Entradas/Salidas en la computación de alto desempeño

Harold Castro

Departamento de Ingeniería de Sistemas y Computación
Universidad de los Andes
A.A. 4976 Bogotá - Colombia
e-mail: hcastro@uniandes.edu.co

Resumen: este artículo presenta un estudio de las Entradas/Salidas (E/S) para los sistemas de cómputo de alto desempeño de propósito general. Las E/S son hoy en día un cuello de botella para el cálculo de alto desempeño, y para eliminarlo se requieren nuevas tecnologías para los subsistemas de E/S. Luego de un análisis de los subsistemas de E/S clásicos, presentamos el principio de las E/S paralelas y de los vectores de discos. Del estudio de las arquitecturas existentes, descartamos aquellas que no pueden ser aplicadas a un sistema de uso general e identificamos las organizaciones que se adaptan mejor para los sistemas de alto desempeño. Finalmente, proponemos una primera clasificación arquitectural de los subsistemas de E/S; basados en esta clasificación, introducimos una arquitectura universal y extensible apropiada para los subsistemas de E/S de alto desempeño.

Introducción

La afanosa búsqueda de potencia de cálculo en los computadores modernos contrasta fuertemente con las escasas mejoras aportadas a los sistemas de Entrada/Salida. En los últimos años la influencia de las E/S en el desempeño de ciertas aplicaciones ha sido ignorada, y las principales innovaciones se han hecho en el campo de la algorítmica, los compiladores y sobre todo en la velocidad de cálculo de los procesadores. El resultado es un desequilibrio penalizante entre las capacidades de los procesadores y la de los sistemas de E/S.

La importancia de un equilibrio entre la potencia de cálculo y el ancho de banda de las E/S es reconocida desde hace tiempo [Am67]. En efecto, para un gran número de aplicaciones, el desempeño global de los sistemas está fuertemente limitados por falta de una transferencia suficientemente rápida entre las unidades de cálculo y los dispositivos de almacenamiento.

Las arquitecturas diseñadas para realizar computación de alto desempeño[1] (máquinas paralelas, masivamente paralelas y supercomputadores) son particularmente sensibles a este desequilibrio. Dada la altísima capacidad de cómputo de estas máquinas, ellas necesitan un subsistema de E/S muy eficaz para que las aplicaciones puedan aprovechar al máximo su rendimiento. Aumentar el tamaño de la memoria principal de los procesadores ya no es suficiente para disimular las actuales limitaciones, y nuevas organizaciones, mejor estructuradas y mejor adaptadas, deben ser propuestas para responder eficazmente a la evolución en la potencia de cómputo.

Aumentar estáticamente la capacidad de un subsistema de E/S no es suficiente. Conceptos como extensibilidad y generalidad de una arquitectura deben ser reconsiderados para tomar en cuenta estos subsistemas. La capacidad de las E/S debe aumentar en función del poder de cómputo de la arquitectura. Para esto, es necesario que el subsistema de E/S acelere la velocidad de acceso a las informaciones almacenadas en disco, en la misma proporción que son aumentadas las capacidades de cálculo. De otra parte, los sistemas deben ser suficientemente flexibles para ofrecer buen rendimiento para diferentes tipos de solicitudes de E/S [Ca95].

El artículo está organizado de la siguiente manera: primero que todo presentamos los dispositivos de almacenamiento, los factores que determinan su desempeño y las soluciones clásicas adoptadas para mejorarlo. Luego estudiamos el principio de las E/S paralelas, los vectores de discos y las arquitecturas RAID. La última sección presenta una clasificación de las arquitecturas de los subsistemas de E/S paralelas e introduce una arquitectura universal para estos subsistemas.

---

[1] High Performance Computing (HPC)

Las unidades de disco

En el corazón de los subsistemas de E/S, las unidades de disco han suscitado el interés de la comunidad científica que se ocupa del desempeño de las E/S. Además de la tecnología de los dispositivos de almacenamiento como unidades independientes, es necesario estudiar más en detalle el agrupamiento de estas unidades para el almacenamiento de informaciones distribuidas.

Centradas principalmente en la celeridad de las operaciones de lectura y escritura, numerosas propuestas han sido implementadas para aumentar la velocidad de los dispositivos estándares de E/S. Aunque hoy es posible, con la ayuda de semiconductores, alcanzar anchos de banda próximos a los 50 Mbits/s, esto representa apenas el doble de la velocidad de un disco clásico.

La diferencia de evolución con respecto a los procesadores es enorme. Mientras los procesadores doblan anualmente su velocidad de cálculo, el tiempo de posicionamiento de las cabezas de lectura/escritura en un disco sólo se reduce a la mitad cada diez años. Hoy se reconoce que a menos que aparezca un invento 'milagroso' en el campo de la tecnología de discos, esta situación continuará así y las limitaciones mecánicas de los dispositivos de E/S no seguirán el ritmo de los desarrollos tecnológicos de las unidades de cómputo.

Factores que determinan el desempeño de un disco

Antes de presentar las mejoras introducidas a la arquitectura de los discos magnéticos, consideramos de gran utilidad comprender bien los elementos que son determinantes en su desempeño. El desempeño de un disco es una función del tiempo de servicio de la unidad. Tres factores dominan el cálculo de este tiempo de servicio: el tiempo de posicionamiento (seek time), la latencia de rotación (rotational time) y el tiempo efectivo de la transferencia (transfer time)[2].

El tiempo de posicionamiento es el tiempo necesario para desplazar las cabezas de lectura y escritura sobre la pista donde se encuentra la información que debe ser transferida. Este tiempo es igual al tiempo de arranque del brazo (aprox. 5 ms) más el tiempo de cruzar las pistas que separan la posición final de la posición actual de las cabezas. En función del número de pistas del disco, este tiempo varía en promedio de 7 a 20 ms.

La latencia de rotación está definida por el tiempo que pasa desde el momento en que la cabeza se sitúa en la pista donde se encuentra la información hasta el momento en que el sector a leer/escribir pasa debajo de la cabeza. Si consideramos que la velocidad de rotación es aproximadamente de 5400 rpm, el tiempo máximo de rotación sería de 10 ms con una media de 5ms.

Finalmente, el tiempo de transferencia es el tiempo necesario para realizar la transferencia efectiva de los bytes solicitados de la memoria al disco o viceversa. Este es el único parámetro que depende del número de bytes a transferir; el tiempo de posicionamiento y el de rotación son constantes independientemente del tamaño de los datos. En consecuencia, la transferencia de grandes bloques de información es mejor amortizada que la de pequeños bloques de datos.

Las soluciones clásicas

Diferentes soluciones han sido propuestas para mejorar el tiempo de servicio de un disco; cada una aporta a uno o varios de los factores estudiados en la sección precedente. La reducción del tiempo de servicio disminuye automáticamente el tiempo de espera de las aplicaciones y en consecuencia se mejora el desempeño global del sistema. Las técnicas presentadas son: los discos de cabezas fijas, el aumento de la densidad de almacenamiento, los discos a conmutación electrónica de circuitos, los cachés de disco y el ordenamiento de solicitudes de E/S [KGP89].

Los discos de cabeza fija. La idea es sencilla: se coloca una cabeza de lectura/escritura sobre cada pista del disco. De esta manera, se elimina la necesidad de posicionar las cabezas y por lo tanto el tiempo de posicionamiento es nulo. Este modelo fue utilizado como memoria de respaldo para la memoria virtual, pero la aparición de discos con varias centenas de pistas, interpuso una barrera económica infranqueable para su implantación.

---

[2] En un sistema sobrecargado, la espera por la liberación de un disco puede ser un factor importante en el cálculo de este tiempo de servicio.

Transferencia en paralelo desde la pila de discos. Hoy en día existen dispositivos capaces de leer o escribir al mismo tiempo sobre las múltiples superficies de un disco. Esto aumenta el ancho de banda de la transferencia pero no tiene ninguna incidencia sobre el tiempo de posicionamiento ni sobre la latencia de rotación. Aunque esta solución ha sido utilizada por diferentes constructores, su realización es muy costosa pues necesita la existencia de varios brazos para desplazar las cabezas independientemente.

Aumentar la densidad de almacenamiento. La densidad de grabación de la información continua a aumentar gracias a la utilización de películas magnéticas de grabación más finas y gracias a las mejoras introducidas en las cabezas de lectura/escritura. Las cabezas actuales son más precisas y capaces de sobre-volar la superficie del disco de más cerca. La técnica utilizada para codificar la información también puede tener un efecto importante sobre la densidad realmente explotada. Entre más pequeño sea el tamaño de la información codificada, menor será el espacio necesario para almacenarla. De esta forma es posible obtener una ganancia de hasta 50%.

Aumentar la densidad de los discos disminuye el tiempo de transferencia porque en cada unidad de tiempo se leen/escriben más bits de información. Sin embargo, esta solución no concierne ni la latencia de rotación ni el tiempo de posicionamiento. En cambio, aparece adecuada para sistemas donde se manipulen grandes volúmenes de información en cada operación de E/S.

Discos de conmutación electrónica. Estos discos son construidos a partir de circuitos de memoria lentos, y pueden ser vistos como una memoria lenta o un disco de alta velocidad. Gracias a su velocidad con respecto a los discos clásicos, las operaciones de E/S sobre estas unidades no son asincrónicas con la unidad de cálculo y en consecuencia no hay que pagar ninguno de los sobrecostos debidos al control del sistema operacional. Estos discos han sido utilizados tradicionalmente como soporte para la memoria virtual.

La estabilidad de las informaciones almacenadas en estas unidades es garantizada por una batería externa, lo que no elimina completamente el riesgo de pérdida de la información ante una baja de tensión. Si las baterías no están suficientemente cargadas en el momento de un corte de electricidad, toda la información será irremediablemente perdida. Esta es la razón por la que estos discos sean utilizados guardar datos temporales exclusivamente. El costo de estas unidades es otro problema: 1 Mbit almacenado en uno de estos discos cuesta aproximadamente 15 veces el valor de almacenar ese mismo Mbit en un disco magnético. Este hecho ha restringido su utilización al dominio de los super-computadores de los grandes centros de cómputo.

Los cachés de disco. Son simplemente zonas de memoria intermedias entre los discos y la memoria principal. Cada vez que un bloque de disco es accedido, el sistema guarda una copia de éste en una zona propia, de manera que toda nueva referencia a este bloque sea resuelta gracias a la copia guardada por el sistema operacional. La eficiencia de estos sistemas depende entonces de la probabilidad de resolver una operación de E/S con las copias de bloques disco mantenidas por el sistema. Esta característica hace que esta solución sea dependiente de la forma como se acceden los datos, es decir, de las aplicaciones. Los cachés de disco pueden ser muy útiles cuando utilizan baterías externas como las de los discos de conmutación electrónica. Un caché no volátil permitiría escrituras rápidas puesto que se eliminaría el riesgo de pérdida de las modificaciones por falla del sistema (a condición por supuesto que las baterías sean seguras).

Los cachés de E/S son una buena solución para las máquinas monoprocesadores, y su administración ha generado una intensa actividad de investigación en la comunidad internacional. Es evidente que la computación de alto desempeño requiere de tales sistemas, pero dentro del contexto de paralelismo introducido por las nuevas necesidades de cálculo, la arquitectura de los sistemas de cachés debe ser estudiada de nuevo para adaptarla a las características específicas de estos sistemas de computación de alto desempeño.

Ordenamiento de solicitudes. Los retardos debidos a los movimientos mecánicos de las unidades de disco pueden ser minimizados con un buen ordenamiento de las solicitudes. Un algoritmo que ordene la lista de solicitudes con el criterio de "primero el tiempo de posicionamiento más pequeño", reduce necesariamente el tiempo de posicionamiento global. Sin embargo, para que el ordenamiento de solicitudes sea eficaz, se necesitan largas listas de solicitudes, y el objetivo de las E/S de alto desempeño es precisamente reducir la longitud de esas listas.

Las E/S paralelas

De la misma forma que un programa es descompuesto en varias tareas para explotar su paralelismo, una operación de E/S puede ser dividida en varios procesos de E/S para ser ejecutados en paralelo [NNSI90]. El modelo está basado en la diversificación de las vías de E/S. De esta forma se obtiene un sistema con varios discos, cada uno con una vía de acceso diferente. Una operación de E/S será entonces dividida en varios procesos de E/S, el número de procesos que se deben generar dependerá del número de vías disponibles, y cada proceso seguirá una vía diferente para transferir los datos entre los discos y la memoria principal.

Para que la división de una operación de E/S sea eficaz, hay que asegurarse que las tareas asociadas a los procesos de E/S sean independientes. Esta independencia es indispensable porque los datos accedidos por cada proceso deben utilizar trayectorias distintas, y esto durante toda la transferencia.

El objetivo de una operación de E/S sobre una unidad de disco es la transferencia de una zona de datos entre la memoria principal y el (los) periférico(s) de almacenamiento (normalmente los discos). La trayectoria utilizada por esta zona comienza (o termina) en las unidades de disco, y si se quiere asegurar la independencia de una vía durante todo el recorrido de los datos, la zona de datos debe ser igualmente distribuida sobre las diferentes unidades de disco.

En un sistema de almacenamiento clásico, un archivo no es almacenado sobre varias unidades, a menos que su tamaño así lo exija. En un sistema de E/S paralelas, los bloques de un mismo archivo son almacenados sobre diferentes unidades de disco. El número de unidades de disco utilizadas determina el grado de paralelismo del sistema y el número de vías diferentes que los datos podrán utilizar durante una transferencia. El paralelismo ofrecido por las diferentes vías de E/S permite también que varias operaciones de lectura o escritura independientes (bloques a transferir, asociados a discos diferentes), sean efectuadas simultáneamente.

La unidad de división de un archivo puede ser escogida en función de los tipos particulares de acceso de cada archivo. Sin embargo, existen configuraciones materiales que pueden definir el tamaño de esta unidad; éste puede ir desde 1 bit hasta el tamaño del bloque físico de los discos.

Los vectores de discos

El aumento de la capacidad de almacenamiento y del desempeño de las unidades de disco para las grandes máquinas no han limitado el ámbito de investigación de los diseñadores de discos. Los computadores personales han creado un mercado para discos menos caros, por parte de un público muy exigente. Estos discos tienen una capacidad de almacenamiento reducida y su desempeño es menos impresionante que el de los discos grandes, pero su precio es abordable por el gran público (cf. tabla 1).

Características	IBM 3380	Fujitsu M2361A	Conners CP3100
Capacidad (Mb)	7500	600	100
Precio/Mb	$18-10	$20-17	$10-7
MTTF[3] previsto (horas)	30.000	20.000	30.000
Número de brazos	4	1	1
Operaciones de E/S por brazo (observadas)	30	24	20
Consumo (W)	6.600	10	10
Volumen (pies[3])	24	0,03	0,03

Tabla 1 Comparación de las diferentes unidades de disco

La tabla 1 compara tres sistemas de discos [PGK88]: el IBM 3380, el Fujitsu M2361A "Super Eagle" y el Conners CP3100, correspondientes a tres tipos de máquinas respectivamente: las grandes máquinas centrales (main frames), los minicomputadores y los computadores personales. Como es de esperar, los discos más caros tienen capacidades

[3] Mean Time To Fail

de almacenamiento y desempeños superiores a los de los discos económicos. En cambio, lo que es sorprendente, es que el ancho de banda por brazo de los grandes discos no sea ni siquiera el doble del de los discos económicos. Sobre la mayoría de los otros parámetros, los valores obtenidos para los discos económicos son superiores o al menos iguales (en calidad) al de los discos grandes. De otro lado, el pequeño tamaño y el bajo consumo de los discos económicos están lejos de ser factores despreciables, sobretodo si se considera que estos discos ya tienen integrados dentro de la unidad la mayor parte de las funciones de los controladores de las grandes máquinas.

Todas estas características condujeron a las proposiciones hechas a mitades de los años 80 que preconizaban la construcción de subsistemas de E/S de gran capacidad y alto desempeño con la agrupación de varios discos económicos [Ki86], [SG86]. Este agrupamiento recibió el nombre de arreglos o vectores de discos. Como lo muestran las informaciones en la tabla 1, un vector de 75 discos CP3100 tiene la misma capacidad de almacenamiento de un IBM 3380, con potencialmente 12 veces su banda pasante en operaciones de E/S. Esto a un costo inferior y por un consumo y espacio físico netamente inferiores.

El punto débil de esta solución es la fiabilidad del subsistema de E/S. El riesgo de un daño aumenta, pues las probabilidades de falla de los discos al interior de un vector son independientes. El tiempo promedio de funcionamiento de un disco (MTTF) es un parámetro de seguridad para la información almacenada. Podemos calcular su valor:

$$\text{MTTF(vector de discos)} = \frac{\text{MTTF de un disco}}{\text{Número de discos del vector}}$$

Por observación de la información en la tabla 1, el valor de MTTF para el vector de 75 discos es de 300 horas, o sea menos de 2 semanas. Es evidente que tal sistema no puede competir con el disco IBM original con un tiempo medio de servicio de 30.000 horas, o sea mas de 3 años antes que la unidad presente un problema. Los vectores de discos no son entonces una solución realista, a menos que se haga un esfuerzo considerable en el dominio de la tolerancia a fallas.

Los sistemas RAID[4] propuestos por D.A Patterson, G. Gibson, y R.H. Katz aparecen para resolver el impase [PGK88]. Gracias a la redundancia de la información, esto sistemas pueden ofrecer una disponibilidad de servicio superior a la de los discos independientes. Hoy existen cinco organizaciones diferentes de sistemas RAID, pero para comprenderlas bien, presentamos primero las diferentes taxonomías posibles de los vectores de discos.

Organizaciones de datos en un vector de discos

Partiendo de un sistema tradicional multidisco, haremos resaltar la idea básica de las otras tres organizaciones: los discos sincrónicos, los archivos entrelazados y los sistemas sincrónicos de archivos entrelazados.

Sistema tradicional. En este tipo de sistemas, cada archivo es almacenado completamente en un sólo disco; el disco es entonces una unidad independiente. Sencillo de implantar, este sistema ofrece un buen nivel de desempeño cuando las solicitudes de E/S hacen referencia a bloques de datos correctamente distribuidos en los diferentes discos. Por el contrario, cuando esas informaciones no están equitativamente distribuidas, ciertos discos pueden recibir muchas más solicitudes que otros. Estos sistemas se degradan entonces fuertemente en condiciones de inestabilidad de equilibrio. Como no se mejora ninguno de los tres factores de desempeño estudiados, el tiempo de servicio de un vector de discos con esta técnica tradicional es el mismo que el de un gran disco. Si se observa una mejora en el desempeño, sólo es debido a la reducción de la longitud de la lista de espera donde debe residir toda solicitud antes de ser tratada.

Discos sincrónicos. Muy útiles cuando el tiempo de transferencia es el factor principal en el tiempo de servicio de las E/S. En una organización de discos sincrónicos, los archivos están entrelazados byte a byte en el conjunto de discos del vector. Esta técnica impone una sincronización de todos los discos, de manera que todas las cabezas de lectura/escritura estén posicionadas sobre el mismo sector en todos los discos. Los discos funcionan así en conjunto, como si fueran uno solo, con un ancho de banda y una capacidad de almacenamiento multiplicados por el número de unidades. El sistema es equitativo con respecto al número de solicitudes pues cada solicitud es tratada simultáneamente por todos los discos [Ki86]. El tiempo de posicionamiento y la latencia de rotación no son modificados. Esta técnica no impone una sincronización total, una sincronización

[4] Redundant Array of Inexpensive Disks, o vector de discos económicos redundantes

parcial por subconjuntos sincrónicos del vector puede ser utilizada. El número de discos sincrónicos que forman una unidad se llama *grado de sincronismo (gs)*.

Archivos entrelazados. Esta técnica difiere de la precedente, no solamente porque no hay sincronización entre los discos, sino también por el tamaño de las unidades de entrelazamiento de los archivos; acá la unidad de división es el bloque físico. Con esta organización de datos, es posible la lectura o escritura en paralelo de varios bloques de un mismo archivo. Dos casos se distinguen cuando una solicitud de E/S es transmitida al sistema de discos, según si la solicitud hace referencia a uno o varios bloques de un archivo. En el primer caso, como sólo un bloque debe ser transferido, la solicitud es simplemente puesta en la lista de espera del disco correspondiente. Si por el contrario la solicitud necesita la transferencia de m bloques, ésta es dividida en m solicitudes, una por bloque, y cada una de ellas se coloca en la lista de espera del disco asociado al bloque que representa. En este caso, el resultado es un acceso paralelo, no sincrónico, a la información.

La ventaja de esta organización es que la carga de los discos es distribuida equitativamente. Aunque, a diferencia de un sistema tradicional, toque pagar el costo del tiempo de posicionamiento y de rotación por cada bloque, y no por cada solicitud, el tiempo de servicio se disminuye porque los tratamientos son efectuados en paralelo y la longitud de las listas de espera de las solicitudes es altamente reducida [LKB87]. El número de accesos simultáneos autorizados para un archivo se llama *grado de entrelazamiento (ge)*.

Los sistemas sincrónicos de archivos entrelazados. Una combinación de las dos técnicas precedentes es posible si se entrelaza la información sobre las diferentes unidades sincrónicas. Aunque las dos técnicas (sincronización y archivos entrelazados) tengan el mismo objetivo de reducción del tiempo de servicio, en el fondo son fundamentalmente diferentes. En los sistemas sincrónicos, un archivo es almacenado por completo en una unidad sincrónica lo que impide el acceso simultáneo a diferentes bloques de un mismo archivo. Su buen rendimiento de debe al principio de localidad de las referencias a un archivo, pues la transferencia de una gran cantidad de información sólo hace intervenir una vez los tiempos de posicionamiento y rotación. Por el contrario, la solución alternativa de archivos entrelazados, aunque ignora el principio de localidad, descompone el acceso a un gran volumen de información en varios accesos a bloques más pequeños, lo que permite la lectura/escritura de varios bloques en paralelo.

La figura 1 resume las diferentes organizaciones de vectores de discos. La notación *f.m* denota el *m-ésimo* bloque del archivo *f*, y *f* solo representa el archivo *f* completo. Los sistemas se caracterizan por la pareja (*gs*, *ge*) y por el número total de discos m del sistema. Una evaluación de estas organizaciones se encuentra en [NB89].

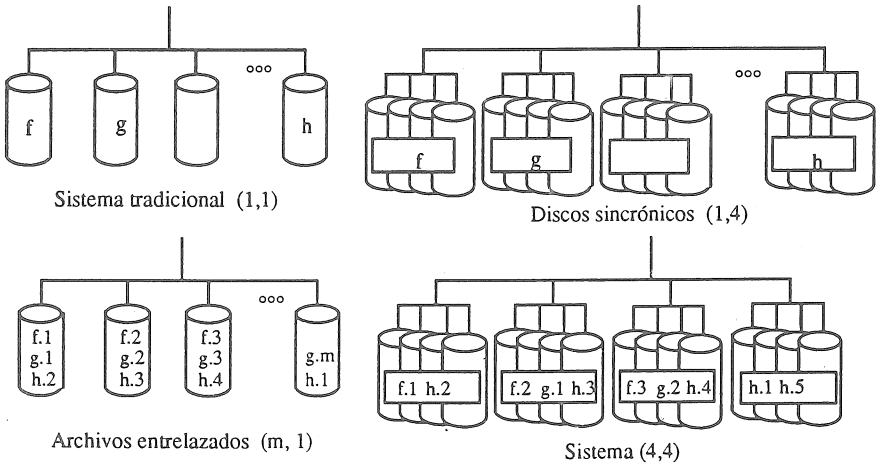

Figura 1 Organizaciones de los vectores de discos

Los sistemas RAID

Los vectores de discos económicos redundantes fueron diseñados para ofrecer una alta capacidad de E/S, garantizando al mismo tiempo una buena fiabilidad del sistema. El principio de estos sistemas es dividir el vector en *grupos fiables* asociando a cada uno de ellos, discos de verificación con información redundante.

Un sistema RAID se caracteriza por cuatro valores:

D el número total de discos de datos (sin los discos de verificación);

G el número de discos de datos en un grupo (sin los discos de verificación);

C el número de discos de verificación en un grupo;

N_G = D/G el número de grupos.

A partir de estos valores podemos medir la eficiencia de cada una de las organizaciones RAID existentes. En esta presentación estudiaremos especialmente la fiabilidad del sistema (MTTF), los sobrecostos incurridos para asegurar esta fiabilidad y el desempeño particular de cada disco del vector.

Habíamos visto que el punto débil de los vectores de disco era su MTTF demasiado corto. Para calcular el MTTF de un sistema RAID debemos tomar en cuenta el tiempo medio de reemplazo de un disco luego de una falla (MTTR[5]), o sea el tiempo necesario para que la información que contenía un disco dañado esté de nuevo disponible. Hay que destacar que este tipo de sistema perdería su fiabilidad si ocurriera una segunda falla antes del reemplazo de la primera unidad dañada, lo que hace necesario tener un valor acotado para MTTR. Una manera de minimizar este valor es mantener siempre un disco de reemplazo disponible, y así proceder a la recuperación de la información inmediatamente después de la falla. En [PGK88], los autores calculan el valor MTTF de un sistema RAID:

Tomando un valor de una hora para MTTR, el valor de MTTF ha sido calculado para diferentes configuraciones de sistemas RAID. Los valores obtenidos superan ampliamente las 30.000 horas de los discos independientes, llegando en varios casos a superar los 30 años de servicio. El problema de fiabilidad de los vectores de disco queda entonces resuelto, y para efectos prácticos no se considera ninguna diferencia a este nivel entre una organización RAID y otra.

El costo de la fiabilidad es fácil de determinar: éste es igual al costo de los discos de verificación. Como lo veremos durante la presentación de cada nivel de organización RAID, hay que tener en cuenta el porcentaje de capacidad de almacenamiento perdida a causa de las informaciones redundantes. En función de la organización RAID este valor puede variar de 4% a 50%.

El desempeño de un sistema de E/S varía de una aplicación a otra; un sistema de E/S puede responder correctamente a las necesidades de una aplicación específica, y ser completamente inadecuado para las exigencias de otra. A medida que avancemos en la presentación iremos destacando las ventajas y desventajas de cada una de las organizaciones.

Nivel 1: discos duplicados[6]

La utilización de discos duplicados es una técnica tradicional para asegurar la disponibilidad de los discos magnéticos. Es evidente que esta solución es muy costosa pues todos los discos son duplicados (G=1 y C=1), lo que resulta en una pérdida del 50% de la capacidad de almacenamiento del vector. Adicionalmente, cada escritura de un dato necesita también la escritura en el disco de verificación correspondiente.

Sin embargo, cuando el número de controladores es suficiente, la capacidad de tratamiento de operaciones de lectura por unidad de tiempo es mucho más importante que en el caso de un disco simple. El desempeño global de estos sistemas es entonces superior al de los discos independientes aun cuando las operaciones de escrituras sean más lentas.

A fin de explotar mejor esta organización, los archivos se entrelazan sector por sector en las N_G parejas de discos, y como en cada grupo sólo hay un disco de datos, las transferencias de grandes volúmenes de datos hace intervenir en paralelo varios grupos de discos.

---

[5] Mean Time To Repair

[6] Mirrored Disks

Nivel 2: código de Hamming

Los archivos son entrelazados bit por bit (un byte es distribuido en 8 discos), y se agregan discos de verificación para detectar y corregir los errores. Un único disco de verificación detecta la ocurrencia de un error durante una operación de E/S, pero para detectar el disco que generó el error y poder así corregirlo, se utiliza una técnica conocida como código de Hamming heredada de los circuitos de memoria. Para un grupo de 10 discos de datos (G=10) se necesitan 4 discos de verificación (C=4), para G=25, C=5 es indispensable.

Como la unidad de lectura/escritura de los discos es el sector, lo ideal para estos sistemas es la transferencia de al menos G sectores. Toda transferencia de tamaño inferior necesita de todas maneras la lectura de G sectores (uno por disco). El nivel 2 es entonces adecuado para sistemas donde las operaciones de E/S se pueden agrupar, pero en cambio no es aconsejable para los sistemas de bases de datos donde las unidades de transferencia son en general muy pequeñas.

Nivel 3: un sólo disco de verificación por grupo

La mayor parte de los discos de verificación del nivel 2 son utilizados para detectar el disco donde se produjo el error, un sólo disco es necesario para la detección del error propiamente dicho. Hoy podemos considerar que la mayoría de los controladores disponen de mecanismos hardware capaces de detectar el mal funcionamiento de un disco (por señales o por informaciones de verificación almacenadas en cada sector). En esta caso los discos de verificación utilizados para detectar los discos dañados son ellos mismos redundantes.

El sobrecosto de un sistema RAID de nivel 3 (C=1) es igual al número de grupos del vector (NG). El desempeño es el mismo que en el nivel 2, pero se mejora el rendimiento por disco. A este nivel el problema de sobrecostos puede considerarse como resuelto; queda por mejorar el desempeño del sistema para acceder a pequeños volúmenes de datos.

Nivel 4: lecturas/escrituras independientes

El nivel 4 intenta mejorar el tiempo de transferencia de pequeñas cantidades de datos gracias al paralelismo. Los datos ya no se entrelazan bit a bit, sino que se utiliza el bloque como unidad para entrelazar los archivos; de esta manera, varias operaciones de E/S pueden tener lugar al mismo momento dentro de un grupo.

Con esta organización una operación de escritura sólo concierne 2 discos: donde se va a escribir la información y el disco de verificación. Una modificación de tamaño inferior o igual a un sector no necesitará sino dos lecturas y dos escrituras, mientras que una lectura del mismo tamaño no necesita sino dos lecturas. Aunque la reducción del número de discos accedidos por operación es muy importante, hay que destacar que el disco de verificación participa en todas las operaciones, y por lo tanto puede volverse un cuello de botella para el sistema.

Nivel 5: integración de los datos y de la información redundante

Las modificaciones del nivel 4 no logran su objetivo de permitir la ejecución de varias operaciones de E/S en paralelo, pues el disco de verificación actúa como un serializador de las operaciones. Para eliminar este inconveniente, el nivel 5 distribuye la información redundante por todos los discos, incluso el disco de verificación.

El impacto de esta modificación sobre el desempeño del sistema es importante porque, ahora, varias lecturas/escrituras pueden ejecutarse en paralelo. De esta manera, el rendimiento por disco de este nivel se aproxima bastante del rendimiento por disco del primer nivel. Como los muestran los resultados de [PGK88], la integración de los datos y la información redundante ofrece buenos resultados para la transferencia de grandes y pequeños volúmenes de información. Cabe anotar sin embargo, que la presencia simultánea de solicitudes de ambos tipos limita fuertemente el desempeño global del sistema.

Comparación de los niveles RAID

Terminamos esta corta presentación de los 5 niveles RAID por una comparación de sus diferentes características. La figura 2 resume esta comparación, y para cada uno de los niveles muestra sus posibilidades según los tres parámetros que tienen la mayor influencia para la concepción de un sistema de E/S.

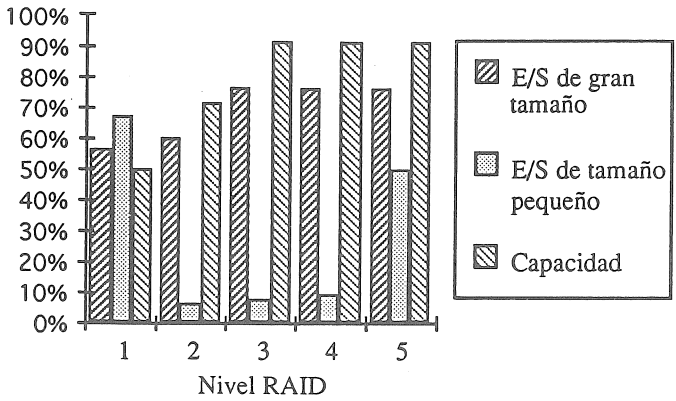

Figura 2 Comparación de los niveles RAID

Con respecto a los criterios de alto desempeño, extensibilidad y generalidad, aparece que sólo los niveles 1 (discos duplicados) y 5 (integración de datos y verificación) responden satisfactoriamente. En efecto, las limitaciones de las otras organizaciones para tratar las solicitudes de pequeñas cantidades de datos las excluyen para el uso general.

La tecnología de sistemas RAID ha seguido evolucionando, y en los últimos años, una nueva organización ha ido tomando fuerza dentro de los sistema que necesitan una altísima fiabilidad. Esta organización, conocida como el nivel 6 [BBBM94], propone una arquitectura de vectores de dos dimensiones con discos de paridad por fila y por columna. Esta arquitectura es evidentemente muy segura pero el costo generado para las escrituras (6 accesos) sólo se justifica por aplicaciones muy específicas.

Arquitectura de los sistemas paralelos de E/S

Los archivos entrelazados ofrecen alto rendimiento a los sistemas de E/S, pero la ganancia efectiva en velocidad de acceso es fuertemente dependiente de la configuración total del sistema, y para el caso de la computación de alto desempeño la limitación de extensibilidad juega un papel decisivo en la escogencia de esta configuración. La ventaja del paralelismo para estos sistemas de alto desempeño es que ofrece una excelente capacidad de extensión, sobretodo en lo que concierne los nodos de cálculo. Si, al menos en teoría, es posible aplicar el mismo principio con las unidades de almacenamiento, por duplicación de unidades (cada una con su vía de acceso independiente) proporcionalmente al número de procesadores, en la práctica problemas coma la distribución de los datos, lá saturación de las redes o la tolerancia a fallas, limitan esta posibilidad.

En lo que sigue proponemos una clasificación de las arquitecturas de E/S existentes, de acuerdo a la manera como están conectados los discos a los procesadores; identificamos así dos grupos arquitecturales: las máquinas que tienen una disposición fija de las unidades de almacenamiento, donde cada procesador dispone de una conexión directa a uno o varios discos; y las configuraciones más generales donde la comunicación entre las unidades de cálculo y las de almacenamiento se hace a través de una red de interconexión (cf. figura 3).

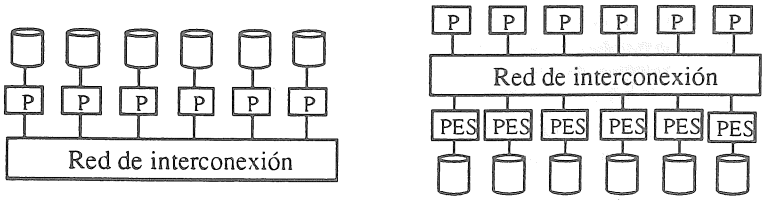

Figura 3 Sistemas paralelos de E/S

Arquitecturas de conexión local procesador-disco(s)

En estas arquitecturas, cada procesador dispone de una conexión directa a su propio disco local. Un mecanismo de DMA[7] controla, localmente al procesador, las transferencias entre la memoria y la unidad de disco, y el alto ancho de banda del sistema está asegurado por grandes bancos de memoria que juegan el papel de cachés de E/S. Estas arquitecturas tienen un propósito bien definido: minimizar el tiempo de transferencia de grandes volúmenes de datos entre los discos y las unidades de cálculo. Por esta razón, estas arquitecturas se han desarrollado principalmente en el ámbito de los supercomputadores de propósito específico.

La agrupación procesador-disco es una consecuencia directa de la pareja procesador-memoria. En efecto, estas arquitecturas son una extensión del modelo MIMD de memoria privada, donde los discos no son otras cosa que un nivel más en la jerarquía de memorias. Ahora cada nodo de la máquina es un computador completo con uno o varios procesadores, una memoria local y uno o varios discos.

Para estas arquitecturas es preferible que la mayor parte de las operaciones de E/S generadas por un procesador sean efectuadas en su disco local. Estos sistemas están dirigidos a las aplicaciones para las cuales se ha realizado de antemano un estudio serio de ubicación de datos en disco, y más crítico aún, donde se espera que los datos mantengan una fuerte estabilidad durante toda la ejecución de las aplicaciones. Este es el caso de aplicaciones como la visualización de imágenes donde cada parte de una imagen es tratada y almacenada por un procesador diferente.

La hipótesis de localidad es respetada solamente por un subconjunto bien reducido de aplicaciones, lo que hace que la utilización de estos sistemas sea bastante restringida. Adicionalmente, construir una máquina en la que cada procesador tiene su propia unidad de almacenamiento limita fuertemente el número de procesadores y por ende la extensibilidad de la máquina (tamaño de la máquina, costo de realización, etc.).

La máquina IBM Many/370 es un ejemplo llevado al extremo de estas arquitecturas [AGHL91]. Many/370 es una máquina paralela construida para ejecutar aplicaciones que hacen uso intensivo de E/S. Su prototipo está constituido de 8 procesadores, 128 discos y un computador frontal. Cada procesador dispone de un adaptador de E/S que le permite controlar 16 discos a partir de 4 buses SCSI[8].

Arquitecturas con red de interconexión para conectar los discos

De la misma forma que las arquitecturas de discos privados se inspiraron del modelo de memoria privada, las arquitecturas de interconexión para el acceso a los discos representan una extensión del modelo MIMD de memoria compartida. El principio de base de estas arquitecturas es que todo los procesadores deben tener un acceso equitativo a toda la información en memoria secundaria. Para alcanzar este objetivo, se utilizan procesadores especializados (procesadores de E/S o PES) que sirven de puentes entre las unidades de cálculo y los dispositivos de almacenamiento. Una unidad de cálculo ya no tiene entonces acceso directo a las unidades de disco, debe dirigir sus solicitudes a un servidor que la pondrá en la lista de espera del disco correspondiente.

Diversas máquinas comerciales como los hipercubos iPSC y nCUBE [AG94], y la CM-5 de Thinking Machines [Le92] utilizan este tipo de organización para sus sistemas de E/S. Nuevas organizaciones han sido propuestas con el objetivo de limitar las interferencias entre el tráfico debido a las comunicaciones entre procesos y el tráfico propio a las E/S. Así, se han ensayado arquitecturas con doble red de comunicación entre los procesadores (una específica para el tráfico de las E/S) pero su alto costo de realización ha dificultado su generalización [Ca95].

Una arquitectura universal para las E/S paralelas

La idea de independencia de redes de comunicación puede ser adaptada para obtener independencia para las comunicaciones entre los procesadores de E/S y de esta forma ganar en flexibilidad para el sistema. En un sistema de E/S de alto desempeño podemos identificar cuatro niveles posibles de paralelismo. Los tres primeros deben ser explotados por el software de E/S pero el cuarto debe ser garantizado por el material (cf. figura 4):

Paralelismo de las solicitudes generadas por las unidades de cómputo (UC): cada UC debe ser capaz de generar sus propias solicitudes de E/S sin tener que dirigirse a un servidor de la aplicación para su tratamiento.

---

[7] Direct Memory Acces

[8] Small Computer System Interface

Servidor paralelo (en los PES) de acceso de datos: el sistema de archivos debe ser una aplicación paralela que se ejecute en los PES, independientemente de cualquier otra actividad en la máquina. Con la existencia de varios procesos para su implementación, el sistema de archivos puede atender simultáneamente solicitudes que vengan de las UC y/o de los controladores.

Acceso en paralelo a los controladores de disco: los PES pueden generar solicitudes que necesiten la participación de varios controladores a la vez. Estas solicitudes serán tratadas en paralelo.

Acceso en paralelo a los dispositivos de almacenamiento: cada controlador puede administrar dispositivos que permitan la transferencia en paralelo de los datos (por medio de vectores de discos o discos multi-cabezas)

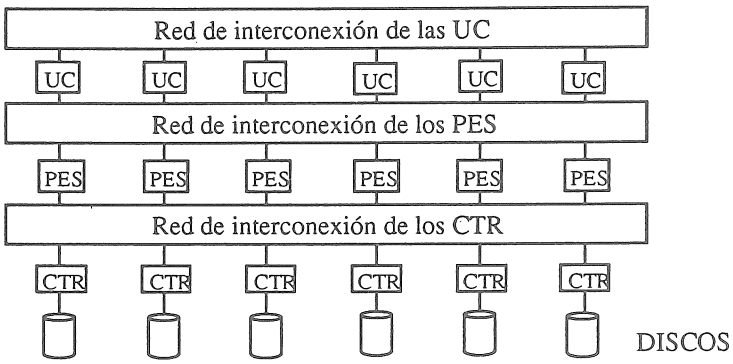

Figura 4 Arquitectura paralela universal de E/S

La figura 4 muestra la estructura general de la arquitectura resultante del análisis de estas características. Se identificaron tres redes de interconexión para aislar cada función del sistema de E/S. La descomposición vertical corresponde a la estructura funcional de las E/S, mientras que la expansión horizontal introduce el aspecto paralelo por la duplicación de las unidades de cómputo, de los procesadores de E/S y de los controladores de disco (CTR) conectados a uno o varios dispositivos de disco.

Conclusión

Las enormes diferencias en la evolución tecnológica entre las unidades de cómputo y los dispositivos de almacenamiento, han hecho de estos últimos una barrera para el desarrollo de las arquitecturas de alto desempeño. Todo parece indicar que esta diferencia va a continuar a aumentar por lo que nuevas organizaciones se hacen indispensables para atenuar esta diferencia.

Los sistemas RAID proponen cinco organizaciones para las E/S a alto desempeño que se apoyan directamente sobre el paralelismo de las operaciones. Nuestro análisis de estas organizaciones nos ha permitido destacar dos de ellas, las de primero y quinto nivel, pues a diferencia de las otras, estas organizaciones no favorecen un tipo de solicitud en desventaja de otros. En este artículo mostramos que para un sistema de propósito general, el nivel 5 ofrece el mejor compromiso con respecto a la capacidad de almacenamiento, sobrecosto del sistema, desempeño para solicitudes de pequeño y gran tamaño, extensibilidad del sistema, etc.

Finalmente propusimos una primera clasificación de los subsistemas de E/S en función de su estructura material para acceder a los datos. A partir de esta clasificación, se identificaron las características propias de los subsistemas de E/S de alto desempeño, lo que nos permitió introducir una arquitectura universal para estos subsistemas. Esta arquitectura maximiza la utilización del paralelismo para ofrecer el mejor desempeño global respetando las condiciones de extensibilidad y generalidad defendidas a lo largo del artículo.

Bibliografía

[Am67] G.M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities", *Proceedings of AFIPS Spring Joint Computer Conference*, Atlantic City, New Jersey, Vol 30, pp. 483-485, Abril 1967.

[AG94] G.S. Almasi, A. Gottlieb, *"Highly Parallel Computing"*, Second Edition, The Benjamin/Cummings Publishing Company Inc., 1994.

[AGHL91] B. Abali, B.D. Gavril, R..W. Hadsell, L. Lam, B. Shimamoto, "Many/370: a parallel computer prototype for I/O intensive applications", *Proceedings of the 6th Distributed Memory Computing Conference*, Portland, Oregon, pp. 728-730, Abril 1991.

[BBBM94] M. Blaum, J. Brady, J. Bruck, J. Menon, "EVENODD: an optimal scheme for tolerating double disk failures in RAID architectures", *Proceedings of the 21th International Symposium on Computer Architecture*, Chicago, Illinois, pp. 245-254, Abril 1994.

[Ca95] H. Castro, "Les Entrées/Sorties dans les architectures massivement parallèles", Tesis de doctorado, Instituto Nacional Politécnico de Grenoble (INPG), Noviembre 1995.

[HNS94] M. Henderson, B. Nickless, R. Stevens, "A scalable high-performance I/O system", *Proceedings of the 1994 Scalable High Performance Computing Conference SHPCC'94*, Knoxville, TN, Mayo 1994.

[Ki86] M.Y. Kim, "Synchronized disk interleaving", *IEEE Transactions on Computers*, Vol C-35, No 11, Noviembre 1986.

[KGP89] R.H. Katz, G.A. Gibson, D.A. Patterson, *"Disk system architecture for high performance computing"*, Technical Report, University of California, Berkeley, csd-89-497, 1989.

[Le92] Charles E. Leiserson et al, "The network architecture of the Connection Machine CM-5", *4th Symposium on Parallel Algorithms and Architectures*, pp. 272-285, Junio 1992.

[LKB87] M. Livny, S. Khoshafian, H. Boral, "Multi-disk management algorithms", *Proceedings ACM SIGMETRICS*, pp 69-77, Mayo 1987.

[NB89] A.L. Narasimha Reddy, P. Banerjee, "An evaluation of multiple-disk I/O systems", *IEEE Transactions on Computers*, Vol 38, No. 12, pp. 1680-1690, Deciembre 1989.

[NNSI90] U. Nagashima, F. Nishimoto, T. Shibata, H. Itoh, M. Gottoh, "An improvement of I/O function for auxialiary storage: parallel I/O for a large scale supercomputing", *Proceedings of the International Conference on Supercomputing*, acm press SIGARCH, Amsterdam, pp. 48-59, Junio 1990.

[PGK88] D.A Patterson, G. Gibson, R.H. Katz, "A case for redundant arrays of inexpensive disks (RAID)", *Proceedings of the International Conference on Management of Data*, SIGMOD, Chicago, Illinois, Junio 1988.

[SG86] K. Salem, H. Garcia-Molina, "Disk Striping", *IEEE 1986 International Conference on Data Engineering*, 1986.