A 9-kbit Associative Memory for High-Speed Parallel Processing Applications

SIMON R. JONES, IAN P. JALOWIECKI, STEPHEN J. HEDGE, AND R. M. LEA

Abstract—Associative memories, by virtue of their regularity and distributed content-addressing capabilities, appear to be promising structures for the VLSI support of parallel processing applications. However, they are best integrated not as stand-alone structures, but rather as components of parallel processing chip architectures.

This paper reports on the design, development, and implementation of the 9-kbit (256-word×37-bit) associative memory used in the single-chip array processing element (SCAPE) chip, a CMOS VLSI associative parallel processor (APP) that integrates 256 associative processing elements (APE's) on a single 68-pad chip to achieve high-speed, cost-effective image and signal processing. This paper demonstrates that a static CMOS content-addressable memory (CAM) design is unsuited to the constraints of the SCAPE chip architecture and that a purely nMOS CAM cell provides the best compromise between the conflicting area, speed, power, and control requirements. Comprehensive details of this design are given together with an evaluation of its performance. Finally, a description of the design methodology used in the construction of the SCAPE chip is presented with a breakdown of circuit areas and operational data.

I. Introduction

IN RECENT YEARS, computer systems engineers have shown renewed interest in the design and integration of VLSI high-speed associative memories [1], [2]. However, it is interesting to note that the approach taken is not to implement a stand-alone associative memory, but rather to incorporate it into a parallel processing chip architecture. For example, the single-chip array processing element (SCAPE) chip [3]–[5] comprises a CMOS associative parallel processor, integrating 256 associative processing elements in a 256-word×37-bit associative memory, together with supporting control logic on a single 68-pad silicon die. This paper reports on the design, development, and implementation of the associative memory in the SCAPE chip.

An associative parallel processor (APP) is an SIMD computational structure, comprising a string of identical associative processing elements (APE's) which is designed to achieve high-speed and cost-effective structured data processing. As shown in Fig. 1, APP's are organized as a two-dimensional associative memory array (AMA) of content-addressable memory (CAM) cells, with microprogrammable bit (BCL) and word (WCL) control logic. In operation, the functions of the BCL are: to transmit search (viz. pattern matching) arguments to the AMA; to transmit WRITE (viz. update the contents of the CAM cell) arguments to the AMA; and to receive and amplify READ data from interrogated CAM words. The search and WRITE data are derived from the input data to the chip modified as specified by control signals.

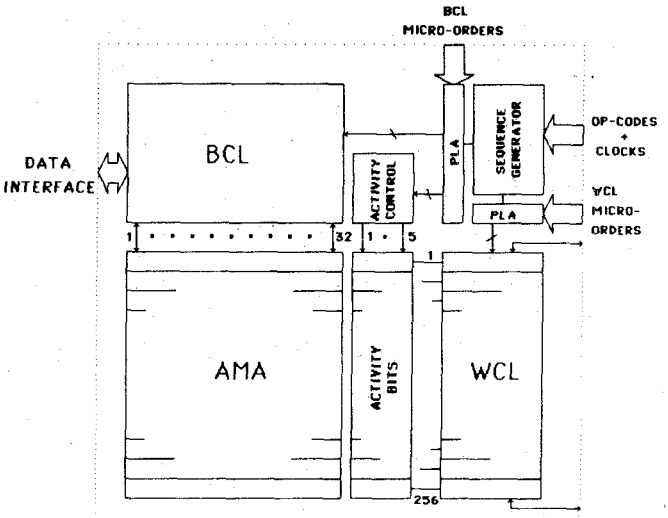

Fig. 1. APP schematic.

In operation, the function of the WCL is to take a bit vector of match/mismatch signals from the AMA, map it onto a new bit vector of AMA READ/WRITE signals, and feed this vector back into the AMA. The mapping function is selected by externally supplied control signals.

II. Associative Memory Structure

The BCL transmits match (search) and WRITE data to all CAM cells per bit column via two lines *Dx* and *Dy*. Valid arguments are ZERO, ONE and *X*. The *X* argument allows complete bit columns to be masked out of search and WRITE operations (viz. during a search, an *X* argument forces all CAM cells in that bit column to signal a match regardless of their data contents; similarly, during a WRITE

Manuscript received March 13, 1986; revised March 2, 1987 and May 27, 1987. This work was supported by RSRE MOD (Procurement Executive) and by the U.K. SERC.

S. R. Jones was with the Department of Electrical Engineering and Electronics, Brunel University, Uxbridge, Middlesex UB8 3PH, England. He is now with the School of Electronic Engineering Science, University College of North Wales, Bangor, Gwynedd, Wales.

I. P. Jalowiecki, S. J. Hedge, and R. M. Lea are with the Department of Electrical Engineering and Electronics, Brunel University, Uxbridge, Middlesex UB8 3PH, England.

IEEE Log Number 8719240.

0018-9200/88/0400-0543$01.00 ©1988 IEEE

operation, an X argument does not alter the data contents of selected CAM cells). The Dx and Dy lines can also be initialized and subsequently sensed during a READ operation. The WCL interface comprises three lines, namely RW, $M0$, and $M1$ for each AMA word row. The RW line selects a word for reading or writing. The $M0$ (match ZERO) and $M1$ (match ONE) lines transmit the match result to the WCL. Each M line in each AMA word row wire-AND's the CAM cell match outputs such that a mismatch in any cell will discharge the line.

Five of the bits in each word row have been designated "activity" bits. These activity bits are used for marking sets of word rows for intermediate processing steps. Furthermore, when the data word length is greater than the AMA word length, data words may be spread over many AMA words. The activity bits can be used in this case to delimit the start and end of data words. This activity "field" has been given its own match line MA. The five activity bits are not connected to $M0$ and $M1$, but wire-AND both their outputs to the MA line. The CAM cell discussed in this paper is a "data-bit" CAM design, but is easily modified to an "activity-bit" CAM design by connecting CAM match outputs to the MA line.

A. CAM Cell Design

Table I shows the functional operation of the CAM cell, as determined by the BCL and WCL interfaces. The performance of the BCL and WCL combined with the target 25-ns internal clock cycle and power dissipation limits, requires the CAM cell to meet or surpass the following performance requirements:

height: 36 μm; width: 46 μm
match/READ/WRITE/refresh: 5 ns
power: < 50 μW ($V_{dd} = 5$ V, $T_a = 27°$C).

As the SCAPE chip was to be fabricated in CMOS, a true CMOS cell (viz. incorporating both n-channel and p-channel transistors) would seem to be the natural choice. Indeed, with 9472 CAM cells to integrate, the inherently low power consumption of CMOS seems to be essential. Consequently, many static CMOS CAM designs were investigated before the circuit shown in Fig. 2 was identified as likely to provide the best combination of circuit complexity and performance. The data cell comprises a pair of cross-coupled inverters, driven by complementary data lines. The match logic uses a dual-NOR structure to discharge the M lines. However, on layout and simulation of this circuit the following problems were identified.

1) As a consequence of the area necessary for p-well separation, the CAM cell was significantly larger than the area allowed by the BCL and WCL pitches. Spacing out the BCL and WCL pitches to match the wider pitch of the CAM cell would have resulted in inefficient usage of silicon, and hence, an unacceptably large chip size.

2) The high switching currents inherent in CMOS designs are compounded by the parallel operation of the AMA (since it is possible in an APP to modify the contents of all CAM cells simultaneously). Moreover, when switching, these cells present a resistive load to the bit-column drivers, and thus to toggle an entire AMA bit column, powerful driver circuitry is required. In turn, the drivers add still further to the switching current drawn and to the chip power dissipation. These large switching currents result in a requirement for a complex power distribution strategy and excessive chip area being consumed by the column drivers.

It is interesting to note that Ogura *et al.* [6] in their design of a 4-kbit associative processor also recognize the above problems. Their solution to these problems uses a more complex static CAM design to reduce CAM switching currents. However, the 9-kbit target of the SCAPE chip, allied to the 2-μm technology, argued against the use of more complex CAM designs.

These factors stimulated further investigations into identifying alternative CAM cells and resulted in the all n-channel transistor pseudostatic CAM cell design, developed from previous work [7]–[9] and shown in Fig. 3. The same dual-NOR match logic is used as in the pure CMOS CAM cell.

The pseudostatic CAM cell offers several advantages over the static CMOS CAM cell, namely: 1) the use of all

TABLE I
CAM FUNCTION TABLE

Operation	RW	Dx	Dy	M	Comment
No WRITE	0	X	X	X	
WRITE ONE	1	1	0	X	RW pulsed to reduce power
WRITE ZERO	1	0	1	X	during all WRITE and
WRITE X/ Refresh	1	1	1	X	refresh operations
Match ONE	0	0	1	P	$M1 \rightarrow$ discharged if CAM = 0
Match ZERO	0	1	0	P	$M0 \rightarrow$ discharged if CAM = 1
Match X	0	0	0	P	$M0/1$ unaffected
READ	1	P	P	X	Dx/y discharged
No READ	0	P	P	X	

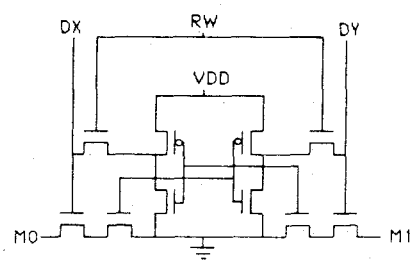

Fig. 2. Static CMOS CAM cell circuit diagram.

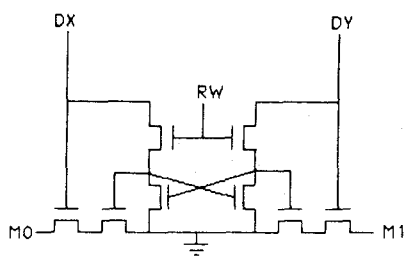

Fig. 3. Pseudostatic CMOS CAM cell circuit diagram.

n-channel transistors avoids the need for well separations and thus permits an efficient and compact layout; 2) n-channel circuits tend to have lower switching currents than the corresponding CMOS equivalents; and 3) unlike the pure CMOS CAM cell design, when a WRITE operation is performed, the n-channel CAM design only presents a resistive load to a high-driven data line. To a low-driven data line, the CAM cell presents a capacitive load. Consequently, while the p-channel transistors in the column drivers have to be substantial (to drive the data lines high against the resistive CAM cells), only modest n-channel drivers are required. This has a twofold effect: first, less area is occupied by the column drivers; and second, the more resistive n-channel transistors reduce the switching currents in the column drivers when driving the data lines.

As a consequence of these advantages, the pseudostatic design was selected for use as the CAM cell in the SCAPE chip.

B. Memory WRITE

Data can be written into the CAM cells by setting opposing values on the data lines (Dx, Dy) and driving the RW line high. This discharges the output of one of the cross-coupled pseudostatic inverters (viz. the one connected to the low-driven data line) which permits the charging up of the output of the other pseudostatic inverter in the CAM cell and accomplishes the toggling of the flip-flop.

The WRITE X (viz. mask the entire bit column during a WRITE operation) is performed by setting both data lines to a high value and then driving the RW line high. The data storage part of the CAM cell then behaves as an enhancement-load nMOS flip-flop. By ratioing the pull-up and pull-down transistors in the flip-flop appropriately, the positive feedback between the cross-coupled inverters can be made sufficient to maintain the data contents of the cell and hence realize a masked WRITE.

During WRITE operations, a direct current path exists between the data lines and the V_{ss} line. In order to reduce power consumption, the RW line is pulsed high for only 5 ns out of the 25-ns clock cycle, reducing the power consumption duty cycle. Furthermore, to allow the BCL to set up the data lines prior to the start of the WRITE operation, this pulsing of the RW line occurs about 20 ns into the 25-ns clock cycle. The delaying and pulsing of the RW line allows a WRITE operation to occupy only one clock cycle and thus significantly speeds up memory WRITE operations.

C. Memory Refreshing

A consequence of the pseudostatic nature of the chosen CAM cell design is a requirement for regular refreshing of the memory if stored data are not to be lost. In essence, the refresh operation is identical to the WRITE X operation as previously outlined: both data lines are set high and the RW line is driven high for about 5 ns. The positive feedback between the cross-coupled pseudostatic inverters that comprise the data storage part of the CAM cell restores the cell voltage levels.

D. Memory Matching

The match operation is implemented by driving the data lines low and simultaneously precharging the match lines ($M0$, $M1$ for the data bits, MA for the activity bits). During the next clock cycle, data lines are driven to the appropriate search argument (viz. search for a ONE, search for a ZERO, search for X). Should a mismatch occur between the search argument on the data lines and the data contents of the CAM cells, then the match line controlled by the high-driven data line discharges. The search X operation drives both data lines low and thus switches off the match discharge paths.

The use of two match lines ($M0$, $M1$) for the data field CAM cells allows not only match/mismatch information to be transmitted to the WCL, but also for the data contents of the CAM cell to be sensed. For example, if a search ONE operation is performed on a CAM cell, then the final state of the $M1$ line is identical to the CAM cell contents. Similarly, after searching for a ZERO, the $M0$ line represents the complement of the CAM cell data contents. This allows two bit values per AMA word row to be transmitted to the WCL simultaneously. By including a bit-serial ALU in each word-row slice, it is then possible to perform arithmetical and logical operations on the AMA data contents in a bit-serial, but word-parallel manner.

E. Memory READ

Reading from the memory is performed by precharging both data lines high and asserting the RW line of the CAM word selected for reading. This results in one of the data lines being discharged via the selected CAM cell at a faster rate than the other line, thereby indicating the data contents of the CAM cell. The dual constraints of the capacitance on the data lines and the small geometries of the CAM cell transistors make it impossible for the data line to be fully discharged to ground within a single 25-ns clock cycle. A voltage-differential sense amplifier has been designed and incorporated in the BCL. The sense amplifier can detect a voltage imbalance of 0.2 V on the data lines and amplifies this signal to set a READ result flip-flop. This allows a single-cycle READ to be implemented.

III. CAM Cell Evaluation

The chip has been fabricated using a 12-μm, two-layer metal, p-well bulk-CMOS process. Fig. 4 shows a microphotograph of one of the AMA quadrants. Each quadrant contains 64 words of 37-bit CAM. In the vertical axis, Dx, Dy, and ground lines are run in first-level metal and adjacent cells share common contacts. In the horizontal axis, second-level metal is used to run the match and RW lines.

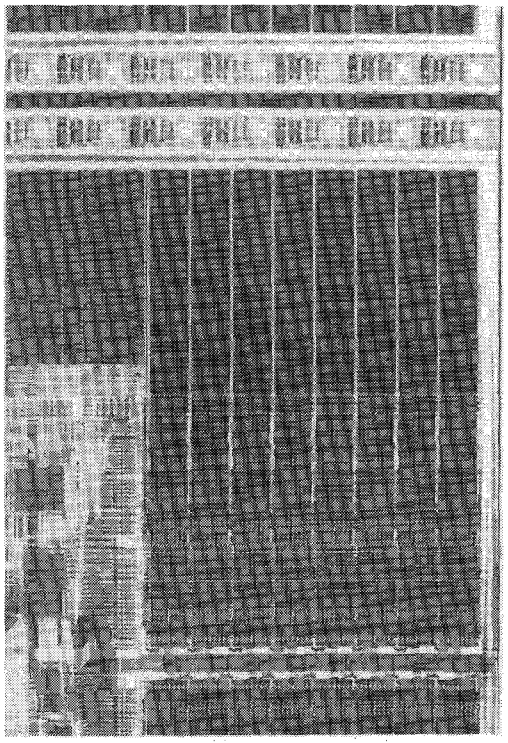

Fig. 4. AMA quadrant microphotograph.

TABLE II
CAM PERFORMANCE (V_{dd} = 5 V, T_a = 27°C)

Transistors	8
Height	36 μm
Width	46 μm
Match	5 ns
WRITE	5 ns
READ	5 ns
Refresh	5 ns
Power	44 μW
Data ONE	3.2 V
Data ZERO	0.0 V

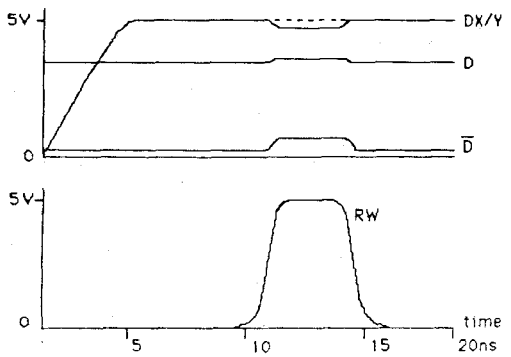

Fig. 5. Simulation of WRITE X operation.

The all n-channel design, combined with the sharing of common contacts between adjacent cells, allows a highly compact layout to be achieved. Indeed, it was found possible to meet the layout constraints dictated by the BCL and WCL pitches and thus achieve efficient usage of silicon.

A. Memory WRITE Evaluation

Table II gives performance details of the CAM cell. The use of an all n-channel transistor design means that the high-voltage level in the data part of the CAM cell can never rise to 5 V since a threshold is lost through the n-channel pull-up transistors. This threshold loss, when combined with the "body effect," results in a data ONE being represented by a 3.2-V differential between the output of the two pseudoinverters in the data part of the CAM cell.

The selection of the transistor sizes and geometries of the data part of the CAM cell proved to be a key part of the design. The ratios of these pull-up and pull-down transistors were chosen as a trade-off between acceptable WRITE speed, power consumption, and maintaining sufficient positive feedback between the two cross-coupled pseudoinverters to maintain the data contents of the CAM cell.

When choosing the transistor channel sizes for the data part of the CAM cell, the use of "large"-channel transistors results in a rugged design (viz. proof against any process spreads). However, the CAM cell power consumption increased dramatically with channel size. Consequently, the transistor channel width chosen was the largest size consistent with meeting the chip power limits. The surprisingly low-power figure in Table II arises from the use of the pulsed RW line for writing, which with a duty cycle of 20 percent (5 ns in every 25-ns cycle) reduces static power consumption.

Fig. 5 shows a simulated WRITE X operation on one CAM cell. The "high" side of the data part of the CAM cell is refreshed to approximately 3.2 V. The opposite side of the CAM cell rises to approximately 0.4 V when the RW line is driven high, only to be restored to ground when the RW line is lowered. These voltages provide an adequate safety margin against the voltage of the "low" side of the data part of the CAM cell rising above the threshold of the transistors in the "high"' side, leading to a possible corruption of the CAM cell contents.

B. Memory Refreshing Evaluation

As with its RAM counterpart, the pseudostatic CAM cell is self-refreshing and this confers two important advantages:

1) no extra unproductive circuitry is required, resulting in low circuit and system control overheads; and
2) the self-refresh capability of the CAM cell allows the entire memory to be refreshed in a single operation and minimizes the time overhead incurred by refresh operations.

The memory is refreshed every fourth clock cycle (i.e., at 100-ns intervals). Moreover, this memory refresh is overlapped with other on-chip operations which do not require memory access and results in no reduction in overall performance due to memory refresh.

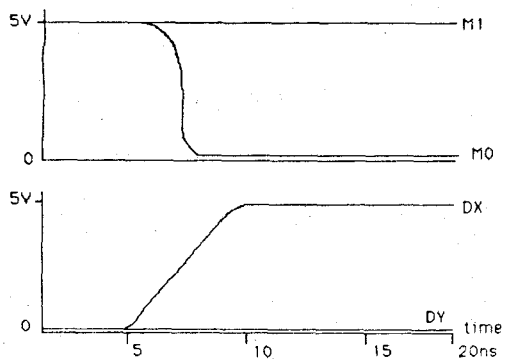

Fig. 6. Simulation of match operation.

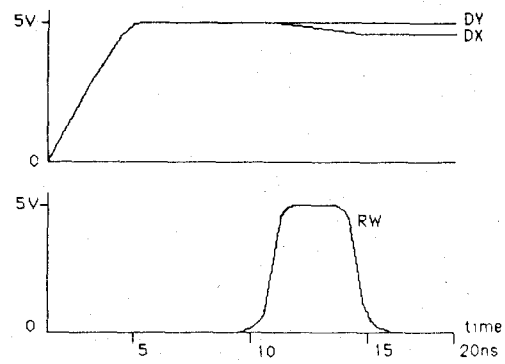

Fig. 7. Simulation of READ operation.

C. Memory Match Evaluation

Fig. 6 shows a simulated match operation. In this case the CAM cell stores a ONE and the search argument is a ZERO (i.e., a search ZERO operation is performed). Consequently, a mismatch occurs and the $M0$ line is discharged. This CAM cell matches more slowly than its static CMOS counterpart due to the high data voltage in the CAM cell having lost a threshold, and hence one of the gates in the dual-NOR discharge structure is not fully ON. In a static CMOS CAM, no threshold is lost and the relevant gate is fully ON. However, the absence of any well separation in the pseudostatic design allows the channels of the match discharge transistors to be enlarged sufficiently to compensate for the degraded gate voltage.

D. Memory READ Evaluation

Fig. 7 shows a simulated READ operation for a CAM cell storing a ZERO where the Dx line is discharged at a faster rate than the Dy line. A voltage differential of 0.2 V, sufficient for the sense amplifier to detect, is achieved within the 5-ns time scale.

An interesting effect, not found in random-access memories, arises out of the fact that with an associative memory it is possible to activate many of the words simultaneously. If this occurs with a READ operation, a situation can sometimes arise (e.g., when the majority of cells read in a bit column store a ONE and a minority store a ZERO) wherein the voltage on one data line can be discharged such that the data contents of minority CAM cells may be overwritten. Many solutions to this problem were considered. However, the complexity and control overheads of these solutions were such as to make it infeasible to incorporate these solutions on chip. Consequently, multiword reading has been proscribed at the system level with this design. Engineers are warned that while no damage to the chip will result if multiword reading is performed, corruption of CAM cell data may ensue. An investigation into the multiword READ problem previously discussed determined that no CAM cell corruption occurred when up to eight CAM cells in any one column were selected.

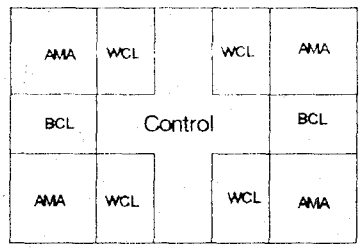

Fig. 8. SCAPE floor plan and microphotograph.

TABLE III
SCAPE CHIP CHARACTERISTICS (V_{dd} = 5 V, T_a = 27°C)

SCAPE die size	75 mm^2
Transistor count	143 000
Memory size	17.5 mm^2
Transistor count	76 000
BCL area	7.3 mm^2
Transistor count	13 200
WCL area	18.4 mm^2
Transistor count	41 500
Package	68 pin
I/O	TTL compatible
Memory organization	(32 + 5) × 256
Internal cycle time	25 ns
Power dissipation (40 MHz)	< 1 W

IV. CHIP DESIGN, FABRICATION, AND PERFORMANCE

The associative memory and its control environment described in this paper have been employed in the SCAPE chip [3]–[5], the floor plan and microphotograph of which are shown in Fig. 8. Table III gives a summary of the SCAPE chip characteristics. The SCAPE design style is based on a cellular design methodology, with highly engineered custom-designed cells being developed within the framework of a common grid. These cells were designed to exactly butt when composed in the chip floor plan. This design style was adopted in order to achieve the design objectives of minimum die size and maximum performance.

The SCAPE design was verified with extensive worst-case SPICE simulation of all cells and all critical timing-path cell compositions. The cell performance data thus obtained were fed back into a complete logical chip model, created using the HILO-2 logic simulator. This enabled verification of the complete chip design both in terms of cell interaction on chip and chip behavior as viewed from its

pins. In addition, using the logic simulator, test sequences were applied to the model that had previously been used on a Pascal simulation of the chip. This Pascal simulation consisted of a behavioral description of each of the major blocks in SCAPE. Thus a comparison between the circuit, logic, and behavioral descriptions of the chip could be performed to verify design correctness.

V. Conclusions

This paper has reported on the design, development, and implementation of a 9-kbit associative memory for use in high-speed parallel processing applications. For reasons of area and controllability, the natural choice (viz. a static CMOS CAM) was found to be unsuitable. Rather surprisingly, an all n-channel transistor pseudostatic CAM cell proved to be better suited to the design requirements of the SCAPE chip.

Careful design of the pseudostatic CAM cell combined with a pulsed WRITE signal resulted in remarkably low power consumption. Furthermore, delaying the arrival of the WRITE pulse allowed a single clock cycle, high-speed WRITE operation to be implemented.

Acknowledgment

The authors would like to acknowledge the support of M. B. E. Abdelrazik and C. Katolean whose contributions to the design and development of the SCAPE chip were invaluable.

References

[1] H. Kadota, J. Miyake, Y. Nishimichi, H. Kudoh, and K. Kagawa, "An 8-kbit content-addressable and reentrant memory," *IEEE J. Solid-State Circuits*, vol. SC-20, no. 5, pp. 951–956, Oct. 1985.

[2] C. Finilla and H. Love, "The associative linear array processor," *IEEE Trans. Computers*, vol. C-26, no. 2, pp. 112–125, Feb. 1977.

[3] R. M. Lea, "SCAPE: A single-chip array processing element for signal and image processing," *Proc. IEE Computers and Digital Techniques* (Pt. E.), vol. 133, pp. 145–151, 1986.

[4] I. P. Jalowiecki and R. M. Lea, "SCAPE—A programmable VLSI chip for signal and image processing," in *Proc. MILCOMP Conf. 1985*. Convex House, Tunbridge Wells, U.K.: Microwave Exhibitions and Publishers Ltd., pp. 155–160.

[5] R. M. Lea, "SCAPE: A VLSI chip for cost-effective image processing," in *Proc. IEE Colloquium VLSI Modules for Image Processing* (IEE Dig. 1983/10), Dec. 1983, pp. 4.1–4.5.

[6] T. Ogura, S. Yamada, and T. Nikaido, "A 4-kbit associative memory LSI," *IEEE J. Solid-State Circuits*, vol. SC-20, no. 6, pp. 1277–1282, Dec. 1985.

[7] S. R. Jones and R. M. Lea, "A CMOS CAM design for the associative memory of the SCAPE chip," Brunel Univ., Uxbridge, England, Tech. Memo. CM/R/136, Oct. 1984.

[8] R. M. Lea, "The comparative cost of associative memory," *Radio Electron. Eng.*, vol. 46, pp. 487–496, Oct. 1976.

[9] H. B. Ismail, "Comparative investigation of nMOS CAM cells," M.Sc. dissertation, Brunel Univ., Uxbridge, England, 1982.

Simon R. Jones was born in Llanelli, South Wales, in July 1958. After obtaining the B.Sc. (Hons) degree in computer science at Hatfield Polytechnic, he was awarded the M.Sc. and Ph.D. degrees in microelectronic systems engineering from Brunel University, Uxbridge, England, in 1983 and 1986, respectively.

Currently, he holds an IT lectureship in computer systems engineering at the University College of North Wales, Bangor, Wales. His research interests include VLSI/WSI computer architecture and the design of computational structures using molecular and biomolecular technologies.

Dr. Jones is a research consultant and an associate member of the Institution of Electrical Engineers.

Ian P. Jalowiecki was born in London, England, on September 1, 1957. He received a first class honours degree in electrical engineering from North East London Polytechnic in 1979, and the M.Sc. degree in microelectronic systems design from Brunel and Southampton Universities in 1983.

In 1978 he joined Marconi Avionics, where he was engaged in defense-related image and signal processing research. From 1983 until 1986 he worked as a Research Fellow at Brunel University, Uxbridge, England, engaged in research into the development of high-speed VLSI parallel processing architectures for image processing. He is currently a Lecturer in the Department of Electrical Engineering at Brunel University, specializing in microelectronics and computer architecture.

Stephen J. Hedge was born on October 4, 1960 in London, England. He graduated from the University of Sussex, Brighton, England, in 1982 with the B.Sc. degree in physics. Subsequently, he received the M.Sc. degree in microelectronics systems design in 1984 and the Ph.D. degree for work on inter-PE communications in microelectronic parallel processors in 1987, both from Brunel University, Uxbridge, England.

His research interests include VLSI architecture and design, wafer-scale integration, fault-tolerance, and parallel computer architecture.

R. M. Lea, photograph and biography not available at time of publication.