HIGHLY-PARALLEL, HIGHLY-COMPACT COMPUTING STRUCTURES IMPLEMENTED IN NANOTECHNOLOGY

D G Crawley, M J B Duff, T J Fountain, C D Moffat, C D Tomlinson

Department of Physics & Astronomy
University College London

Abstract

In this paper, we describe work carried out as part of the ARPA ULTRA program, in which we are evaluating how the evolving properties of nano-electronic devices could best be utilized in highly parallel computing structures. Because of their combination of high performance, low power and extreme compactness, such structures would have obvious applications in spaceborne environments, both for general mission control and for on-board data analysis. However, the anticipated properties of nano-devices mean that the optimum architecture for such systems is by no means certain. Candidates include SIMD arrays, neural networks and MIMD assemblies.

We explain that, because the propagation of signals through large arrays of nano-devices is almost certain to be a source of difficulty, our initial investigations center on the most regularly structured locally-connected architecture, the SIMD mesh-connected processor array. This structure offers the additional advantages of minimum external interfacing, conceptually simple redundancy schemes for fault-tolerance, and moderate memory requirements.

We describe the current phase of our program, in which we are simulating structures based on both resonant tunnelling devices and quantum cellular automata. In addition, we describe a novel architecture (the propagated instruction processor) which removes the requirement for other than near-neighbour connections for both data and control lines.

We calculate that the minimum anticipated device dimensions (in the order of a few tens of nanometres) would allow an array of 1000x1000 processors to be constructed on a few square mms of semiconductor. This would be equivalent in general performance to a system of at least 1000 DEC Alpha devices but would be optimally suited to the on-board analysis of sensor data, particularly in the form of images. Such a structure would offer ample computing power for the autonomous control of robotic systems.

We report initial results from the performance evaluation of our simulated structures using a comprehensive suite of algorithms. Our future program involves completing this evaluation for all of the SIMD-based systems and then, hopefully, extending our investigations to include the two other promising architectural configurations - neural networks and MIMD systems.

1. Introduction

In the next decade we are likely to see the emergence of several families of semiconductor devices with characteristic dimensions of the order 10^{-9}m. These so-called nanoelectronic devices will provide circuit elements which are several orders of magnitude smaller and several orders of magnitude faster than are currently available and will therefore present new challenges to computer architects. The potential use of these nanoelectronic devices to construct highly-parallel, highly-compact computing structures is being studied as part of the ARPA Ultra program [1,2].

The extremely small size of the devices will make it possible to incorporate millions of logic gates in a single chip and, whilst this offers enormous potential for high-speed, low-power computing, it does at the same time present major difficulties of organisation and control which will need to be studied with extreme care.

For ease of fabrication and in order to provide a comprehensible computer architecture which can be both tested and programmed intelligibly, a regular structure is to be preferred. Fortunately, a natural candidate architecture is available and has been studied in depth for the last twenty-five years: the Single Instruction stream, Multiple Data stream (SIMD) mesh-connected processor array [3].

SIMD arrays consist of assemblies of identical, simple processor elements (PEs), usually connected each to its nearest neighbours in a square array and each capable of only simple logic operations on single-bit data. Instructions are fed in a parallel stream to every PE and each instruction is executed simultaneously by every PE. Currently, arrays are in operation with between 32^2 and 256^2 PEs. Memory is distributed uniformly across the array so that each PE has access to, typically, several kilobytes of local data storage. Memory in this form is clearly well suited to image data structures (each pixel residing in the local memory of one PE) and both this and the parallel processing strategy which can easily be implemented in arrays have led to SIMD mesh-connected arrays being applied particularly successfully to image processing tasks [4,5].

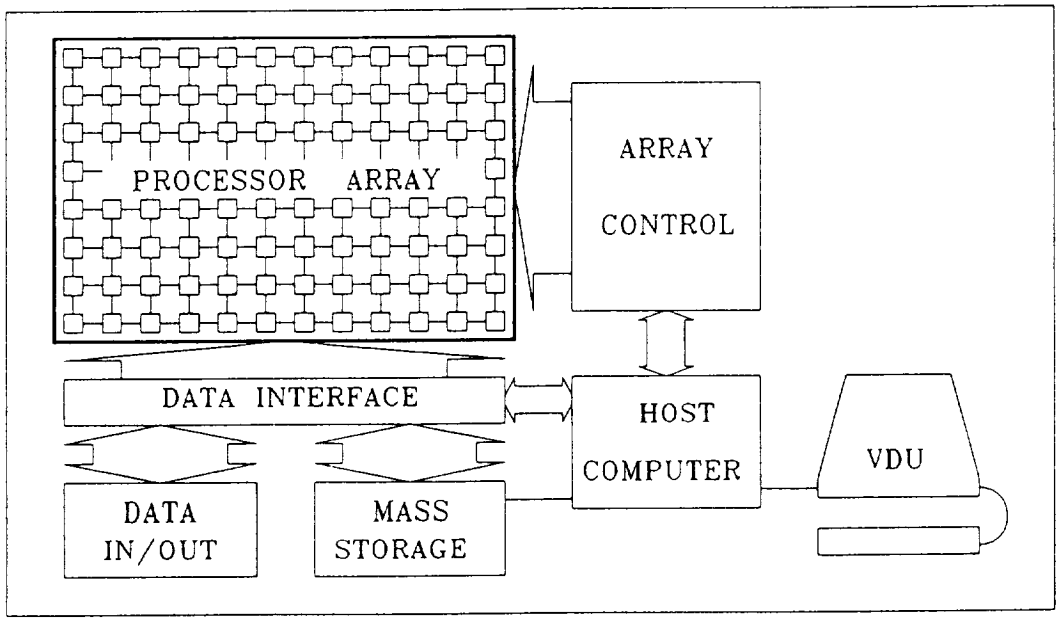

Figure 1 General architecture of SIMD systems

The large number of PEs in even the smaller arrays leads to a high degree of parallelism and hence to potentially large computing speed gains over single processors. However, the PEs are inherently simple and individually much less powerful than the processors found in, for example, workstations, so the gains from parallelism are offset against the losses due to simplicity of the PEs. Further losses are the result of underlying lack of parallelism in many of the tasks to be executed which make it difficult to make effective use of a major fraction of the PEs at any one time. The successful use of arrays in image processing stems from the fact that many image processing tasks are basically parallel in nature, especially considering those derived from convolution, which can be decomposed into local neighbourhood operations (i.e. operations in which the result value at any address (i,j) in the array is a function of the

original values at all addresses in the immediate neighbourhood of (i,j), typically for all points (x,y) in which $i-1 \leq x \leq i+1$ and $j-1 \leq y \leq j+1$).

One disadvantage of SIMD arrays is the comparatively inefficient manner in which data has to be transported across the array when the algorithm so requires it. A good example of this problem is image rotation in which, ultimately, it might be necessary to move pixels from one side of the array to the other. If the only connections available are those between neighbouring PEs, then this process has to be executed in a sequence of steps from PE to PE along the required path and may therefore involve a very large number of system clock cycles. A similar problem occurs in relation to inputting data to and outputting data from the array, although in conventional arrays, this is usually achieved by providing additional data paths along the principal array directions.

2. Nanoelectronic Arrays

Arrays constructed from nanoelectronic devices will have, in general, all the benefits and disadvantages of the arrays constructed using present day technology. The expected additional benefits are larger array dimensions (permitting, for example, the processing of larger images and other data sets) [6], higher speeds (leading to increasingly complex programs which will run in real time) and lower power consumption and weight (resulting in higher system portability). As the technology matures, it can be expected that the benefits will be obtained with no significant increase in cost.

On the other hand, it can also be anticipated that larger array sizes and the nanoelectronic technology will introduce additional operational difficulties illustrated in Figure 2. The current indications are that it will not be possible to incorporate many (if any) 'wire-like' connections to and between PEs over distances much greater than those between PEs; the dimensions of the active elements are so small that loss-free metal links between them would occupy too great a fraction of the substrate space. Reducing the dimensions of the wire connections to the nanometre range would introduce unacceptable losses. This implies that the SIMD instruction stream could not be carried on a word parallel bus running to every PE.

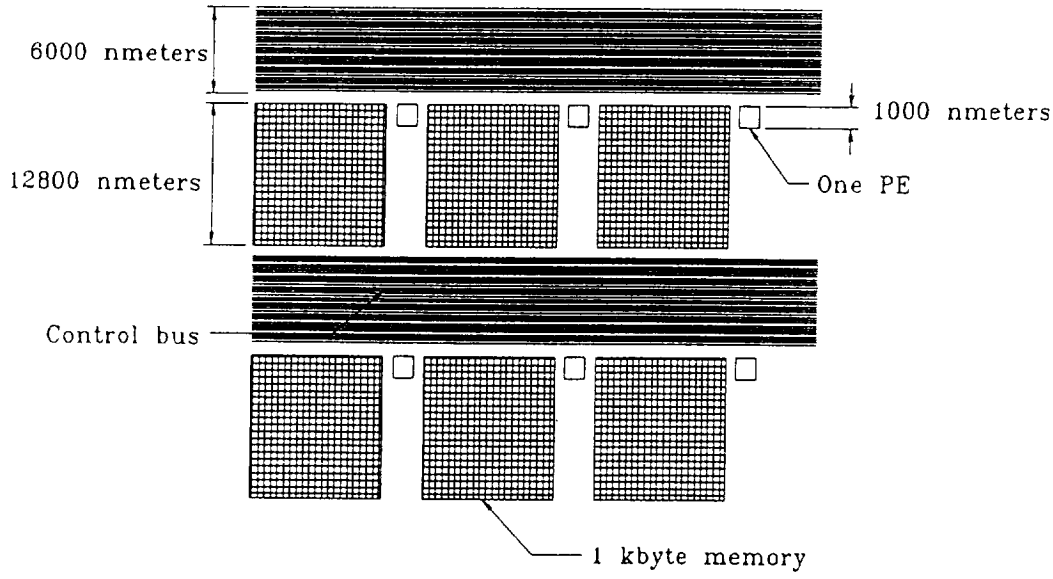

Figure 2 A conceptual floor plan illustrating the relative areas occupied by PEs, memory and control bus

In the same way, data transport from outside the array into the local memory in the array would also have to be serialised. Under these circumstances, in which the PEs are used as the instruction or data paths, then PE control would no longer be simultaneous and *O(N)* clock cycles (for an N x N array) would be needed for each step of a parallel operation [6].

A further problem which might be expected to arise in large arrays is that material and process defects could make it almost impossible to construct an array with all its PEs fully operational. Test techniques will have to be devised to pinpoint the defective PEs and circuit redundancy (with error correction) incorporated into the design.

Finally, preliminary design calculations have indicated that for an array with conventional computing capability, the area occupied by the local memory completely dwarfs that taken up by the PEs. Unless between-chip optical interconnect can be achieved, it will not be feasible to remove local memory from the array chip and array sizes will not be as large as had been hoped.

3. Circuit Design and Simulation

Nanoelectronic technology is in a very early stage of development and it is by no means clear which of the various proposed types of device will be the first to become available for practical use in computing circuits. The current indications are that both *resonant tunnelling devices* (RTDs) and *quantum cellular automata* (QCAs) are worth serious consideration, although the former are likely to become available rather sooner than the latter [7].

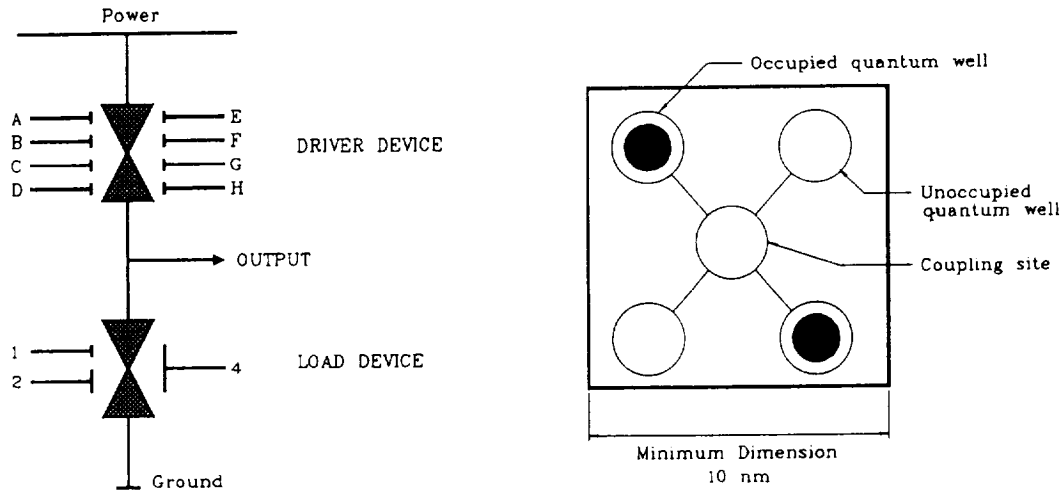

Figure 3 Nanoelectronic devices: (a) A resonant tunelling threshold logic element, (b) a quantum cellular automaton

3.1 RTD-based circuits

Devices based on resonant tunnelling effects can be scaled down to very small dimensions since their operation depends on the ability of current to tunnel through potential barriers; it is this same tunnelling effect which causes ordinary transistor operation to break down (through current leakage) when attempts

are made to scale them to comparably small dimensions. In the present programme, an element has been devised by placing two RTDs in series, one of which acts as a driver and the other as a load. By suitably adjusting bias currents on the two devices, the elementary circuit can be switched between two stable states. A further extension of this proposal is to provide the biasing by means of multiple contacts to both RTDs so that the circuit produced behaves as a general purpose threshold gate shown in Figure 3(a). This can then be tuned to implement thresholding functions which, in turn, can be used as the basis for producing a comprehensive selection of Boolean logic functions. The universal nature of these implementations leads to the possibility of the fabrication of complex circuits with a high degree of regularity, whilst the reduction in number of devices (from the hundred or so required in a conventional implementation) to two offers further benefits of compactness.

3.2 QCA arrays

An alternative nanoelectronic structure is based on single electrons occupying quantum wells. These also scale to extremely small dimensions, an element being a small as 10^{-8}m square. Two state logic elements can be envisaged which comprise five quantum wells in a configuration shown in Figure 3(b).These elements can be assembled into processor circuits such as that shown in Figure 4, which will execute all the computational functions required in a general purpose array processor.

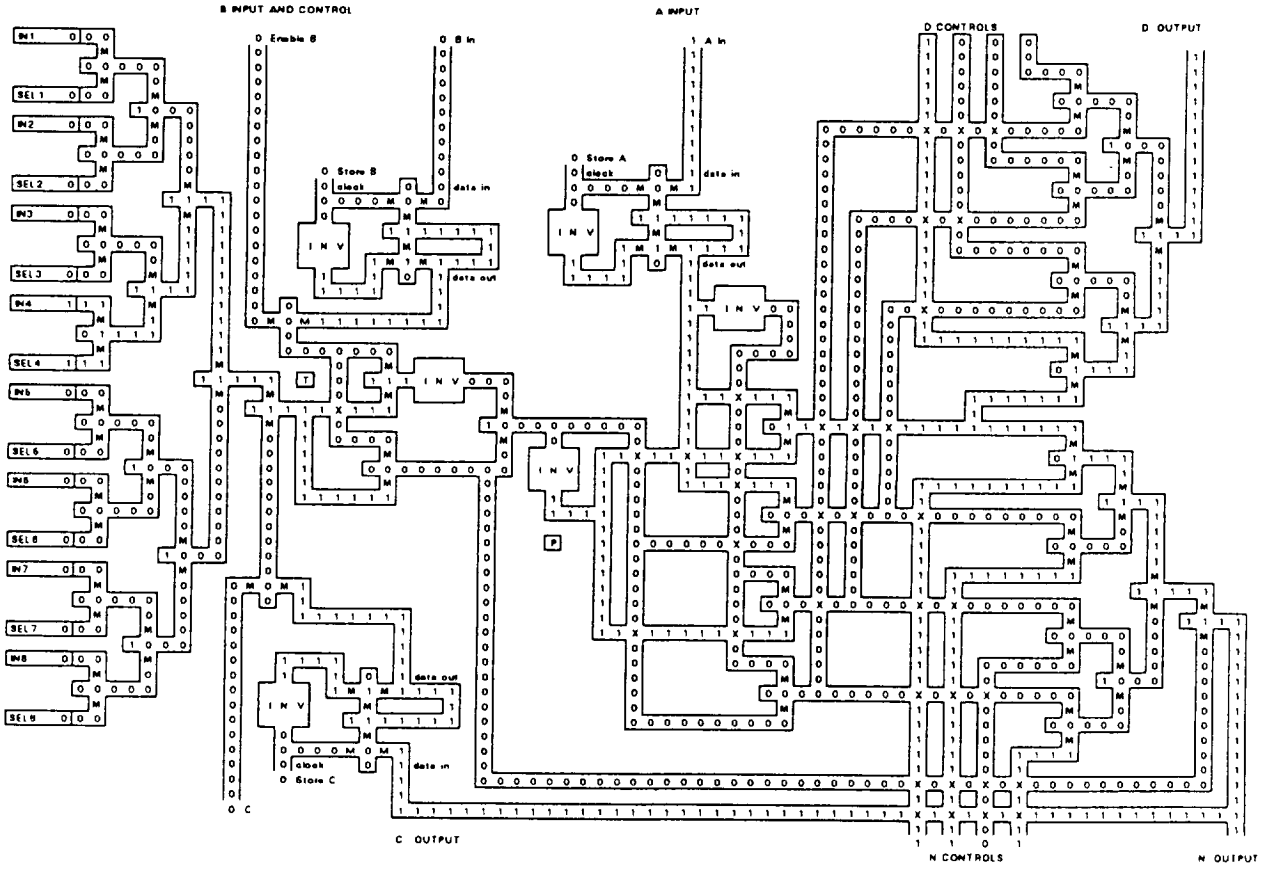

Figure 4 An EXCEL simulation of the CLIP4 processing element

In the case of QCA arrays, the technology is less well advanced than that for RTDs and, at this stage, it is only possible to explore potential circuit designs without attempting to obtain hard information as to

expected performance or device size. A simulation based on an EXCEL spreadsheet is being used for this purpose and yields some performance information in terms of clock cycles [8]. The simulation also gives some idea as to the probable complexity of the proposed circuits (i.e. the number of devices required) and indicates the type of layout that might be involved.

3.3 Array simulation

Enough is known about the expected properties of RTDs to make it worthwhile simulating RTD arrays both in hardware (employing large scale RTDs and heterojunction bipolar transistors) and in software in order to evaluate candidate array architectures. A versatile, semiconductor technology independent, software array simulator has been written as part of this project and is being used in conjunction with a

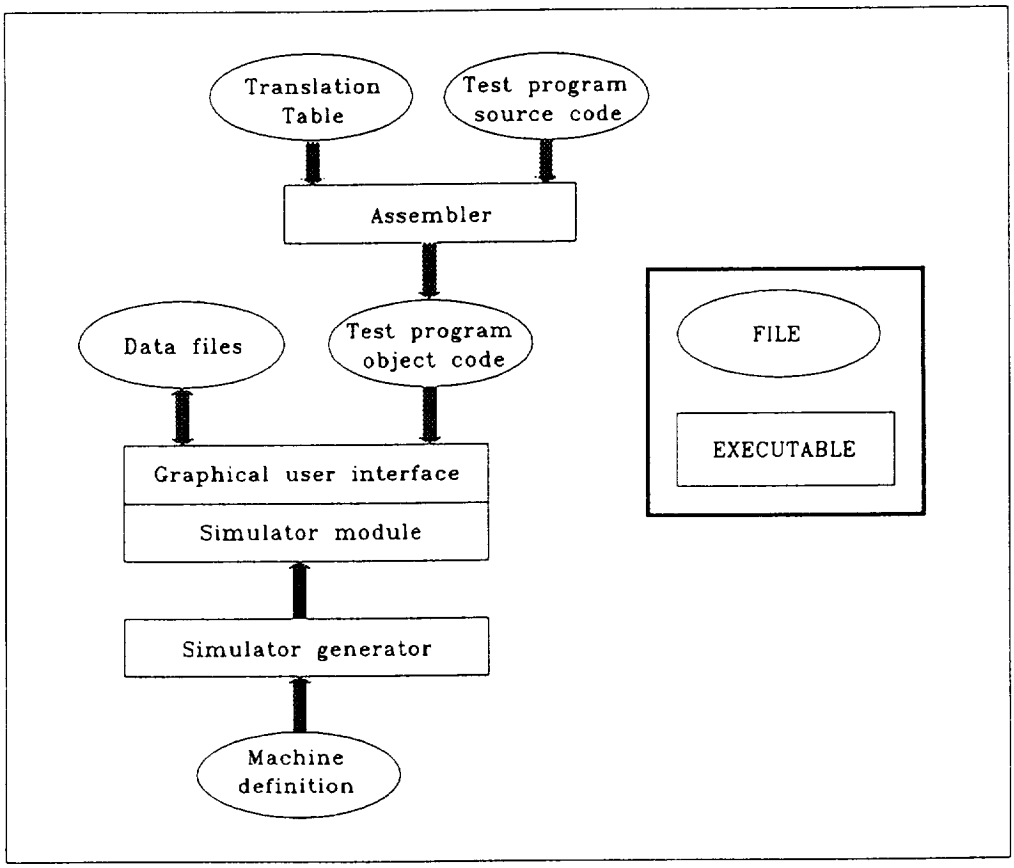

Figure 5 The array simulator software

specially selected suite of image processing algorithms [9]. This allows an assessment to be made of the performance to be expected for appropriate combinations of technology, PE circuit design, array architecture and detailed algorithmic implementation, noting here that efficient algorithms defined at the task level (e.g a Fourier Transform) can be expected to require a detailed design which takes into account the computer structure on which the computation is to be performed. At the very least, full account must be taken of the parallelism to be encountered. The structure of this simulator is illustrated in Figure 5.

As a first step, simulations are being based on the CLIP3 array processor which was developed by some of the authors in 1973. This array was constructed and, in a slightly modified form (CLIP4), used for image processing continuously over a ten year period [10]. It is therefore well understood and a good

vehicle for the studies now being made. Experience in programming such arrays is invaluable in order to maximise the probability that the best possible array performance is being achieved for a given algorithmic task.

4. The Propagated Instruction Processor

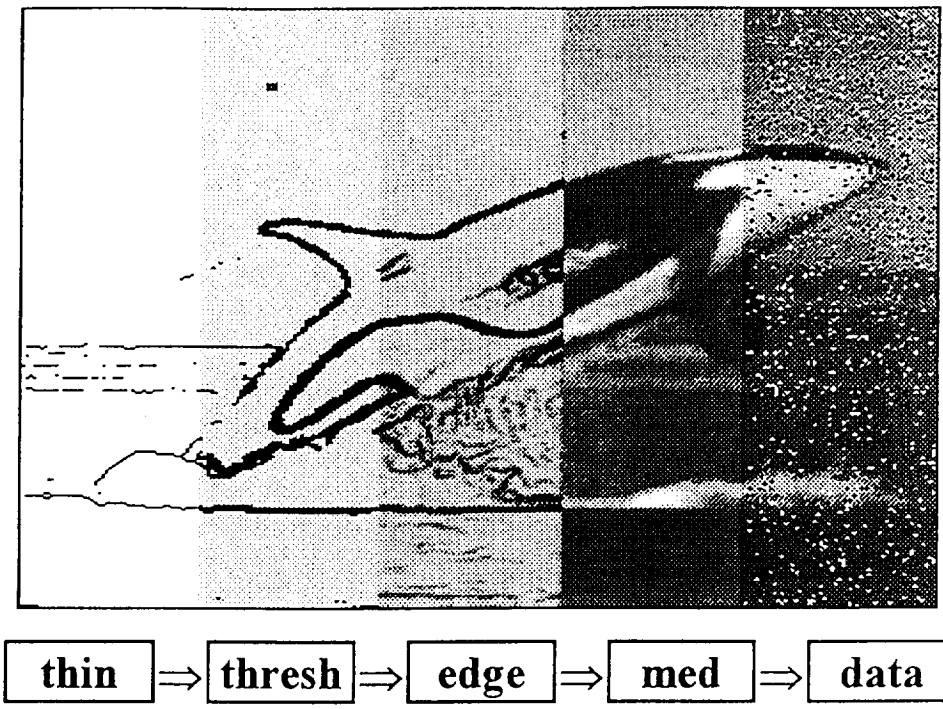

Figure 6 An illustration of propagated instruction processing

If the difficulty of distributing PE instructions (due to the non-availability of wide instruction busses) is as severe as has been feared, it might be necessary to consider a very different way of controlling the processor arrays. Since it will be argued that data (and therefore instruction) paths will have to be stepped from PE to PE via the nearest neighbour connections, then it is important to investigate whether this process can be carried out without losing large numbers of clock cycles. One solution is suggested: the instructions are clocked into the array from left to right along each row. As each PE receives an instruction (or as each group of columns of PEs receive an instruction), the instruction is executed and the results stored in local memory. A sequence of instructions is therefore executed in a travelling wave across the array and the array behaves as a type of pipeline in the manner illustrated in Figure 6 [6]. With careful design of the instruction stream, it has been shown that apart from some latency in the pipeline, programs can be executed in a time which is not significantly longer than that achieved in a conventional SIMD array. An exception to this occurs when the operations being performed are not confined to local neighbourhoods and involve propagating data in various directions across the array. Strategies for dealing with this situation are being evolved in simulation and will be evaluated later in the project. At the worst, instructions involving propagation of data can be executed by first propagating the instruction across the entire array and then allowing time for global propagation as the processed data steps through PEs to all parts of the array. Once again, $O(N)$ time is involved but it is considered that this type of operation forms only a small fraction of the total in typical programs.

5. Applications for Nanoelectronic Computing Arrays

The minimum anticipated device dimensions (in the order of a few tens of nanometres) would allow an array of 1000 x 1000 PEs to be constructed on a few square mm of semiconductor. Although it is too early to state with certainty what will be the switching frequency of any particular device, a target figure of around 100 GHz (for both RTDs and QCAs operating at room temperature) would seem to be feasible [11,12]. The functionality obtainable with the more complex nanoelectronic devices can be compared with simple circuits constructed in present day technology and requiring possibly 10 clock cycles at switching frequencies of 100MHz. This would indicate a speed gain of at least 10^4.

If these preliminary figures are considered in relation to the use of nanoelectronic SIMD arrays for image processing, the potentially large array size would allow high resolution images to be processed at television frame rates at which the complexity of the processes being performed could be increased from the currently attainable simple level, such as skeletonizing binary images, to the complex vision level involving the detection and recognition of objects in an uncontrolled environment. A more precise evaluation of the performance to be expected from such systems is now being carried out, using the simulator and the test suite of image processing algorithms discussed above.

It is clear that the use of nanoelectronic based vision systems in conjunction with robotic systems would lead to the possibility of autonomous navigation of exploratory vehicles under conditions in which response rates are of importance and where the transmission delay for signals sent back to a control station would prevent the possibility of real-time steering and vehicle control. More generally, in any circumstances in which the conditions are hazardous to human life, one of the main reasons for requiring human operators to be present can be eliminated by equipping robotic systems with a vision capability. Current research in this area [13,14] illustrates that neither presently available on-board systems nor remote control by powerful off-line computers offers sufficient intelligence for this task.

Although the main emphasis of this discussion has been on the use of nanoelectronic SIMD arrays for image analysis, conventional technology SIMD arrays have already been applied to a wide range of non-image problems. However, in general, SIMD arrays do not perform optimally on high-level tasks in vision systems and in more general control applications where many varying data types from a range of sensors may be involved. In this type of application and as part of the Ultra programme, the authors are also considering the use of nanoelectronic devices in other computer architectures, especially **M**ultiple **I**nstruction stream, **M**ultiple **D**ata stream (MIMD) systems and neural networks.

MIMD systems can be envisaged as small arrays (typically 32 - 128 PEs) of relatively complex PEs, sometimes assembled on a high speed bus to which is also connected memory, some of it being local to each PE and some accessible by all the PEs. Alternative connection structures are also employed in which direct paths are supplied between selected subsets of PEs (thus enabling especially high performance over a chosen range of commonly occurring algorithms). Systems of this type do not assume data to have an intrinsic two-dimensional structure (as is ideally the case with SIMD array processors) and each PE will usually operate with relatively little direct communication with other PEs (compared with mesh-connected arrays). SIMD arrays obtain their parallelism by operating in parallel on different parts of the data whereas MIMD arrays can also obtain parallelism by splitting the program into parallel sections. Provided the task to be performed is not essentially serial in its structure, MIMD systems can achieve $O(P)$ performance gains (compared with single processors), where P is the number of PEs in the system. The lack of dependence of MIMD systems on the data structure makes them most suitable for processes in which many different data types are involved.

The compact nature of nanoelectronic systems gives the opportunity, in principle, of constructing, and employing in a space environment, hybrid systems combining both SIMD and MIMD subsystems, thereby acquiring the special capabilities of both subsystems with negligible increase in loading on the spacecraft power supplies or payload weight. The problems encountered in operating hybrid systems are being studied by the authors and their collaborators in another research programme [15].

The use of nanoelectronic devices to construct neural networks is also being investigated by the authors. Neural networks have been shown to provide powerful solutions to a broad spectrum of problems ranging from natural speech recognition to financial forecasting. These computing circuits are trained by example, rather than explicitly programmed. On the negative side, whereas the networks exhibit the capability of generalisation (e.g. by classifying objects *similar* to ones in the training set but not *included* in the training set), each network is designed to work on a specific problem or, at the most, a specific class of problems. However, in a situation in which autonomous behaviour in a previously unexplored environment is wanted, it could be that a neural network solution might be the most likely to succeed.

Two alternative approaches are being considered. In the first, neural network algorithms can be embedded in mesh-connected SIMD arrays [16] so the array is programmed to act as a neural network (which then has to be trained for its specific function). In the second approach, nanoelectronic devices directly implementing the thresholding function employed in neural networks are connected together in a network structure. This architecture may well prove difficult to construct as many neural networks which are currently being investigated require large numbers of interconnections within the assembly of neural elements. The difficulties already described in attempting to provide long-distance connections between nanoelectronic components of a circuit may well render this approach unworkable. Nevertheless, the well publicised successes of neural networks in some areas of application make it worthwhile not to discard these studies too readily.

6. Conclusion

The potential offered by nanoelectronic devices for the construction of highly-compact computing structures with extremely high clock frequencies, coupled with lower power consumption and low weight, makes them ideal candidates for space-borne computing systems in which communication times back to the control centre are too long to permit effective real-time control.

Circuits are envisaged in which millions of nanoelectronic devices will be combined on one semiconductor substrate and it is realised that the control and programming of such circuits will present new challenges to software engineers and programme designers. The experience gained over the past two decades in working with SIMD arrays is expected to be invaluable in the studies which must now be made and is influencing the design of nanoelectronic circuits in the direction of array architectures.

Although no fully scaled nanoelectronic devices operating at room temperature have yet been fabricated, the progress in that direction is sufficiently encouraging to stimulate plans for employing the devices which will emerge in the next decade and to consider them as excellent candidates for space applications.

7. References

[1] ARPA ULTRA Electronics Program Review, Presentation Summaries, Santa Fe New Mexico, (1993)

[2] ARPA ULTRA Electronics Program Review, Presentation Summaries, Santa Fe New Mexico, (1994)

[3] Flynn M. J., Very high-speed computing systems, Proc. IEEE 54, pp. 1901-1909, (1966)

[4] Cellular Logic Image Processing, Eds. Duff M. J. B. and Fountain T. J. , Academic Press, (1986)

[5] Strong J. P., Computations on the MPP at the Goddard Space Flight Center, in Proc. IEEE Vol. 79 No. 4, Eds. Maresca M. and Fountain T. J., pp. 548-558, (1991)

[6] Fountain T. J. and Tomlinson C., The Propagated Instruction Processor, in Proc of BMVC 95, pp. 563-572, BMVA Press, (1995)

[7] Moffat C., A Survey of Nanoelectronics, Internal Report 94/2, Image Processing Group, Dept. of Physics, University College London, (1994)

[8] Tougaw P. D. and Fountain T. J., The Design of Computer Memory Elements using Adiabatically-switched Quantum Cellular Arrays, in preparation, (1995)

[9] Crawley D., Flexible Simulation of Processor Arrays, submitted to 4th Euromicro Workshop on Parallel & Distributed Processing, Braga, Portugal (1996)

[10] Fountain T. J., Processor Arrays: Architectures and Applications, Academic Press, (1987)

[11] Tougaw P. D. and Lent C. S., Dynamic Behaviour of Coupled Quantum Dot Cells, in Proc. 3rd Int. Conf. on Computational Electronics, Ed. Goodnick S., (1994)

[12] Chang C. E. et al., Analysis of heterojunction bipolar transistor/resonant tunneling diode logic for low-power and high-speed digital applications, IEEE Trans. on Electron Devices 40 4, pp. 685-691, (1993)

[13] Carnapete L. S. et al., A miniature retina-based AGV called VAMPIRE, in Proc.CAMP95, IEEE Computer Society Press, (1995)

[14] Danese G. et al., PAVIA: A Control System for Active Vision, in Proc. CAMP95, IEEE Computer Society Press, (1995)

[15] Architecture and Programming Model of an SIMD-MIMD system for Image Processing, Esprit BRA Project 8894 Deliverable A45, (1994)

[16] Forshaw M. R. B., Implementation of Neural Network and Optimisation Algorithms on Mesh Arrays, Internal Report 95/1, Image Processing Group, Dept. of Physics, University College London, (1995)