Test Resource Partitioning Based on Efficient Response Compaction for Test Time and Tester Channels Reduction

Yinhe Han† **Yongjun Xu**† **Huawei Li**† **Xiaowei Li**† **Anshuman Chandra**§
† *Institute of Computing Technology, CAS, Beijing, 100080* §*Synopsys, Inc.*
† *Graduate School of the CAS, Beijing, 100039* *Mountain View, CA 94043*
Email: {yinhes,lxw}@ict.ac.cn *Email: Anshuman.Chandra@synopsys.com*

Abstract

This paper presents a test resource partitioning technique based on an efficient single-output response compaction design called quotient compactor (q-Compactor). Some design theorems of quotient compactor are presented to achieve full diagnostics ability, minimize error cancellation and handle the X bits in the outputs of the CUT. The quotient compactor can also be moved to the load-board to reduce the number of ATE channels required. Our experimental results on the ISCAS89 benchmark circuits and an MPEG 2 decoder SOC show that the proposed compaction scheme is very efficient.

1. Introduction

Core-Based system-on-a-chip design and 0.13 micro process technology have raised several test challenges. The rapidly increasing test cost is a major component. There are some major factors that contribute to test cost: available test memory, available number of tester channels, test time, test frequency, available number of pins for scan-in and scan-out purpose [1]. Test Resource Partitioning (TRP) offers a promising solution to these problems by moving some test resources from ATE to chip or load-board. Three main on-chip TRP techniques descried in the literature are: Test vector compaction and compression technique [2][3][19]; Test Pins Reduction Technique; and Test Frequency Improvement Technique [4].

On-board TRP is a comparatively new research topic [4]. The new test resource placed on load-board can be reconfigured and require no modifications to the chip design. Hence, no additional area overhead is an inherent advantage of the on-board TRP. It provides a solution to tradeoff between the cost of chip manufacture and test.

In this paper, we will present an on-chip and on-board TRP based on response compaction that offers a number of important advantages. On-chip compactor leads to reduced testing time and test data volume. On-board TRP can compact the output of chips to a single output, as can help us to save the tester channels significantly.

The rest of the paper is organized as follows: Section 2 presents an overview of previous work on response compaction. In Section 3, we describe our q-Compactor and some design theorems are presented. In the Section 4, we present experimental results on the effectiveness of q-Compactor for aliasing and test time reduction when used as a part of DFT. We verified our response compactor in a FPGA prototype environment and some experimental results are also presented in the Section 5, followed by conclusion in Section 6.

2. Previous Work on Response Compaction

Response compaction is a very well researched topic and a number of techniques have been presented in the literature [1][6][7][8][9][11][12]. The first compaction scheme using error codes and XOR network was proposed in [7]. X-compact technique is presented in [1]. It is based on a XOR-network and guarantees error detection in the presence of unknown logic values and diagnosis capabilities.

Recently, new sequential compactors based on MISR have been presented. The OPMISR technique is presented in [6] which can double the number of scan chains using a on-product MISR. [9] presents a design technique of space compactor and MISR which can avoid the two errors cancellation. However, the above two techniques don't

* This paper is supported by the National Natural Science Foundation of China (NSFC) under grant No.90207002

Proceedings of the 12th Asian Test Symposium (ATS'03)
1081-7735/03 $17.00 © 2003 **IEEE**

have an effective way to eliminate the X-bit in the scan chain whose outputs are not known during simulation.

J. Rajski presents a convolutional compactor for test response in [8]. This compactor can provide high compaction ratio and low aliasing probability. It also has a ability to handle unknown states. But [8] does not give a direct full diagnostic capability of compactor. Some general design theorems of response spread network are also absent in [8].

3. Proposed Test Response Compaction Technology

3.1. The on-chip architecture of q-Compactor

Our proposed compactor is called as q-Compactor which is inserted between the scan chains outputs and the pads of chip. It consists of two parts: Quotient Monitored Shift Registers (QMSR) and Response Spread Network. The QMSR compacts the multiple output response into a single bit output, which is monitored by the ATE. The response spread network is used to spread the outputs of the scan chain into different registers of QMSR and helps to partially avoid the error cancellation.

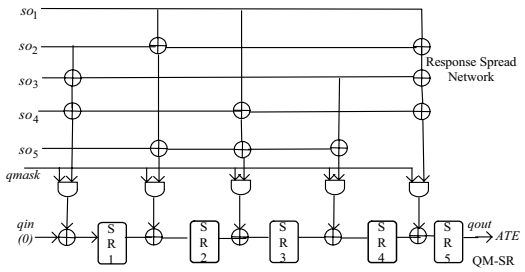

Figure 1. An example of q-Compactor

An example of q-Compactor is shown in Figure 1. The compactor has five inputs which are spread into five registers. Each scan output is propagated to three registers. For example, so_1 is mapped to 2nd, 3rd, and 5th register. If so_1 is erroneous, the syndrome can be spread to these registers.

The QMSR contains two inputs: *qin* and *qmask* and a output *qout*. The signal *qin* is used to seed the QMSR or provide the test vectors to test the QMSR. The signal *qmask* is used to control modes of the QMSR for compaction, diagnosis and freeze the scan chain during diagnostics. It selects between *compaction mode* and *scan chain mode* of QMSR. For the control signal *qmask= 1*, the QMSR is configured as a compactor that accumulates and compresses the scan chains contents. When the *qmask=0*, the QMSR will become a serial scan chain, which can unload the value in the registers and load the seed of the QMSR. The *scan chain mode* of QMSR supports a diagnostic data dump feature. This diagnostic operation can be re-run on the tester to collect the successive failing data. It may be written as a test protocol in a STIL's format which can be automatically executed in the tester [5]. After the interval data is collected, a linear equation is set up to determine the values of scan chains. Then a diagnostic tool can be used to determine faults that best explain the failing scan cells.

3.2. Design techniques of Response Spread Network

If our compactor is implemented using exclusive-or gates, it can be represented by a binary *n*k* matrix Π (matrix with only 0s and 1s) which is called as 'spread matrix'. In this matrix, each row of the matrix corresponds to an input of spread network and each column corresponds an output of network. The row i and column j of the spread matrix is 1 if and only if the *i*th spread network input is related to the *j*th output of the spread network; otherwise, the entry is 0. The spread matrix of Figure 1 is shown as:

$$\Pi = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

3.2.1. Design Theorems for error cancellation and handling X-state

For a conventional sequential compactor, such as Multiple Input Signature Register (MISR), there are two sources of aliasing: error masking and error cancellation. Our QMSR is a sequential compactor. But unlike the MISR, there is no first class aliasing source in our QMSR since no feedback path exists. A ingenious design of spread network can help us reduce the second class aliasing. For analyzing the error cancellation, we define an "equivalent relation" as:

Two k-dimensional vectors R1,R2, "R1, R2 are equivalent ", if and only if:

1) the number of 1s in R1 and R2 is equal.

2) the number of 1s in R1 and R2 is equal with one

or $\forall i, j \quad |P_{R1}(i) - P_{R2}(i)| = |P_{R1}(j) - P_{R2}(j)|$

(i,j<=k) where $P_{R1}(i)$ *and* $P_{R1}(j)$ *represent position of the i-th and j-th 1s in the vector R1.* $P_{R2}(i)$ *and* $P_{R2}(j)$ *represent position of the i-th and j-th 1s in the vector R2.*

It is easy to prove that the next two theorems are true:

[Theorem 1. for 2 errors cancellation] *One or two errors from any scan chain at the same clock cycle or different clock cycles within a k-cycle window are guaranteed to be*

detected if and only if no two rows in $\prod$ *are equivalent .*

[Theorem 2. for error cancellation] *One□two or any odd number of errors from any scan chain at the same clock cycle or different clock cycles within a k-cycle window are guaranteed to be detected if no two rows in* $\prod$ *are equivalent and every row contains an odd number of 1s.*

Given the number of registers in QMSR *k,* spread coefficient *r* (the same odd 1s every row contains), the bound on the maximum number *n* of inputs of spread network satisfying Theorem 2 is as follows:

$$n = \sum_{i=2}^{k-1} \binom{i-1}{r-2}$$

It can be concluded from partition laws of positive integer in combinational algebra [17].

When the number of scan chains *(n)* is determined, a spread matrix satisfying Theorem 2 can be designed in various selections of *k*. Design of lager *k* means the lower error cancellations but the larger area overhead. The tradeoff between the error cancellation probability and area overhead is evaluated by some extensive experiments. Table 1 lists the results of "cancellation probability/area overhead (equivalent NAND gate)". 4 errors were randomly inserted in a k-depth pattern to evaluate the cancellation probability. Every measurement was done by conducting 10^8 simulations.

Table 1. Analysis of cancellations probability/ area overhead

100 inputs 4 errors cancellation			
K	**r=3**	**r=5**	**r=7**
30	3722/1250	542/2050	98/2850
40	360/1400	43/2200	33/3000
50	55/1550	24/2350	11/3150
60	32/1700	12/2500	0/3300
70	7/1850	4/2650	0/3450
80	3/2000	8/2800	0/3600
90	0/2150	2/2950	1/3750
100	0/2300	1/3100	0/3900

The best balances between the error cancellation probability and area overhead are shadowed in the Table 1 for all cases. We can see the selection of lager *k* and small *r* is a better choice since it can handle the best balance of cancellation probability and area overhead. Another advantage of larger *k* is to allow us to optimize the XOR network.

The X-bit in the response is a problem for the conventional sequential compactor such as MISR [14]. However, since our compactor doesn't have a feedback, the X bits mapped to registers will be shifted out in at most *k* cycles. If a spread network is designed satisfying Theorem 2, it will provide a strong capacity to handle the X bit:

[Property for handling X-state] *If both conditions of Theorem 2 are satisfied, a single error from one scan chain with a single X-bit produced by another scan chain is guaranteed to be detected.*

This capacity of handling X-state is verified in our FPGA prototype environment. A case is analyzed as follows: The number of X bits is 0.015% (3527) in a 0.5M-depth and 47-input test vectors. The errors produced from the outputs of chip are 736. If in absence of X bit, they will be observed 2071 times through the quotient bits. When these X bits are produced, these errors are observed 1926 times and only 3 errors are masked entirely.

3.2.2. Design Theorems for diagnosis

To make failure diagnosis of the failing gate possible, the exact data of every erroneous bit unloaded from the scan chains must be known. Our q-Compactor can provide two ways to get this information: direct diagnosis and full diagnosis. The direct diagnosis uses the quotient bits stored in the ATE fail buffer to determine the single one fault from the scan chains. The full diagnosis adds some unloading intervals when the QMSR is configured as *scan chain mode* to collect the failing data. The full diagnosis can be run on-line through a test protocol.

The direct diagnosis of single error is guarded by next theorem:

[**Theorem 3 for 1 error diagnosis**] *One error from any scan chain can be identified if and only if no two rows in* $\prod$ *are equivalent.*

Sine no two rows in $\prod$ are equivalent, we can match the row vector of matrix with the erroneous bits in the quotient sequence to identify the syndrome.

Example 1: If there are errors in 1st, 3rd, 5th bits of a five-bit fraction in failing data for the q-Compactor in Figure 1, we can search in $\prod$ and find 4th row matched. So we can confirm that the 4th scan chain has produced error.

However, this matching operation isn't always true if we can't confirm that the only one error occurs. In actual case, confirmation of single error is same difficult as knowing the full information of faults. So the full information must be first provided in the practical diagnosis case.

The *serial scan chain mode* of QMSR has provided a possibility of getting the full information of scan chains. The following theorem will change the possibility to reality:

[Theorem 4 for full diagnosis] *The q-Compactor can have the capacity to identify all errors, which can be produced by any scan chain at any time, if the design of spread matrix guards:*

$$Rank\ (\Pi^T) = n$$

(Rank (Π^T) *is the rank of the matrix* Π^T*)*

If the input vector from scan chains to spread network at *t cycle* can be expressed as: $X_t = [x_1(t), x_2(t), x_3(t), \cdots\cdots, x_n(t)]^T$; state of registers in QMSR : $S_t = [s_1(t), s_2(t), s_3(t), \cdots\cdots, s_m(t)]^T$. $\prod$ is the spread matrix, we can get:

$$\Pi^T * X_t \oplus S_{t-1} = S_t$$

If we seed the QMSR to 0, the S_{t-1} will be 0 and the equation can be simplified as:

$$\Pi^T * X_t = S_t$$

The capacity of getting the full information about scan chains is equivalent with guarding that there is a unique solution in above system equation. Then the Theorem 4 can be concluded from Cramer Law [16].

Example 2: For the compactor of Figure 1, if we found the errors occur *at t cycle,* unloading intervals are inserted and get the symptomatic states: $S_t^S = [1\ 1\ 0\ 1\ 0]$, it is different with the expected state: $S_t^E = [1\ 0\ 0\ 0\ 0]$. We can calculate the contents of output of scan chains in *t* cycle:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

A Gauss-Jordan Elimination can be used to solve this equation and get symptomatic input vector: $X_t^S = [\ 1\ 1\ 0\ 0\ 1\ ]$. Compared to the expected input vector: $X_t^E = [\ 1\ 0\ 1\ 0\ 1\ ]$,we can determine that the 2nd and 3rd scan chains have produced the errors.

Given the number of input of spread network *n*, the minimal number of outputs of spread network satisfying the Theorem 4 must be guaranteed:

k = n

4. Experiment Results

In this section, we experimentally evaluate the proposed test response compaction techniques for large circuits of the ISCAS'89 benchmarks. These circuits are full-scan and test vectors are generated by the Mintest[20]. For each ISCAS'89 benchmark circuit, we have generated a q-Compactor. Table 2 presents the design properties of ISCAS'89 circuit and corresponding q-Compactor. Column 2,3,4 provide the Vectors Count(*vec*) of Mintest, the number of D Flip-Flop and PIs(*DFFs*), assumed maximum available Pins Count *pins* (in practical case ,this limitation is from the physical design or the tester). Column 5,6,7,8 list INput Count (*INC*), spread coefficient *r*, the number of stage in QMSR *k* and *area* overhead (equivalent NAND gate), *capacity* of q-Compactor. We design the spread network satisfying the Theorem 2 and 4. We can see, to a 62 inputs q-Compactor, the area is 1426 equivalent NAND gates. However, the size of our compactor doesn't depend on the test set and the size of CUT, if the CUT is not a small benchmark circuit but a 2M gates current SOC design, the area overhead of q-Compactor is only 0.07%.This ratio is low compared to the average 15% for the single-output compactor in [11].

Table 3. Aliasing probability of actual stuck-at faults

Circuit	Faults	Injected faults	Num. of random simulations	Aliasing ?
S13207	9815	1	9815	No.
		2	10^5	No.
		10	10^5	No.
S15850	11727	1	11727	No.
		2	10^5	No.
		10	10^5	No.
S35932	39094	1	39094	No.
		2	10^5	No.
		10	10^5	No.
S38417	31180	1	31180	No.
		2	10^5	No.
		10	10^5	No.
S38584	36305	1	36305	No.
		2	10^5	No.
		10	10^5	No.

In the following series of experiments, we will evaluate the effectiveness of error detection. The flow of experiment is: We first generate faults (include the redundant faults) list for every benchmark circuit. Then the special number of faults is randomly selected from this list and injected to the circuit. A fault simulator HOPE [18] and a software simulator of q-Compactor are called to obtain the erroneous quotient result. If the erroneous quotient result is equal with the expected correct quotient result, we will state the count as aliasing count. Table 3 shows the result of experiments. We inject 1 or 2 faults per simulation to simulate the mild defect, and 10 faults per simulation to serious defect. From the Table 3, we can see, it is zero-aliasing to all these cases listed in the table. The possible reasons we consider as follows: In the single one faults scene, the fault response are centralized in a single scan chain since the circuit is partitioned by the scan chain. So it is always detected as is guaranteed by theorems. The

two faults scene is similar. In the many faults scene, registers of QMSR mapped by these errors are numerous. So the entire cancellation probability is low.

5. On-Board TRP based on Test Response Compaction

Functional test needs a large number of pins to transform the data. In order to do functional test, the tester has to have the large I/O bandwidth. This will increase the test cost. So a similar idea of test response compaction of functional test is inspired. We can use the q-Compactor to compact the functional outputs, as reduces the monitoring tester channels. To verify our idea of reducing tester channels and the efficiency of handling X-bit of the q-Compactor during functional test, we set up a prototyping environment. We select a commercial SOC of MPEG 2 decoder as DUT, which contains 3M transistors with 42 input pins, 71 I/O (bi-direct) pins and 51 output pins. The tester is ETS 760 from High-Level Corp, which contains 196 tester channels and maximum test frequency is 35M. The q-Compactor is implemented by a quick FPGA. It satisfies the Theorem 2, Theorem 4 and has 51 pins. The environment is described in Figure 2. Table 4 lists that the tester channels we need provide when we directly test the chip and after we compact the outputs.

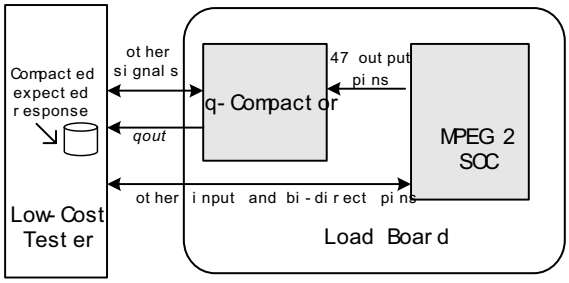

Figure 2. On-Board TRP Architecture with q-Compactor

Table 4. Tester Channels Reduction for MPEG- 2 SOC

Tester	Before Compaction	After Compaction
Tester Channels as driver	42	42+3
Bi-direction Tester Channels	71	71
Tester Channels as receiver	47	1
Total Tester Channels	160	117

Table 2. q-Compactor for ISCAS'89 benchmark

Circuit Name	Properties of CUT			Properties of q-Compactor				
	Vec.	DFFs	pins	INC	*r*	k	Area	Capacity
S13207	233	700	20	18	3	18	414	Theorem 2,4
S15850	96	611	20	18	3	18	414	Theorem 2,4
S35932	12	1763	64	62	3	62	1426	Theorem 2,4
S38417	68	1664	64	62	3	62	1426	Theorem 2,4
S38584	110	1464	64	62	3	62	1426	Theorem 2,4

Table 5. Capacity of handling X bits of q-Compactor during function test

Pattern	ratio. of X bit		num. of errors	times of errors observed			
				0	1	2	3
Pattern 1	0.0075%		3674	2	12	98	3562
Pattern 2	0.015%		736	3	13	175	545
Pattern 3	0.08%		5632	25	1126	2872	1609
Pattern 4	0.13%		7568	798	1984	3793	993
Pattern 5	1	0.31%	6524	1341	3113	1655	415
	2	0.09%	6524	19	1235	3718	1552

During function test, five test patterns to test different functional modules are applied to four bad chips. Each pattern is about 0.5M cycles test depth and has many X bits. These patterns are run and the erroneous data is collected in Table 5 (only give information of one typical chip for each pattern). Because the spread coefficient of our compactor is

3, one error is observed at most 3 times. From the Table 5, we can see, with increase of the ratio of X-bit, the error masking probability increases. But in the worst case (0.31%), there are only about 20.6% errors masked. It is acceptable since other errors can help us to determine the defective chip. Another point we must notice, to the pattern 5, if one pin that yields major X bits is gated by the *qin* and *qmask* when it is in X state, the ratio of X bit will drop quickly(from 0.31% to 0.09%). As the result of this enhancement, the error masked probability by X bits becomes low (about 0.3%).

Acknowledgments

The first author would like to thank Hair Corp. for providing the device in the section 5, Professor YiHe Sun and his students of Tsinghua Univ. to convert the test vectors for us**.**

6. Conclusion

Test Resource Partitioning is an effective technique to reduce test cost. It can be used to save test time, improve test frequency, reduce test data volume. In this paper, we present a TRP technique which can compact test response. If the compactor is placed on chip, it will shorten the scan chain and reduce the scan-in and scan-out shift cycles. Because only one output is needed to be monitored, 50% of test time can be saved compared to conventional full scan design. If our compactor is placed on load board, it can significantly reduce the necessary tester channels because of high compact ratio provided by the compactor. To guard the effectiveness of q-Compactor, some theorems are presented and these theorems can avoid the 1, 2, 3 or any odd number of errors cancellation and handle the X-bits in the outputs of scan chains. Moreover, the q-Compactor offers a full diagnosis capacity if required. Experimental results have shown the efficiency of the proposed technique.

References

[1] S.Mitra and K. S. Kim, "X-Compact an efficient response compaction technique for test cost reduction", *Proc. of ITC.*, pp. 311-320, 2002.

[2] S. Reda and A. Orailoglu, "Reducing test application time through test data mutation encoding", *Proc. of DATE,* 2002.

[3] E. H. Volkerink, A. Khoche, S. Mitra, "Packet-based input test Data compression techniques", *Proc. of ITC,* pp. 154-163, 2002.

[4] M. Nourani, J. Chin, " Testing high-speed SoCs using low-speed ATEs", *Proc. of VTS,* 2002.

[5] P. Wohl and J. A. Waicukauski, S. Patel, G. Maston, "Effective diagnostics through interval unloads in a BIST environment", *Proc. of DAC,* pp. 54-63, 2002.

[6] C. Barnhart, V. Brunkhorst, F. Distler, O. Farnsworth, A. Ferko, B. Keller, D. Scott, Koenemann, and T. Onodera, "Extending OPMISR beyond 10x scan test efficiency" *IEEE Design and Test ,* Vol. 19, No.5, pp. 65-73, 2002.

[7] K. K. Saluja and M. Karpovsky, "Testing computer hardware through data compression in space and time", *Proc. of ITC,* pp. 83-88, 1983.

[8] J. Rajski, J. Tyszer, C. Wang, S. M. Reddy, "Convolutional compaction of test response", *A paper submitted to ITC,* 2003.

[9] P. Wohl and J. A.Waicukauski, T.W.Williams, "Design of compactors for Signature-Analyzers in Built-IN SELF-TEST", *Proc. of ITC,* pp. 54-63, 2001.

[10] P. H Bardell, W. H. McAnney, J. Savir, "Built-In Test for VLSI: Pseudorandom Techniques", *John Wiley&Sons,* 1987.

[11] B. B. Bhattacharya, A. Dmitriev, M. Goessel and K. Chakrabarty, "Synthesis of single-output space compactor for scan-based sequential circuits", *IEEE Trans. on CAD&CS*, Vol.6, No. 2 , pp. 290-294, Oct. 2002.

[12] A.Ivanov, B. Tsuji, Y. Zorian, "Programmable Space Compactors for BIST", IEEE Trans. Computers, Vol.45, No. 12, Dec. 1996, pp. 1393-1405.

[13] E. J. Marinissen, V. Iyengar and K. Chakrabarty. "A set of benchmarks for modular testing of SOCs", *Proc. of ITC,* pp. 519-528, 2002.

[14] I. Pomerantz, S. Kundu, S. M. Reddy, "On output response compression in the presence of unknown output values", *Proc. of DAC,* pp. 255-258, 2002.

[15] S. M. Reddy, K. K. Saluja, and M. G. Karpovsky, "A data compression technique for built-in self-test", *IEEE Trans. on CAD & CS*, Vol. 37, No 9, pp. 1151-1156, Oct. 1988.

[16] Cullen and C. G. *Line,* " *Algebra with Application", Addison-Wesley, ISBN 0-673-99386-8, 1997.*

[17] R. A. Brualdi, "Introductory Combinatorics", *Third Edition, Pearson Education, Inc,* 1999.

[18] H. K. Lee and D. S. Ha, "On the generation of test patterns for combinational circuits", *Tech. Report No,12-93, Department of Electrical Engineering, Virginia Tech.*

[19] A. Chandra, K. Chakrabarty,"Test Resource Partitioning for SOCs", *IEEE Design and Test of Computers,* Vol 18 , pp 80-91, Sep.-Oct. 2001.

[20] I. Hamzaoglu and J. H. Patel, " Test set compaction algorithms for combinational circuits", *Proc. of ITC,* pages 283-289, 1998.