}$ (1)

En la figura $\mathbf{1}$ se puede observar que la secuencia de la serie es impredecible, lo cual obliga a tratarla como un proceso o serie estocástica.

A lo largo del tiempo se ha pensado que el número de eventos que ocurren en un intervalo de tiempo se distribuyen como un proceso de Poisson, bajo el supuesto de que

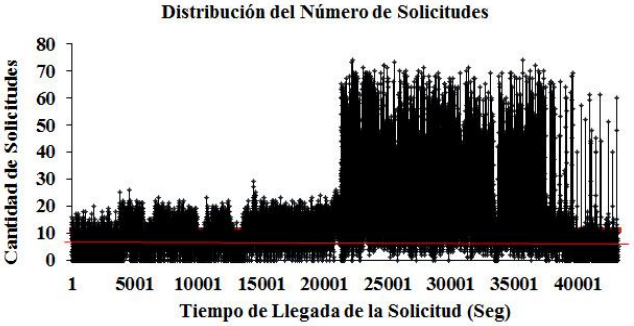

Fig. 1. Distribución del número de solicitudes a disco

el tiempo entre dos llegadas sucesivas es independiente. Sin embargo, en muchas aplicaciones informáticas y de telecomunicaciones actuales, este supuesto es inválido.

Pues en muchos casos no se puede esperar independencia entre los diversos tiempos entre llegadas (Alzate. 2001).

Para el tratamiento de nuestra serie de tiempo $\boldsymbol{x(t)}$, como un modelo de tráfico, es necesario que la misma sea una serie de tiempo estacionaria

Definición 1

Se dice que un proceso estocástico $\boldsymbol{x\ t\ , t \epsilon Z}$ es estrictamente estacionario, si la distribución conjunta de las variables aleatorias $\boldsymbol{x\ t_1\ , x\ t_2\ , x\ t_3\ , \ldots\ldots, x\ t_n}$ es la misma distribución conjunta de las variables aleatorias, $\boldsymbol{x\ t_{1+k}\ , x\ t_{2+k}\ , x\ t_{3+k}\ , \ldots\ldots, x\ t_{n+k}}$ para todo $\boldsymbol{n \in N,\ t_1, t_2, \ldots\ldots, t_n,\ k \in Z}$. (Alzate. 2001).

Definición 2:

Se define un proceso estocástico como estacionario, si sus propiedades estadísticas son invariantes ante una traslación del tiempo. Es decir, si el mecanismo físico que genera el experimento no cambia con el tiempo (González. 2012).

Dicho de otra forma, la distribución conjunta de ésta serie, debe ser la misma distribución de la serie de tiempo original, desplazada $\boldsymbol{k}$ unidades de tiempo.

Estas definiciones son muy restrictivas, pues obligan a estudiar demasiadas funciones, de acuerdo a la longitud de la serie. Por tanto en este estudio se usará una definición menos restrictiva de *estacionariedad*.

Se aplicará la definición de débilmente estacionaria, que establece lo siguiente:

Definición 3:

La serie $\boldsymbol{x\ t}$ es débilmente estacionaria si $\boldsymbol{E\ x\ t\ = \alpha\mu}$ (independientemente del tiempo). (González. 2012).

Es decir, que el promedio de la serie es invariante ante desplazamientos en el tiempo.

En otras palabras, se puede decir que una serie de tiempo es estacionaria si fluctúa alrededor de un mismo valor durante todo su recorrido.

Al calcular el promedio de la serie de tiempo $\boldsymbol{x(t)}$, se obtiene el valor $\mathbf{11,07} \approx (\mathbf{11})$, es decir que en promedio ocurren 11 solicitudes a disco por segundo.

Como se puede observar en la Fig. $\mathbf{1}$, la serie de tiempo $\boldsymbol{x\ t}$ fluctúa alrededor de $\mathbf{11}\ \boldsymbol{Solicitudes/seg}$. a lo largo de todo su recorrido. Por tanto, de acuerdo con la Definición 3 se puede afirmar que esta serie de tiempo es débilmente estacionaria.

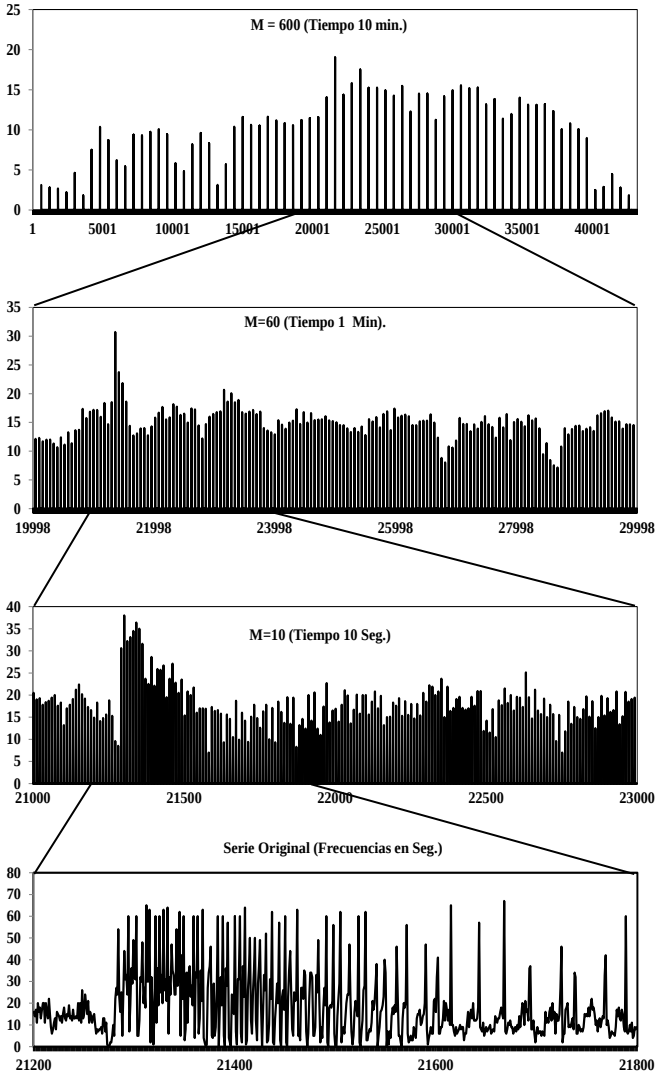

Fig. 2.- Número de solicitudes por períodos

Esta serie, tal y como se encuentra definida, se puede modelar mediante un proceso de Poisson. Pero, como se ha dicho, este tipo de procesos (Poisson), establece que existe independencia en el tiempo entre llegadas de dos eventos sucesivos. Este supuesto no necesariamente ocurre en la realidad. Recientemente se han estudiado otros modelos que se ajustan mejor a procesos con estas características. Toman en cuenta la relación existente entre los tiempos de llegadas, como es el caso de las series autosimilares.

Una vez que se ha demostrado la estacionariedad de la serie de tiempo, (en nuestro caso, la serie es débilmente estacionaria), se procede a determinar la Autosimilaridad.

Los modelos o procesos autosimilares permiten capturar la relación de dependencia entre un evento y los eventos anteriores. De ahí la importancia que han cobrado estos procesos en el mundo actual.

Informalmente, la autosimilaridad, auto-similitud o fractalidad, significa que un proceso estocástico se ve "más o menos" igual en cualquier escala.

Observando los gráficos mostrados en la Fig. 2, es de destacar que las figuras se parecen entre sí de una manera aproximada. Si se restringe la similitud sólo a algunas estadísticas de estas series de tiempo a diferentes escalas, se puede decir que éstas son muy similares entre sí. Casi se podría decir que la serie de tiempo se replica a sí misma a diferentes escalas del tiempo. Dicho de otra forma, las series se comportan de manera similar para cada período de tiempo de longitud t.

Es conveniente resaltar que cada gráfica que conforma la Fig. **2**, presenta los datos a diferente granularidad, (Riska et al. 2009), es decir, con unidades de tiempo a distintas escalas. La figura con $\boldsymbol{M = 600}$, corresponde a las frecuencias de accesos a disco cada 10 minutos. La figura con $\boldsymbol{M = 60}$, representa las frecuencias de los accesos a disco cada minuto $\boldsymbol{60\ segundos}$). De igual forma, la figura con $\boldsymbol{M = 10}$, corresponde a las frecuencias de los accesos a disco cada $\boldsymbol{10}$ segundos y por último, en la figura que muestra la serie original se representan las frecuencias de accesos a disco por segundo. Estos agrupamientos de las frecuencias, se realizaron con la intención de estudiar y comparar el comportamiento de la serie de tiempo a diferentes escalas.

Para verificar si nuestra serie de tiempo x(t) es un modelo autosimilar en el tiempo discreto, se tiene que considerar el proceso agregado mostrado en la ecuación $(\boldsymbol{2})$, que se muestra a continuación:

$$X^{m}\ i\ = \frac{1}{m}\ \ \sum_{t=1+m(i-1)}^{m_i} X\ t \qquad (2)$$

Es decir, la serie original $\boldsymbol{x(t)}$ se particiona en bloques no solapados de tamaño m y se promedian los valores de cada bloque, de manera que cada promedio será el promedio de llegadas o solicitudes en el bloque $\boldsymbol{i-}$ é$\boldsymbol{simo}$.

Si las medidas descriptivas tales como la media, se conservan en las distintas compresiones, entonces existe evidencia de que la serie $\boldsymbol{x(t)}$ definida anteriormente, se comporta como un proceso autosimilar.

En este estudio se establecieron tres procesos agregados, es decir, se particionó la serie original en distintas series conformadas por bloques no solapados. Estos bloques se definieron en los siguientes tamaños $\boldsymbol{10}$, $\boldsymbol{60}$, $\boldsymbol{y\ 600}$, y se calcularon los promedios por cada bloque. Cada resultado representa el número de solicitudes promedio en el bloque i obteniéndose tres series distintas de acuerdo al tamaño del bloque.

En la Fig. 3 se puede observar que los valores de los promedios de las solicitudes para los distintos tamaños de m se preservan. Por tanto se puede concluir que estamos ante una serie o proceso autosimilar.

La autosimilaridad se puede expresar matemáticamente de diversas formas equivalentes: una de ellas expresa que la varianza de las muestras de los procesos agregados de tamaño m , decrecen o tienden a cero de forma más lenta que el recíproco del tamaño del bloque.

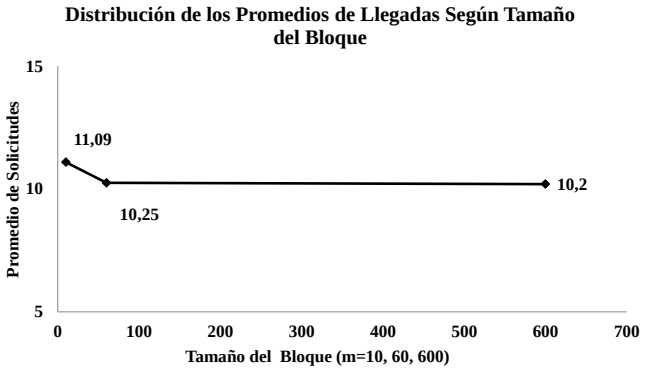

Fig. 3.- Distribución de los Promedios de llegadas según Tamaño del Bloque

Esto se puede expresar de la siguiente manera:

$$\left.\begin{array}{l} \boldsymbol{V\ x^m\ \rightarrow 0} \\ \boldsymbol{m \rightarrow \infty} \\ \boldsymbol{\frac{1}{m} \rightarrow 0} \end{array}\right\} \begin{array}{l} \text{Más Lento} \\ \\ \text{Más Rápido} \end{array} \qquad (3)$$

Para verificar la relación anterior, se realizaron tres agregaciones. Cada una representa uno de los procesos agregados descritos anteriormente. Cada proceso se encuentra determinado por el tamaño del bloque. En la tabla **1** se puede observar que las varianzas de los procesos agregados $\boldsymbol{Var\ X^m}$ tienden a cero, de igual forma que el inverso del tamaño del bloque $\boldsymbol{1/m}$. Pero en el caso de las varianzas, en efecto, se muestra que la tasa de decrecimiento es más lenta que la del inverso del tamaño del bloque. Se puede observar que, mientras la primera decrece en $\boldsymbol{29,67\%}$ cuando se aumenta el tamaño del bloque de **10** a **60**, el inverso del tamaño del bloque decrece en un $\boldsymbol{83,33\%}$. Por otro lado, cuando el tamaño del bloque aumenta de **60** a **600** la varianza del proceso agregado decrece en $\boldsymbol{19,01\%}$, mientras que el reciproco del tamaño del bloque decrece en **90%**. Esto nos conduce a otra evidencia de que nos encontramos con una serie de tiempo que tiene características de autosimilaridad.

El tráfico autosimilar se caracteriza por la presencia permanente de ráfagas a través de diferentes escalas del tiempo.

Así mismo, las propiedades que definen a los procesos autosimilares son las dos siguientes:

Dependencia a corto plazo y
Dependencia a largo plazo

Tabla 1. Estadísticas Descriptivas de los Procesos Agregados

Grupos	*Datos*	$Var(X^m)$	*Decrecimiento* $\%(Var\ X^m)$	$1/m$	*Decrecimiento* $\%\ 1/m$
$Seg\ m = 10$	4319	33,741	– –	0,1	– –
$Min\ m = 60$	719	23,731	29,67	0,0167	83,33
Diez Min $m = 600$	71	19,011	19,89	0,0017	90

Estas propiedades se hacen presentes en el tráfico de datos, y es posible determinarlas mediante la estimación del parámetro de Hurst $\boldsymbol{0.5 < H < 1}$), , que indica el grado de autosimilaridad de una serie, (Moreno. 2007).

Las propiedades mencionadas anteriormente se refieren a la correlación que existe entre los diversos eventos que componen la serie. La Dependencia a corto plazo, significa que la auto-correlación entre el tiempo de llegada de dos eventos sucesivos es muy cercana a cero. Mientras que la Dependencia a largo plazo, expresa que la auto-correlación entre los sucesivos eventos va decayendo en forma exponencial.

Es conveniente recordar que la autosimilitud y la fractalidad describen el fenómeno en el que cierta propiedad de un objeto se preserva con respecto a la escalización en el tiempo o en el espacio. (Alzate. 2001).

Según ese mismo autor, se pueden observar dos tipos de autosimilitud, Determinística y Estocástica, las cuales se explican a continuación:

Autosimilitud Determinística:

Esta es la forma más sencilla de autosimilitud, que se puede obtener por simple construcción, mediante la iteración de cierto comportamiento. En este caso la serie de tiempo se replica de manera exacta y basta con tomar una porción de la misma y replicarla, para poder reproducirla a una mayor escala.

Autosimilitud Estocástica

Similar al caso anterior, pero ahora cada segmento de la serie de tiempo se reproduce de forma más o menos exacta, es decir de forma aproximada. Dentro de la autosimilitud estocástica a su vez se pueden distinguir varios tipos, que se listan a continuación:
Autosimilitud de Segundo Orden.
Autosimilitud Asintótica de Segundo Orden.
Dependencia de Largo Rango.

Una vez que se ha determinado que la serie de tiempo es autosimilar, es necesario verificar qué tipo de autosimilaridad presenta la misma.

Definición 4:

Un proceso es Exactamente Autosimilar de Segundo Orden si:

$$Var\ x^m = \frac{Var\ x}{m^\beta} \quad (4)$$

Es decir que la varianza del proceso agregado es igual a la varianza de la serie original entre el tamaño del bloque elevado a la potencia beta.

Donde:

$$\beta = 2\ 1 - H \ . \quad (5)$$

$$H: Parámetro\ de\ Hurst$$

Este parámetro, es un valor tal que $(0.5 < H < 1)$, de manera que mientras más cercano a uno se encuentre este valor, mayor será el grado de autosimilaridad de la serie. Sin embargo, es conveniente mencionar también aquí, que en el caso cuando el parámetro toma valores entre $(0.5 < H < 1)$, a distintas granularidades o agrupamientos, el comportamiento de la serie de tiempo es similar, en el sentido que sus medidas descriptivas, como la media y la varianza, son parecidas en cada granularidad. Las series con este tipo de comportamiento son denominadas series o procesos persistentes. En este tipo de series o procesos, es de esperar que las secuencias de eventos se encuentren correlacionadas. En consecuencia, un proceso de crecimiento debería ser seguido por otro período análogo. Por otra parte, un valor de $(H = 0.5)$ en el parámetro, refleja que la serie de tiempo puede ser considerada como un proceso aleatorio y por tanto no presenta grado de dependencia alguno entre la ocurrencia de los eventos, por lo que en este caso, la autocorrelación en dicha serie de tiempo es cero. Por último $(0 \leq H < 0.5)$ indica que la serie de tiempo presenta antipersistencia es decir, que se espera que un período de crecimiento venga seguido por otro de decrecimiento. Este tipo de comportamiento es caracterizado por señales irregulares.

Por lo denotado en 5 , en un primer paso se deberá estimar el parámetro β, y luego en función de este valor se estimará el parámetro de Hurst H .

Otra definición que debe cumplirse de forma conjunta con 4 es la siguiente:

Definición 5:

La Autocorrelación del proceso agregado debe ser igual a la autocorrelación de la serie original (de retardo k)

$$R\ x^m, k = R\ x, k \quad (6)$$

Para verificar 4 se estimará el parámetro β. Aunque existen diversas técnicas para la estimación de β y de H, en este estudio se utilizará la técnica del diagrama de Varianza - Tiempo. La cual consiste en graficar $Log\ Var\ x^m \quad Vs\ Log\ m$. Este gráfico se genera a partir de una serie de datos $x(t)$ para distintos valores de m.

En la literatura especializada, referida a las series autosimilares, existen varios métodos que sirven para la estimación del parámetro de Hurst, entre los que se puede mencionar la Gráfica R/S, el Periodograma y el método de la Gráfica de la Varianza. En este trabajo se utilizó este último, pues es de fácil aplicación y sencilla comprensión para el lector. Yan, (2009), afirma que la técnica del diagrama de Varianza es uno de los más utilizados para la estimación del Parámetro de Hurst.

Dado que en este trabajo se generaron los tres procesos agregados descritos anteriormente, el resultado es una línea recta de regresión cuya pendiente es $-\beta$, quedando el gráfico como se muestra en la Fig. 4.

En este caso se puede observar que la pendiente es de

$$\beta = -0.13$$

En consecuencia el parámetro de Hurst (H) será:

$$H = 1 - \frac{\beta}{2} \qquad H = 0.93 \quad (7)$$

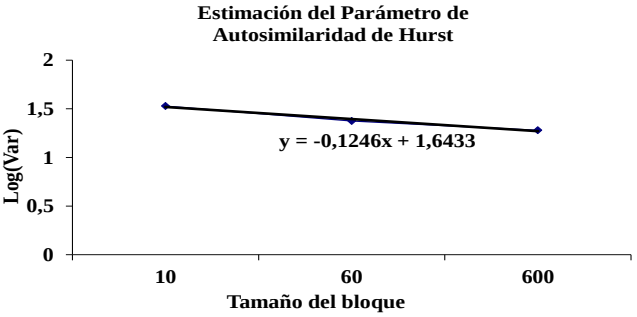

Fig. 4.- Estimación del Parámetro de Autosimilitud de Hurst

Como se puede observar en la Tabla 2, no se cumple la condición 4 para ninguno de los procesos agregados. Se puede inferir que la serie originalmente establecida, no es exactamente autosimilar de segundo orden.

Tabla 2.
Cálculos para determinar Autosimilitud Exacta de Segundo Orden

M	β	$Var(X^m)$	$Var(X)$	$\frac{Var(X^m)}{m^\beta}$
10		33,741		48,503
60	0,13	23,731	65,429	38,425
600		19,011		28,485

Ahora para verificar que se cumple la definición 5 , se realizará el estudio de los gráficos de autocorrelación, tanto de la serie original como la de cada uno de los procesos agregados.

En las figuras $5a$ a $5d$, se puede observar que las

autocorrelaciones de los diferentes procesos agregados no tienden a parecerse a la autocorrelación de la serie original. Por tanto, dado que no se cumple ninguna de las definiciones **4** y **5**, se puede concluir que la serie de tiempo no presenta un comportamiento exactamente autosimilar de segundo orden.

En consecuencia, es necesario entonces determinar la Autosimilitud Asintótica de Segundo Orden.

Definición 6:

$\boldsymbol{X(t)}$ es asintóticamente autosimilar de segundo orden si:

$$\lim_{m\to\infty} r\ k^m = r(k) \quad (8)$$

Es decir que la función de autocorrelación de los procesos agregados será igual, en la medida que $\boldsymbol{m}$ crece, a la función de autocorrelación de la serie original.

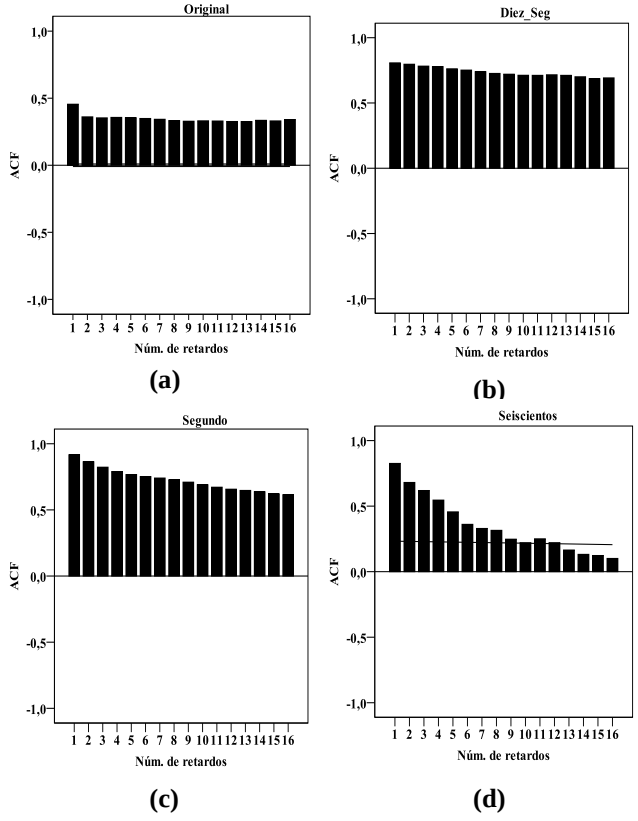

Fig. 5.- Autocorrelación. (a) Serie Original; (b) Serie m = 10; (c) Serie m = 60 y (d) Serie m=600

En la Fig. **5** se puede notar claramente que la función de autocorrelación no tiende a parecerse a la autocorrelación de la serie original, a medida que el tamaño de los procesos agregados se incrementa, es decir, cuando $\boldsymbol{m} => +\infty$. Esto nos proporciona evidencia para concluir que la serie de datos $\boldsymbol{x(t)}$ tampoco es asintóticamente autosimilar de segundo orden.

Por tanto, se debe determinar ahora si existe Dependencia de Largo Rango.

Definición 7:

Un proceso estocástico presenta Dependencia de Largo Rango si su autocorrelación decae a un ritmo lento, (inferior al de la exponencial).

La Fig. 6 muestra la distribución de la función de autocorrelación de la serie de tiempo X(t) a distintas granularidades. Adicionalmente, esta función de autocorrelación es comparada con el comportamiento de la función exponencial, que en este caso en particular, es ($\boldsymbol{e^{-x}}$).

Es conveniente resaltar que para las distintas granularidades que se propusieron en este estudio, se puede observar en la figura mencionada que en todos los casos, la función de autocorrelación decae a un ritmo más lento que el de la exponencial considerada. Por lo que se puede concluir que existe evidencia que indica que la serie original $\mathbf{X\ t}$: Número de solicitudes en el intervalo de tiempo $(\mathbf{t-1}, \mathbf{t})$ tiene características de autosimilaridad con Dependencia de Largo Rango.

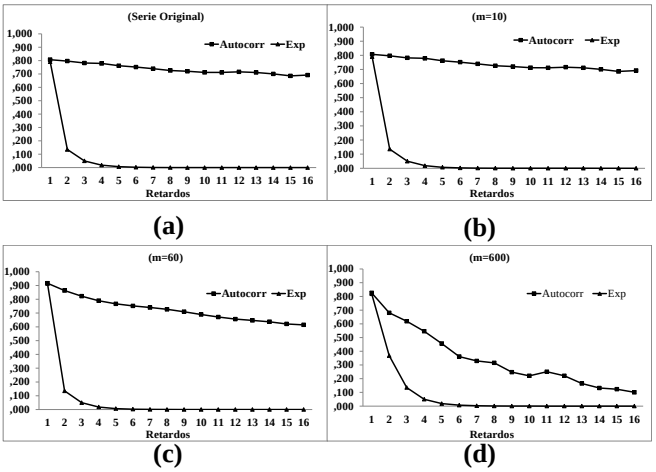

Fig. 6.- Comparación entre la distribución Exponencial y: la Función de Autocorrelación a distintas granularidades

4 Conclusiones

Las series de tiempo que tienen características de autosimilaridad, tienen la ventaja de tomar en cuenta las estructuras de correlación, que no son consideradas por los modelos de tráfico tradicionales, aplicados a las redes modernas.

En una serie autosimilar, el parámetro de Hurst **H** determina el grado de autosimilaridad de la serie de datos. Mientras más cercano a 1 se encuentre el valor del paráme-

tro, mayor será el grado de autosimilaridad. En el caso de que $0.5< H <1$ la serie de tiempo es catalogada como proceso persistente. Si $\mathbf{H = 0.5}$ existe una dependencia de corto rango. En ese caso la autocorrelación entre un evento y el siguiente es estadísticamente cero. Cuando ($0\leq H <0.5$) es un proceso antipersistente. Y si $\mathbf{H > 1}$, se pierde la estacionariedad de la serie. Este es un supuesto importante en los procesos autosimilares.

Para la estimación del parámetro de autosimilaridad $\boldsymbol{H}$ a pesar de que existen varias técnicas de estimación, en este proyecto se utilizó la técnica del diagrama Varianza - Tiempo, por su sencillez y facilidad para realizar los cálculos.

En el presente artículo se trabajó con una serie de tiempo estocástica y se demostró que tenía características de autosimilaridad. Se pudo determinar que la varianza de las muestras de los procesos agregados de tamaño $\boldsymbol{m}$, decrecen o tienden a cero de forma más lenta que el recíproco del tamaño del bloque. Esta serie autosimilar tiene un parámetro de autosimilaridad de Hurst de $\boldsymbol{H = 0.93}$ (Proceso Persistente).

Adicionalmente se pudo demostrar que dentro de la autosimilaridad, esta serie se puede caracterizar como una de Dependencia de Largo Rango, puesto que la función de autocorrelación de la serie de los procesos agregados, decae a un ritmo inferior al de la exponencial. Esta conducta de las diversas autocorrelaciones se observa de forma más notoria, a medida que aumenta el tamaño de los procesos agregados. Una de las ventajas de trabajar con procesos o series autosimilares, es que para estimar algunas de sus medidas descriptivas, como la media, basta con calcularla a partir de un segmento de la serie y ésta tiene que ser bastante similar a la media del total de la serie.

Los datos de las series que se encuentran referidos al tráfico en redes, se adaptan muy bien cuando se modelan mediante procesos autosimilares, dado que estos capturan la autocorrelación existente entre los tiempos entre llegadas de eventos sucesivos.

El objetivo principal de este artículo ha sido el de comprobar la hipótesis de que la serie de tiempo $\mathbf{X(t)}$ presenta características de serie autosimilar y en efecto se ha comprobado dicha hipótesis, una vez que se ha verificado la presencia de autosimilaridad en la serie de tiempo. Luego, el próximo paso sería, para un trabajo futuro, determinar el(los) modelo(s) que mejor se ajuste(n) al conjunto de datos, de manera que se pueda simular o predecir el comportamiento de la serie de tiempo. En esa etapa de simulación, se podrán obtener las herramientas necesarias para validar y calibrar los modelos que se hayan propuesto. Vale la pena mencionar que estos pasos de validación y calibración de los modelos tienen cabida solo después de que se hayan propuestos los mismos, y esto escapa al alcance del artículo presentado actualmente. Por lo que, la calibración y validación del modelo han sido consideradas como propuesta para trabajos posteriores.

Una importante contribución de este trabajo de investigación es que ayudar a definir un modelo estadístico que permita generar trazas sintéticas a partir de los datos estudiados, para luego tener datos disponibles por un período mayor al que se trabajó en la serie de datos y de esta forma poder simular los procesos durante lapsos mucho mayores sin la necesidad de utilizar de forma directa los dispositivos para realizar las mediciones.

Agradecimientos

Este trabajo se llevó a cabo, gracias al financiamiento del Centro de Desarrollo Científico, Humanístico, Tecnológico y de las Artes, CDCHTA, de la Universidad de Los Andes con el proyecto identificado bajo el número I-1366-13-02-B.

Referencias

Alzate M, 2001, Introducción al Tráfico Autosimilar en Redes de Comunicaciones, Revista INGENIERIA, Universidad Distrital, Vol. 6(2), pp. 6 – 17.

Ganger G. R, 1995 Generating Representative Synthetic Workloads. An Unsolved Problem, Proceedings of the Computer Measurement Group (CMG) Conference, December, pp. 1263-1269.

González O, Estudio del tráfico del nodo central de REDUNIV usando los métodos Whittle local y gráfico R/S, Fecha de Consulta: 28 febrero 2012. Disponible en: http://www.monografias.com/trabajos68/estudio-trafico-nodo-central-reduniv/estudio-trafico-nodo-central-reduniv2.shtml

Kavalanekar S, B. Worthington, Q. Zhang, V. Sharda, 2008, Characterization of Storage Workload Traces from Production Windows Servers,Workload Characterization IISWC,. IEEE International Symposium on. ISBN: 978-1-4244-2778-9/08/ 2008. pp. 119–128.

Moreno J, J. Padilla V. Escobar A. Correo, 2007, Caracterización y Simulación del Tráfico de Redes LAN mediante el modelo MMPP, Revista de la Facultad de Ingeniería de Antioquia.. Nro: 42. pp: 7-29.

Riska, A., E. Riedel, 2009, Evaluation of Disk-level Workloads at Different Time-scales, IEEE. 978-1-4244-5152-2/09. pp. 158-167.

(UMass 2011). UMassTraceRepository. 2007. Se encuentra en http://traces.cs.umass.edu/index.php/Storage/Storage. Fecha de Consulta: 20 Septiembre 2011.

Yan W, (2009, February). Revealing Self-similarity in NTFS File Operations. Inposter paper, Proceedings of the 7th USENIX Conference on File and Storage Technologies, San Francisco, CA.

Recibido: *11 de diciembre de 2012*

Revisado: *13 de junio de 2013*

Borrero Molina, Armando: *Doctor en Informática, Ingeniero de Sistemas.*

Quintero Gull, Carlos*: Lic. En Estadística. Correo-electrónico cgull@ula.ve.*