D. P. Gaver
P. A. W. Lewis
G. S. Shedler

Analysis of Exception Data in a Staging Hierachy

Abstract: This paper is an analysis of program address trace data in a demand-paged computer system with a three-level staging hierarchy. Our primary objective is to explore the data both graphically and numerically, using methods that may be useful when other data traces become available. In addition, plausible point-process type models are fit to the data. Such an approach, combining data-analytic procedures with probability modeling, should prove useful in understanding program behavior and thus will aid in the rational design of complex computer systems.

1. Introduction

Although a number of stochastic (queuing) models for the structure of multiprogrammed computer systems operating under demand paging have been proposed and studied, e.g., [1-4], the probabilistic representation of the behavior of programs running in such systems has received relatively less attention (however, see [5, 6] for examples of two complementary approaches to the formulation of program behavior models). In view of the necessity of understanding the referencing patterns of programs in order to improve the decision algorithms in current and future systems, further studies of program behavior leading to the mathematical characterization of computer system workloads are appropriate.

This paper reports results of a study aimed at better understanding program reference patterns in a demand-paged computer system with a three-level staging hierarchy. The approach taken is, first to represent the actual program-address trace data and, second, to fit the latter with plausible stochastic models. The point of view taken is rather similar to that of [5] concerning the modeling and analysis of page exceptions in a two-level memory. However, in this paper we have emphasized the use of simple graphical methods of statistical data analysis and modeling rather than more formal and complex statistical techniques such as spectral analysis (cf. [7]). Throughout this study we have used an interactive APL/360 computing system, a tool which we have found well suited to this type of statistical analysis and modeling.

The stochastic processes studied here occur in a three-level staging hierarchy. A description of the hierarchy that we assume and of the stochastic process models of interest that suggest themselves therein are given in section 2. The sense in which the exception process in the staging hierarchy can be viewed as a bivariate point process (see [8]) is discussed briefly in section 3, and section 4 contains a description of the data available for analysis. The data analysis and modeling of exception processes are given in sections 5–8. Estimates of values of the parameters in the exception process model are presented in section 9. Section 10 contains an assessment of the fit of the model, and section 11 gives a summary of the results and conclusions.

2. Description of the staging hierarchy

The data sequences studied in this paper occur in a demand-paged computer system having as a storage structure a three-level staging hierarchy, as described in [9]. In such a system all information that is explicitly addressable is divided into units of equal size called *blocks*, each of which is further divided into units of equal size called *pages*. Level 1 of the hierarchy (the execution store) is similarly divided into page-size sections called *page frames* and levels 2 and 3 of the storage hierarchy are divided into block-size sections called *block frames*. In such computer systems it is possible to execute a program by supplying it with only a few page frames and block frames of storage. When the page containing the first executable instruction has been loaded into a page frame, execution begins and continues until some required information is not in the execution store. An attempt to reference information not currently contained in level 1 is termed an *exception*. When a "demand" for

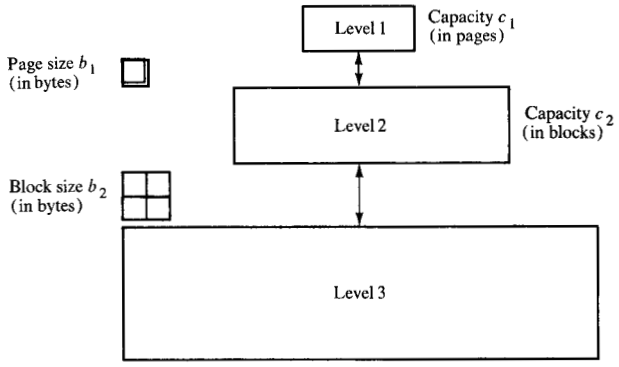


Figure 1 Staging hierarchy.

a page occurs, the page containing the required information is staged through adjacent storage levels up to the execution store in accordance with a prescribed regime for hierarchy management. Thus the hierarchy is *linear*, i.e., there are data paths only between adjacent levels, and data in level 3 must pass through level 2 before going into level 1.

In the three-level staging hierarchies we consider (see Fig. 1), information is transmitted between levels 1 and 2 of the hierarchy in page-size units. The size of a page, in bytes, is denoted by b_1 and the page capacity of level 1 is denoted by c_1. Thus b_1c_1 is the total number of bytes of information which can be contained in level 1. Similarly, information is transmitted between levels 2 and 3 of the hierarchy in block-size units, consisting of an integral number of pages. The byte size of a block is denoted by b_2. The block capacity of hierarchy level 2 is denoted by c_2, and thus b_2c_2 is the byte capacity of level 2. Level 3 of the hierarchy is considered to be of sufficient capacity to contain all information that is explicitly addressable in the computer system.

The staging hierarchy is assumed to be managed under the least recently used (LRU) replacement policy in which reference is broadcast to all levels of the hierarchy. Specifically, if the page containing the required information is currently in level 2 of the hierarchy, the page is fetched (page pull) from level 2, overwriting some page currently in level 1. In general, this replaced page must be written down into level 2 (page push). The page selected to be overwritten is the page in level 1 which has been referenced the least recently of those in this execution storage level i.e., the least recently used page. An instance of an exception of this type, in which the required page is found at level 2 of the hierarchy, is termed a "hit to level 2."

Upon an attempt to reference information neither in level 1 nor in level 2, the page containing the required information is staged up from level 3 as follows. The block containing the desired page is fetched (block pull) from level 3, overwriting the least recently used block in level 2. Again, in general, this replaced block must be written down into level 3 (block push). Now that the block containing the required page resides in level 2, the required page can be transmitted to level 1 (via a page pull and a LRU page push) and the reference can be executed. An instance of an exception of this type, in which the required page is not found at level 2 but only at level 3, is termed a "hit to level 3." In the sequel, we shall use the term *staging hierarchy* to mean a three-level staging hierarchy managed under the LRU replacement policy.

There are many related (but not necessarily equivalent) data sequences that describe page reference patterns in a staging hierarchy; important examples are listed below.

1. *References* $\{R_i(b)\}$, i.e., sequences of page references for pages of size b, where $R_i(b)$ is the name of the page referenced at (discrete) time i.
2. *Distances* $\{D_i(b)\}$, i.e., sequences of stack distances for LRU replacement, as defined in [10], where $D_i(b)$ is the total number of distinct pages (of size b) referenced since the last reference to $R_i(b)$.
3. Sequences corresponding to exceptions to either level for various capacities at levels 1 and 2. We denote such a sequence by $\{T_j(c_1, c_2; b_1, b_2)\}$, where $T_j(c_1, c_2; b_1, b_2)$ is the time (in references) of the jth exception in a three-level staging hierarchy in which level 1 contains c_1 pages of size b_1 and level 2 contains c_2 blocks of size b_2.
4. Sequences corresponding to exceptions of two types. We denote such a sequence by $\{T_j(c_1, c_2; b_1, b_2); h_j(c_1, c_2; b_1, b_2)\}$, where $T_j(c_1, c_2; b_1, b_2)$ is the time of the jth exception and $h_j(c_1, c_2; b_1, b_2)$ equals 2 if the jth exception is a hit to level 2 and $h_j(c_1, c_2; b_1, b_2)$ equals 3 if the jth exception is a hit to level 3. This is a complete description of the pattern of exceptions.

We note here that this last sequence of exceptions of two types is related to distance sequences by the following result (derived in [9]). The relationship facilitates the collection of data on the sequence of exceptions of two types. The bivariate sequence of exceptions of two types can be obtained from two distance sequences $\{D_i(b_1)\}$ and $\{D_i(b_2)\}$.

Proposition 1

Provided that $c_2 \geq c_1$, if $T_j(c_1, c_2; b_1, b_2) = i$, then

$$h_j(c_1, c_2; b_1, b_2) = \begin{cases} 2 \text{ if } D_i(b_1) > c_1 \text{ and } D_i(b_2) \leq c_2 \\ 3 \text{ if } D_i(b_1) > c_1 \text{ and } D_i(b_2) > c_2. \end{cases} \quad (1)$$

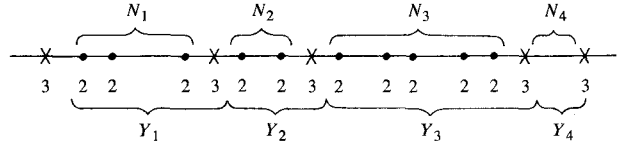
Table 1 Sample characteristics of the exception process for tape B with LRU replacement. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes; number of references is 34,723,105.

	Parameters		
	$c_1 = 32$ $c_2 = 32$	$c_1 = 32$ $c_2 = 64$	$c_1 = 64$ $c_2 = 64$
Number of exceptions	983,596	983,596	475,920
Number of hits to level 3	598	380	380
Number of hits to level 2	982,998	983,216	475,540
Estimated mean time between hits to level 3	58,138.5	91,552.8	91,552.8
Estimated variance of times between hits to level 3	2.564×10^{10}	3.898×10^{10}	3.898×10^{10}
Estimated coefficient of variation of times between hits to level 3	2.754	2.156	2.156
Minimum time between hits to level 3	2	2	2
Maximum time between hits to level 3	1,631,317	1,631,317	1,631,317
Estimated mean number of hits to level 2 between hits to level 3	1,645.3	2,617.1	1,252.8
Estimated variance of number of hits to level 2 between hits to level 3	6.396×10^6	1.472×10^7	3.811×10^6
Estimated coefficient of variation of number of hits to level 2 between hits to level 3	1.537	1.466	1.558
Minimum number of hits to level 2 between hits to level 3	0	0	0
Maximum number of hits to level 2 between hits to level 3	27,898	29,019	16,578

3. Exception processes and point processes

This paper is concerned with the derivation, via the statistical analysis of actual program traces, of empirically valid stochastic models for sequences of exceptions in a three-level staging hierarchy as described in the previous section. The point of view taken in the analysis and modeling is that the exception processes are bivariate point processes [8]. Assuming that the page size b_1 and block size b_2 in the hierarchy have been fixed, and given capacities c_1 and c_2 for levels 1 and 2 of the hierarchy, along with a sequence of page references, an exception can occur on any of the successive page references. If these references are considered to occur at equidistant time points and the interval of time between successive references is taken to be the unit of time, the exceptions constitute a (univariate) point process (series of events) in discrete time, and the exceptions along with their type (hit to level 2 or hit to level 3) constitute a bivariate point process [8]. A realization of this bivariate point process of exceptions is illustrated in Fig. 2, where exceptions that are hits to level 2 are indicated by a dot denoted 2, and exceptions that are hits to level 3 are indicated by a cross denoted 3. Throughout, this bivariate point process of exceptions is termed the *exception process*. The length (in references) of the generic interval between successive hits to level 3 is denoted by Y. We take Y to be the number of references following a hit to level 3 until the next hit to level 3 (including this next hit); thus $Y \geq 1$. The number of hits to level 2 in the interval Y is denoted by N, or by $N(Y)$ when we wish to emphasize the dependence on Y; necessarily for all Y, $0 \leq N(Y) < Y$.

Figure 2 Realization of the process of exceptions of two types: N_i is the number of hits to level 2 and Y_i is the number of references, both counted between the $(i-1)$th and ith hits to level 3.



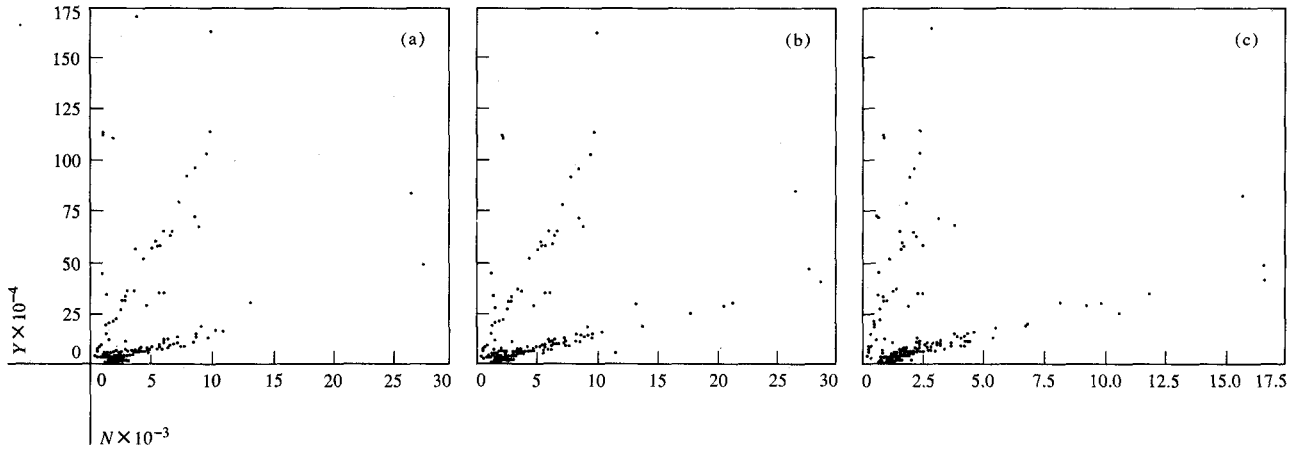


Figure 3 Scatter diagrams of points in the $[Y, N(Y)]$ plane for $b_1 = 256$, $b_2 = 32{,}768$, and (a) $c_1 = 32$, $c_2 = 32$; (b) $c_1 = 32$, $c_2 = 64$; (c) $c_1 = 64$, $c_2 = 64$.

4. Data and preliminaries

Although data from several programs have been examined, results displayed in this paper are for a particular address trace referred to as tape B. From the sequence of addresses traced, the sequence of LRU distances were derived by stack processing techniques [10] for various page sizes b in the range 256 to 32,768 bytes. Also, for various choices of pairs of capacities c_1 and c_2 in the range 16 to 64, sequences of exceptions were obtained from pairs of distance sequences by the techniques described in [9]. Results are given here for three pairs of capacities $c_1 = 32$ pages, $c_2 = 32$ blocks; $c_1 = 32$, $c_2 = 64$; and $c_1 = 64$, $c_2 = 64$, all for the case $b_1 = 256$ bytes and $b_2 = 32{,}768$ bytes. The trace data consisted of 34,723,105 references to 166 distinct 32,768-byte blocks.

In a staging hierarchy encountered in practice, the number of hits to level 2 is typically several orders of magnitude larger than the number of hits to level 3. This is, in fact, the case for the data obtained from tape B (see Table 1) for which there are several hundred hits to level 3, but hundreds of thousands of hits to level 2 (over the range of b_1, b_2, c_1, and c_2 considered). Thus hundreds of thousands of intervals and point-type pairs would be required for a complete description of the exception process. Such a voluminous amount of data is not only difficult to comprehend, it is also expensive to manipulate. As a result, the statistical analysis and modeling described in this paper was based solely on the $\{Y\}$ and $\{N(Y)\}$ sequences—respectively, the *intervals* between successive hits to level 3 and the *counts* of hits to level 2 between successive hits to level 3. Much potentially informative data can be obtained by display and analysis of the *time positions* of level 2 hits.

Some sample characteristics of the data obtained from tape B are displayed in Table 1. Sample characteristics for four non-overlapping sections of the data were examined. No indication of gross departure from stationarity was observed. Accordingly, the assumption of stationarity was made in our analysis, details of which are given in the next three sections.

5. Graphical study of points in the $[Y, N(Y)]$ plane

Our analysis of the available data began with a set of scatter diagrams of points in the $[Y, N(Y)]$ plane (see Fig. 3). These three scatter diagrams reveal the apparent existence, in the material under scrutiny, of *two* distinct kinds of referencing behavior. For each of the three pairs of capacities ($c_1 = 32$, $c_2 = 32$; $c_1 = 32$, $c_2 = 64$; and $c_1 = 64$, $c_2 = 64$) there is a striking two-line relationship in the graphical display of the observed values of Y and the corresponding values of $N(Y)$. By this we mean that points in the $[Y, N(Y)]$ plane appear to be of two types, and in each of the two types, points of that type seem to be clustered about a straight line (through the point $Y = 1$, $N(Y) = 0$) in the plane. The data analysis that has been done proceeds from this observed double linearity of points in the $[Y, N(Y)]$ plane. The discovery of this empirical relationship suggested the probability models that we formulated. Work remains to explain the phenomenon in terms of program peculairities and to establish its generality, or lack thereof. The existence of a two-state phenomenon is not surprising in view of the data-analytic results of [5]; the linearity in the two states, is, however, quite striking.

As a further step, for each pair of capacities, the m observed points (y_j, n_j) in the $[Y, N(Y)]$ plane were

partitioned into two disjoint sets by means of a *separation line* $\mathscr{L}$. One method for separating the points is by a least-squares line through the point (1, 0). It is easily shown that the slope s of the least squares line is

$$s = \sum_{j=1}^{m} n_j y_j \Big/ \sum_{j=1}^{n} n_j^2. \tag{2}$$

Alternatively, we can take as the separation line the line determined by the points (1, 0) and (y_r, n_r) where (y_r, n_r) is the vector sum resultant of the set of observed points $\{(y_j, n_j)\}$. The latter method is equivalent to taking as the inverse of the slope of the separation line the maximum likelihood estimate of the rate of occurrence of events for a Bernoulli process model of the hits to level 2 of the hierarchy as described in Section 7. This estimate for the rate of occurrence is simply

$$\frac{1}{s} = \sum_{j=1}^{m} n_j \Big/ \sum_{j=1}^{m} (y_j - 1). \tag{3}$$

The resulting classification of the points is somewhat sensitive to the method of separation because of the relatively large number of points that are close to the origin (see Table 2). On purely empirical grounds, we chose as the separation line $\mathscr{L}$ the Bernoulli process maximum likelihood (vector sum resultant) line. The equation for this separation line $\mathscr{L}$ is

$$Y = sN(Y) + 1. \tag{4}$$

The resulting separation of points is summarized in Table 3.

Table 2 Slopes of fitted separation lines in the $[Y, N(Y)]$ plane for tape B with LRU replacement. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes.

	Parameters		
	$c_1 = 32$ $c_2 = 32$	$c_1 = 32$ $c_2 = 64$	$c_1 = 64$ $c_2 = 64$
Least-squares fit	36.875	29.474	48.922
Maximum likelihood fit (Bernoulli process)	35.336	34.982	73.078

Table 3 Separation of points in the $[Y, N(Y)]$ plane for tape B with LRU replacement. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes.

	Parameters		
	$c_1 = 32$ $c_2 = 32$	$c_1 = 32$ $c_2 = 64$	$c_1 = 64$ $c_2 = 64$
Total number of points	597	379	379
Number of upper points (Bernoulli maximum likelihood separation)	65	60	56
Number of lower points (Bernoulli maximum likelihood separation)	532	319	323
Proportion of upper points	0.1089	0.1583	0.1478

6. Properties of the intervals between successive hits to level 3

In this section we study the sequence of Y intervals between successive hits to level 3. In the next section, based on this analysis, we consider models for the counts of the number of hits to level 2 occurring in an interval between successive hits to level 3.

Note that for fixed page size b_1 and block size b_2, the sequence of Y intervals is determined by the value of c_2. For $c_2 = 32$ as well as $c_2 = 64$, the sequences of Y intervals have been examined. (Sample characteristics of the Y intervals are given in Table 1. In both cases, the marginal distributions of Y are quite positively skewed, having estimated coefficients of variation in excess of 2 (see Table 1). Moreover the sequence of Y values is correlated; the three estimated first serial correlation coefficients $\hat{\rho}_1$ all differ significantly from zero under the normal approximation. [Asymptotically, $\hat{\rho}_1$ is normally distributed with $\mathrm{E}[\hat{\rho}_1] \sim 0$ and $\mathrm{Var}[\hat{\rho}_1] \sim (n-1)^{-1}$ under fairly general conditions when $\rho_1 = 0$; cf. [7], p. 92.]

In view of the two types of behavior suggested by the observed double linearity in the $[Y, N(Y)]$ plane and the point classification we have made, it is reasonable to examine the sequence of Y components, conditional on point type. We consider two sequences of intervals, Y_1 and Y_2 intervals, derived from the original sequence of Y components. Sample characteristics of the marginal distribution of these two sequences are given in Table 4; corresponding characteristics of the counts of hits to level 2 between hits to level 3 are given in Table 5. Evidently, the marginal distributions of Y_1, the type 1 Y components, and of Y_2, the type 2 Y components, are quite different, with the mean of Y_1 being approximately ten times larger than the mean of Y_2.

Histograms of the distributions of the Y_1 intervals suggest a mixture of random variables. Plots of the logarithm of the empirical survivor function $\tilde{R}_{Y_1}(y)$, where

$$\tilde{R}_{Y_1}(y) = \frac{\text{number of } Y_1 \text{ intervals greater than } y}{\text{number of } Y_1 \text{ intervals}}, \qquad y = 1, 2, \cdots \tag{5}$$

are generally convex with a linear tail. This suggests the use of a mixture of two geometric (plus one) distributions as a model for the marginal distribution of Y_1; i.e., to assume that for $0 < \pi_1 < 1$ and $0 < q_{11}, q_{12} < 1$,

$$\Pr\{Y_1 = y\} = \pi_1 q_{11}^{\,y-1}(1 - q_{11}) + (1 - \pi_1) q_{12}^{\,y-1}(1 - q_{12}), \qquad y = 1, 2, \cdots. \tag{6}$$

Table 4 Sample characteristics of intervals between successive hits to level 3 conditioned on point type for tape B with LRU replacement. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes.

	Parameters					
	$c_1 = 32$, $c_2 = 32$		$c_1 = 32$, $c_2 = 64$		$c_1 = 64$, $c_2 = 64$	
	Type 1	*Type 2*	*Type 1*	*Type 2*	*Type 1*	*Type 2*
Number of hits	65	532	60	319	56	323
Estimated mean time between hits	323,256	25,746	350,378	42,871	376,424	42,163
Estimated variance of times between hits	1.196×10^{11}	1.32×10^{9}	1.311×10^{11}	3.126×10^{9}	1.410×10^{11}	3.049×10^{9}
Estimated coefficient of variation of times between hits	1.145	1.98	1.068	1.701	0.995	1.715
Maximum time between hits	1,631,317	850,172	1,631,317	850,172	1,631,317	850,172
Minimum time between hits	2	27	2	27	2	27

Table 5 Sample characteristics of number of hits to level 2 between hits to level 3 conditioned on point type for tape B with LRU replacement. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes.

	Parameters					
	$c_1 = 32$, $c_2 = 32$		$c_1 = 32$, $c_2 = 64$		$c_1 = 64$, $c_2 = 64$	
	Type 1	*Type 2*	*Type 1*	*Type 2*	*Type 1*	*Type 2*
Estimated mean	2,797.5	1,504.5	3,078.2	2,530.4	1,048.1	1,288.3
Estimated variance	8.842×10^{6}	5.931×10^{6}	9.120×10^{6}	1.576×10^{7}	8.742×10^{5}	4.315×10^{6}
Estimated coefficient of variation	1.063	1.619	0.9810	1.59	0.892	1.612
Maximum number	9,940	27,898	9,940	29,019	3,724	16,578
Minimum number	0	2	0	2	0	2

Denoting by Y_{11} and Y_{12} random variables having these geometric (plus one) distributions, we have for $j = 1, 2$, $\mathrm{E}[Y_{1j}] = (1 - q_{1j})^{-1}$, $\mathrm{Var}[Y_{1j}] = q_{ij}(1 - q_{ij})^{-2}$, and $C^2(Y_{1j}) = q_{1j}$. Similar considerations suggest taking a mixture of two geometric (plus one) distributions as a model for the marginal distribution of Y_2; i.e., to assume that for $0 < \pi_2 < 1$ and $0 < q_{21}, q_{22} < 1$,

$$\Pr\{Y_2 = y\} = \pi_2 q_{21}{}^{y-1}(1 - q_{21}) + (1 - \pi_2) q_{22}{}^{y-1}(1 - q_{22}). \quad (7)$$

Geometric (plus one) random variables with parameters q_{21} and q_{22}, respectively, are denoted by Y_{21} and Y_{22}.

The Y_1 and Y_2 sequences has been examined by estimating serial correlation coefficients (see Table 6). In the case of the Y_2 intervals, the first two estimated serial correlation coefficients do not differ significantly from zero. In fact, no indication of dependence in the Y_2 sequences has been found.

Although the Y_2's can apparently be considered as independent, identically distributed, random variables having a mixture distribution, there is evidence that the Y_1's are a dependent mixture. In the case of the Y_1 sequences, the sample sizes may be too small to justify the use of the normal approximation for the estimated first serial correlation coefficients. On the basis of this approximation, however, only for the sequence corresponding to $c_1 = 64$, $c_2 = 64$ is there evidence that the first serial correlation differs from zero. Note (in Table 6) that estimates of second-order correlation coefficients indicate dependence in the Y_1 sequences, although the nature of the dependence is difficult to assess from the available data. Accordingly, in our model we consider the Y_1's to be an independent mixture.

7. Models for counts of hits to level 2

In Section 5, a double-line relationship between Y and $N(Y)$ was observed by means of a scatter diagram in the

$[Y, N(Y)]$ plane, this observation leading us to partition the set of data points into two disjoint sets: those of type 1, the points clustered about the upper line, and type 2, the points clustered about the lower line. In this section we concentrate on the phenomenon of linearity in the $[Y, N(Y)]$ plane, with the aim of postulating a point process model that accounts for the observed linearity.

Recalling that for $i = 1, 2$ we denote the generic Y component of a type i point by Y_i, we denote the corresponding N component by $N(Y_i)$. Then, for $y = 1, 2, 3, \cdots$, the observed linear relationship for type i points can be summarized as

$$\mathrm{E}[N(Y_i)|Y_i = y] = p_i(y-1), \tag{8}$$

where $0 < p_i < 1; \quad i = 1, 2.$

We ultimately seek a plausible stochastic mechanism for generating a bivariate series of events corresponding to the hits to level 2 and hits to level 3. We begin by considering the interval between successive hits to level 3. Consideration of models for the sequence of intervals between hits to level 3 and its relationship to the two point types is deferred to later sections of the paper. Perhaps the simplest way to account for the observed linearity is to assume that the hits to level 2 occur "at random" in the interval between hits to level 3. More specifically, given a point in the $[Y, N(Y)]$ plane of type i, within an interval Y_i between successive hits to level 3, hits to level 2 occur according to a Bernoulli process. Thus the number of hits to level 2 between successive hits to level 3 is conditionally binomial; i.e., for $y \geq 1$ and $0 < p_i < 1$,

$$\Pr\{N(Y_i) = n | Y_i = y\} = \binom{y-1}{n} p_i^{\,n}(1-p_i)^{y-1-n},$$

$$0 \leq n \leq y-1. \tag{9}$$

It is quite easy to show that under this assumption, $\mathrm{E}[N(Y_i)|Y_i = y] = p_i(y-1)$.

In view of our finding (in section 6) that the marginal distributions of Y_1 and Y_2 are mixtures, it seems plausible to consider a mixed Bernoulli process model in which the parameter of the process depends on the distribution from which the interval between hits to level 3 is generated. The number of hits to level 2 is conditionally binomial; i.e., for $1 \leq i, j \leq 2$,

$$\Pr\{N(Y_{ij}) = n | Y_{ij} = y\} = \binom{y-1}{n} p_{ij}^{\,n}(1-p_{ij})^{y-1-n},$$

$$0 < p_{ij} < 1; \; n = 0, 1, \cdots, y-1. \tag{10}$$

(Recall that Y_i is a mixture of Y_{i1} and Y_{i2}.) It is easy to show that linearity in the $[Y, N(Y)]$ plane is retained.

Proposition 2

Let $y \geq 1$. If for $1 \leq i, j \leq 2$ and some $0 < p_{ij} < 1$,

$$\Pr\{N(Y_{ij}) = n | Y_{ij} = y\} = \binom{y-1}{n} p_{ij}^{\,n}(1-p_{ij})^{y-1-n},$$

$$n = 0, 1, \cdots, y-1,$$

and for $0 < \pi_i < 1$,

$$\Pr\{Y_i = y\} = \pi_i \Pr\{Y_{i1} = y\} + (1-\pi_i)\Pr\{Y_{i2} = y\},$$

then

$$\mathrm{E}[N(Y_i)|Y_i = y] = [\pi_i p_{i1} + (1-\pi_i)p_{i2}](y-1)$$

$$\equiv p_i(y-1). \tag{11}$$

Table 6 Estimated serial correlation coefficients for intervals between hits to level 3 for tape B with LRU replacement. Normalized values are given in parentheses. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes.

	Parameters		
Coefficient/points	$c_1 = 32$ $c_2 = 32$	$c_1 = 32$ $c_2 = 64$	$c_1 = 64$ $c_2 = 64$
First/all	0.228	0.201	0.201
	(5.565)	(3.902)	(3.902)
Second/all	0.633	0.592	0.592
	(15.432)	(11.486)	(11.486)
First/type 1	−0.165	−0.241	−0.307
	(−1.849)	(−1.849)	(−2.275)
Second/type 1	0.431	0.459	0.492
	(3.419)	(3.500)	(3.612)
First/type 2	−0.046	0.016	0.001
	(−1.049)	(0.282)	(0.174)
Second/type 2	−0.0075	−0.021	−0.001
	(−0.172)	(−0.377)	(−0.205)

Table 7 Estimated transition probabilities for Markov chain models for the sequence of point types for tape B with LRU replacement. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes.

	Parameters		
	$c_1 = 32$ $c_2 = 32$	$c_1 = 32$ $c_2 = 64$	$c_1 = 64$ $c_2 = 64$
Zeroth-order model			
$\hat{p}(1)$	0.1089	0.1583	0.1478
$\hat{p}(1)$	0.8911	0.8417	0.8522
First-order model			
$\hat{p}(1, 1)$	0.5231	0.5667	0.6250
$\hat{p}(1, 2)$	0.4769	0.4333	0.3750
$\hat{p}(2, 1)$	0.0565	0.0786	0.0621
$\hat{p}(2, 2)$	0.9435	0.9214	0.9379
Second-order model			
$\hat{p}(1, 1, 1)$	0.8529	0.8824	0.8571
$\hat{p}(1, 1, 2)$	0.1471	0.1176	0.1429
$\hat{p}(1, 2, 1)$	0.3548	0.3462	0.1905
$\hat{p}(1, 2, 2)$	0.6452	0.6538	0.8095
$\hat{p}(2, 1, 1)$	0.1667	0.1600	0.2500
$\hat{p}(2, 1, 2)$	0.8333	0.8400	0.7500
$\hat{p}(2, 2, 1)$	0.0380	0.0548	0.0532
$\hat{p}(2, 2, 2)$	0.9620	0.9452	0.9468

Table 8. Estimated serial correlation coefficients $\hat{\rho}_j$ of lag j for process of runs of point types for tape B with LRU replacement. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes.

	Parameters		
j	$c_1 = 32$ $c_2 = 32$	$c_1 = 32$ $c_2 = 64$	$c_1 = 64$ $c_2 = 64$
1	−0.2374	−0.2404	−0.2383
2	0.2464	0.2198	0.0501
3	−0.2410	−0.2234	−0.2330
4	0.0804	−0.0345	0.0493
5	−0.2078	−0.1763	−0.0887
6	0.1426	0.0817	0.3482
7	−0.1421	−0.0952	−0.1949
8	0.1781	0.3332	0.0257
9	−0.1706	−0.1642	−0.1192
10	0.3611	0.2759	0.1129
11	−0.1964	−0.1508	−0.1099
12	0.0393	−0.0745	0.0641
13	−0.2272	−0.1799	−0.2142
14	0.1394	0.1467	0.3451
15	−0.2107	−0.1626	−0.1318
16	0.4173	0.4035	0.0532
17	−0.1853	−0.1229	−0.1415
18	0.2284	0.2683	−0.0617
19	−0.2031	−0.1851	−0.0849
20	0.1051	−0.0374	0.3043
21	−0.1549	−0.1213	−0.1155
22	0.0008	−0.0228	
23	−0.1508	−0.0388	
24	0.0370	0.1765	
25	−0.1304	−0.1708	
26	0.1644	0.2118	
27	−0.1009		
28	0.1350		
29	−0.1428		
30	0.0701		
31	−0.1326		

Thus, a mixed Bernoulli process model for the hits to level 2 between successive hits to level 3 is consistent with the examination of the data that we have made so far.

Such a mixed Bernoulli process model for the type i hits to level 2 is attractive since it is specified by but two parameters which, as we show in section 9, can be estimated conveniently by the method of moments. To complete the formulation of a model for the exception process, it remains to examine the sequence of point types. This is the topic of the next section.

8. Analysis of the process of point types

In seeking a model for the process of point types, a natural choice (by virtue of its simplicity) is a two-state (Markov) chain, including independent trials as a special case. Estimates of conditional probabilities of transition for zeroth-, first-, and second-order chains have been obtained and are given in Table 7. Denoting the process of point types by $\{\tau_i\}$, $i \geq 1$, where

$$\tau_i = \begin{cases} 1 \text{ if the } i\text{th point is of type 1} \\ 2 \text{ if the } i\text{th point is of type 2,} \end{cases} \tag{12}$$

we let, in the case of the zeroth-order (independent trials) chain,

$$p(i) = \Pr\{\tau_m = i\},\ i = 1, 2. \tag{13}$$

For the first-order chain we let

$$p(i, j) = \Pr\{\tau_m = j | \tau_{m-1} = i\},\ 1 \leq i, j \leq 2; \tag{14}$$

and for the second-order chain we let

$$p(i, j, k) = \Pr\{\tau_m = k | \tau_{m-1} = j,\ \tau_{m-2} = i\},\quad 1 \leq i, j, k \leq 2. \tag{15}$$

Evidently, the estimated conditional probabilities indicate gross departures from an independent trials model, and give little support for a first-order Markov chain model.

An alternative to a Markov chain for the process of point types is an interval model, i.e., one based on the sequence of runs of points of the same type. We now consider this process of intervals, the alternating sequence of lengths L_1 of runs of points of type 1 and lengths L_2 of runs of points of type 2. Estimates of the serial correlation coefficients $\hat{\rho}_j$ of lag j have been computed for the sequence of runs and appear in Table 8. The striking feature of these estimated correlation coefficients is the almost strict alternation of signs, suggesting an alternating renewal process model for the sequence of runs of point types. For an alternating renewal process the serial correlation coefficient of lag j [7, p. 196] is $\alpha(-1)^j$, where

$$\alpha = \left[\frac{(\sigma_1^2 + \sigma_2^2)}{2(\mu_2 - \mu_1)^2} + 1\right]^{-1}, \tag{16}$$

and μ_i and σ_i^2 are, respectively, the mean and variance of the marginal distribution of the lengths of runs of type i. Estimates of these quantities along with estimates of α are given in Table 9. Note, however, that the estimated serial correlation coefficients are consistently smaller in magnitude than the estimated α. Estimates of the serial correlation coefficients have been computed for the sequence of lengths of runs of type 1 points and the sequence of lengths of runs of type 2 points. Although interpretation of these estimated correlation coefficients is difficult because of the small sample sizes involved, there is no strong indication of dependence. Accordingly, we adopt an alternating renewal process model of the sequence of runs of point types.

With respect to appropriate forms for the distributions of the lengths of runs, note (Table 9) that the runs of type 2 points have relatively large estimated means and estimated coefficients of variation close to one. Runs of type 2 points have much smaller estimated means, but larger estimated coefficients of variation.

Plots of the estimated log-survivor functions $\ln R_{L_1}(x)$ and $\ln R_{L_2}(x)$ have been examined. From the general shape of these plots we are led to entertain as the distribution of the length L_1 of runs of type 1 points a negative binomial (plus one) distribution with scale parameter r_1 and shape parameter ℓ_1; i.e.,

$$\Pr\{L_1 = i\} = \binom{\ell_1+i-2}{i-1} r_1^{\,i-1}(1-r_1)^{\ell_1},$$
$$0 < r_1 < 1;\ \ell_1 > 0;\ i = 1, 2, \cdots. \qquad (17)$$

For the distribution of the length L_2 of runs of type 2 points, we also assume a negative binomial (plus one) distribution with parameters r_2 and ℓ_2; i.e.,

$$\Pr\{L_2 = i\} = \binom{\ell_2+i-2}{i-1} r_2^{\,i-1}(1-r_2)^{\ell_2},$$
$$0 < r_2 < 1;\ \ell_2 > 0;\ i = 1, 2, \cdots. \qquad (18)$$

9. Estimates of parameters

In the proposed model of the exception process, using (17) and (18) a sequence of point types (type 1 or 2) is generated according to an alternating renewal process for the lengths of runs of point types. Given a point of type i in this sequence, an interval Y_i between successive hits to level 3 of the hierarchy is generated from the appropriate mixture distribution. The distribution from which the interval between hits to level 3 was sampled (Y_{i1} or Y_{i2}) determines a Bernoulli process and $N(Y_i)$ hits to level 2 are generated within the interval Y_i according to Eq. (10).

There are thus 14 parameters to estimate:

Bernoulli processes of hits to level 2

$\hat{p}_{11}$ for type 1 hits to level 2 $[N(Y_{11})]$
$\hat{p}_{12}$ " 1 " 2 $[N(Y_{12})]$
$\hat{p}_{21}$ " 2 " 2 $[N(Y_{21})]$
$\hat{p}_{22}$ " 2 " 2 $[N(Y_{22})]$

Renewal processes of intervals between hits to level 3

$\hat{q}_{11}$ for intervals Y_{11} (geometric (plus one))
$\hat{q}_{12}$ " Y_{12} "
$\hat{\pi}_1$ mixing probability for Y_1 intervals
$\hat{q}_{21}$ for intervals Y_{21} (geometric (plus one))
$\hat{q}_{22}$ " Y_{22} "
$\hat{\pi}_2$ mixing probability for Y_2 intervals

Alternating renewal processes of runs of point types

$\hat{r}_1$ scale parameter } for lengths L_1 of runs of points of type 1 (negative binomial (plus one))
$\hat{\ell}_1$ shape parameter }
$\hat{r}_2$ scale parameter } for lengths L_2 of runs of points of type 2 (negative binomial (plus one))
$\hat{\ell}_2$ shape parameter }

Table 9 Sample characteristics of sequence of runs of point types for tape B with LRU replacement. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes.

	Parameters		
	$c_1 = 32$, $c_2 = 32$	$c_1 = 32$, $c_2 = 64$	$c_1 = 64$, $c_2 = 64$
$\hat{\alpha}$	0.5087	0.5058	0.4787
Mean of runs of points of type 1	2.097	2.308	2.667
Variance of runs of points of type 1	11.49	17.66	21.43
Coefficient of variation of runs of points of type 1	1.617	1.82	1.74
Maximum of runs of points of type 1	17	17	17
Minimum of runs of points of type 1	1	1	1
Mean of runs of points of type 2	17.13	12.23	15.38
Variance of runs of points of type 2	424.9	174.7	330.5
Coefficient of variation of runs of points of type 2	1.203	1.081	1.18
Maximum of runs of points of type 2	81	42	63
Minimum of runs of points of type 2	1	1	1

Parameters of the model were estimated from the data in an ad hoc manner. The parameters q_{i1} were estimated from the slopes of the linear tails of the log-survivor functions of Y_i; this involved a visual judgment of where the linearity set in. For $c_1 = c_2 = 32$ these points were taken to be 350,000 for Y_1 and 75,000 for Y_2 (cf. [5, p. 95]). The parameters π_i and q_{i2} in the geometric (plus one) distribution mixture were then obtained by matching the estimated mean and variance of the marginal distribution of Y_i. This was accomplished by using the following construction for a mixture of this kind.

Let $\mu > 0$ and $\sigma > 0$ be given such that $\sigma^2 > \mu^2 + \mu$. It is easily verified that for $0 < \mu_q < \mu$, if

$$\mu_p = \mu + \frac{\sigma^2 - \mu - \mu^2}{2(\mu - \mu_q)} \quad \text{and} \qquad (19)$$

$$\pi = \frac{2(\mu - \mu_q)^2}{\sigma^2 - \mu - \mu^2 + 2(\mu - \mu_q)^2}, \qquad (20)$$

then $0 < \pi < 1$, $\mu_q < \mu < \mu_p$, $\pi\mu_p + (1-\pi)\mu_q = \mu$, and $\pi(2\mu_p^{\,2} + \mu_p) + (1-\pi)(2\mu_q^{\,2} + \mu_q) = \sigma^2 + \mu^2$. Thus the mixture of two geometric (plus one) distributions specified by

$$\Pr\{X = x\} = \pi p^{x-1}(1-p) + (1-\pi)q^{x-1}(1-q),$$
$$x = 1, 2, \cdots, \qquad (21)$$

Table 10 Estimated parameters for exception process model for tape B with LRU replacement. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes; level 1 capacity $c_1 = 32$ pages; level 2 capacity $c_2 = 32$ blocks. The raised number indicates how many times the preceding digit is repeated.

Parameter	*Value*
$\hat{p}_{11}$	0.00728
$\hat{p}_{12}$	0.0409
$\hat{p}_{21}$	0.0461
$\hat{p}_{22}$	0.1208
$\hat{\mu}_{11}$ (conditional mean > 350,000)	386,788
$\hat{\mu}_{12}$	66,750
$\hat{q}_{11}$	$0.9^{5}74$
$\hat{q}_{12}$	$0.9^{4}85$
$\hat{\pi}_1$	0.801
$\hat{\mu}_{21}$ (conditional mean > 75,000)	74,801
$\hat{\mu}_{22}$	5,982
$\hat{q}_{21}$	$0.9^{4}86$
$\hat{q}_{22}$	$0.9^{3}83$
$\hat{\pi}_2$	0.287
$\hat{r}_1$	0.90
$\hat{l}_1$	0.12
$\hat{r}_2$	0.96
$\hat{l}_2$	0.64

with p and q determined by $\mu_p = (1-p)^{-1}$ and $\mu_q = (1-q)^{-1}$, has mean $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$. The estimated parameters were obtained according to (19) and (20) using $\mu = \hat{\mathrm{E}}[Y_i]$ and $\sigma^2 = \widehat{\mathrm{Var}}[Y_i]$, $\hat{q}_{i2}$ being chosen such that $\log \hat{q}_{i1}$ is equal to the estimated slope of the linear tail of the log-survivor function of Y_i.

We now consider the estimation of the parameters in the Bernoulli processes of hits to level 2. It can be shown that if Y_{ij} has a geometric (plus one) distribution with parameter q_{ij} and the number of hits to level 2 is conditionally binomial with parameter p_{ij}, then $N(Y_{ij})$ has a geometric distribution with $p_{ij}q_{ij}[1 - q_{ij}(1-p_{ij})]^{-1}$ as a parameter. Using this fact, estimates of the p_{ij} can be obtained by matching the first moment of the marginal distribution of $N(Y_i)$ and the first moment of the product $N(Y_i)Y_i$. Specifically, having values for $\hat{\pi}_i$, $\hat{q}_{i1}$ and $\hat{q}_{i2}$, estimated parameters $\hat{p}_{i1}$ and $\hat{p}_{i2}$ were obtained as the solution of the simultaneous equations

$$\hat{\mathrm{E}}[N(Y_i)] = \pi_i \frac{p_{i1}q_{i1}}{(1-q_{i1})} + (1-\pi_i)\frac{p_{i2}q_{i2}}{(1-q_{i2})}; \tag{22}$$

$$\hat{\mathrm{E}}[N(Y_i)Y_i] = \pi_i \frac{2p_{i1}q_{i1}}{(1-q_{i1})^2} + (1-\pi_i)\frac{2p_{i2}q_{i2}}{(1-q_{i2})^2}. \tag{23}$$

For the runs L_1 and L_2 of point types, the scale and shape parameters of the assumed negative binomial (plus one) distributions were obtained by the method of moments, i.e., $\hat{r}_i$ and $\hat{\ell}_i$ were obtained, for $i = 1, 2$, as the solution of the simultaneous equations

$$\hat{\mathrm{E}}[L_i] = 1 + \ell_i r_i/(1-r_i); \tag{24}$$

$$\widehat{\mathrm{Var}}[L_i] = \ell_i r_i/(1-r_i)^2. \tag{25}$$

The estimates of the parameters in the exception process model are given in Table 10 for $c_1 = 32$, $c_2 = 32$. We denote by μ_{ij} the quantities $(1-q_{ij})^{-1}$.

10. Tests of the fit of the model

We now consider the fit of the proposed model by examining computed and estimated characteristics of the model for $c_1 = c_2 = 32$. The marginal distribution of intervals Y between successive hits to level 3 in the model can be easily obtained. For $y = 1, 2, \cdots$,

$$\begin{aligned}\Pr\{Y = y\} = \beta[&\pi_1 q_{11}^{\,y-1}(1-q_{11}) \\ &+ (1-\pi_1)q_{12}^{\,y-1}(1-q_{12})] \\ &+ (1-\beta)[\pi_2 q_{21}^{\,y-1}(1-q_{21}) \\ &+ (1-\pi_2)q_{22}^{\,y-1}(1-q_{22})],\end{aligned} \tag{26}$$

where

$$\beta = \frac{\mathrm{E}[L_1]}{\mathrm{E}[L_1] + \mathrm{E}[L_2]} = \frac{(1-r_1)(1-r_2) + (1-r_2)\ell_1 r_1}{2(1-r_1)(1-r_2) + (1-r_2)\ell_1 r_1 + (1-r_1)\ell_2 r_2} \tag{27}$$

is the stationary probability in the alternating renewal process of point types that a point is of type 1.

In Fig. 4(a) the empirical log-survivor function (dots) for the intervals Y is shown with the corresponding theoretical log-survivor function (solid line) computed from (26) using the estimated parameters in Table 9. Note that we are validating or testing using the same data that were used for fitting parameters. Although this procedure is convenient, it is questionable and provides a relatively weak measure of goodness-of-fit. It would be desirable to validate the model using other data.

Proceeding, similarly, we can obtain the marginal distribution of counts $N(Y)$ of hits to level 2 between hits to level 3. For $n = 0, 1, 2, \cdots$,

$$\begin{aligned}\Pr\{N(Y) = n\} = \beta\Big[&\pi_1\Big(\frac{p_{11}q_{11}}{1-q_{11}(1-p_{11})}\Big)^n\Big(\frac{1-q_{11}}{1-q_{11}(1-p_{11})}\Big) \\ &+ (1-\pi_1)\Big(\frac{p_{12}q_{12}}{1-q_{12}(1-p_{12})}\Big)^n\Big(\frac{1-q_{12}}{1-q_{12}(1-p_{12})}\Big)\Big] \\ &+ (1-\beta)\Big[\pi_2\Big(\frac{p_{21}q_{21}}{1-q_{21}(1-p_{21})}\Big)^n\Big(\frac{1-q_{21}}{1-q_{21}(1-p_{21})}\Big) \\ &+ (1-\pi_2)\Big(\frac{p_{22}q_{22}}{1-q_{22}(1-p_{22})}\Big)^n\Big(\frac{1-q_{22}}{1-q_{22}(1-p_{22})}\Big)\Big].\end{aligned} \tag{28}$$

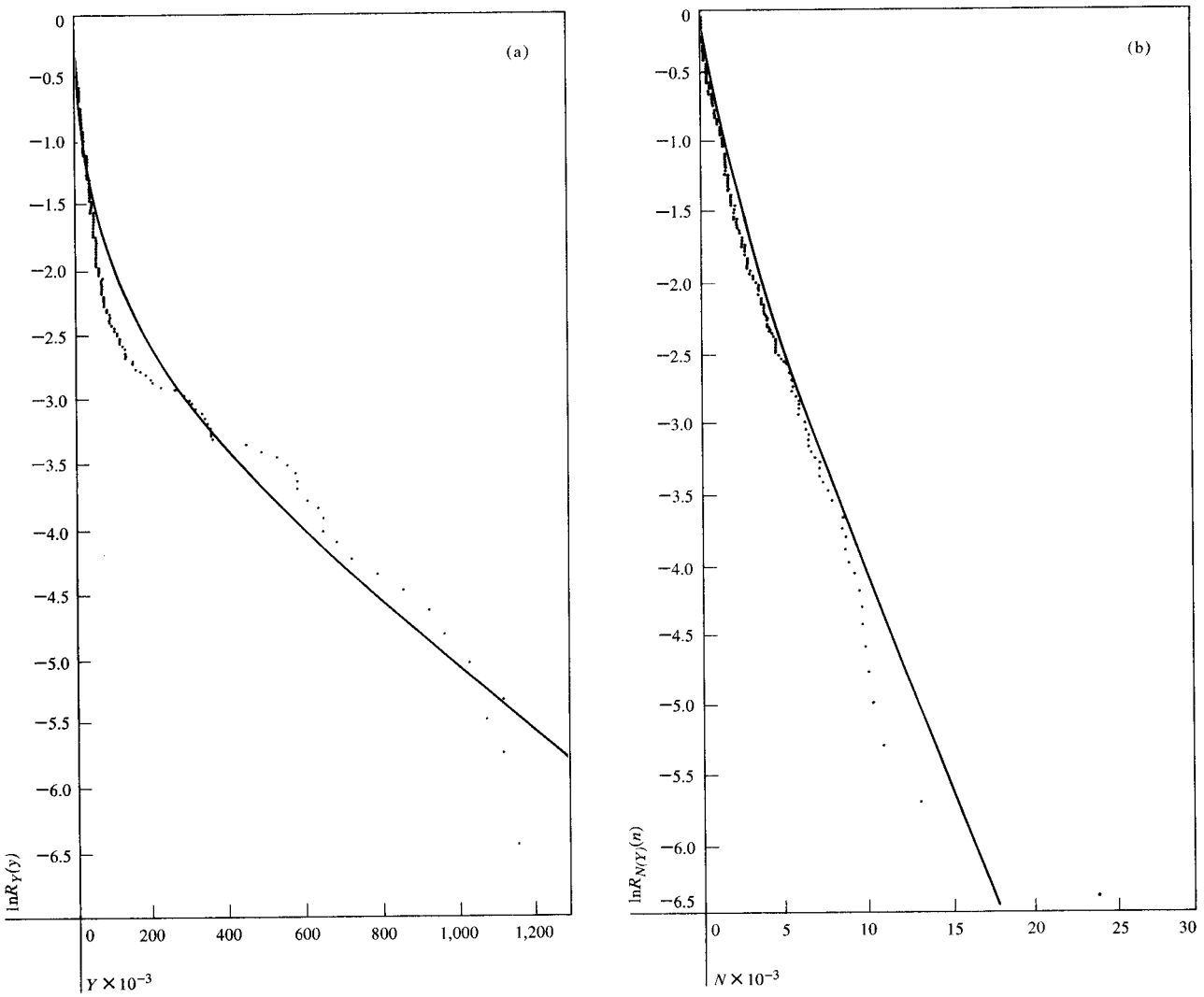


Figure 4 Log-survivor functions for (a) intervals Y between hits to level 3 and (b) counts $N(Y)$ of hits to level 2 between hits to level 3; $b_1 = 256$, $b_2 = 32{,}768$, $c_1 = 32$, $c_2 = 32$.

The empirical log-survivor function for the counts $N(Y)$ is shown in Fig. 4(b) with the corresponding theoretical log-survivor function.

To get some idea of the extent to which the dependence structure of the model is consistent with the observed dependence between Y and $N(Y)$, we consider the cross-correlation between the variables defined by

$$\rho(N(Y), Y) = \frac{\mathrm{E}[N(Y)Y] - \mathrm{E}[N(Y)]\mathrm{E}[Y]}{\{\mathrm{Var}[N(Y)]\mathrm{Var}[Y]\}^{\frac{1}{2}}}. \qquad (29)$$

We sketch the computation of this quantity for the proposed exception process model. Expressions for the first and second moments of $N(Y)$ are given by

$$\mathrm{E}[N(Y)] = \beta\left[\frac{\pi_1 p_{11} q_{11}}{(1-q_{11})} + \frac{(1-\pi_1)p_{12}q_{12}}{(1-q_{12})}\right] + (1-\beta)\left[\frac{\pi_2 p_{21} q_{21}}{(1-q_{21})} + \frac{(1-\pi_2)p_{22}q_{22}}{(1-q_{22})}\right]; \qquad (30)$$

$$\mathrm{E}[N^2(Y)] = \beta\left[\pi_1\left(\frac{p_{11}q_{11}}{(1-q_{11})} + \frac{2p_{11}^2 q_{11}^2}{(1-q_{11})^2}\right) + (1-\pi_1)\left(\frac{p_{12}q_{12}}{(1-q_{12})} + \frac{2p_{12}^2 q_{12}^2}{(1-q_{12})^2}\right)\right] + (1-\beta)\left[\pi_2\left(\frac{p_{21}q_{21}}{(1-q_{21})} + \frac{2p_{21}^2 q_{21}^2}{(1-q_{21})^2}\right) + (1-\pi_2)\left(\frac{p_{22}q_{22}}{(1-q_{22})} + \frac{2p_{22}^2 q_{22}^2}{(1-q_{22})^2}\right)\right]. \qquad (31)$$

Table 11 Cross correlation between Y and $N(Y)$ for tape B. Page size $b_1 = 256$ bytes; block size $b_2 = 32{,}768$ bytes. Estimated variances of $\hat{\rho}[N(Y), Y]$ are given in parentheses.

	Parameters		
	$c_1 = 32$ $c_2 = 32$	$c_1 = 32$ $c_2 = 64$	$c_1 = 64$ $c_2 = 64$
Computed $\rho[N(Y), Y]$	0.62	0.55	0.38
Estimated $\hat{\rho}[N(Y), Y]$	0.59 (0.01)[a]	0.52 (0.05)[a]	0.38 (0.07)[a]

[a]Values obtained from four sections of the data.

An expression for $\mathrm{Var}[N(Y)]$ is obtained from (30) and (31). The corresponding expression for $\mathrm{Var}[Y]$ is obtained from (26) via

$$\mathrm{E}[Y] = \beta\left[\frac{\pi_1}{(1-q_{11})} + \frac{(1-\pi_1)}{(1-q_{12})}\right] + (1-\beta)\left[\frac{\pi_2}{(1-q_{21})} + \frac{(1-\pi_2)}{(1-q_{22})}\right]; \tag{32}$$

$$\mathrm{E}[Y^2] = \beta\left[\frac{\pi_1(q_{11}+1)}{(1-q_{11})^2} + \frac{(1-\pi_1)(q_{12}+1)}{(1-q_{12})^2}\right] + (1-\beta)\left[\frac{\pi_2(q_{21}+1)}{(1-q_{21})^2} + \frac{(1-\pi_2)(q_{22}+1)}{(1-q_{22})^2}\right]. \tag{33}$$

Finally

$$\mathrm{E}[N(Y), Y] = \beta\left[\frac{\pi_1 2p_{11}q_{11}}{(1-q_{11})^2} + \frac{(1-\pi_1)2p_{12}q_{12}}{(1-q_{12})^2}\right] + (1-\beta)\left[\frac{\pi_2 2p_{21}q_{21}}{(1-q_{21})^2} + \frac{(1-\pi_2)2p_{22}q_{22}}{(1-q_{22})^2}\right], \tag{34}$$

and $\rho[N(Y), Y]$ is obtained from (30) – (34) according to (29).

The computed values of $\rho[N(Y), Y]$ along with the estimated values $\hat{\rho}[N(Y), Y]$ are given in Table 11. Estimates of the variance of the $\hat{\rho}[N(Y), Y]$ obtained from four sections of the data are given in parentheses.

11. Summary and concluding remarks

0. We have shown how the process of exceptions in a three-level LRU staging hierarchy may be represented graphically in the $[Y, N(Y)]$ plane. For the particular program analyzed an unexpected two-line configuration appeared.
1. A tentative model has been proposed for the bivariate point process of exceptions in the staging hierarchy, based on the observation for realizations of the process of intervals between hits to level 3 of the hierarchy and counts of hits to level 2. The graphical display of 0. above was instrumental in suggesting the model.
2. Parameters of the model have been estimated from the available data in an ad hoc manner.
3. The fit of the model has been examined by comparing the empirical log-survivor functions of intervals between hits to level 3 and counts of hits to level 2 between hits to level 3 with the computed theoretical log-survivor functions and also by comparing the estimated cross-correlation of intervals between hits to level 3 and counts of hits to level 2 with the computed theoretical value. On the basis of these measures, the fit is reasonably good.
4. A striking indication of the existence of two types of paging behavior was observed – a double linear relationship between intervals between hits to level 3 of the hierarchy and counts of hits to level 2.

Several limitations of the study should be mentioned.

1. It would be desirable to formalize the procedure for estimating parameters and also to obtain estimates of parameters from sections of the data in order to examine the sensitivity of the estimation procedure: Error estimates, such as rough confidence limits, for the parameters are also needed.
2. The study should be done for more page and block sizes, as well as capacities, to yield more information on how the parameters change.
3. Relatively weak measures of goodness-of-fit have been used. For example, the marginal distribution of intervals between hits to level 3 does not depend on detailed assumptions about the distributions of runs of point types of the conditional process of hits to level 2.
4. More program tapes should be examined to confirm (or deny) double linearity (or multiple linearity) of intervals between hits to level 3 and counts of hits to level 2. Explanations for this behavior should be deduced in the hope that they will lead to improved hierarchy designs.
5. It would be desirable to relate the parameters of the model directly to the basic hierarchy design parameters (page size, block size, and capacities). Some work has been done on this problem and will be reported elsewhere.

Acknowledgment

The work of D. P. Gaver and P. A. W. Lewis was partially supported by National Science Foundation grant AG 476 at the Naval Postgraduate School, Monterey, California. These authors are also consultants to the IBM Research Division.

References

1. D. P. Gaver and G. S. Shedler, "Processor Utilization in Multiprogramming Systems Via Diffusion Approximations," *Oper. Res.* **21**, 569 (1973).
2. D. P. Gaver and G. S. Shedler, "Approximate Models for Processor Utilization in Multiprogrammed Computer Systems," *SIAM J. Computing* **2**, 183 (1973).
3. S. S. Lavenberg, "Queuing Analysis of a Multiprogrammed Computer System Having a Multilevel Storage Hierarchy," *SIAM J. Computing* **2**, 232 (1973).
4. P. A. W. Lewis and G. S. Shedler, "A Cyclic-Queue Model of System Overhead in Multiprogrammed Computer Systems," *J. ACM* **18**, 199 (1971).
5. P. A. W. Lewis and G. S. Shedler, "Empirically Derived Micromodels for Sequences of Page Exceptions," *IBM J. Res. Develop.* **17**, 86 (1973).
6. G. S. Shedler and C. Tung, "Locality in Page-Reference Strings," *SIAM J. Computing* **1**, 218 (1972).
7. D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Methuen Ltd., London, and Barnes and Noble, Inc., New York, 1966.
8. D. R. Cox and P. A. W. Lewis, "Multivariate Point Processes," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 3, edited by L. LeCam, J. Neyman and E. L. Scott, University of California Press, Berkeley, 1972, p. 401.
9. D. R. Slutz and I. L. Traiger, "Determination of Hit Ratios for a Class of Staging Hierarchies," Research Report RJ 1044, IBM Research Laboratory, San Jose, California, 1972.
10. R. L. Mattson, J. Gecsei, D. R. Slutz and I. L. Traiger, "Evaluation Techniques for Storage Hierarchies," *IBM Syst. J.* **9**, 78 (1970).

Received April 15, 1974

D. P. Gaver and P. A. W. Lewis are located at the Naval Postgraduate School, Monterey, California 93940; G. S. Shedler is at the IBM Research Laboratory, Monterey and Cottle Roads, San Jose, California 95193.