Partial Match Retrieval in Two-headed Disks

Yannis Manolopoulos and Athena Vakali

Department of Informatics, Aristotle University, Thessaloniki, Greece, 54006.
email: manolopo,vakali@eng.auth.gr

Abstract. The performance of a disk with two heads per surface separated by a fixed number of cylinders is examined. We derive the probability distribution of arm stops, the expected number of stops as well as the expected number of cylinder clusters, i.e. the number of sets of consecutive compound cylinders. In comparison with a single-headed disk, it is shown that the performance gain may reach 50% on the average.

1 Introduction

According to the problem of partial match retrieval, the aim is to group the file records into pages to facilitate partial match queries in such a way that queries are answered by accessing the minimal number of pages on the average [15]. On the other hand, according to the problem of multidisk files, the purpose is to allocate the file pages to specific disks in order to exploit the fact that the disks may be accessed concurrently [6]. Since the multidisk file problem has been proven to be NP-hard, the effort is directed towards proposing an efficient heuristic. The known heuristics can be summarized in three categories:

- the Random Data Allocation and the Partition Data Allocation methods,
- the Modulo Allocation methods, such as the Disk Modulo [6], the Generalized Disk Modulo and the Binary Disk Modulo Allocation methods, and finally,
- two methods based on the Minimal Spanning Trees and the Shortest Spanning Paths algorithms.

For a short comparative survey on these methods see [8], and for a more recent collection of relevant citations see [10].

These traditional problems have been broadly examined over the last two decades. A more general term that has appeared in the literature more recently is declustering. Declustering is a technique aiming at improving the I/O performance which is the bottleneck in database processing environments, a fact that is more important nowadays in the light of very fast processors. This purpose is achieved by distributing a file over many disk units in order to maximize the parallelism and minimize the response time for partial match as well as for range queries. This technique may be useful in the context of database machines [4, 5], multiprocessor systems [16], or in multiple disks [3].

Let us clarify some terminology. A file is a collection of records which are comprised of a set of attributes denoted by A_i, where $0 \leq i \leq n$ (n is the number of attributes). Partial match queries are queries of the form $q : (A_1 =$

$a_1, A_2 = a_2, ..., A_n = a_n$), where a_i is either a value from the domain of the i-th attribute, or is unspecified (denoted by *). For example, multi-attribute hashing is a method which maps attribute values to bit sequences by the use of a hash function, and concatenates these bit sequences to form a binary vector for each file record. Then the structure is used to satisfy partial match queries by concatenating attribute hash values of the query to identify the relevant binary vectors.

The number of unspecified attributes in a partial match query is x, where $0 \leq x \leq n$. Given a query q, the query response set $R(q)$ is the set of pages that qualify for the query q. The response time on a query q is $\max\{N_1, N_2, ..., N_m\}$, where m is the number of disks and N_i is the number of qualifying pages on the i-th disk (where $1 \leq i \leq m$). Strict optimal declustering scheme is a scheme which answers a query by accessing $\lceil |(R(q)|/m \rceil$ pages at maximum.

disk	Stored records
1	(a,a) (a,b) (a,c) (a,d)
2	(b,a) (b,b) (b,c) (b,d)
3	(c,a) (c,b) (c,c) (c,d)
4	(d,a) (d,b) (d,c) (d,d)

disk	Stored records
1	(a,a) (b,b) (c,c) (d,d)
2	(a,b) (b,c) (c,d) (d,a)
3	(a,c) (b,d) (c,a) (d,b)
4	(a,d) (b,a) (c,b) (d,c)

Example.
Suppose that $n = 2$ attributes, i.e. A_1 and A_2, whereas the domains of each attribute are $\{a, b, c, d\}$. Suppose, also, that we have $m = 4$ conventional disks. If the records were stored in the four disks according to the left part of the above scheme, then the response time for a query of the form $(y, *)$ or $(*, y)$ (where $y \in \{a, b, c, d\}$) would be $\max\{N_1, .., N_4\} = 4$. However, if the records were stored as depicted at the right part of the same scheme, then the response time for a query of the form $(y, *)$ or $(*, y)$ would be $\max\{N_1, .., N_4\} = 1$. We understand that the second allocation is optimal. □

In the present work we assume that a declustering technique spreads a file over a number of two-headed disks. For our study, the specific declustering technique is transparent, that is, any declustering technique could be adopted. In other words, we are concerned with intraparallelism (parallelism on a disk per se) but not with interparallelism (parallelism between specific disks). We examine the time performance of partial match queries and calculate the gain in answering such a query in a two-headed disk with heads separated by a fixed number of cylinders over answering it using a single-headed one. More specifically, we provide an analysis for the probability distribution of the arm stops for answering a partial match query. Also, we calculate the expected number of arm stops as well as the expected number of cylinder clusters, i.e. the number of sets of consecutive compound cylinders. For both measures (either stops or clusters), it is shown that the performance gain, in comparison with a single-headed disks, may reach 50% on the average.

2 Analysis for the number of stops

The aforementioned problems have been examined in the context of conventional disks systems with one head per surface. These systems have been studied extensively. The two most crucial factors which affect the performance of magnetic disk storage devices during input/output operations are seeking and latency [17]. More specifically, seeking is a heavier cost than latency and depends on the workload, which at a given point in time is issued by the operating system or by the database management system. In particular, heavier workload results in a larger number of cylinders being visited; consequently, a larger distance has to be traveled on the disk, which is equal to the number of cylinders lying between the first and the last hit cylinder involved in the request.

Assume that at a certain point in time, a number of N requests arrive. If these requests are serviced by a SCAN based scheduling policy, then the following linear equation approximates the expected seek time, T, as a function of these two cost metrics:

$$T = S \times T_{min} + D \times (T_{max} - T_{min}) \; / \; (C-1) \tag{1}$$

In this equation C is the number of cylinders, T_{min} (respectively, T_{max}) is the seek time when the disk heads are moving a distance of 1 (respectively, $C-1$) cylinder(s), S (where $S < N$) is the actual number of cylinder hits and D is the distance traveled by the read/write heads. It is evident that the first product expresses the time due to the inertia of the moving mechanism, while the second product expresses the actual time for traveling. Among other works concerning conventional disks, we note that references [13] and [11] derive formulae for the S and D quantities respectively.

However, disks with two heads per surface have been introduced and studied over the last few years. The two heads may move independently of each other but for the moment only systems with two heads per surface separated by a fixed number of cylinders do exist as commercial products [14]. In the sequel, we focus on systems having two heads per surface separated by a fixed number of cylinders. For such systems, it has been proved that the optimum separation distance is $0.44657 \times C$ if the FCFS scheduling policy is adopted [1], but if the SCAN scheduling algorithm is applied then this distance should be equal to $\lfloor C/2 \rfloor - 1$ or $\lceil C/2 \rceil - 1$ [12]. It is noted that the re-examination of analytical issues in two-headed disk systems is not trivial due to the additional physical constraints. Otherwise stated, it is certain that such a system with C cylinders is not equivalent performance-wise to a conventional disk having $C/2$ cylinders with double capacity.

The performance of two-headed disks is superior than that of conventional ones because both products of Equation (1) are reduced. The reduction of the distance traveled by the disk heads has been analyzed in the literature [12]. The reduction in terms of the number of arm stops has not yet been evaluated. This reduction is a consequence of the fact that when the moving mechanism stops to service a request from a given cylinder, it may then service a request from the cylinder which lies under the other disk head without moving at all. We say that any two cylinders which may be visited in this way (i.e. one after the other at no extra seeking) form a compound cylinder. It is evident that in a conventional disk with one head per surface the number of arm stops equals the number of cylinder hits. However, this observation does not hold in the case of the modern two-headed systems because once the arm stops, it may service at no extra seek movement both cylinders of a compound cylinder, which lie under the two heads at the same time.

In the present study we examine how the response time can be reduced in a two-headed disk, regardless of the specific allocation method or whether it is a strictly optimal or not. We adopt the following work assumptions:

- The disk has C cylinders, where C an even number without loss of generality, therefore the disk has $C/2$ compound cylinders,
- The fixed head separation distance between them is $C/2$ cylinders, and
- The request involves N distinct cylinders, out of the C ones,

and proceed to the following theorem.

Theorem.

The probability distribution of the number of arm stops of a two-headed disk is:

$$P(N-i) = 2^{N-2i} \frac{\binom{C/2-i}{N-2i}\binom{C/2}{i}}{\binom{C}{N}} \tag{2}$$

where $N \leq C/2$ and $0 \leq i \leq \lfloor N/2 \rfloor$, or

$$P(N-i) = 2^{C-N-2i} \frac{\binom{C/2-i}{N-C/2+i}\binom{C/2}{i}}{\binom{C}{N}} \tag{3}$$

where $N > C/2$ and $N - C/2 \leq i \leq \lfloor N/2 \rfloor$.

Proof of the first part.

Let a request involving N distinct cylinders, where $N \leq C/2$. Under a certain probability the number of arm stops may be N, or even less, i.e. $N-1, N-2, \ldots, \lceil N/2 \rceil$. First, the calculation of the number of instances that exactly N arm stops will take place follows. This case may be interpreted by viewing that the N requests are related with N compound cylinders with an one-to-one correspondence. Suppose that all the required cylinders fall in the first half of the disk and may be serviced only by the left head; therefore, the number of stops is N. The first disk half consists of $C/2$ cylinders, therefore the number of ways that these N cylinders may be selected out of the $C/2$ ones is $\binom{C/2}{N}$. However, there is a chance that a cylinder of the second disk half is hit but at the same time the symmetric cylinder belonging to the same compound cylinder lying at a distance of $C/2$ cylinders is not hit. In this instance, again, N arm stops will take place. Such an occurrence exists for each of the $\binom{C/2}{N}$ ones, therefore this number of ways should be multiplied by 2^N.

The case, that $N-1$ arm stops will take place, is treated as follows. The fact that $N-1$ arm stops will take place is explained by accepting that each of the $N-2$ compound cylinders receives one request, while one compound cylinder receives two requests. By using the previous reasoning, it is derived that the number of ways that the one compound cylinder, which receives two requests, may be selected out of the $C/2$ cylinders is $\binom{C/2}{1}$. The number of ways that the $N-2$ compound cylinders, each one receiving one request, may be selected out of the $(C/2-1)$ ones is $\binom{C/2-1}{N-2}$. In an analogous manner, to derive the total number of ways that this fact may occur we must multiply the previous quantities by 2^{N-2}.

The first part of the proposition is derived by generalizing on the concept of compound cylinders receiving one or two requests and considering that the total number of ways that N cylinders may be selected out of the C ones is $\binom{C}{N}$. □

Proof of the second part.

Let, now, a request involving N distinct cylinders be given, where $N > C/2$. In the worst case and under some probability the number of arm stops may be $C/2$, or may be less, i.e. $C/2-1, C/2-2, \ldots, \lceil N/2 \rceil$.

First, the calculation of the number of instances that exactly $C/2$ arm stops will take place follows. This case may be interpreted by viewing that the N requests concern all the $C/2$ cylinders of the first half of the disk and the remaining $N-C/2$ requests fall in the second half of the disk. The number of ways that these $N-C/2$ requests may be selected out of the $C/2$ cylinders of the second half is $\binom{C/2}{N-C/2}$. However, there are chances that some cylinders of the first disk half will not be hit, therefore this number of ways should be multiplied by 2^{C-N}.

The case, that $C/2-1$ arm stops will take place, is treated as follows. Let us suppose that $C/2-1$ cylinders of the first half of the disk are hit, while the remaining $N-C/2+1$ required cylinders belong to the second half of the disk. These $N-C/2+1$ cylinders may be selected in $\binom{C/2-1}{N-C/2+1}$ ways. In addition, by using the reasoning of the first part of the proof, the total number of ways is derived by multiplying the previous quantity by $2^{C-N-2} \times \binom{C/2}{1}$. The second part of the proposition is derived by generalizing on the concept of compound cylinders which may receive one or two requests and considering that the total number of ways that the N cylinders may be selected out of the C ones in $\binom{C}{N}$. □

Corollary.
The expected number of arm stops of a two-headed disk is:

$$\mathrm{E}[S] = \sum_{N\leq C/2}^{0\leq i\leq \lfloor N/2\rfloor} (N-i)\times P(N-i) + \sum_{N>C/2}^{N-C/2\leq i\leq \lfloor N/2\rfloor} (N-i)\times P(N-i) \quad (4)$$

3 Analysis for the number of clusters

In this section we carry out an analysis deriving the expected value of clusters. As a *cluster* we define the set of consecutive compound cylinders which are hit when resolving a partial match query. In such a way the number of clusters is actually the number of random disk accesses to resolve the query. The notion of clusters is important when examining the performance of a disk, since cylinder clustering affects the value of distance D traveled by the disk heads, as shown in relation (1).

Suppose that a partial match query consisting of n bits, one bit per attribute, is posed against a multiattribute hashed file residing in a two-headed disk with $C = 2^n$ cylinders. Assume, also, that x (out of n) bits are unspecified. Therefore, in order to identify the compound cylinder where a certain record/request resides, we use the formula "cmod2^{n-1}", where c is the cylinder number.

Theorem.
The expected number of clusters satisfying a partial match query is:

$$\mathrm{E}[clusters] = \frac{1}{\binom{n-1}{n-x}} \sum_{i=1}^{x} \binom{n-i-1}{n-x-1} 2^{x-i} \quad (5)$$

Proof [7].
Let us number the bits of a partial match query vector from right to left. If the n-th (=leftmost) bit of the partial match query vector is unspecified, i.e. it is a "don't care" bit (*), then the $n-x$ specified bits will lie among the first (=rightmost) $n-1$ bits. In case that i is the rightmost specified bit position, then all of the $n-x$ specified bits will lie in positions from i to $n-1$. Also, a set of $x-i$ unspecified characters will lie in this range. Since the $n-x$ specified bits lie up to the $(n-1)$-th position, i will range between 1 and x positions. The partial match query will be resolved by 2^{x-i} clusters and since one of the specified bits lies in the i-th bit position, the rest $n-x-1$ of the specified characters range in the $n-1-i$ positions left. So, there are $\binom{n-1-i}{n-x-1}$ ways to choose $n-x-1$ characters in that range. Then, the total number of clusters for all possible partial match queries with $(n-x)$ specified bits is $2^{n-x}\sum_{i=1}^{x}\binom{n-1-i}{n-x-1}2^{x-i}$, where 2^{n-x} is the number of different bit vectors for each set of $n-x$ bits. If all possible partial match queries with $(n-x)$ specified bits are considered to be equiprobable, then the expected number of clusters is given by relation (5). □

4 Results and Discussion

Figure 1 depicts the probability distribution function for the number of arm stops for some values of the parameter N. For all curves it is accepted (for computational reasons) that the disk consists of C=100 cylinders (only). It is observed that this probability distribution seems rather skew. For example, if the number of the requested distinct cylinders equals 50 (N=50), then theoretically the curve should range from 25 to 50. However, practically non infinitesimal values of the curve exist from 33 to 42 arm stops.

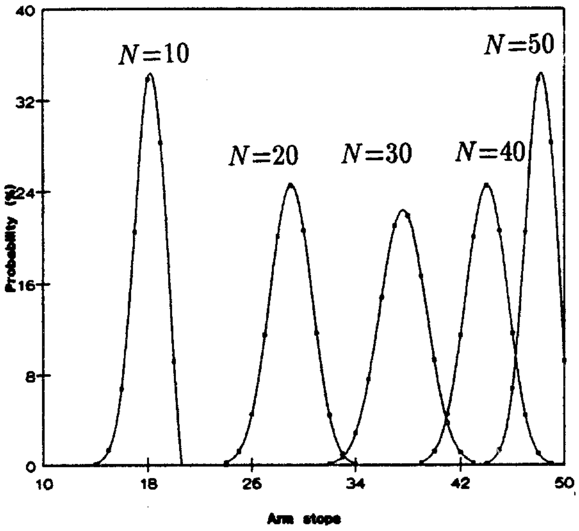

Fig. 1. Probability distributions of the number of arm stops as a function of the number of the requested cylinders (N).

A second observation is that the distribution functions for values of the parameter N, which are equidistanced from the central disk cylinder, are identical in shape. In other words, the distribution function for a request of N_1 distinct cylinders, where $N_1 < C$, is shifted $C/2 - N_1$ positions to give the distribution function of a request of N_2 distinct cylinders, where $N_2 = C - N_1$. This may be verified by reconsidering the two parts of the probability distribution at Section 2. More specifically, each term of one part is equal to the corresponding term of the other part. Therefore, the following Corollary holds.

Corollary.

If $N_1 = C/2 - a$ and $N_2 = C/2 + a$, where a is an integer number smaller than $C/2$, then $P(N_1 - i) = P(N_2 - a - i)$ for $0 \leq i \leq N_1/2$. □

In Figure 2 the expected number of arm stops is depicted as a function of the number of the requested cylinders, while Figure 3 illustrates the expected gain in arm stops in comparison with single-headed disks as a function of the same parameter. For these figures, too, it is assumed that the disk consists of 100 cylinders. The endpoints of both curves are obvious; more specifically in Figure

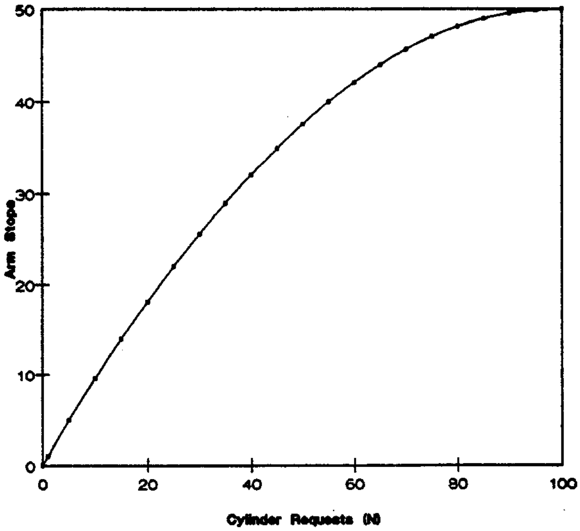

Fig. 2. Expected number of arm stops as a function of the number of the cylinder requests (N).

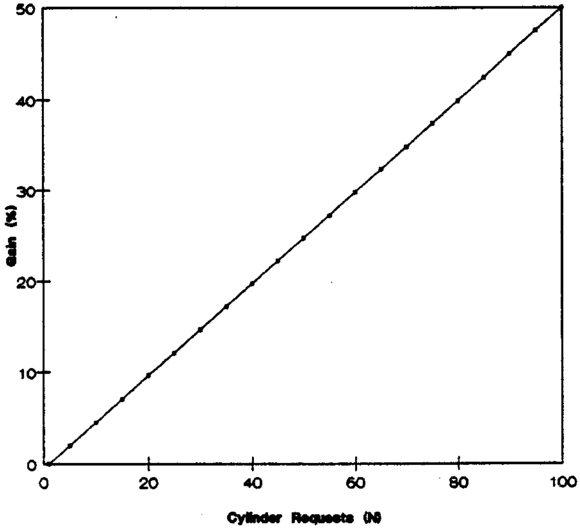

Fig. 3. Percentage of the expected gain as a function of the number of cylinder requests.

2 for N=1 and N=100 the expected number of stops is 1 and 50 respectively, while in Figure 3 for the same values of N the expected gain is 0% and 50% respectively. It is evident, that in practice a disk workload of C pending requests is not realistic and, therefore, the value of 50% is only of theoretical interest. From the last figure, it is remarked that the expected gain shows a linear nature on the number of cylinder requests. This is easily explained by the following reasoning. Suppose that N out of C cylinders are requested; according to some head scheduling policy the arm moves and finally stops on top of a certain cylinder.

The probability that the buddy cylinder of the present compound cylinder must be visited is $(N-1)/(C-1)$, since all cylinders are hit equiprobably. Therefore, greater is the number N of requests, greater is the probability, and consequently the gain, that another cylinder will be visited with no move of the disk arm. The tangent of the linear curve of Figure 3 is $50/(C\text{-}1)$, therefore the expected gain for any number of cylinder requests, N, is $50 \times (N-1)/(C-1)\%$. It is noted, also, that this observation is valid for whatever is the scheduling policy applied.

Table 1 reports some results on the performance of one- and two-headed disks with $C = 2^n = 1024$ cylinders, i.e. $n = 10$. Obviously, the two-headed disk has 512 compound cylinders. The table gives the expected number of clusters for partial match queries with a varying number of unspecified bits, given that the queries are posed against a multiattribute hashed file. We see that the expected gain over the conventional disk is more than 47%. More specifically, if the n-th attribute of the query is unspecified, the two-headed disk behaves almost twice as good as the one-headed disk, whereas in case that the n-th attribute is specified both one- and two-headed disks behave similarly. This is due to the fact that if the n-th bit is specified, then the derived compound cylinder has already been accessed from all the possible combinations of the first $n-1$ bits.

x	1-headed disk	2-headed disk	gain (%)
1	1.9	1	47.368
3	6.65833	3.52778	47.017
5	21.8373	11.6746	46.538
7	61.8583	33.5119	45.825
9	102.3	56.77778	44.499

Table 1. Expected number of clusters for one-headed and two-headed disks.

In conclusion, in the present report we examine magnetic disks with two heads per surface separated by a fixed number of cylinders. These systems exist commercially and it has been proved that they perform better than conventional disks in terms of the distance traveled by the arm mechanism. For example, in [12] it has been shown that the gain in the distance traveled is near 50% if the heads are optimally spaced. This work's contribution is the derivation of the probability distribution for the number of stops of the moving arm. Moreover, we calculate the expected number of arm stops as well as the expected number of cylinder clusters. This information is of use for the design of a declustering/allocation method which will exploit the specific hardware characteristics. We anticipate that examining error correcting codes [7, 8], or techniques based on access patterns [9], or even techniques which have been designed specifically for two-disk sets [2], could result in promising schemes specifically designed for multiple two-headed disks.

In addition, it is worth noting that in some recent disks with voice-coil actuators, the seek time does depend linearly on the seek distance traveled, D, but it is a function of the square root of this quantity. Thus, the impact of latency time on the response time may dominate (at least, when compared to the seek time it may become of a similar time cost) and, therefore, latency has to be re-examined in the context of two-headed disks. In addition, declustering, locality and placement issues have to be examined afresh.

References

1. Calderbank, A.R., Coffman, E.G., Flatto, L.: Optimum Head Separation in a Disk System with Two Read/Write Heads. Journal of the ACM **31** (4) (1984) 826-838
2. Chang, C.C., Chen, C.Y.: Performance of Two-Disk Partition Data Allocations, BIT **27** (1987) 306-314
3. Chen, P.M., Lee, E.K., Gibson, G.A., Katz, R.H., Patterson, D.A.: RAID - High-performance, Reliable Secondary Storage. ACM Computing Surveys **26** (2) (1994) 145-185
4. Copeland, G., Alexander, W., Boughter, E., Keller, T.: Data Placement in Bubba. Proceedings, ACM SIGMOD Conference, Chicago, Illinois (1988) 99-109
5. DeWitt, D., Gerber, R.H., Graefe, G., Heytens, M.L., Kumar, K.B., Muralikrishna, M.: GAMMA - a High Performance Database Machine. Proceedings, 12th VLDB Conference, Kyoto, Japan (1986) 228-237
6. Du, H.C., Sobolewski, J.S.: Disk allocation for Cartesian Product Files on Multiple-disk Systems. ACM Transactions on Database Systems **7** (1) (1985) 82-101
7. Faloutsos, C.: Multiattribute Hashing using Gray Codes. Proceedings, ACM SIGMOD Conference, Washington D.C., (1986) 228-238
8. Faloutsos, C., Metaxas, D.: Disk Allocation Methods using Error Correcting Codes. IEEE Transactions on Computers **40** (8) (1991) 907-914
9. Kim, J.U., Chang, H., Kim, T.G.: Multidisk Partial Match File Design with Known Access Pattern. Information Processing Letters **45** (1993) 33-39
10. Li, J., Srivastava, J., Rotem, D.: CMD - a Multidimensional Declustering Method for Parallel Database Systems. Proceedings, 18th VLDB Conference, Vancouver, Canada (1992) 3-14
11. Manolopoulos, Y., Kollias, J.G.: Estimating Disk Head Movements in Batched Searching. BIT 28 (1988) 27-36
12. Manolopoulos, Y., Kollias, J.G.: Performance of a Two-headed Disk System when Serving Database Queries under the SCAN Policy. ACM Transactions on Database Systems **14** (3) (1989) 425-442
13. Manolopoulos, Y.: Probability Distributions for Seek Time Evaluation. Information Sciences **60** (1-2) (1992) 29-40
14. Page, I.P., Wood, R.T.: Empirical Analysis of a Moving Headed Disk Model with Two Heads Separated by a Fixed Number of Tracks. The Computer Journal **24** (4) (1981) 339-342
15. Rivest, R.L.: Partial Match Retrieval Algorithms. SIAM Journal on Computing **5** (1) (1976) 19-50
16. Stanfill, C., Kahle, B.: Parallel Free Text Search on the Connection Machine System, Communications of the ACM **29** (12) (1986) 1229-1239
17. Wong, C.K.: Algorithmic Studies in Mass Storage Systems, Computer Science Press, (1983)