Advance into the past

Ivor Catt looks at the inhibitions imposed on designers of computers by the conventional mythology of devices and architecture.

by Ivor Catt

Rational forward progress in computer technology could only be achieved if a significant proportion of computer scientists had some mastery of most of the technologies and disciplines involved. Unfortunately this is not the case, because the necessary spread of knowledge and understanding – from semiconductor physics at one extreme to complex software and computer applications at the other – is too broad.

Computer scientists habitually assume that the conventional wisdom, or myth, imposed on other specialities than their own, is true. They find it convenient to base their views on the state of the art in other fields on information supplied by amateurs rather than those actually working in them. A specialist in any one field tends to see his professional survival as depending on the stabilization of the conventional-wisdom straight-jacket which at one time or another has been imposed on every other speciality. This is because change in these other fields would make his own speciality too fluid, and he would not survive . . . a point of view which, although usually subconscious, sometimes comes out into the open.

For example, around 1970 it was commonly said, "We are having so much difficulty mastering the software of present computers that it is important, *if we are to progress*, that computer hardware be frozen for a decade or more." Some readers will see the irony implicit in this comment, which was often made by programmers with no knowledge of engineering, which meant virtually all programmers. There followed an explosion of complex software techniques, including list processing, which could have been much more easily achieved by hardware modification; but this option had been outlawed. The result was an increase in the complexity and confusion of already over-complex software, and a deterioration in the overall position.

In general, all other disciplines ganged up on each individual discipline and forced it to remain essentially static, at least in its perceived structure when it interfaced with other disciplines. Examples are:

- The blocking of any blending of memory and processing, any move away from absolute von Neumann, and strict adherence to the 'von Neumann bottleneck', even though at one extreme the technology was demanding it and at the

C.A.M.

other extreme almost all applications were demanding it.

- The blocking of any deviation from the traditional drift from fully serial machines to fully word-parallel machines, even though (a) the technology demanded a reversal towards serial working, (b) the change in the relative cost of circuit and interconnection demanded it, and, strongest of all, (c) a strong mythology had developed that the computer industry was combining with an avowedly serial industry, telecommunications (citing the appointment of a Minister for Information Technology as evidence). Here we see one myth combating and overcoming another, unfortunately the wrong one, "fully parallel fetishism", being the victor.
- The imposition for all time of the t.t.l. logic signal as industry standard. This occurred even though t.t.l. logic, which came into general use in spite of its weaknesses in design (including the heavy standing current in signal lines, the high signal swing, etc), had given way as the industry standard circuit to c.m.o.s., which had much greater circuit density, in which a very different logic signal standard would have been more efficient.
- The maintenance of a key feature of the thermionic valve – the idea that hermetic seal was necessary to stop the cathode from burning up – well beyond the disappearance of the cathode through drastic changes in the technology towards silicon semiconductor l.s.i. Few computer engineers realise that the 'hermetic seal fetishism' which continues today in v.l.s.i. chips dates back to the danger of allowing oxygen to reach a hot cathode, and has nothing to do with semiconductor technology.
- The inexplicable standardization, without a murmur, on the use of Kovar as the metal for the leads coming from an integrated circuit chip, even though every parameter of Kovar except one is bad in this application. The one good parameter is that Kovar wets to glass, so allowing the formation of a hermetic seal. Kovar's bad features include the following:

– It has rather high electrical resistivity, so degrading performance by creating extra voltage drop in the signals entering a chip.

– It is magnetic, so that signals into a chip are delayed, and energy wasted, while the magnetic field is built up.

– It is not ductile, and work hardens fast, so that there is an unnecessarily large risk of fracture due to bending or vibration.

– Worst of all, it does not wet to solder. In order to make it possible to solder to a Kovar lead, the lead has to be gold plated. However, during the soldering process, the gold dissolves into the solder, creating a brittle alloy and also, should soldering and de-soldering be repeated, the dissolving away of all the gold and the creation of a dry, non-wetted joint between solder and virgin Kovar.

- Microprocessor manufacturers have displayed ignorance of the mechanism of digital signal propagation and voltage decoupling. Placing voltage pins at opposite corners of the package, thus introducing a large single-turn inductor in series with the voltage supply, is the worst possible pin choice, limiting the speed of microprocessors and also making them pattern-sensitive. Although only marginally significant in the old 14 or 16-pin dil integrated circuit, the problem created increases rapidly as the square of the package length, making the large microprocessor chip slow (only 4 MHz), pattern sensitive, and depdendent on the layout of the host printed-circuit board in a manner not understood (and so not predicted) by system designers.

● Looking at another aspect of the standard package, I suppose that I should be relieved that the industry did not standardize on the even more absurd IBM SLT package, 1965 vintage, which had a line of pins down all four sides of a square package. When deciding how the pins should exit from an integrated circuit package, the decisive aim should be to minimize the obstruction of printed circuit conductors in the host p.c.b. The two, unrelieved, lines of pins are about as obstructive, and therefore as inefficient, as it is possible to devise (*pace* the IBM SLT). Alternate pins should have been staggered, and this is a simple operation (which would not have created significant problems in the manufacture of i.c. sockets). I only mention this to show how thoughtless and casual developments have been, not to propose change at this late stage.

● The standardization by the industry of t.t.l. with its totem pole (push-pull) output was based on the mistaken idea that the load seen by an i.c. output is capacitive. This was true for thermionic-valve logic gates, with their high impedance, low current outputs, but ceased to be true when we used transistors, at which point the load seen by a fast output became resistive; either a transmission line characteristic impedance (resistance) or a t.t.l. input load (also essentially resistive). Whereas a capacitive load could helpfully be driven push-pull, today a resistive load can perfectly well be driven by one transistor, as is demonstrated by the fact that the fastest existing circuit, 1 ns e.c.l. has a single transistor output.

Speed of logic

Generally, the limiting factor in the speed of logic is not the time taken for a transistor to switch on or off, but rather the time taken thereafter for the switched current to charge or discharge the stray capacitance in the line connecting this transistor to the next. A good measure of the delay involved, i.e. the gate delay, is gained by multiplying the resistance of the drive transistor when switched on by the stray capacitance that it has to drive.

When a bipolar transistor, as used in a t.t.l. circuit, is switched on, its resistance is less than 10 ohms. The capacitance of the line, or wire, on the printed circuit board joining this output to the next logic element is of the order of 20 picofarads. Multiplying these two together gives us a time delay of 200 picoseconds. This shows us that, from this point of view at least, sub-nanosecond logic speeds are possible and we do not pay a speed penalty if our logic signals skip from chip to p.c.b. to chip to p.c.b. and so on.

In stark contrast, the smallest possible unipolar, or mos transistor, when switched on, still has a resistance of 10,000 ohms. If it drives 20 picofarads of capacitance on a printed circuit board, the delay, or signal rise time, resulting would be 20 pico multiplied by 10,000, that is, 200 nanoseconds. So if the physically smallest possible (i.e. square) cmos output transistor has to drive a signal off the chip onto the printed-circuit board, the achievable speed is only 200 nanoseconds, that is, one thousand times slower than bipolar t.t.l. This dire situation can be improved by making the drive transistor bigger and so reducing its resistance. Actually, we might put ten square transistors in parallel to reduce the resistance from 10,000 ohms to

1,000 ohms. However, the price we pay is that these drive transistors have to be made very big, consuming large areas on the surface of the silicon chip. This undermines the reason for using mos which is that an mos circuit takes up less area on the chip than does a bipolar. By the way, if we make the output transistor more beefy, we can make the mos output t.t.l. compatible, and this is usually done.

Let us now consider the situation when a cmos signal on an l.s.i. chip goes from one logic stage to the next without leaving the chip. In this case, the stray capacitance which must be driven is only one tenth of a picofarad, and if the drive transistor is the smallest possible, i.e. 10,000 ohms resistance, the time constant, or delay, is only 1 nanosecond. From this we can deduce that it is not true that cmos is slow. Cmos signals across the chip have a high intrinsic speed, and so inter-chip circuitry should be serial, since this will reduce the amount of circuitry required for each function. (It is ridiculous for operations inside current microprocessor chips to be fully parallel. However, if someone made a serially operating microprocessor chip, probably nobody would buy it because, although its performance might be the same as its parallel competitors, the news would get out that the serial microprocessor contained very little hardware; there would be nothing for the salesmen to boast about.)

Note that if, by increasing the size of an output transistor by putting a number of square transistors in parallel, the output resistance of one bit of a 16-bit bus leaving the chip is brought down to 1000 ohms, so that the speed (rise time) is reduced to 20 nanoseconds, but sixteen such large transistors are needed to handle the sixteen-bit parallel word, using up valuable area on the integrated circuit surface. The same output data rate could be achieved by combining all 160 transistors in parallel to drive the sixteen bits serially down only one wire leaving the chip. In this case, assuming the same amount of chip area for the single drive transistor, a resistance of one sixteenth of 1,000 ohms could be achieved, leading to a bit rate of nearly 1,000 megabits down the single line. The point being made here is that *parallel working does not enhance speed if the circuits used are cmos*. On the other hand, a heavy price is paid when we go fully parallel – extra cost in wiring and extra pins in the i.c. package leading to extra failure (since the main cause of failure is the interconnections) and also far more failure due to pattern sensitivity with parallel data busses. Also, parallel working increases the pysical size of the resulting system, because size is largely dictated by number of interconnecting wires. It also forces us to use extremely complex, expensive test and debugging equipment including logic analysers with their awkward, octopus-like probe pods. By comparison, it is trivially easy to attach a single oscilloscope probe to a point where serial data is passing.

The Nub of computation

The heading of this section is purposedly inappropriate, to illustrate the problem at the very start. The 'computer science' discipline has come to think that its objective is 'computation', 'information processing' or some such. This is not true, or alternatively, if it is true, then 'computer science' is getting in the way of a much more important discipline, which is the application of technology to society's needs.

In our society or culture, certain historical necessities arise. It is usually thought that whether or not a certain development was a historical necessity is proven after the event by whether such a thing in fact came to pass. I think this is wrong. For instance, the wheel and axle was clearly a historical necessity in both Europe and the Americas, and the fact that the natives of the Americas never used the wheel and axle does not prove that it was not a historical necessity. More generally, we can see the extreme cases where a tribe or genus dies out because it evades a step which is a historical necessity.

Our society may well avoid historical necessity in the development of computer science, but that does not in my opinion negate the fact that what follows is a historical necessity.

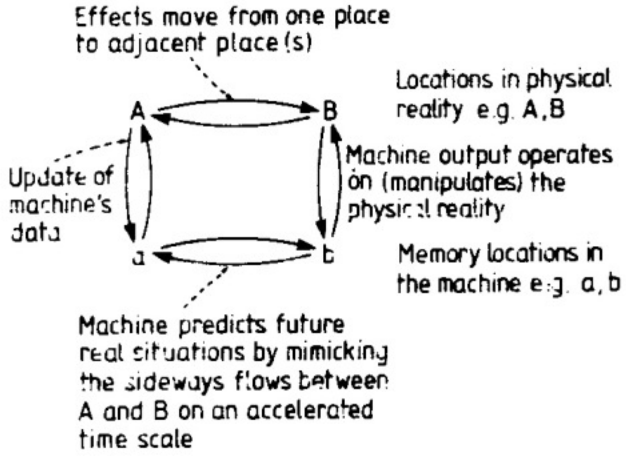
The proper objective for computer science or digital electronics is to apply technology to meeting human or sociological needs. (This is a quote from my 1969 *New Scientist* article[1]') I would probably limit the broad range of application to physical, not intellectual, needs.

Any physical situation which our technology can usefully be applied to will be a multi-dimensional array of values which need (a) analysis and (b) manipulation. Digital electronics won over analogue twenty years ago, I believe for ever, and so our machinery needs to contain a *digital analogue of reality*, and in fact always does so. One measure of the elegance of our machinery, and probably of its efficiency and simplicity is the ease with which the analogue in our machine maps onto the reality of which it is an analogue. The design of an elegant (and also one suspects efficient) machine requires of the designer knowledge of the physical reality which is the target of our machine; of the nature of data manipulation and computation; and of the physical nature of the machine.

Since the ideal seems to be a machine which can be regarded as a *physical analogue of reality*, and the closeness with which the machine's structure and information mimics the physical reality, the 'computer scientist' must have competence in all fields above.

The problem is that today, programmers, calling themselves computer scientists but having no competence in anything except the second (with perhaps a little competence in the first), think they can usefully contribute to the design and development of our future machines.

A second measure of the elegance of our machinery is the degree to which changes in the physical reality we are mimicking (or recording) in our machine can be easily effected in our machine. This is why a machine is very bad if it does not have content-addressable memory, and in fact it needs more than that. It needs processing capability *in situ* in the memory. This is because values or parameters in physical reality change *in situ*, influenced only by parameters which are physically nearby. This leads to the next requirement of a good machine, which is that since in physical reality there is not action at a distance but all interaction is local, our machine should have superior (or even only) interaction capability between values (vectors, scalars, etc.) which relate to physically close points in the physical reality. Further, it appears that the ability to effect interaction between values which relate to


Effects move from one place to adjacent place(s)
Locations in physical reality e.g. A,B
A
B
Update of machine's data
Machine output operates on (manipulates) the physical reality
a
b
Memory locations in the machine e.g. a, b
Machine predicts future real situations by mimicking the sideways flows between A and B on an accelerated time scale

points which are physically distant in reality may not be necessary *at all* in our machine, although this is pushing the point rather far.

A further requirement of our machine is that updating, or interaction, capability between points in physical reality and the related points in our machine where the digital analogue for that region of physical reality is stored, should be as efficent as possible.

The task of the machine architect is to exploit the potential of his technologies to meet these requirements. I believe I have done the best compromise in the Property 1 a invention[2], but it is not ideal, considering the above criteria. The above criteria are not merely a *post hoc* rationalization tending to show that my architectures are the best.

From the point of view of the above analysis, the reigning computer architecture theorists are doomed to failure. They (e.g. Petri nets) concentrate on the mechanism of computation in the machine, the bottom horizontal lines between a and b.

However, this has no value if the thinker does not bear in mind the dualism; that the arrows between a and b are a reflection of arrows between A and B; that computation only has value to the extent that it mimics events which occur in the real world. (This relates to my statement in the fourth paragraph above that the 'broad range of application', by which I mean the main field of application for our machine, and therefore the paradigm which should control their architecture, is directed towards practical rather than intellectual applications.

References

1. I. Catt, Dinosaur Among The Data?, *New Scientist*, 6 March, 1969
2. For Property 1 a invention, see "Wafer Scale Integration", *Wireless World*, July, 1981, pages 57-59.

Page 129 is blank.

WAVES IN SPACE

I would like to make one comment on T. C. Webb's letter in the August issue. My co-author Malcolm Davidson experimented with sending data (highs and lows for 1s and 0s) in both directions down a 1 kilometre length of twisted pair. He found that the losses experienced by the signals travelling in one direction were less when pulses were being sent in the other direction.

Conventional theory would say that during the time when one positive pulse passes through another going the other way in a transmission line, the i^2R losses drop to zero. The total current is zero during this time.

Ivor Catt
St Albans
Herts

In his letter, published in *WW* November, 1983, W. M. Dalton hit a nasty land-mine that I first noticed some years ago. Let me first quote the moment when he hits it.

> "Let us start from known facts. (1) Light is an electromagnetic phenomenon: demonstrated by Faraday and Kerr. (2) Light is not a static problem: it is ocillatory (Hertz). (3) The electric and magnetic fields are at right-angles and *always* 90 degrees out of phase. Some recent textbooks show these in-phase – an unparadonable error."

I am anxious that Mr Dalton expands on why this error is unpardonable, and what disasters this error might lead us into.

First let me list some non-recent textbooks which show these in-phase.

G. W. Carter, Professor of Electrical Engineering in the University of Leeds, in his book The Electromagnetic Field in its Engineering Aspects, (Longman 1954) draws the B and E fields in-phase on page 271. Significantly, although he emphasises that E and B are at right angles (page 274) he never seems to say in the test that B and E are in phase.

A. F. Kip, Professor of Physics, University of California, Berkeley, in his book Fundamentals of Electricity and Magnetism, (McGraw-Hill 1962) draws the H and E fields in-phase on page 322. On that same page the text says that the two fields are perpendicular to each other, but does not state that they are in-phase. Again significantly, I cannot find mention in the text that they are in-phase.

O. Heaviside F.R.S., in his book Electromagnetic Theory Vol 3, 1912, in art. 452, page 4, wrote

> "The General Plane Wave . . . the slab may be of any depth and any strength, and there may be any number of slabs by side behaving in the same way, all moving along independently and unchanged. So E=μvH expresses the general solitary wave, where, at a given moment, E may be an arbitary function of x . . "

'Replace μv by √μ/ϵ – I. Catt]

Whereas some books (Carter and Kip) vaguely indicate that E and H are in-phase, other books seem to fail to discuss relative phase at all see for example Gullwick 1959, Bewley 1933. The trap was nicely set for Dalton, and he has my sympathy.

Now let us turn to my article in *Wireless World*, July 1979, entitled The Heaviside Signal.

> "We have shown that the passage of a TEM wave and all the mathmatics that has mushroomed around it does not rely on a causality relationship (or interchange) between the electric and magnetic field. Rather, they are co-existent, co-substantial, co-eternal."

In that article I compare and contrast two mutually contradictory versions of the transverse electromagnetic wave. I believe that the full realisation that E and H are in-phase deals a death-blow to one of those versions, the rolling wave, and leaves the other, the Heviside signal, the victor.

Because the differential of sin is cos and the differential of cos is minus sin, half-witted mathematicians have invaded the physics of the TEM wave and imposed a spurious story that E causes H causes E. Since sin, cos and −sin are 90 degrees out of phase, part of their phoney baggage is to imply that E and H are 90 degrees out of phase. (See my article in *WW* in March 1980.) Because the sine wave is amenable to mathematical high jinks, another part of their baggage is to imply that a TEM wave is sinusoidal. It's time we cleaned the claptrap out of electromagnetic theory.

Ivor Catt
St. Albans
Hertfordshire

WIRELESS WORLD FEBRUARY 1984

TEM-WAVE PHYSICS

Lest the fierceness of Mr Catt's response to Mr Dalton (February 1984 issue) obscures what he said, could I diplomatically support all that was contained in his letter while at the same time describe a situation where E and H are 90 out of phase. This should please Mr Dalton.

But first let me remind Mr Dalton that the opposite of "static" is "dynamic" and not "oscillatory". The last is just one of many modes of motion which need not even be periodic. This is particularly important because the example I propose to give for E and H being 90° out of phase is static. This should please Mr Catt.

Starting from Maxwell's equations it is easy to derive equations of wave propagation for E and H, the solutions of which are

$$E=f(x-ct)$$

and

$$H=\frac{1}{c\mu}f(x-ct)$$

where f can be any function, not just sinusoidal or even periodic e.g. a digital (level) change, a single pulse – square or any other shape.

The variation (f) of H matches precisely the variation of E (also f) whatever whatever f happens to be. There is no delay between E and H or, in the case of f being sinusoidal, no phase difference. As Mr Catt states there is no causality between E and H. However, and this may be part of the origin of Mr Dalton's error, there is a rotation of 90° from E and H which is right handed about (not along) the direction of propagation. Thus if f is sinusoidal E and H are in phase but at right angles to each other in space, not time.

If the equations above are divided one into the other then

$$\frac{E}{H}=\frac{1}{c\mu}=\sqrt{\frac{\mu}{\epsilon}}=Z_o$$

where Z_o is the wave impedance of free space (about 375 ohms) which is independent of f.

If E and H were sinusoidal and 90° out of phase as Mr Dalton suggests, then Z_o would be the tangent i.e. from minus infinity to plus infinity. This would make it difficult for a wave to propagate. At the very least it would imply causality if one knew which occurred first and at worst would mean changing the title of your illustrious magazine.

This brings me to the example of E and H being out of phase and possibly the other half of Mr Dalton's confusion.

Suppose that a sinusoidal wave described by

$$E_1=E_0\sin\left\{\frac{2\pi}{\lambda}(x-ct)\right\}$$

has superimposed on it an equal wave but travelling in the opposite direction, say by reflection, described by

$$E_2=E_0\sin\left\{\frac{2\pi}{\lambda}(x+ct)\right\}$$

Some trigonometry reduces the sum of these to

$$E_1+E_2=2E_0\sin\frac{2\pi x}{\lambda}\cos\frac{2\pi ct}{\lambda}$$

$$\text{or } 2E_0\sin\frac{2\pi x}{\lambda}\cos\omega$$

$$\text{Similarly}\quad H_1+H_2=-2H_0\cos\frac{2\pi x}{\lambda}\sin\omega t$$

This results in the well-known standing wave where the nodes of H correspond with the peaks of E and vice versa i.e. 90° out of phase. When E is a maximum, H is zero everywhere. Then H grows and E decreases until it is a maximum and E is zero, and so on cyclically. Thus the standing wave has all the appearance of transforming itself from an entirely electric form to an entirely magnetic one and vice versa. But it is just an illusion, for as Mr Catt states, there is no causality between E and H for a single wave, still less is there any between two in which we only observe their interference pattern.

This, I hope, explains the source of Mr Dalton's confusion.

Finally I would like to disagree with Mr Catt (only in a very minor way) concerning his references. Carter in his book "The Electromagnetic Field in its Engineering Aspects" pages 266 to 276 is quite specific about there being a delay (or phase difference in the sinusoidal case) between E and H, both in his diagrams and text, and of which the above is, I trust, an accurate paraphrase. They correspond, though in different words, with the views expressed by Mr Catt.

E. O. Richards
Hitchin
Herts

PS: For those who share Mr Catt's disgust with sin and cos I commend a closer look at Walsh functions, an introduction to which appeared in these pages in January 1982. An excellent book on the subject is "Walsh Functions and the Engineering Applications".

WIRELESS WORLD APRIL 1984

C.A.H.