FU JEN STUDIES

SCIENCE AND ENGINEERING

NO. 34, DEC. 2000

CONTENTS

輔仁學誌—理工類

中華民國八十九年十二月　　第三十四期

目　錄

Rotation-Based Interpolation Lattice Filters for Interference Suppression in Spread Spectrum Communication Systems

Jenq-Nan Lee
Department of Computer Science
National Taiwan Ocean University
Keelung, Taiwan 202, R.O.C.

Jenq-Tay Yuan
Department of Electronic Engineering
Fu Jen Catholic University
Taipei, Taiwan 242, R.O.C.

Abstract

In this paper we employ a rotation-based QR-decomposition-based least squares lattice (QRD-LSL) interpolation filter to overcome the difficulties associated with the least-mean-square (LMS) algorithm in the suppression of the narrowband interferences in direct-sequence (DS) spread spectrum systems. Computer simulations compare the results of using the QRD-LSL interpolation algorithm with the LMS algorithm in terms of SNR improvement and rate of convergence. The results show that the LMS algorithm exhibits extremely slow convergence speed and the QRD-LSL algorithm that has the merits of lattice structure displays excellent results in terms of rate of convergence as well as SNR improvement.

Key Words: Narrow-band interference suppression, QR-decomposition, least squares lattice, spread spectrum communication systems

I. INTRODUCTION

The digital whitening technique has been used to enhance the spread spectrum communications performance in the presence of narrowband jamming and interference. This is due to the fact that the interference is generally assumed to be nonwhite and is therefore can be predicted by use of whitening techniques that can be implemented as a transversal filter (or tapped-delay-line realizations) [1] - [11]. On the other hand, both the spread spectrum signal and the thermal noise are wide-band processes and hence they cannot be predicted (or interpolated) accurately by using the neighboring samples. A one-sided prediction filter (or two-sided interpolation filter) is thus used to suppress the interference in order to improve the performance of a direct sequence (DS) spread spectrum. It has been shown in [3], [4], [12], and [13] that better interference rejection may be obtained by using the interpolation filter rather than the prediction filter with the same number of taps. This is due to the fact that the former makes better use of the correlation between the nearest neighboring samples than the latter.

Many articles have been published on the use of prediction filter employing the lattice structure for suppression of narrowband interference [14] - [20]. The lattice structure is known to have the advantages of modularity, robustness, and better numerical conditioning compared to the transversal structure [21] - [23]. Simulation results and experimental results in [14] - [20] indicate that the lattice structure has great potential in narrowband interference suppression especially in complex jamming environments (e. g., multiple jammers). However, to our knowledge, previous references to the use of lattice filters for suppression of narrowband interference have been primarily concerned with one-sided prediction only. For two-sided interpolation, the literature has so far only considered the LMS algorithms. However, the LMS algorithm exhibits an extremely slow convergence rate. Consequently, the two-sided interpolation using the LMS algorithm may be unable to effectively suppress narrowband interference with a fast time-variant frequency. Even for one-sided prediction lattice filters considered in the available literature, except for [20] in which least squares lattice (*LSL*) has been considered, almost all the previous publications

use the *gradient adaptive lattice* (*GAL*). The convergence behavior of the GAL algorithm is more rapid than that of the LMS algorithm, but inferior to that of exact recursive LSL algorithms [21].

This paper proposes the use of the order-recursive *QR-decomposition-based least squares lattice* (*QRD-LSL*) *interpolation* filters developed in [24] for suppression of narrowband interference in the DS spread spectrum system. As well known, the QRD-LSL filter is the most important order-recursive adaptive filter since it represents the most fundamental form of an order-recursive adaptive filter. It is also well known that the performance of the QRD-LSL algorithm in a limited-precision environment is always superior to that of recursive LSL and recursive least-squares (RLS) algorithms due to the fact that the QRD-LSL algorithm is a "*square-root*" type algorithm and consequently, this results in a significantly reduced dynamic range of data [21]. The QRD-LSL algorithm combines highly desirable features of recursive least-squares estimation, QR-decomposition, and lattice structure, thus offering fast rate of convergence, good numerical properties, low sensitivity to round off error and parameter perturbation, high level of computational efficiency, and order-recursive property. In this work, the performance of the QRD-LSL interpolation filter in rejecting narrowband interference is evaluated when the narrowband interference is modeled as a single tone, multiple tones, and a second-order autoregressive (AR) process respectively. Simulation results demonstrate that the QRD-LSL interpolation filters can substantially outperform the conventional prediction and interpolation filters in terms of rate of convergence and SNR improvement.

II. SYSTEM MODEL

The low-pass equivalent of a DS spread spectrum modulation waveform is given by [10]

$$m(t) = \sum_{k=0}^{L-1} c_k q(t - k\tau_c) \tag{1}$$

where L is the number of pseudonoise (PN) chips per bit, τ_c is the chip interval, $\{c_k\}$ is the PN chip sequence used to spread the transmitted signal, and $q(t)$ is a rectangular pulse of duration τ_c. The transmitted signal is expressed as

$$s(t) = \sum_k b_k m(t - kT_b) \tag{2}$$

where $\{b_k\}$ is the binary information sequence and $T_b = L\tau_c$ is the bit duration. The received signal is then of the form

$$z(t) = s(t) + w(t) + i(t) \tag{3}$$

where $w(t)$ is wideband Gaussian noise, and $i(t)$ is narrowband interference. After the received signal has been processed by a matched filter and sampled at the chip rate $\frac{1}{\tau_c}$, the received signal becomes

$$z(n) = s(n) + w(n) + i(n) \tag{4}$$

where $\{s(n)\}$, $\{w(n)\}$, and $\{i(n)\}$ are the discrete-time sequences due to $\{s(t)\}$, $\{w(t)\}$, and $\{i(t)\}$, respectively and $\{s(n)\}$, $\{w(n)\}$, and $\{i(n)\}$ are assumed to be mutually independent. If the PN sequence is assumed to be truly random, the sequence $\{s(n)\}$ can be considered to be a sequence of independent and identically distributed (i. i. d.) binary random variables taking on values +1 or -1 with equal probability.

III. QRD-LSL INTERPOLATION FILTERS

A. Least-Squares Interpolation Filters of Order (p, f)

An estimate of the interference $i(n)$ can be formed from the received signal $z(n)$ in (4). When using the (p,f) thorder linear interpolation, we linearly estimate the present interference $i(i)$ from p past and f future neighboring received signal samples as shown in Fig. 1, viz.,

$$\hat{i}_{p,f}(i) \cong \hat{z}_{p,f}(i) = -\sum_{\substack{k=-p \\ k\neq 0}}^{f} b_{(p,f),k}(n-f)z(i+k), 1-f \leq i \leq n-f \tag{5}$$

where $b_{(p,f),k}(n\text{-}f)$ is the interpolation coefficient at time n which remains fixed during the observation interval $1\text{-}f \leq i \leq n\text{-}f$ with the prewindowing condition on the received signal samples. The reason that $\hat{i}_{p,f}(i) \cong \hat{z}_{p,f}(i)$ is because the spread spectrum signal $s(i)$ and the thermal noise $w(i)$ are wideband processes and hence

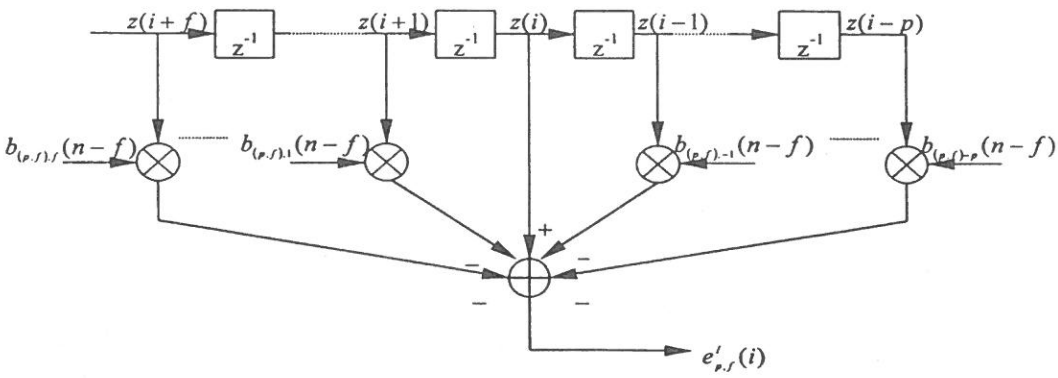

Fig. 1. $(p,\ f)^{\text{th}}$ **order linear interpolation transversal filter**

they cannot be interpolated accurately using the neighboring samples, whereas the interference $i(i)$ is generally assumed to be nonwhite and therefore can be accurately interpolated. The length of the signal, n , is variable. The order, $N = p + f$. By using (5), the $(p,f)^{th}$ order interpolation error at each time unit can be defined as

$$e^I_{p,f}(i) = \hat{s}(i) = z(i) - \hat{i}_{p,f}(i) \cong z(i) - \hat{z}_{p,f}(i)$$
$$= z(i) + \sum_{\substack{k=-p \\ k\neq 0}}^{f} b_{(p,f),k}(n-f)z(i+k), 1-f \leq i \leq n-f \qquad (6)$$

Herein, we refer to any N^{th} order interpolation filter that operates on the present received signal sample as well as p past and f future received signal samples to produce the $(p,f)^{th}$ order interpolation error at its output as a $(p,f)^{th}$ order interpolation filter where $N = p + f$ is assumed implicitly. If the most recent received signal sample used is $z(n)$, then (6) can be written in a matrix form for $1\text{-}f \leq i \leq n\text{-}f$ as in (7), or more compactly in (8).

$$\begin{bmatrix} e^I_{p,f}(1-f) \\ \vdots \\ e^I_{p,f}(-1) \\ e^I_{p,f}(0) \\ e^I_{p,f}(1) \\ \vdots \\ e^I_{p,f}(n-f) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ z(1) \\ \vdots \\ z(n-f-1) \\ z(n-f) \end{bmatrix} - \begin{bmatrix} z(1) & \cdots & 0 & 0 & \cdots & 0 \\ z(2) & & \vdots & \vdots & & \vdots \\ z(3) & \cdots & 0 & 0 & \cdots & 0 \\ & \cdots & z(1) & z(1) & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ & \cdots & & & \cdots & \\ \vdots & & \vdots & \vdots & & \vdots \\ z(n) & \cdots & z(n-f+1) & z(n-f-1) & \cdots & z(n-N) \end{bmatrix} \begin{bmatrix} -b_{(p,f),f}(n-f) \\ -b_{(p,f),f-1}(n-f) \\ \vdots \\ -b_{(p,f),1}(n-f) \\ -b_{(p,f),-1}(n-f) \\ \vdots \\ -b_{(p,f),-p}(n-f) \end{bmatrix} \qquad (7)$$

$$e^{I}_{p,f}(n-f) = z(n-f) - Z_{p,f}(n)b_{p,f}(n-f) \tag{8}$$

The least-squares solution to (8) can be determined by using the projection approach [22] - [23], [25] - [26]. The projection operator that projects vectors onto the subspace spanned by the columns of $Z_{p,f}(n)$ (i. e., of p past received signal samples and f future received signal samples) can be expressed by

$$Pz = Z_{p,f}(n)[Z^{T}_{p,f}(n)Z_{p,f}(n)]^{-1}Z^{T}_{p,f}(n) \tag{9}$$

Its orthogonal complement is denoted by $P^{\perp}_{Z} \equiv (I - P_{Z})$ where I represents the identity matrix. The interpolation error vector of (8) can therefore be determined by $e^{I}_{p,f}(n-f) = P^{\perp}_{Z}z(n-f)$.

B. Order-Recursive LSL Interpolation Filters and Exact Decoupling Property of LSL Interpolation Filters.

Order-recursive LSL interpolation filters that require only $O(N)$ operations have been developed in [24] by employing a modified version of linear forward and backward predictions referred to therein as the *intermediate forward and backward predictions*.

When a sequence of p past and f future signal samples is taken into consideration to estimate the present received signal sample $z(n\text{-}f)$, appropriate combinations of f delayed intermediate forward prediction errors and p delayed intermediate backward prediction errors, followed finally by the interpolation error, $e^{I}_{p,f}(n-f)$, form C^{f}_{N} sets of LS orthogonal bases [24]. Each of the C^{f}_{N} orthogonal basis sets can be shown to provide an orthogonal basis for $[\, z(n\text{-}N), \cdots, z(n\text{-}f\text{-}1), z(n\text{-}f), z(n\text{-}f+1), \cdots, z(n)\,]$, spanned by the present received signal sample, $z(n\text{-}f)$, p past received signal samples, and f future received signal samples. The LS orthogonality among all the elements within each of these orthogonal bases is referred to as the *exact decoupling property of the LSL interpolation filters*. The exact decoupling property of the LSL interpolation filters has been verified in [24]. Owing to the exact decoupling property, a $(p, f)^{th}$ order LSL interpolation filter automatically generates all N of the outputs that would be provided by N separate transversal interpolation filters of length 1, 2,

…, N , (i. e., a multiple order filter) where $N = p + f$. Higher order lattice filters are obtained from lower order ones by simply adding more stages, leaving the original stages unchanged. This modular structure permits dynamic assignment, and rapid automatic determination of the most effective filter length. Consequently, optimum removal of strong narrowband interferences of unknown or time-varying bandwidth may be achieved. The QRD-LSL interpolation filters are implemented by combining the exact decoupling property of the LSL interpolation filters with the well-conditioned and numerically stable QR-decomposition technique. The QRD-LSL interpolation algorithm has been summarized in [24]. Fig. 2 presents a signal-flow graph that shows the $(2, 2)^{th}$ order QRD-LSL interpolation filter using the sequence BFBF.

IV. SIMULATION RESULTS

Computer simulations have been carried out to compare the performances for single tone interference, multiple tone interference, and a second-order AR interference, respectively. The power of all the interferences is chosen such that the jammer-to-signal power ratio (J/S) is 20dB. Each learning curve was obtained by ensemble-averaging the squared value of [$s(n)$-$\hat{s}(n)$] over 200 independent trials of the experiment. Meanwhile, the power of the wideband Gaussian noise $w(n)$ is kept constant at $\sigma_w^2 = 0.01$. Fig. 3 compares the learning curves for a multiple tone sinusoidal interferer, $i(t) = A\cos(0.15t + \theta_1) + A\cos(0.4t + \theta_2) + A\cos(0.65t + \theta_3)$, where A is the amplitude of all sinusoidal interferences, and θ_1, θ_2, and θ_3 are random phases uniformly distributed from 0 to 2π. This figure illustrates that the rate of convergence of the QRD-LSL algorithm is much faster than that of the LMS algorithm in the presence of a multiple tone sinusoidal jammer. Table I calculates the average SNR improvement over last 500 data samples for both multiple tone and AR jammers. The latter jammer was obtained by passing white noise through a second-order IIR filter with two poles at 0.9557 + j0.1914 and 0.9557-j0.1914, i. e., $i(n) = 1.9114i(n-1) - 0.95i(n-2) + e(n)$ where $\{e(n)\}$ is white Gaussian noise. Fig. 4 shows results for a second-order AR jammer. The simulation results also show that the QRD-LSL interpolation filter has much faster convergence and smaller steady-state MSE compared with the LMS interpolation filter in the presence of the AR jammer.

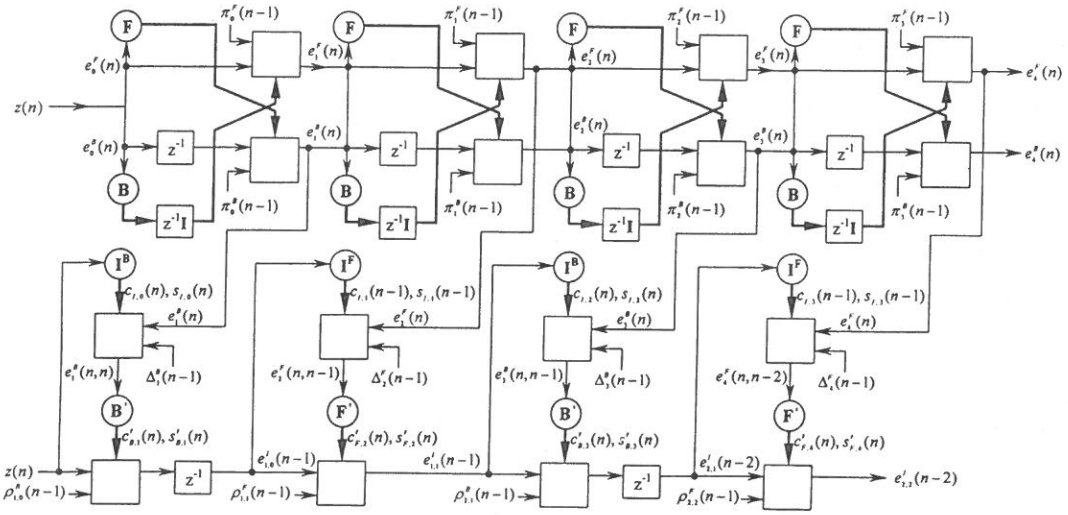

Fig. 2. Signal-flow graph $(2, 2)^{th}$ order QRD-LSL interpolation filter using sequence BFBF [24]

V. CONCLUSIONS

This paper presents an order-recursive rotation-based QRD-LSL interpolation filter as a solution to the narrowband interference rejection in a DS spread spectrum system. The proposed N^{th} -order interpolation lattice filter automatically generates all N of the outputs that would be provided by N separate transversal interpolation filters. This feature can effectively suppress the narrow-band interference of time-varying bandwidth in DS spread spectrum systems. Simulation results indicate that the QRD-LSL interpolation filters can substantially outperform the conventional prediction and interpolation filters in terms of rate of convergence and SNR improvement for single tone, multiple tone, and AR interferences.

REFERENCES

(1) F. M. Hsu and A. A. Giordano, "Digital whitening techniques for improving spread-spectrum communications performance in the presence of narrow-band jamming and interference," IEEE Trans. Commun., vol. COM-26, pp. 209-216, Feb. 1978.

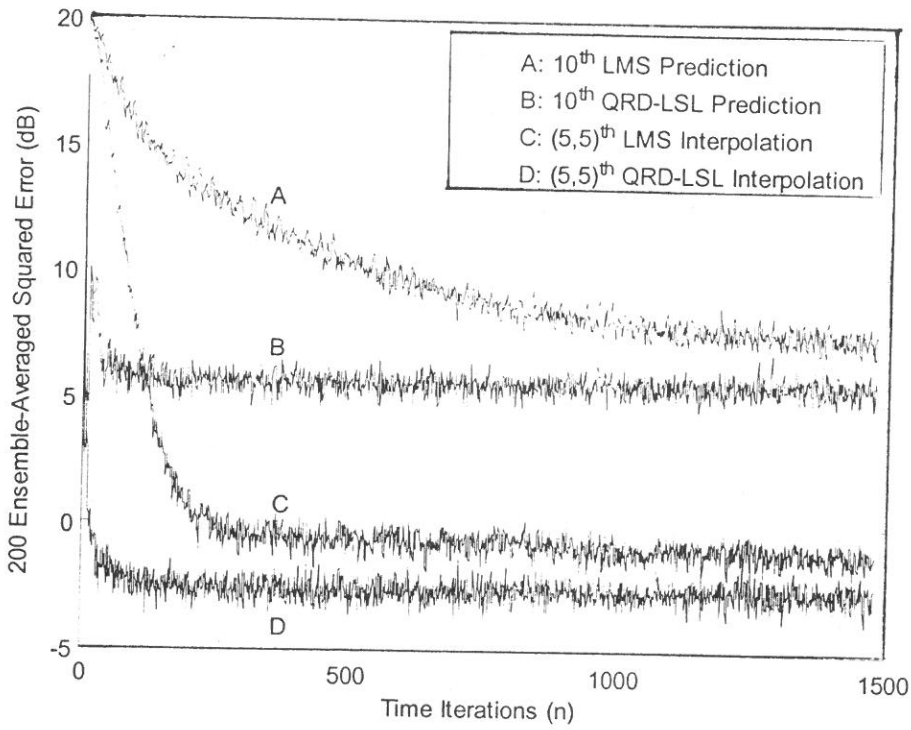

Fig. 3. Multiple tone jammers with random phase (forgetting factor $\lambda = 0.99$ and step-size parameter $\mu = 0.00003$)

Table I. SNR IMPROVEMENT FOR MULTIPLE-TONE AND AR JAMMERS.

	SNR improvement (dB)	
	Multiple-tone	AR
LMS prediction	12.28	14.78
QRD-LSL prediction	14.45	17.25
LMS interpolation	21.35	21.28
QRD-LSL interpolation	22.66	23.24

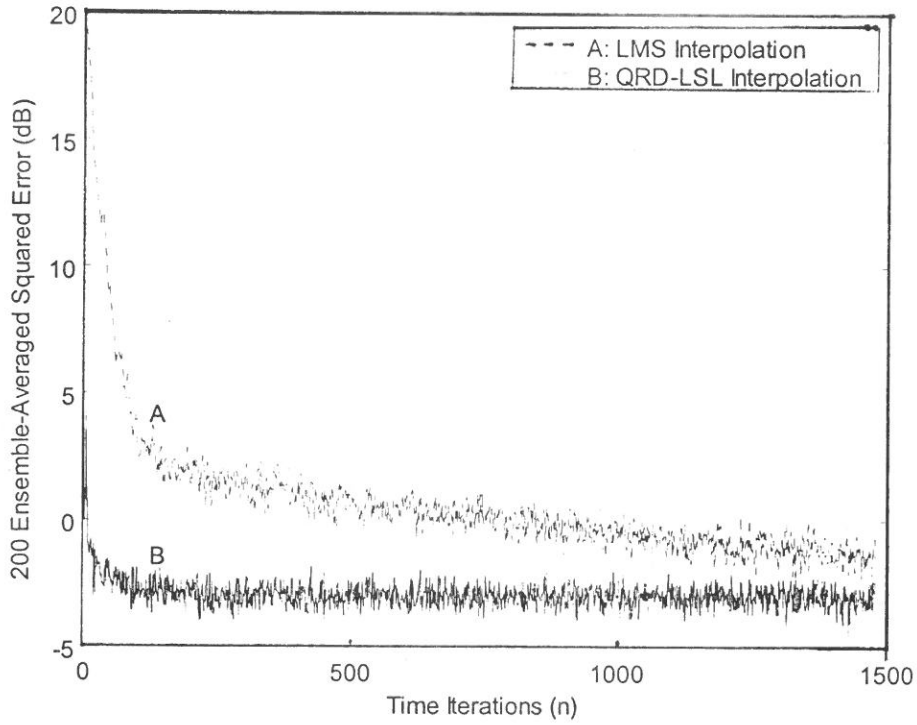

Fig. 4. Second-order AR jammer (forgetting factor $\lambda = 0.99$ and step-size parameter $\mu = 0.00003$)

(2) J. W. Ketchum and J. G. Proakis, "Adaptive algorithms for estimating and suppressing narrow-band interference in PN spread-spectrum systems," IEEE Trans. Commun., vol. COM-30, pp. 913-924, May 1982.

(3) L. Li, and L. B. Milstein, "Rejection of narrow-band interference in PN spread-spectrum systems using transversal filters," IEEE Trans. Commun., vol. COM-30, pp. 925-928, May 1982.

(4) L. Li, and L. B. Milstein, "Rejection of pulsed CW interference in PN spread-spectrum systems using complex adaptive filters," IEEE Trans. Commun., vol. COM-31, pp. 10-20, Jan. 1983.

(5) L. B. Milstein, "Interference rejection techniques in spread spectrum communications," Proceedings of the IEEE, vol. 76, no. 6, pp. 657-671, June 1988.

(6) E. Masry, "Closed-form analytical results for the rejection of narrow-band interference in PN spread-spectrum systems - Part I: Linear prediction filters,"

IEEE Trans. Commun., vol. COM-32, pp. 888-896, Aug. 1984.

(7) R. A. Iltis and L. B. Milstein, "Performance analysis of narrow-band interference rejection techniques in DS spread-spectrum systems," IEEE Trans. Commun., vol. COM-32, pp. 1169-1177, Nov. 1984.

(8) R. A. Iltis and L. B. Milstein, "An approximate statistical analysis of the Widrow LMS algorithm with application to narrow-band interference rejection," IEEE Trans. Commun., vol. COM-33, pp. 121-130, Feb. 1985.

(9) G. J. Saulnier, P. K. Das, and L. B. Milstein, "An adaptive digital suppression filter for direct-sequence spread-spectrum communications," IEEE J. Select. Areas Commun., vol. SAC-3, no. 5, September 1985.

(10) R. Vijayan and H. V. Poor, "Nonlinear techniques for interference suppression in spread-0spectrum systems," IEEE Trans. Commun., vol. COM-38, pp. 1060-1065, July 1990.

(11) W. R. Wu and F. F. Yu, "New nonlinear algorithms for estimating and suppressing narrowband interference in DS spread spectrum systems," IEEE Trans. Commun., vol. COM-44, pp. 508-515, April 1996.

(12) E. Masry, "Closed-form analytical results for the rejection of narrow-band interference in PN spread-spectrum systems - Part II: Linear interpolation filters," IEEE Trans. Commun., vol. COM-33, pp. 10-19, January 1985.

(13) L. A. Rusch and H. V. Poor, "Narrowband interference suppression in CDMA spread spectrum communications," IEEE Trans. Commun., vol. COM-42, pp. 1969-1979, February/March/April 1994.

(14) J. Guilford and P. Das, "The use of the adaptive lattice filter for narrowband jammer rejection in DS spread spectrum systems," Proc. IEEE Int. Conf. on Commun., pp. 822-826, June 22-26, 1985.

(15) G. J. Saulnier, K. Yum, and P. Das, "The suppression of tone jammers using adaptive lattice filtering," Proc. IEEE Int. Conf. on Commun., pp. 24. 4. 1-24. 4. 5, June 1987.

(16) Y. He, S. -F. Lei, P. Das, and G. J. Saulnier, "Suppression of narrowband jammers in a DS spread spectrum receiver using modified adaptive filtering technique," Proc. Globecom'88, vol. 1, pp. 540-545, 1988.

(17) J. R. Zeidler, K. C. Chew, R. A. Axford, and W. H. Ku, "Frequency tracking performance of adaptive lattice filters," Proc. 25th Asilomar Conf. on Signals, Systems and Computers, vol. 2, pp. 643-649, 1991.

(18) F. Takawira, "Adaptive lattice filters for narrowband interference rejection in DS spread spectrum systems," Proc. 1994 IEEE South Africa Symp. on Commun. and Signal Processing, pp. 1-5, 1994.

(19) G. J. Saulnier, W. A. Haskins, and P. Das, "Tone jammer suppression in a direct sequence spread spectrum receiver using adaptive lattice and transversal filters," Proc. MILCOM'87, pp. 123-127, 1987.

(20) W. S. Hodgkiss, Jr., and J. A. Presley, Jr., "Adaptive tracking of multiple sinusoids whose power levels are widely separated," IEEE Trans. Circuits and Systems, vol. CAS-28, no. 6, pp. 550-561, June 1981.

(21) S. Haykin, Adaptive Filter Theory, Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1996, 3rd ed.

(22) B. Friedlander, "Lattice Filters for Adaptive Processing," Proceedings of the IEEE, vol. 70, no. 8, pp. 829-867, August 1982.

(23) Peter Strobach, Linear Prediction Theory - A Mathematical Basis for Adaptive Systems, Springer - Verlag, 1990.

(24) J. -T. Yuan, "QR-decomposition-based least-squares lattice interpolators," IEEE Trans. Signal Processing, vol. 48, no. 1, pp. 70-79, Jan. 2000.

(25) B. Proat, B. Friedlander, and M. Morf, "Square root covariance ladder algorithms," IEEE Trans. Automat. Contr., vol. AC-27, pp. 813-829, Aug. 1982.

(26) D. T. L. Lee, M. Morf, and B. Friedlander, "Recursive least squares ladder estimation algorithms," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp. 627-641, June 1981.

received September 29, 2000
revised October 12, 2000
accepted November 2, 2000

旋轉式插值格式網狀濾波器用來做展頻通訊系統之窄頻干擾抑制

李 正 男
海洋大學資訊科學系

袁 正 泰
輔仁大學電子工程系

摘 要

本篇論文我們利用 QR 分解架構的格狀插值濾波器來抑制直接序列展頻通訊中的窄頻干擾，並與傳統的最小均方濾波器做比較。在電腦模擬方面，我們分別利用 QR 分解架構的格狀插值濾波器與最小均方濾波器，比較其收斂速率與訊號對干擾比的改善率。結果顯示，無論是在收斂速率或訊號對干擾比的改善率，我們所提出的 QR 分解架構格狀插值濾波器都較優於傳統的最小均方濾波器。

Estimators of Process Capability Index $C_{Np}(u, v)$ for Symmetric Processes with Known Center of Symmetry

Sy-Mien Chen
Department of Mathematics
Fu Jen Catholic University
Taipei, Taiwan 242, R.O.C.

Yu-Sheng Hsu
Department of Mathematics
National Central University
Chungli, Taiwan 320, R.O.C.

Abstract

Pearn and Chen (1997) proposed a class of capability indices $C_{Np}(u, v)$ which generalize the process capability indices $C_p(u, v)$ (Vännman (1995)) to accommodate cases where the underlying distributions may not be normal. The current indices $C_{Np}(u, v)$ are functions of quantiles which may not be known. In this article, we propose four estimators based on some quantile estimators which are given by Cohen et al. (1985) for symmetric processes when the center of symmetry is known. It is shown that these four estimators of $C_{Np}(u, v)$ are all asymptotically normally distributed, and are asymptotically equivalent. When we compare these four estimators with the estimator proposed by Pearn and Chen (1997), we found that our estimators are asymptotically more efficient.

Key Words: asymptotic distribution; specification limits.

INTRODUCTION

A process capability index is a numerical measure which provides information on

whether a production process is capable of producing items within the specification limits predetermined by the designer. Recent research and advances made in this subject are neatly summarized in Sullivan (1984), Kane (1986), Chan et al. (1988a, 1988b), Spiring (1991), Pearn et al. (1992), Kotz and Lovelace (1998), and Vännman (1995), among others. In addition to their typical applications in industry, process capability indices are particularly useful in chemical batch processes (Morris and Watson (1998)), rubber-edge production lines (Chen and Pearn (1997)) and academic fields for studying the classroom performance (Johnson and Smith (1997)).

Vännman (1995) presented a unified approach to capability indices research by introducing a class of capability indices $C_p(u, v)$ that generalizes the fourbasic indices, C_p, C_{pk}, C_{pm} and C_{pmk}. It has been shown that $C_p(u, v)$ are appropriate for normal processes, but not for non-normal process data. To deal with this difficulty, Pearn and Chen (1997) and Chen and Pearn (1997) generalized $C_p(u, v)$ to a class of capability indices $C_{Np}(u, v)$. However, the current indices $C_{Np}(u, v)$ are functions of quantiles which may not be known. Chen and Hsu (2000) proposed two estimators of the indices $C_{Np}(u, v)$ based on empirical and kernel quantile estimators. It was shown that estimators of $C_{Np}(u, v)$ they proposed and the one proposed by Pearn and Chen (1997) are all consistent, asymptotically normally distributed and asymptotically equivalent.

In this article, we restrict our discussion on symmetric processes when the center of symmetry is known. We propose four estimators based on some quantile estimators given by Cohen *et al*. (1985). It is shown that these four estimators of $C_{Np}(u, v)$ are all asymptotically normally distributed and are asymptotically equivalent. When we compare these four estimators with the estimator proposed by Pearn and Chen (1997), we found that our estimators are asymptotically more efficient. In section two, the definition of $C_{Np}(u, v)$ is given. In section three, four estimators of the the family of indices are defined, and the asymptotic distributions of the proposed estimators are derived. The asymptotic distribution of the estimator proposed by Pearn and Chen (1997) is also derived. Section 4 provides the conclusions.

2. Process Capability Indices $C_{Np}(u, v)$

Let $X_1, \cdots, X_n$ be a sample of measurements from a process which has

distribution F with the λ-th quantile F_λ, mean μ and variance σ^2. Let L and U be the lower and the upper specification limits of a product characteristic, respectively. Denote by $d=\frac{U-L}{2}$ the half length of the specification limits, $m=\frac{U+L}{2}$ the mid-point of the specification interval, T the target value. A class of process capability indices proposed by Pearn and Chen (1997) is defined as follows:

$$C_{Np}(u,v)=\frac{d-u\mid F_{.50}-m\mid}{3\sqrt{(\frac{F_{.99865}-F_{.00135}}{6})^2+v(F_{.50}-T)^2}}, \tag{1}$$

where u, $v\geq 0$.

When the process has symmetric tolerances (i.e. $T=m$) and assume the underlying distribution is normal, this new class of indices defined also generalizes the four basic indices, C_p, C_{pk}, C_{pm} and C_{pmk}.

3. Asymptotic Distributions of Estimators of $C_{Np}(u,v)$

3.1 Estimator based on Cohen et al.

It has been pointed out by Cohen et al. (1985) that quality control data often times entails repeated measurements with instruments. Frequently these data collected regularly (weekly say) over many years reflect symmetry. For this reason, Cohen et al. (1985) proposed some quantile estimators of a symmetric distribution and derived some important statistical properties. Let $X_1, \cdots, X_n$ be a random sample of size n from a distribution with c.d.f. F. Let ξ_λ denote the λth quantile, which is uniquely defined under the conditions imposed on F. Assume F is symmetric with θ as the center of symmetry. In the case of known center, let $\theta=\theta_0$. The order statistics are $X_{(1)}\leq X_{(2)}\leq\cdots\leq X_{(n)}$, and $Y_{(1)}\leq Y_{(2)}\leq\cdots\leq Y_{(n)}$ represent the ordered values of $Y_j=|X_j-\theta_0|$, $j=1, 2, \cdots, n$. Cohen et al. (1985) proposed the following four estimators of the quantile:

1. $T_1(\lambda)=(T^+(\lambda)-T^-(1-\lambda))/2$, where $T^+(\lambda)=(T-\theta_0\vee 0)$, $T^-=(T-\theta_0\wedge 0)$, and $T(\lambda)=\alpha X_{([n\lambda]+1)}+(1-\alpha)X_{([n\lambda])}$, *for* $\alpha=n\lambda-[n\lambda]+1/2$ and $[a]$ is the largest integer less thanor equal to a.

2. (Flipped-Over estimator) $T_2(\lambda)=\alpha\, Y_{([n(2\lambda-1)]+1)}+(1-\alpha)Y_{([n(2\lambda-1)])}$, where $\alpha=n(2\lambda-1)-[n(2\lambda-1)]+1/2$.
3. $T_3(\lambda)$ is the Rao-Blackwellized version of $T_1^+(\lambda)$ and it depends on the Y sample.
4. $T_4(\lambda)$ is a nonparametric Bayes estimator using a Dirichlet prior and a symmetric linear loss function.

By assuming that F is twice differentiable in a neighborhood of ξ_λ; that is $F'(x)=f(x)$ is bounded away from 0 in the neighborhood and that F'' is bounded in the neighborhood, $T_i(\lambda)$ are asymptotically equivalent and are all asymptotically normal with mean ξ_λ and with variance $\lambda(1-2\lambda)/2nf^2(\xi_\lambda)$ if $0<\lambda<0.5$; with asymptotic variance $(2\lambda-1)(1-\lambda)/2nf^2(\xi_\lambda)$ if $0.5<\lambda<1$, for $i=1, 2, 3, 4$. And T_i are asymptotically optimal in the sense of minimum asymptotic variance in the class of regular estimators, $i=1, 2, 3, 4$. [Cohen et al. (1985)]

For a symmetric distribution with center of symmetry θ_0, $F_{0.135}=2\theta_0-F_{99.865}$. Therefore, the indices can be reduced to

$$C_{Np}(u,v)=\frac{d-u\mid\theta_0-m\mid}{3\sqrt{(\frac{F_{99.865}-F_{0.135}}{6})^2+v(\theta_0-T)^2}}=\frac{d-u\mid\theta_0-m\mid}{\sqrt{(F_{99.865}-\theta_0)^2+9v(\theta_0-T)^2}}.$$

Consider the following estimator of $C_{Np}(u,v)$

$$\hat{C}_{Np}(u,v)=\frac{d-u\mid\theta_0-m\mid}{3\sqrt{(\frac{\hat{F}_{99.865}-\hat{F}_{0.135}}{6})^2+v(\theta_0-T)^2}}=\frac{d-u\mid\theta_0-m\mid}{\sqrt{(\hat{F}_{99.865}-\theta_0)^2+9v(\theta_0-T)^2}}$$

where $\hat{F}_{99.865}$ is a quantile estimator of $F_{99.865}$.

Denote $T_i=\hat{F}_{99.865}$ be the four estimators of $F_{99.865}$ proposed by Cohen et al., $i=1, 2, 3, 4$.

Theorem 1:

Let

$$\hat{C}_{Np}(u,v) = \frac{d - u \mid \theta_0 - m \mid}{\sqrt{(T_i - \theta_0)^2 + 9v(\theta_0 - T)^2}}, i = 1,2,3,4.$$

Then

$$\sqrt{n}(\hat{C}_{Np}^{(i)}(u,v) - C_{Np}(u,v)) \xrightarrow{L} N(0,\phi^2\sigma^2), i = 1,2,3,4,$$

where $\sigma^2 = (2\lambda - 1)(1-\lambda)/2f^2(\xi_\lambda)$, $\lambda = .99865$, and

$$\phi = -\frac{(d - u \mid \theta_0 - m \mid)(F_{99.865} - \theta_0)}{\{(F_{99.865} - \theta_0)^2 + 9v(\theta_0 - T)^2\}^{3/2}}.$$

Proof:

Let $h(t) = \dfrac{d - u(\theta_0 - m)}{\sqrt{(t-\theta_0)^2 + 9v(\theta_0 - T)^2}}$, then

$$h'(t) = -\frac{(d - u|\theta_0 - m|)(t - \theta_0)}{[(t-\theta_0)^2 + 9v(\theta_0 - T)^2]^{3/2}}.$$

Implies that $h'(F_{99.865}) \neq 0$. Denote $\phi = h'(F_{99.865})$, then by Theorem on page 24 in Serfling,

$$\sqrt{n}(\hat{C}_{Np}^{(i)}(u,v) - C_{Np}(u,v)) \xrightarrow{L} N(0,\phi^2\sigma^2), i = 1,2,3,4,$$

where $\sigma^2 = (2\lambda - 1)(1-\lambda)/2f^2(\xi_\lambda)$, $\lambda = .\ 99865$, and

$$\phi = -\frac{(d - u \mid \theta_0 - m \mid)(F_{99.865} - \theta_0)}{\{(F_{99.865} - \theta_0)^2 + 9v(\theta_0 - T)^2\}^{3/2}}. \square$$

From above theorem, by comparing the asymptotic variance of the four estimators $\hat{C}_{Np}^{(i)}(u,v)$ we see that these four estimators are asymptotically equivalent.

3.2 Estimator based on Pearn and Chen

For unknown quantiles, Pearn and Chen (1997) applied the method in Clement (1989) to find the quantile estimators for the generalizations $C_{Np}(u,v)$, and obtained a superstructure for the estimators as follows:

$$\breve{C}_{Np}(u,v) = \frac{d - u \mid \breve{\xi}_{.50} - m \mid}{3\sqrt{(\frac{\breve{\xi}_{.99865} - \breve{\xi}_{.00135}}{6})^2 + v(\breve{\xi}_{.50} - T)^2}},$$

where

$$\check{\xi}_{.00135} = X_{([R_1])} + (R_1 - [R_1])(X_{([R_1]+1)} - X_{([R_1])}),$$

$$\check{\xi}_{.5} = X_{([R_2])} + (R_2 - [R_2])(X_{([R_2]+1)} - X_{([R_2])}),$$

$$\check{\xi}_{.99865} = X_{([R_3])} + (R_3 - [R_3])(X_{([R_3]+1)} - X_{([R_3])});$$

$$R_1 = .00135n + .99865, \quad R_2 = .5n + .5; \quad R_3 = .99865n + .00135.$$

In this setting, the notation $[R]$ is defined as the greatest integer less than or equal to the number R, and $X_{(i)}$ is the i-th order statistics of $X_1, X_2, \cdots\cdots, X_n$.

It was shown by Chen and Hsu (2000) that

$$\sqrt{n}(\check{C}_{Np}(u,v) - C_{Np}(u,v)) \to \begin{cases} N(0, D\Sigma D'), if\ \xi_{.5} \neq m, \\ -\dfrac{1}{3G} \mid W_1 \mid - \dfrac{d}{3G^2} W_2, if\ \xi_{.5} = m, \end{cases}$$

where

$$(W_1, W_2) \sim N((0,0), \Phi\Sigma\Phi'),$$

$$\Sigma = \begin{bmatrix} \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_3 \end{bmatrix},$$

and

$$\sigma_1 = \frac{(.5)^2}{f^2(\xi_{.5})},$$

$$\sigma_2 = \frac{\xi_{.99865} - \xi_{.00135}}{18}\left[\frac{(.5)(.00135)}{f(\xi_{.5})f(\xi_{.99865})} - \frac{(.00135)(.5)}{f(\xi_{.00135})f(\xi_{.5})}\right],$$

$$\sigma_3 =$$

$$\left(\frac{\xi_{.99865} - \xi_{.00135}}{18}\right)^2\left[\frac{0.0135(.99865)}{f^2(\xi_{.00135})} + \frac{.99865(.00135)}{f^2(\xi_{.99865})} - 2\frac{(.00135)^2}{f(\xi_{.00135})f(\xi_{.99865})}\right].$$

$$\Phi' = \begin{bmatrix} u & \dfrac{v(\xi_{.5} - T)}{G} \\ 0 & \dfrac{1}{2G} \end{bmatrix}.$$

$$G = G(\xi_{.00135}, \xi_{.5}, \xi_{.99865}, v, T) = \left\{\left(\frac{\xi_{.99865} - \xi_{.00135}}{6}\right)^2 + v(\xi_{.5} - T)^2\right\}^{1/2},$$

$$D = (sgn(m - \xi_{.5})\frac{u}{3G} - \frac{dv(\xi_{.5} - T)H}{3G^3}, -\frac{dH}{6G^3}).$$

For a symmetric distribution with center of symmetry θ_0, the estimator proposed by Pearn and Chen (1997) is reduced to the following form:

$$\hat{C}^*_{Np}(u,v) = \frac{d - u \mid \theta_0 - m \mid}{\sqrt[3]{(\frac{\hat{F}_{99.865} - \hat{F}_{0.135}}{6})^2 + v(\theta_0 - T)^2}}$$

$$= \frac{d - u \mid \theta_0 - m \mid}{\sqrt{(\hat{F}^*_{99.865} - \theta_0)^2 + 9v(\theta_0 - T)^2}}$$

where $\hat{F}^*_{99.865} = X_{([R_3])} + (R_3 - [R_3])(X_{([R_3]+1)} - X_{([R_3])})$;

Theorem 2:

$$\sqrt{n}(\hat{C}^*_{Np}(u,v) - C_{Np}(u,v)) \xrightarrow{L} N(0, \phi^2\sigma^2_*), i = 1,2,3,4$$

where $\sigma^2_* = \frac{\lambda(1-\lambda)}{f^2(\xi_\lambda)}$, and ϕ is defined as in theorem 1 above.

Proof:

By Bahadur representation (page 93, Serfling (1980)),

$$\sqrt{n}\{X_{([R_n]+1)} - X_{([R_n])}\} = O_p(n^{-1/5}).$$

It follows that

$$\sqrt{n}(R_n - [R_n])\{X_{([R_n]+1)} - X_{([R_n])}\} = o_p(1).$$

Therefore, $\sqrt{n}\{\breve{\xi}_\lambda - \hat{\xi}_\lambda\} = o_p(1)$, for all $\lambda \in (0, 1)$. Hence, the theorem holds by the same argument in the proof of Theorem 1. □

Theorem 3:

$\hat{C}^{(i)}_{Np}$ (u, v) are asymptotically more efficient than $\hat{C}^*_{Np}$ (u, v).

Proof:

The result follows by Theorem 1 and Theorem 2. □

4. Conclusions

By assuming that the process is symmetric and the center of symmetry is known, we proposed four estimators $\hat{C}_{Np}^{(i)}(u,v)$ of the index $C_{Np}(u,v)$ based on Cohen et al. (1985). And it is shown that the four estimators are all asymptotically equvalent, and are asymptotically normally distributed. When we compare the asymptotic variance of the estimator based on Pearn and Chen (1997), $\hat{C}_{Np}^{*}(u,v)$, with that of the estimators $\hat{C}_{Np}^{(i)}(u,v)$, it shows that $\hat{C}_{Np}^{(i)}(u,v)$ are asymptotically more efficient than $\hat{C}_{Np}^{*}(u,v)$. i.e. Clements' method is not proper in this case.

Acknowledgements

The research for the first author was partially supported by National Science Council Grants, Republic of China, NSC 88－2118－M030－001.

BIBLIOGRAPHY

(1) Chan, L. K., Cheng, S. W. and Spring, F. A. (1988a). "The Robustness of the Process Capability Index, C_p, to Departures from Normality," *Statistical Theory and Data Analysis II*, Elsevier Science Publishers B. V., North-Holland, 223－239.

(2) Chan, L. K., Cheng, S. W. and Spring, F. A. (1988b). "New Measure of Process Capability Indices: C_{pm}," *Journal of Quality Technology*, 20, 162－175.

(3) Chen, K. S. and Pearn, W. L. (1997). "An Application of Non-normal Process Capability Indices," *Quality and Reliability Engineering International*, 13, 355－360.

(4) Chen, Sy-Mien and Hsu, Yu-Sheng (2000). "Asymptotic Analysis of Estimators for $C_{Np}(u,v)$ Based on Quantile Estimators," Accepted by *Journal of Nonparametric Statistics*.

(5) Cohen A., Lo S. H. and Singh K. (1985). "Estimating a Quantile of a

Symmetric Distribution," *Ann. Statist.*, Vol. 13, No. 3, 1114－1128.

(6) Johnson F. C. and Smith G. S. (1997). " Process Stability and Process Capability in Public Schools," *Quality Engineering*, 9 (3), 503－520.

(7) Kane, V. E. (1986). "Process Capability Indices," *Journal of Quality Technology*, 18, 41－52.

(8) Kotz, S. and Lovelace C. R. (1998). *Process Capability Indices*, Arnold.

(9) Morris R. A. and Watson E. F. (1998). "Determining Process Capability in a Chemical Batch Process," *Quality Engineering*, 10 (2), 389－396.

(10) Pearn, W. L., Kotz, S. and Johnson, N. L. (1992). "Distribution and Inferential Properties of Process Capability Indices," *Journal of Quality Technology*, 24, 216－231.

(11) Pearn, W., Chen, K. (1997). "Capability Indices for Non-Normal Distribution with an Application in Electrolytic Capacitor Manufacturing," *Micro-elctronics and Reliability*, Vol 37, No. 12, 1853 - 1858.

(12) Serfling, R. J. (1980). "Approximation Theorems of Mathematical Statistics," John Wiley and Sons, New York, 1-125.

(13) Spiring, F. A. (1991). "The C_{pm} Index," *Quality Progress*, 24 (2), 57－61.

(14) Sullivan L. P. (1984). "Reducing Variability: A New approach to Quality," *Quality Progress*, 17, 15－21.

(15) Vännman, K. (1995). "A Unified Approach to Capability Indices," *Statistica Sinica*, 5, 805－820.

received September 14, 2000
revised October 9, 2000
accepted October 27, 2000

對稱點已知之製程其 C_{NP}（u, v）估計式之研究

陳思勉
輔仁大學數學系

許玉生
中央大學數學系

摘　要

Pearn and Chen（1997）提出一個新的製程能力指標族 $C_{Np}(u,v)$，該製程能力指標族解除了由 Vännman（1995）所提出製程能力指標族 $C_p(u,v)$ 中有關常態分配之限制。此新的製程能力指標族 $C_{Np}(u,v)$ 爲分位點之函數，而分位點通常是未知的。本文根據 Cohen et al.（1985）所提分位點估計式提出四個 $C_{Np}(u,v)$ 之估計式，並證明此四個估計式在大樣本之下均服從常態分配且均等價。當我們把此四個估計式與 Pearn and Chen（1997）所提之估計式做比較時，發現我們所提之四個 $C_{Np}(u,v)$ 之估計式在大樣本之下較有效。

A Modified M/M/1/N Queueing Model for the Data Buffering of a Dual Channel Microcomputer-based Oscilloscope System

Ying-Wen Bai and Hong-Gi Wei
Department of Electronic Engineering
Fu Jen Catholic University
Taipei, Taiwan 242, R.O.C.

Abstract

Traditionally, oscilloscopes were used to show analog voltage waveforms of electronic circuits. Currently, analog voltage waveforms can be digitized for easy processing, storing and transmitting of the measured signal by computers and their networks. In this paper, we propose a modified M/M/1/N queueing model for the analysis of characteristics of sampled data arrival and departure from the data buffer in a digital oscilloscope implemented by a microcomputer system with interface circuits. The primary results from this model and the measurement of the experimental system show that the smaller the ratio between the sampled data arrival rate and departure rate for the data buffer mechanism, the smaller is the loss waveform interval for the dual measured signals.

Key Words: Microcomputer, Oscilloscope, Queueing Model, Data Buffer, Loss Waveform Interval

1. Introduction

An oscilloscope is a measuring tool for analog waveforms and is used to verify the design of electronic circuits. Over the past few decades, numerous researchers have been working in different aspects in order to improve the functions of oscilloscopes.

One of them is to digitize analog waveforms and to store the sampled signal for further processing by computers. However, due to the speed limitation of the sampling circuits, data buffers, and I/O channels of a microcomputer system, the bandwidth of a microcomputer-based oscilloscope can be very limited. Fortunately, because of the progress of VLSI technology, the sampling rate of sampling circuits can reach beyond several GHz. In addition, the speed of data buffers has improved. It can reach a couple of hundred MHz. However, the speed of an I/O channel of a microcomputer system can still be a bottleneck from a continuous data flow point of view, although the speed of I/O channels can reach several hundred KHz [1-5].

From the previous observation of the speed comparison of major modules of a microcomputer-based oscilloscope, it is learned that either one must specify the lowest bandwidth of all modules as the bandwidth of the whole digital oscilloscope system, or avoid the bandwidth limitation by tolerating the existence of waveform loss mechanisms. Due to the speed of I/O channels being slower than the one of arriving data from a sampled signal, part of the waveform is neglected, causing waveform loss as a result of the overflow of the data buffer during the data input to the microcomputer. If the dual measured signals are periodic, then the missing part of waveform is just an interval of repetitive parts of waveforms. Therefore, the missing waveform problem would not cause a serious distortion from the point of view of the signal amplitude. Thus, in this paper, we propose a M/M/1/N queueing model, based on the random signal sampling method in order to model and estimate the loss probability of the sampled data for input into the data buffer. In addition, by transformation of the data loss probability into the relationship of stored and lost waveform intervals of the dual measured signals, we can estimate the bandwidth limitation for an individual module, which causes the length of waveform loss. Conclusively, the primary results from the queueing model let the designers know the relationship between the missing intervals of the measured waveforms and the bandwidth difference among different modules of a microcomputer-based oscilloscope system [6-11].

The rest of this paper is organized as follows. In section 2, the dual channel microcomputer-based oscilloscope system and its modified M/M/1/N queueing model

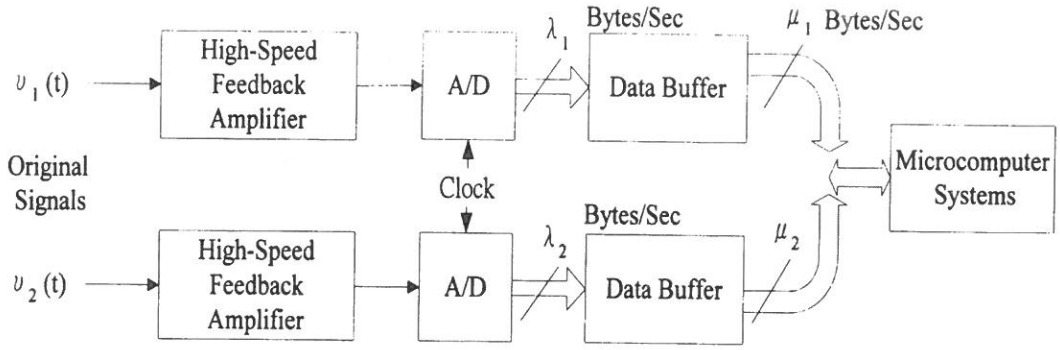

Fig. 1. A dual channel design for a microcomputer-based oscilloscope system

to analyze the characteristics of the data buffering are described. In section 3, the estimation of waveform missing intervals is discussed. In section 4, the measurement of an experimental platform for the microcomputer-based oscilloscope system is presented. The last section draws conclusions.

2. The Dual Channel Microcomputer-based Oscilloscope and Modified M/M/1/N Queueing Model for Data Buffering

Fig. 1 shows a dual channel design for a microcomputer-based oscilloscope system. In Fig. 1, the original signals $\upsilon_1(t)$ and $\upsilon_2(t)$ are input to the wide-band amplifier and then reach a analog to digital conversion module to receive the sampled signal $\upsilon_{s1}(t)$ and $\upsilon_{s2}(t)$ in a digital format. The data from the sampled signal is temporarily stored in a fast first-in first-out (FIFO) data buffer and then alternately read into a microcomputer for storing as a computer data file. The data then can be processed by curve fitting algorithms and displayed.

As a result of the speed difference between the arrival rate λ and the service rate μ of the data buffer in Fig. 1, arrival data may be lost if the data buffer is filled up and the service rate μ is not fast enough. Therefore, the data loss in the data buffer can cause the measured waveform loss. In this section, we use a modified M/M/1/N queueing model for the data buffering to estimate the data loss probability due to the speed difference between the data arrival rate and the data service rate of the data

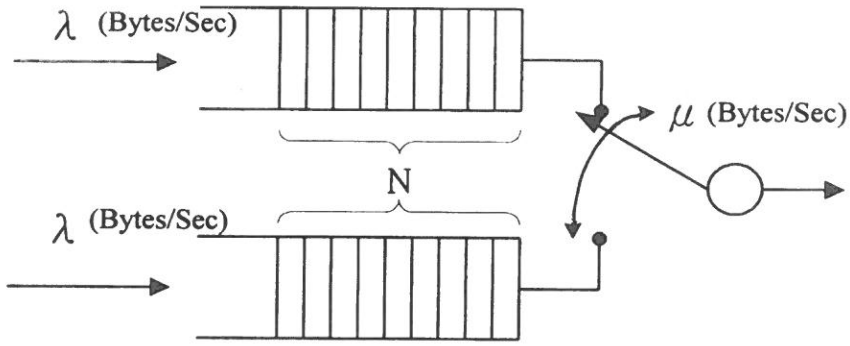

Fig. 2. A modified M/M/1/N queueing model for the data buffering in a microcomputer-based oscilloscope system

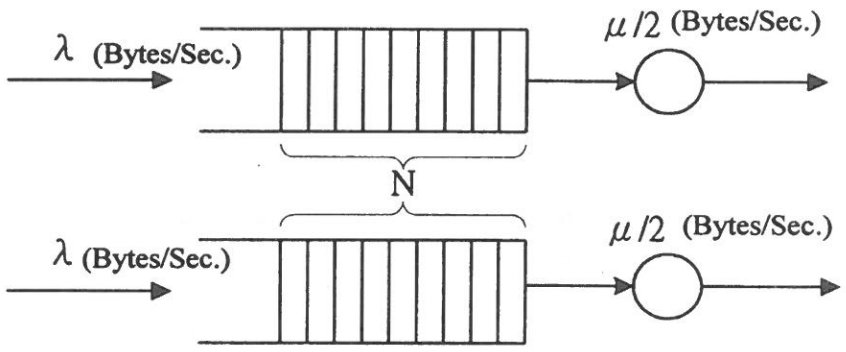

Fig. 3. An equivalent M/M/1/N queueing model for dual data buffering

buffer as shown in Fig. 2 [9-11]. For the simplicity of representation, we set $\lambda_1 = \lambda_2 = \lambda$ and $\mu_1 = \mu_2 = \mu$.

There are a couple of ways to analyze the modified M/M/1/N queueing model shown in Fig. 2. A simple way, we propose in this paper, is to separate the service rate of the server into a set of two half of service rates. The equivalent queueing model is shown in Fig. 3 where the two data buffer are working independently.

Because of the independent operations of the dual data buffer, the dual M/M/1/N queues can be analyzed separately. To analyze a M/M/1/N queue in Fig. 3, we use the state diagram as shown in Fig. 4 and find out the state probability for state 0 to state N.

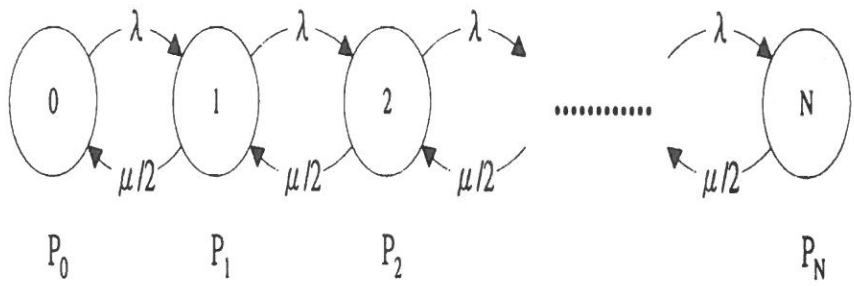

Fig. 4. State diagram of a M/M/1/N queue

Let P_n, $0 \leqq n \leqq N$, denote the state probability that there are n data in the data buffer. Based on the state probability model of a stationary non-time-varying system, the rate-equality principle yields the following set of balanced equations:

State	Rate at which the process leaves = rate at which it enters
0	$\lambda P_0 = (\mu/2)\ P_1$
$1 \leqq n \leqq N-1$	$(\lambda+\mu/2)\ P_n = \lambda P_{n-1} + (\mu/2)\ P_{n+1}$
N	$(\mu/2)\ P_N = \lambda P_{N-1}$

The argument for state 0 in the queue has no data in the buffer, namely, when in state 0, the process will leave only via an arrival (which occurs at rate λ) and hence the rate at which the process leaves state 0 is λP_0. On the other hand, the rate at which the process can enter state 0 is $(\mu/2)\ P_1$.

Rewriting the proceeding system of equations we get,

$$P_1 = \frac{\lambda}{\mu/2} P_0$$

$$P_{n+1} = \frac{\lambda}{\mu/2} P_n + \left(P_n - \frac{\lambda}{\mu/2} P_{n-1} \right) \qquad 1 \leq n \leq N-1$$

$$P_N = \frac{\lambda}{\mu/2} P_{N-1}$$

Which, solving in terms of P_0, yields

$$P_1 = \frac{\lambda}{\mu/2} P_0$$

$$P_2 = \frac{\lambda}{\mu/2}P_1 + \left(P_1 - \frac{\lambda}{\mu/2}P_0\right) = \frac{\lambda}{\mu/2}P_1 = \left(\frac{\lambda}{\mu/2}\right)^2 P_0$$

$$P_3 = \frac{\lambda}{\mu/2}P_2 + \left(P_2 - \frac{\lambda}{\mu/2}P_1\right) = \frac{\lambda}{\mu/2}P_2 = \left(\frac{\lambda}{\mu/2}\right)^3 P_0$$

$$\vdots$$

$$P_{N-1} = \left(\frac{\lambda}{\mu/2}\right)^{N-1} P_0$$

$$P_N = \left(\frac{\lambda}{\mu/2}\right)^N P_0 \tag{1}$$

By using the fact that $\sum_{n=0}^{N} P_n = 1$, we obtain,

$$P_n = \frac{\left(\frac{\lambda}{\mu/2}\right)^n \left(1 - \frac{\lambda}{\mu/2}\right)}{1 - \left(\frac{\lambda}{\mu/2}\right)^{N+1}}, \quad n = 0,1,\cdots N$$

Note that in this case, there is no need to impose the condition that $\frac{\lambda}{\mu/2} < 1$. The queue size or the buffer size, by definition is constant so there is no possibility of increasing [9,10,11].

When the queue is in state N, the arriving data will not enter the queue or the data buffer, so the sampled data of the measured data in this interval will be discarded. In other words, $P_n|_{n=N}$ is the data loss probability.

$$P_n|_{n=N} = \frac{\left[\frac{\lambda}{\mu/2}\right]^N \left[1 - \frac{\lambda}{\mu/2}\right]}{1 - \left[\frac{\lambda}{\mu/2}\right]^{N+1}} \tag{2}$$

Fig. 4 shows the characteristics of Eq. (2). Assume the utilization $\rho = \frac{\lambda}{\mu/2} = 0.1$, 0.5, 1, 2, 5, 10, $0 \leqq n \leqq N$, then we can obtain Fig. 5 in which the data loss probability increases with the increase of the utilization ρ. Also the data loss probability increases when the size of data buffer decreases. Therefore, to reduce the data loss probability, one can reduce the utilization ρ, or increase the data service rate $\mu/2$ to a certain level of the data arrival rate. In order to increase the data service rate $\mu/2$, we can reduce the I/O time delay by using faster I/O ports, assembly code and direct

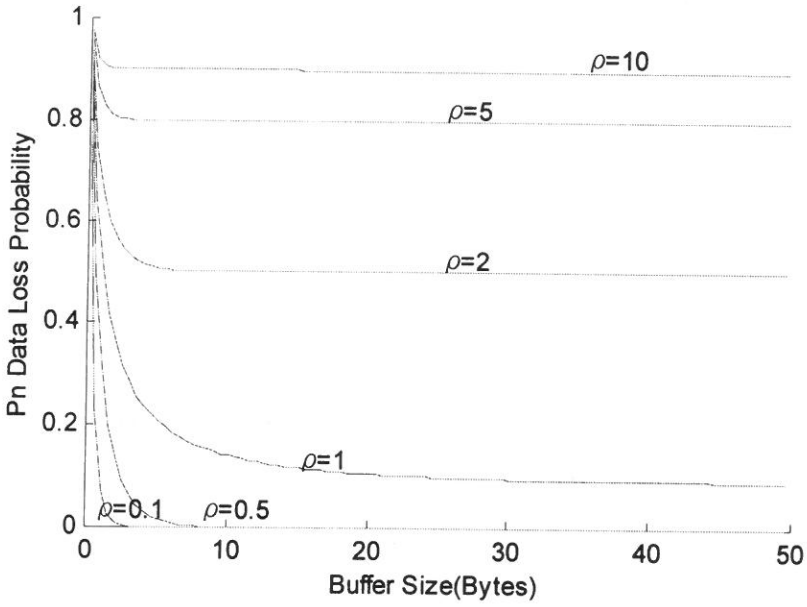

Fig. 5. The data loss probability of the data buffer

memory access mechanism. Another way to reduce the data loss probability is to increase the size of the data buffer.

3. The Missing Waveform Intervals of Dual Measured Waveforms

From the random sampling data point of view, the data buffer in Fig. 1 can be modeled as a M/M/1/N queue as shown in Fig. 2. Given a data arrival rate λ and a data service rate $\mu/2$, from the analysis of an equivalent M/M/1/N queueing model in Fig. 3, we can estimate the data loss probability of the queue, or the data buffer in Fig. 1. Furthermore, by transferring data loss probability, we can also estimate the measured and loss waveform intervals in a dual measured signal as shown in Fig. 6.

The data loss probability P_N shown in Eq. (2) can be approximately equal to T_{Loss}/T_{Total}, where TLoss is a loss waveform interval and T_{Total} is the measured waveform interval in Fig. 6 and $\rho=\frac{\lambda}{\mu/2}$. We can obtain Eq. (3) equivalently.

$$\frac{T_{LOSS}}{T_{Total}}=\frac{(\rho)^N\ (1-\rho)}{1-\ (\rho)^{N+1}} \tag{3}$$

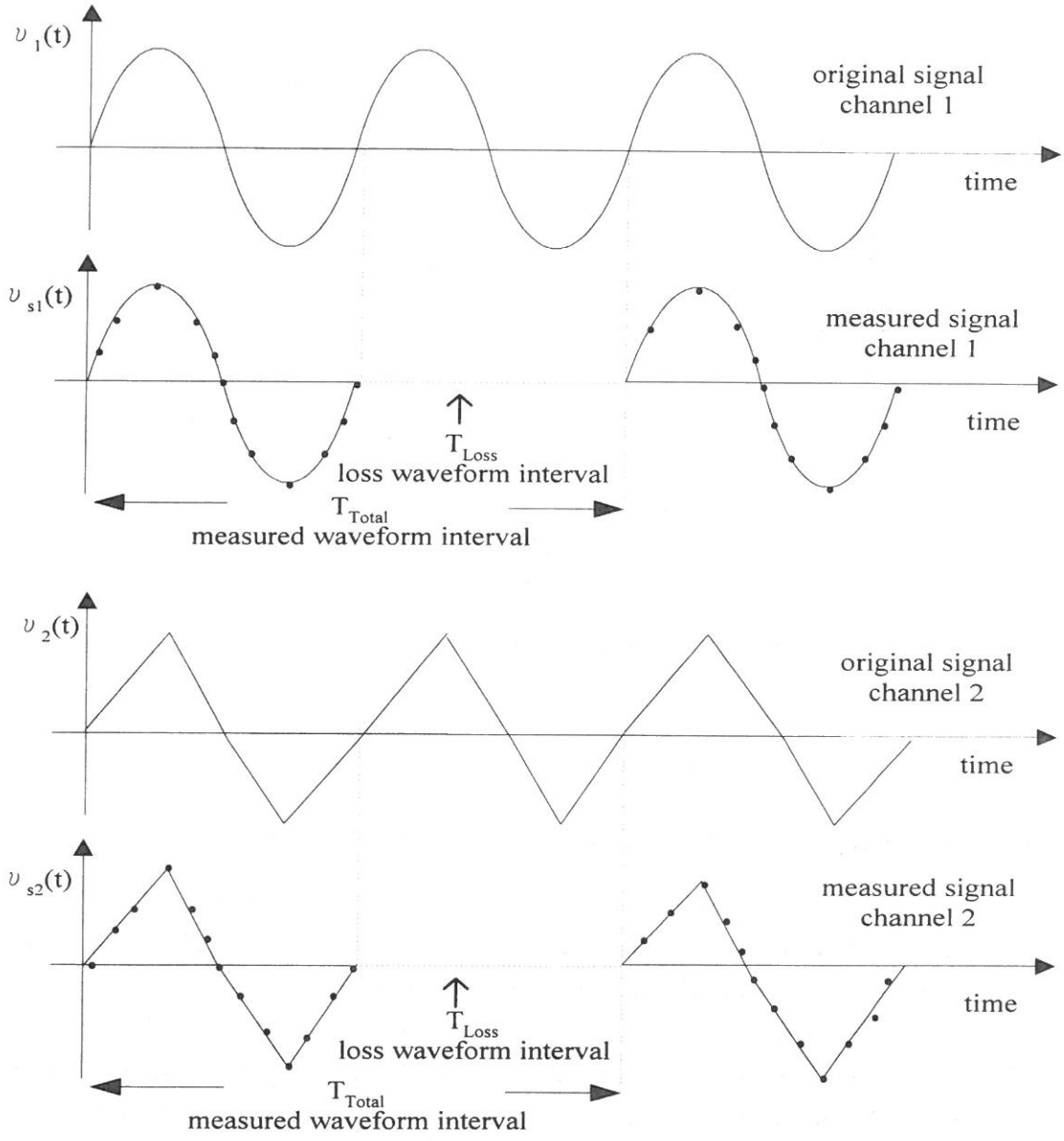

Fig. 6. Measured and loss waveform intervals in a dual measured signals.

Fig. 7 shows the characteristic curves for Eq. (3) with data buffer size N = 1, 10, 100, 1000 bytes, and the utilization ρ varies from 0 to 5. The curves show that the

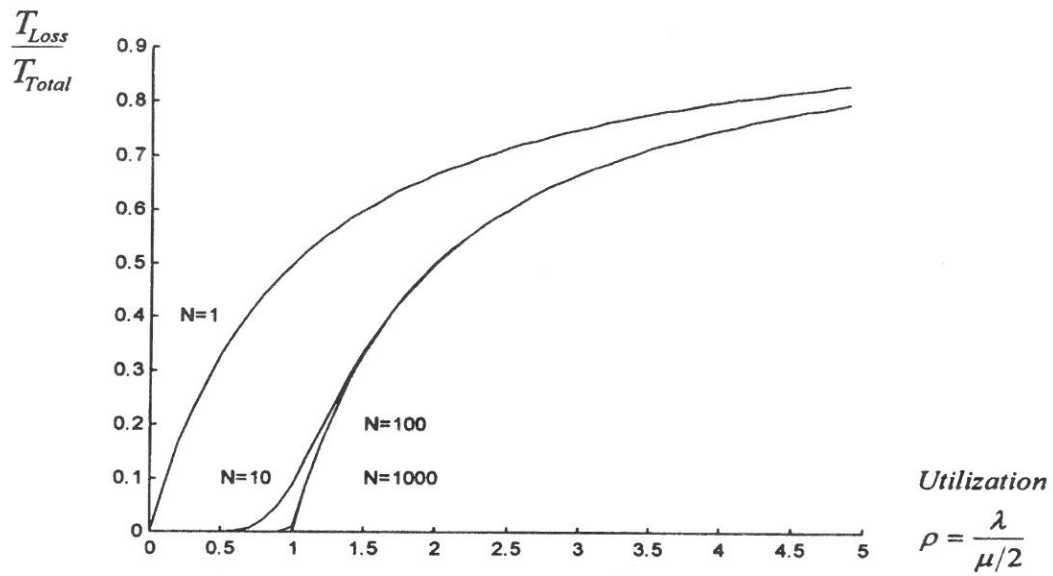

Fig. 7. The ratio between the loss and the measured waveform intervals

smaller the utilization ρ, the smaller is the loss waveform interval in the measured waveform interval, or the smaller is the distortion. In addition, the larger the buffer size for data buffer in Fig. 1, the smaller is the loss waveform interval.

Although there exists a missing waveform interval of the measured waveform, the microcomputer-based oscilloscope can still show part of the repetitive periodic waveforms. If the missing interval of the measured waveform belongs to part of the repetitive waveform, then a missing mechanism may not cause a serious distortion in comparison with the functions of a traditional analog oscilloscope.

4. Implementation and Measurements of the Experimental System

In order to fulfill the dual channel microcomputer-based oscilloscope system shown in Fig. 1, we have designed and implemented an experimental system based on the hardware block diagram and software flowchart as shown in Fig. 8 and Fig. 9, respectively. Fig. 10 shows the physical implementation of a dual channel microcomputer-based oscilloscope system circuit board. Fig. 11 is the program segment translated into assembly code, which is shown in Table 1.

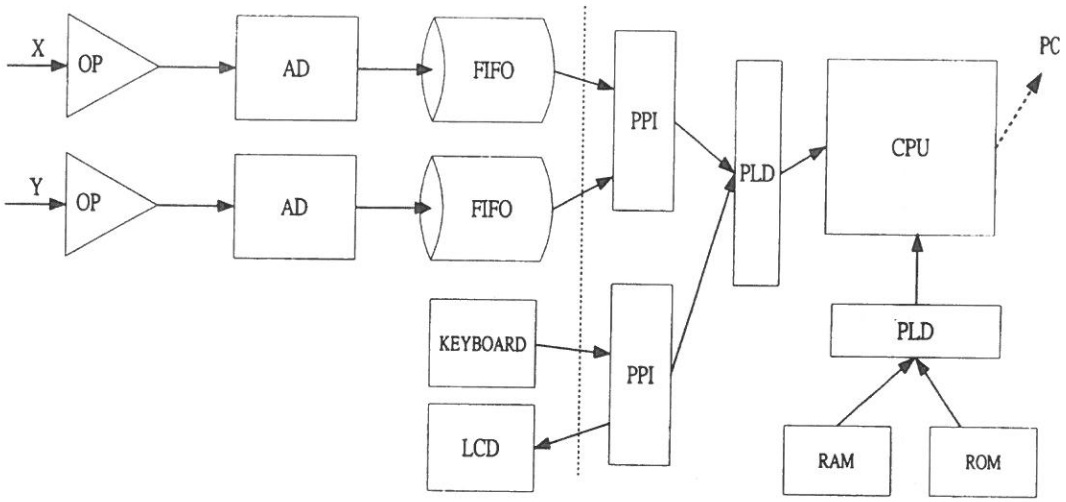

Fig. 8. Hardware block diagram of the dual channel microcomputer-based oscilloscope system

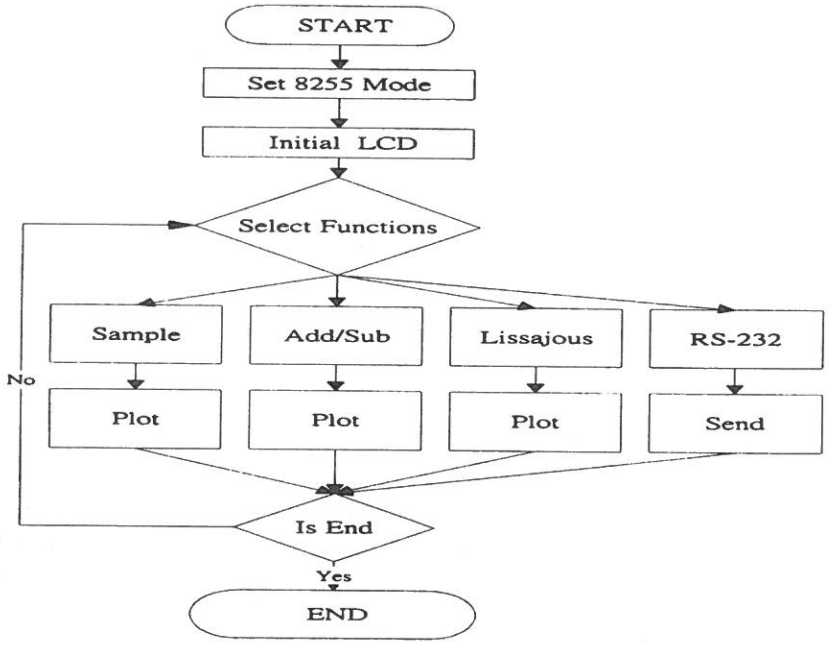

Fig. 9. Software flowchart of the dual channel microcomputer-based oscilloscope system

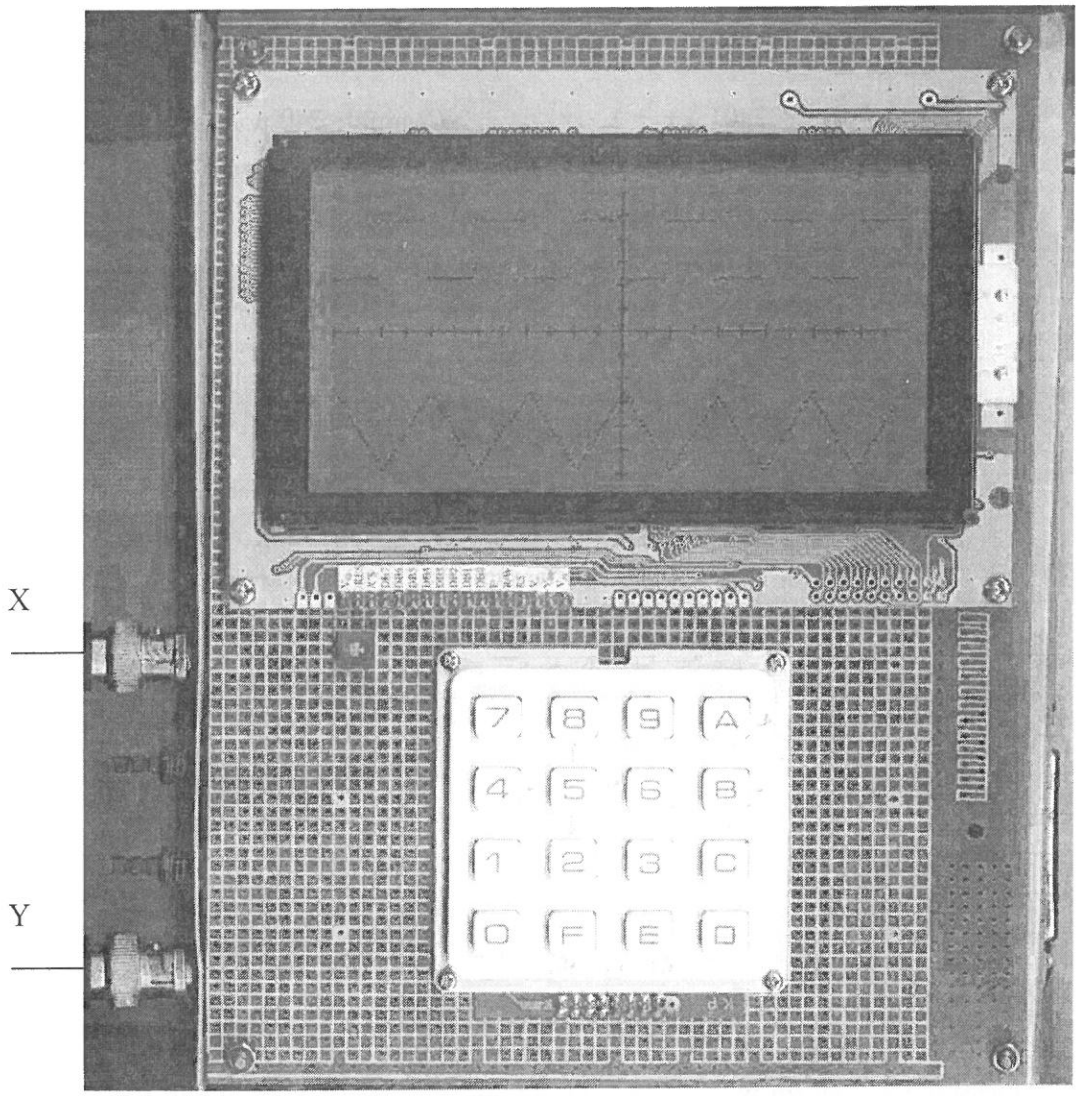

Fig. 10. The circuit board of the dual channel microcomputer-based oscilloscope system

Since the quartz crystal frequency of the 8051 is 12MHz in this system, the clock cycle period is 1/12us. Each machine cycle period is 12 times the clock cycle period in a 8051 and a machine cycle period is 1us. The 28 machine cycles in Table 1 need 28us to read a byte, or a sampled point of the measured waveform. A 28us data input period is equivalent to 36KHz bandwidth for the sampled point of the measured waveform. If

```
void Save-X-RAM (void)
{    ...
     for (i=0; i<240; i++)              /*sample 240 points */
      { FIFO1-CTL=FIFO1-RD-1; /*send out a READ signal */
      FIFO1-CTL=FIFO1-RD-0; /*to read a byte */
      FIFO1-CTL=FIFO1-RD-1;
      FIFO1-RAM [i] =FIFO1-DB; /* save to RAM */
      }
      ...
}
```

Fig. 11. The program segment for input data into a microcomputer

Table 1. The assembly code for input data into a microcomputer

Operator	Machine Cycle	Operation Times	Required Machine Cycles
MOV	1	12	12
MOVX	2	5	10
ADD	1	1	1
ADDC	1	1	1
INC	1	2	2
DEC	1	1	1
CLR	1	1	1
			28

we need ten sampled points for each period, we may only have 3.6KHz bandwidth for the measured waveform without missing waveform interval as shown in Fig. 12.

If we can tolerate the missing waveform interval for the measurement of the periodic waveforms, then we can obtain a scalable bandwidth of the measured signals. From the scalable relationship shown in Eq (3), we learn that if the equivalent service rate is $\mu/2=18$K byte/sec., and the arrival rate is from $\lambda=1$K byte/sec. to 20M byte/sec., then, we can obtain the ratio between the loss waveform interval and the measured waveform. From Fig. 13, it can be seen that if we can tolerate the loss waveform at 99% of the measured waveform, then we can have 350KHz of measured

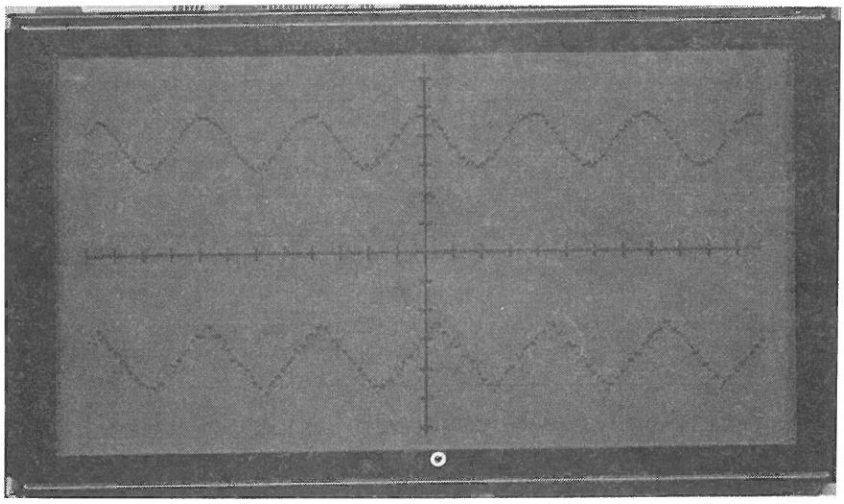

Fig. 12. A measured waveform without missing waveform interval shownin the experimental system

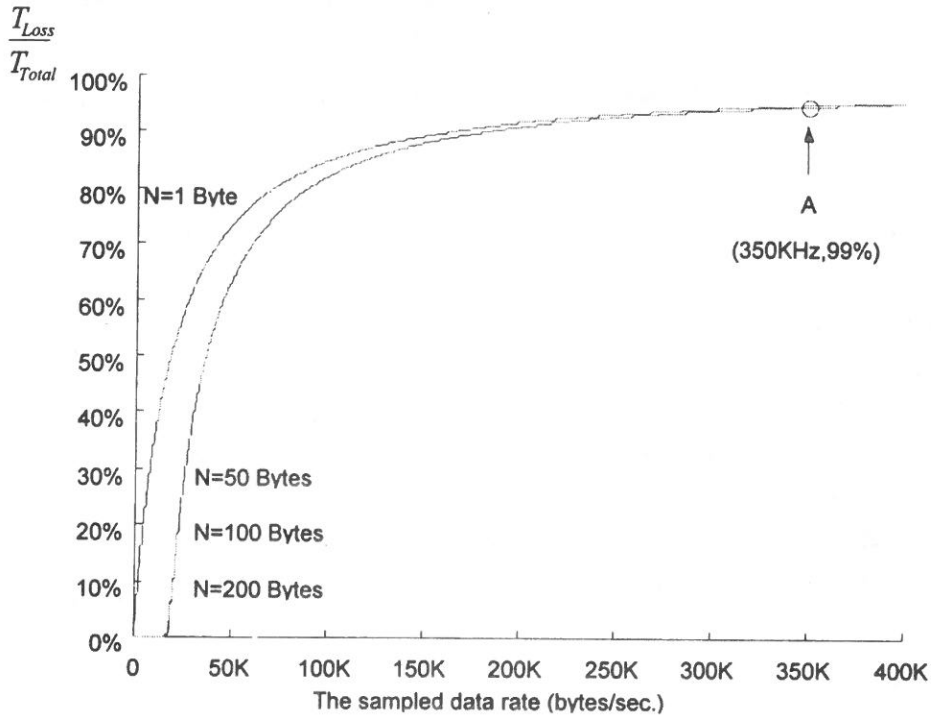

Fig. 13. The relationship between sampled data rate and loss waveform interval.

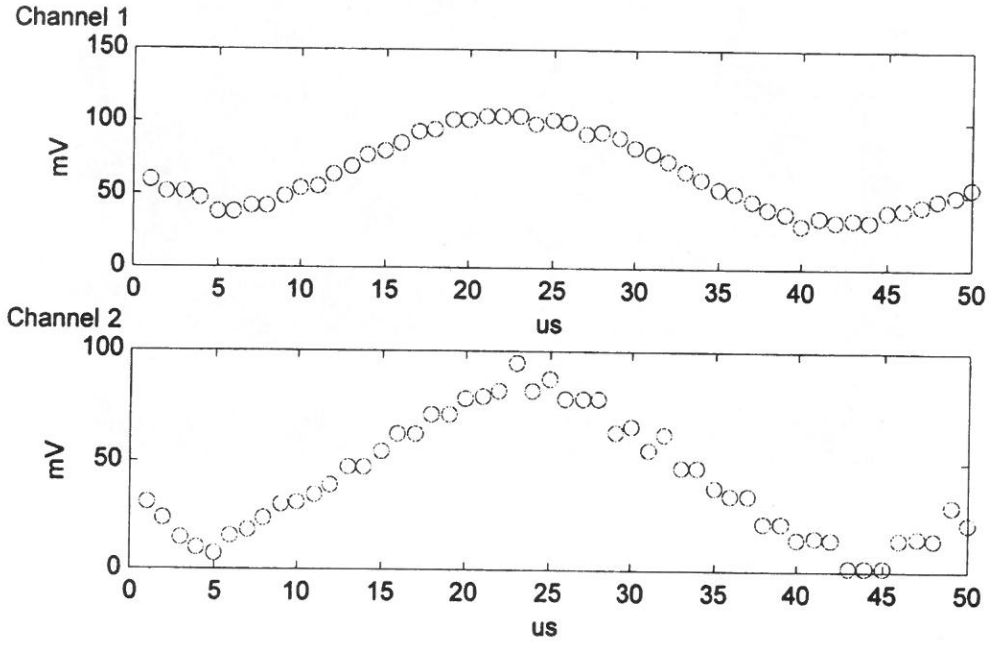

Fig. 14. A sampled signal shown in the dual channel microcomputer-based oscilloscope system.

```
y1=interp1 (x1, s1, i1, 'cubic');
y2=interp1 (x2, s2, i2, 'cubic');
subplot (2, 1, 1);    plot (i1, y1, '-');
xlabel ( 'us'); ylabel ( 'mV'); axis ( [0,50,0,150]);
subplot (2, 1, 2);    plot (i2, y2, '-');
xlabel ( 'us'); ylabel ( 'mV'); axis ( [0,50,0,100]);
```

Fig. 15. A curve-fitting program for dual sampled signals

bandwidth shown as point "A" for the experimental system.

Fig. 14 shows a sampled signal in the dual channel microcomputer-based oscilloscope system. This sampled signal is processed by the curve-fitting program as shown in Fig. 15 with which we can obtain a reconstructed signal that is shown in the display of our computer as Fig. 16.

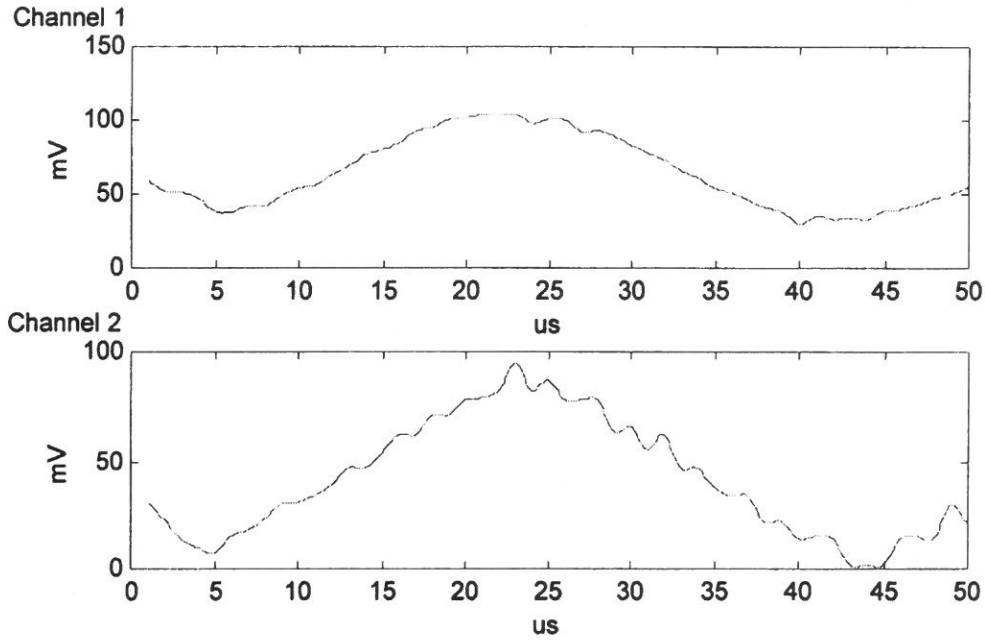

Fig. 16. A dual signal after the curve-fitting program

5. Conclusions

In this paper a modified M/M/1/N queueing model for the data buffering of a dual channel microcomputer-based oscilloscope system has been proposed. In this model, the arrival process and the departure process can represent the sampled signal input and data output process of a data buffer, respectively. From steady-state behavior analysis, we can obtain the approximate results for the data loss probability and missing waveform intervals. This probability is dependent on the data arrival rate and the data service rate of the data buffer in a dual channel microcomputer-based oscilloscope system. The higher the bandwidth of the measured signals we need, the higher will be the data arrival rate and the data loss probability. At the same time, the higher the bandwidth of the measured signals we need, the higher will be the missing waveform interval. Fortunately, the oscilloscope is used to measure repetitive periodic waveforms, so the missing waveform intervals may not cause any serious problem from

a steady-state point of view. We have also designed and implemented an experimental microcomputer-based oscilloscope. The measured data is in very good agreement with the result of the theoretical analysis. Furthermore, the post processing in a computer for the measured signal can be done by many ways. A typical curve-fitting program to reconstruct the measured signals has also been presented.

References

(1) M. J. Riezenman, "Test and Measurement," IEEE Spectrum, Vol. 32. 1, PP. 52-52, Jan. 1995.

(2) S. A. Chickamenahalli and A. Hall, "Interfacing a Digital Oscilloscope to a Personal Computer Using GPIB," Frontiers in Education Conference, 27th Annual Conference, Teaching and Learning in an Era of Change, Proceedings Vol. 2, PP. 904, 1997.

(3) S. Matsukura, T. Asaka, Y. Sugihara, and N. Tonosaka, "Fast up-date Techniques for Digital Oscilloscope," 8th IEEE Instrumentation and Measurement Technology Conference, PP. 341-346, 1991.

(4) H. B. Crawley, R. Mckay, W. T. Meeyer, E. I. Rosenbergg, and W. D. Thomas, " Using IEEE Standard 1057 for Testing Analog-to-digital Converters," Nuclear Science Symposium and Medical Imaging Conference Record, Vol. 1, PP. 272-276, 1995

(5) A. Luh, "Instrumentation Integrated System," WESCON/94, Idea/Microelectronics Conference Record, pp. 482-486, 1994

(6) Y. Lembeye, J. P. Keradec and G. Cauffet, "Improvement in the Linearity of Fast Digital Oscilloscopes Used in Averaging Mode," IEEE Transactions on Instruction and Measurement, Vol. 43, Issue: 6, PP. 922-928, Dec. 1994.

(7) E. R. Pratico and M. A. Eltzmann, "A Microcomputer-based Data Acquisition System for Transient Network Analyzer Operation," IEEE Transactions on Power Systems, Vol. 9, No. 2, PP. 812-818, May 1994.

(8) Jian Qiu Zhang, Zhao Xinmin, Hu Xiao and Sun Jinwei, "Sinewave Fit Algorithm Based on Total Least-Squares Method with Application to ADC Effective Bits Measurement, Vol. 464, PP. 1026-1030, Aug. 1997.

(9) Sheldon M. Ross, Introduction to Probability Models, Sixth Edition, Academic Press, 1997.

(10) Guntter Bolch, Stefan Grener, Hermann DeMeer, and Kishor S. Trivedi, Queueing Networks and Markov Chains, John Wiley & Sons, Inc., 1998.

(11) Randolph Nelson, Probability, Stochastic Processes, and Queueing Theory, Springer-Verlag New York. Inc., 1995.

received September 27, 2000
accepted November 6, 2000

A Modified M/M/1/N Queueing Model for the Data Buffering of
a Dual Channel Microcomputer-based Oscilloscope System

雙通道微算機示波器系統資料緩衝之修正 M/M/1/N 佇列模型分析

白 英 文　　魏 鴻 祺
輔仁大學電子工程系

摘　　要

傳統上，示波器是用來顯示電子電路類比電壓波型之特用儀器，最近類比電壓波型可以被數位化，量測信號因而易於被計算機處理、儲存並且可以透過計算機網路傳輸。這篇論文提出以微算機爲基礎之雙通道數位化示波器系統設計與建構，並以修正 M/M/1/N 佇列模型分析數位化電壓波型資料抵達資料緩衝器之資料漏失機率，進一步將資料緩衝器之資料漏失機率轉換成量測信號波型漏失區間，從模型分析和實驗平台量測結果，顯示我們獲得了相同的趨勢，也就是量測信號之資料抵達率和資料緩衝器服務率比值愈小時，雙通道量測信號波型漏失區間也就愈短。

關鍵詞：微算機，示波器，佇列模型，示波器，資料緩衝，漏失波型區間

Design and Implementation of a Microcomputer 8051 System Powered by Dual Batteries Charged by Solar Cells

Ying-Wen Bai and Cheng-Lun Chang
Department of Electronic Engineering
Fu Jen Catholic University
Taipei, Taiwan 242, R.O.C.

Abstract

Single chip microcomputer systems are becoming increasingly popular in today's control and information applications. But, due to the limitation of battery energy, these systems only have very limited operation time or recharge cycles, if their power is supplied by rechargeable battery. Thus, in this paper, we propose a design and implementation for the software and hardware of a microcomputer 8051 system powered by a dual rechargeable battery that is charged by solar cells. From the feasibility analysis of the queueing model for the stochastic charging and discharging process of the dual battery system, we learn that the operation time in our design can be much longer than that of a single rechargeable battery power supply. With a two third utilization ratio, we can obtain an operation time four times higher than with a single rechargeable battery power supply. In addition, the technology trend shows the power consumption rate for a typical microcomputer system is decreasing and the power generation efficiency for typical solar cells is increasing. Hence, solar cells as the power charging sources of the microcomputer 8051 system supplied by a dual rechargeable battery can be feasible in the near future.

Key Words: Microcomputer, Solar Cells, Queueing Model, Operation Time.

INTRODUCTION

Over the past few years, numerous researchers have been working with different levels of low power technology for microcomputer system design. In terms of system, circuits, and device power saving, the results show that the average power consumption of a microcomputer computer has more than a twenty-percent and a ten-percent reduction every year from 1992 to 1997 and 1998 to 2001, respectively. Reducing power consumption is attractive due to its potential to extend the recharge period of portable information applications. Further, the longer the battery operation time before recharge, the more convenient it is for mobile users to operate a portable microcomputer system [1,2,3,4].

Eventually, the power consumption of a single chip microcomputer system will be small enough to be supplied or recharged by other power sources. One of the proposed power sources is mechanic vibration [5]. Among others, we previously proposed solar cells that can be used as power supply sources [9,10]. Although mc-Si solar cell power generation efficiency at present is not high enough, their efficiency increased from 14.2 percent to 16.8 percent during the years of 1990 to 1997. This improvement can also reduce the gap between the charging and discharging rate of the power supply of a microcomputer system, so the power exhaustion probability within a certain operation time will be reduced year by year [6,7,8].

To prolong the battery operation time before recharging, in this paper, we design and implement the software and hardware module for a single chip microcomputer 8051 system with a dual battery, charged by solar cells. Based on the design and implementation, this paper also shows the estimation for the power exhaustion probability and the experimental measurement for the operation time that depend on the power generation efficiency of solar cells and the power consumption rate of a microcomputer. In addition, due to the overlapping of charging and discharging period, if the ratio between the charging and discharging rate is two third, then the operation time can potentially be prolonged four times in comparison with the case of a single rechargeable battery.

The rest of this paper is organized as follows. In section 2, the technology trend

about the power consumption of a microcomputer and the power generation efficiency of solar cells are discussed. In section 3, the queueing model for the stochastic charging and discharging behavior for the dual rechargeable battery in a single chip microcomputer system is presented. In addition, the feasibility estimation for the dual rechargeable battery in a single chip microcomputer system is given. In section 4 and 5, the design and implementation of the software and hardware modules for this system are provided. In section 6, the experimental results of this system are given. The last section draws some conclusions.

2. Technology Trend about Power Consumption of Microcomputers and Power Generation of Solar Cells

In order to extend the operation time before recharging the battery in a microcomputer system, recently some research about low-power technology has been conducted [1,2]. The typical low-power technology includes low-power electronic devices, storage devices, displays and power management methods of communication and computation systems. The results show that the power consumption of a typical notebook computer, one type of microcomputers, has decreased over the last decade [3,4]. Fig. 1 shows the technology trend. From this figure it can be seen that from 1992 to 1994, there was an improvement of more than 100 percent, but from 1994 to 1997 there was an improvement of 30 percent and from 1998 to 2002 there will be an improvement of only 10 percent. Due to the computation performance consideration, the curve in Fig. 1 shows that the power efficiency improvement is slowing down but still gaining substantial value. In addition, the low power VLSI circuit technology is improving year by year, so the power consumption of a notebook computer can be below 5W around 2002. Table 1 shows the recent technology level for the power consumption efficiency of computers including microcomputer 8051 systems. Currently, a microcomputer 8051 consumes less than 0.1W.

For the charging capability, the power generation efficiency of typical solar cells has increased more than 20 percent in the last five years as shown in Fig. 2. With the case of a typical area of 100 cm^2, we learn that the power generation efficiency of solar cells is still improving each year up to now, although there is a theoretical limitation

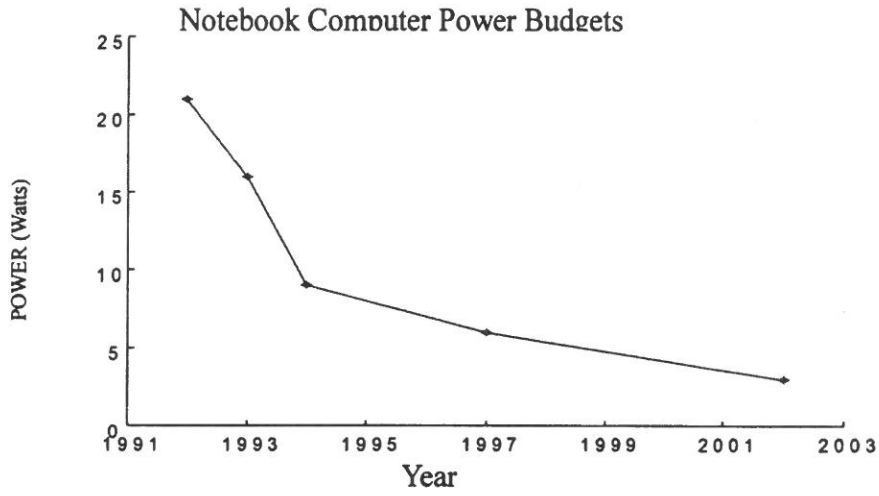

Fig. 1. The power consumption for a typical notebook computer

Table 1. The power consumption efficiency of microcomputers

	Power (W)	Energy (Wh)	Weight (lb)	Length (in)	Width (in)	Height (in)
Notebook	5 to 15	18 to 36	5.5 to 7.5	11 to 12	7.5 to 9	1.5 to 2
Subnotebook	2 to 10	5 to 24	3 to 5	9.5 to 11	6 to 8	1.5 to 2
Palmtop	<.5	5.5	0.7	6.3	3.5	1
PDA	<.5	5	0.75	7	4.5	1
8051 Microcomputer	Normal mode <0.1W Idle mode <25mW Low power mode <0.25mW			7	3.5	0.5

for the power efficiency of photo-volt transformation. Form the above characteristics of typical solar cell products, we can derive the power generation density (W/cm^2), as shown in Table 2. Currently, the area of a portable microcomputer is around 100 that is estimated from the physical size of a handset. Combining the area of the handset and power generation efficiency, we learn that the average power generation of typical solar cells is at about the level of 0.384W under the sunshine.

Fig. 3 shows a typical design for solar cells powered microcomputers. The energy

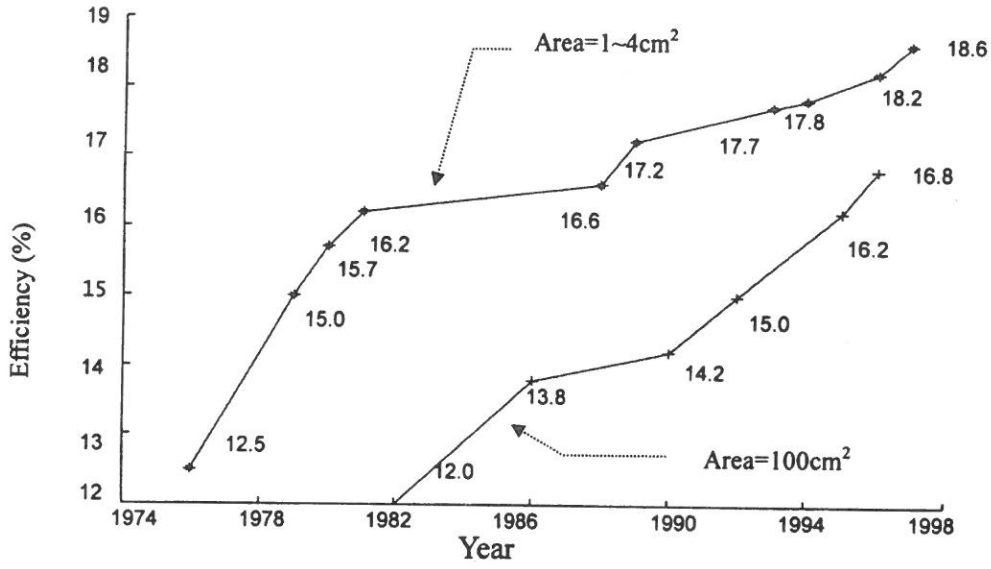

Fig. 2. Efficiency progress of 1~4 cm² and 100 cm² mc-Si solar cells over the past 20 years [6-8].

Table 2. The power density of typical solar cells.

Type model	Size, cm			Power		Power density
	Width	Length	Thickness	V_{op}, v	I_{op}, mA	mW/cm²
SP 1506	15.24	5.50	0.235	7.5	48.0	4.30
SP 1508	15.24	7.62	0.235	7.5	60.0	3.88
SP 1512-2A	15.24	15.24	0.235	7.5	130.0	4.20
SP 1512-2B	15.24	15.24	0.235	7.5	106.0	3.42
SP 1530	15.24	29.00	0.235	7.5	200.0	3.40
					Average	3.84

Condition FL: 200 Lux, 25℃

is generated by solar cells and is used to charge up the rechargeable battery. Usually, the power generation should be greater than or equal to the power consumption, so that the microcomputer can work without exhausting battery power. However, because the power generation can be less than the power consumption, especially indoor, the microcomputer will work with a power exhaustion probability. As the rechargeable battery provides the energy storage mechanism, the power generated from solar cells can be used to charge up the rechargeable battery even if the machine is

turned off. The stochastic characteristics of power charging and discharging behavior can be modeled by a queueing model [9]. The previous result shows that under certain operational conditions there is only a 0.46 hours operation time in a 24 hours period with a power exhaustion probability of less than 1% [11].

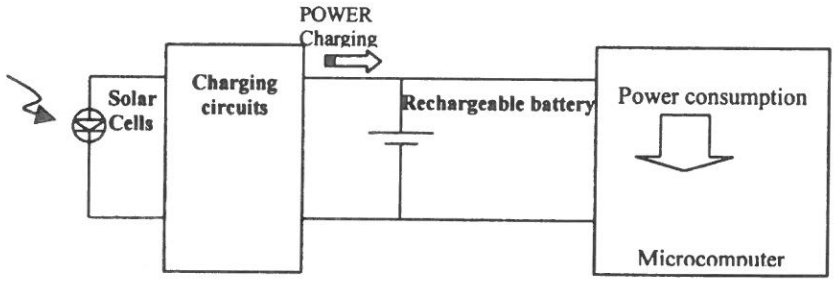

Fig. 3. A typical design for solar cell powered microcomputer

However, as the operation time or the utilization rate of a microcomputer is not long enough, the solar cell powered microcomputer may be not a feasible solution nowadays. There are several ways to extend the operation time. One of them, proposed in this paper, is equipped with dual rechargeable batteries as shown in Fig. 4. Its power charging and discharging behavior are discussed in the next section.

3. The Queueing Model for the Power Charging and Discharging Behavior of Dual Rechargeable Battery

As shown in Fig. 4, the energy is generated by solar cells and is used to charge up the dual rechargeable batteries alternatively. When the first battery power is low, the second battery will be switched on as a power supply, and vice versa. The power control mechanism includes voltage level detection; interrupt handling mechanism, and sending control data to a switching box by an output port of the portable microcomputer [10-12].

In a typical power charging and discharging model, solar cells can be viewed as the charge sources for the rechargeable battery. The battery power increases when the charges enter it at a higher rate than they leave. In other words, the battery power decreases when the charges leave it at a higher rate than they enter. From a typical

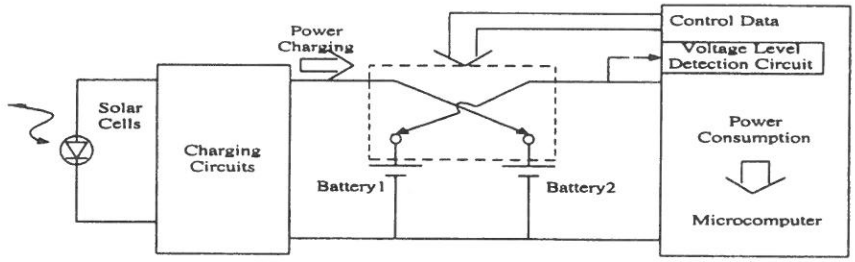

Fig. 4. A dual rechargeable battery design for a solar cell powered portable microcomputer

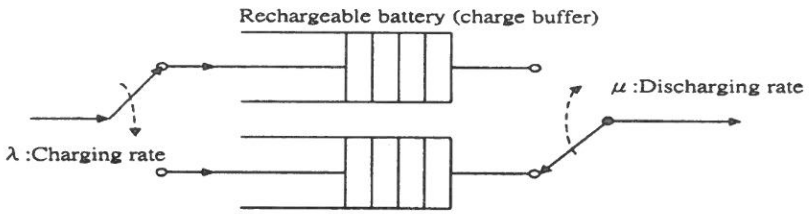

Fig. 5. A modified M/M/1 queueing model for a dual rechargeable battery system

random operating point of view, we use a modified M/M/1 queueing model, shown in Fig. 5, for charging and discharging of the dual battery module in a portable microcomputer [9].

In addition, the probability distribution for the charge arrival process and the charge departure process is dependent on the operation characteristics of mobile users. The behavior of the power charging and discharging for every microcomputer can be very different.

Because of the individual difference in operational behavior, it can be very difficult to give a precise analysis for the queueing model. However, an approximate analysis can be made, which can provide certain estimated characteristics for the charge arrival and departure processes. For the sake of simplicity, the approximate analysis is based on the non-time-varying and stationary queueing model. Recently, the steady-state M/M/1 queueing model and the dynamic M/M/1 queue model have been analyzed [9-13]. Both the models represent a Poisson charge arrival process and exponential departure process. In addition, for the special user operation characteristics, the

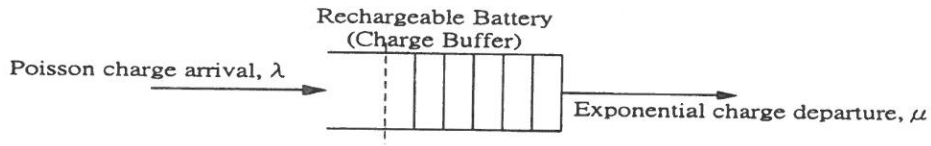

Fig. 6. An equivalent M/M/1 queueing model for a dual rechargeable battery

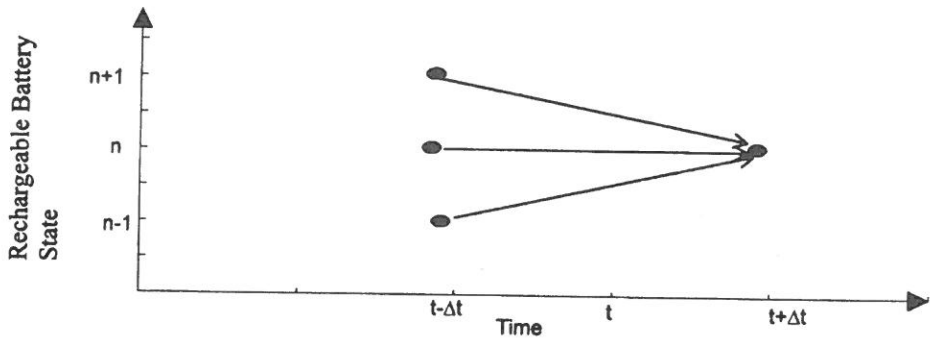

Fig. 7. M/M/1 state-time diagram for the model of a rechargeable battery

characteristics for the specific charge arrival process and the departure process can be used for the similar analysis.

For the approximate analysis of the charge arrival process and the departure process, we use a single equivalent M/M/1 queueing model as shown in Fig. 6 in order to represent the dual M/M/1 model shown in Fig. 5. The equivalent model is similar to a single queue with a single-server type with double charge buffer, a Poisson charge arrival rate of λ, and an exponential charge departure rate of μ.

As noted in the reference [13], concerning the static properties of the M/M/1 queue, the average rechargeable battery energy occupancy, and the probability of using-up the energy of the rechargeable battery are readily determined once we find the probabilities of state P_n of the rechargeable battery. P_n represents the probability that there are n units of charges in the battery and P_0 represents the probability that there are zero units of charges in the battery. In other words, P_0 is the probability of using-up the energy in a rechargeable battery. If P_0 is small enough, based on a certain operating environment, then the solar cell powered handheld microcomputer system

can be acceptable from the power supply point of view. Specifically, we assume that the charge arrival process to the rechargeable battery be a Poisson distribution, with a charge rate λ. Assume the charge departure rate to be exponential, with a discharging rate μ, and then the probability $P_n\ (t+\Delta t)$ means that there are n units of charges in the rechargeable battery at time $(t+\Delta t)$. Referring to the state-time diagram shown in Fig. 7, it is assumed that if the rechargeable battery is in state n at time $t+\Delta t$, it can only have been in states n−1, n, or n+1 at time t.

The charging and discharging process of a rechargeable battery can be represented by a typical queue. Therefore, the probability of the queue in state n at time $t+\Delta t$ can be the sum of the state probability in state n-1, n, or n+1 at time t, each multiplied by the probability of charges arriving at state n in the intervening Δt units of time. We thus have, a generating equation for$P_n\ (t+\Delta t)$,

$$\begin{aligned} P_n(t+\Delta t) &= P_n(t)[(1-\lambda\Delta t)(1-\mu\Delta t)+\mu\Delta t\lambda\Delta t+O(\Delta t)] \\ &\quad + P_{n-1}(t)[(\lambda\Delta t)(1-\mu\Delta t)+O(\Delta t)] \\ &\quad + P_{n+1}(t)[(\lambda\Delta t)(1-\mu\Delta t)+O(\Delta t)] \end{aligned} \tag{1}$$

Since $O\ (\Delta t)$ includes terms of orders $(\Delta t)^2$ and higher, the terms involving $(\Delta t)^2$ in Eq (1) should be incorporated in O (Δt). Simplifying Eq (1) by dropping O (Δt) terms and higher order terms, one can obtain $P_n\ (t)$ shown in Eq (2) by solving differential-difference equations.

$$P_n(t)=\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^2\left[1-\sum_{k=0}^{\infty}\left(\frac{(\lambda t)^k e^{-\lambda t}}{k!}e^{-\mu t}\sum_{m=0}^{n+i+1+k}\left(\frac{\mu t}{m!}\right)^m\right)\right] \tag{2}$$

In taking the limit as t→∞, we see the terms of the formulation $e^{-(\lambda+\mu)tm}$ which go to 0. Hence, when t→∞ the result is,

$$\lim_{t\to\infty} it\ P_n(t)=P_n(t)=P_n=\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \tag{3}$$

When $\lambda/\mu<1$ we get a valid steady-state probability distribution and when n=0, we obtain the power exhaustion probability P_0.

$$P_0=1-\frac{\lambda}{\mu},\quad 0\le\frac{\lambda}{\mu}<1 \tag{4}$$

Where λ is the charging rate for the power generation of the solar cells and μ is the

discharging rate for the power consumption of a microcomputer system. Because common users are not using the microcomputer system all the time, the utilization of a microcomputer could be smaller than 100%. Hence, we can consider the utilization (u. t.) and then obtain an approximate equivalent power consumption rate as,

$$\hat{\mu} = (u.t.)\mu \tag{5}$$

In addition, because the solar cells may not generate power all day due to the lack of sunshine, the duty cycle of the generating power could be smaller than 100 %. Hence, we can consider the duty cycle ($d.\ c.$), and then obtain an approximate equivalent power generation rate as

$$\hat{\lambda} = (d.c.)\lambda \tag{6}$$

If we rewrite Eq (4) with the equivalent power consumption rate $\hat{\mu}$ and power generation rate $\hat{\lambda}$, we obtain

$$P_0 = 1 - \frac{\hat{\lambda}}{\hat{\mu}} = 1 - \frac{(d.c.)\lambda}{(u.t.)\mu} \tag{7}$$

During daylight, the duty cycle (d. c.) can be assumed to have a typical value, for example, d. c. = 0.3. Fig. 8 shows the relationship between the power using up probability and the utilization (u. t.) of a microcomputer system. Fig. 8 shows the

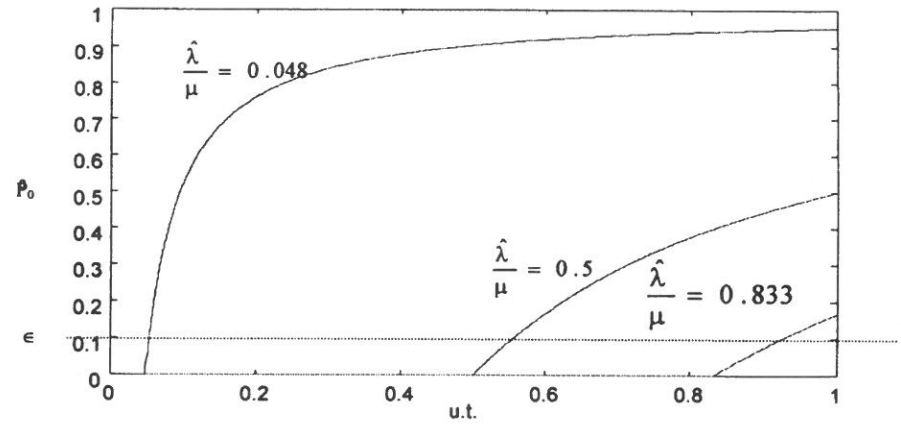

Fig. 8. The relationship between the power using up probability and the utilization ration of microcomputer system.

smaller utilization with respect to the smaller power using-up probability for a given set of parameters such as: the charging rate (λ), the duty cycle (d. c.), and the discharging rate (μ).

For the current technology level, if we assume $P_0 \leqq 0.01$, a very small probability for using up power, then we can obtain Eq (8),

$$u.t. \leq \frac{0.3\lambda}{0.99\mu} = \frac{0.3 \times 0.384\ W}{6W} = 1.92\% \tag{8}$$

u. t. $\leqq 1.92\%$ means the microcomputer operation time can be 0.46 hours per day. Gradually, when the technology improves, the ratio $\hat{\lambda}/\hat{\mu}$ increases, and the operation time will be prolonged. Within five years, $\hat{\lambda}/0.99\mu$ can reach (0.3) 0.5W/0.99·2W, and u. t. $\leqq 7.6\%$, which means the operation time can be 1.82 hours in 24 hours, a reasonable operation time for a common mobile user per day. By applying the same procedure, Table 3 is derived and it shows, sooner or later, that the solar cell powered microcomputer systems will be feasible from a viewpoint of the power supply.

Table 3. The relationship among ratio, utilization u. t., and operation time

$(\frac{\hat{\lambda}}{0.99\ \mu})$	$\frac{(0.3)\ (0.384w)}{(0.99)\ (6w)}$	$\frac{(0.3)\ (0.4w)}{(0.99)\ (4w)}$	$\frac{(0.3)\ (0.5w)}{(0.99)\ (2w)}$	$\frac{(0.3)\ (0.6w)}{(0.99)\ (1w)}$
Utilization (u. t.)	1.92%	3.03%	7.6%	18.2%
Operation time in 24 hours	0.46 hours	0.72 hours	1.82 hours	4.3 hours

3.1 The Relationship between Operation Time and Utilization for the Dual Rechargeable Battery System

To prolong the operation time, we have designed and implemented a microcomputer 8051 system powered by a dual battery charged by solar cells. To estimate the total operation time for the dual battery system, we assumed that both the rechargeable batteries are full at the starting point and summed up each operation time interval for using the battery 1 and 2 alternatively.

$$T_{\text{total operation time}} = t_1 + t_2 + t_3 + t_4 + \cdots$$

When $t_1 = \frac{Q}{\mu}$, Q: Energy capacity of a Rechargeable Battery

$t_2 = \frac{Q}{\mu}$ t_i: Single battery operation time

$$\because \lambda t_2 = \mu t_3 \qquad \therefore t_3 = \frac{\lambda}{\mu} t_2 \tag{9}$$

$$\because \lambda t_3 = \mu t_4 \qquad \therefore t_4 = \frac{\lambda}{\mu} t_3 = \frac{\lambda}{\mu}\frac{\lambda}{\mu} t_2 \tag{10}$$

Following the same procedure, one can will obtain the total operation time for the dual rechargeable battery of a microcomputer system as shown in Eq (11)

$$\begin{aligned} T_{totaloperation\ time} &= \frac{Q}{\mu} + \frac{Q}{\mu} + \frac{\lambda}{\mu}\frac{Q}{\mu} + \frac{\lambda}{\mu}\frac{Q}{\mu}\frac{\lambda}{\mu} + \cdots \\ &= 2\frac{Q}{\mu} + \frac{Q}{\mu}\left(\frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \cdots\left(\frac{\lambda}{\mu}\right)^n\right) + \cdots \\ &= 2\frac{Q}{\mu} + \frac{Q}{\mu}\frac{\lambda}{\mu}\left(\frac{1}{1-\frac{\lambda}{\mu}}\right) \end{aligned} \tag{11}$$

$$= 2\ T_{single\ battery\ operation\ time} + T_{single\ battery\ operation\ time}\frac{\lambda}{\mu}\left(\frac{1}{1-\frac{\lambda}{\mu}}\right) \tag{12}$$

$$T_{total\ operation\ time} = \left(2 + \frac{\rho}{1-\rho}\right) T_{single\ battery\ operation\ time} \quad 0 \le \rho = \frac{\lambda}{\mu} < 1 \tag{13}$$

Fig. 9 shows the characteristic curve of Eq (14) that has the relationship between the ratio of the single battery operation time to the total operation time and the utilization ratio (ρ). When $\rho = 2/3$, shown at point A in Fig. 9, i. e. the charging rate is equal to two third of the discharging rate, then, from Eq (14), we can obtain an operation time four times longer than a single rechargeable battery power supply. In other words, if there is only a 0.46 hours operation time in a 24 hour interval for a single rechargeable battery, then the operation time can potentially be prolonged four times. Or the operation time can be extended to 1.84 hours in a 24 hours interval, due to the overlapping charging period for a dual rechargeable battery.

In addition, the higher the battery charging rate and utilization ratio (ρ), the longer the operation time prior recharging. Hence, the solar cells powered portable

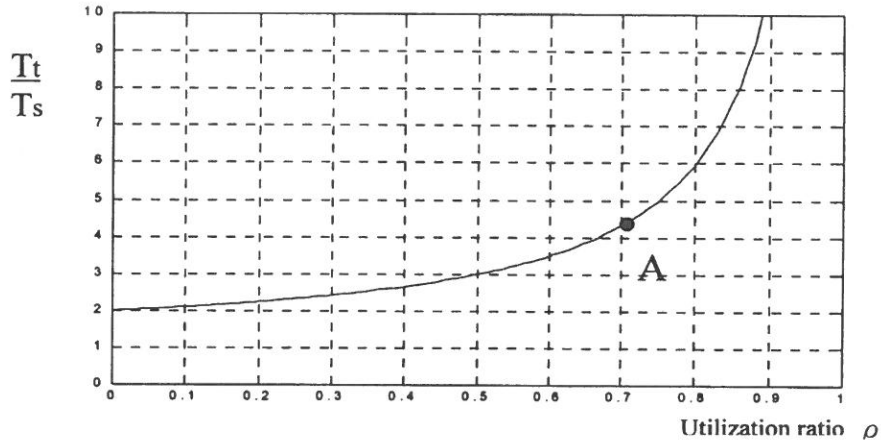

Fig. 9. The relationship between the total operation time and the utilization ratioρ

microcomputer system with dual rechargeable battery can be feasible in the near future.

$$\frac{T_{\text{total operation time}}}{T_{\text{single battery operation time}}} = \frac{T_1}{T_s} = 2 + \frac{\rho}{1-\rho} \tag{14}$$

4. The Software Design for a Microcomputer 8051 System Powered by Solar Cells

Fig. 10 shows the flowchart for controlling the switching activity between the dual batteries in a microcomputer 8051 system. The software from this flowchart is implemented by an assembly language program, which can provide input/output, arithmetic and data moving functions for the microcomputer system. In this flowchart, at the beginning, the control program initiates the microcomputer 8051 system and executes an application program and LCD display program. If the battery still has enough power, then the LCD display program will be kept in execution. If the battery power is low, then the event will generate an interruption in the microcomputer 8051 system. This system will record the CPU time or number of clock cycles used as the message for further control. In addition, the microcomputer 8051 will send out a control signal for switching the batteries. The second battery will be set to fit in and take the place of the first battery. The first battery will be switched to the

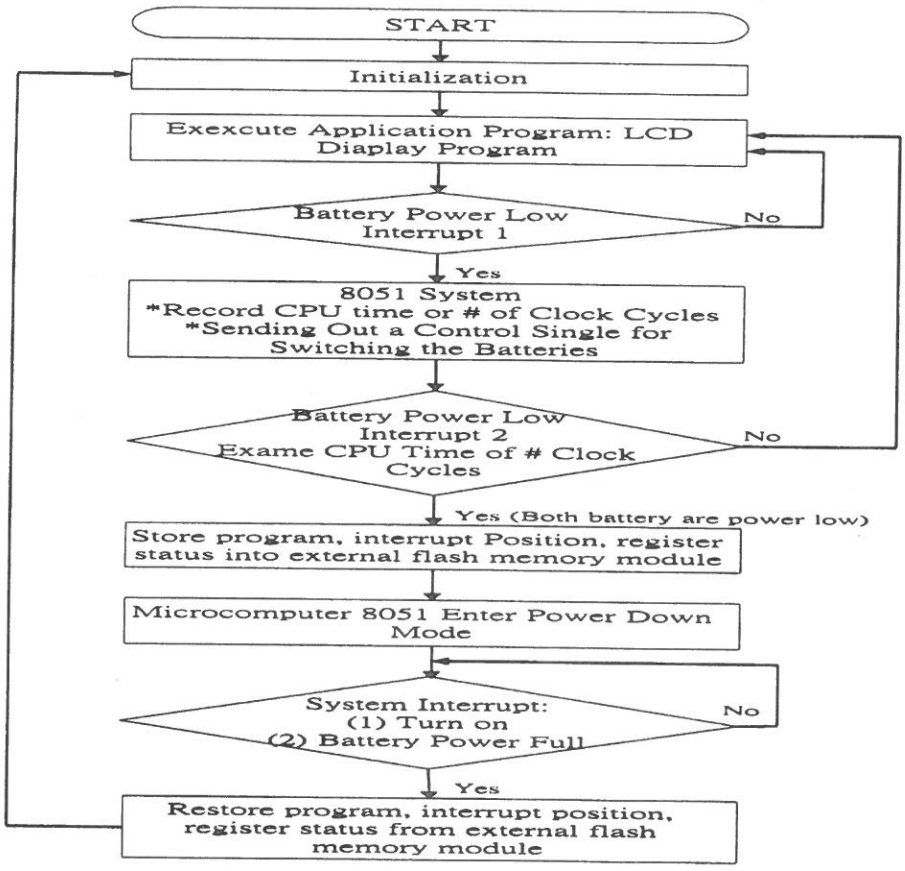

Fig. 10. The software flowchart for a microcomputer 8051 system powered by solar cells

charging state. If the power of battery 2 is low then the second interrupt happens. When both the batteries are low, the system will store the system status into an external flash memory module. The system will enter power down mode. If the user turns on the system again or the batteries are charging up again, then the microcomputer 8051 system will switch back to normal operation state. It will restore the data from the external flash memory and continue the execution that was interrupted by the power low event. Overall, the controlling mechanism will preserve the data integrity during state transition from the low battery power and charging up events.

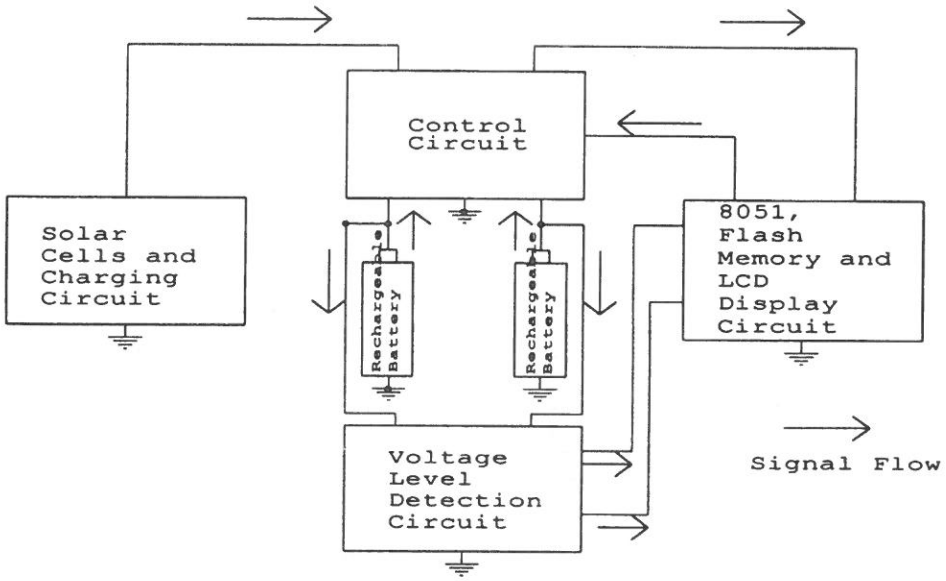

Fig. 11. The block diagram of the hardware design for the system used

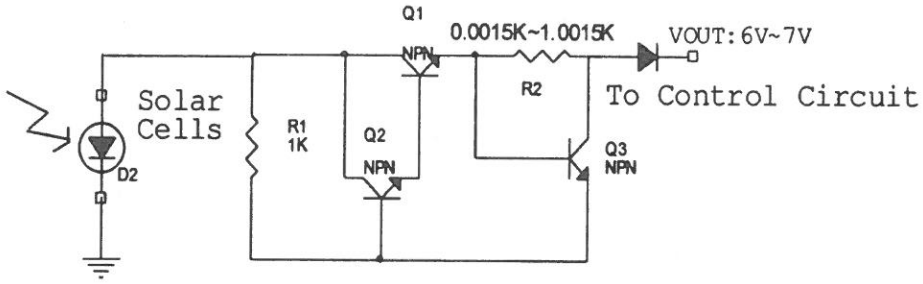

Fig. 12. Solar cells and charging circuits

5. The Hardware Design of Microcomputer 8051 System Powered by Solar Cells

Fig. 11 shows the block diagram of the hardware design of a microcomputer 8051 system powered by solar cells. Five major blocks are designed for controlling the system operation according to the software flowchart shown in Fig. 10.

Fig. 12 shows the circuit for solar cells and charging circuit.

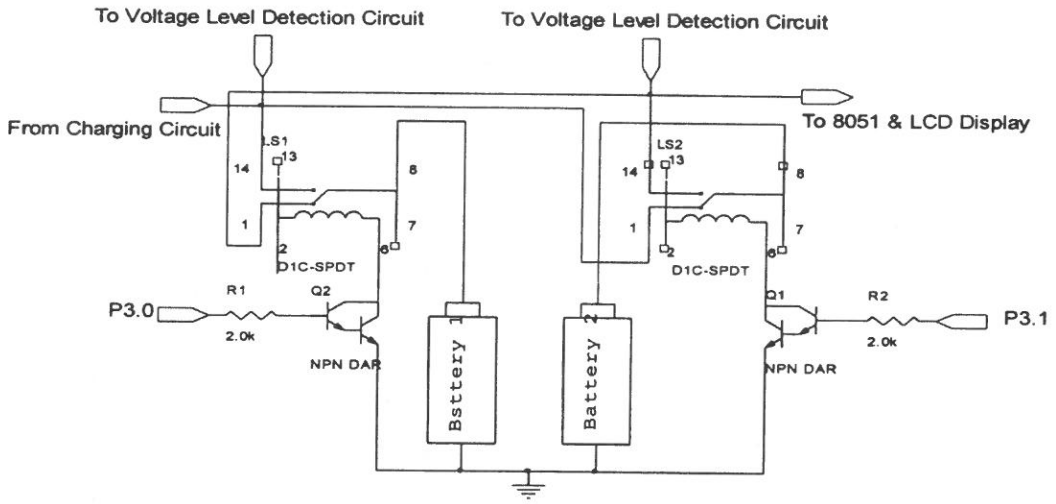

Fig. 13. Switching control circuit

The solar cells we use in this design can generate a voltage of 7V～9V and a current of 0mA～44mA. The generated voltage from solar cells is input in a voltage regulator circuit, and its output voltage is from 6V～7V and current is 0mA～40mA.

Fig. 13 shows the controlling circuit for selecting batteries as the power supply. This selecting action is based on the output to the voltage level detection circuit from battery 1 or battery 2. When the voltage level is lower than the reference voltage level, 4.5 volts, then the voltage level detection circuit will send out the interrupt signal to 8051 system. When 8051 sends control signal 0 to P3.0 and 1 to P3.1, respectively, battery 1 is at charging state by solar cells and battery 2 is used as the power supply for the microcomputer 8051 system.

Fig. 14 shows the voltage level detection circuit that can detect battery's low power and send out the interrupt signal to the 8051 system. The interrupt service routine in the 8051 system will send out the control signal to switch the battery into charging state, and then select the second battery as power supply.

The Microcomputer 8051 system is used for controlling the LCD display. In addition, it receives the signal from voltage level detection and sends out the control signal to switch the battery. If both batteries are low power, the 8051 system will enter power down state and store the current system status into an external flash

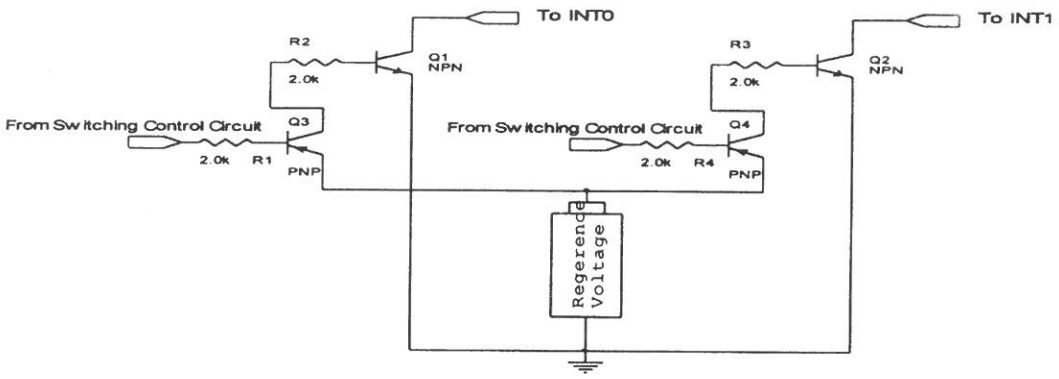

Fig. 14. Voltage Level Detection Circuit

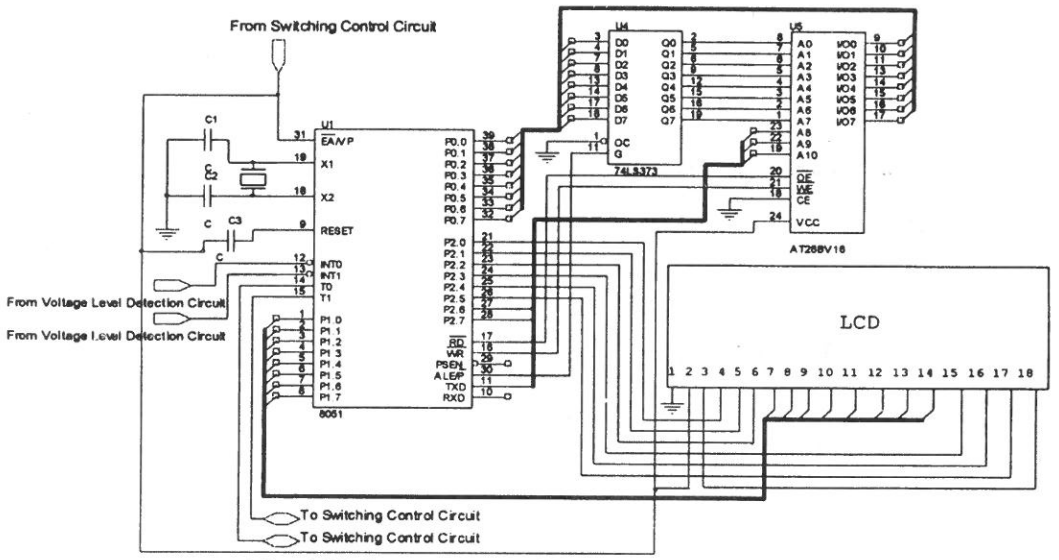

Fig. 15 Microcomputer 8051 system and LCD

Fig. 15. Microcomputer 8051 system and LCD

memory module. When the battery power charges up, the 8051 system will recover from power down state and restore the previous system status from the external flash memory module.

Fig. 16 shows the whole hardware design for microcomputer 8051 system

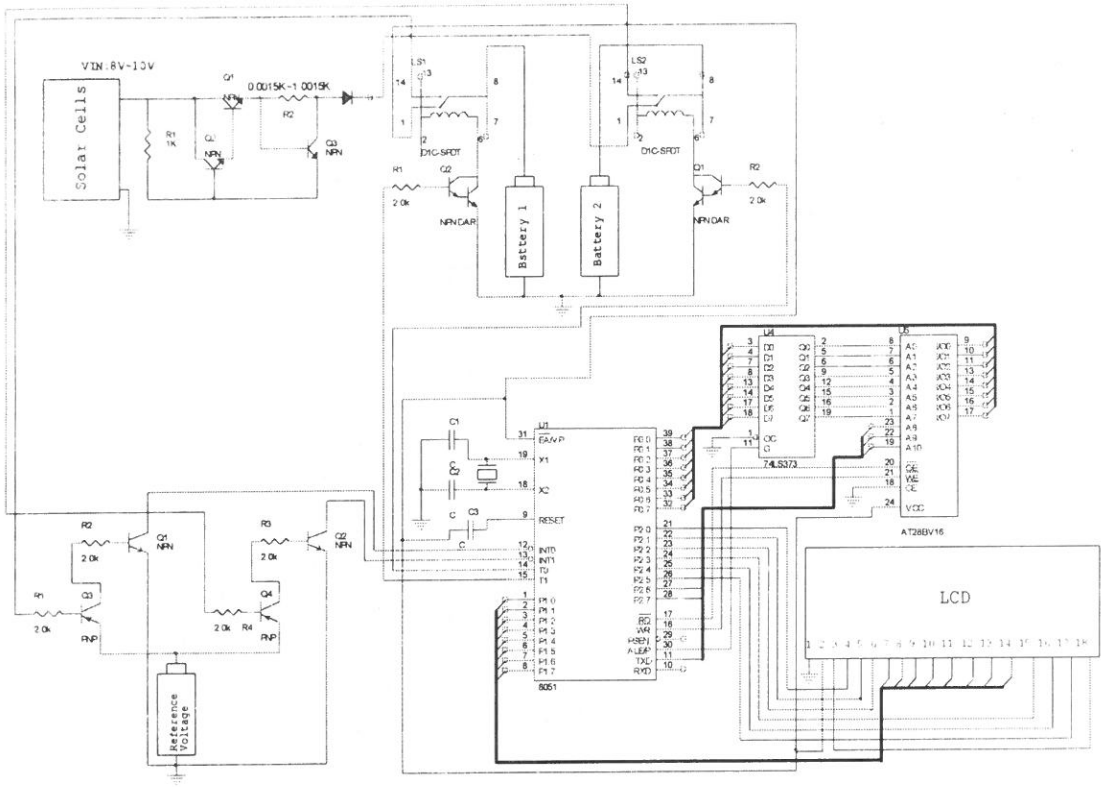

Fig. 16. The complete hardware design used in our microcomputer 8051 system

powered by dual battery charging by solar cells. In cooperation with the software flowchart shown in Fig. 10, the system can operate by itself. The experimental measurement data are recorded and shown in the next section.

6. Experimentally Measured Data

According to the hardware design shown in Fig. 16, we built the circuit board as shown in Fig. 17. The operation of our circuit board is controlled by the program based on the design from the flowchart shown in Fig. 10. The operation of our experimental platform shows that the fundamental functions are presented. These fundamental functions include, demo software operation, battery charging by solar cells, and dual battery switching due to exhaust one of two batteries.

The energy generation of solar cells that we use is from two pieces of solar cells that are not indicated in Fig. 17. The area of each solar cell is 29 cm × 15.24 cm. The

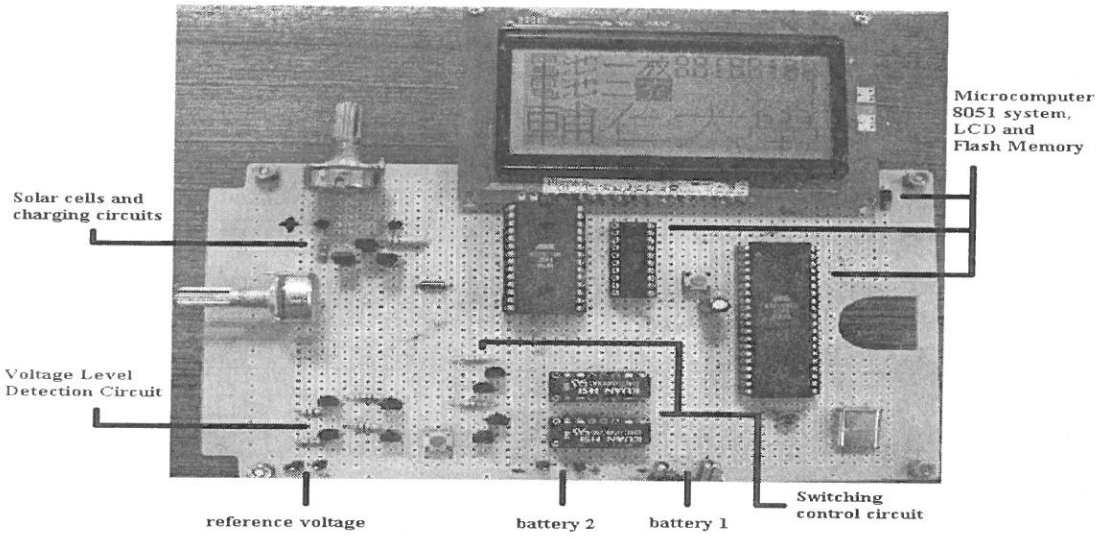

Fig. 17. The circuit board for the 8051 system powered by solar cells

solar cells are exposed to indoor light or outdoor sunshine. From the experiment, we learn that the energy generation from the outdoor light is much higher than the need of the 8051 system. Therefore, there will not be any problem of battery power exhaustion and battery switching. To demo the battery switching, we need to use the indoor light of two 60W light bulbs placed within 10cm. Under this experimental condition, we can obtain the energy generation from the solar cells that is slightly less than the need of our 8051 system. Therefore, there will be problems of battery power exhaustion and battery switching. The current generation and dissipation of each module in Fig. 17 is shown in Table 4 and the supply voltage is shown in Table 5.

Table 4. The current generation and dissipation of each module

	Solar Cells and Charging Circuits		Switching Control Circuit	Battery 1	Battery 2	Voltage Level Detection Circuit	Microcomputer 8051, LCD and Flash Memory
	Dissipation Current	Supporting Current					
Battery 1 discharge Battery 2 charge	4 mA	0~40 mA	0 mA	19.5 mA	0 mA	0 mA (Note 1)	19.5 mA
Battery 1 charge Battery 2 discharge	4 mA	0~40 mA	34.7 mA	0 mA	19.5 mA	0 mA (Note 1)	19.5 mA

Note 1: The instantaneous current during battery switching is 2.82 mA

Table 5. The voltage supply of each module

	Solar Cells and Charging Circuits		Switching Control Circuit	Battery 1	Battery 2	Voltage Level Detection Circuit	Microcomputer 8051, LCD and Flash Memory
	Before Regulator	After Regulator					
Battery 1 discharge Battery 2 charge	7.0V ~9.3V	6.24V ~7.0V	Battery 1 Supporting	5.36V	5.48V	4.72V (Note 2)	Battery 1 Supporting
Battery 1 charge Battery 2 discharge			Battery 2 Supporting				Battery 2 Supporting

Note 2: Reference voltage for voltage detection

Fig. 18 shows the relationship between the charging current and the battery voltage. When the voltages of the rechargeable battery increase, the charging current will decrease. Fig. 19 shows the relationship between the charging time and the battery voltage. When it takes more than 1000sec. charging time then the battery voltage will increase to above 5.15 volts.

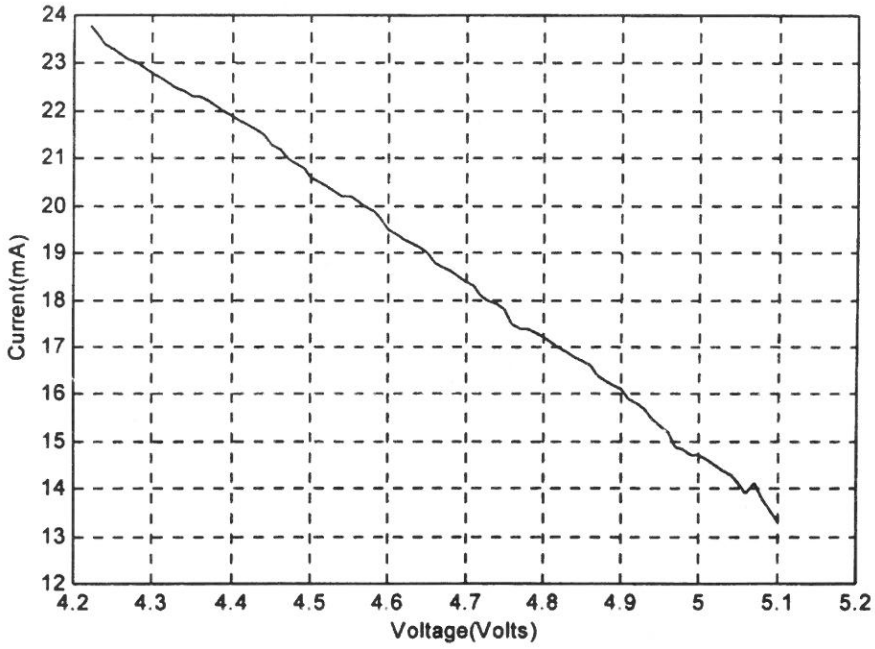

Fig. 18. Relationship between the charging current and battery voltage of rechargeable battery

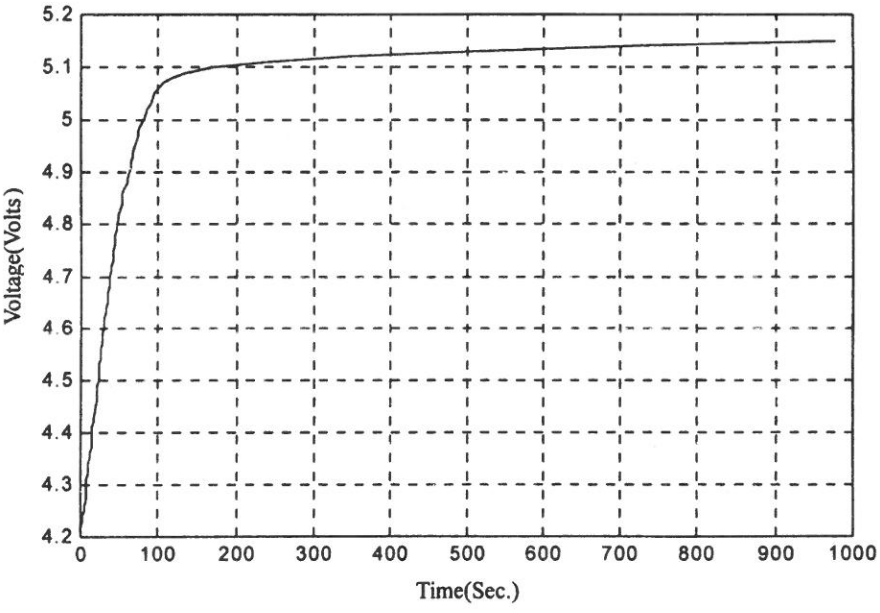

Fig. 19. Relationship between the charging time and battery voltage of rechargeable battery

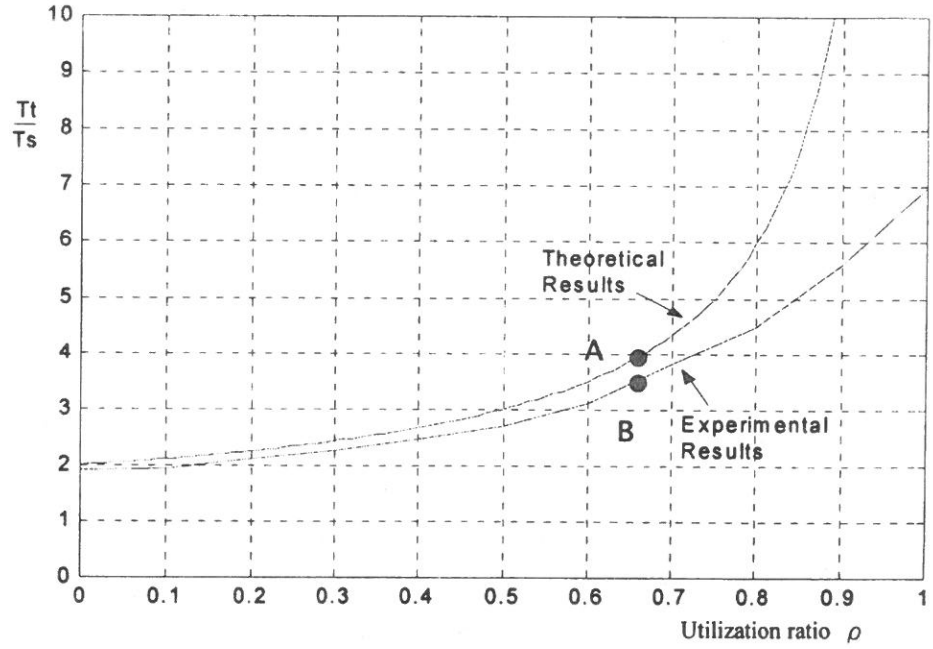

Fig. 20. The comparison of the operation time between the theoretical results and the experimental result

According to the provided experimental data, we have recorded every operation time interval of the dual battery switching and found the total operation time. Fig. 20 shows both the theoretical results from Eq (14) and experimental results from Fig. 17 for the total operation time. These results agree with each other although there is a slight difference. The difference may be from the error in the battery charging process.

7. Conclusions

The design and implementation of a microcomputer 8051 system powered by dual battery charged by solar cells has been presented. The hardware components we have used are very common and are of low cost. The control program designed uses very common assembly language. The experimental system has shown a very stable operation. The dual battery design has also shown the potential that the total operation

time can be extended more than twice as observed from the theoretical results and the experimental results. For a two third utilization ratio, we can gain four times the total operation time of the single battery design from theoretical results and three and a half times the operation time from experimental results. The difference between the theoretical results and the experimental results is from the error of the battery charging process. In addition, when our system operates at the daytime outdoors sunshine, it can work all the time without battery exhaustion because the energy generation by solar cells is larger than the energy consumption of our 8051 system.

REFERENCES

(1) E. P. Harris, S. W. Depp, W. E. Pence, S. Kirkpatrick, M. Sri-Jayantha, and R. R. Trountman, "Technology Directions for Portable Computers," Proceedings of the IEEE Vol. 83, No. 4, PP. 636-658, April 1995.

(2) J. M. C. Stork, "Technology Leverage for Ultra-Low Power Information System," Proeedings of the IEEE, Vol. 83, No. 4, PP. 607-610, April 1995.

(3) Z. J. Limnios and K. L. Gabriel, "Low-Power Electronics," IEEE Design & Test of Computer, Vol. 11, No. 4, PP. 8-13, Winter 1994.

(4) C. A. Waraick, "Trends and Limits in the " Talk Time " of Personal Communications," Proceeding of the IEEE, Vol. 83, No. 4, April 1995.

(5) J. D. Meindl, "Low Power Microelectronics: Retrospect and Prospect," Proceedings of the IEEE, Vol. 83, No. 4, April 1995.

(6) J. Nijs, S. Sivoththaman, J. Szluficik, K. De Clercg, F. Duerinckx, E. Van Kerschaever, R. Einhaus, J. Poortmans, Tom Vermeulen, R. Mertrns, "Overview of Solar Cell Technologies And Results on High Efficiency Multicrystalline Silicon Substrates," Solar Energy Materials on Solar Cell 48. PP. 199-217, 1997.

(7) A. Rohatgi, S. Narasimha, "Design, Fabrication, and Analysis of Greater Than 18% Efficiency Multicrytalline Silicon Solar Cells," Solar Energy Materials and Solar Cells 48, P. P. 187-197, 1997.

(8) Willian Sweet, "Technology 1999 Analysis & Forecast: Power & Energy," IEEE Spectrun, P. P. 62-67, January 1999.

(9) Ying-Wen Bai, "A Queueing Model for Feasibility Analysis of the Solar Cell Powered Handheld Information System, " Proceeding of IASTED International Conference Applied Modelling and Simulation, P. P. 270-275, Sep. 1999.

(10) Rajeevan Amirtharajah and Anantha P. Chandraksan, "Self-Powered Signal Processing Using Vibration-Based Power Generation," IEEE Journal of Solid-State Circuits, Vol. 33, No. 5, P. P. 687-695, May 1998.

(11) Rulnick, John, M. and Bambos, Nicholas, "Mobile Power Management for Maximun Battery Life in Wireless Communication Network," Proceeding-IEEE INFOVOM V2, PP. 443-450, 1996.

(12) P. Berg and D. Lawuence, "System Power Management for PC Card," IC Card System & Design Vol. 4, PP. 31-33, May-Jun 1994.

(13) Ying-Wen Bai and Cheng-Lun Chang, "Feasibility Analysis a Queueing Model for a Dual Rechargeable Batteries in Solar Cells Powered Mobile Computers," Proceeding of IASTED International Conference Applied Modelling and Simulation, P. P. 136-142, May. 2000.

received September 27, 2000
revised October 24, 2000
accepted November 8, 2000

微算機 8051 系統太陽能電池供電與雙充電電池之設計與建構

白 英 文　　張 政 崙
輔仁大學電子工程系

摘　　要

單晶片微算機系統在控制和資訊上的應用日益重要，如果因爲操作上的需要，有一些微算機系統可能經由充電電池供電，由於充電電池能源有限，這些系統僅能擁有非常有限的操作時間和再充電週期。因此，本論文提出微算機8051系統太陽能電池供電與雙充電電池之軟硬體設計與建構，從雙充電電池之隨機充放電程序，我們嘗試採用佇列模型進行可行性分析，其分析結果和實驗量測結果都顯示雙充電電池設計之操作時間比單充電電池設計超過兩倍以上，如果這個 微算機系統利用度爲三分之二時，雙充電電池設計之操作時間比單充電電池設計超過四倍。除此之外，現今技術趨勢顯示典型微算機系統電源消耗量正逐年下降，而太陽能電池能源產生效率正逐年提昇，因此，以太陽能電池供電與雙充電電池設計之微算機8051系統在不久的將來會成爲一種實用可行之方案。

關鍵詞：微算機，太陽能電池，佇列模型，操作時間

A Profit Evaluation System (PES) for VLSI Systems at Early Design Stage

Shyue-Kung Lu and Tsung-Ying Lee
Department of Electronic Engineering
Fu Jen Catholic University
Taipei, Taiwan 242, R.O.C.

Abstract

In this paper, we propose a Profit Evaluation System (PES) for IC designers. This system will help designers to determine the yield and test plan when specified quality level is given. Type of circuit fabric and raw manufacturing data (i.e., wafer size, wafer cost, defect density and distribution) are given for the system. The outputs of the system are the values of yield and fault coverage that will generate maximal profit. Different yield models and cost models are selectable for the users. Experimental results show that the system can find the optimal yield and test plan for generating the maximal profit.

Key Words: Profit evaluation system, Yield, Fault corerage, defect

INTRODUCTION

The profitability of integrated circuits manufacturing depends on the fabrication yield defined as the probability that a manufactured circuit chip is defect-free. Because of the rapid increase in the complexity of VLSI integrated circuits and the ever increasing quest for more functions on a single chip, the shrinkage of device geometry is reasonable. However, both of these effects have caused chips to become more susceptible to various yield loss mechanisms. This in turn increases the defect density

per unit area. Therefore, a yield of 100% is virtually unlikely. In general, low manufacturing yields also lead to high DPM (Defect Per Million) level and high level of testability is inevitable. In particular, the emphasis on the quality of shipped products makes testing an even more challenging topic. Let Y denote the manufacturing yield, DL be the defect level, that is, the probability of shipping a defective product. The fault coverage of the test used to check for manufacturing defects is denoted as FC. Williams [2] has proposed the relationship between these variables as

$$DL = 1 - Y^{1-FC} \tag{1}$$

From this equation, we can see that the quality level depends heavily on manufacturing yield (Y) and test quality (FC). In order to meet the quality constraint, we can either add some form of redundancy to improve the yield or perform high-quality test to increase fault coverage. However, defect-tolerant chips will increase the chip area and lead to less dies per wafer. Therefore, the cost-per-die will increase. Similarly, the test cost increases exponentially with the fault coverage. The problem rises here is to determine the optimum number of redundancy and test quality that will maximize the profits. In this paper, a Profit Evaluation System (PES) is proposed and implemented with C language. The analyzed results are helpful for IC designers to determine the redundancy and test plan in the early design phase. Experimental results show that our system is valuable for gaining the maximal profit.

Flowchart of Profits Evaluation System

The flowchart of our Profit Evaluation System is shown in Fig. 1. In general, the quality level (measured in DPM) is a mandatory constraint that must be met. Therefore, one of the inputs to the system is the quality level constraint (1). According to the constraint, we can estimate yield and fault coverage combinations that meet the constraint (2) by using the above equation. Let (Y_i, FC_i) be a combination that meets the quality level, $i \in N$. Our PES system makes tradeoffs between yield and fault coverage to maximize the profit function P (Y_i, FC_i). Once a combination is computed, redundancy analysis (4) is performed according to the adopted yield model (3) and the raw manufacturing data (5). The number of spares then can be determined. Similarly, the fault coverage required can be used to compute

the cost of testing according to the test cost model (6).

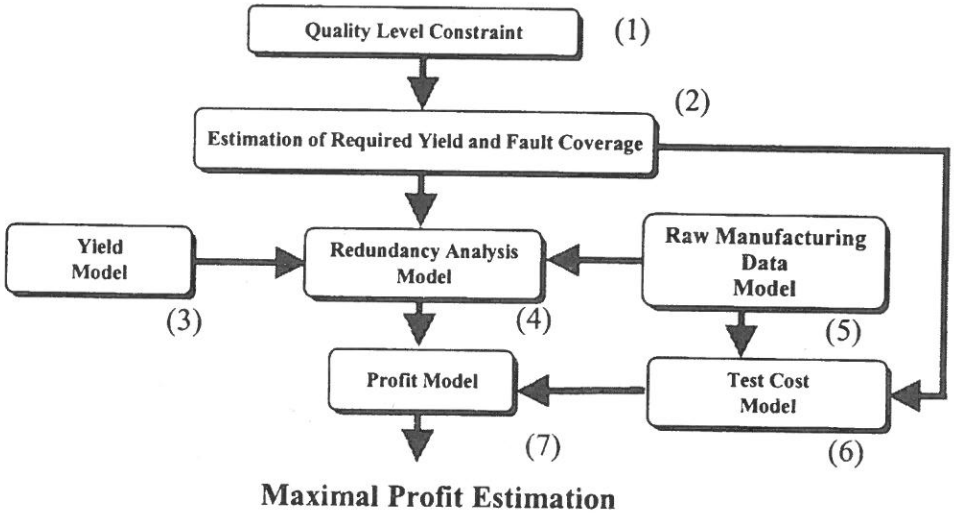

Fig. 1. Flowchart of Profit Evaluation System.

Table 1. Relative statistics and parameters.

Manufacturing Statistics (including testing data)		Circuit Parameters	
DL	Defect level	L	Chip length
Do	Defect density	W	Chip width
α	Clustering parameter	CKC	Test pattern clock cycle
α_{FC}	Fault coverage factor	FFs	No. of flip-flops
NWL	Net wafer per lot	k	Fraction of total chip occupied by modules
R	Wafer radius	m	Identical modules
SL	Scribe channel length	c	Control circuit overhead
SW	Scribe channel width		
TST	Test setup time		
WLT	Wafer loading time		
DST	Die stepping time		

Finally, profit model (7) is used to compute the profit. After finite number of iterations, we can find the optimum combination of FC and Y that will maximize profit. Table 1 list the raw manufacturing data including testing environment data and circuit parameters in our system. The details of these models used in the flowchart are illustrated in the following.

Yield Model and Redundancy Analysis

Recently, manufacturing technologies have brought out steady and significant improvement regarding contamination and defect density and have forced us to re-evaluate the economical advantages of circuits with redundancy. Many yield models have been proposed in previous works. The most common statistical yield models are the Poisson model and its derivative—the compound Poisson model. The Poisson model can be expressed as

$$Y = e^{-DA}, \tag{2}$$

where D is the fabrication defect density, and A is the critical area [2]. Since we are dealing with large area ICs, it would not be realistic to use Poisson distribution, because it has been clearly demonstrated that it does not reflect very well the yield of large chips. Most recent models are based on the well-known negative-binomial distribution. The expression for yield of a standard chip is:

$$Y = G(1 + \sigma A/\alpha)^{-\alpha}, \tag{3}$$

where G, σ, α denote the gross defect yield, the defect density, and the clustering factor, respectively.

In very large chips, if the whole chip is expected to be fault free, the yield will be very low. Therefore, spares are added to the design and accepting these chips if they have the required number of fault-free modules. To ease our discussion, we first deal with chips, which contain one type of module. In other words, we consider an idealized chip with a global area A divided in two main parts. The first part has m identical modules occupying an area kA (k = fraction of total chip occupied by modules) and the second part contains the rest of the circuit. To make the first part fault tolerant, we add "r" redundant modules and some control circuit. We define "c"

as the control circuit overhead. Then the compound Poisson yield model with redundant can be modified as [7]:

$$Y = G \cdot \left(\sum_{i=0}^{r} \binom{m+r}{i} \sum_{j=0}^{i} (-1)^j \binom{i}{j} \left(1 + \frac{(m+r-i+j)\sigma kA}{am} \right)^{-\alpha} \right) \cdot \left(1 + \frac{\sigma(1-k+c)A}{\alpha} \right)^{-\alpha} \tag{4}$$

If the required yield is computed, then the value of r can be determined. Similarly, the users can select different yield models and the level of redundancy. Algorithms proposed in [11] can evaluate the optimum number of spares in defect-tolerant ICs. In PES, the conventional algorithm is implemented to find the suitable number of redundancy.

In order to make our model more accurate, we have to evaluate Yo [6] that is the fraction of chips already working after the first wafer test. Yo can be show as:

$$Y_0 = G \left\{ \left(\sum_{i=0}^{r} \binom{r}{i} \sum_{j=0}^{i} (-1)^j \binom{i}{j} \left(1 + \frac{(m+r-i+j)\sigma kA}{\sigma m} \right)^{-\alpha} \right) \cdot \left(1 + \frac{\sigma(1-k+c)A}{\alpha} \right)^{-\alpha} \right\}. \tag{5}$$

It also follows that (Y′-Yo) represents the number of circuits that can be repaired and that need reconfiguration.

IV Test Cost Model

Once the test quality (fault coverage) is determined, test cost model is used to evaluate the cost of testing. In general, the test cost is proportional to the testing time, which is inversely proportional to the throughput. Test throughput is defined as the average number of chips processed by a testing platform during a time unit. Time model proposed in [5, 6] can be used to estimate the testing time. Test throughput is determined by many parameters. These include the *test setup time* (TST), the unit die *stepping time* (DST), the wafer *loading time* (WLT), the number of wafers per processing lot (NWL), the number of dies per wafer (N_W), the chip yield (Y), the chip testing time for a good die, T_{PASS}, and the chip testing time, T_{FAIL}, for a bad chip. In general, TST, DST, WLT, and NWL can be easily obtained for a given

test platform and test environment. T_{PASS} is proportional to the fault coverage. In our system, we first estimate the yield model and apply for the given specification (i. e., defect level). Then, we use these factors to gain fault coverage from the equation:

$$FC = 1 - log(1 - DL)/log(Y) \tag{6}$$

After knowing the fault coverage, the number of test patterns can be determined by using the following equation [6]:

$$FC = 1 - e^{-\alpha_c \times V}$$
$$V = \frac{ln(1 - Fc)}{-\alpha_c} \tag{7}$$

where V and α_c denote the test length and the increase rate, respectively. In general, we have to know the number of flip-flops (FFs) and the clock cycle (CKC) from the design. Then, the relationship of V and T_{PASS} can be expressed as [12]:

$$T_{PASS} = V \times (FFs + 1) \times CKC \tag{8}$$

Hirase [10] approximated T_{FAIL} to a quarter of T_{PASS}. When we know the test parameters used in our model, the average testing time can be estimated as [6]:

$$T_{CHIP} = \frac{TST + WLT \times NWL + DST \times N_w \times NWL + T_{PASS} \times NPD + T_{FAIL} \times NFD}{N_w \times NWL}, \tag{9}$$

where the net passed die, $NPD = N_w \times NWL \times Y$, and the net failed die, $NFD = N_w \times NWL\ (1 - Y)$. The numerator of Eq. (9) is the total testing time for a wafer lot, and the denominator is the total number of chips tested.

V Profit Model

The profit of a design can be expressed as the difference between revenue (R) and the total manufacturing cost (MC). That is, $Profit = R - MC$. Manufacturing cost counts for silicon cost, testing cost, packaging cost, and the final test cost. In Sec. IV, we estimate the total testing time. In this section, we can evaluate test cost for chips from the determined test time. We first define the test cost rate (TCR, unit:

\$ /sec), which may be changed with different test platforms. Moreover, the wafer lot cost of testing (C_{chip}) can be estimated by using T_{CHIP} and TCR. It can be expressed as:

$$C_{chip} = T_{CHIP} \times TCR. \tag{10}$$

Therefore, the total cost of testing then can be evaluated with the following equation:

$$C_T = \frac{C_{chip}}{Y_{pack} \times Y} \times N_w \times NWL, \tag{11}$$

where Y_{pack} denote the yield after packaging. The number of dice per wafer N_W is an important parameter for cost evaluation that can be expressed as [7]:

$$N_w = TRUNC\left[\left(1 - \frac{\sqrt{2 - (L + S_L)(W + S_w)\pi}}{2R}\right)\left(\frac{\gamma\pi R^2}{(L + S_L)(W + S_w)}\right)\right], \tag{12}$$

where L and W denote the die length and width, respectively. "S_L" and "S_W" denote scribe channel length and width and "R" is the wafer radius. The parameter "γ" represents the useful inner portion of a wafer. If spares are included in the fault-tolerant design, the total cost of testing then expressed as:

$$C'_T = \frac{(1 + Y' - Yo)C_{chip} + (1 - Yo)D}{Y_{pack}(Y_r(Y' - Yo) + Yo)} \times N_w' \times NWL, \tag{13}$$

where D, Y_{pack}, Y_o, N_W' and Y_r, denote the diagnostic test cost, the yield after packaging, the yield after the first wafer test, the number of fault tolerant dice per wafer, and the restructuring yield. The diagnostic test identify possible defective areas before reconfiguration. The silicon cost can be estimated from the cost of each wafer [4, 7], C_W,

$$C_s = \frac{C_w}{Y_{pack} \times Y} \times NWL \tag{14}$$

the number of wafer per lot, NWL, and the manufacturing yield. It is given by

$$C_s' = \frac{C_w}{Y_{pack}(Y_r(Y' - Y_o) + Y_o)} \times NWL \tag{15}$$

Similarly, the silicon cost for the fault-tolerant design can be modified as

Whether design with redundancies or not, the packaging cost can be expressed as [7]:

$$C_p = \frac{P}{Y_{pack}} \times N_w \times NWL, \tag{16}$$

where P denotes the packaging cost per die. To lump under on hand, the final testing cost can be estimated the same as [7]:

$$C_{FT} = \frac{Tf}{Y_{pack}} \times N_w \times NWL$$
$$Tf = (DST + T_{PASS} \times Y) \times TCR, \tag{17}$$

where Tf [5] denotes the final testing cost per die. Similarity, N_W, Y is replaced by N_W', Y' when considering design with redundancy. The manufacturing costs for either the design with or without redundant modules must count for package cost (C_P) or the final test cost (C_{FT}). That is

$$MC = C_T + C_s + C_p + C_{FT} \tag{18}$$

After knowing the manufacturing cost, MC, we evaluate the revenue, R later to compute the profit. The revenue relative to the product price, yield, and the number of passed dies is given by:

$$R = P_r \times N_W \times Y, \tag{17}$$

where P_r denote the chip price. When considering fault-tolerant designs, N_W and Y are replaced with N_W' and Y'. If the chip area is increased by adding redundancy, the number of dies, N_W', will be less than N_W. On the contrary, Y' is higher than Y by adding redundancy. According to the revenue and the manufacturing cost, the profit can be estimated as:

$$Profit = R - MC \tag{18}$$

The trade-off between the number of redundancies we add at least and the profit from yield improvement we gain the most is our system's purpose. Some experiment results will be shown in next section.

Experiment Results

A practical example with manufacturing data [6] as shown in Table 1 is analyzed

with our PES system. For practical applications, users can change these data to meet their design.

Table 1. Manufacturing Data

Name	Data
TST	2400s
WLT	30s
DST	0.1s
NWL	15
9R	15cm
Pr	$ 25
C_w	$ 3000
α	2.5

Now, we input manufacturing data and compute Y = 62.8% (without redundancy). In this case, it is assumed that $m = 32$, and the added redundancy can not be more than $TRUNC\ (\sqrt{m})$. So we can get 5 different yield levels. After input these yield values to our system, we can evaluate distinct profits (as shown in Fig. 2).

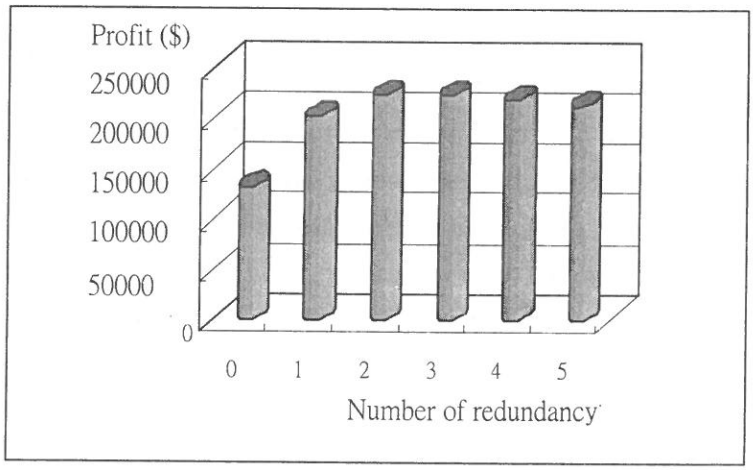

Fig. 2. The relationship between profits and redundancy.

From Fig. 2 we can find that the optimal profit does not occur at the highest yield value. This is because that the growth rate of yield may saturate after some threshold value is achieved. When beyond this value, the profit resulted from the improved yield will be offset by the silicon cost. Fig. 2 shows that the threshold value occurs when 3 redundancies are added into our design and gains the highest profit. As mentioned before, the chip area in this example is 0.3 cm^2. Fig. 3 shows the profit improvement when the chip area is reduced to 0.2 cm^2.

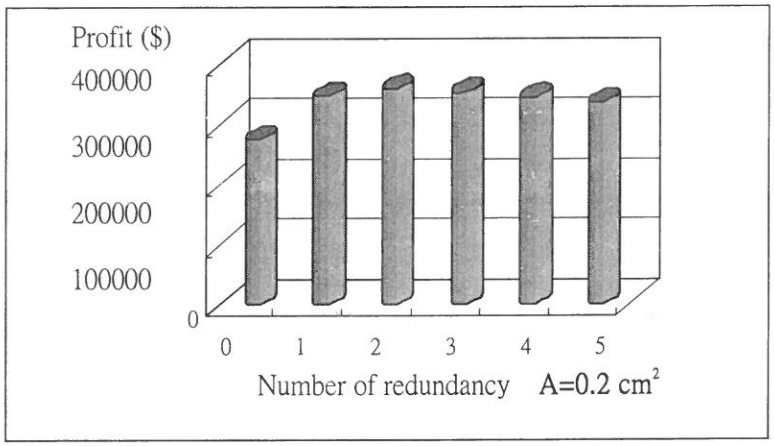

Fig. 3. The relationship between profits and redundancy.

The profits are more than that shown in Fig. 2 (because N_W increases). It is evident that profits are proportional to N_W. In fact, profit is relative to yield and fault coverage. In our system, high yield usually combines with low fault coverage to achieve specified quality level, and profits are proportional to yield but inversely proportional to fault coverage (lower testing cost). Fig. 4 shows the relationship between profit, yield, and fault coverage. With our PES system, we can find the coordinate in the (Y, FC) plane that will result in the optimum profit as shown in this figure.

CONCLUSIONS

In this paper , we propose a Profit Evaluation System for IC designers . This

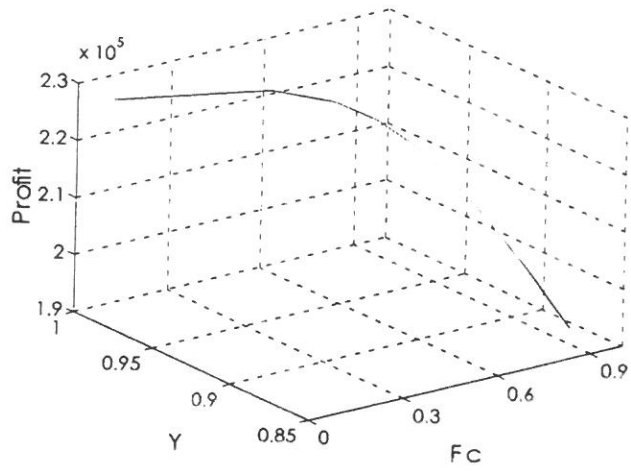

Fig. 4. Economics analysis of profit, Y, and FC

system will help designers to determine the yield and test plan when specified quality level is given. Type of circuit fabric and raw manufacturing data (i.e., wafer size, wafer cost, defect density and distribution) are sent into the system. The outputs of the system are the values of yield and fault coverage that will generate maximal profit. Different yield models can be implemented in the system. The PES system is a prototype for profit analysis of VLSI systems. Experimental results show that the system can find the optimal yield and test plan for generating the optimum profit. In the future, the building blocks for SOC (system-on-a-chip) designs may be a controller core, embedded SRAM memory, and some dedicated logic. Therefore, we will find the yield models and the cost models for these blocks to make our system suitable for SOC systems.

REFERENCES

(1) T. W. Williams and N. C. Brown, "Defect level as a function of fault coverage", *IEEE Trans. Computers*, Vol. C-30, No. 12, pp. 987-988, Dec. 1981.

(2) V. Kim, M. Tegethoff, and T. Chen, "ASIC yield estimation at early design cycle," in *Proc. Int. Test Conf. (ITC)*, pp. 590-594, 1996.

(3) A. V. Ferris-Prabhu. "Modeling of critical area in yield forecasts," *IEEE Journal of Solid State Circuits*, SC-20 (4): 874-880, Aug. 1985.

(4) G. A. Allan, and A. J. Walton, "Fast yield prediction for accurate costing of ICs," *IEEE Innovative Systems in Silicon Conference*, pp. 279-787, 1996.

(5) K. Nakamae, H. Sakamoto, and H. Fujioka, "How ATE planning affects LSI manufacturing cost." *IEEE Design & Test of Computers*, pp. 66-73, 1996.

(6) V. Kim, T. Chen, and M. Tegethoff, "ASIC manufacturing test cost prediction at early edsign stage," in *Proc. Int Test Conf. (ITC)*, pp. 356-361, 1997.

(7) Y. Gagono, Y. Savaria, M. meunier, and C. Thibeault, "Are defect-tolerant circuits with redundancy really cost-effective? Complete and realistic cost model," *IEEE International Symposium*, pp. 157-165, 1997.

(8) J. D. Lin, J. M. Lu, and C. W. Wu, "An improved VLSI test economics analysis system," in *Proc. 9th VLSI/CAD Symp.*, pp. 149-152, 1998.

(9) R. Ross, J. Bailey, N. Atchison, and M. Effron, "A comprehensive sequential yield analysis methodology and the financial payback for higher yields," *IEEE/SEMI Advanced Semiconductor Manufacturing Conf.*, pp. 80-86, 1999.

(10) J. Hirase, "Study on the Costs of on-site VLSI testing," in *Proc. Int Test Conf. (ITC)*, pp. 438-443, 1995.

(11) C. Thibeault, Y. Savaria, and J. L. Houle, "A fast Method to Evaluate the Optimum Number of Spares in Defect-Tolerant Integrated Circuits," *IEEE Trans. Computers*, vol. 43, no. 6, pp. 687-697, June 1994.

(12) S. Narayanan, and M. A. Breuer, "Reconfigurable scan chains: A novel approach to reduce test application time," *Computer-Aided Design*, 1993. *ICCAD-93*. Digest of Technical Papers., 1993 IEEE/ACM International Conference on, Page (s): 710 -715, 1993.

received September 22, 2000
revised October 27, 2000
accepted November 14, 2000

超大型積體電路初期設計階段之收益評估系統

呂學坤 李宗穎
輔仁大學電子工程系

摘 要

在這篇論文中，我們針對積體電路設計者提供了一個收益評估系統，這個系統將可以幫助設計者在設計規格的限制與一些製程參數資料的要求下計算出產品的良率與測試計劃，而我們根據此系統輸出結果的良率與錯誤函蓋率便可求得產品的最大收益。在本系統中，不同的良率模型與成本模型可以隨使用者的需要做調整。文中實驗結果也顯示本系統可以找到最好的良率與測試計劃以求得最高的收益，未來也要將此系統應用在系統單晶片的設計上。

An Incremental Updating Technique for Discovering Sequential Patterns in Large Databases

Show-Jane Yen and Chung-Wen Cho
Department of Computer Science and Information Engineering
Fu Jen Catholic University
Taipei, Taiwan 242, R.O.C.

Abstract

Mining sequential patterns is to discover sequential purchasing behaviors of most customers from a large amount of customer transactions. The strategy of mining sequential patterns focuses on discovering *frequent sequences*. A frequent sequence is an ordered list of the itemsets which purchased by a sufficient number of customers.

The previous approaches for mining sequential patterns need to repeatedly scan the database, and take a large amount of computation time to find frequent sequences, which are very time consuming. Therefore, it is important to reduce the number of database scans to improve the efficiency of those mining algorithms. Moreover, because the new transactions progress continuously and the obsolete transactions need to be removed, the discovered sequential patterns may be changed. In order to keep new information, it is very costly to periodically rediscover the sequential patterns from the updated transaction database.

In this paper, we present an algorithm *SLP* to find sequential patterns, which can significantly reduce the number of the database scans. Besides, we also propose an incremental updating algorithm *USLP* to maintain the discovered sequential patterns when the database is updated. *USLP* makes use of the discovered information and considers the updated customer transactions to find all the new sequential patterns. The experimental results show that our algorithms are more efficient than the other algorithms for mining

sequential patterns.

Key Words: Data mining, Knowledge discovery, Transaction database, Sequential pattern, Incremental updating technique

INTRODUCTION

Because the capacity of the storage is getting larger, large amount of data can be stored in the database. Potential useful information may be embedded in the large databases. Hence, how to discover the useful information exists in such databases is becoming the popular field in the computer science. The purpose of data mining [1, 2, 5, 6, 7, 12] is to discover the useful information from the large databases such that the quality of decision making can be improved.

A *transaction database* consists of a set of *transactions*. A transaction typically consists of the transaction identifier, the customer identifier (the buyer), the transaction date (or transaction time), and the items purchased in this transaction. Mining sequential patterns [2, 3, 14] is to find the sequential purchasing behaviors of most customers from a transaction database. For example, there is a sequential pattern "< {shirts, necktie} {jacket} {shoes} >70%" discovered from the transaction database in a department store, which means that seventy percent of the customers buy jacket after buying both the shirts and necktie, and then they buy shoes after buying jacket.

The definitions about mining sequential patterns are presented as follows: An *itemset* is a non-empty set of items. An itemset X is *contained* in a transaction T, if $X \subseteq T$. A sequence is an ordered list of the itemsets. A sequence s is denoted as $<s_1, s_2, \cdots, s_n>$, where s_i is an itemset. A sequence $<a_1, a_2, \cdots, a_n>$ is contained in another sequence $<b_1, b_2, \cdots, b_m>$ if there exist integers $i_1 < i_2 < \cdots < i_n$, $1 \leqq i_k \leqq m$ such that $a_1 \subseteq bi_1, \cdots, an \subseteq bi_n$, and $<a_1, a_2, \cdots, a_n>$ is a *subsequence* of sequence $<b_1, b_2, \cdots, b_m>$. A *maximal sequence* is a sequence that is not contained in any other sequence.

A *customer sequence* is the list of all the transactions of a customer, which is ordered by increasing transaction-time. A customer sequence *c supports* a sequence *s* if *s* is contained in c. The *support* for a sequence s is the number of customer sequences that supports *s*. If the support for a sequence s satisfies the user-specified minimum support threshold, then s is called *frequent sequence*. Otherwise, s is a *non-frequent sequence*. The *length* of a sequence *s* is the number of itemsets in the sequence. A sequence of length k is called a *k-sequence*, and a frequent sequence of length *k* a *frequent k-sequence*. In general, before generating the frequent sequences, we need to generate the *candidate sequences*, and scan the database to count the support for each candidate sequence to decide if it is a frequent sequence. A candidate sequence of length *k* is called a *candidate k-sequence*.

A *sequential pattern* is a frequent sequence that is not contained in any other frequent sequence, that is, a maximal frequent sequence. *Rakesh Agrawal*, and et al. [2] proposed an approach to find sequential patterns. This algorithm needs to make multiple passes over the database. For the k^{th} pass, a lot of candidate *k*-sequences need to be counted to generate all the frequent k-sequences. The algorithm *DSG* (*Direct Sequential pattern Generation*) [14] needs to scan the database only once to generate all the frequent sequences. For this database scan, *DSG* records the related information, constructs an association graph, and then all the frequent sequences can be generated by traversing the constructed association graph. However, when the database is getting larger, the related information may not fit in the main memory. In this paper, we present an efficient algorithm *SLP* to discover all the sequential patterns, which can significantly reduce the number of database scans.

Because the business is going, the new transactions are generated anytime and some obsolete transactions need to be removed to keep the freshest information. The previous algorithms for mining sequential patterns [2, 3, 14] do not consider the situation described above. Thus all the sequential patterns need to be re-discovered by applying those algorithms when the database is updated. However, the action of the database updation is frequent. It is costly to re-find all the sequential patterns. Hence, we present an incremental updating technique to discovering sequential patterns when the database is updated. This technique uses the discovered information and consider

the updated part of the database to discover all the sequential patterns.

The rest of this paper is organized as follows: Section 2 presents the algorithm to discover sequential patterns, and section 3 presents the algorithm of incremental updating for mining sequential patterns. The performance evaluation and the experimental results are shown in section 4. Finally, we conclude this paper and present directions for future research in section 5.

2. Mining Sequential Patterns

In this section, we present our approach for mining sequential patterns. We refer to [2] to decompose our approach into five phases:

1. **Sort Phase.** The transaction database-D (e. g., Table 1.) is ordered by customer identifier (CID), in which the transactions are ordered by increasing transaction time, to form a customer-oriented transaction database-Db (e. g., Table 2.).
2. **Fitemset Phase. In this phase, all the frequent 1-sequences are generated from Db.**

Table 1. The transaction database (D)

TID	CID	Itemset	TID	CID	Itemset
1	1	{a, b}	8	2	{b}
2	3	{a, b}	9	3	{a}
3	2	{a}	10	3	{b}
4	4	{a, c}	11	3	{c}
5	1	{a}	12	2	{a, b, c}
6	1	{b}	13	4	{c}
7	4	{b, c}			

Table 2. The customer-oriented transaction database (Db)

CID	Customer sequence
1	{a, b} {a} {b}
2	{a} {b} {a, b, c}
3	{a, b} {a} {b} {c}
4	{a, c} {b, c} {c}

3. **Transformation Phase.** In this phase, each frequent 1-sequence is encoded and the customer-oriented transaction database *Db* is transformed. Each transaction in *Db* is transformed as all the frequent 1-sequences contained in this transaction, and each frequent 1-sequence is replaced by its code. If none of the frequent 1-sequence is contained in a customer sequence, then this customer sequence is removed. For example, consider Db in Table 2, and assume that the minimum support is 3. From the preceding phase, we can find all the frequent 1-sequences $<\{a\}>$, $<\{b\}>$, $<\{c\}>$ and $<\{a,b\}>$, and their codes are shown in Table 3. Finally, Db is transformed into a transformed transaction database-DB which is shown in Table 4.
4. **Sequence Phase.** From the transformed transaction database-DB, all the frequent sequences are generated in this phase.
5. **Maximal Phase.** This phase generates all the sequential patterns. If a frequent sequence s is not contained in any other frequent sequence, then this sequence is a maximal frequent sequence, that is, a sequential pattern.

Table 3. Mapping table.

Frequent 1-sequence	Support	Code
a	4	A
b	4	B
c	3	C
a, b	3	D

Table 4. The transformed transaction database

CID	Transformed customer sequence
1	$<\{A,B,D\}\{A\}\{B\}>$
2	$<\{A\}\{B\}\{A,B,C,D\}>$
3	$<\{A,B,D\}\{A\}\{B\}\{C\}>$
4	$<\{A,C\}\{B,C\}\{C\}>$

The major work of mining sequential patterns is to find all the frequent sequences. Hence, in the Sequence Phase, we present an efficient algorithm SLP to generate all the frequent sequences from a transformed transaction database.

2. 1 Algorithm SLP (Smallest and Largest Position)

Let L_k be the set of the frequent k-sequences, and C_k be the set of the candidate k-sequences.

Definition 1.

Let a sequence s = $<s_1, s_2, \cdots, s_n>$, and a customer sequence $c = <c_1, c_2, \cdots, c_m>$. If there exists $i1 < i_2 < \cdots < in$, $1 \leqq i_k \leqq$m, such that $s_1 \subseteq ci_1, \cdots, s_n \subseteq ci_n$, then the *position* of s in c is i_n. Otherwise, the position is 0.

For example, consider Table 4. The positions of 1-sequence $<\{A\}>$ in customer sequence 3 are 1 and 2, the position of 1-sequence $<\{C\}>$ is 4, and the positions of 2-sequence $<\{A\}\{B\}>$ is 3 and 3.

Definition 2.

The *position pair* of a sequence s in a customer sequence c is denoted as (F, L), where F and L are the smallest position and the largest position of s in c, respectively. If s is not contained in c, then the position pair of s in c is $(0, 0)$. If the position of s in c is p, then the position pair of s is (p, p), where $1 \leqq p \leqq$the number of the transactions in c.

For example, consider Table 4. The position pair of 1-sequence $<\{A\}>$ in customer sequence 3 is $(1, 2)$ and the position pair of 2-sequence $<\{A\}\{B\}>$ is $(3, 3)$.

SLP scans the database pass by pass and each pass includes three steps. In the $(k/2)^{th}$ $(k = 2, 4, \cdots)$ pass, *SLP* generates L_k and L_{k+1}. In the following, we describe the three steps for the $(k/2)^{th}$ pass $(k = 2, 4, \cdots)$ of the *SLP* algorithm, and use the transformed transaction database *DB* in Table 4 with minimum support = 3 as a running example throughout this paper.

Step 1. Generate candidate k-sequences C_k

SLP generates candidate k-sequences from the frequent $(k\text{-}1)$ -sequences that are generated in the previous pass. For every two frequent $(k\text{-}1)$ -sequences $s_1 = <a_1, \cdots, a_{k\text{-}2}, a_{k\text{-}1}>$ and $s_2 = <b_1, \cdots, b_{k\text{-}2}, b_{k\text{-}1}>$, if $a_1 = b_1, \cdots, a_{k\text{-}2} = b_{k\text{-}2}$ then the two candidate k-sequences $<a_1, \cdots, a_{k\text{-}2}, a_{k\text{-}1}, b_{k\text{-}1}>$ and $<a_1, \cdots, a_{k\text{-}2}, b_{k\text{-}1}, a_{k\text{-}1}>$ are generated.

After generating candidate k-sequences C_k, if C_k is not empty, then the candidates in C_k need to be further pruned. Otherwise, SLP terminates.

Lemma 1. *If a candidate k-sequence s is a frequent sequence, then all the* $(k\text{-}1)$*-subsequences of s must be frequent* $(k\text{-}1)$*-sequences.*

According to Lemma 1, if one of the $(k\text{-}1)$ -subsequences of the candidate k-sequence s is not in $L_{k\text{-}1}$, then s is pruned from C_k. For example, from Table 4, we can find $L_1 = \{<A>, <B>, <C>, <D>\}$ in the Fitemset Phase. The set C_2 of the candiate 2-sequences generated from L_1 is: $\{<\{A\}\{B\}>, <\{A\}\{C\}>, <\{A\}\{D\}>, <\{B\}\{A\}>, <\{B\}\{C\}>, <\{B\}\{D\}>, <\{C\}\{A\}>, <\{C\}\{B\}>, <\{C\}\{D\}>, <\{D\}\{A\}>, <\{D\}\{B\}>, <\{D\}\{C\}>\}$.

Step 2. Scan the transformed transaction database DB to record the position pair for each candidate k-sequence in each customer sequence, and count the support for each candidate k-sequence to generate Lk.

In this step, *SLP* scans *DB* to record the position pair of each candidate k-sequence in each customer sequence, and count the support for each candidate k-sequence. If the support for a candidate k-sequence is no less than the minimum support, then the candidate k-sequence is a frequent k-sequence.

For example, the position pairs of candidate 2-sequences in each customer sequence in Table 4 are shown in Table 5, in which the candidates with underline are frequent 2-sequences.

Step 3. Generate C_{k+1} from L_k and compute the simple position of each candidate in C_{k+1} to generate L_{k+1}.

Definition 3.

Let a sequence $s = <a_1, \cdots, a_k>$, and the two $(k\text{-}1)$ -subsequences of s be $s_1 = <a_1, \cdots, a_{k-2}, a_{k-1}>$ and $s_2 = <a_1, \cdots, a_{k-2}, a_k>$. Assume that the position pairs of s_1 and s_2 in the customer sequence c are (f_1, l_1) and (f_2, l_2), respectively. The *simple position* of s in the customer sequence c is denoted as $Sim_s(c)$, where

$$Sim_s(c) = \begin{cases} 0, if\ f_1 \geq l_2 \\ l_2, \ if\ f_1 < l_2\ and\ f_1 \neq 0,\ f_2 \neq 0 \end{cases} \quad (1)$$

Consider Table 4. The position pair of $<\{A\}\{B\}>$ in customer sequence 4 is (2, 2), and the position pair of $<\{A\}\{C\}>$ is (2, 3). Hence, the simple position of $<\{A\}\{B\}\{C\}>$ in customer sequence 4 is 3.

Table 5. The position pairs and the supports for candidate 2-sequences.

Candidate 2-sequences	Position pairs				support
	Customer sequence 1	Customer sequence 2	Customer sequence 3	Customer sequence 4	
$<\{A\}\{B\}>$	(3, 3)	(2, 3)	(3, 3)	(2, 2)	4
$<\{A\}\{C\}>$	(0, 0)	(3, 3)	(4, 4)	(2, 3)	3
$<\{A\}\{D\}>$	(0, 0)	(3, 3)	(0, 0)	(0, 0)	1
$<\{B\}\{A\}>$	(2, 2)	(3, 3)	(2, 2)	(0, 0)	3
$<\{B\}\{C\}>$	(0, 0)	(3, 3)	(4, 4)	(3, 3)	3
$<\{B\}\{D\}>$	(0, 0)	(4, 4)	(0, 0)	(0, 0)	1
$<\{C\}\{A\}>$	(0, 0)	(0, 0)	(0, 0)	(0, 0)	0
$<\{C\}\{B\}>$	(0, 0)	(0, 0)	(0, 0)	(2, 2)	1
$<\{C\}\{D\}>$	(0, 0)	(0, 0)	(0, 0)	(0, 0)	0
$<\{D\}\{A\}>$	(2, 2)	(0, 0)	(2, 2)	(0, 0)	2
$<\{D\}\{B\}>$	(3, 3)	(0, 0)	(3, 3)	(0, 0)	2
$<\{D\}\{C\}>$	(0, 0)	(0, 0)	(4, 4)	(0, 0)	1

Table 6. The simple positions and the supports for candidate 3-sequences.

Candidate 3-sequences	Simple positions				support
	Customer sequence 1	Customer sequence 2	Customer sequence 3	Customer sequence 4	
<{A}{B}{C}>	0	3	4	3	3
<{B}{A}{C}>	0	0	1	0	1

Lemma 2. *A customer sequence c supports a sequence s if and only if the simple position of s in c is not equal to* 0.

Rationale:

Suppose that c supports $s = <a_1, \cdots, a_k>$. Hence, c supports the two $(k\text{-}1)$ -subsequences of s: $s_1 = <a_1, \cdots, a_{k\text{-}2}, ak\text{-}1>$ and $s_2 = <a_1, \cdots, a_{k\text{-}2}, a_k>$. Assume that the position pairs of s_1 and s_2 in the customer sequence c are (f_1, l_1) and (f_2, l_2), respectively. If the simple position of s in c is 0, then the position f_1 is greater than or equal to the position l2 according to Definition 3. This means that all the positions of s_2 in c are less than all the positions of s_1. Hence, s is not contained in c, that is, c does not support s. It is contradictory to the supposition. So, the simple position of s corresponding to c is not equal to 0.

For example, *SLP* generates $C_3 = \{< \{A\}\ \{B\}\ \{C\} >, < \{B\}\ \{A\}\ \{C\} >\}$ from L_2, and computes the simple position of each candidate 3-sequence in each customer sequence in Table 4. According to Lemma 2, if the simple position of a candidate sequence s in c is not equal to 0, then c supports s. Hence, the supports of all the candidate 3-sequences can be found by computing the simple positions, and the frequent 3-sequence $< \{A\}\ \{B\}\ \{C\} >$ can be generated. Table 6 shows the simple position of each candidate 3-sequences in each customer sequence, which with underline are frequent 3-sequences.

Finally, *SLP* terminates because there is no candidate 4-sequence generated in our example.

3. Incremental Updating for Mining Sequential Patterns

In this section, we describe our incremental updating approach for mining sequential patterns. We decompose our approach into five phases:

Table 7. The new transactions

TID	CID	itemset
14	1	{b, c}
15	3	{d}
16	1	{d}
17	5	{a, b, d}

1. **InsDel Phase.** In this phase, the new transactions are inserted into the original transaction database-D, and the obsolete transactions which are deleted from D are marked. After the updating, we have the updated transaction database D'.
2. transaction database-D' is ordered by customer identifier (CID) in which the transactions are ordered by increasing transaction time to form a customer-oriented transaction database-Db'. For example, consider Table 1. The two transactions TID 1 and TID 2 are marketed (deleted), and the transactions TID14TID17 in Table 7 are inserted into Table 1. The customer-oriented transaction database-Db' is shown in Table 8, where the transactions with dotted underline are the deleted transactions.
3. **ReFitemset Phase.** In this phase, we discover all the frequent 1-sequences from Db'. For example, assume that the minimum support is 3, and all the frequent 1-sequences found from Table 8 are shown in Table 9

Table 8. The customer-oriented transaction database (Db')

CID	Customer sequence
1	{a, b} {a} {b} {b, c} {d}
2	{a} {b} {a, b, c}
3	{a, b} {a} {b} {c} {d}
4	{a, c} {b, c} {c}
5	{a, b, d}

Table 9. all the frequent itemsets in Db'

frequent 1-sequences	support
{a}	5
{b}	5
{c}	4
{d}	3
{b, c}	3

4. **Retransformation Phase.** In this phase, each frequent 1-sequence is encoded and the customer-oriented transaction database (*Db'*) is transformed. If the frequent 1-sequence is a frequent 1-sequence before the database being updated, the original code is remained. Otherwise, the frequent 1-sequence is encoded and its code cannot appear in the original mapping table. Finally, the 1-sequence in the original mapping table, which turn out to be a non-frequent 1-sequence is removed. For the above example, the original mapping table is shown in Table 3, and the new mapping table is shown in Table 10, in which the 1-sequence {a, b} in Table 3 is removed since 1-sequence {a, b} turns out to be a non-frequent 1-sequence after updating the database.

After encoding each frequent 1-sequences, the customer-oriented transaction database Db' is transformed into the *transformed updated database DB'*. The method of the transformation is the same as the method described in the Transformation Phase for mining sequential patterns in section 2. For example, Table 8 can be transformed into Table 11, where the inserted transactions also are marked with underline.

Besides, we decompose *DB'* into two parts. The customer sequences, which do not contain the inserted or deleted transactions, are included in the first part which is called unchanged part. The other customer sequences, which contain the inserted or deleted transactions, are included in the second part called *changed part*.

During the database decomposition, the frequent 1-sequences in the changed part but not in the unchanged part, which are not the frequent 1-sequences before the database updating, are put in the set Bseed. For example, after the database decomposition, the decomposed updated database DDB' is shown in Table 12 and B_{seed}

$= \{<\{E\}>\}$.

Table 10. The mapping table after updating the transaction database.

Frequent itemset	Codes
{a}	A
{b}	B
{c}	C
{d}	E
{b, c}	F

Table 11. The complete updated database-DB'

CID	Customer sequence
1	<{A}{B}>,<{A}>,<{B}>,<{B}{C}{F}>,<{E}>
2	<{A}>,<{B}>,<{A}{B}{C}{F}>
3	<{A}{B}>,<{A}>,<{B}>,<{C}>,<{E}>
4	<{A}{C}>,<{B}{C}{F}>,<{C}>
5	<{A}{B}{E}>

Table 12. The complete updated database DB' after the decomposition

CID	Customer sequence	CID	Customer sequence
2	<{A}>,<{B}>,<{A}{B}{C}{F}>	1	<{A}{B}>,<{A}>,<{B}>,<{B}{C}{F}>,<{E}>
4	<{A}{C}>,<{B}{C}{F}>,<{C}>	3	<{A}{B}>,<{A}>,<{B}>,<{C}>,<{E}>
		5	<{A}{B}{E}>
Unchanged part		Changed part	

5. **Resequence Phase.** From the decomposed updated database-DDB', all the frequent sequences are generated in this phase.
6. **Remaximum Phase.** This phase generates all the sequential patterns.

The major work of the incremental updating for mining sequential patterns is to find all the frequent sequences after updating the transaction database. Hence, in the

Resequence Phase, we present an efficient algorithm *USLP* to generate all the frequent sequences from an updated transaction database.

3.1 Algorithm USLP (Updating of Smallest and Largest Position)

Let be the set of the frequent k-sequences, and be the set of the candidate k-sequences after updating the database.

USLP scans the database pass by pass. For the $(k/2)^{th}$ ($k = 2, 4, \cdots$) pass, *USLP* generates L_k' and L_{k+1}'. In the following, we use Table 12 as our example database throughout this section and describe the steps for the $(k/2)^{th}$ pass ($k = 2, 4, \cdots$) of *USLP*.

Step 1. Generate C_k' from L_{k-1}'.

This step is the same as the step 1 of *SLP*. For example, from Table 11, we can find $L_1' = \{<A>, <B>, <C>, <E>, <F>\}$ in the Retransformation Phase. The candidate 2-sequences generated from L_1' are $<\{A\}\{B\}>, <\{A\}\{C\}>, \cdots$, and $<\{F\}\{E\}>$.

Step 2. Decompose C_k' into $ClassA_k$, $ClassB_k$, and $ClassC_k$, respectively.

Definition 4.

$ClassA_k = \{Candi \mid Candi \in C_k' \cap L_k\}$

$ClassB_k = \{Candi \mid Candi \in C_k',\ Candi\ at\ least\ contains\ a\ 1\text{-}sequence\ x \in B_{seed}.\}$

$ClassC_k = \{Candi \mid Candi \in C_k';\ Candi \notin ClassA_k,\ ClassB_k\}$

Because each candidate k-sequence s $ClassA_k$ was a frequent sequence before updating the database, we need to check if it is still a frequent sequence after updating the database. In the unchanged part of *DDB'*, the number of the customer sequences which support a sequence s is unchanged after updating the database. Hence, we only need to scan the changed part of *DDB'* to determine the supports for the candidate sequences in $ClassA_k$.

For each candidate k-sequence $s \in ClassB_k$, we also only need to scan the changed part of *DDB'* to determine if s is a frequent sequence, since s is not contained in the unchanged part of *DB'*.

For each candidate k-sequence $s \in ClassC_k$, we need to scan all the customer sequences in DDB' to determine the support of s, because s may be contained in the changed part and unchanged part of DDB'. According to Definition 4, we can see that $C_k' = ClassA_k \cup ClassB_k \cup ClassC_k$, and $ClassA_k \cap ClassB_k = ClassA_k \cap ClassC_k = ClassB_k \cap ClassC_k = \phi$

For above example, before updating the database, the frequent 2-sequences $L_2 = \{<\{A\}\{B\}>, <\{A\}\{C\}>, <\{B\}\{A\}>, <\{B\}\{C\}>\}$. From Table 12, we can obtain $B_{seed} = \{<\{E\}>\}$. Thus, according to Definition 4, C_2' can be decomposed into: $ClassA_2 = \{<\{A\}\{B\}>, <\{A\}\{C\}>, <\{B\}\{A\}>, <\{B\}\{C\}>\}$, $ClassB_2 = \{<\{A\}\{E\}>, <\{B\}\{E\}>, <\{C\}\{E\}>, <\{E\}\{A\}>, <\{E\}\{B\}>, <\{E\}\{C\}>, <\{E\}\{F\}>, <\{F\}\{E\}>\}$, $ClassC_2 = \{<\{A\}\{F\}>, <\{B\}\{F\}>, <\{C\}\{F\}>, <\{F\}\{A\}>, <\{F\}\{B\}>, <\{F\}\{C\}>\}$.

Step 3. Scan the changed part of DDB', and record the position pairs of each candidate k-sequence in each customer sequence.

In the following, we first redefine the position for a sequence.

Definition 5.

Let sequence $s = <s_1, s_2, \cdots, s_n>$ and customer sequence $c = <c_1, c_2, \cdots, c_m>$. Suppose s is contained in c and there exists $i_1 < i_2 < \cdots < i_n$, $1 \leqq i_k \leqq m$, such that $s_1 \subseteq ci_1, \cdots, s_n \subseteq ci_n$. If ci_1 is not contained in the deleted transactions of c, then the *position* of s in c is $+i_n$. Otherwise, the position of s in c is $-i_n$. If s is not contained in c, then the position of s in c is 0. For example, consider Table 12, the positions of 1-sequence $<\{A\}>$ in customer sequence 3 are -1 and +2, the position of $<\{C\}>$ is +4, and the positions of 2-sequence $<\{A\}\{B\}>$ are -3 and +3.

Definition 6.

Given a subsequence r of the customer sequence c, the *position pair* of s at r in c is denoted as (F, L), where F and L are the smallest position and the largest position of s at r in c, respectively. If s is not contained in c, then the position pair of s at r in c is $(0, 0)$. If the position of s at r in c is p, then the position pair of s at r in c is (p, p). If the positions of s at r in c are p, q, $|p| = |q|$, and $p > q$,

then the position pair of s is (p, p).

For example, consider Table 12. Suppose the subsequence r of customer sequence 3 is ranging from transaction 1 to transaction 4, that is $\{<\{A\}\{B\}>, <\{A\}>, <\{B\}>, <\{C\}>\}$. The position pair of 1-sequence $<\{A\}>$ at r in customer sequence 3 is (-1, +2), and the position pair of 2-sequence $<\{A\}\{B\}>$ is (+3, +3).

Table 13. The related information of candidate 2-sequences.

Candidate 2-sequence	Position pairs			Support
	CID 1	CID 3	CID 5	
< {A} {B} >	(+3, +3) (+3, +4)	(+3, +3) (+3, +3)	(0, 0) (0, 0)	2
< {A} {C} >	(0, 0) (+4, +4)	(+4, +4) (+4, +4)	(0, 0) (0, 0)	2
< {B} {A} >	(-2, -2) (0, 0)	(-2, -2) (0, 0)	(0, 0) (0, 0)	0
< {B} {C} >	(0, 0) (+4, +4)	(+4, +4) (+4, +4)	(0, 0) (0, 0)	2
< {A} {F} >	(0, 0) (+4, +4)	(0, 0) (0, 0)	(0, 0) (0, 0)	1
< {B} {F} >	(0, 0) (+4, +4)	(0, 0) (0, 0)	(0, 0) (0, 0)	1
< {C} {F} >	(0, 0) (0, 0)	(0, 0) (0, 0)	(0, 0) (0, 0)	0
< {F} {A} >	(0, 0) (0, 0)	(0, 0) (0, 0)	(0, 0) (0, 0)	0
< {F} {B} >	(0, 0) (0, 0)	(0, 0) (0, 0)	(0, 0) (0, 0)	0
< {F} {C} >	(0, 0) (0, 0)	(0, 0) (0, 0)	(0, 0) (0, 0)	0
< {A} {E} >	(5, 5)	(5, 5)	(0, 0)	2
< {B} {E} >	(5, 5)	(5, 5)	(0, 0)	2
< {C} {E} >	(5, 5)	(5, 5)	(0, 0)	2
< {E} {A} >	(0, 0)	(0, 0)	(0, 0)	0
< {E} {B} >	(0, 0)	(0, 0)	(0, 0)	0
< {E} {C} >	(0, 0)	(0, 0)	(0, 0)	0
< {E} {F} >	(0, 0)	(0, 0)	(0, 0)	0
< {F} {E} >	(5, 5)	(0, 0)	(0, 0)	1

For each candidate k-sequence $s \in (ClassA_k \cup ClassC_k)$, two position pairs in each customer sequence c in the changed part of DDB' are recorded. For the first position pair, we record the position pair of s at r_1 in c, where r_1 is ranging from the first transaction to the last transaction before updating the database. For the second position pair, we record the position pair of s at r_2 in c where r_2 is ranging from the first transaction to the last transaction after updating the database. For example,

consider the customer sequence 1 in Table 12, where r_1 is $<\{A,B\}\{A\}\{B\}>$, and r_2 is $<\{A\}\{B\}\{B,C,F\}\{E\}>$. Thus, the two position pairs of the candidate 2-sequence $<\{A\}\{B\}>$ in customer sequence 1 are $(+3, +3)$ and $(+3, +4)$, respectively.

For each candidate k-sequence $s \in ClassB_k$, we just record a position pair at r_3 in each customer sequence c in the changed part of DDB', where r_3 is ranging from the first transaction to the last transaction after updating the database. For example, consider Table 12. The candidate 2-sequence $<\{A\}\{E\}> \in ClassB_2$, and the position pair of $<\{A\}\{E\}>$ in customer sequence 1 is $(+5, +5)$.

Hence, in the changed part of DDB' scan, the position pairs of each candidate k-sequence in each customer sequence is recorded. For the above example, *USLP* scans the changed part in Table 12 and records the position pairs of each candidate 2-sequence in each customer sequence. Table 13 shows the related information about candidate 2-sequences after scanning the changed part. In Table 13, we use double line to distinguish the candidate sequences that belong to different classes.

Table 14. The increase count and the decrease count of some candidates $\in (ClassA_2 \cup ClassC_2)$

Candidate 2-sequence	Increase count	Decrease count
$<\{A\}\{B\}>$	0	0
$<\{A\}\{C\}>$	0	0
$<\{B\}\{A\}>$	0	2
$<\{B\}\{C\}>$	1	0
$<\{A\}\{F\}>$	1	0
$<\{B\}\{F\}>$	1	0

For each candidate k-sequence $s \in (ClassA_k \cup ClassC_k)$, we also need to compute the *increase count* and the *decrease count*. From scanning the customer sequence c in the changed part of DDB', if c supports s after database updating, then the increase count needs to be increased by 1. If c does not support s after database updating but supports s before database updating, then the decrease count is increased by 1. If $s \in$

$ClassA_k$, after scanning the changed part of DDB', we can obtain the support of s as follows:

The support of s = the support of s before database updating + the increase count of s - the decrease count of s (2)

Table 14 shows the increase count and the decrease count of some candidate 2-sequences$\in$ ($ClassA_k \cup ClassC_k$) after scanning the changed part of Table 12.

Step 4. Find the frequent sequences from ClassAk and ClassBk , and prune the candidates which cannot be frequent sequences from ClassCk, .

For each candidate sequence $s \in ClassA_k$, the number of the customer sequences which support s in the unchanged part of DDB' is unchanged after database updation. However, in the changed part of DDB', the number of the customer sequences which support s may be changed. In step 3, after scanning the changed part of DDB', the increase count and the decrease count for candidate sequence s can be obtained. We use expression (2) to compute the support of candidate sequence s to determine whether it is a frequent sequence. For the above example, sequence $<\{B\}\{A\}> \in ClassA_2$ is a frequent sequence and its support is 3 before database updating. After scanning the changed part of DDB', the increase count and the decrease count of $<\{B\}\{A\}>$ are 0 and 2, respectively. By expression (2), the support of $<\{B\}\{A\}>$ is $3+(0-2)=1$, which is less than the minimum support threshold. Hence, $<\{B\}\{A\}>$ is turn out to be a non-frequent sequence after updating the database.

For the candidate sequence $s \in ClassB_k$, because s is only contained in the changed part of DDB', the support of s can be obtained after scanning the changed part of DDB'. Hence, we can determine if s is a frequent sequence. In our example, all the candidates $\in ClassB_2$ are non-frequent sequences after scanning the changed part of DDB'.

Lemma 3. *Let sequence S* $ClassC_k$, *R = the increase count of S - the decrease count of S* , C_{before} *be the number of the customer sequencesbefore updating the database* , *and* C_{after} *be the number of customer sequences after updating the database. S is a non-frequent sequence , if one of the following conditions holds*:

(1) $C_{after} < C_{before}$, $R < 0$, and $|R| \geqq |C_{after} - C_{before}|$ (the minimum support

threshold/C_{after})

(2) $C_{after} \geqq C_{before}$, and $R \leqq 0$

(3) $C_{after} > C_{before}$, $R > 0$, and $R \leqq (C_{after} - C_{before})$ (the minimum support threshold/C_{after})

By using Lemma 3, we can prune the candidate sequences $\in ClassC_k$, which are the non-frequent sequences. For the above example, candidates $<\{C\}\{A\}>$, $<\{C\}\{B\}>$, $<\{C\}\{F\}>$, $<\{F\}\{A\}>$, $<\{F\}\{B\}>$ and $<\{F\}\{C\}>$ can be pruned from $ClassC_2$ according to Lemma 3, and the candidates $<\{A\}\{F\}>$ and $<\{B\}\{F\}>$ are remained in $ClassC_2$.

Step 6. Scan the unchanged part of DDB', record the position pair of each candidate k-sequence ClassCk in each customer sequence and find the frequent sequences from ClassCk.

In this step, for each candidate k-sequence $s \in ClassC_k$, we record the position pair of s in each customer sequence c in the unchanged part of DDB' according to Definition 2. After scanning the unchanged part of DDB', the support of s can be obtained. For example, Table 15 shows the related information of the candidate 2-sequences $\in ClassC_2$ after scanning the unchanged part of DDB' (i.e., customer sequences CID 2 and CID 4), and the candidate $<\{A\}\{F\}>$ is a frequent 2-sequence. Finally, *USLP* generates all the frequent 2-sequences $L_2' = \{<\{A\}\{B\}>, <\{A\}\{C\}>, <\{B\}\{C\}>\}$.

Step 7. *Generate C_{k+1}' from L_k', and Decompose C_{k+1}' into $ClassA_{k+1}$, $ClassB_{k+1}$, and $ClassC_{k+1}$.*

After generating C_{k+1}', *USLP* decomposes C_{k+1}' into $ClassA_{k+1}$, $ClassB_{k+1}$, and $ClassC_{k+1}$ according to Definition 4. For example, before updating the database, the frequent 3-sequence $L_3 = <\{A\}\{B\}\{C\}>$. Hence, L_3' can be decomposed into $ClassA_3 = <\{A\}\{B\}\{C\}>$ and $ClassB_3 = ClassC_3 = \phi$.

Step 8. Compute the *simple position* of each candidate (k + 1)-sequence in the changed part of DB', and then find the frequent 3-sequences from $ClassA_k$ and $ClassC_k$.

In the following, we first redefine the simple position for a sequence needs to be redefined first.

Definition 7.

Let sequence $S = < s_1, \cdots, s_{k+1} > (ClassA_{k+1} \cup ClassC_{k+1})$, and the two k-subsequence of S be $a = < s_1, \cdots, s_{k-1}, s_k >$ and $b = < s_1, \cdots, s_{k-1}, s_{k+1} >$, where $a, b \in (ClassA_k \cup ClassC_k)$. If the position pairs of a and b in the customer sequence c are $(f_{a1}, l_{a1})(f_{a2}, l_{a2})$ and $(f_{b1}, l_{b1})(f_{b2}, l_{b2})$, respectively, then the simple positions of S in c denoted as (sim_1, sim_2, sim_3) are defined as follows :

(1). If $f_{a2} \neq 0$, $l_{b1} \neq 0$, and $f_{a2} < l_{b1}$, then $sim_1 = l_{b1}$. Otherwise, $sim_1 = 0$

(2). If $f_{a1} \neq 0$, $l_{b1} \neq 0$, the sign of f_{a1} is the same as the sign of l_{b1}, and $f_{a1} < l_{b1}$, then $sim_2 = l_{b1}$. If $f_{a1} \neq 0$, $l_{b1} \neq 0$, the sign of f_{a1} is different from the sign of l_{b1}, and $f_{a1} < l_{b1}$, then $sim_2 = -l_{b1}$. Otherwise, $sim_2 = 0$.

(3). If $f_{a2} \neq 0$, $l_{b2} \neq 0$, and $f_{a2} < l_{b2}$, then $sim_3 = l_{b2}$. Otherwise, $sim_3 = 0$.

For example, the position pairs of $<\{A\}\{B\}>$ in customer sequence 1 in Table 13 are (+3, +3) (+3, +4), and the position pairs of $<\{A\}\{C\}>$ are (0, 0) (+4, +4). Thus, the simple positions of 3-sequence $<\{A\}\{B\}\{C\}>$ in customer sequence 1 are (0, 0, +4).

Definition 8.

Let $(k+1)$ -sequence $S = < s_1, \cdots, s_{k+1} > \in ClassB_{k+1}$, and the two k-subsequences of S: $a = < s_1, \cdots, s_{k-1}, s_k >$, and $b = < s_1, \cdots, s_{k-1}, s_{k+1} >$. The simple position of s in the customer sequence c is defined as follows:

(1). If $a \in ClassB_k$, $b \in (ClassA_k \cup ClassC_k)$, the position pair of a in c is (f_a, l_a), and the position pairs of b in c are $(f_{b1}, l_{b1})(f_{b2}, l_{b2})$, then the simple position sim of S in c is : If $f_a \neq 0$, $l_{b2} \neq 0$, and $f_a < l_{b2}$, then $sim = l_{b2}$. Otherwise, sim = 0.

(2). If $a \in (ClassA_k \cup ClassC_k)$, $b \in ClassB_k$, the position pairs of a in c are (f_{a1}, l_{a1}) (f_{a2}, l_{a2}), and the position pair of b in c is (f_b, l_b), then the simple position sim of S in c is: If $f_{a2} \neq 0$, $l_b \neq 0$, and $f_{a2} < l_b$, then $sim = l_b$. Otherwise, $sim = 0$.

(3). If a, $b \in ClassB_k$, the position pairs of a and b in c are (f_a, l_a), and (f_b, l_b), respectively, then the simple position sim of S in c is: If $f_a \neq 0$, $l_b \neq 0$, and $f_a < l_b$, then $sim = l_b$. Otherwise, $sim = 0$.

For example, the position pairs of 2-sequence $<\{A\}\{B\}>$ in customer sequence 1 in Table 13 are $(+3, +3)$ $(+3, +4)$, and the position pair of 2-sequence $<\{A\}\{E\}>$ in customer sequence 1 is $(5, 5)$. Thus, the simple position of 3-sequence $<\{A\}\{B\}\{E\}>$ in customer sequence 1 is 5.

Lemma 4. *For a $(k+1)$-sequence $S \in (ClassA_{k+1} \cup ClassC_{k+1})$, let the simple positions of S in the customer sequence c be (sim_1, sim_2, sim_3). If sim_1 is not equal to 0, then c supports S before and after updating the database.*

Rationale :

Suppose the subsequence r of the customer sequence c is ranging from the first transaction of c after updating the database to the last transaction of c before updating the database. If r contains $S = <s_1, \cdots, s_{k+1}>$, then c supports S before and after updating the database. Thus, c also supports the two k-subsequence $a = <s_1, \cdots s_{k-1}, s_k>$, and $b = <s_1, \cdots, s_{k-1}, s_{k+1}>$ of S. Let the position pairs of a and b at r in c are $(f_{a1}, l_{a1})(f_{a2}, l_{a2})$ and $(f_{b1}, l_{b1})(f_{b2}, l_{b2})$, respectively. If the simple position sim_1 of S in c is 0, then f_{a2} is greater than or equal to l_{b1} according to the Definition 8. That is, the positions of b are smaller than all the positions of a at r in c. Thus, r does not contain S. It is contradictory to the supposition. So, sim_1 must be not equal to 0.

Lemma 5. *Let the simple positions of the $(k+1)$-sequence $S = <s_1, \cdots, s_{k+1}> \in (ClassA_{k+1} \cup ClassC_{k+1})$ in the customer sequence c be (sim_1, sim_2, sim_3). If sim_2 is less than 0, then s_1 is contained in the deleted subsequence of c. If sim_3 is greater than or equal to the position of the inserted subsequence of c, then s_{k+1} is contained in the inserted subsequence.*

Lemma 6. *For a sequence $S \in ClassB_{k+1}$, if the simple position sim of S in c is greater than 0, then customer sequence c supports S.*

Hence, in this step, we first compute the simple positions of each candidate $(k+1)$-sequence in each customer sequence in the changed part of *DDB'* according to Definition 7 and Definition 8. From Lemma 4 and Lemma 5, we can obtain the increase count and the decrease count for a candidate $(k+1)$-sequence $s \in (ClassA_{k+1} \cup ClassC_{k+1})$. From Lemma 4, we can find whether the customer sequence c still supports s after updating the database. According to Lemma 5, we can check whether s is contained in the deleted subsequence or the inserted subsequence of c. After this checking, the increase count and the decrease count of s can be obtained.

After computing the simple positions of each candidate $(k+1)$-sequence corresponds to each customer sequences in the changed part of *DDB'*, the frequent $(k+1)$-sequences can be obtained from $ClassA_{k+1}$ and $ClassB_{k+1}$, and some candidates can be pruned from $ClassC_{k+1}$. For the candidates $ClassA_{k+1}$, *USLP* computes the supports of them by using expression (2) to determine whether they are frequent sequences. For the candidates $ClassB_{k+1}$, the supports of them can be obtained by computing the simple positions of them in the changed part of *DDB'* according to Lemma 6. For example, the increase count and the decrease count of $<\{A\}\{B\}\{C\}>$ are 1 and 0, respectively. The support of $<\{A\}\{B\}\{C\}>$ before updating the database is 3. By expression (2), the support of $<\{A\}\{B\}\{C\}>$ is $3+(1-0)=4$, which is a frequent 3-sequence.

Step 9. Compute the simple position of each candidate $(k+1)$-sequences $\in$ $ClassC_{k+1}$ in each customer sequence in the unchanged part of *DDB'*, and then find the frequent $(k+1)$-sequences from *Class* C_{k+1}.

USLP computes the simple positions of $s \in ClassC_{k+1}$ in the unchanged part by using expression (1) in Definition 3. The number of the customer sequences which support s in the unchanged part of *DDB'* can be obtained according to Lemma 2

According to above expression, the simple position of candidate $s \in ClassC_{k+1}$ can be obtained, and the frequent sequences can be discovered from $ClassC_{k+1}$. In our example, because of $ClassC_{k+1} = \phi$, there are no frequent sequences generated in this step.

Finally, *USLP* generates $L_3' = \{<\{A\}\{B\}\{C\}>\}$, and stops running.

Table 15. The related information for some candidate 2-sequences $\in$ ClassC_k.

Candidate 2-sequence	Position pair		Support
	CID 2	CID 4	
<{A}{F}>	(3, 3)	(2, 2)	3
<{B}{F}>	(3, 3)	(0, 0)	2

5. Experimental results

We refer to [2] to generate three synthetic transaction databases. We evaluate the performance of *SLP* by comparing this algorithm with *Apriori*all [2]. Suppose the maximum length of the candidate sequences generated by *Apriori*all is q. Hence, *Apriori*all has to scan the given transaction database q-1 times, but *SLP* needs only to scan the database $\lceil q-1/2 \rceil$ times. Figure 1 shows the relative execution times for *Apriori*all and *SLP* over the minimum support ranging from 20.2% to 19%, SLP outperforms Apriori*all*, and the performance gap increases as the minimum support decreases because the number of database scans increases for *Apriori*all.

We evaluate the performance of *USLP* by inserting transactions into and deleting transactions from the three tested transaction databases, and then compare this algorithm with *Apriori*all [2].

Assume that the number of the transactions in the original transaction database is n. we delete $n \times 1\%$ transactions from and insert $n \times 1\%$ transactions into the three tested databases, and then execute the two algorithms *USLP* and *Apriori*all. In the situation of a few data updating, most of the frequent sequences can be found efficiently by using discovered information by *USLP*, but Apriori*all* not. After inserting 2500 transactions into and deleting 2500 transactions from the tested databases, Figure 2 shows that *USLP* outperforms *Apriori*all. When the number of the inserted transactions and the deleted transactions increase, the relative execution time for *Apriori*all and *USLP* decrease, since the number of the customer sequences in the changed part of *DDB'* increase. *USLP* needs to take more effort to record the position pairs of candidate k-sequences and compute the simple positions of candidate

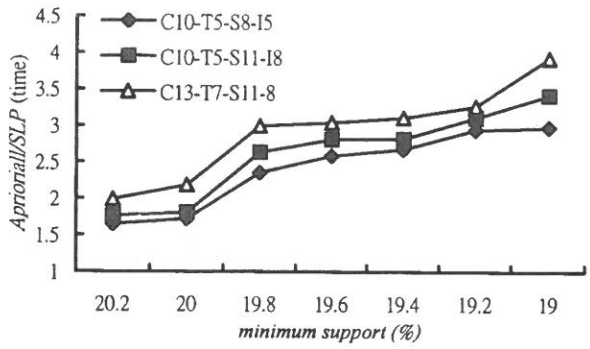

Fig. 1. Relative execution time

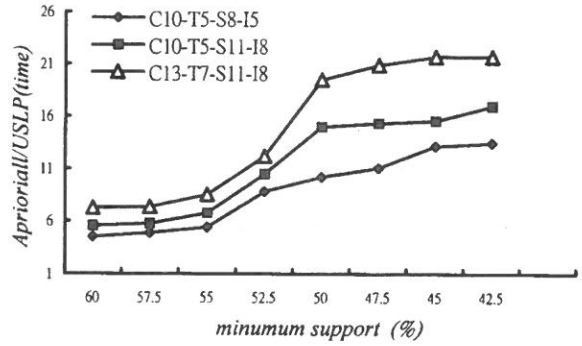

Fig. 2. Relative execution time

$(k+1)$ -sequences in the changed part of *DDB'*. Moreover, the relative execution time for Aprioriall and USLP decrease more quickly in the case of deleting transactions than the case of inserting transactions, since Aprioriall needs to scan fewer transactions than USLP when the transactions are deleted. However, *USLP* needs to scan the deleted transactions to compute the supports of some candidates. Figures 3 and 4 show the relative execution time for *Aprioriall* and *USLP* when the transactions are deleted and inserted, respectively.

6. Conclusion

When the customers are getting larger, in order to record all the position pairs of

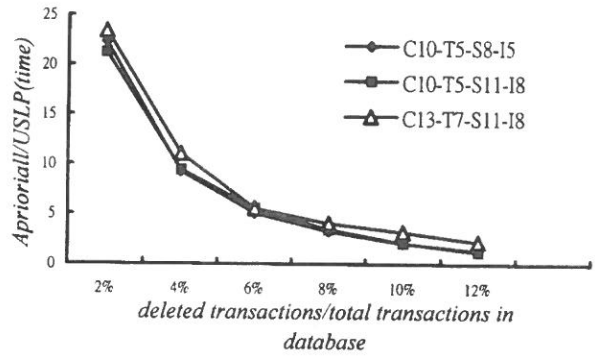

Fig. 3. Relative execution time

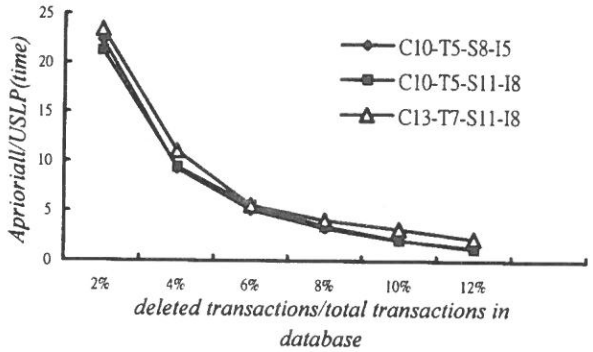

Fig. 4. Relative execution time

each candidate in the database, *SLP/USLP* may run out of memory. For such a situation, we can segment the database into segments. For each segment, only the customer sequences which related the segment are processed, and only the position pairs of each candidate in the segment are recorded.

Mining traversal patterns [15] is to discover the sequential traversing behaviors of most users in the network from a data log. Therefore, when a user is visiting a web page, we can use this knowledge to predict what webs the user wants to traverse next, and then pre-fetch the web pages. In the future, we shall consider to apply our approaches to discover the frequent traversal patterns in the world-wide-web.

Acknowlegement

This work is partially supported by the Republic of China National Science Council under Cortract No. 89-2213-E-030-022. The author wishes to express his thanks for the financial support of the Societas Verrbi Divini.

REFERENCES

(1) R. Agrawal and et al. Fast Algorithm for Mining Association Rules. In *Proceedings of International Conference on Very Large Data Bases*, pages 487-499, 1994.

(2) Rakesh Agrawal, and et al. Mining Sequential Patterns. In *Proceedings of International Conference on Data Engineering*, pages 3-14, 1995.

(3) Rakesh Agrawal, and et al. Mining Sequential Patterns: Generalizations and Performance Improvements. In *Proc. of the Fifth Int' l Conference on Extending Database Technology* (*EDBT*), Avignon, France, March 1996.

(4) Rakesh Agrawal, and et al. Mining Generalized Association Rules. *In Proceedings of the* 21*st VLDB* Conference Zurich, swizerland, 1995.

(5) Rakesh Agrawal, Johannes Gehrke: Dimitrios Gunopulos, Prabhakar Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Minig Applications. In *Proc. of the ACM SIGMOD Int' l Conference on Management of Data*, Seattle, Washington, June 1998.

(6) R. Agrawal, K. Lin, H. S. Sawhney, K. Shim. Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. *Proc. of the* 21*st Int' l Conference on Very Large Databases*, *Zurich*, Switzerland, September 1995.

(7) R. Agrawal, R. J. Bayardo Jr. and R. Srikant. Athena: Mining-based Interactive Management of Text Databases. *IBM Research Report* RJ10153, July 1999.

(8) R. J. Bayardo Jr. and R. Agrawal. Minig the Most Interesting Rules. In *Proc. of the* 5*th ACM SIGKDD Int' l Conf. on Knowledge Discovery and Data Mining*, August 1999.

(9) C. H. Cai, Ada Wai-Chee Fu, C. H. Cheng, W. W. Kwong: Mining

Association Rules with Weighted Items. IDEAS 1998: 68-77.

(10) David W. Cheung, Jiawei Han, Vincent T. Ng and C. Y. Wong. Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique. In *ICDE* 1996, page 106-114.

(11) David W. Cheung, S. D. Lee and Benjamin Kao. A General Incremental Technique for Maintaining Discovered Association Rules. In *DASFAA* 1997, page 185-194.

(12) Anthony K. H. Tung, Hongjun Lu, Jiawei Han, Ling Feng : Breaking the Barrier of Transactions : Mining Inter-Transaction Association Rules. KDD 1999: 297-301.

(13) Shiby Thomas, Sreenath Bodagala, Khaled Alsabti and Sanyay Ranka. An Efficient Algorithm for the Incremental Updation of Association Rules in Large Databases. In *KDD* 1997, page 263-2.

(14) Show-Jane Yen and Arbee L. P. Chen. An Efficient Approach to Discovering Knowledge from Large Databases. In *PDIS*, page 8-18, 1996.

(15) Show-Jane Yen Mining Frequent Traversal Patterns in a Web Environment. In *Proceedings of International Symposium on Intelligent Data Engineering and Learning*, pages. 219-224, 1998.

(16) Suh-Ying Wur, Yungho Leu: An Effective Boolean Algorithm for Mining Association Rules in Large Databases. *DASFAA* 1999: 19-30.

received September 29, 2000
revised October 12, 2000
accepted November 2, 2000

從大型資料庫中挖掘序列型樣的更新技術

顏 秀 珍　　左 聰 文
輔仁大學資訊工程學系

摘　　要

挖掘序列型樣是要從顧客交易的資料庫中找出顧客購物的循序行爲，挖掘序列型樣的主要工作是要找出大部份顧客常常依照某種順序購買的商品序列，過去已有不少學者提出挖掘序列型樣的演算法，這些演算法都需要重複掃描交易資料庫，而且需要花很多時間做大量的計算才能找出所有的序列型樣，因此，利用這些方法來找出我們所感興趣的購物順序是非常費時的。

另外，隨著新的交易被加入資料庫中，有些我們之前所找出的購物順序可能會有所改變，所以我們必須從更新後的資料庫中，重新找出所有我們感興趣的購物順序。挖掘序列型樣的更新技術能利用之前已找出的資訊，花費較少的時間，對更新過後的資料庫，重新找出我們所感興趣的購物順序。如此，便不用費時的去重新執行挖掘序列型樣的演算法。

在這篇論文中，我們提出一個挖掘序列型樣演算法-SLP，SLP能有效率的減少資料庫掃描的次數。另外，我們也提出一個更新序列型樣的演算法-USLP，USLP能根據既有的資訊，有效率的從更新之後的交易資料庫中，找出所有更新之後的購物順序。實驗結果顯示，我們所提出的演算法比其他挖掘序列型樣的演算法來找尋所感興趣的購物順序還要有效率。

關鍵詞： 資料探勘，知識發掘，交易資料庫，序列型樣，遞增式的更新技術

Universal Cascadable Current-Mode Biquad Using Only Three CCII+s and One CF+

Yung-Chang Yin
Department of Electronic Engineering
Fu Jen Catholic University
Taipei, Taiwan 242, R.O.C.

Abstract

A novel configuration using three plus-type second-generation current conveyors (CCII+s) and one plus-type current follower (CF+) is proposed to realize current-mode lowpass, highpass, bandpass, notch and allpass filters. These current-mode filters have high-output impedance, so they can be cascaded without any additional matching circuit. The proposed filters in the paper required less active and passive components than the previous filters. The use of only CCII+s and CF+ simplifies the configuration expectation for class-A operation. Finally, experimental results are included to certify the theoretical prediction.

Key Words: current conveyor, current follower, sensitivity

INTRODUCTION

Some recent literature has mentioned that the circuits that utilize a current-mode amplifier will operate at higher signal bandwidths, with greater linearity and with a larger dynamic range, than their voltage-mode counterparts [1]. Hence, several generic current-mode biquads using current conveyor have been described [2～8]. However, all of these that have been proposed have some disadvantages, such as high input impedance, low output impedance, an excessive number of active and passive components and the inclusion of some minus-type current conveyors (CCII−s).

Basically, the CCII+ is a simple current mirror and the CCII− is used two cross-coupled current mirrors. However, the CCII− must suffer from the limitations imposed by the frequency response and mismatch of the transistor mirrors. In this paper, a new configuration using only three CCII+s and one CF+ for realizing the cascadable current-mode filters is developed. The proposed circuit can realize highpass, lowpass, bandpass, notch and allpass filters from the same configuration. The major goal is to reduce the number of the required CCII− elements for the current-mode filters. Finally, experimental results are given to confirm the afore mentioned theoretical analysis.

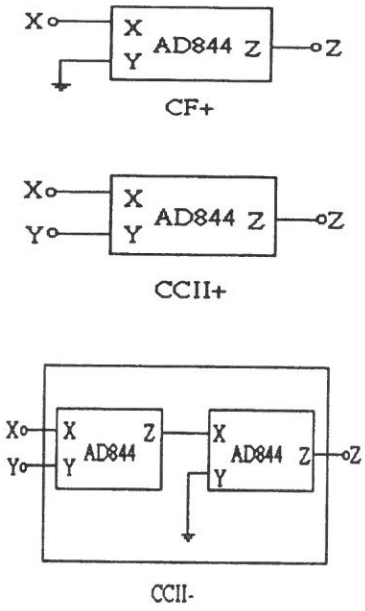

Fig. 1. CF+, CCII+ and CCII− implementation using AD844s

(a). CF+ implementation using AD844

(b). CCII+ implementation using AD844

(c). CCII− implementation using AD844s

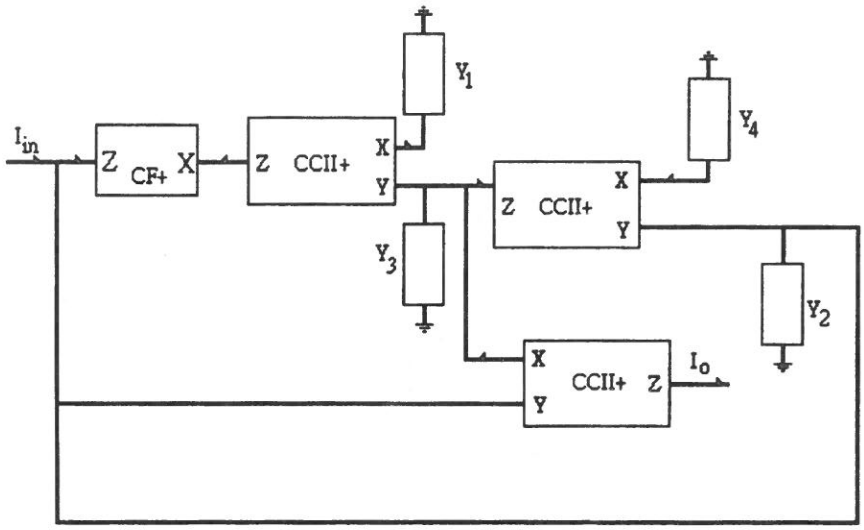

Fig. 2. Proposed universal active current circuit using three CCⅡ+s and one CF+

CIRCUIT DESCRIPTION

The plus-type CF, plus-type CCⅡ and minus-type CCⅡ implementation using AD844 are shown in Fig. 1 (a), (b) and (c), respectively. The proposed filter circuit based on and employing three CCⅡ+s and one CF+ is shown in Fig. 2. Using standard notation, the port relations of a CCⅡ+ and CF+ can be characterized by $i_z = i_x$, $v_x = v_y$, and $i_y = 0$ for CCⅡ+ and $v_x = 0$ and $i_z = i_x$ for CF+, respectively. From a routine analysis, the transfer function of the Fig. 2 can be given by:

$$\frac{I_0}{I_{in}} = \frac{Y_3 - Y_4}{Y_1 + Y_2} \tag{1}$$

where $Y_1 - Y_4$ are admittances. If $Y_2 = Y_3$ and $Y_4 = 2Y_1$, equation (1) will be

$$\frac{I_0}{I_{in}} = \frac{Y_2 - 2Y_1}{Y_1 + Y_2} \tag{2}$$

If the admittances are chosen as $Y_2 = sC_2 + (1/R_2)$ and $Y_1 = 1/(R_1 + 1/sC_1)$, equation (2) will be

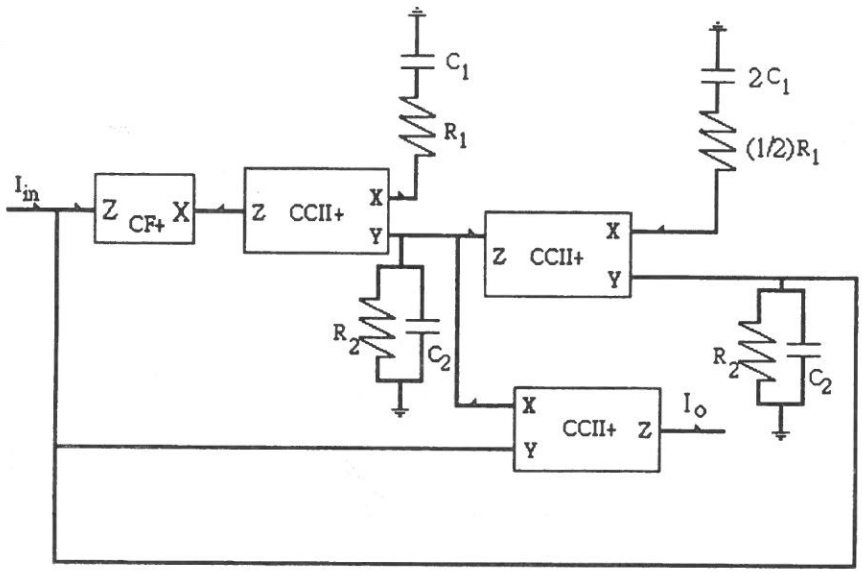

Fig. 3. A current-mode second-order notch/allpass filter

$$\frac{I_0}{I_{in}} = \frac{s^2C_1C_2R_1R_2 + s(C_1R_1 + C_2R_2 - 2C_1R_2) + 1}{s^2C_1C_2R_1R_2 + s(C_1R_1 + C_2R_2 + C_1R_2) + 1} \tag{3}$$

Hence, if $C_1R_1 + C_2R_2 = 2C_1R_2$, a second-order notch filter shown in Fig. 3 can be realized. If $2(C_1R_1 + C_2R_2) = C_1R_2$, a second-order allpass filter shown in Fig. 3 can be also achieved. Furthermore, if $Y_3 = 0$ (open-circuited), $Y_4 = Y_1 = sC_1$ and $Y_2 = G_2$, equation (1) can be expressed as

$$\frac{I_0}{I_{in}} = \frac{-sC_1}{sC_1 + G_2} \tag{5}$$

Thus, a highpass filter can be obtained.

If $Y_3 = 0$ (open-circuited), $Y_1 = sC_1$ and $Y_4 = Y_2 = G_2$. Equation (1) can be written as

$$\frac{I_0}{I_{in}} = \frac{-G_2}{sC_1 + G_2} \tag{6}$$

Thus, a lowpass filter can be obtained.

To realize a bandpass filter, the admittances in Fig. 2 might be chosen as follows:

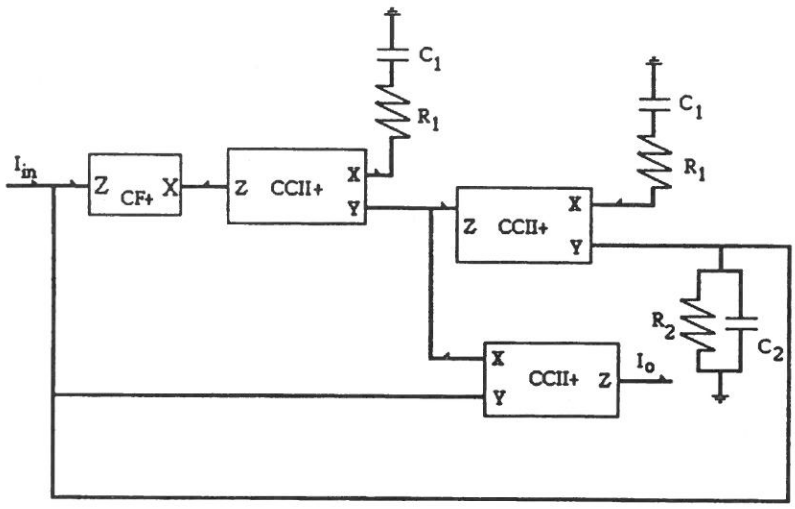

Fig. 4. A current-mode second-order bandpass filter

$Y_3 = 0$ (open-circuited), $Y_4 = Y_1 = 1/(R_1 + 1/sC_1)$ and $Y_2 = sC_2 + (1/R_2)$. Equation (1) can be expressed as

$$\frac{I_0}{I_{in}} = \frac{-sC_1R_2}{s^2C_1C_2R_1R_2 + s(C_1R_1 + C_2R_2 + C_1R_2) + 1} \tag{7}$$

Thus, a second-order bandpass filter shown in Fig. 4 can be obtained. The resonance angular frequency ω_0 and the quality factor ϑ of the current mode bandpass filter shown in Fig. 4 can be expressed as

$$\omega_0 = \frac{1}{(C_1C_2R_1R_2)^{1/2}} \quad \text{and} \quad \vartheta = \frac{(C_1C_2R_1R_2)^{1/2}}{C_1R_1 + C_2R_2 + C_1R_2}$$

By relating a sensitivity parameter F to the element of variation X_i

$$S_{X_i}^F = \frac{X_i}{F}\frac{dF}{dX_i}$$

The passive sensitivities are given by

$$S_{R_1}^{\omega_0} = S_{R_2}^{\omega_0} = S_{C_1}^{\omega_0} = S_{C_2}^{\omega_0} = -1/2$$

$$S_{C_1}^{\vartheta} = \frac{1}{2}\left[1 - \frac{R_1 + R_2}{\Delta(C_2R_1R_2)}\right]$$

$$S_{C_2}^{\vartheta} = \frac{1}{2}[1 - \frac{R_2}{\Delta(C_1R_1R_2)}]$$

$$S_{R_1}^{\vartheta} = \frac{1}{2}[1 - \frac{C_1}{\Delta(C_1C_2R_2)}]$$

$$S_{R_2}^{\vartheta} = \frac{1}{2}[1 - \frac{C_1 + C_2}{\Delta(C_1C_2R_1)}]$$

where $\Delta = (C_1R_1 + C_2R_2 + C_1R_2)$. All the passive sensitivities are quite small. From the above mentioned, these filters provide the following advantage: a saving of the active components (AD844), low input impedance, and they are capable of cascading without current buffers. The comparison between this paper and the previous one (Hou and Wu 1997) is shown in the table. From the table, the required number of active components (AD844) is indeed reduced.

Table 1. Number of AD844 components

	lowpass	highpass	notch	bandpass	allpass
Previous paper (Hou and Wu 1997)	7	7	7	7	8
Present paper	4	4	4	4	4

Comparison of the active components (AD844)

EXPERIMENTAL RESULTS

To demonstrate the feasibility of the of the proposed circuits, the circuits in Fig. 2 are implemented by using the commercial available CCII+ and CF+ IC (AD844). A second-order notch filter is shown in Fig. 3 with $C_1 = C_2 = 1nF$ and $R_1 = R_2 = 10k\Omega$. A second-order bandpass filter is shown in Fig. 4 with $C_1 = C_2 = 1uF$ and $R_1 = R_2 = 10k\Omega$. The Matlab has carried out simulating the ideal curves of the proposed filters. The experimental results above are measured using the Hewlett Packard network/spectrum analyzer 4195A. Figure 5 shows the experimental results for notch and bandpass filters. The theoretical analysis correlated with the measured results with few errors which were due to the errors of the use of active and passive elements. However, the measured frequency response of the filter is in good agreement with the theory.

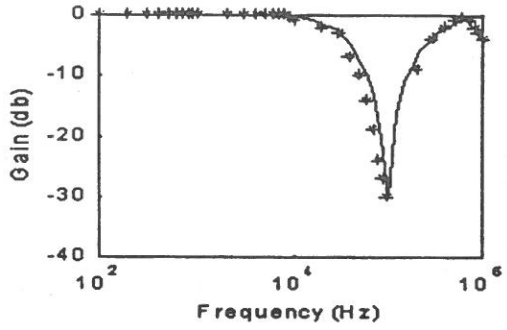

Fig. 5 (a). Amplitude-frequency response for notch

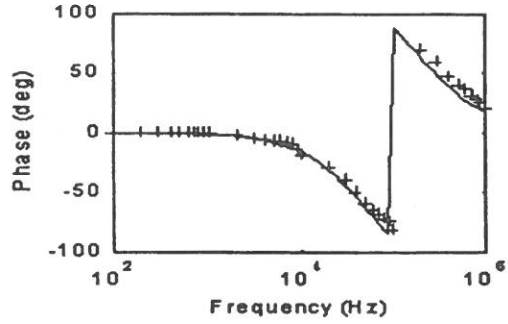

Fig. 5 (b). phase-frequency response for notch

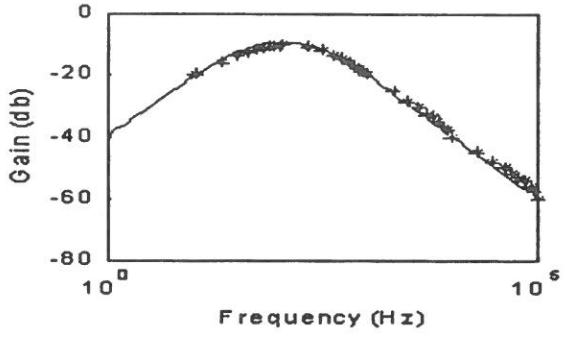

Fig. 5 (c). Amplitude-frequency response for bandpass

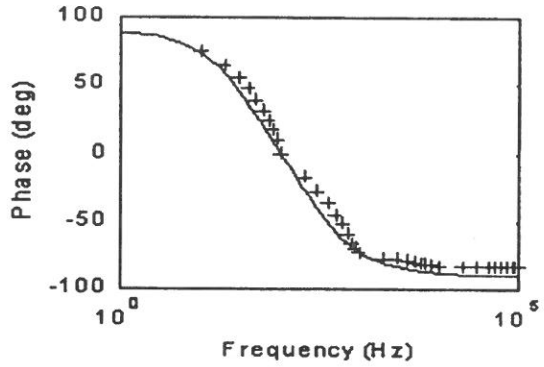

Fig. 5 (d). phase-frequency response for bandpass

Fig. 5. Amplitude-frequency response and phase-frequency response

(a) Amplitude-frequency response for notch

(b) phase-frequency response for notch

(c) Amplitude-frequency response for bandpass

(d) phase-frequency response for bandpass

—— ideal response

***, + measured response**

CONCLUSION

A new configuration using only three CCII+s and one CF+ for realizing various types of cascadable current-mode filters is presented. The lowpass, highpass, bandpass, notch and allpass filters can be achieved from the same configuration. This paper has improved the previous one of Hou and Wu (1997). Finally, experimental results confirmed the theoretical analysis. The results will be useful in analogue signal processing applications.

ACKNOWLEDGMENT

The author wishes to express his thanks for the financial support of the Societas Verrbi Divini.

REFERENCES

(1) B. Wilson, "Recent Developments in Current Conveyors and Current-Mode Circuits", IEE Proc-G, vol. 137, no. 2, pp. 63-77, 1990.

(2) G. W. Rober and A. S. Sedra, "All Current-Mode Frequency Selective Circuits", Electron. Lett., Vol. 25, pp. 759-761, 1989.

(3) C. M. Chang, "Universal active current filters using single second-generation current conveyors", Electron. Lett., vol. 27, no. 18, pp. 1614-1617, 1991

(4) C. M. Chang, C. C. Chein and H. Y. Wang "Universal active current filters using single second-generation current conveyors", Electron. Lett., vol. 29, no. 13, pp. 1159-1160, 1993.

(5) C. M. Chang and P. C. Chen "Universal active current filter with three inputs and one output using current conveyors", Int. J Electron., vol. 71, no. 5, pp. 817-819, 1991.

(6) R. Senani "New current-mode biquad filter", Int. J Electron., vol. 73, no. 4, pp. 735-742, 1992.

(7) Y. C. Yin "A Current-mode Biquad filter with multi-output using CC Ⅱ s", Journal of the Chinese Institute of Electrical Engineering., vol. 3, no. 4, pp. 367-370, 1996.

(8) C. L. Hou and J. S. Wu "Universal cascadable current-mode biquad using only four four CC Ⅱ s", Int. J Electron., vol. 82, no. 2, pp. 125-129, 1997.

received September 22, 2000
accepted November 23, 2000

使用三個正型電流傳輸器和一個電流隨耦器合成泛用電流式濾波器

鄞 永 昌
輔仁大學電子工程系

摘 要

本論文提出使用三個正型電流傳輸器和一個正型電流隨耦器合成新結構之低通，高通，帶通，帶拒和全通濾波器，此濾波器具有高的輸出阻抗，所以不需任何的匹配電路，同時本濾波器所使用之主動與被動元件比先前的濾波器少，最後以實驗證明本新結構濾波器的理論預測。

關鍵詞：電流傳輸器，電流隨耦器，靈敏度

Preparation and Properties of Poly(Aniline-*co-ortho*-Ethoxyaniline) -Clay Nanocomposite Materials

Jui-Ming Yeh*, Shir-Joe Liou, Chiung-Yu Lai, Yung-Su Chang

Department of Chemistry
Chung-Yuan Christian University
Chung Li, Taiwan 320, R.O.C.

Abstract

A series of lamellar nanocomposite materials that consisted of poly(aniline-*co-ortho*-ethoxyaniline) and layered montmorillonite (MMT) clay were prepared by effectively dispersing the inorganic nanolayers of MMT clay in organic poly(aniline-*co-ortho*-ethoxyaniline) matrix via *in-situ* polymerization. Both organic aniline and ethoxyaniline monomers at specific feeding ratio were first intercalated into the interlayer regions of organophilic clay hosts and followed by an one-step oxidative polymerization. The as-synthesized poly(aniline-*co-ortho*-ethoxyaniline) -clay materials were characterized by infrared spectroscopy, wide-angle powder X-ray diffraction and transmission electron microscopy. The effects of material composition on the properties of anticorrosion and molecular barrier of bulk poly(aniline-*co-ortho*-ethoxyaniline) along with a series of PCN materials, in the form of coating and free-standing film, were also studied by electrochemical corrosion potential measurements and gas permeability measurements, respectively.

Key Words: Nanocomposite, Clay, Aniline, Barrier, Corrosion

* To whom correspondence should be addressed.

INTRODUCTION

Conducting polymers, such as polypyrrole, polythiophene, polyaniline, etc., consisted of conjugated electronic structures have received considerable attention lately because of many promising technological applications. Some specific conducting polymers, e. g., polyaniline and its derivatives, have been found to display interesting corrosion protection properties. In the past decade, the use of polyanilines as anticorrosion coatings had been explored as the potential candidates to replace the chromium-containing materials, which have adverse health and environmental concerns. [1-6] Wei et. al. [6] demonstrated the corrosion protection effect of polyaniline through a series of electrochemical measurements on the doped or undoped polyaniline-coated cold rolled steel (CRS) under various conditions. Wessling [5] proposed a full mechanism that the corrosion protection of polyaniline on steel is attributed to an increase in the corrosion potential and to the redox catalytic property of polyaniline in the formation of passive layer of metal oxide.

Layered materials, such as smectite clays (e. g., montmorillonite, MMT), attracted intense research interest for the preparation of polymer-clay nanocomposites in the past decade, because of their lamellar elements displayed high in-plane strength, stiffness as well as high aspect ratio. [7] The polymer-clay nanocomposites were found to enhance the gas barrier [8], thermal stability [9], mechanical strength [10] and fire retardant [11] properties of polymers. Recently, there are a number of reports on the preparation and properties for the lamellar nanocomposites of polyaniline with various layered materials. [12-18]

In this paper, we present the preparation of poly(aniline-*co-ortho*-ethoxyaniline) -clay nanocomposite (PCN) materials by incorporating nanolayers of MMT clay into soluble polyaniline derivative matrix, i.e., poly(aniline-*co-ortho*-ethoxyaniline), and study their anticorrosion performance on cold-rolled steel coupons in comparison with that of the emeraldine base of poly(aniline-*co-ortho*-ethoxyaniline) by performing electrochemical corrosion potential measurements in 5 wt-% aqueous NaCl electrolyte. The PCN materials were characterized by infrared spectroscopy, wide-angle powder X-ray diffraction and transmission electron microscopy. The effect of material

composition on the molecular barrier property was also investigated by the gas permeability measurements.

Experimental Section

Chemicals and Instrumentations Aniline (99%, Fluka) was doubly distilled under a reduced pressure. 1.0 M HCl and 1.0 M NH_4OH were prepared from diluting the concentrated ammonia and hydrochloric acid (Riedel-de Haen) with distilled water for preparing the acidic and basic aqueous media, respectively. 1-Methyl-2-pyrrolidinone (NMP) (99.97%, Tedia) and o-ethoxyaniline (99%, Acros) were used as received without further purification. Ammonium persulfate (98%, SHOWA) was employed as oxidant. The used montmorillonite clay consisted of a CEC value of 114 meq/100 g and an unit cell formula $Ca^{+2}_{0.084}Na^{+}_{0.143}[Al_{1.69}Mg_{0.31}Si_4O_{10}(OH)_2 \cdot 2H_2O$ was provided by ITRI. The used intercalating agent was tetradecyltrimethylammonium chloride (Fluka).

Wide-angle X-ray diffraction study of the samples was performed on a Rigaku D/MAX-3C OD-2988N X-ray diffractometer with cooper target and Ni filter at a scanning rate of 4°/min. The samples for transmission electron microscopy (TEM) study was first prepared by putting powder of PCN materials into epoxy resin capsules followed by curing the epoxy resin at 100℃ for 24 hours in a vacuum oven. Then the cured epoxy resin containing PCN materials were microtomed with Reichert-Jumg Ultracut-E into 60～90-nm-thick slices. Subsequently, one layer of carbon about 10 nm thick was deposited on these slices on mesh 100 copper nets for TEM observations on a JEOL-200FX with an acceleration voltage of 120 KV.

Electrochemical measurements of sample-coated CRS coupons were performed on a VoltaLab 21 Potentiostat/Galvanostat in a standard corrosion cell equipped with two graphite rod counter electrodes and a saturated calomel electrode (SCE) as well as the working electrode. FTIR spectra were recorded on pressed KBr pellets using a BIO-RAD FTS-7 FTIR spectrometer. A Yanagimoto Co., Ltd gas permeability-measuring apparatus (model GTR 10) was employed to perform the permeation experiment of air.

1. Synthesis of Poly(aniline-*co-ortho*-Ethoxyaniline)

In a typical procedure, both 10 ml of doubly distilled aniline (0.107 mole) and 1.4 ml of ortho-ethoxyaniline monomers (0.012 mole) were dissolved in 600 ml of 1.0 M HCl and the mixture was precooled to ~0℃ in an ice bath. A solution (200 ml) of 5.6 g (0.025 mole) $(NH_4)_2S_2O_8$ in 1.0 M HCl was added to the solution dropwise over a period of 15 minutes with vigorous magnetically stirring. After ~ 2 hours, the precipitate was collected on a Buchner funnel. Upon drying under dynamic vacuum at room temperature, the HCl-doped poly(aniline-*co-ortho*-ethoxyaniline) was obtained as a green powder. The HCl-doped poly(aniline-*co-ortho*-ethoxyaniline) was subsequently converted into emeraldine base form by stirring ~ 3 g of the poly(aniline-*co-ortho*-ethoxyaniline) fine powder in an excess amount (500 ml) of 1.0 M ammonium hydroxide at room temperature for 3 hours. Upon filtering and drying under dynamic vacuum for 48 hours, the emeraldine base (EB) form of poly(aniline-*co-ortho*-ethoxyaniline) was obtained as a blue powder.

2. Preparation of Organoplilic Clay

The organophilic clay was prepared by a cation-exchange reaction between the sodium cations of MMT clay and alkylammonium ions of intercalating agent. Typically, 5 g of MMT clay with a CEC value of 114 meq/100 g was stirred in 500 ml distilled water (beaker A) at room temperature overnight. A separate solution containing 2.5 g of intercalating agent in another 100 ml of distilled water (beaker B) under magnetically stirring, followed by adding 1.0 M HCl aqueous solution to adjust the pH value to about 3-4. After stirring for 1 hour, the protonated amino acid solution (beaker B) was added at a rate of approximately 10 ml/min with vigorous stirring to the MMT suspension (beaker A). The mixture was stirred overnight at room temperature. The organophilic clay was recovered by ultracentrifugating (9000 rpm, 30 min) and filtering the solution in a Buchner funnel. Purification of products was performing by washing and filtering samples repeatedly for at least three times to remove any excess of ammonium ions.

3. Preparation of poly(aniline-*co-ortho*-ethoxyaniline) /Clay Nanocomposite Materials

As a typical procedure for the preparation of the PCN materials, first, an appropriate amount of organophilic clay was introduced into 400 ml of 1.0 M HCl under magnetically stirring overnight at room temperature. Aniline monomer (0.1 mol) and *ortho*-ethoxyaniline monomer (0.012 mol) were subsequently added to the solution which was stirred for another 24 hours. Upon addition of ammonium persulfate (0.025 mol) in 20 ml of 1.0 M HCl, the solution was stirred for 3 hours at 5 ℃ in an ice bath. The as-synthesized HCl-doped lamellar nanocomposite precipitates were then obtained by filtering and drying under dynamic vacuum at room temperature for 48 hours. The final PCN products in base form were obtained by immersing the HCl-doped nanocomposites into 400 ml of 1.0 M NH_4OH under magnetically stirring for 4 hours at room temperature, followed by filtration and drying under vacuum at room temperature for 48 hours.

4. Preparation of Coatings and Electrochemical Measurements

The emeraldine base of poly(aniline-*co-ortho*-ethoxyaniline) and PCN fine powder were dissolved in NMP to give typically 1-wt% solutions. The solutions were cast dropwise onto the cold rolled steel (CRS) coupons (1.0×1.0 cm) followed by drying in air for 48 hours to give coatings of ca. ~ 10 μm in thickness. The coated and uncoated coupons were then mounted to the working electrode so that only the coated side of the coupon was in direct contact with the electrolyte. The edges of the coupons were sealed with super fast epoxy cement (SPAR*). All the electrochemical measurements were made at room temperature and repeated at least three times. The electrolyte was NaCl (5 wt-%) aqueous solution. The open circuit potential (OCP) at the equilibrium state of the system was recorded as the corrosion potential (E_{corr} in V vs. SCE).

5. Preparation of Free-Standing Films and Barrier Property Measurements

0.3 g of poly(aniline-*co-ortho*-ethoxyaniline) and PCN materials in base form was dissolved in 10 ml NMP under magnetically stirring at room temperature for 4 hours.

The solution was cast onto a substrate (e. g., a microscope glass slide). The solvent was allowed to evaporate at 90 ~ 100 ℃ under the hood for 24 hours. The sample-coated glass substrate was then immersed into the distilled water for 5-6 hours to give the free-standing film of poly (aniline-*co-ortho*-ethoxyaniline) and PCN materials. Molecular barrier property for free-standing film of samples was determined by air transmission through the nanocomposite films according to ASTM standard E 96. A Yanagimoto Co., Ltd gas permeability-measuring apparatus (model GTR 10) was employed to perform the permeation experiment of air. The pressure on one face of the free-standing film was kept at 200 kPa and the other face was at zero pressure initially to allow the air to permeate through the free-standing film. The rate of transmission of air was obtained by gas chromatography, from which the air permeability was calculated.

Results and Discussions

To synthesize the PCN materials, organophilic clay was first prepared by a cation-exchange reaction between the sodium cations of MMT clay and alkylammonium ions of intercalating agent. Organic aniline and *ortho*-ethoxyaniline monomers were subsequently intercalated into the interlayer regions of organophilic clay hosts and followed by an one-step oxidative polymerization. The composition of the PCN materials was varied from 0 to 1 wt-% of clay with respect to poly(aniline-*co-ortho*-ethoxyaniline) content as summarized in Table 1.

表 1 Sample composition and electrochemical corrosion potential measurement data.

Compound Code	feed composition (wt-%)		Ecorr (mV)
	copolymer	MMT	
Bare	--	--	-641
PANI	100	0	-556
CLAN025	99.25	0. 75	-554
CLAN05	99.50	0.50	-540
CLAN075	99.25	0.75	-371
CLAN1	99.00	1.00	-361

Characterization

The representative FTIR spectra of the organophilic clay, bulk poly(aniline-*co-ortho*-ethoxyaniline) and PCN materials are shown in Figure 1. The characteristic vibration bands of poly(aniline-*co-ortho*-ethoxyaniline) are at 1498 cm^{-1} and 1583 cm^{-1}, and those of MMT clay are shown at 1040 cm^{-1} (Si-O), 600 cm-1 (Al-O) and 420 cm^{-1} (Mg-O) [17]. As the loading of MMT clay is increased, the intensities of MMT clay bands become stronger in the FTIR spectra of PCN materials.

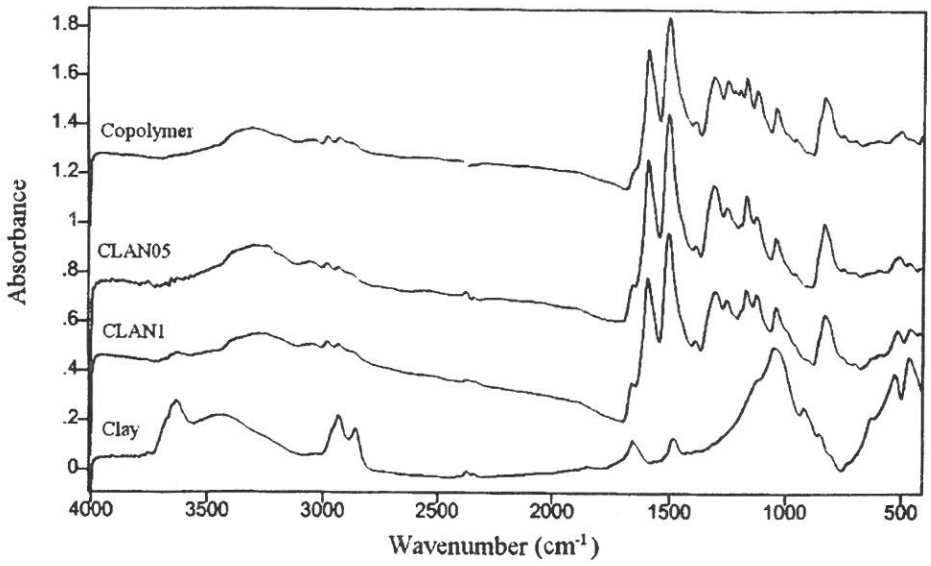

Fig. 1. Infrared spectra of copolymer, clay and nanocomposite materials.

Figure 2 shows the wide-angle powder X-ray diffraction patterns of organophilic clay and a series of PCN materials. There is a lack of any diffraction peak in 2θ = 2-10° as opposed to the diffraction peak at 2θ = 5.0° (d spacing = 1.65 nm) for organophilic clay, indicating the possibility of having exfoliated silicate nanolayers of organophilic clay dispersed in poly(aniline-*co-ortho*-ethoxyaniline) matrix.

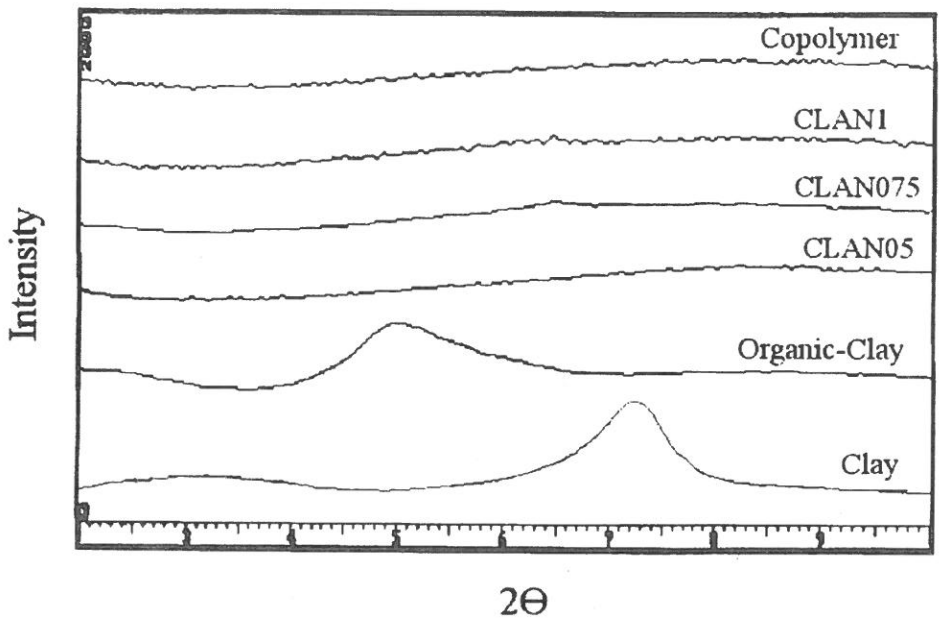

Fig. 2. Wide-angle powder X-ray diffraction patterns of copolymer, clay and Nanocomposite materials.

In Figure 3, the TEM of PCN materials with 1 wt-% clay loading shows that the lamellar nanocomposite has a mixed nanomorphology. Individual silicate layers, along two, three and four layer stacks, were found to be exfoliated in the poly(aniline-*co-ortho*-ethoxyaniline) matrix. In addition, some larger intercalated tactoids (multiplayer particles) can also be identified.

Corrosion Protection Properties of Coatings

Corrosion protection of sample-coated CRS coupons can be observed from the values of corrosion potential (E_{corr}), as listed in Table 1. The CRS coupon coated with emeraldine base of poly(aniline-*co-ortho*-ethoxyaniline) shows a higher E_{corr} value than the uncoated CRS. However, it exhibits a lower Ecorr value than the specimen coated with PCN materials. For example, the CLAN075-coated CRS has a high corrosion potential of ca. -372 mV at 30 min. Even after 5 hours measurement, the potential remained at ca. -390 mV. Such a E_{corr} value implies that the CLAN075-coated CRS is

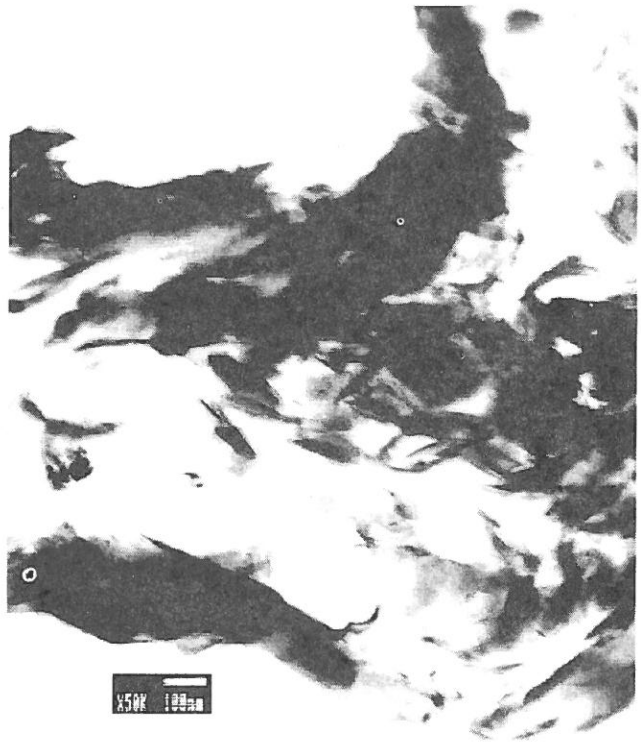

Fig. 3. Transmission electron microscopy of CLAN1..

more noble towards the electrochemical corrosion compared to the emeraldine base of poly (aniline-*co-ortho*-ethoxyaniline). Visual observation of the corrosion products clearly reveals that the PCN samples exhibiting corrosion protection have a grayish oxide layer form over the bare exposed CRS surface, similar to what was observed by Wessling under the polyaniline dispersion coatings on steel. [5] Enhanced corrosion protection effect of poly(aniline-*co-ortho*-ethoxyaniline) -clay nanocomposite materials compared to bulk poly(aniline-*co-ortho*-ethoxyaniline) might be resulted from dispersing silicate nanolayers of clay in poly(aniline-*co-ortho*-ethoxyaniline) matrix to increase the tortousity of diffusion pathway of H_2O molecules. This is further evidenced by the studies of the H_2O molecular barrier effect as discussed in the following section.

H_2O Molecular Barrier of Free-Standing Films

The free-standing films of PCN materials and bulk poly (aniline-*co-ortho*-ethoxyaniline) used for the molecular barrier measurements were prepared to have film thickness of ~ 70 μm. Compared to bulk copolymer, free-standing film of PCN materials at low clay loading (e. g., 0.50 wt-%) shows about 40% reduction of H_2O molecular permeability, as shown in Fig 4. This is caused by the barrier properties of the nanolayers of clay dispersed in the composite. [8] Furthermore, it should be noted that a further increase of clay loading results in a further enhanced molecular barrier property of bulk PCN materials.

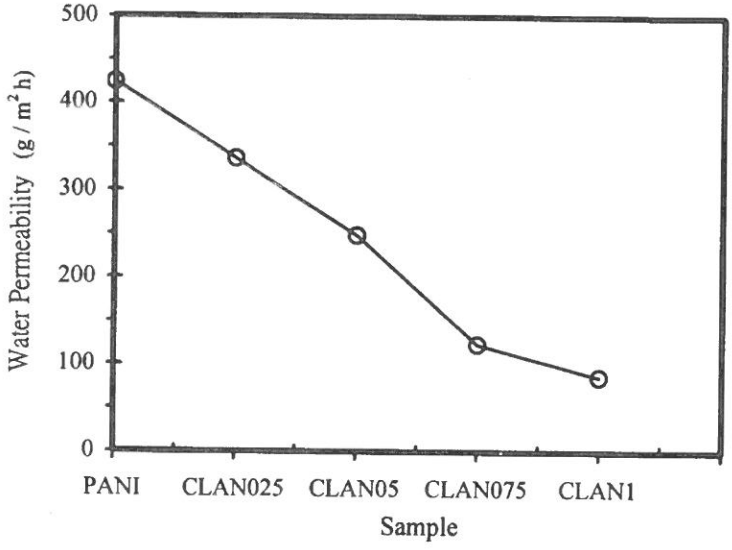

Fig. 4. Water permeability of PANI and a series of nanocomposite materials.

CONCLUDING REMARKS

A series of lamellar nanocomposite materials that consisted of emeraldine base of poly(aniline-*co-ortho*-ethoxyaniline) and layered montmorillonite (MMT) clay were prepared by effectively dispersing the inorganic nanolayers of MMT clay in organic

poly (aniline-*co-ortho*-ethoxyaniline) matrix via *in-situ* polymerization. The as-synthesized PCN materials were characterized by infrared spectroscopy, wide-angle powder X-ray diffraction and transmission electron microscopy. Anticorrosion effect of emeraldine base of poly (aniline-*co-ortho*-ethoxyaniline) -clay nanocomposite (PCN) materials at low clay loading compared to conventional poly (aniline-*co-ortho*-ethoxyaniline) was demonstrated by electrochemical measurements of corrosion potential on CRS in 5 wt-% aqueous NaCl electrolyte. The coatings of PCN materials were found to offer good corrosion protection and showed a better anticorrosion performance than conventional poly (aniline-*co-ortho*-ethoxyaniline) coating. The H_2O molecular barrier of PCN materials in the form of free-standing film (e. g., CLAN05) exhibits a 40% reduction in permeability compared to the emeraldine base of bulk poly (aniline-*co-ortho*-ethoxyaniline).

ACKNOWLEDGMENT

The financial support of this research by the NSC 89-2113-M-033-001 is gratefully acknowledged.

REFERENCE

(1) Deberry, D. W. J. Electrochem. Soc. 1985, 132, 1027.
(2) Wessling, B. Synth. Met. 1991, 907, and 1057.
(3) Elsenbaumer, R. L., Lu, W. K., Wessling, B. Int. Conf. Synth. Met. Seoul, Korea, Abstract No. APL (POL) 1-2, 1994.
(4) Wrobleski, D. A.; Benicewicz, B. C.; Thompson, K. G.; Byran, C. J. Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.) 1994, 35 (1), 265.
(5) Wessling, B. Adv. Mater. 1994, 6, 226.
(6) Wei, Y.; Wang, J.; Jia, X.; Yeh, J. -M. and Spellane, P. Polymer 1995, 36, 4535.
(7) Pinnavaia, T. J. Science 1983, 220, 365.
(8) Lan, T.; Kaviratna, P. D.; Pinnavaia, T. J. Chem. Mater. 1994, 6, 573.
(9) Tyan, H. -L.; Liu, Y. -C.; Wei, K. -H. Chem. Mater. 1999, 11, 1942.
(10) Wang, Z.; Pinnavaia, T. J. Chem. Mater. 1998, 10, 3769.

(11) Gilman, J. W.; Jackson, C. L.; Morgan, A. B.; Hayyis, R., Jr.; Manias, E.; Giannelis, E. P.; Wuthenow, M.; Hilton, D.; Phillips, S. H. Chem. Mater. 2000, 12, 1866.

(12) Biswas, M.; Ray, S. S. J. Appl. Polym. Sci. 2000, 77, 2948.

(13) Wu, C. -G.; DeGroot, D. C.; Marcy, H. O.; Schindler, J. L.; Kannewurf, C. R.; Liu, Y. -J.; Hirpo, W.; Kanatzidis, M. G. Chem. Mater. 1996, 8, 1992.

(14) Wang, L.; Brazis, P.; Rocci, M.; Kannewurf, C. R.; Kanatzidis, M. G. Chem. Mater. 1998, 10, 3298.

(15) Chao, K. -J.; Ho, S. -Y.; Chang, T. -C. US patent 5, 340, 500 (1994)

(16) Giannelis, E.; Mehrota, V. US Patent 5, 032, 547 (1991)

(17) Choi, H. J.; Kim, J. W.; Kim, S. G.; Kim, B. H. Joo, ; J. Polym. Mater. Sci. Eng. 2000, 82, 245.

(18) Wu, Q.; Xue, Z.; Qi, Z.; Wang, F. Polymer 2000, 41, 2029.

received October 27, 2000
accepted November 29, 2000

苯胺/鄰位取代乙氧基苯胺共聚物與黏土之納米複合材料合成與性質研究

葉瑞銘 劉時州 賴瓊玉 張永素
中原大學化學系

摘 要

藉由陽離子界面活性劑來改質具親水性無機層狀黏土，再進一步將兩有機苯胺單體依適當比例滲透進入親油化改質處理過之黏土層間，於適當之氧化劑量添加下進行典型高分子之氧化共聚合反應，以合成一系列有機無機混成之納米複合材料。所合成之納米複合材料以紅外光譜儀，X-光繞射儀及穿透式電子顯微鏡進行觀察與鑑定。所合成之材料將其溶於適當之溶劑中可做成自由立膜與塗料來進行阻水氣及防金屬腐蝕之應用研究。在阻水氣的研究上，發現無機層狀黏土的導入可有效降低水氣在高分子自由立膜的透過速率，並且隨著無機層狀黏土含量在納米複合材料中的增加，阻水氣的效果更加明顯。另一方面，在防金屬腐蝕之應用研究上，隨著無機層狀黏土的導入可有效提升整體塗料在金屬防腐蝕上的效果，並且隨著無機層狀黏土含量在納米複合材料中的增加，金屬防腐蝕的效果更加明顯。

關鍵詞：納米複合材料， 黏土， 苯胺， 阻氣， 腐蝕

Document Structure Analysis using Back-Propagation Network

Yuan-Kai Wang
Department of Electronic Engineering
Fu Jen Catholic University
Taipei, Taiwan 242, R.O.C.

Abstract

The digitalization of paper-based documents is an important issue in information era. The digitalization means the process of scanning paper documents, analyzing the layout of document image, and converting it into texts by Optical Character Recognition (OCR) technique. Due to that texts recognized by OCR do not contain structural information, digitized document cannot derive hierarchical structure. In this paper, we propose a document structure analysis technique by means of layout information and neural networks. The objective is to classify blocks in a document, such as title blocks and paragraph blocks, into a structural hierarchy and transfer it into an XML (eXtensible Markup Language) document automatically. Conference paper is chosen in our experiments. In the experiments, we can achieve 94.44% correctness rate using the back-propagation network (BPN) model. Experimental results validate the feasibility of the proposed scheme.

Key Words: Document Analysis, neural networks, back-propagation networks

INTRODUCTION

The objective of document structure analysis is to classify text blocks in document image and extract logical structure automatically. This is an interesting research topic since a fully automated document digitalization cannot be achieved without considering recovering document structure. In previous studies, traditional approaches are based on bottom-up method [1], both characteristics and threshold values obtained from statistical analysis of the pixel distribution of digital image data are applied. These approaches strongly depend on heuristic information derived from locally analyzed samples. They often failed to recognize various kinds of documents. Recently, approaches characterized as the top-down method had been proposed for the recognition of documents. The capability of recognition mechanism is based on global and conceptual knowledge, rather than on the local characteristics used by the statistical recognition approaches .

Knowledge-based methods play a very important role. The main characteristic is to define the knowledge about documents as document models and then interpret document images analytically using the document model [2].

To sum up, various methods [1, 3-10] are compared in Table 1. These methods have learning mechanism but they cannot learn automatically.

In this paper, we propose a recognition approach using layout knowledge and neural networks for understanding document blocks. Neural networks have the capabilities of mapping, fault-tolerance, learning, and parallel processing. These capabilities make neural network suitable for designing a complex classification methodology. Some neural networks use supervised network learning strategies to adjust weight. With the optimal weight matrixes, the neural network can analyze a new document.

Table 1. Comparisons of various document understanding methods

Characteristic Method	Approach	Learning
Esposito et al., 1990 [3]	Rule-based	O
Watanabe et al., 1993 [1]	Rule-based	X
Watanabe et al., 1995 [4]	Rule-based	X
Parmentier et al., 1997 [5]	Model-based	X
Walischewski, 1997 [6]	Heuristics-based	O
Lin et al., 1997 [7]	Heuristics-based	X
Watanabe et al., 1997 [8]	Rule-based	X
Brugger et al., 1998 [9]	Model-based	O
Chang, 1998 [10]	Model-based	X

2. The Framework

To analyze document structure, it means to identify the role of each text block in document structure. Since every kind of document has typical structure, a template about structure definition is needed. The template is a priori knowledge that includes all necessary and optional elements in that kind of document, and hierarchy, sequence, and occurrence information of elements. The template, of course, should not contain layout information.

In addition to template structure, our framework utilizes neural network to recognize document item blocks. The system is illustrated in Figure 1. A paper document processed by image segmentation, block identification, and OCR, will then subsequently processed by our approach, which includes geometry data transformation and neural learning and recognition. The logical structure tree in Figure 1 is a knowledge store that contains template structure. It is important that neural networks are usually adopted for "flat" (vector-based) or sequential representation.

3. Structure Analysis of Conference Paper

To learn the layout information about one sort of documents, we have to provide priori knowledge . The priori knowledge is the template structure of the kind of

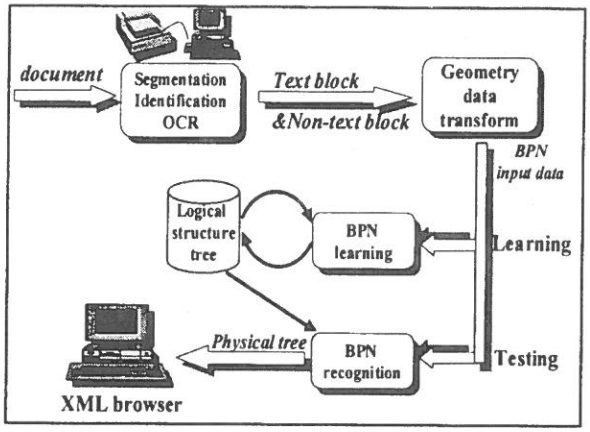

Fig. 1. The process of the document structure analysis with backpropagation network

documents with tree representation. Conference papers from the proceedings of International Conference on Document Analysis and Recognition (ICDAR) in 1997 are taken as our target for structure analysis.

Conference papers, with the very similar structure with journal papers, are composed of title, author, affiliation, abstract, section, sub-section, paragraph, reference, footnote, page number, graph, table and equation etc. Some elements are necessary, some can be ignored, some can appear with many times, and some must appear after some other elements.

The structure and relation of these elements are illustrated in the structure diagrams of Figure 2. It depicts where elements fit in terms of hierarchy, order, and occurrence. The notation "+" represents that the element can appear one time or more than one time (occurrences (1), and "*" represents none or more than one time (occurrences (0).

According to the structure diagram shown in Figure 2, it is easy to derive a document type definition (DTD). Shown in Table 2 is the DTD following XML specification. The terms in Figure 5 and Table 2, such as "+", "*", PCDATA, Empty, ",", and "|", are adopted from XML specification. In addition to "+" and "*", we use EMPTY to represent graph, table and equation elements in conference papers; PCDATA represents composite elements of document; "," represents order

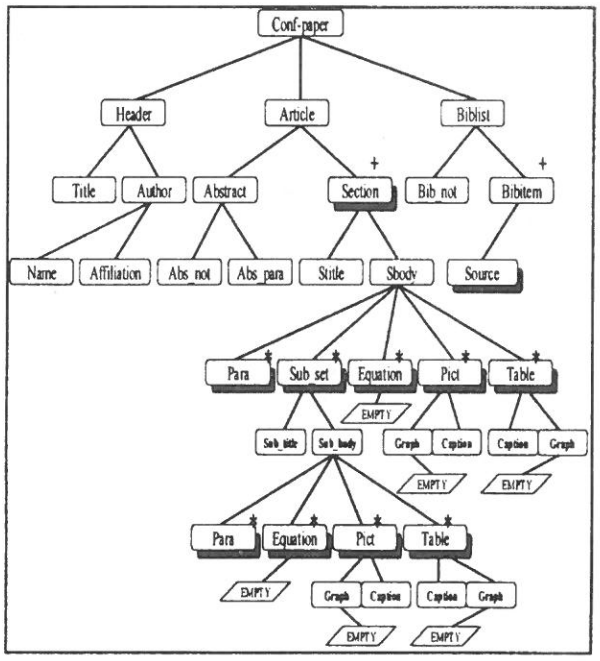

Fig. 2. Structure diagram of conference paper

sequence; and " | " represents logic OR relation. Footnotes and page numbers are ignored in our analysis and experiments for simplification.

Every element in table 2 corresponds to a text block in document image. The title element usually corresponds to the text block that has large font size with bold fontface, and is always at the top of a page. Other elements also have similar layout information. Since every element has corresponding and characteristic layout information, the information can help identify unknown text block.

4. Learning and Recognition

Learning is the process of automatically obtaining layout information for each element in document type definition. Recognition is the process of identifying an unknown text block as a characterized DTD element. Here we use backpropagation network to accomplish both processes.

Table 2. Document Type Definition of conference paper.

```
<! ELEMENT Paper (Header, Article, Biblist) >
<! ELEMENT Header (Title, Author) >
<! ELEMENT Article (Abstract, Section+) >
<! ELEMENT Biblist (Bib-not, Bibitem+) >
<! ELEMENT Author (Name, Affiliation) >
<! ELEMENT Abstract (Abs-not, Abs-para) >
<! ELEMENT Section (Stitle, Sbody) >
<! ELEMENT Bibitem (Source) >
<! ELEMENT Sbody (Para| Sub-set| Equation| Pict| Table) * >
<! ELEMENT Sub-set (Sub-title, Sub-body) >
<! ELEMENT Sub-body (Para| Equation| Pict| Table) * >
<! ELEMENT Pict (Graph, Caption) >
<! ELEMENT Table (Caption, Graph) >
<! ELEMENT Graph EMPTY>
<! ATTLIST Graph content ENTITY #REQUIRED>
<! ELEMENT Caption (#PCDATA) >
<! ELEMENT Equation EMPTY>
<! ATTLIST Equation content ENTITY #REQUIRED>
<! ELEMENT Table EMPTY>
<! ATTLIST Table content ENTITY #REQUIRED>
<! ELEMENT Title (#PCDATA) >
<! ELEMENT Name (#PCDATA) >
<! ELEMENT Affiliation (#PCDATA) >
<! ELEMENT Abs-not (#PCDATA) >
<! ELEMENT Abs-para (#PCDATA) >
<! ELEMENT Stitle (#PCDATA) >
<! ELEMENT Para (#PCDATA) >
<! ELEMENT Sub-title (#PCDATA) >
<! ELEMENT Bib-not (#PCDATA) >
<! ELEMENT Source (#PCDATA) >
```

4. 1 Geometry information conversion

The purpose of this stage is to adjust the input parameters of backpropagation network in both learning and recognition processes. To feed layout information into backpropagation network, coordinates of block are one possible information. However, it is too sensitive to be a good feature. A conversion stage is then proposed to preprocess the coordinates of block.

In this stage, document image is divided into $m \times n$ areas which is obtained by equally splitting the document image horizontally and vertically. Each area has an area

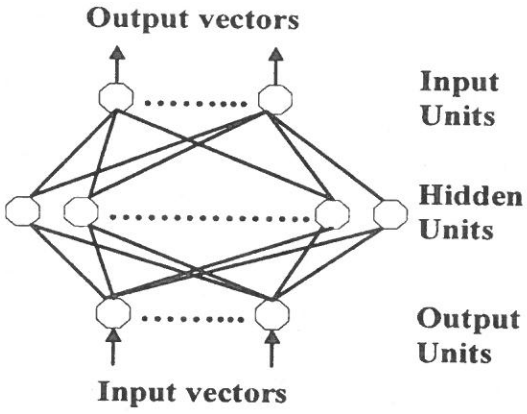

Fig. 3. The conversion of coordinates of block.

number. The two coordinates of each block, left-top and right-bottom corner, are converted into two area numbers. An example is shown in Figure 3. Document image is divided into 4×3 blocks. A block with left-top coordinate (196, 2162) and right-bottom coordinate (1164, 2394) will have two area numbers 2 and 6, since its left-top corner falls in the number 2 area and the bottom-right corner in the number 6 area.

4. 2 Backpropagation network

Backpropagation network (BPN) belongs to supervised learning network and suit to classification, and prediction. A BPN contains an input layer, hidden layer, and output layer. Input layer includes input units with given input characteristics. Hidden layer shows interactive effects of input units. Output layer contains output units for derived output of recognition.

BPN in common uses a nonlinear transfer function named sigmoid function which is

$$F(X) = \frac{1}{1+e^{-x}}$$

To evaluate the learning performance of backpropagation network, we often classify learning and testing examples during collection. In network learning period, we can input testing examples while learning cycle. We evaluate the convergence of BPN

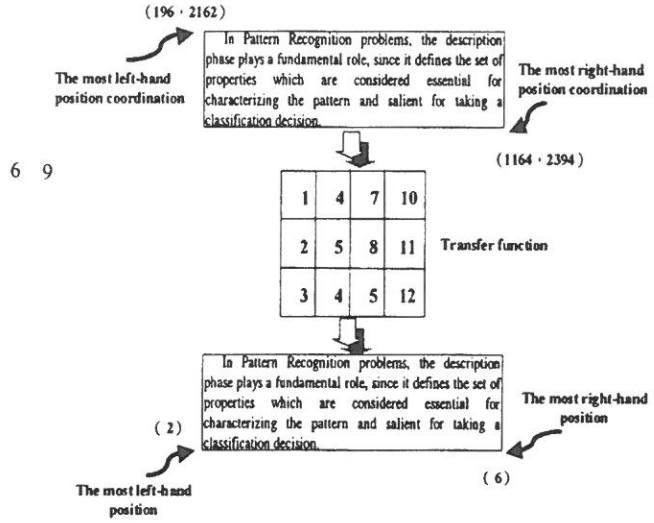

Fig. 4. Backpropagation network.

classification by total error rate. Its value range between 0 and 1. The equation is following:

$$\text{total} - \text{error} - \text{rate} = \frac{\text{total error examples}}{\text{sum of examples}}$$

It is necessary to analyze geometry characteristics and mutual relation of document elements, and then learn and classify by using neural network to map logic structure of document for document understanding.

5. Empirical Experiments and Analysis

Our hardware includes a Pentium 166MMX PC and a Mustek 600CP scanner. Programs are implemented by C. Experimental samples contain ten conference papers from ICDAR' 97. Each paper has many blocks such as title block, name block, and so on. Nine papers are taken for training, and one for testing. The number of block for learning and testing is show in Table 3.

Although conference papers have formal and fixed format, the following questions still complicate the structure analysis problem:

Table 3. Number of blocks for learning and testing.

Purpose Block	Learning	Testing
Title	9	1
Name	10	1
Affiliation	10	1
Abs-not (Abstract notation)	9	1
Abs-par (Abstract paragraph)	9	1
Stitle (Section title)	49	4
Paragraph	389	46
Sub-title (Sub-section title)	28	3
Caption	53	3
Bib-not (Bibliography notation)	9	1
Bibitem (Bibliogrpahy item)	87	14
Source (Bibliogrpahy content)	89	14
Sum	751	90

* Because "paragraph" is the most popular block in learning examples, testing blocks are easy to be recognized as "paragraph".
* "paragraph" can appear anywhere such that the position information of "paragraph" will confuse the recognition of other blocks.
* There is the same parameter for different elements such as line height such that "Stitle" and "Sub-title" can have the same line height.
* Scanned paper image may be inclined.
* A complete paragraph may cross pages.

There are three tests with different neural parameters. The parameters and results are explained as follows.

For the first testing, there are five nodes in the input layer of backropagation network. The five nodes represent five parameters: block width, block length, average line height in the block, the left-top area number, and the right-bottom area number. Output layer has twelve nodes representing DTD elements such as Title, Name, Affiliation, Abs-not, Abs-para, Stitle, Para, Sub-title, Caption, Bib-not, Bibitem, and Source. There are eight nodes in hidden layer.

Recognition rate of the first test is 86.67%. The convergence performance is illustrated in Figure 5. A confusion matrix is tabulated in Table 4 to analyze errors.

There are twelve errors among 90 testing examples. Title is recognized as Affiliation that appears in header of paper. Name is recognized as Stitle that has one line. Abs-para is recognized Para that is paragraph text of paper. Para is recognized as Stitle or Caption because Para may appear one line. Sub-title is recognized as Stitle that has one line. Caption is recognized as Para.

Table 4. Confusion matrix of test 1.

TARGET INFERENCE	SOURCE	BIB-ITEM	BIB-NOT	CAPT-ION	SUB-TITLE	PARA GRAPH	STITLE	ABS-PARA	ABS-NOT	AFFILI-ATION	Name	Title	SUM
SOURCE	14	0	0	0	0	0	0	0	0	0	0	0	14
BIBITEM	0	14	0	0	0	0	0	0	0	0	0	0	14
BIB-NOT	0	0	1	0	0	0	0	0	0	0	0	0	1
CAPTION	0	0	0	0	0	2	1	0	0	0	0	0	3
SUB-TITLE	0	0	0	0	1	0	0	0	0	0	0	0	1
PARAGRAPH	0	0	0	3	0	43	0	1	0	0	0	0	47
STITLE	0	0	0	0	2	1	3	0	0	0	1	0	7
ABS-PARA	0	0	0	0	0	0	0	0	0	0	0	0	0
ABS-NOT	0	0	0	0	0	0	0	0	1	0	0	0	1
AFFILIATIOON	0	0	0	0	0	0	0	0	0	1	0	1	2
Name	0	0	0	0	0	0	0	0	0	0	0	0	0
Title	0	0	0	0	0	0	0	0	0	0	0	0	0
Sum	14	14	1	3	3	46	4	1	1	1	1	1	90

For the second test, the input layer has two more nodes: one for bold typeface of block text, and another for italic typeface. The hidden layer has nine parameters, and the output layer still has twelve nodes.

The recognition rate of test 2 is 92.22%. The convergence performance is also illustrated in Figure 5. The test 2 has better performance than test 1. Confusion matrix is shown in Table 5. There are seven incorrect testing examples. Name is recognized as Para. Stitle is recognized Sub－title or Caption that may appear one line and bold posture. Para is recognized as Caption that may appear one line and normal posture. Caption is recognized as Title that appears bold posture.

The third test fixes the typeface of title element as bold, name and affiliation as normal, abstract as italic, paragraph as normal, and bibliography as normal. The recognition rate is 94.44%. Its result is also shown in Figure 5. Confusion matrix is shown in Table 6. There are five incorrect testing examples. The incorrect results of

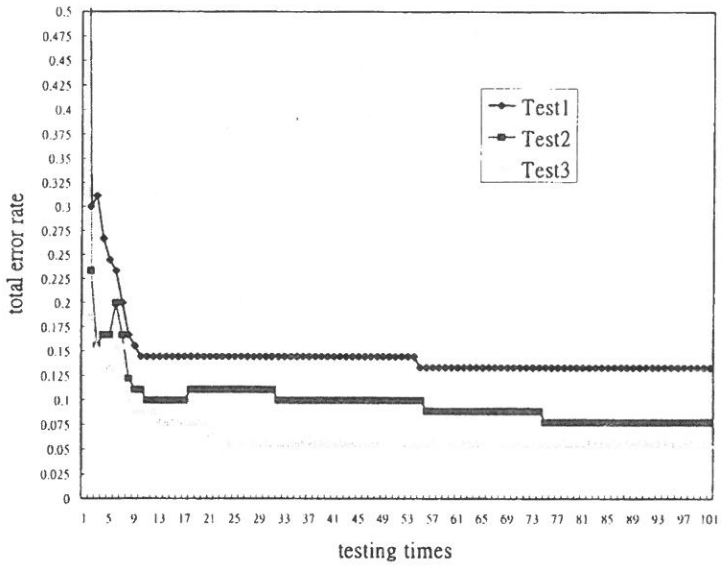

Fig. 5. Convergence performance.

Table 5. Confusion matrix of test 2.

TARGET INFERENCE	SOURCE	BIB-ITEM	BIB-NOT	CAPT-ION	SUB-TITLE	PARA GRAPH	STITLE	ABS-PARA	ABS-NOT	AFFILI-ATION	Name	Title	SUM
SOURCE	14	0	0	0	0	0	0	0	0	0	0	0	14
BIBITEM	0	14	0	0	0	0	0	0	0	0	0	0	14
BIB-NOT	0	0	1	0	0	0	0	0	0	0	0	0	1
CAPTION	0	0	0	1	0	1	1	0	0	0	0	0	3
SUB-TITLE	0	0	0	0	2	0	1	0	0	0	0	0	3
PARAGRAPH	0	0	0	0	0	45	0	0	0	0	1	0	46
STITLE	0	0	0	0	1	0	2	0	0	0	0	0	3
ABS-PARA	0	0	0	0	0	0	0	1	0	0	0	0	1
ABS-NOT	0	0	0	0	0	0	0	0	1	0	0	0	1
AFFILIATIOON	0	0	0	0	0	0	0	0	0	1	0	0	1
Name	0	0	0	0	0	0	0	0	0	0	0	0	0
Title	0	0	0	2	0	0	0	0	0	0	0	1	1
Sum	14	14	1	3	3	46	4	1	1	1	1	1	90

Name and Sub－title are the same as Part II. Stitle is recognized as Caption that may appear one line and bold posture. Bib-not is recognized as Abs－not that appears one line and bold posture.

Table 6. Confusion matrix of test 3.

TARGET INFERENCE	SOURCE	BIB-ITEM	BIB-NOT	CAPT-ION	SUB-TITLE	PARA GRAPH	STITLE	ABS-PARA	ABS-NOT	AFFILI-ATION	Name	Title	SUM
SOURCE	14	0	0	0	0	0	0	0	0	0	0	0	14
BIBITEM	0	14	0	0	0	0	0	0	0	0	0	0	14
BIB-NOT	0	0	0	0	0	0	0	0	0	0	0	0	0
CAPTION	0	0	0	3	0	0	2	0	0	0	0	0	5
SUB-TITLE	0	0	0	0	2	0	0	0	0	0	0	0	2
PARAGRAPH	0	0	0	0	0	46	0	0	0	0	1	0	47
STITLE	0	0	0	0	1	0	2	0	0	0	0	0	3
ABS-PARA	0	0	0	0	0	0	0	1	0	0	0	0	1
ABS-NOT	0	0	1	0	0	0	0	0	1	0	0	0	2
AFFILIATIOON	0	0	0	0	0	0	0	0	0	1	0	0	1
Name	0	0	0	0	0	0	0	0	0	0	0	0	0
Title	0	0	0	0	0	0	0	0	0	0	0	1	1
Sum	14	14	1	3	3	46	4	1	1	1	1	1	90

6. Conclusions and Future Works

Layout information is useful for document element classification and structure analysis. We analyze document structure structure diagram that implies the hierarchy, order, and occurrence of element, and construct document type definition for conference paper. Neural Networks provide inference and learning capabilities and experimental results to validate the feasibility of the proposed scheme. The more fixed the layout characteristics present, the more correct the recognition result is.

Technology progresses rapidly, and the capability of software and hardware advance substantially. As a result, the recognition result of OCR will be more correct. We get geometry information of text blocks easily in the process of segmentation and these are useful to document image process. Our preliminary study focuses on text blocks. However, it will be useful if we add the information from non-text blocks.

In the future, we will develop XML editor and tools that automatically edit document DTD according to document logical structure, then we use text blocks and non-text blocks of OCR recognition and neural networks to learn and classify document elements, and automatically tag and produces XML documents according to DTD. We will also develop a robust XML parser that automatically correct recognized wrong elements from neural network.

REFERENCE

(1) T. Watanabe, Q. Luo and N. Sugie, "Structure recognition methods for various types of documents," International Journal on Machine Vision and Applications, pp. 163-176, 1993.

(2) T. Watanabe, Q. Luo and N. Sugie, "A Cooperative Document Understanding Method among Multiple Recognition Procedures," Proc. of 11th ICPR, pp. 689-692, 1992.

(3) F. Esposito, D. Malerba and G. Semeraro, "An Experimental Page Layout Recognition System for Office Document Automatic Classification: An Integrated Approach for Inductive Generalization," Proc. of 10th ICPR, pp. 557-562, 1990.

(4) T. Watanabe, Q. Luo and N. Sugie, "Layout Recognition of Multi-kinds of Table-form Documents," IEEE Tran. On PAMI, Vol. 17, No. 4, pp. 432-445, 1995.

(5) F. Parmentier and A. Belaid, "Logical Structure Recognition of Scientific Bibliographic References" Proc. of the 5th ICDAR, 1997.

(6) Hanno Walischewski, "Automatic Knowledge Acquisition for Spatial Document Interpretation," Prco. of the 5th ICDAR, 1997.

(7) C. C. Lin, Y. Niwa and S. Narita, "Logical Structure Analysis of Book Document Images Using Contents Information", Proc. of the 5th ICDAR, 1997.

(8) T. Watanabe and X. Huang, "Automatic Acquisition of Layout Knowledge for Understanding Business Cards" Proc. of the 5th ICDAR, 1997.

(9) R. Brugger, F. Bapst and R. Ingold, "A DTD Extension for Document Structure Recognition". 7th Int. Conf. on Electronic Publishing, Lecture Notes in Computer Science 1375, Springer, 1998.

(10) T. Y. Chang, Automatic Document of Technical Manuals in CALS Strategy, Master thesis, Graduate School of Defence Information, National Defence Management College, Taiwan, R. O. C., 1998.

received October 28, 2000
accepted December 6, 2000

以倒傳遞類神經網路進行文件結構分析

王 元 凱
輔仁大學電子工程系

摘 要

書面文件數位化在資訊化的時代是一個重要的研究問題。透過影像處理與圖形識別的技術，可大量將書面文件掃瞄成文件影像，運用光學文字識別將影像轉換爲文字檔。然而由於文件中的文字含有結構化的資訊，若能分析文件結構並抽取出階層式的資訊，將可以進一步提供處理全文檢索。本篇論文提出一個運用類神經網路來進行文件結構分析的方法，將根據圖文分離後得到的區塊分類訊息，進行辨認而獲得結構資訊，並將此結構資訊轉換爲 XML 文件。本研究方法經過實驗證明，對於科技會議論文可以有94.44%的正確率。

關鍵詞： 文件分析，類神經網路，倒傳遞神經網路

Hypo-Osmotic Stress Induced Immune Suppression and Apoptosis in Mice

Shiu-Huey Chou, Chih-Wei, Chen,
Chien-Chang Chiu, Huei-Ling Lian
Department of Biology
Fu Jen Catholic University
Taipei, Taiwan 242, R. O. C.

Abstract

The effect of hypo-osmotic stress on immune changes was examined in this study. Male, inbred BALB/c mice were received the intraperitoneal administration of 0.5 or 0.2% saline solution to induce acute hypo-osmotic stress response. Mice received 0.9% saline injection as control. Results showed that stressed animals have higher hemoglobin release in plasma than control animals. Hypo-osmotic stress significantly suppressed T- and B- lymphocytes proliferation to mitogen Con A and LPS, but increased the proliferation of bone marrow derived granulocyte-monocyte precursors in stress mice. There was no change of nitric oxide production between stress and control groups. Additionally, hypo-osmotic stress enhanced the percentage of corticosterone-induced apoptoic cells and dead cell population in the thymus of stressed mice. Ladder-like DNA fragmentation was also increased in the thymus of mice received 0.5% or 0.2% saline injection. Thus, this study suggests that the hypo-osmotic stress-induced immune suppression of spleen and thymus may be caused by stress-mediated apoptosis *in vivo*.

Key Words: hypo-osmotic stress, hemoglobin, apoptosis, corticosterone, and epinephrine, DNA fragmentation

INTRODUCTION

It is well accepted the concept that stress contributes to pathophysiology of nearly all psychiatric illnesses and a range of medical illness (Joasod and McKenzie, 1976; Keller et al., 1981; Kiecolt-Glaser et al., 1984; Cohen and Williamson, 1991). Many illnesses are exacerbated by improper dietary intake or water output (Schuppan et al., 1995). Osmosis is major cellular mechanism to maintenance the body physiology homeostasis. Recent investigations have shown that change of cell osmolarity can significantly affect the cellular mediated immune function and exacerbate the sensibility to pathogen invading. For example, hypertonic saline injection after trauma can increased human peripheral blood monocyte number, T-cell proliferation, and the delayed-type hypersensitivity reaction (Junger et al., 1994). Dehydration after exercise can decrease the ratio of $CD4^+$ to $CD8^+$ in man peripheral blood (Greenleaf et al., 1995). Moreover, Exposure to hyperosmotic condition increased the proinflammatory cytokines such as IL-1α, IL-1β, TNF-α, and IL-8 production (Shapiro and Dinarello; 1997).

The balance between cell proliferation and programmed cell death is a key factor to regulate the cellular homeostasis of the immune system. It has been reported that programmed cell death or apoptosis plays an important role in development and regulation of the immune response (Wyllie et al., 1980; Raff, MC, 1992). The study of osmotic stress indicated those stress signal induced Ras activation and O_2^- formation and subsequently induced apoptosis in Jurkat T cell line (Gulbins et al., 1997). In addition, Sorbitol-induced osmotic stress can caused P38 protein and protein kinase C mediated apoptosis in rat alveolar type II cells *in vitro* (Yasmin et al., 1998). However, most reporters of the effect of osmotic stress on immune function were done *in vitro* estimation. There are not many reports on *in vivo* animal study.

Therefore, the purpose of this research was to characterize the effect of the hypo-osmotic stress on multiple measures of immune function using mice. To evaluate immune function, this study included a measure of the proliferative response of lymphocytes to T and B ell mitogens, nitric oxide production from mitogen-stimulated cell culture, and thymus and spleen apoptosis induced by stress hormones such as

corticosterone or epinephrine. Additionally, measurement of plasma hemoglobin level was used as indicator of a stress response. This study suggested that the out of balance in cell osmolarity can induce immune suppression *in vivo* system and this suppression may be caused by stress-mediated apoptosis.

MATERIALS AND METHODS

Mice

Inbred male mice of the BALB/c strain (10-12 weeks) were purchased from Laboratory animal center in National Taiwan University College of Medicine (Taipei, Taiwan R.O.C.). Upon arrival, mice were caged in a colony room where a 12-hour light-dark cycle was maintained through artificial illumination. They received free access to both food and water throughout the experiment except as noted and a 2-week acclimation period prior to the experimental manipulation. Additionally, animals were handled on a bidaily basis to prevent the hyperactivity that occurs when they are left untouched for a prolonged period. All animal manipulation were followed in the rule of *Guide for the Care and Use of Laboratory* (NRC, 1996).

Reagents and Chemicals

Complete culture medium (CCM) consisted of RPMI 1640 supplemented with L-glutamine (2mM), Hepes (10mM), gentamycin (50 μg/ml) and 10% heat-inactive fetal bovine serum (HIFBS). The mitogens Con A and lipopolysaccharide (LPS) from *Escherichia coli* and Herchest 33342 and properdium iodide were purchased from Sigma (St. Louis, MO, USA). Thymocytes, splenocytes and bone marrow cells were collected used collecting buffer that consisted of Dulbecco's Phosphate Buffer Saline Solution without calcium and magnesium (cmf-DPBS) supplemented with 10 mM Hepes, 2mM EDTA, 50 μg/ml gentamycin and 2% HIFBS. Red blood cells were removed used osmotic lysing buffer: 150mM ammonium chloride, 10 mM sodium bicarbonate, and 1 mM EDTA.

Experimental Paradigm

Mice were randomly assigned to one control or two experimental groups (n = 4/

group). Control mice received an injection of sterile 0.9 % saline solution. The experimental groups of animals received either 0.5% saline or 0.2% saline solution injection to induce hypotonic osmotic stress. Each injection was 3 ml / per mice administrated intrapertoneal (IP). For 2 h following each injection of 0.2, 0.5, 0.9% saline, all animals were deprived of food and water. Two hours following the injection, each subject was anesthesia by ether gas. One ml blood was harvested from heart. After blood collected, Mice were rapidly sacrificed by cervical dislocation.

Tissue Collection and Preparation of Leukocytes

Tissue and blood were collected as described previously (Chou et al., 1997).

Blood. After sacrifice, the animal was placed on its back. Whole blood (1 ml) was collected from each subject's heart using heart puncture through 22-gauge needle into 3.8% sodium citrate contained syringe. Blood was transferred into 1.5 ml microcentrifuged tube and storage in the ice until assay.

Spleen and Thymus. Each subject's spleen and thymus were dissociated into a single-cell suspension by grinding the spleen between the frosted ends of two glass slides. The single-cell suspension was collected after 8 min of unit gravity sedimentation to remove cell debris and clump. Subsequently, the splenocytes and thymocytes were enumerated and diluted to 5×10^6/ml with completed culture medium.

Bone marrow cells. Bone marrow cells were harvested by flushing the femoral and tibia shafts with collecting buffer. Red blood cells were lysed with osmotic lysing buffer and washed three times with collecting buffer and CCM. The cell number and viability of bone marrow cells were determined by 0.2% trypan blue. The differentation of bone marrow cells to granulocyte-monocyte precursors (GM-p) was induced by cultured bone marrow cells with optimal concentraiton L929 cells culture-derived conditional medium. One hundred μl of 5×10^6/ml bone marrow cells mixed with equal volume conditional medium and culture at 37 ℃ for 48 hours. The proliferation of GM-p was determinated by MTT assy (detail referred in mitogen proliferation assay).

Hemoglobin determination

The effective of hypotonic solution induced stress response was estimated by relative hemoglobin level in the blood. The method to measure hemoglobin in blood was modified by Boyle et al. (1979). Brief, 100 μl fresh whole blood was mixed with 900 μl D-PBS, then centrifuged at 1500 rpm for 10 min. 600 μl of supernatant was collected from each treated mice and the hemoglobin release level was determined on spectrophotometer (U-2000, Hitachi, Japan) by measuring optical density at a wavelength of 410 nm.

Mitogen Proliferation Assay

The mitogen proliferation assay was performed with thymocytes and splenocytes to assess lymphocyte proliferation. The method was modified from Chou et al. (1996). Con A was used as T -lymphocyte mitogen and LPS as a B-lymphocyte mitogen. The proliferation response was evulated with the reduction ratio of MTT (Sigma, Chemical Co, USA) by mitochondrial dehydrogenase of viable cells to a blue formzan product which can be measured spectrophotometrically. Brief, 100 μl of 5×10^6 splenocytes or thymocytes were incubated with Con A (1.5 μg/ml for spleen, 3 μg/ml for thymus) or LPS (2.5 μg/ml for spleen, 5 μg/ml for thymus) for 48 hours. In the end of incubation, 10 μl of stock MTT solution (5 mg/ml) was added to all wells of an assay and plated were incubated at 37 ℃ for 30 min -1 hour. Acid-isopropanol (100 μl of 0.04 N HCl in isopropanol) was added to all wells and mixed thoroughly to dissolve the dark blue crystals. After a few minutes at room temperature to ensure that all crystals were dissolved, the optical density was measured in an ELISA microtiter plate reader (BioRad 550, USA) at 570 nm.

Nitric Oxide Assay

Supernatants were harvested from cultures of mitogen-stimulated lymphocytes prior to add with MTT. The nitric oxide concentrations were determined as nitrite by a microplate assay (Stuehr and Nathan, 1989). One hundred μl sample aliquots, were mixed with 100 μl of Griess reagent (0.5% sulfanilamide / 0.05% naphthylethylene

diamine dihydrochloride / 2.5% H_3PO_4) and incubated at room temperature for 10 min. The optical density was measured in an ELISA microtiter plate reader at 550 nm against PBS as a blank. A sodium nitrite ($NaNO_2$) standard curve (0-100 μmole/ml) was generated in parallel. Results are expressed as the mean of nitrite concentrations in the supernatants of triplicate wells and analyzed as repeated-measures ANOVA by using the no-mitogen control and all concentrations of a given mitogen.

Corticosterone and Epinephrine-Induced Apoptosis Assay

Apoptotic cells were quantified by fluorescent microscope to measure the incorporation of Hoechst 33342 (Sigma, St. Louis, USA) fluorescent dye in thymocytes (Sun et al., 1992). A single cell suspension of thymocytes (2×10^6) were incubated with or without corticosterone (0, 0.1, & 1.0 μM) and splenocytes were incubated with or without epinephrine (0 & 10 μM) for four hours and then stained for 10 min. with 1 μg/ml Hoechst 33342 at 37 ℃. The cells were rapidly centrifuged; (5 min. 450 xg), and all media and Hoechst dye were removed. The thymocytes and splenocytes were resuspended in propidium iodide (PI, 5 μg/ml; Sigma) in PBS and analyzed on phase-contrast fluorescent microscopy (Nikon Labophot-212A, Japan). The cells staining positive for PI were excluded from the analysis because they constitute dead cells. The apoptotic cell was identified by a decrease in mean size and bright blue staining with Hoechst dye. Data was reported as percentage apoptotic cells and dead cells.

Gel Electrophoresis

The apoptosis induced DNA fragment was expressed by agarose gel electrophoresis. The method to run gel was modified from Perandones et al., 1993. Brief, at the end of corticosterone or epinephrine incubation time, 1×10^6 thymocytes or splenocytes were harvested and wash with D-PBS twice, then cell pellets was centrifuged and collected into the tube. Cell pellets were resuspended in 400 μl hypotonic lysis buffer (0.2% Triton X-100, 10mM Tris and 1 mM EDTA, PH=8) and centrifuged 15 min at 13800 xg. The supernatant was collected and treated with an equal volume of absolute isopropyl alcohol and 0.5 M NaCl to precipitate the DNA. The samples were stored at -20 ℃ overnight, and centrifuged at 13800 x g for 15 min.

The pellet was washed with 200 μl of 70% ethanol, then allowed to dry at room temperature. The DNA was resuspended with 12 μl of TE solution (10 mM Tris Cl, 1mM EDTA, pH7.4) 3 μl loading buffer (50% glycerol, 1X TAE), incubated at 37 ℃ for 20 min, then electrophoresed on 1% agarose gel containing ethidium bromide for 1 hour. Gels were photographed using UV transillumination (Image master -VDS Pharmacia Biotech, USA).

Tumor Necrosis Factor (TNF) Assay

The concentration of TNF in the supernatants of Con A- or LPS-stimulated spleen culture was measured. Whole spleen leukocytes (2.5×10^6/ml) were incubated in complete media and Con A (3μg/ml) or LPS (5μg/ml) for 48-hr at 37℃ in 5% CO_2. Culture supernatants were harvested and stored at − 70℃ until the assay. TNF production by stimulated splenocytes was modified after that described by Wang et al. (1985) and Hogan and Vogel (1988). The ability of splenocytes to produce TNF was determined by using a TNF-sensitive, actinomycin D-treated murine L929 fibroblasts cell line. This assay, however, cannot distinguish between the α and β form of TNF. Brief, L929 cells were collected and washed once in complete media. The cells were diluted to 2.5×10^5/ml and 100μl of suspension added to a 96 wells plate. The plate was incubated 18-24hr. At this time, the growth of L929 cells were stopped by adding 10μl of actinomycin-D. Later, 100μl of diluted supernatants were add to wells in triplicate and incubated for 24-hr at 37℃ in CO_2 incubator. The plate was washed twice with phosphate-buffered saline (PBS) and stained with 0.2% crystal violet / 2% formalin in PBS. The fixed cells were washed several times with tap water and allowed to dry overnight at room temperature. Ethanol (50%) in PBS was added to each well of plates to dissolve the stain-fixed cells. The optical density was measured in an ELISA microtiter plate reader at 550-595 nm. A TNF standard curve (0-200 units/ml) was generated in parallel on each plate. The TNF content of each sample was calculated from the standard curve and reported as units/ml.

Statistical Treatment of Data

A computerized program for analysis of variance (ANOVA; Statistix, NH

Analytical Software) was used to assess differences among experimental and control groups. For the present experiment to characterize the effects of osmotic stress an ANOVA to compare 3 stress-levels (0.9 % saline injection/non-stress, 0.5% and 0.2% saline injection/stress) was used to assess significance of the various measures. The level of significance for the F test was set at P less and equal than 0.05.

RESULTS AND DISCUSSION

Hypotonic-induced osmotic stress increased the release of hemoglobin in the blood plasma

An elevated level of hemoglobin in blood plasma provided an indicator of stress (Figure 1). The plasma supernatant in mice which received 0.5% and 0.2% saline injection had shown significantly higher absorbency at 410 nm than mice that were received 0.9% saline injection (0.74 for 0.5% saline treated mice, 0.695 for 0.2% saline injected mice, and 0.373 for control mice), $p < 0.05$. However, there was no difference between groups that received 0.5% and 0.2% saline injection. The leak or rupture of leukocytes in the peripheral induced increased hemoglobin level in the plasma. The morphology of leukocytes was shown the increase of cell size and shape of cells under microscopic observation (data was not shown). This result indicated that administration of hypotonic solution into mice induced a effectively osmotic stress response in experiment subjects.

Hypotonic-induced osmotic stress decreased the mitogen-stimulated lymphocyte proliferation in spleen and thymus, but increased the granulocyte-monocyte precursor (GM-p) proliferation in bone marrow

The proliferation response of splenic leukocytes and thymocytes to various mitogens was measured as an indicator of functional potential. The result indicated that the hypotonic-induced osmotic stress significantly decreased the splenic (Figure 2-A) and thymic (Figure 3-A) T - lymphocytes proliferation activity to mitogen Con A, $P < 0.05$. However, there was no significantly difference between groups that were received 0.5% or 0.2% saline injection for spleen and thymus T-lymphocyte proliferation . Suppression of Con A mitogen suggested that *in vivo* cell - mediated

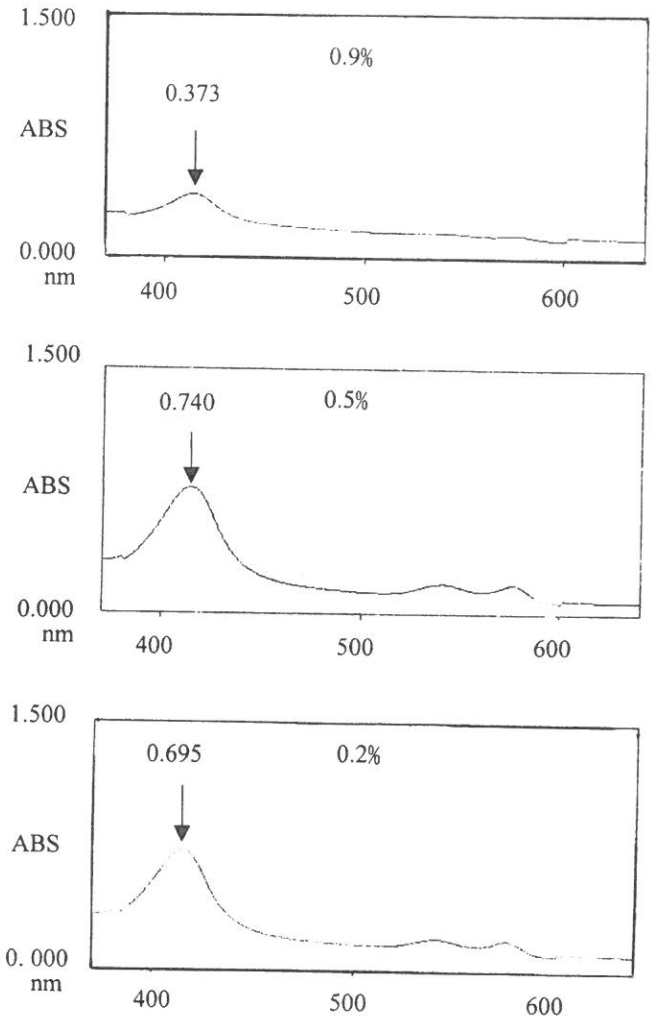

Fig. 1. The change of relative hemoglobin levels in heart blood for mice received 0.9 (top), 0.5 (middle), and 0.2% (button) saline injection. The absorbency value of OD 410 was indicated by arrow symbol. Data only presented for one mice from each treated group.

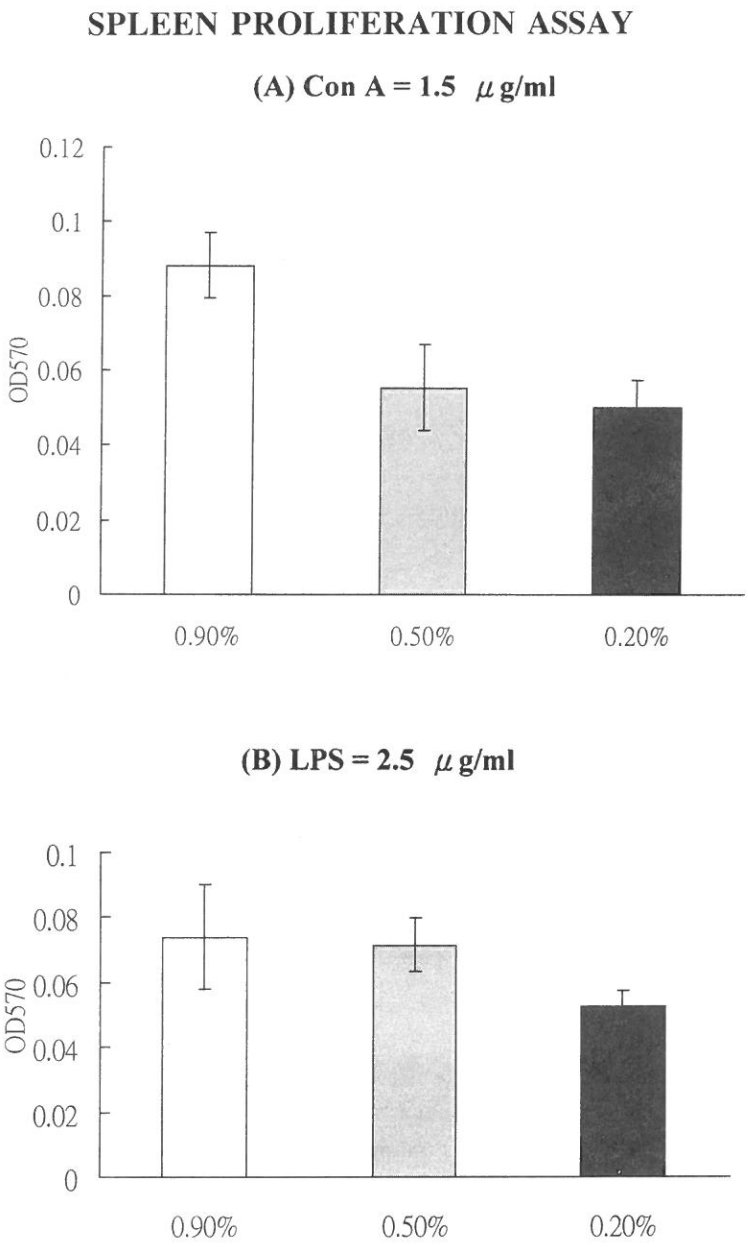

Fig. 2. Proliferative responses of spleen leukocytes to the optimal concentrations of mitogen Con A (panel A) and LPS (panel B) as noted in A - B. Mice received 0.9 or 0.5 or 0.2% saline injection. Results are expressed as the mean (±SE) in absorbance (OD); n = 4/group.

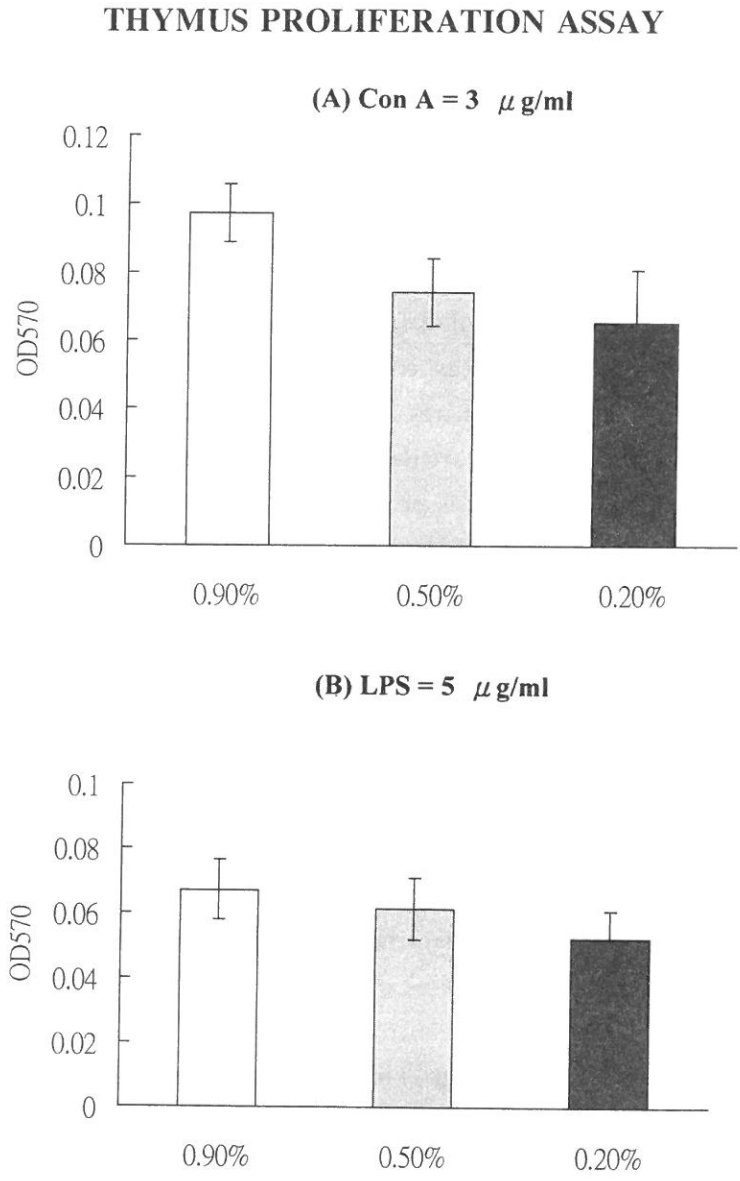

Fig. 3. Proliferative responses of thymus leukocytes to the optimal concentrations of mitogen Con A (panel A) and LPS (panel B) as noted in A – B. Mice received 0.9 or 0.5 or 0.2 saline injection. Results are expressed as the mean (±SE) in absorbance (OD); 4/group.

immunity can be suppressed by osmotic stress. In addition, the LPS-stimulated proliferation of spleen and thymus were also significantly decreased in mice that were received 0.2% saline injection, $p < 0.05$ (Figure 2B and 3B). Suppression of LPS mitogen suggested that humoral immunity also was affected by osmotic stress. Inversely, the hypo-osmotic stress increased the proliferation of bone marrow-derived granulocyte-monocyte precursor, $p < 0.05$ (Figure 4). Stressed-mice received 0.5 or 0.2 % saline injection presented higher GM-p proliferation response than control mice did. In general, stress can induce leukopenia because of production of stress hormones, in addition, stress hormones such as corticosterone released during physiologically conditions plays an important role in regulating the trafficking of leukocytes between the blood and other immune compartments (Dhabhar et al., 1995; Chou et al., 1996). Therefore, result of increase proliferation in marrow-derived GM-p can be explained that body compromises to the decrease in thymus and spleen lymphocytes proliferation for maintenance of homeostasis.

The hypo-osmotic stress-induced immune suppression was not caused by nitric oxide but induced by DNA apoptosis.

In conjunction with mitogen-induced blastogenesis, the production of nitric oxide (NO) was measured in the same culture prior to the addition of MTT. The NO production of mononuclear cells in spleen and thymus was measured as an indicator of macrophage function. As nitric oxide was rapidly reduced to the stable products nitrite and nitrate, NO production was quantified from the concentration of nitrite. The result indicated that hypo-osmotic stress did not change the production of NO in the spleen and thymus in comparison to control mice (Table 1). There is no significantly difference in control and stress mice. This result may suggest that the decrease of lymphocyte proliferation to mitogen stimulation may not induce by nitric oxide. It is well known that stress affects the cytokine production (Chou et al., 1996). Hence, it is not surprising that hypo-osmotic stress also disturbs the balance of TNF production. TNF is an inflammatory and antiviral cytokine. It has been reported that TNF plays important roles in cell activation, proliferation and apoptosis (Baker and Reedy, 1996). Results from this study indicated that hypo-osmotic stress decreased the production of TNF in Con A-stimulated splenocytes culture of mice that received 0.5%

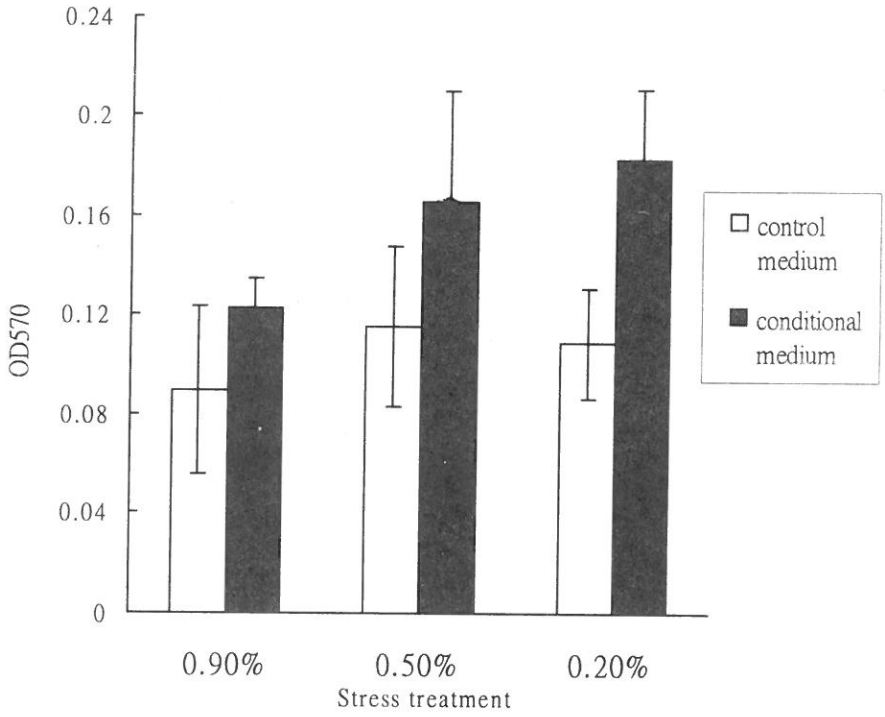

Fig. 4. Proliferation response of bone marrow cells to non-conditional (open bar) or conditional medium (solid bar) for mice received 0.9, 0.5, 0.2% saline solution injection. Conditional medium was prepared by collected culture supernatant from 2-day L929 cell culture. Results are expressed as the mean (±SE) in absorbance (OD); n = 4/group.

or 0.2% saline injection (Figure 6-A), but mild decreased in LPS-stimulated splenocytes culture of mice that received 0.2% saline injection (Figure 6-B). However, hypo-osmotic stress significantly increased the percent of thymic apoptoic cells in mice received 0.5% or 0.2% saline injection, $p < 0.05$ (Table 2). Moreover, the percent of dead cells in thymus mild increased in stress mice. DNA fragmentation and nuclear shrinkage are one of characteristic morphologic changes in apoptosis. Hence, DNA fragmentation from control and stress mice thymocytes supernatant was

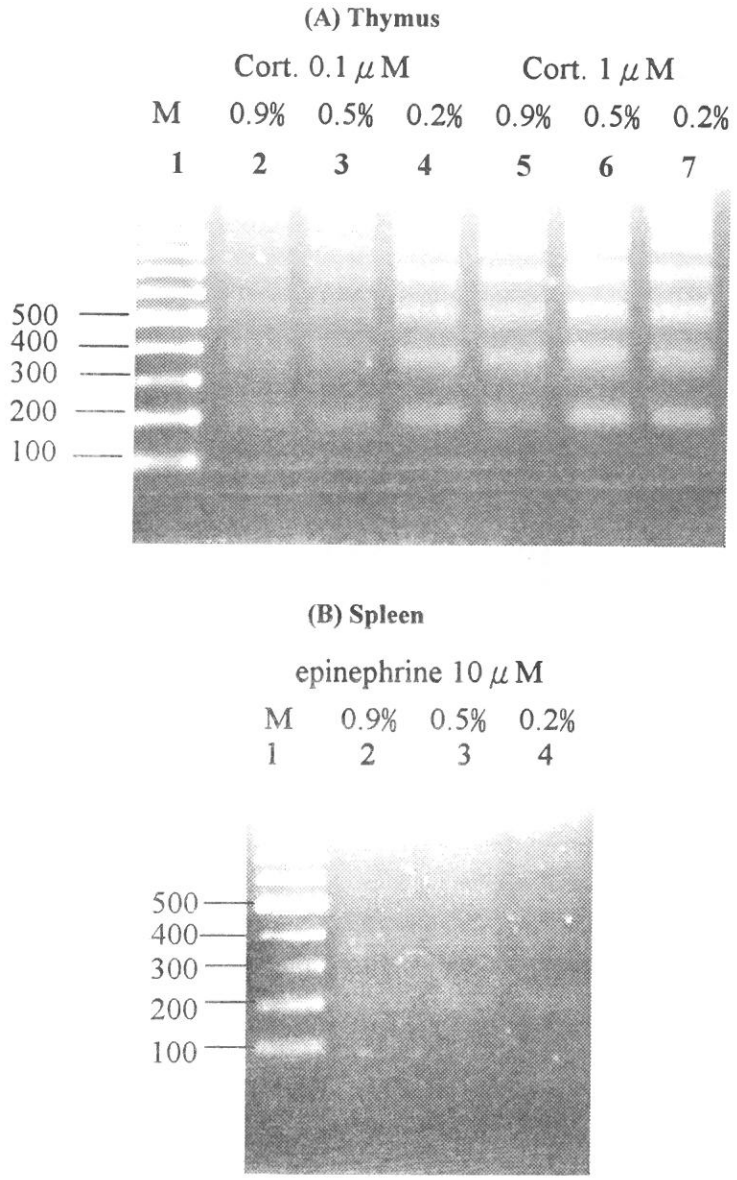

Fig. 5. The degree of DNA fragment in thymus (panel A) and spleen (panel B) for mice received 0.9, 0.5, or 0.2% saline injection. 1-2 × 10^6 thymocytes and splenocytes were isolated from each group mice and cultured with corticosterone (0.1 and 1 μM) or epinephrine (10μM) for 4 hours. Small DNA fragments from supernatant were harvested and estimated with 1% agarose gel electrophoresis. A 100-bp ladder was run as a size standard.

SPLEEN TUMOR NECROSIS FACTOR PRODUCTION

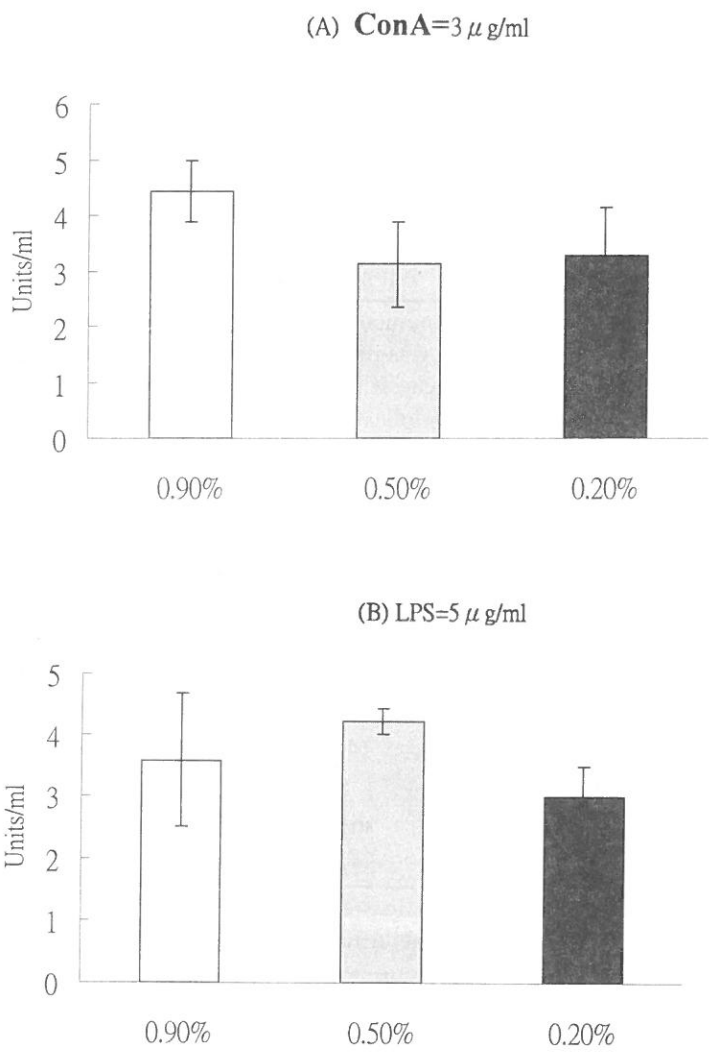

Fig. 6. Effects of hypo-osmotic stress on the TNF production by Con A (A) and LPS (B) stimulated splenocytes from stressed and non-stressed mice. Stressed mice received 0.5 or 0.2% saline solution injection, and non-stressed mice received an equal volume of saline. Results of TNF production were expressed as the mean (±SE) of units/ml (1unit = 3.7 pg.); n = 3/group.

Table 1. Effects of 0.9, 0.5, 0.2% Saline Injection on the Nitric Oxide Production of Splenic and Thymi Leukocytes.

	Injection manipulation		
	0.9%	0.5%	0.2%
Spleen[a]			
Con A	3.99±1.74	3.98±1.61	3.45±1.21[b]
LPS	3.29±1.18	3.06±0.82	3.44±1.38
Thymus[a]			
Con A	2.74±0.03	2.81±0.29	2.59±0.09
LPS	2.80±0.21	2.57±0.17	2.80±0.29

[a] Production of nitric oxide by splenocytes and thymocytes cultures incubated with 1.5 or 3 μg/ml Con A or 2.5 or 3 μg/ml LPS. Nitric oxide was measured as nitrite in supernatant by use of the Griess reagent in the supernatant of blastogenesis cultures.
[b] Nitrite is expressed as means ± SE of average triplicate umole/ml; n = 4 per group.

Table 2. Effects of 0.9, 0.5, 0.2% saline Injection on corticosterone-induced apoptosis in thymus and epinephrine induced apoptosis in spleen.

	Injection manipulation		
	0.9%	0.5%	0.2%
Thymus[a]			
% dead[b]	43.95±3.23	45.09±3.78	47.26±4.31[c]
% apoptosis[b]	9.12±0.79	14.23±7.46	14.16±1.32
Spleen[a]			
% dead[b]	32.11±1.73	30.29±10.97	38.74±5.70
% apoptosis	9.92±6.77	7.52±1.67	7.56±4.84

[a] thymocytes from each mice were exposed to corticosterone (0.1 or μM) for 4 hours. Splenocytes from each mice wrer exposed to epinephrine 10 μM for 4 hours.
[b] Percentage of apoptotic cells and dead cells were calculated based on the results of Hochest 33342 and PI, stain.
[c] Percent of apoptotic cells and dead cell are expressed as means ± SE; n = 4/group.

determined by gel electrophoresis (Figure 5, panel A). Results indicated that corticosterone-induced a ladder-like DNA fragmentation in thymus was significantly higher in stress mice than control mice. The percent of apoptoic cells in spleen was mild decreased in stress mice (Table 2). However, the result of DNA fragmentation in epinephrine-induced spleen apoptosis assay had showed mild increased in the stress

mice (Figure 5, panel B). The density of DNA fragmentation in the spleen was very weak. This may be explained that the insensitivity of splenocytes to epinephrine may be affected by osmotic stress. This finding may suggest that the hypo-osmotic stress-induced immune suppression may be caused by the induction of apoptosis response in stress animal's thymus and spleen. However, the signal to induce apoptosis in thymus or spleen may not directly relate with TNF production.

ACKNOWLEDGMENTS

This work was supported by the grant from the SVD section, FU-JEN Catholic University, Taiwan, R. O. C.

REFERENCES

(1) Baker, SJ and Reedy, EP (1996) Transducers of life and death: TNF receptor superfamily and associated proteins. Oncogene. 12: 1-9.

(2) Boyle, M; Gee, A, and Borsos T (1979) studies on the terminal stages of immune hemolysis. VI. Osmotic blockers of differing stokes' radii detect complement-induced transmembrane channels of differing size. J. of Immunology, 123 (1): 77-82.

(3) Chan, WH; Yu, JS; Yang, SD. (1999) Shock-induced apoptosis via a caspase-dependent mechanism: evidence for the involvement of oxidative stress. Journal of Cellular Physiology, 178 (3): 397-408.

(4) Chou, SH; Kojic, L; Nordyke-Messingham, K., and Cunnick, J (1996). Characterization of the effect of 2-deoxy-D-glucose (2-DG) on the immune system. Brain, Behavior, and Immunity. 10: 399-416.

(5) Chou, SH; Kojic, and Cunnick, J (1997). Evidence for the involvement of catecholamines in 2-DG-induced immunomodulatory effects in spleen. Brain, Behavior, and Immunity. 11: 79-93.

(6) Cohen, S., & Williamson, GM. (1991). Stress and infectious disease in humans. Psych Bull 109, 5-24.

(7) Dhabhar, FS; Miller, AH; Stein, M; McEwen, BS and Spencer, RL (1995) Effect of stress on immune cell distribution: Dynamics and hormonal

mechanisms. J. Immunol., 1554: 5511-5527.

(8) Greenleaf JE; Jackson CG; Lawless D. (1995) CD4+/CD8+ T-lymphocyte ratio: effects of rehydration before exercise in dehydration men. Med. Sci. Sports Exerc, 27 (2): 194-9.

(9) Gulbins, E, Welsch, J, Lepple-Wienhuis, A, Heinle, H, and Lang F. (1997) Inhibition of Fas-induced apoptotic cell death by osmotic cell shrinkage. Bioc. Bioph. Research communications 236, 517-521.

(10) Hogan, MM and Vogel, SN (1988) Production of TNF by rINF-γ-primed C3H/HeJ (Lpsd) macrophages requires the presence of lipid-A-associated proteins. J. Immunol, 141: 4196-4202.

(11) Joasod, A,. & McKenzie, J. M., (1976). Stress and the immune response in rats. Int. Arch. Allergy Appl. Immunol. 50, 659-63.

(12) Junger WG; Liu FC; Lomis WH; Hoyt DB. (1994) Hypertonic saline enhances cellular immune function. Circ Shock, 42 (4): 190-6.

(13) Keller, S. E., Weiss, J. M., Schleiffer, S. J., Miller, N. E., & Stein, M. (1981). Suppression of immunity by stress. effect of a graded series of stressors on lymphocyte stimulation in the rat. Science Wash. DC, 213, 1397-1400.

(14) Kiecolt-Glaser, J. K., Garner, W., Speicher, C., Penn, G. M., Holliday, J., & Glaser, R. (1984). Psychosocial modifiers of immunocompetence in medical students. Psychosom. Med. 46, 7-14.

(15) Raff, MC (1992) Nature 356: 397-400.

(16) Schuppan, D., Atkinson, J., Ruehl, M., & Riecken, E. O. (1995). Alcohol and liver fibrosis--pathobiochemistry and treatment. Z-Gastroenterol. Sep 33 (9): 546-50.

(17) Shapiro, L and Dinarello C. (1997) Hyperosmotic stress as a stimulant for proinflammatory cytokine production. Experimental Cell Research 231: 354-362.

(18) Stuehr, D. and Nathan, C. (1989) Nitric oxide: a macrophage product responsible for cytosis and respiratory inhibition in tumor target cells. J. Exp. Med. 169, 1543-1555.

(19) Sun, XM; Snowden, RT; Skilleter, DN et al., (1992) Anal. Biochem. 204: 351.

(20) Wang, AM; Creasey, AA; Ladner MB et al., (1985) Molecular cloning of the complementary DNA for human tumor necrosis factor. Science 228: 149-154.

(21) Wyllie, AH; Kerr, JFR, Currie, AR (1980) Int. Rev. Cytol. 68: 251-306.

(22) Yasmin SE, Southerland LM, Power J, Nicholas, T. and Murray A. (1998) Osmotic stress induces both secretion and apoptosis in rat alveolar type II cells. Am. J. Physiol. 275: L670-678.

received *Octorber 30, 2000*
revised *November 20, 2000*
accepted *December 14, 2000*

低張滲透壓緊迫引發小白鼠免疫功能低下和細胞的凋亡

周秀慧　陳智維　邱健彰　廖慧玲
輔仁大學生物系

摘　要

滲透壓是細胞用來維持內外恆定的機制之一。在壓力狀態下，滲透壓改變也最易影響細胞的活性，進而降低功能的執行。本實驗主要是採用小白鼠實驗動物模式來探討活體的狀況下，低張滲透壓緊迫（hypo-osmotic stress）對免疫功能的影響。將雄性純種BALB/c老鼠分成三組，第一組與第二組小白鼠爲實驗組，各自接受等量0.5%或0.2%低張鹽水的腹腔注射，第三組的小白鼠則接受0.9%等張鹽水注射作爲對照組。實驗結果顯示接受低張鹽水注射之小白鼠的血清存有較高濃度的血色素並且骨髓顆粒性-單核性先驅細胞顯著增植，但其胸腺與脾臟中的T淋巴細胞與B淋巴細胞對分裂原引發的增殖作用呈現抑制作用，而脾臟與胸腺所產生之一氧化氮濃度並無顯著的改變。然而，低張滲透壓緊迫卻造成小白鼠胸腺細胞凋亡與死亡數目顯著增加，因此，此研究結果建議胸腺細胞凋亡可能是活體中低張滲透壓緊迫導致二級免疫器官免疫活性低下的主要誘因。

關鍵詞： 低張滲透壓緊迫，血色素，細胞凋亡，醣皮質激素，腎上腺素，去氧核糖核酸片段化

過氧化氫水溶液緊急排放行爲的研究

鄭 吉 豐
中原大學化學系

摘 要

在蒸氣及非凝結氣體混合（hybrid）系統，溫度壓力須達到緩和狀態（tempering condition），才會有明顯兩相排放，在本研究推導出排放設定壓力（P_S, Pa）與緩和溫度（T_{TP}, K）之關係可表示爲：

$$P_s = (1+0.5\frac{h_{vl}Mwv}{\Delta H})\ exp\left(a-\frac{b}{T_{TP}}\right)$$

排放設定壓力與緩和溫度之間關係，與溶液汽化熱、溶質分解熱及安東尼方程式中常數有關，與反應活化能無關，因此此系統兩相排放溫壓條件，可由此式求得。當升溫速率（在 100 ℃）於低於 1 ℃/min 狀態下，溫度壓力無轉折現象（turnaround），也就是排放過程溫壓持續下降，對 H_2O_2 槽體無任何危害，且蒸汽壓力爲主要壓力來源；隨著升溫速率升高（1℃/min ＞ dT/dt ≥ 30 ℃/min），於排放過程中，溫壓出現轉折現象，非凝結性氣體產生速率快速增加，非凝性氣體占有總排放壓力比率快速增加，但總體排放壓力低於容器設計壓力 3bar 表壓力，對 H_2O_2 槽體尚無立即危害；當升溫速率高於或等於 40 ℃/min，由於熱產生速率高於熱移除速率，因此溫度快速增加，導致蒸汽壓力大幅增加，超過槽體設計壓力，有立即危害。在不同排放口徑對壓力排放行爲之影響，當排放口徑縮小時，排放時最高壓力逐漸增加，有效排放時間增長，主要是因爲當排放口徑縮小時，排放速率減慢，同時移除熱速率較慢，同時導致溶液溫度增加，並經由一系列計算，排放時最高壓力大致和排放面積關係如下：$LogA = 41.88/P_{max}$ + 常數，在不同升溫速率條件下，計算出排放時最高壓力，建議所使用適當尺寸安全閥。

簡 介

近年美國化學工程師學會（AIChE）所屬之緊急排放系統設計協會（Design Institute for Emergency Relief Systems, DIERS）對於兩相排放有相當深入之探討，其研究已逐漸被歐美日工業界所重視。DIERS 研究發現反應性製程之緊急氣液兩相流動之排放，爲達到壓力排放安全所需要之排放面積，較純氣相之排放所需面積可大至數十倍。換句話說，若工廠當初對於反應製程是以 API 之純氣相排放作業爲設計依據時，雖然爲了更安全而將排放實際所需面積增加 1～2 倍，然而在緊急情況下若實際上是兩相排放的話，仍然極有可能無法將壓力有效排放而造成危害。

若以洩放之壓力來源估算排放口徑，失控反應可規劃成三大項類，包括了（1）蒸汽（vapor）系統；（2）氣體（gassy）系統；（3）蒸汽與氣體之混合（hybrid）系統。蒸汽或沸騰系統係指反應過程中反應物或產物會生成蒸汽之現象，洩放時反應熱可藉由蒸氣潛熱釋放方式移除，隱含在這個系統的現象爲排放過程中溫度之反折（turnaround）；反觀氣體系統並無此溫度反折之現象，其原因是由於氣體的產生必須在臨界（critical）溫度以上，而排放過程中無法達到溫度反折點；而蒸汽及氣體之混合系統則發生於這兩種系統之間。目前大部分之論文只對於蒸汽系統有詳細之研究[1-4]，但氣體系統及蒸汽與氣體之混合系統（hybrid）於工業界之應用日趨重要，事實上已有多起嚴重意外事件，因過氧化物的反應失控導致反應器或儲槽爆炸及人員傷亡。

利用 Leung 實驗室所發展出來理論［5-6］及其理論所寫成 VSSPH（Vent Sizing Software Program for Hybrid System）軟體，本研究進一部推導出在無大量不可壓縮氧氣干擾下，安全閥在什麼條件下，會進行明顯兩相排放。並分成三部分探討，（1）不同升溫速率（反應溫度爲 100 ℃時）對兩相排放行爲之影響，（2）不同排放面積對兩相排放行爲之影響，（3）不同反應升溫速率時安全閥截面積與最高壓力之關係，進而對混合（hybrid）系統之緊急排放設計做有效之建議。

兩相排放時設定壓力與溶液溫度之關係推導

蒸汽與非凝結氣體混合物於排放時之主要方程式可由排放容器之巨觀平衡後導得，這些隨時間變動之非穩定方程式包括

$$\frac{dm}{dt} = - W \tag{1}$$

$$\frac{dm_g}{dt} = m\dot{m} - Wx_g \tag{2}$$

$$\frac{dT}{dt} = [mq - Wv_i\frac{h_{vl}}{v_{vl}}]/(mC_p) \tag{3}$$

方程式（1）至（3）分別代表了總質量平衡、非凝結氣體成份之質量平衡及能量平衡，其中 W 是排放速率（relief vent rate），也就是排放質量通率 G（relief mass flux）與其排放面積的乘積，$\dot{m}g$ 是非凝結氣體產生率與總質量之比例（specific gas generation rate, Kg gas/Kg·s），q 是比能量排放速率（specific energy release rate, J/Kg·s），v_{vl}是蒸汽與溶液的比容差，h_{vl}爲溶液汽化熱。

本系統採用均勻容器排放模式（homogeneous -vessel or uniform-froth venting regime），也就是在排放過程容器內部氣液移動速度是一樣，氣液是不分離（no disengagement of vapor/gas from the liquid），排放口的流動，亦假設爲均勻兩相排放，關於排放流動參數評估及演繹方程式請參閱 Leung 實驗室［5-6］及 Bird 實驗室論文［7］。

壓力容器內含過氧化氫水溶液，失控造成快速分解，隨著時間的增加過氧化氫水溶液會分解爲水及非凝結氣體（氧氣），並同時伴隨著熱量之釋放。由於實驗設備中壓力容器假設是完全絕熱，故溶液釋放出的熱量被保留並繼續提供給溶液作爲分解反應所需之能量；加速分解反應是依循阿累尼爾斯定律（Arrhenius law），q 是比能量排放速率（J/Kg·s）可表示如下：

$$q(J/kg \cdot s) = \left[\frac{dT}{dt}\right]_{373.15} C_p exp\left[\frac{Ea}{R}\left(\frac{1}{373.15} - \frac{1}{T}\right)\right] \tag{4}$$

當 H_2O_2 分解時，放出熱量導致溶液快速升溫，$\left[\frac{dT}{dt}\right]_{373.15}$爲在 373K 升溫速率（K/s），$C_p$ 爲 $H_2O_{2(aq)}$比熱（J/kg·K），Ea 爲 $H_2O_{2(aq)}$分解活化能（cal/mole），R 爲氣體常數，每莫耳 H_2O_2 分解，放出 98050J（$\Delta H = -98050$J/mole），生成 16 克氧氣，因此 $\dot{m}_g$ 比非凝結氣體產生率（Kg gas/Kg·s）可表示如下：

$$\dot{m}_g(kg\ O_2/kg \cdot s) = \frac{0.016}{\Delta H} C_p\left[\frac{dT}{dt}\right]_{373.15} exp\left[\frac{Ea}{R}\left(\frac{1}{373.15} - \frac{1}{T}\right)\right] \tag{5}$$

過氧化氫排放實驗的緩和狀態（tempered condition）是指溶液的反應熱產生與

排放時的蒸發熱移除達到平衡的狀態［8］在緩和狀態達到時系統內的升溫速率停止，根據 Leung 實驗室［6］研究結果顯示，當槽中氣體壓力與排放設定壓力相等時，溶液溫度等於或高於緩和溫度，經由洩壓閥所進行的排放爲兩相排放，而此溫度下溶液的分壓是正比於莫耳產生率，非凝結氣體與蒸汽的分壓比可表示爲

$$\frac{P_g}{P_v} = \frac{X_g/M_{wg}}{X_v/M_{wv}} = \left(\frac{\dot{m}_g}{M_{wg}}\right)/\left(\frac{q}{h_{vl}M_{wv}}\right) \tag{6}$$

結合方程式（4）、（5）、（6），可得下列方程式

$$\frac{P_g}{P_v} = 0.5\frac{h_{vl}M_{mv}}{\Delta H} \tag{7}$$

若考量排放口徑，當系統壓力與排放設定壓力相等時，緩和狀態的溫度可求出，同時系統的蒸汽壓力可表示爲

$$P_s = P_g + P_v = (1 + 0.5\frac{h_{vl}M_{mv}}{\Delta H})P_v \tag{8}$$

$$P_s = (1 + 0.5\frac{h_{vl}M_{mv}}{\Delta H})exp\left(a - \frac{b}{T_{TP}}\right) \tag{9}$$

P_s爲洩壓閥設定壓力（Pa），h_{vl}爲溶液汽化熱（2259kJ/kg），M_{wv}爲水蒸汽分子量（0.018Kg/mole），a、b 爲安東尼方程式常數（24.27、4795），因此對過氧化氫水溶液，進行兩相排放時設定壓力與溶液溫度之關係，P_S與 T_{TP}（°K）關係如下：

$$P_s = 1.2074\ exp\left(24.27 - \frac{4795}{T_{TP}}\right) \tag{10}$$

T_{Tp}爲所謂緩和溫度，從方程式（9）可以推論排放設定壓力與緩和溫度之間關係，與溶液汽化熱、溶質分解熱及安東尼方程式中常數有關，與反應活化能及溶液濃度無關，也就是在固定洩放壓力下，會進行兩相排放時的溫度與是否添加催化劑（如金屬離子、酸鹼）無關，爲 P_S（Pa）與 T_{TP}（°K）關係如圖 1，隨著洩壓閥設定壓力升高，進行兩相排放時的緩和溫度逐漸增加。

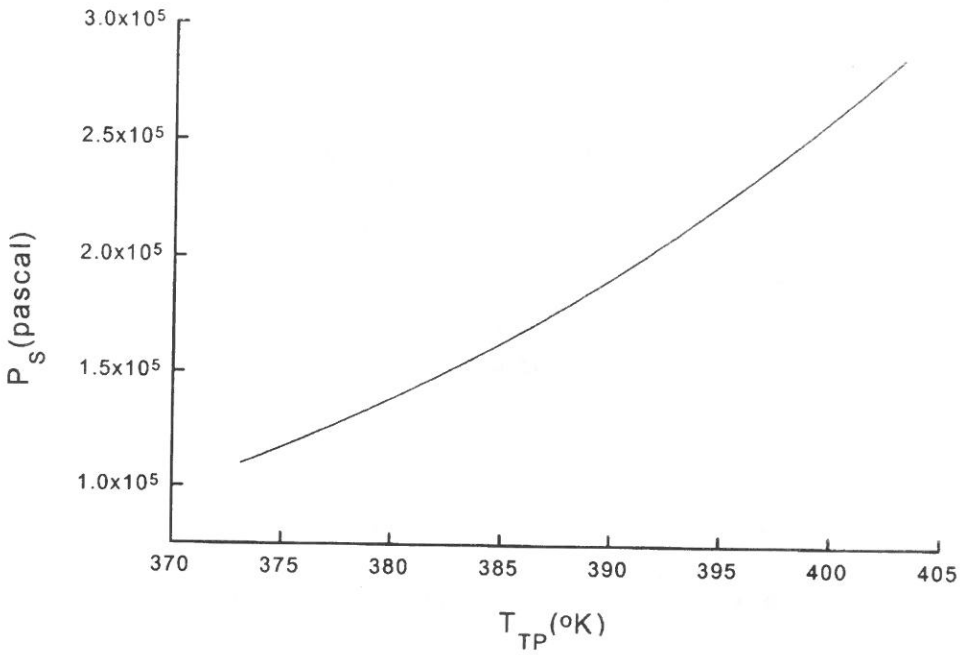

圖 1 洩壓閥設定壓力（PS）與緩和溫度（TTP）關係

結果與討論

I. 不同升溫速率（反應溫度爲 100℃時）對兩相排放行爲之影響

一般在過氧化氫水溶液儲槽在設有洩壓安全閥，但適當排放尺寸是必須的，以避免排放時最高壓力（P_{max}）超過槽體設計壓力，工業上廠商所使用 $H_2O_{2(aq)}$ 濃度及所含不純物都不儘相同（酸鹼及金屬離子都可催化 $H_2O_{2(aq)}$ 分解），因此 $H_2O_{2(aq)}$ 失控行爲及升溫速率都不相同，導致排放尺寸設計不盡相同，在本節將探討過氧化氫水溶液不同升溫速率時安全閥排放面積之設計及不同升溫速率（反應溫度爲 100 ℃）、不同排放面積對排放行爲之影響，所處理狀況如下：

設計壓力容器具有 0.546 m^3 的容量，容器設計壓力爲 3 bar·g（58.2 psia），爲了避免壓力容器過壓所設計的壓力洩放裝置的設定壓力爲 0.5 bar·g（T_{TP} = 109.5

℃），而爲了設計需求，容器的比熱含量被設定爲零以確定完全絕熱的狀態能夠達成，其餘參數詳列於表一：

表 1 過氧化氫水溶液反應相關參數

反應式	$H_2O_{2(l)} \rightarrow H_2O_{(l)} + \frac{1}{2} O_2$
容器體積	$V_o = 0.546\ m^3$
溶液質量	$m_o = 454\ Kg$
起始濃度	15 wt%（9.03 mol %）H_2O_2
液位高	80 %
蒸汽分子量	$M_{wv} = 18$（大部分爲水蒸氣）
氣體分子量	$M_{wg} = 32$（氧氣）
溶液密度	$\rho = 0.001039\ m^3/Kg$（$1.039\ g/cm^3$）
液體比熱	$C_p = 4\ KJ/Kg \cdot K$
水蒸氣壓	$P_v = \exp(24.27 - 4795/T)\ N/m^2$
水蒸氣比容	$v_v = RT/P_v/M_{wv}$
氣體比容	$v_g = RT/P_g/M_{wg}$
蒸發熱	$h_{vl} = bR/M_{wv} = 2259\ KJ/Kg$

在本研究所使用排放口徑爲 R-type（16 in^2），反應溫度爲 100 ℃，圖 2 爲過氧化氫水溶液在不同升溫速率時排放壓力（P_{tot}，psia）對時間（s）之變化，當反應升溫速率（dT/dt，℃/min）≤ 1 ℃/min 時，溫度壓力無轉折現象（turnaround），也就是排放過程溫壓持續下降，對 $H_2O_{2(aq)}$ 槽體無任何危害；隨著升溫速率升高（1 ℃/min ＞ dT/dt ≥ 30 ℃/min），於排放過程中，溫壓出現轉折現象，但排放時最高壓力（P_{max}）低於容器設計壓力 3bar 表壓力，對 H_2O_2 槽體尚無立即危害；當升溫速率高於或等於 40 ℃/min，P_{max} 快速增加，超過槽體設計壓力，有立即危害，當 dT/dt = 45 ℃/min 時（圖 3），壓力增加幅度，已到失控地步，其中 P_v 爲蒸汽分壓，P_g 爲氧氣分壓。

排放時總壓力爲非凝結性氧氣壓力和蒸汽壓力之和（$P_{tot} = P_g + P_v$），詳細觀察過氧化氫水溶液反應升溫速率對排放時非凝結性氧氣分壓（圖 4）影響，當 dT/dt ≤ 1 ℃/min 時，氧氣分壓占總壓比率低於 20%，隨著升溫速率增加，氧氣分壓逐漸增加，壓力增加程度和升溫速率幾乎呈正比，因爲升溫速率越高產生氧氣速率越快、非凝結性氧氣壓力越高；圖 5 爲升溫速率對排放時蒸汽分壓影響，可觀察到當 dT/dt

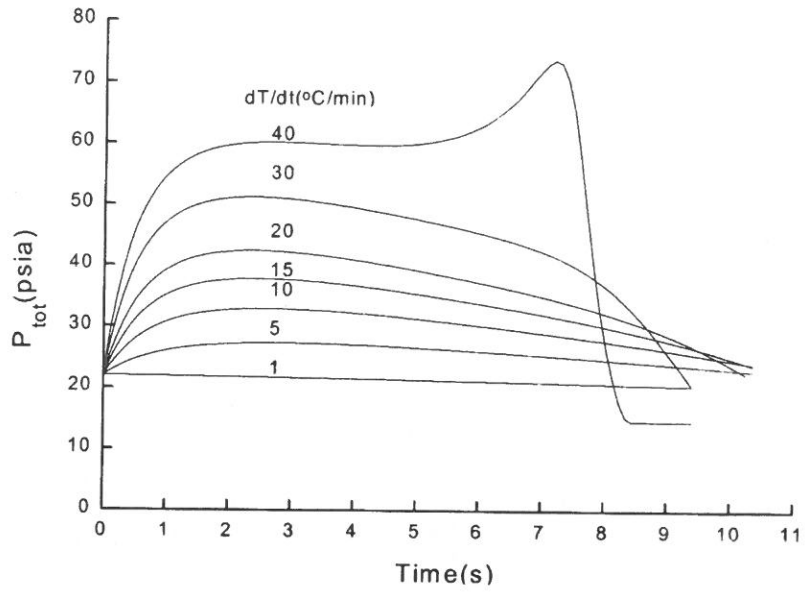

圖 2　不同反應升溫速率時排放壓力對時間之變化

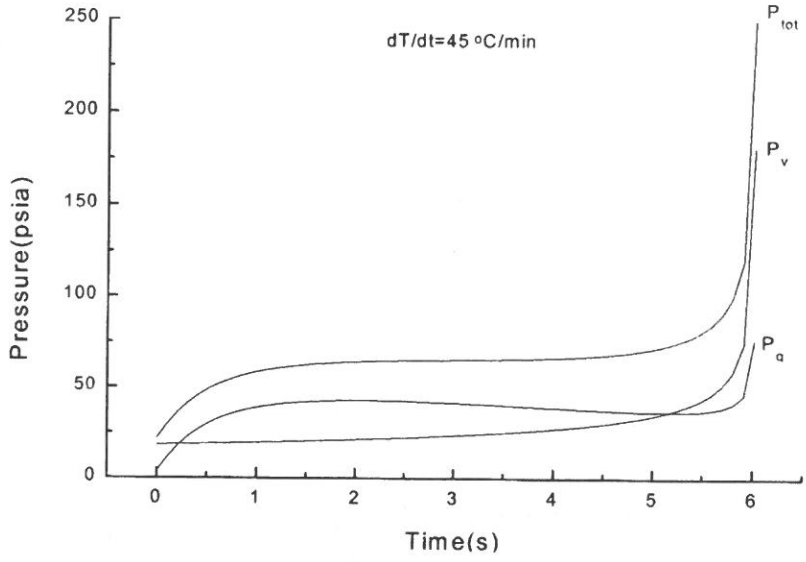

圖 3　反應升溫速率爲 45℃/min 時排放壓力對時間之變化

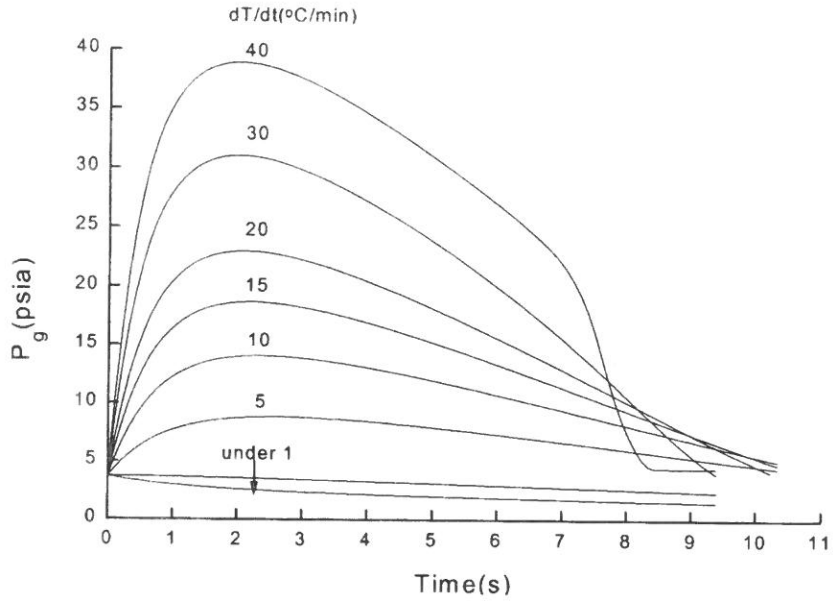

圖 4　反應升溫速率對排放時非凝結性氧氣分壓關係

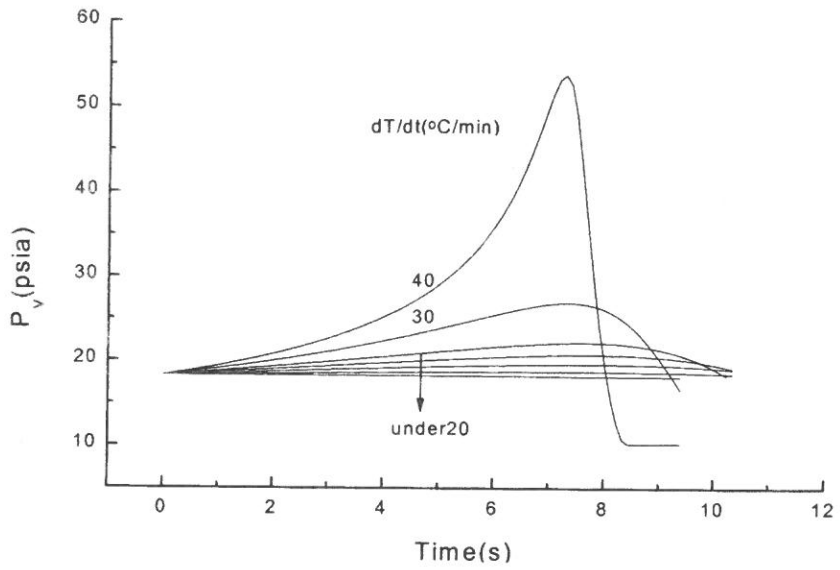

圖 5　反應升溫速率對排放時蒸汽分壓影響

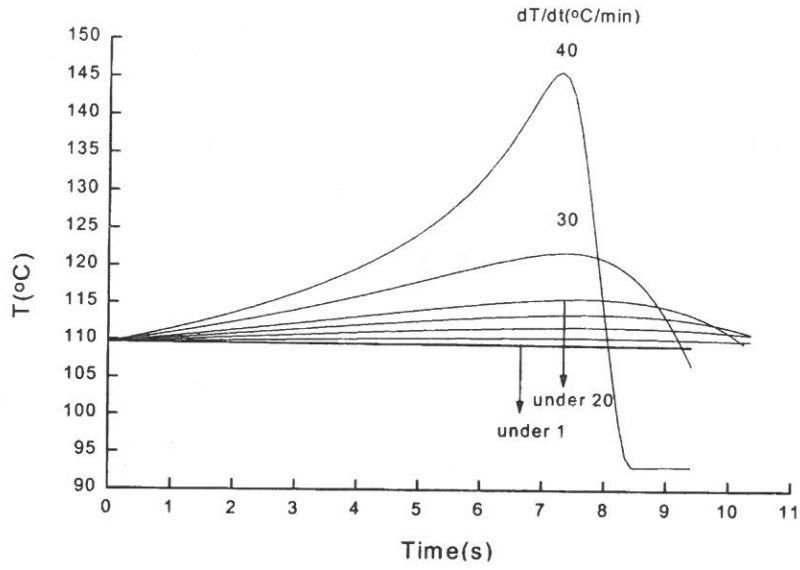

圖 6　反應升溫速率對排放時溶液溫度之影響

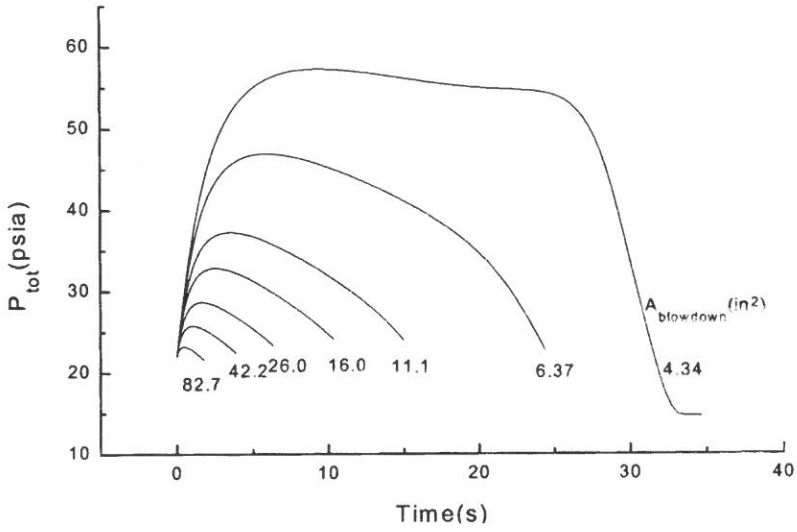

圖 7　不同排放口徑對壓力排放行爲之影響

≤ 20 ℃/min 時，蒸汽分壓增加非常緩慢，但升溫速率高於 30 ℃/min 時，蒸汽分壓增加非常快速，因爲當 dT/dt < 30℃/min，反應所放出熱量足以被排放蒸汽所帶走（口徑爲 R-type（16 in^2）），然而升溫速率繼續增加時，熱量逐漸累積，使得過氧化氫水溶液溫度快速增加（圖八），因蒸汽壓力和溫度是呈指數關係，所以蒸汽分壓快速增加，最後即使在洩放閥全開，亦到達失控地步（圖 3）。

II. 不同排放面積對兩相排放行爲之影響

本節所採用反應升溫速率爲 10 ℃/min（在溫度爲 100 ℃），圖 7 顯示在不同排放口徑對壓力排放行爲之影響，當排放口徑縮小時，排放時最高壓力（P_{max}）逐漸增加，有效排放時間增長，主要是因爲當排放口徑縮小時，排放速率（lb/s）減慢（圖 8），同時移除熱速率較慢，導致溶液溫度增加（圖 9），當口徑縮小爲 3.6 in^2 時，熱移除速率小於熱生成速率，最後即使在洩放閥全開，亦到達失控地步（圖 10）。

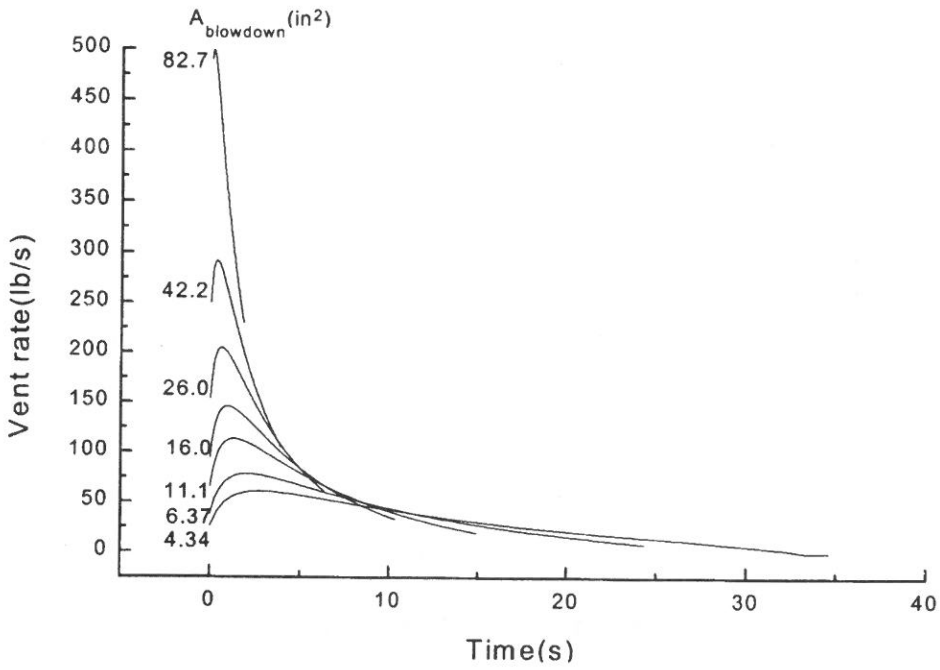

圖 8 不同排放口徑對排放速率之影響

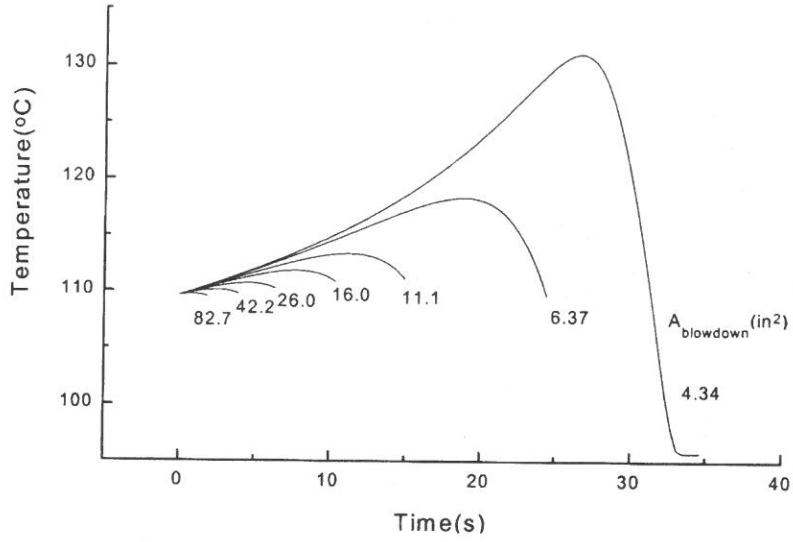

圖 9　不同排放口徑對排放時溶液溫度之影響

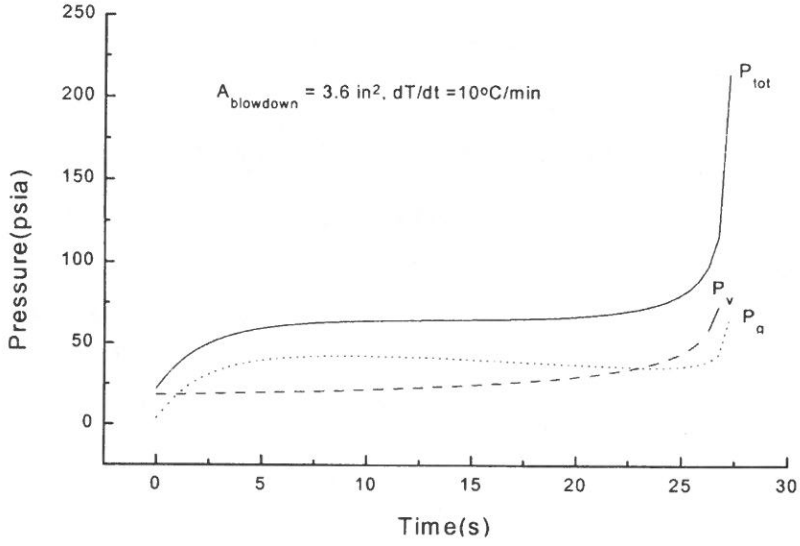

圖 10　排放口徑爲 3.6in² 時排放壓力對時間之變化

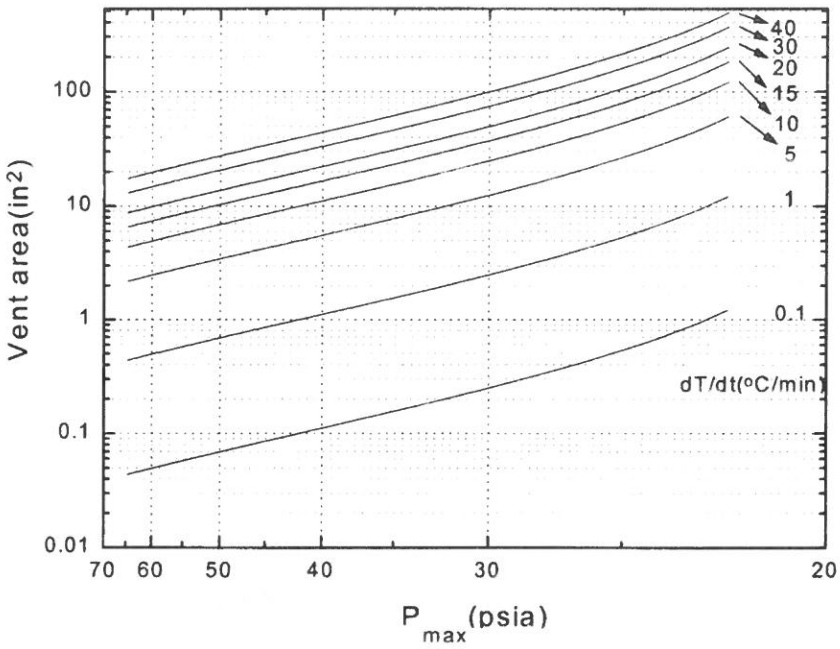

圖 11　不同反應升溫速率時安全閥截面積與最高壓力之關係

III.　不同反應升溫速率時安全閥截面積與最高壓力之關係

圖 11 爲過氧化氫水溶液在不同反應升溫速率時，安全閥截面積（A，in^2）與排放時最高壓力（P_{max}，psia）之關係（解析解結果［6］），其中截面積取對數，壓力取倒數，可觀察到當最高壓力高於 27psia，兩者線性關係良好，關係如下：$LogA = 41.88/P_{max}$ + 常數，常數與升溫數率有關，即在不同升溫速率時，圖形呈平移的，斜率具有一致性，此升溫速率和濃度及催化劑有關係，必須由實驗得知。

表 2 在不同反應升溫速率時截面積與最高壓力之關係矩陣表

安全閥規格 (截面積, in^2)	P_{max} psia	P_{max} (0. 39)	P_{max} (1)	P_{max} (5)	P_{max} (10)	P_{max} (15)	P_{max} (20)	P_{max} (30)	P_{max} (40)
J (1.287)	P_{set}	27.6	37.7	101.9	380.8	失控	失控	失控	失控
K (1.838)	P_{set}	25.0	33.1	74.0	158.2	失控	失控	失控	失控
L (2.853)	P_{set}	22.5	28.8	55.3	91.9	149.8	270.7	失控	失控
M (3.600)	P_{set}	P_{set}	26.9	48.8	75.2	110.0	163.8	失控	失控
N (4.340)	P_{set}	P_{set}	25.6	44.6	65.7	90.7	124.3	261.0	失控
P (6.379)	P_{set}	P_{set}	23.2	37.9	52.0	66.6	83.1	127.8	206.3
Q (11.050)	P_{set}	P_{set}	P_{set}	31.1	40.1	48.3	56.4	73.9	94.8
R (16.00)	P_{set}	P_{set}	P_{set}	27.8	34.8	40.7	46.4	57.6	69.5
T (26.00)	P_{set}	P_{set}	P_{set}	24.4	29.6	33.8	37.6	44.6	51.5
T_2 (27.87)	P_{set}	P_{set}	P_{set}	24.0	29.0	33.0	36.6	43.3	49.7
V (42.190)	P_{set}	P_{set}	P_{set}	P_{set}	25.8	28.9	31.6	36.5	40.9
W (60.75)	P_{set}	P_{set}	P_{set}	P_{set}	23.5	26.0	28.2	32.1	35.4
Y (82.68)	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	24.0	25.9	29.1	31.8
Z (90.95)	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	23.5	25.3	28.3	30.9
Z_2 (108.86)	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	22.5	24.1	26.8	29.2
AA (136.69)	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	22.8	25.2	27.3
BB (168.74)	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	23.9	25.8
BB_2 (185.0)	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	P_{set}	23.4	25.1

表二爲圖 11 計算結果應用於目前標準安全閥尺寸，所得到關係矩陣表，表中括弧爲升溫速率（℃/min），必須經由實驗得知，洩壓閥設定壓力 P_{set}爲 0.5bar 表壓力 (21.8psia)，因此表格中出現 P_{set}，即表示排放過程中壓力無轉折現象，是非常安全，當 Pmax 低於槽體設計壓力 (58.2psia)，表示在此升溫速率反應條件下是安全，也就是表下半區是安全，上半區是超過槽體設計壓力，灰色格代表可允許最小安全閥尺寸，從表中可看出，當 $dT/dt \leq 1℃/min$，所列的安全閥都是安全的。

結　論

根據 Leung 實驗室的理論，在混合（hybrid）系統，溫度壓力須達到緩和狀態（tempered condition），才會有明顯兩相排放，在此狀態下，排放時溶液的蒸發熱移除與反應熱釋放達到平衡的狀態，也就是在固定設定壓力下（P_S, pascal），溫度在高於緩和溫度（T_{TP}, K）下排放，有明顯兩相排放行爲，在本研究推導出排放設定壓力與緩和溫度之關係可表示爲

$$P_s = (1 + 0.5\frac{h_{vl}M_{wv}}{\Delta H})exp\left(a - \frac{b}{T_{TP}}\right)$$

排放設定壓力與緩和溫度之間關係，與溶液汽化熱、溶質分解熱及安東尼方程式中 a、b 常數有關，與反應活化能無關，因此，此系統兩相排放溫壓條件，可由此式求得。

根據 Leung 實驗室的理論及此理論所轉換 VSSPH 軟體，1000 磅 15 wt% H_2O_2 在容器體積爲 19.27 ft^3 約占 80 %液位，排放口徑爲 R-type（16 in^2），此容器設計壓力爲 3bar 表壓力，排放設定壓力爲 0.5 bar 表壓力，在此條件下計算及研究放熱升溫速率及液位對兩相排放行爲之影響。當升溫速率（在 100 ℃）於低於 1 ℃/min 狀態下，溫度壓力無轉折現象（turnaround），也就是排放過程溫壓持續下降，對 H_2O_2 槽體無任何危害，且蒸汽壓力爲主要壓力來源；隨著升溫速率升高（$1 > dT/dt \geq 30$ ℃/min），於排放過程中，溫壓出現轉折現象，非凝結性氣體產生速率快速增加，非凝性氣體占有總排放壓力比率快速增加，但總體排放壓力低於容器設計壓力 3bar 表壓力，對 H_2O_2 槽體尚無立即危害；當升溫速率高於或等於 40 ℃/min，由於熱產生速率高於熱移除速率，因此溫度快速增加，導致蒸汽壓力大幅增加，超過槽體設計壓力，有立即危害。不同反應升溫速率時安全閥截面積與最高壓力之關係，在此研究中計算，排放時最高壓力（P_{max}, psia）大致和排放面積（A, in^2）關係如下：$LogA = 41.88/P_{max}$ + 常數，常數與升溫數率有關，可用此關係式選擇適合尺寸之安全閥。

符號說明表：

A	排放面積（venting area）
a, b	安東尼常數（Antoine constants）

C_p	溶液比熱
Ea	H_2O_2 分解活化能
h_{vl}	溶液汽化熱
G	排放質量通率（relief mass flux）
ΔH	H_2O_2 分解熱
M_{wg}	非凝結氣體分子量
M_{wv}	蒸汽分子量
m_g	非凝結氣體質量
$\dot{m}_g$	非凝結氣體產生率與總質量之比例（specific gas generation rate）
P_g	非凝結氣體分壓
P_v	蒸氣氣體分壓
P_s	排放時設定壓力
P_{max}	最高排放壓力
q	比能量排放速率（specific energy release rate）
R	氣體常數
T_{TP}	緩和溫度
t	時間
T	溫度
v_{vl}	蒸汽與溶液的比容差
W	排放速率（relief vent rate）
x_g	非凝結氣體質量分率
x_v	蒸氣氣體質量分率

參考文獻

(1) Duxbury, H. A., "Relief System Sizing for Polymerization Reactors" The Chem. Engr., 31 (1980).

(2) Fauske, H. K., "Generalized Vent Sizing Monogram for Runaway Chemical Reactions," Plant/Operations Prog., 3, 213 (1984).

(3) Leung, J. C., "Simplified Vent Sizing Equations for Emergency Relief Requirements in Reactors and Storage Vessels," AIChE J., 32, 1622 (1986).
(4) Leung, J. C., "Overpressure During Emergency Relief Venting in Bubbly and Churn-Turbulent Flow," AIChE J., 33, 952 (1987).
(5) Leung, J. C., and M. Epstein, "Flashing Two-Phase Flow Including the Effects of Noncondensable Gases," ASME Trans. J. of Heat Transf., 113, 269 (1991).
(6) Leung, J. C., "Venting of Runaway Reactions with Gas Generation" AIChE J., 38, 723 (1992).
(7) Bird, R. B., W. E. Stewart, and E. N. Lightfoot, Transport Phenomena, Example 15. 5-3, p. 480, Wiley, New York (1960.)
(8) Huff, J. E., "Emergency Venting Requirements," Plant/Operations Prog., 1, 211 (1982).

89 年 10 月 27 日 收稿
89 年 12 月 2 日 修正
89 年 12 月 21 日 接受

Studies of Emergency Relief of Hydrogen Peroxide Aqueous Solution

C. -F. Cheng

Department of Chemistry

Chung Yuan Christian University

Chung Li, Taiwan 320, R.O.C.

Abstract

In hybrid system of gas and vapor, temperature and pressure must reach so-called tempering condition and significant two-phase discharge would occur. In this study, the following relation between the vessel set pressure (P_S) and tempering temperature (T_{TP}) was derived and related to vaporization heat of solution, decomposed heat of solute and Antoine constant a, b but not activation energy. Thus, the temperature and pressure of two-phase venting can be obtained from following equation.

$$P_S = (1+0.5\frac{h_{vl}M_{wv}}{\Delta H})\ exp\left[a-\frac{b}{T_{TP}}\right]$$

When temperature rise rate at 100 ℃ was less than 1 ℃/min, no temperature and pressure turnaround occur, thus temperature and pressure of process descend without hazard to vessel of hydrogen peroxide solution and vapor was major pressure source. However, temperature rise rate was between 1 and 30 ℃/min, temperature and pressure turnaround occur. Release rate of uncondensed gas increased quickly and uncondensed gas was major pressure source but total venting pressure was less than 3 bar·g without immediate hazard to vessel. When temperature rise rate was more than 40 ℃/min, the vapor pressure of generation would increase very quickly and be more than the design pressure of vessel with immediate hazard to vessel because the heat release rate was more than the evaporative heat

removal rate. Venting behaviors at different venting areas were studied. When the venting area became smaller, maximum pressure was increased and venting time was elongated because the venting rate decreased and evaporative heat removal rate decreased. By a series of evaluation, the relation between maximum pressure and venting area was Log A = $41.88/P_{max}$ + constant. Maximum pressures could be calculated at different temperature rise rates and suggest suitable sizes of safety relief valves.

Key Words: 緊急排放 (emergency relief), 兩相排放 (two-phase discharge), 混合系統 (hybrid system), 緩和狀態 (tempering condition), 過氧化氫 (hydrogen peroxide)