Analysis of replication in distributed data bases

François Baccelli, E.G. Coffman

INRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. 954 90 20

Rapports de Recherche

N° 150

ANALYSIS OF REPLICATION IN DISTRIBUTED DATA BASES

François BACCELLI
Edward G. COFFMAN Jr.

Août 1982

ANALYSIS OF REPLICATION IN DISTRIBUTED DATA BASES

by

F. Baccelli
I.N.R.I.A.
B.P. 5 - Rocquencourt
LeChesnay, France

and

E. G. Coffman, Jr.
Bell Laboratories
Murray Hill, New Jersey

I. Introduction

A study of file replication policies for distributed data bases will be approached through the analysis of an M/M/m queue subjected to state-independent, preemptive interruptions of service. The durations of periods of interruption constitute a sequence of independent, identically distributed random variables. Independently, the times measured from the termination of one period of interruption to the beginning of the next form a sequence of independent, exponentially distributed random variables. Preempted customers resume service at the terminations of interrupt periods.

Such a model is readily interpreted within systems representing certain machine repair or retrofit problems. However, although our analysis will be conducted initially in these more general terms, the application stimulating the present paper corresponds to a system providing two modes of service under a preemptive-resume regime.

ANALYSIS OF REPLICATION IN DISTRIBUTED DATA BASES

François BACCELLI
Edward G. COFFMAN Jr

Résumé

On analyse le problème de la duplication des fichiers dans des systèmes tels que les bases de données réparties ou certains mécanismes de section critique. Dans ces systèmes, la duplication des fichiers permet un accès parallèle des requêtes de lecture. Cependant, ces dernières sont périodiquement interrompues par l'arrivée de mises à jour devant être exécutées prioritairement sur chaque copie des fichiers. La durée de mise à jour de l'ensemble des copies est usuellement une fonction croissante du nombre m de ces copies (à cause des algorithmes de maintien de leur cohérence mutuelle). Ainsi, le taux global de traitement des lectures augmente et celui des mises à jour diminue avec m. Un premier objectif dans ce type de problèmes, fut de déterminer pour un tel système, la valeur de m maximisant le débit en lectures sous des hypothèses de saturation. L'approche plus générale développée ici permet d'aborder l'optimisation du temps de réponse des lectures en fonction de m.

Abstract

We model replicated files in distributed data bases and certain critical section problems. In such systems, the file replication allows parallel access for reads. However these last are periodically interrupted by the arrival of updates which have to be processed with a preemptive priority on all the file copies. The update time is modeled as an increasing function of the number of copies, owing to such effects as limited parallelism, coherence control mechanisms, etc. Thus, since the read service rate increases with m, but that for the updates decreases with m, a primary objective of the analysis has been to find optimum values of m that maximize read throughput under saturation assumptions. Our more general approach will allow such an optimization under the expected read-waiting time measure.

In our specific application, we model replicated files in distributed data bases [1,2] and certain critical section problems [1]. In such systems we have an m-server "read" queue that is interrupted preemptively by the arrival of "updates" (or "writes") whose processing constitutes an M/G/1 queue in which a service period consists of the time required to execute write operations updating all m servers. The update time is modeled as an increasing function of m, owing to such effects as limited parallelism, coherence control mechanisms, etc. (See [1,2] for a full discussion.) Thus, since the service rate at the read queue increases with m, but that at the write queue decreases with m, a primary objective of the analysis has been to find optimum values of m that maximize throughput at a saturated read queue. Our more general approach will allow such an optimization under the expected read-waiting time measure.

In the next two sections we develop a general analysis of the M/M/m queue with service interruptions. In section IV we specialize the terminology and results to our data-base application.

II. Notation

Let λ and μ be the parameters of the exponential interarrival and service times, respectively, in the M/M/m

system, and let γ be the parameter of the exponential distribution governing the times between periods of interrupted service. Periods of interruption have the distribution $S(x)$ which is assumed to have the transform $S^*(s)$, density $s(x)$, a finite first moment $1/\sigma$, and a hazard rate $\sigma(x) = s(x)/(1-S(x))$ that exists for all $x \geq 0$.

The state of the system is represented by the triple $Z_t = (N_t, X_t, Y_t)$ where $N_t \geq 0$ is the number in the read queue at time t; X_t is a binary macrostate variable signifying an available ($X_t = 0$) or interrupted ($X_t = 1$) system at time t; and if $X_t = 1$, $Y_t \geq 0$ is equal to the time elapsed since the beginning of the current interruption, and if $X_t = 0$, Y_t is equal to 0. Clearly, $\{Z_t, t \geq 0\}$ is a Markov process whose behavior can be described by the functions

$$(1) \quad \begin{cases} p(t,n,0,0) = \Pr\{N_t=n, X_t=0, Y_t=0 \mid N_0=0, X_t=0, Y_t=0\}, & n \geq 0 \\ p(t,n,1,y) = \dfrac{\partial}{\partial y} \Pr\{N_t=n, X_t=1, Y_t \leq y\}, & n \geq 0,\ y \geq 0. \end{cases}$$

We assume complete convergence of Z_t as $t \to \infty$, and hence the existence of the limits $p(n,0) = \lim_{t\to\infty} p(t,n,0,0)$ and $p(n,1,y) = \lim_{t\to\infty} p(t,n,1,y)$.

Next, we calculate the limiting state probabilities with the eventual objective being an analytical expression

for the joint distribution, $\Pr\{N_t = n, Y_t \leq y\}$, of the number in the queue and the time elapsed since the beginning of the current period of interruption.

III. Analysis

By a routine application of Kolmogorov's Forward Equations the limiting probabilities must satisfy

$$(2)\qquad \begin{cases} [\lambda+\gamma+\mu \min(m,n)]p(n,0) = \lambda p(n-1,0) \\ + \mu \min(m,n+1)p(n+1,0) + \int_0^\infty p(n,1,y)\sigma(y)dy, \quad n \geq 1 \end{cases}$$

$$(3)\qquad (\lambda+\gamma)p(0,0) = \mu p(1,0) + \int_0^\infty p(0,1,y)\sigma(y)dy$$

$$(4)\qquad \frac{\partial}{\partial y} p(n,1,y) + [\lambda+\sigma(y)]p(n,1,y) = \lambda p(n-1,1,y), \quad n \geq 1, \ y > 0$$

$$(5)\qquad p(n,1,0) = \gamma p(n,0), \quad n \geq 0$$

$$(6)\qquad \frac{\partial}{\partial y} p(0,1,y) + [\lambda+\sigma(y)]p(0,1,y) = 0.$$

Defining the transforms

$$(7)\qquad G(z,0) = \sum_{n\geq 0} p(n,0)z^n, \quad |z| \leq 1$$

$$(8)\qquad G(z,1,y) = \sum_{n\geq 0} p(n,1,y)z^n, \quad |z| \leq 1,$$

equations (2)-(6) yield in the usual way

$$(9)\qquad \begin{cases} G(z,0)\left[\lambda(1-z)+\gamma+m\mu\left(1-\frac{1}{z}\right)\right] = \\ \mu\left(1-\frac{1}{z}\right)\sum_{j=0}^{m-1}(m-j)z^{j}p(j,0) = \int_{0}^{\infty}G(z,1,y)\sigma(y)dy \end{cases}$$

and

$$(10)\qquad G(z,1,y) = \gamma G(z,0)\exp -\left[\lambda y(1-z) + \int_{0}^{y}\sigma(x)dx\right]$$

Thus, on substitution and simplification we find

$$(11)\qquad G(z,0) = \frac{\mu\left(1-\frac{1}{z}\right)\sum_{j=0}^{m-1}(m-j)p(j,0)z^{j}}{\lambda(1-z) + m\mu\left(1-\frac{1}{z}\right) + \gamma[1-S^{*}(\lambda(1-z))]}$$

For the joint distribution that we seek, it is readily seen that the transform

$$H(z,y) = \lim_{t\to\infty}\sum_{n\geq 0}\Pr\{N_t = n, Y_t \leq y\}z^{n}$$

is given by

$$(12)\qquad H(z,y) = G(z,0)\left[1 + \int_{0}^{y}\exp\left\{-\left[\lambda x(1-z) + \int_{0}^{x}\sigma(u)du\right]\right\}dx\right],$$

so that the generating function for the number of customers in steady state is

$$(13)\qquad H(z) = G(z,0)\cdot\left[1 + \frac{\gamma}{\lambda}\cdot\frac{1-S^{*}(\lambda(1-z))}{1-z}\right].$$

Thus, the solution to (11) that we now develop completely determines our joint distribution in equilibrium.

First, let $C(z) = [1-S^*\lambda(1-z))]/(1-z)$, so that (11) can be rendered as

$$(14)\quad \left(1 - \frac{z}{m\mu}\,[\lambda+\gamma C(z)]\right) G(z,0) = \sum_{j=0}^{m-1} \left(\frac{m-j}{m}\right) p(j,0) z^j$$

Expanding $C(z)$ in power series yields

$$C(z) = \left[\sum_{n\geq 0} z^n\right]\left[1 - \sum_{\ell\geq 0} z^\ell \int_0^\infty \frac{(\lambda t)^\ell}{\ell!}\, e^{-\lambda t}\, dS(t)\right]$$

Selecting the coefficient, $c(k)$, of z^k we find

$$(15)\quad c(k) = \sum_{n\geq k+1} \int_0^\infty \frac{(\lambda t)^n}{n!}\, e^{-\lambda t}\, dS(t)$$

Next, the $p(j,0)$ may be calculated as follows.

Lemma. Define the recurrence

$$q(0) = 1$$

$$(16)\quad q(j) = \frac{m}{j} \sum_{\ell=0}^{j-1} q(\ell) f(j-1-\ell),\quad 1 \leq j \leq m-1,$$

where

$$f(0) = \frac{1}{m\mu}(\lambda+\gamma c(0))$$

(17)

$$f(j) = \frac{\gamma}{m\mu} c(j),\ 1 \le j \le m-1.$$

Then the coefficients $p(j,0)$, $0 \le j \le m-1$, are given by

$$(18)\quad p(0,0) = \frac{\frac{\sigma}{\sigma+\gamma} - \frac{\lambda}{m\mu}}{\sum_{j=0}^{m-1} (\frac{m-j}{m})q(j)}$$

and

$$(19)\quad p(j,0) = p(0,0)q(j),\ 0 \le j \le m-1.$$

Proof. Consider the (m-1)-degree polynomials

$$P(z) = \sum_{j=0}^{m-1} p(j,0)z^j$$

$$R(z) = \frac{1}{m} \sum_{j=0}^{m-1} jp(j,0)z^j$$

Separating out the first m-1 terms of its expansion, $G(z,0)$ may be written

$$G(z,0) = P(z) + z^m L(z)$$

where $L(z)$ is analytic in the unit circle. Substituting into (14) and re-arranging, we obtain

$$(20) \quad R(z) - zP(z)F(z) = z^m L(z)[zF(z)-1]$$

where

$$F(z) = \frac{1}{m\mu}\,[\lambda+\gamma C(z)]$$

Thus, since the right-hand side of (20) and each of its first m-1 derivatives vanishes at $z = 0$, the coefficient of z^j in $R(z)$ must be equal to the coefficient of z^{j-1} in $P(z)F(z)$ for each $1 \le j \le m-1$. Furthermore, $F(z)$ is analytic in the unit circle and the coefficients of its expansion are the $f(j)$ in (17). This verifies that, when considering $p(0,0)$ as an undetermined constant, the $g(j) = p(j,0)/p(0,0)$, $0 \le j \le m-1$, satisfy the recurrence in (16).

The remaining unknown parameter, $p(0,0)$, is determined from

$$G(1,0) = \frac{p(0,0)\sum_{j=0}^{m-1}(\frac{m-j}{m})q(j)}{1 - \frac{\lambda}{m\mu}\,(1+\gamma/\sigma)}$$

and by (12),

$$H(1,\infty) = 1 = G(1,0)(1+\gamma/\sigma).$$

Substitution results in (18). ■

Accumulating the results so far and addressing the stability question, we obtain our main result.

Theorem. The function $H(z) = \lim_{t\to\infty} \sum_{n\geq 0} \Pr\{N_t = n\} z^n$, $|z| \leq 1$, for the number in the system in equilibrium is the generating function of a proper probability distribution if and only if the condition

$$\lambda < \frac{m\mu}{1+\gamma/\sigma} \tag{21}$$

is fulfilled. In this case, the function $H(z)$ is given by

$$H(z) = \frac{\left[1 + \frac{\gamma}{\lambda} \cdot \frac{1-S^*(\lambda(1-z))}{1-z}\right]\left[\sum_{j=0}^{m-1} \frac{m-j}{m} p(j,0) z^j\right]}{1 - \frac{\lambda z}{m\mu}\left[1 + \frac{\gamma}{\lambda} \cdot \frac{1-S^*(\lambda(1-z))}{1-z}\right]} \tag{22}$$

where the m unknown coefficients $p(j,0)$, $0 \leq j \leq m-1$, are determined from the lemma with the $c(j)$ given by (15).

Proof. Informally, the stability condition follows from the fact that $\frac{1/\gamma}{1/\gamma+1/\sigma}$ is the fraction of time the higher priority interrupt process is idle, and hence the fraction of time available to the M/M/m queue. More precisely, the necessity of the condition is obtained from (18): If (21) is not satisfied $p(0,0) \leq 0$ and the function $H(z)$ cannot be a generating function. For sufficiency we first prove that if $\lambda < m\mu/(1+\gamma/\sigma)$ then $|zF(z)| < 1$ for all $|z| \leq 1$.

We have

$$\frac{\partial S^*(\lambda(1-z))}{\partial z} = \lambda\int_0^\infty ye^{-\lambda y(1-z)}s(y)dy.$$

For $|z| \leq 1$ and $z \neq 1$ classical inequalities yield $\left|\frac{\partial S^*(\lambda(1-z))}{\partial z}\right| < \lambda/\sigma$, whereupon integration yields $|S^*(\lambda(1-z))-1| < \frac{\lambda}{\sigma}|z-1|$. Furthermore, $[S^*(\lambda(1-z))-1](z-1) \to \lambda/\sigma$ as $z \to 1$. Thus $|[S^*(\lambda(1-z))-1]/(z-1)| \leq \lambda/\sigma$, $|z| \leq 1$, and (21) implies $|zF(z)| < 1$ for all $|z| \leq 1$. Hence, when (21) holds, the denominator in (22) has no root in the unit circle, and H(z) is therefore analytic in this domain.

From (15) the c(k), $k \geq 0$, are positive. Thus, use of the lemma and the fact that (21) implies $p(0,0) > 0$, shows that the p(j,0), $0 \leq j \leq m-1$, are positive. Expressing H(z) in the form

$$H(z) = \left[1 + \frac{\gamma}{\lambda}C(z)\right]\left[\sum_{j=0}^{m-1}\frac{m-j}{m}p(j,0)z^j\right]\left[\sum_{n\geq 0}(zF(z))^n\right],$$

it is readily seen that the coefficients in the expansion of H(z) must be positive. H(z) also satisfies H(1) = 1, so that H(z) is the generating function of a proper probability distribution. The remainder of the theorem simply combines (13) and (14) with the lemma. □

From the results of this section expected values can be found in the usual ways. Of course, the computational

complexity will be determined primarily by the form of $S^*(z)$ and the calculation of the $c(k)$. In connection with the data base application mentioned earlier, these calculations along with related computational issues are illustrated in the next section.

IV. Optimal Data-Base File Replication

In our specific model of the data base problem described in the Introduction, we assume that the service interruptions of the M/M/m queue correspond to the busy periods of an M/M/1 queue. The M/M/m queue services read requests in parallel on m identical file replicas, and the M/M/1 queue services the high priority update (write) requests, i.e., requests to update all m file replicas. Accordingly, the expected value, $1/\delta$, of the exponential update distribution will be treated as an increasing function of m.

Consistent with our earlier definitions, the intervals between consecutive periods of interruption of the read queue are now to be interpreted as the update interarrival periods in a Poisson process with parameter γ. Clearly, as m grows the read queue benefits from the greater parallelism, but suffers from the reduced availability of the file replicas caused by the increased update times and busy periods. In assessing this trade-off we shall look at both the equilibrium expected read time, W, and the

maximum (saturated) read throughput rate, T. Past analyses [1,2] have been restricted to the latter performance measure.

T is easily found as $m\mu$ times the fraction of time in equilibrium that the update queue is idle, i.e. the probability, $1-\gamma/\delta$, that the M/M/1 queue is empty. Thus,

$$(23) \quad T = \frac{m\mu}{1-\gamma/\delta}$$

To obtain W from our earlier results, it is convenient to make use of the busy-period equation for the update queue

$$(24) \quad S^*(z) = B^*(z+\gamma(1-S^*(z)))$$

where $B^*(s) = \delta/(\delta+s)$ is the transform of the exponential update distribution. From $C(z) = [1-S^*(\lambda(1-z))]/(1-z)$ we obtain after substitution and routine manipulations

$$(25) \quad C(z)^2[\gamma(1-z)] + C(z)[\delta+\lambda(1-z)-\gamma] = \lambda$$

Accordingly,

$$(26) \quad \gamma c(0)^2 + (\delta+\lambda-\gamma)c(0) - \lambda = 0$$

Since $\delta < \gamma$ is implied by a stable update queue, the positive real solution to (26) is given by

$$(27) \qquad c(0) = \frac{\sqrt{(\delta+\lambda-\gamma)^2+4\lambda\gamma} - (\delta+\lambda-\gamma)}{2\gamma}$$

For the coefficient $c(k)$ in the expansion of $C(z)$ it is routine to verify that the relation

$$(28) \qquad c(k) = \frac{1}{2\gamma c(0)+\delta+\lambda-\gamma} \left\{ c(k-1)(\lambda+\gamma c(0)) + \gamma \sum_{\ell=1}^{k-1} c(\ell)[c(k-1-\ell)-c(k-\ell)] \right\}, \quad k \geq 1,$$

is implied by (25).

Next, we can compute W by applying Little's result, $W = \bar{n}/\lambda$, where $\bar{n}$ is the expected number of read requests in the system in equilibrium. Letting $E(S^2)$ be the second moment of the busy period distribution, $S(t)$, we differentiate (13) and obtain

$$\bar{n} = H'(z)\Big|_{z=1} = G'(z,0)\Big|_{z=1}(1+\gamma/\sigma) + G(1,0)\,\frac{\gamma\lambda}{2}\,E(S^2)$$

where

$$G(1,0) = \frac{1}{1+\gamma/\sigma}$$

$$G'(z,0)\Big|_{z-1} = \frac{G(1,0)[F(1)+F'(z)\Big|_{z=1}] + \sum_{j=1}^{m-1} \frac{m-j}{m}\, jp(j,0)}{1 - F(1)}$$

$$F(1) = \frac{\lambda}{m\mu}\,(1+\gamma/\sigma)$$

$$F'(z)\Big|_{z=1} = \frac{\gamma\lambda^2}{2m\mu}\,E(S^2).$$

For the busy period of the update queue we have the standard results

$$1/\sigma = \frac{1}{\delta-\gamma}$$

$$E(S^2) = \frac{2\delta}{(\delta-\gamma)^3}.$$

Thus

$$(29)\quad \bar{n} = \frac{\lambda\gamma}{(\delta-\gamma)^2} + \frac{\frac{\lambda\delta}{m\mu}\left[1 + \frac{\lambda\gamma}{(\delta-\gamma)^2}\right] + \delta \sum_{j=1}^{m-1} \frac{m-j}{m}\, jp(j,0)}{\delta(1-\lambda/m\mu) - \gamma}$$

V. Conclusions

Our basic objective, reached in the last section, has been the analysis necessary for supporting a study of optimal file replication in distributed data bases, using more general measures than heretofore. With parameter values corresponding to specific systems, it can be seen that (27), (28) and (29) provide for an easily implemented, efficient computation requiring $O(m^2)$ time and $O(m)$ space.

In a numerical study of the optimization problem the key structural parameter to be specified is the average update time $1/\delta$. A basic assumption made in [1] is $1/\delta = m$, which applies when the m file replicas have to be updated in strict sequence, each according to a (normalized) unit exponential distribution.

However, in general the time to update the m file replicas will not be such a simple function of m. Parallelism may be possible in "broadcasting" updates to distributed data bases, thus reducing update times. On the other hand, coherence control mechanisms designed to synchronize the times when updated data bases become effective will tend to increase update times (see [2]). Depending on which effects dominate, the growth of $1/\delta$ with m may therefore be either faster or slower than linear.

To provide an idea of what can be gathered from numerical results, Figure 1 illustrates performance as a function of m, in terms of the measure $W = \bar{n}/\lambda$, assuming that $1/\delta = m$. (Note that log W is graphed.) Note that departures from the minimum expected read times are worse for underestimates than for equal overestimates of the optima. Also, for λ sufficiently small m = 1 is optimum, and as λ approaches saturation, optimum values of m will be so large that architectural issues other than those we have discussed may limit how many replicas can be provided.

In Figure 2 we illustrate the comparison of throughput and expected-read-time optimizations. Remarkable in this figure is how the optimal values of m depend on the performance measure. As can be seen the value of m maximizing read throughput exceeds the corresponding value minimizing expected read times. Moreover, W is a more rapidly varying function of m.

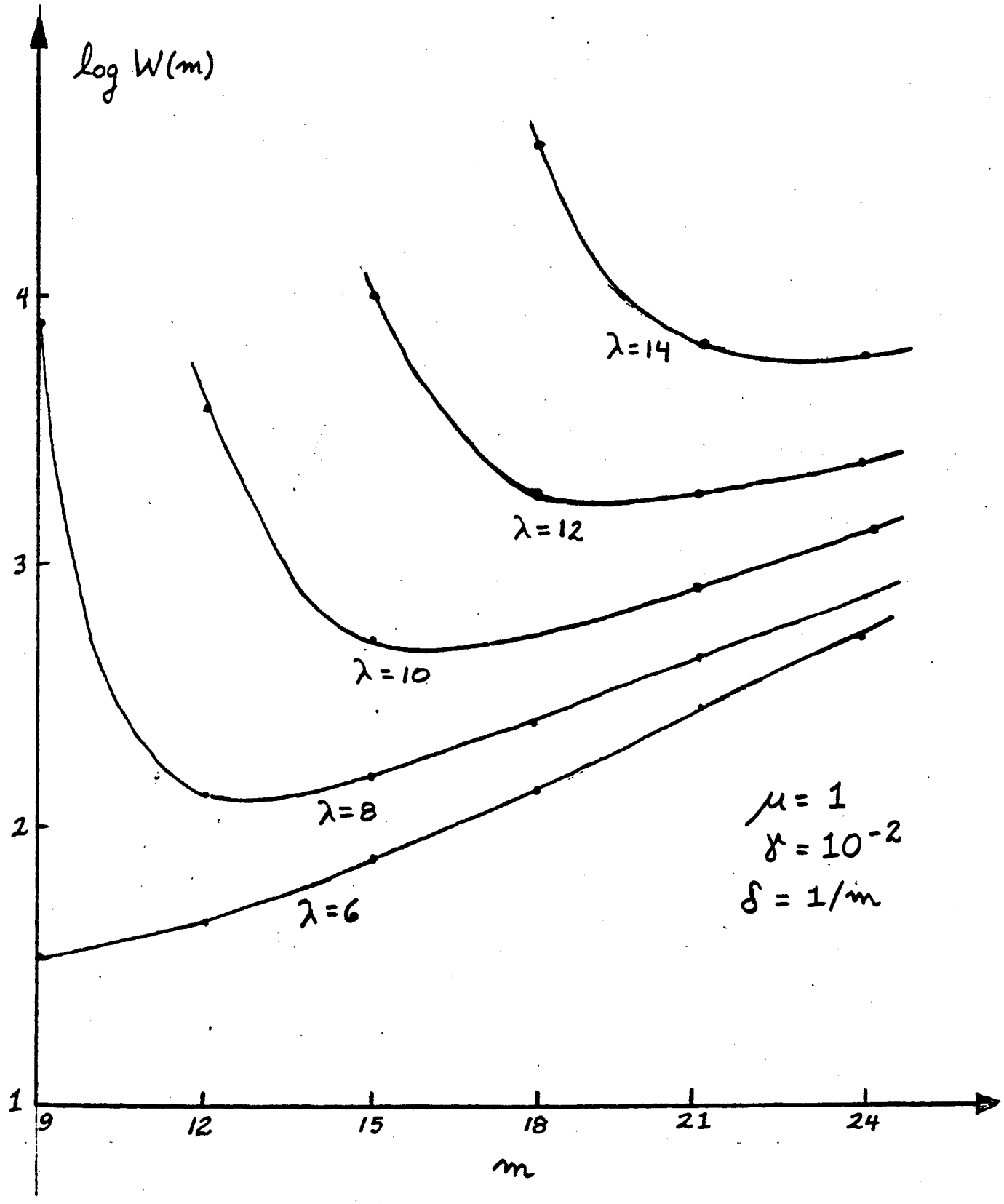

Figure 1. Logarithm of average READ response time vs. m.

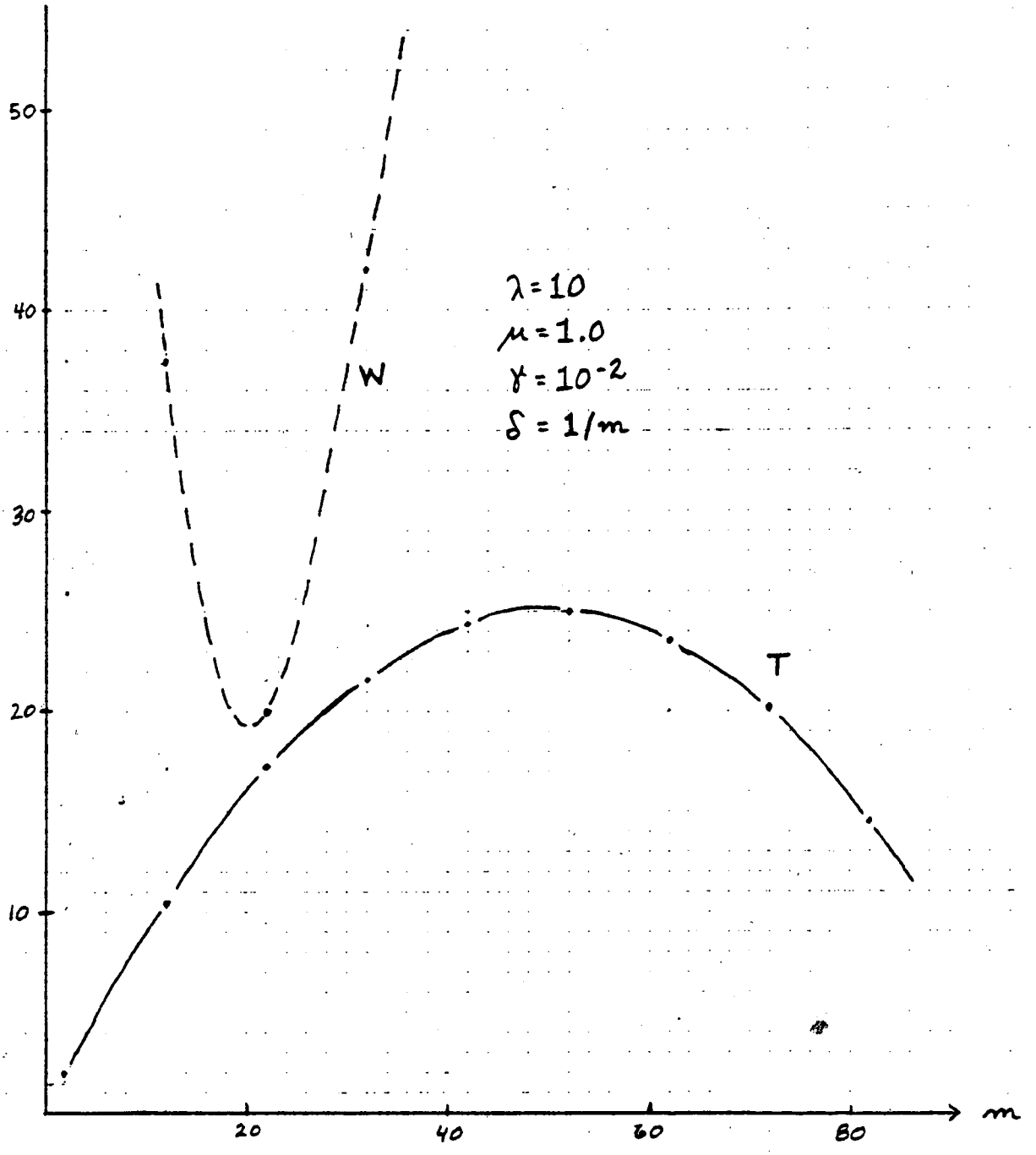

Figure 2. Illustrating the Optimization Problem

REFERENCES

1. Coffman, E. G., Jr., E. Gelenbe, H. O. Pollak and R. C. Wood, "An Analysis of Parallel-Read Sequential Write Systems", Performance Evaluation, Vol. 1, No. 1, Jan. 1981, pp. 63-70.
2. Coffman, E. G., Jr., Erol Gelenbe and Brigitte Plateau, "Optimization of the Number of Copies in a Distributed Data Base", IEEE Trans. on Software Engineering, Vol. SE-7, No. 1, Jan. 1981, pp. 78-84.

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique