A Study of Dynamic Reconnection

Alexandre Brandwajn

Amdahl Corporation

1250 East Arques Avenue

Sunnyvale, California 94086

ABSTRACT

The recently introduced Extended Architecture (XA) of large IBM computer systems includes, in the disk I/O area, the ability for an access to be resumed and completed on a path different from the one on which it has been initiated. The expected disk performance improvement due to this feature - known as Dynamic Path Reconnection - is investigated in this note.

Two popular double pathing connection schemes are considered: switched substrings (as in IBM 3380 Dynamic Path Selection) and dual ported devices (as in Amdahl 6280 Dynamic Performance Pathing). A simple classical queueing system is used to model missed revolutions when transfer paths are found busy. The numerical results obtained indicate that, depending on load, a substantial reduction in the average I/O time can be expected with the Dynamic Path Reconnection feature. This reduction can top 30% under moderately heavy loads. The accuracy of the analytical model for missed reconnections has been checked using discrete-event simulation.

1.0 INTRODUCTION

Direct Access Storage Devices (DASDs), i.e., disks, are an important part of large commercial computer installations. Their importance stems both from their potential impact on system performance, especially in transaction oriented applications, and their total cost within an installation which is predicted to top that of Central Processors in a very near future [1].

Typically, several disks share a transfer path (consisting of a channel, a control unit and a string controller) to the CPU. In recent years, there has been a clear move to improve DASD performance by multiplying access paths to devices. This increases the degree of concurrency of operation within the DASD farm. Two popular connection schemes with two paths, illustrated in Figures 1 and 2, are dual ported devices (e.g., Amdahl 6280 Dynamic Performance Pathing) and switched substrings (e.g., IBM 3380 Dynamic Path Selection).

A disk I/O operation consists of several phases only some of which require the availability of the transfer path. Therefore, devices with Rotational Position Sensing disconnect from the transfer path when it is not needed so that it can be used by some other disk, and reconnect to it when they are ready to transfer. If the path happens to be busy when the device attempts to reconnect, a missed reconnection occurs, and the device will try another time after a full revolution when it reaches again the correct angular position [2].

In the classical IBM I/O architecture, an I/O operation must proceed and complete on the transfer path on which it has been initiated. With multiple access paths, this introduces some inefficiency in terms of missed device reconnections, since it is possible for a missed reconnection to occur even when a tranfer path is available if this path is not the one on which the disk access has started. Being able to complete an I/O operation on any path available requires a specific design of the control unit and of the operating system software. Such a design is part of the recently announced IBM Extended Architecture (XA) and is known as the Dynamic Path Reconnection feature [3].

The goal of this note is to assess the impact of Dynamic Path Reconnection on average I/O time in the case of dual ported devices and of switched substrings (IBM Dynamic Path Selection scheme). Previous work on multiple access paths to disks includes papers by Bard [4,5] and Brandwajn [6]. [4] and [6] both assume classical channels, and [5] is concerned mainly with the application of the maximum entropy principle to the evaluation of missed reconnections with multiple paths rather than their impact on the performance of specific pathing schemes.

The method employed here to evaluate missed reconnections is different. A simple loss-system model [7], classical in telephone applications of queueing theory, is used. Numerical results produced by this method have been found to compare favorably with those of discrete-event simulations.

In the next section, we briefly describe the approach used to evaluate average I/O times. Section 3 is devoted to numerical results illustrating the expected improvement in performance provided by Dynamic Path Reconnection.

2.0 EVALUATION OF AVERAGE I/O TIME

The average I/O time may be viewed as consisting of two parts: queueing until the I/O operation is initiated, and actual I/O service time (see Figure 3). The queueing time comprises a wait for the requested device to become free and a wait for a transfer path to the device to initiate the operation. The actual I/O service comprises a seek (arm positioning), latency (until the device reaches in its rotation the appropriate angular position), a delay due to missed reconnections and the data transfer (search, if any, is treated here the same as data transfer). Additionally, depending on the type of DASD, some path wait may be involved if the path is found busy following the completion of the seek portion. In most cases, the path wait is considerably smaller than other elements of the average I/O time. We neglect it here.

Let

W : average I/O time

Q : average queueing time

B : average I/O service time (no queueing)

S : average seek time

L : average latency time

M : average missed reconnection delay

T : average transfer time

R : device rotation period

We have

$$W = Q + B , \tag{1}$$

where

$$B = S + L + M + T. \tag{2}$$

We assume that the average seek time does not vary considerably with load so that the missed reconnection delay is the main load dependent element of the I/O service time. For simplicity, we also assume that the I/O load is balanced, i.e., all disks are equally utilized and so are the elements of the transfer paths (channels, control units).

We now evaluate the missed reconnection delay. Consider first a set of dual ported disks connected to two control units each control unit having one channel hooked to it as shown in Figure 1. Denote by d the number of devices (actuators) in this system. Assume that reconnection requests on a first attempt (just after latency) constitute a Poisson process. Let p(j) be the stationary probability that j paths are busy (j=0,1,2). Let also a(j), j=0,1, be the probability that a reconnection attempt will succeed given that it finds j paths busy.

Without Dynamic Path Reconnection, a specific path must be found available so that $a(j)=1-j/2$. With Dynamic Reconnection, any free path will do, and $a(j)=1$, $j=0,1$.

We assume that the utilization of a channel (or, equivalently, its throughput) is known, and we denote it by U. Treating first attempt reconnection requests as "calls" subject to loss, we readily obtain (cf. [7])

$$p(j) = \frac{1}{G} x^j \prod_{i=1}^{j} (d-i+1)a(i-1)/i, \quad j=0,1,2; \quad (3)$$

where G is a normalization constant. This equation involves an unknown parameter x which may be determined from the condition that the total path utilization must be equal to 2U

$$\sum_{j=1}^{2} jp(j) = 2U . \quad (4)$$

Let p_1 be the probability of a missed reconnection on a first attempt. We readily have, for $d>1$,

$$p_1 = 1 - \sum_{j=0}^{1} (d-j)p(j)a(j) / \sum_{j=0}^{2} (d-j)p(j). \quad (5)$$

Note that $(d-j)p(j)$ is proportional to the rate of reconnection requests arriving to find j paths busy ($j=0,1,2$), and $(d-j)p(j)a(j)$ is proportional to the rate of successful reconnection requests with j paths busy ($j=0,1$). Thus the fraction in (5) represents the probability of a successful reconnection.

The evaluation of the missed reconnection probability for the Dynamic Path Selection arrangement (Figure 2) is quite similar in principle. The main difference is that here there are two reconnection points at which a miss can occur: the substring to which a given actuator belongs (string controller) and the control unit. Let s be the number of actuators in a substring, and h the number of such substrings connected to the two paths of the control unit (in our case $h=4$). For a first reconnection attempt, the probability that the substring is found free is denoted by p_s, and is readily seen to be

$$p_s = (1-U_s)/(1-U_s/s) , \quad (6)$$

where U_s is the utilization of a substring path, and is assumed known. The reconnection of substrings to the control unit is modelled again as a loss system. Denote by $p(j)$ the stationary probability that j paths ($j=0,1,2$) are busy at the control unit. Let as previously $a(j)$, $j=0,1$, be the probability that a reconnection attempt from a substring to the control unit is successful given that j paths are busy at the control unit. In absence of Dynamic Path Reconnection, $a(j)=1-j/2$, since a specific path must be free. With Dynamic Reconnection, $a(j)=1$ ($j=0,1$). Hence,

$$p(j) = \frac{1}{H} y^j \prod_{i=1}^{j} (h-i+1)a(i+1)/i, \quad j=0,1,2, \quad (7)$$

where H is a normalization constant, and the unknown parameter y can be easily determined from the condition

$$\sum_{j=1}^{2} jp(j) = 2U. \quad (8)$$

U is the channel utilization.

The probability of a successful first attempt reconnection of a substring to the control unit, denoted by p_u, is thus

$$p_u = \sum_{j=0}^{1} (h-j)p(j)a(j) / \sum_{j=0}^{2} (h-j)p(j) , \quad (9)$$

and the miss probability on a first attempt is simply

$$p_1 = 1 - p_s p_u . \quad (10)$$

It is generally recognized (see, e.g., [4]) that subsequent reconnection attempts experience a higher miss probability than the first attempt. From the several corrections proposed in the literature to account for this phenomenon, we choose the one that yields for the average number of missed reconnections per transfer, denoted by m,

$$m = p_1/[(1-p_1)(1-q)] , \quad (11)$$

where $q=\exp(-R/T)$ (cf [9]).

The average missed reconnection delay is simply $M=mR$, and we now know all the elements of the average I/O service time. In order to

estimate the amount of queueing incurred prior to service, we treat each actuator as a single M/G/1 queue. Additionally, we assume that the various phases of device service time, viz., seek, latency etc, are mutually independent. Knowing the variance of the seek and service times, this allows us to estimate the coefficient of variation of the I/O service time which we denote by c. Using the Pollaczek-Khinchin formula [8], we obtain for the average total I/O time W

$$W = B \{ 1 + U (1+c^2)/[2(1-U)] \} , \qquad (12)$$

where U is the device utilization given by the product of B, the average device service time, by the successful I/O rate sustained by the device.

In the next section, we use these results to examine numerically the expected improvement in average I/O time provided by the Dynamic Path Reconnection feature.

3.0 NUMERICAL RESULTS

Our first set of numerical results is devoted to the impact of Dynamic Path Reconnection on the performance of dual ported devices. As an example, we choose device parameters corresponding to the Amdahl 6280 disk with transfer rate of 1.86 Mbyte/s. We assume that only 60% of disk accesses experience the manufacturer specified average seek time. We also assume that a transfer is preceded by a search of .7 ms, and that the protocol overhead adds an average of 1 ms to the path busy time. We have S=10.8 ms (average seek time), R=15.15 ms (revolution time) and L=R/2 (latency).

In Figure 4, we have plotted the average I/O time versus the I/O throughput sustained by a single channel, with and without Dynamic Reconnection, for a system with 16 actuators and average transfer block length of 4,096 bytes (T=2.9 ms). Similar results for a system with 32 actuators are shown in Figure 5. Figures 6 and 7 show analogous results with an average block length of 10,000 bytes (T=6.38 ms).

We observe that, even with the relatively short transfers of 4,096 bytes, the predicted relative improvement may exceed 15%. With the larger transfer length, the relative improvement in average I/O time is predicted to attain 30%. The predicted improvement is lower for the higher number of actuators.

Our second set of numerical results deals with the Dynamic Path Selection (switched substrings) pathing scheme. Here, we set device parameters to those of the IBM 3380 DASD with a transfer rate of 3 Mbytes/s. Again, we assume that 40% of I/Os experience no arm movement while the remaining 60% of accesses are subject to the manufacturer specified average seek time. As previously, we assume a .7 ms search and 1 ms path overhead. We have here S=9.6 ms (average seek time), R=16.67 ms (rotation period) and L=R/2.

Figure 8 illustrates the improvement in the average I/O time for a system with a total of 8 actuators (4 sustrings of 2) and average block length of 4,096 bytes (T=2.07 ms). The results of Figure 9 pertain to a system with 16 actuators (4 substrings of 4) and other parameters kept unchanged. The results obtained for an average block length of 10,000 bytes (T=4.03 ms) are displayed in Figures 10 and 11.

We observe that the Dynamic Path Reconnection feature may be expected to substantially reduce the total I/O time, by up to 40% under moderately heavy loads.

Note that although the total storage capacity in both sets of results is comparable (the IBM 3380 is a double density disk), the results presented are not meant to be used to compare the expected performance of the two disk products. Indeed, the values taken for path overheads do not necessarily match those of the real devices, and additional considerations such as device lead time and geometry have not been taken into account.

As whole, we conclude that Dynamic Path Reconnection can be expected provide a substantial improvement in the average I/O time. This improvement appears to be relatively larger for more congested systems.

The results obtained rely, to a large extent, on the loss system approach used to model missed reconnections. We have run a number of discrete-event simulations to gain confidence in this approach. As an example, we have represented in Figure 12 simulation and analytical results for the set of model parameters corresponding to Figure 10 with Dynamic Reconnection. We have plotted p_1, the first attempt probability of finding the substring free, p_2, the first attempt probability that a substring finds a control

unit path available, and m, the overall expected number of missed reconnections per transfer. Each simulation point corresponds to 100,000 transfer completions. The following distributional assumptions were used in this example: total of seek and latency times - uniform distribution, transfer time - exponential. The agreement between simulation and analytical results appears good.

It is interesting to note that the simulation results seem quite robust vis a vis distributional assumptions. E.g., taking constant transfer times results, in most cases, in a less than 5% change in the average number of missed reconnections. Also, typically a 90% confidence interval for the latter quantity has a width of about 2% of the point estimate.

4.0 CONCLUSION

We have studied the expected improvement in the performance of disks provided by the ability to resume an I/O operation on any transfer path available (Dynamic Path Reconnection). We have considered two popular bi-path connection schemes: dual ported devices (as in the Amdahl 6280 Dynamic Performance Pathing) and switched substring (as in the IBM 3380 Dynamic Path Selection).

A simple classical queueing model has been applied to represent missed revolutions occurring when transfer paths are found busy. The validity of this model has been checked by simulation. The numerical results obatined indicate that a substantial reduction in the average I/O time - up to 30% under moderately heavy loads - can be expected from the Dynamic Path Reconnection feature.

5.0 REFERENCES

1. Artis, H. P.: Predicting the Behavior of Secondary Storage Management Systems for IBM Computer Systems, in The 8-th International Symposium on Computer Performance Modelling, Measurement and Evaluation, November 4-6, 1981, Amsterdam, Holland, North-Holland Publ. Co., pp. 416-435.

2. Brown, D. T., Eibsen, R. L., Thorn, C. A.: Channel and Direct Access Device Architecture, IBM Systems J. 11, 186-198 (1972).

3. IBM 3380 Direct Access Storage Description and User's Guide, IBM Publication GA26-1664-1, San Jose, California (1982).

4. Bard, Y.: A Model of Shared DASD and Multipathing, Comm. ACM 23, 564-572 (1980).

5. Bard, Y.: I/O Systems With Dynamic Path Selection, and General Transmission Networks, Performance Evaluation Review 11, 118-129 (1982).

6. Brandwajn, A.: Multiple Paths Versus Memory For Improving DASD Subsystem Performance, in The 8-th International Symposium on Computer Performance Modelling, Measurement and Evaluation, November 4-6, 1981, Amsterdam, Holland, North-Holland Publ. Co., pp. 401-414.

7. Cooper, R. B.: Introduction to Queueing Theory, The Macmilan Co., New York, 1972.

8. Kleinrock, L.: Queueing Systems, Vol.1: Theory, John Wiley & Sons, New York, 1975.

9. Brandwajn, A.: Models of DASD Subsystems with Multiple Access Paths: A Throughput-Driven Approach, IEEE Transactions on Computers, May 1983.

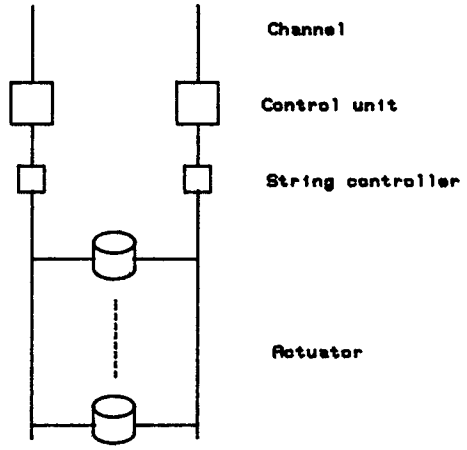

FIGURE 1: Dual ported devices

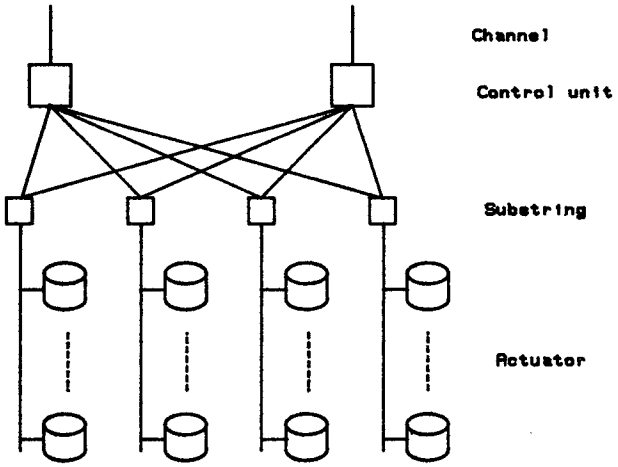

FIGURE 2: Switched substrings

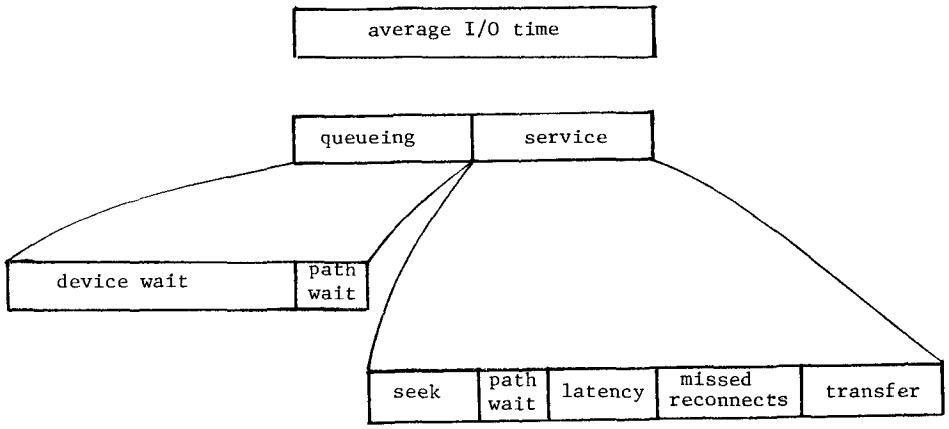

FIGURE 3: Elements of average I/O time

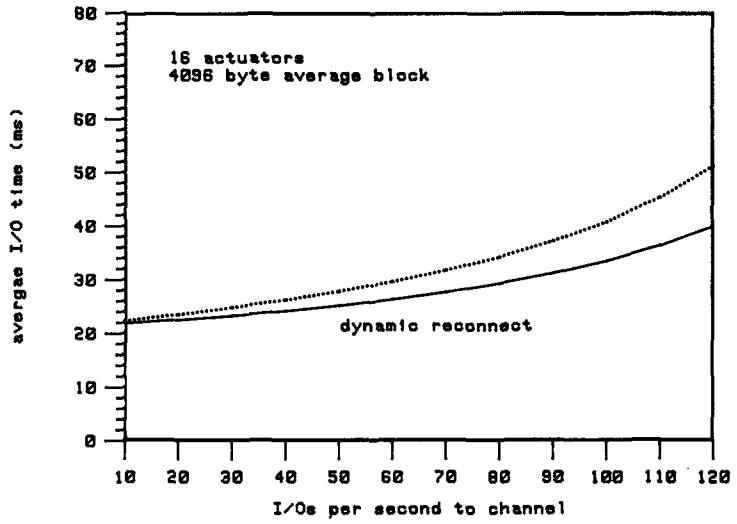

FIGURE 4: Performance with dual ported disks

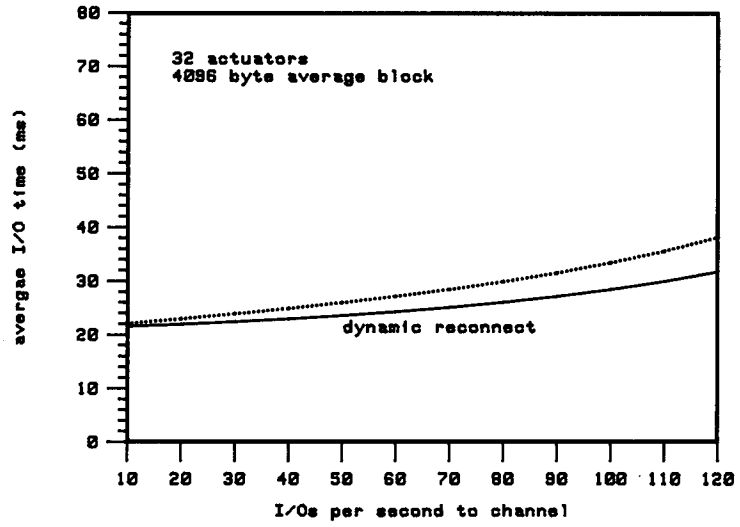

FIGURE 5: Performance with dual ported disks

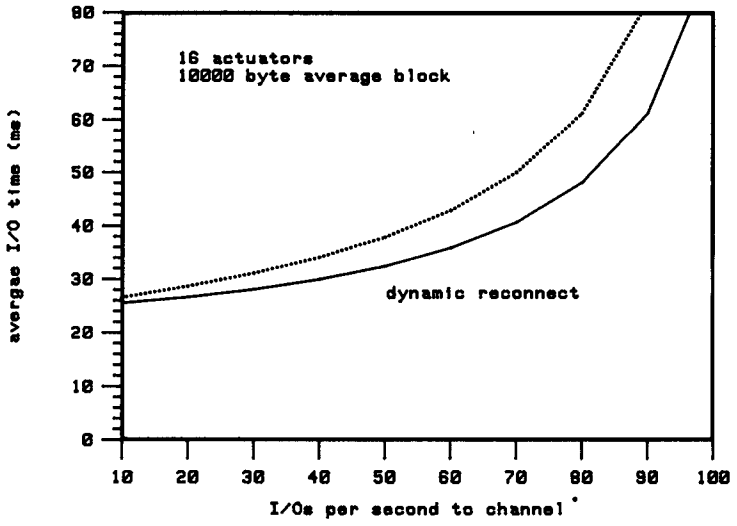

FIGURE 6: Performance with dual ported disks

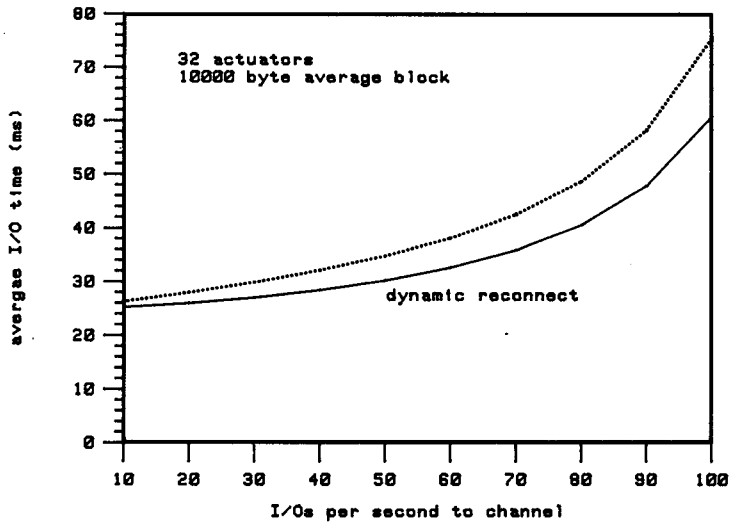

FIGURE 7: Performance with dual ported disks

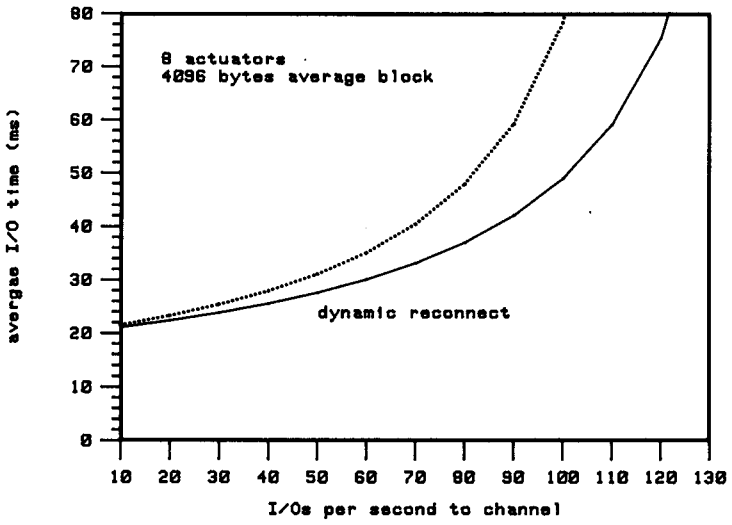

FIGURE 8: Performance with switched substrings

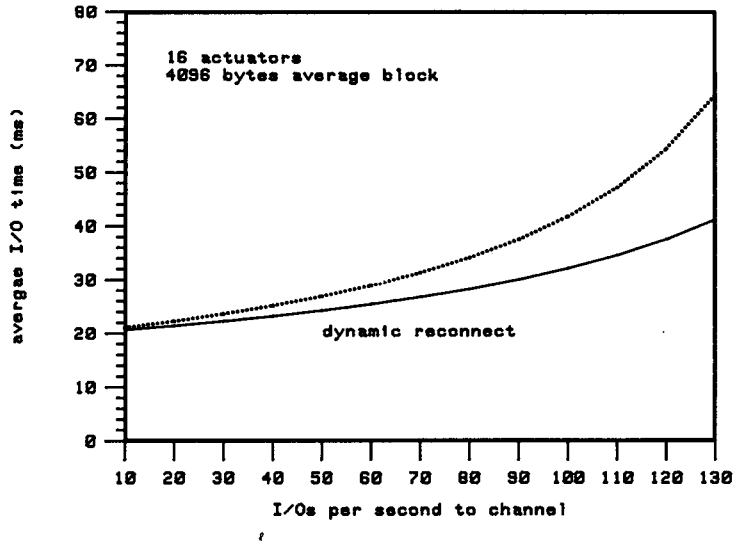

FIGURE 9: Performance with switched substrings

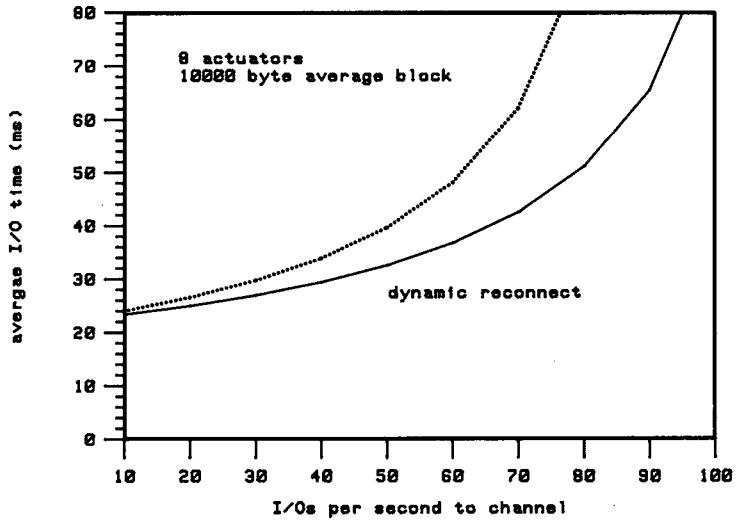

FIGURE 10: Performance with switched substrings

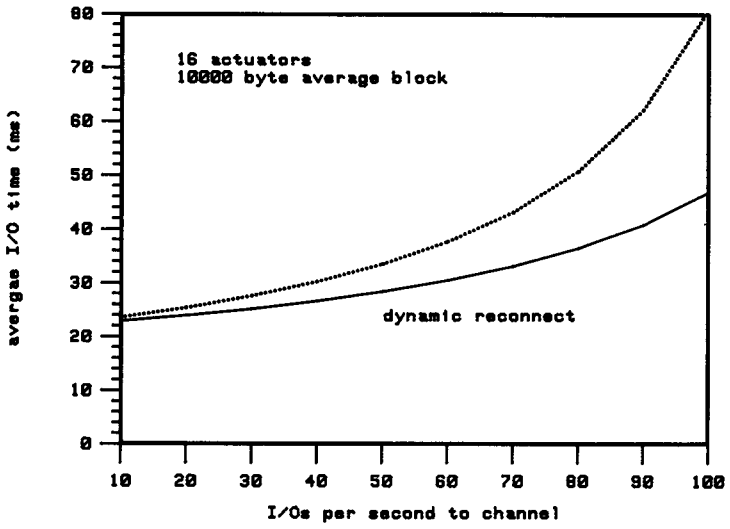

FIGURE 11: Performance with switched substrings

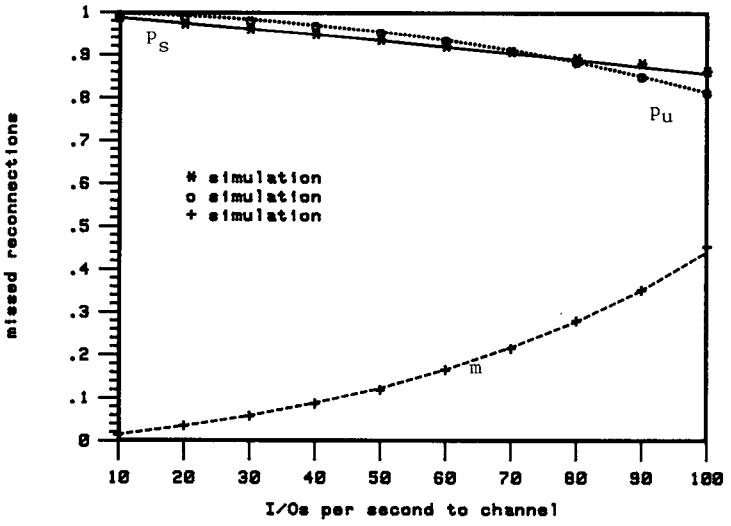

FIGURE 12: Comparison of simulation and analytical results