Modelado estocástico de las operaciones de entrada/salida sobre un disco

Laura Prada, José Daniel García, Alberto Núñez, Javier Fernández y Jesús Carretero

Ramón Flores

Departamento de Informática

Departamento de Estadística

Universidad Carlos III de Madrid

28270 Colmenarejo, Madrid

{laura.prada,josedaniel.garcia,alberto.nunez,javier.fernandez,

jesus.carretero,ramonjesus.flores}@uc3m.es

Resumen

Una de las técnicas más habituales de evaluación del rendimiento de los subsistemas de entrada/salida de un computador es la utilización de modelos de simulación detallados que tienen en cuenta características específicas de los dispositivos de almacenamiento como la geometría del disco, la división en zonas y los algoritmos utilizados por las políticas de la memoria caché del disco.

No obstante, dichos modelos requieren una mayor carga computacional y además están basados en discos que difieren en sus características y prestaciones de los discos más actuales.

Otra alternativa es el modelado del dispositivo de almacenamiento, incluyendo su caché, como un proceso estocástico. Esto permite la generación de los tiempos requeridos por las operaciones realizadas sobre el dispositivo con una menor carga computacional y con un alto nivel de fidelidad. De esta manera se puede alcanzar una mayor escalabilidad en la evaluación del rendimiento de subsistemas de entrada/salida mediante simulación.

En este trabajo se presenta una evaluación y modelado de un disco de tamaño moderado (20 GB). Los resultados obtenidos demuestran que los tiempos de acceso al disco evaluado constituyen un proceso estocástico estacionario y que este hecho es independiente de la activación o no de la caché de disco. Asi mismo los resultados sugieren que, en ausencia de caché, los tiempos de acceso siguen una distribución normal. En cambio, en presencia de caché, el tiempo de acceso se puede modelar como la composición de varias funciones de distribución, debido al efecto de la propia caché. En ambos casos nuestro modelo se ajusta a los datos experimentales con un error cuadrático medio menor al 1 %.

1. Introducción

Una de las técnicas más habituales en la evaluación del rendimiento de sistemas es la evaluación por simulación estocástica. La simulación presenta indudables ventajas puesto que no hace necesaria la disponibilidad del sistema real y permite comparar distintas alternativas de diseño. No obstante, un aspecto fundamental es el uso de parámetros que sean representativos del sistema a evaluar.

En muchos sistemas, el rendimiento del sistema está muy influenciado por el rendimiento del subsistema de entrada/salida, del que los dispositivos de almacenamiento son un elemento clave. Tradicionalmente, los discos se han modelado utilizando modelos analíticos bastante detallados basados en la geometría del dispositivo [5], incluyendo detalles de zo-

nificación [10, 1] llegando en algunos casos al nivel de emulación. En algunos casos estos modelos incluyen herramientas de caracterización automática [6].

No obstante, estos modelos se basan en discos de cierta antigüedad y con capacidades de almacenamiento bajas en comparación con los disponibles hoy día. Por ejemplo, la base de datos de parámetros de disco [2] de DiskSim contiene modelos de disco de entre los años 1992 y 1999 con capacidades de almacenamiento de entre 1 GB y 9 GB. Por otra parte, la mayoría de los modelos se basan en una simulación muy detallada tanto del comportamiento físico del disco como del comportamiento de la caché incorporada al mismo [7].

Cuando se plantea la evaluación de sistemas de almacenamiento escalables, los recursos computacionales consumidos por le modelo de disco crecen con el número de discos que se incluyen en el sistema simulado. En este caso, el uso de modelos muy detallados puede tener un impacto negativo en el tiempo necesario para poder llevar a cabo simulaciones arbitrariamente largas. El uso de un modelo tan detallado solamente se justifica si un modelo más sencillo no puede ofrecer una precisión similar.

Un enfoque alternativo es modelar el comportamiento del tiempo necesario para ejecutar las operaciones sobre un disco como un proceso estocástico. En un modelo de este tipo se puede considerar un disco, incluyendo su caché, como un fenómeno a caracterizar del que se desconocen los detalles.

En este artículo se presenta el método seguido para realizar la caracterización del tiempo de acceso a un disco como un proceso estocástico. El disco seleccionado para su caracterización es un disco de tamaño moderado (20 GB). Si bien, el objetivo es la caracterización de discos de gran tamaño, el trabajo se ha realizado primero con un disco de menor tamaño con objeto de poner a punto el método de trabajo.

2. Metodo de evaluación

Para realizar la evaluación se ha procedido a aplicar una carga de trabajo a un disco. La carga de trabajo utiliza un formato compatible con el definido por el *Storage Performance Council* [9]. El disco evaluado en este trabajo es un Maxtor 2B020H1, que tiene una capacidad de 20 GB, con un total de 39851760 sectores de 512 bytes direccionables mediante LBA.

Con objeto de medir el tiempo de acceso al disco evaluado, se ha montado como disco adicional en un sistema con un disco primario. En el disco primario se encuentra el sistema operativo y se utiliza para volcar el archivo con la traza de la evaluación.

Un aspecto importante en la evaluación es la necesidad de evitar que el sistema de ficheros o el gestor de bloques interfieran en las medidas. Es decir, es necesario que todas las operaciones que se realizan sobre el disco, se realicen físicamente sobre el dispositivo. En Linux esto se puede conseguir definiendo un dispositivo de caracteres (mediante `mknod`) y asociándolo al dispositivo a evaluar (mediante `raw`). De esta manera se puede acceder directamente al dispositivo mediante el uso de las llamadas `read` y `write`. Además se consigue que toda la entrada salida se realice directamente en el espacio de direcciones del proceso de evaluación, mediante DMA. Este método obliga a que las direcciones tanto en el disco como en memoria se encuentren alineadas a un múltiplo del tamaño de sector del dispositivo (512 bytes).

Otro aspecto a considerar es el efecto que tienen sobre el rendimiento del disco la caché de escritura y el búfer de prelectura. Para comprobar este efecto se han realizado las evaluaciones primeramente desactivando ambas opciones del dispositivo (mediante `hdparm`). Posteriormente se han repetido las evaluaciones activando dichas opciones.

Para minimizar las posibles interferencias de otros procesos en las mediciones realizadas el proceso de evaluación se ha ejecutado con un número mínimo de procesos activados (usando el modo `init 1`).

Para realizar la evaluación, se ha utilizado una carga de trabajo que forma parte de la especificación para ASU-1 del benchmark SPC-1 [8]. Las características de la carga de trabajo utilizada se muestran en la tabla 1.

Parámetro	Valor
Tamaño de transferencia	4 bloques
Porcentaje de lecturas	50 %
Dirección de transferencia	Uniforme
Número de operaciones	10^6

Cuadro 1: Características de la carga de trabajo

Para generar la carga de trabajo se ha utilizado el generador de números aleatorios Mersenne Twister [4] que tiene un periodo de $2^{19937} - 1$. Esto garantiza que no hay autocorrelación entre los datos generados ni hay correlación entre el tipo de operación y el número de bloque lógico de cada operación.

3. Resultados de la evaluación

La evaluación y posterior modelado se ha realizado primero inhibiendo la caché de disco. Aunque se podría argumentar que esta condición no es realista, la comparación de los resultados obtenidos con los que se obtienen en el caso en que la caché esta activada permite una mejor comprensión del efecto que tiene la caché sobre el comportamiento global del disco.

3.1. Evaluación sin caché de disco

Si se consideran las operaciones que se realizan sobre un disco, los tiempos de acceso de cada una de las peticiones constituyen un proceso estocástico. Para valorar si el proceso estocástico presenta un período transitorio inicial, hemos efectuado un análisis de las distintas realizaciones del experimento para la carga de trabajo propuesta a las que hemos aplicado el procedimiento de Welch [12, 3]. El procedimiento implica la evaluación de la media móvil con distintos tamaños de ventana y la verficiación de la convergencia de la misma. Las figuras 1, 2 y 3 muestran que la media móvil converge desde el principio de la serie para valores relativamente pequeños de tamaño de ventana. Por tanto, podemos concluir que el proceso no presenta período transitorio inicial y se trata de un proceso estocástico estacionario.

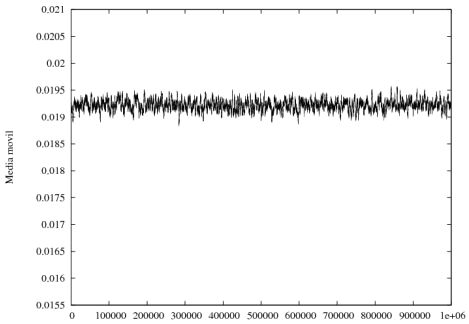

Figura 1: Media móvil para peticiones sin caché con tamaño de ventana 1000

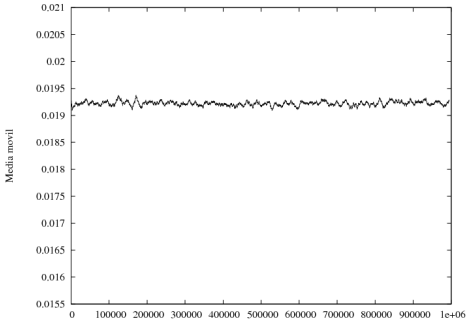

Figura 2: Media móvil para peticiones sin caché con tamaño de ventana 5000

El proceso presenta un tiempo medio de $19{,}225 \pm 0{,}008525 ms$ con un nivel de confianza del 95 %. Un intervalo de confizanza tan preciso supone un apoyo a la hipótesis de que el proceso es totalmente estacionario.

El estudio de los histogramas de los datos experimentales (figura 4), sugiere que se pueden ajustar bastante bien a una distribución normal. La tabla 2 muestra los estimadores de máxima verosimilitud para dicha distribución.

La coincidencia entre los datos experimentales y la distribución normal estimada (figura 5) es muy alta. Para comparar los datos experimentales con las mencionadas distribuciones se ha realizado una comparación de cuatiles

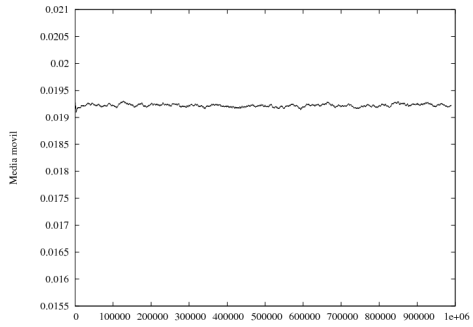

Figura 3: Media móvil para peticiones sin caché con tamaño de ventana 10000

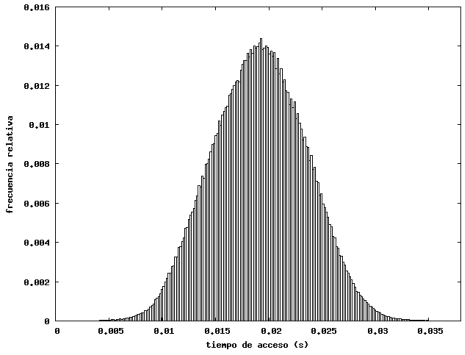

Figura 4: Histograma promedio para peticiones sin caché.

(figura 6).

Para medir el grado de ajuste de la distribución normal a los datos experimentales, se utiliza el error cuadrático medio. Además de ser un estimador del ajuste, esto permite comparar nuestros resultados con otros trabajos previos como el de Ruemmler y Wilkes [5]. En este caso el error cuadrático medio obtenido es de 146 microsegundos. Esto representa un error del 0,75 %. Este error es más que aceptable si se compara con el 2,6 % del trabajo anteriormente citado.

3.2. Con caché de disco

Las evaluaciones realizadas se han repetido activando las cachés de disco (tanto de lectura como de escritura). Aplicando el procedimiento de Welch [12, 3] a los resultados de las evaluaciónes se han obtenido representaciones de la media móvil para distintos valores de ventana (figuras 8 y 9).

Parámetro	Valor
$\hat{\mu}$	0.019221
$\hat{\sigma}$	0.004346

Cuadro 2: Estimadores de máxima verosimilitud de operaciones sin caché para la distribución normal.

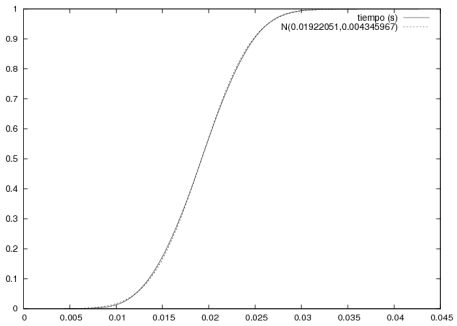

Figura 5: Destribución experimental y distribución normal estimada.

El proceso presenta un tiempo medio de $15{,}61943 \pm 0{,}01804 ms$ con un nivel de confianza del 95 %. Un intervalo de confizanza tan preciso supone un apoyo a la hipótesis de que el proceso es totalmente estacionario.

El histograma de los datos experimentales (figura 10) sugiere que la distribución del tiempo de acceso se deriva de la composición de varias distribuciones individuales. Esto se debe, sin duda, al comportamiento de la caché de disco. Por una parte, existe un número elevado de peticiones (más del 16 %) para el que el tiempo de acceso es muy pequeño (menor de 5 ms). Por otra parte, el resto de la distribución podría representarse como la suma de la distribución para el caso sin caché más una distribución que incorpora frecuencias altas para valores ligeramente menores y que puede deberse al comportamiento de la caché en el tratamiento de las peticiones de escritura.

Como el objetivo de este trabajo es la ge-

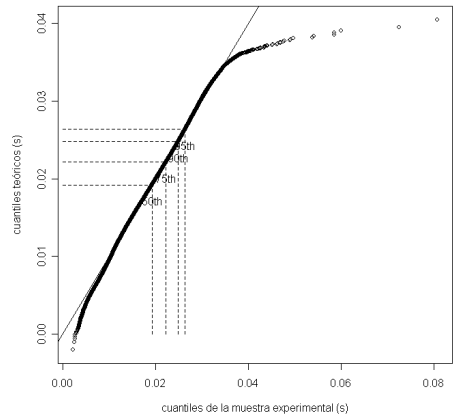

Figura 6: Gráfico cuantil-cuantil para la distribución de normal sin caché.

neración de valores aleatorios que sigan la distribución del tiempo de acceso, se ha optado por utilizar un generador de números aleatorios basado en una muestra empírica. No obstante, se considera interesante el modelado de la distribución obtenida como una mixtura o una composición de distribuciones, lo que queda para un trabajo posterior.

Para generar valores aleatorios que sigan una cierta distribución experimental se ha utilizado como base el entorno de simulación OMNET++ [11] que ofrece la posibilidad de generar números aleatorios a partir de un conjunto de datos representados en forma de histograma. Este enfoque hace que el tiempo de iniciación de una simulación se incremente, al ser necesario cargar una muestra experimental al iniciar el simulador.

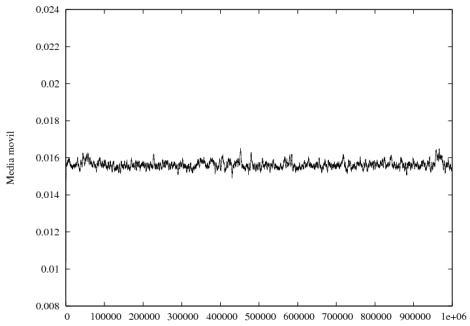

Figura 7: Media móvil para peticiones con caché con tamaño de ventana 1000

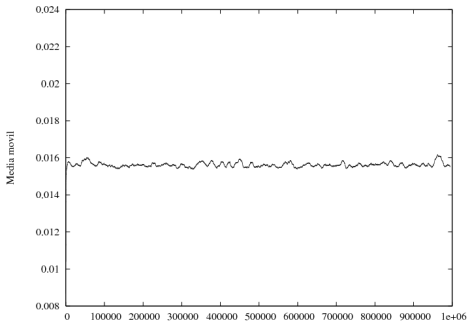

Figura 8: Media móvil para peticiones con caché con tamaño de ventana 5000

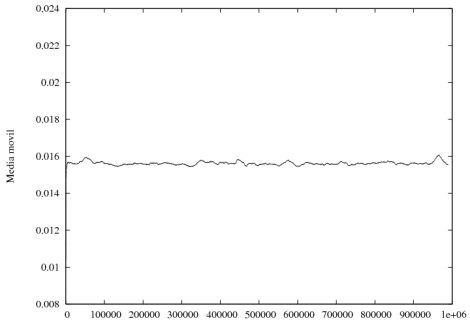

Figura 9: Media móvil para peticiones con caché con tamaño de ventana 10000

La figura 11 muestra una comparación de los datos experimentales con los datos suministrados por el generador de números aleatorios basado en histograma. Puede apreciarse que la coincidencia de cuantiles es prácticamente total.

Para medir el grado de ajuste de la distribución generada a los datos experimentales, se utiliza el error cuadrático medio. El error cuadrático medio obtenido es de 126 microsegundos. Esto representa un error del 0,65 %.

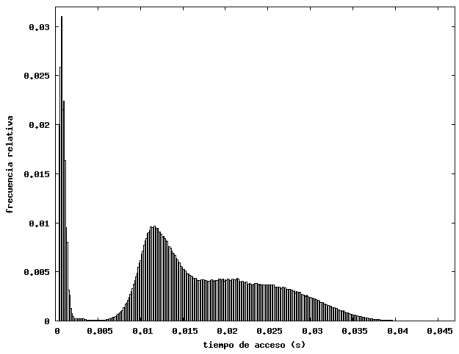

Figura 10: Histograma promedio para peticiones con caché.

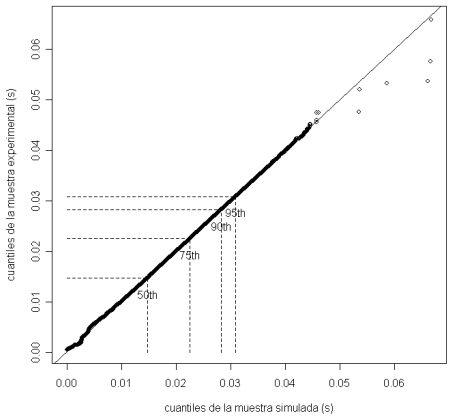

Figura 11: Gráfico cuantil-cuantil para la generación de valores aleatorios basada en histograma.

4. Conclusiones

En este trabajo se ha presentado un método para la evaluación y modelado del tiempo de acceso a un dispositivo de almacenamiento como un proceso estocástico. Las evaluaciones se han realizado tanto inhibiendo como permitiendo el uso de la caché de disco. Si bien no parece tener sentido, al menos de forma general, la inhibición de la caché de disco, el modelado de este caso supone una ayuda en la comprensión del comportamiento del dispositivo.

En los dos casos analizados se ha observado que el tiempo de acceso al dispositivo constituye un proceso estocástico completamente estacionario en el que no se observa la existencia de un período transitorio inicial.

En el caso en el que la caché de disco se encuentra inhibida, se ha comprobado que el tiempo de acceso se puede modelar bastante bien mediante la distribución normal, siendo el error relativo cometido de 0.75 %.

Al activar la caché de disco, el modelo que sigue la distribución no se ajusta directamente a ninguna distribución conocida. Más bien parece que el efecto del uso de la caché provoca la superposición de varias funciones de distribución, lo que podría modelarse mediante una mixtura de distribuciones o una combinación lineal de distribuciones. Para este caso se ha procedido a la generación de números aleatorios a partir de los datos de un histograma, lo que permite construir una fuente de números aleatorios aun cuando su función de distribución sea conocida. El error relativo cometido usando este método es del 0.65 %.

5. Tabajo futuro

Las evaluaciones realizadas en este trabajo corresponden a la utilización de una de las trazas sintéticas del estándar SPC-1. En un próximo trabajo pretendemos confirmar los resultados obtenidos utilizando una carga de trabajo complete formada por la combinación de todos los flujos de peticiones de SPC-1 que son más representativos de la utilización real de un subsitema de entrada/salida.

Por otra parte, somos conscientes de la necesidad de validar los resultados obtenidos frente a discos de mayor tamaño. Por esta razón el siguiente paso será la evaluación y construcción de un modelo para un disco de 300 GB.

Otro aspecto en el que se hace necesario avanzar es el modelado del tiempo de acceso para el caso de activación de cachés de disco. Si bien es cierto que hemos sido capaces simular con bastante precisión el tiempo de acceso al disco, el método se basa en cargar inicialmente en el simulador un conjunto de datos

representativo del histograma de los datos experimentales. La descomposición de la función de distribución, ya sea como una mixtura o como una combinación lineal de funciones de distribución, permitiría alcanzar un método de generación de números aleatorios sin requisitos de tiempo de iniciación ni de consumo de memoria.

Por último pretendemos incluir los resultados obtenidos en la herramienta de simulación de sistemas de entrada/salida escalables que se está desarrollando en el grupo ARCOS de la Universidad Carlos III de Madrid.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Educación y Ciencia mediante el proyecto TIN2004-02156 (*Almacenamiento de altas prestaciones en Entornos GRID*) y por la Comunidad Autónoma de Madrid mediante el proyecto *Técnicas de optimización de la entrada/salida en aplicaciones para entornos de computación de altas prestaciones*.

Referencias

[1] The DiskSim Simulation Environment. `http://www.pdl.cmu.edu/DiskSim/index.html`.

[2] Database of Validated Disk Parameters for DiskSim v2.0 `http://www.pdl.cmu.edu/DiskSim/diskspecs.html`.

[3] Law, A. M. y Kelton, W. D. *Simulation Modeling and Analysis*, McGraw-Hill, 2000.

[4] Matsumoto, M. y Nishimura T. *Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudorandom Number Generator*. ACM Trans. on Modeling and Computer Simulation, 8(1):3–30, 1998.

[5] Ruemmler, C. y Wilkes, J. *An introduction to disk drive modeling*. Computer, 27(3):17–28. 1994.

[6] Schindler, J. y Ganger, G.R *Automated disk drive characterization* CMU SCS Technical Report CMU-CS-99-176, Diciembre, 1999.

[7] Shiver, E., Merchants,A. y Wilkes, J. An analytic behavior model for disk drives with readahead caches and request reordering ACM SIGMETRICS Performance Evaluation Review, 26(1):182–191. 1998.

An analytic behavior model for disk drives with readahead caches and request reordering

[8] Storage Performance Council *Open Benchmark-1 (SPC-1) Official Specification*. Revisión 1.10.1. 2006.

[9] Storage Performance Council, *SPC Trace File Format Specification. Revisión 1.0.1.* 2002.

[10] Triantafillou, P., Christodoulakis, S., y Georgiadis, C.A. *A comprehensive analytical performance model for disk devices under random workloads* IEEE Transactions on Knowledge and Data Engineering, 14(1):140–155. 2002.

[11] OMNeT++ `http://www.omnetpp.org`.

[12] Welch, P. D., *The statistical analysis of Simulation Results* en *Computer Performance Modeling Handbook*, Academic Press, 1983.