RC 20057 (5/10/95)
Engineering and Technology

IBM Research Report

Empirical Relations and Design

R. W. Keyes
IBM Research Division
T.J. Watson Research Center
Yorktown Heights, NY 10598

IBM Research Division
Almaden . T.J. Watson . Tokyo . Zurich

Empirical relations and design

R. W. Keyes[+]

IBM Research Division

Yorktown Heights, NY 10598

Abstract

Large, complex circuits that can perform the logical operations required in a computer have been designed in many places for over three decades. Experience has shown that complex interconnection patterns are needed to implement the logic. The physical constraints of a wiring technology must allow designers and design tools a large degree of choice among possible interconnection patterns. Certain generalizations concerning quantitative aspects of the interconnections have been recognized. Several of the empirical observations are reviewed and related where possible to models and used to illustrate trends. These generalizations give early guidance to design, and also may be regarded as targets against which designs may be measured. [+] Research Staff Member Emeritus

Empirical relations and design

R. W. Keyes

Introduction

The introduction of the representation of information in computing machinery by electrical signals several decades ago made large machines possible. Electrical pulses can rapidly carry information through long distances from one part of a machine to another without the friction and wear associated with mechanical parts. The earliest electrical computers used relays to perform logic operations. They proved their worth by greatly speeding the performance of a variety of calculations and stimulated a search for faster, more reliable components. These were found, first in vacuum tubes, which operated much faster than relays, and soon after in transistors. All of these devices have as a common ingredient an extensive network of connecting wires.

The use of transistors has led to the extremely low cost that allows large computers, large in the sense that they contain many millions of components, to be made widely available. The low cost has been achieved through integration, putting a great many components on single piece of silicon. Manufacturing operations are performed on large silicon wafers that contain many components, and the cost of an operation is divided among all of the components. The wafers are eventually divided into much smaller chips. The silicon wafers that are handled in the fabrication of 4 megabit semiconductor memory chips may contain 200 chips, a total of about 10^9 bits of information on one wafer. The cost of producing the wafer is shared among all of these bits. The measure of integration is the number of components on a single chip. An essential element of integration is miniaturization, making everything on the chip smaller. Integration has also been advanced by increases in the size of chips. The thrust of automatic information-processing hardware technology since the introduction of the transistor has been one of making transistors and everything associated with them smaller.

Besides low cost, miniaturization and integration have led to higher speed, lower power, and higher reliability. Improvements in all of these measures have continued to this day and have made it possible to construct machines containing tens of millions of components. Two of these properties, low power per component and high reliability, are essential to the existence of very large systems. High speed is a welcome fallout of miniaturization and integration.

There is another source of progress in integration: improvements in design that allow devices to occupy less space without a reduction in dimensions. [1] We may call this contribution "compaction." For example, a capacitor built as two conducting planes separated by a thin insulator on a surface occupies much more area than a trench capacitor, in which a trench is etched into a substrate and a capacitor built from layers formed on the walls of the trench.

The number of devices in a useful computational system ranges from a few thousand upwards. Figure 1 shows how the number of logic gates in the most powerful computers has increased with time. The amount of information storage that is available to these powerful systems has grown even faster. An enormous amount of communication between the many parts of a system is necessary.

The cost factors that drive miniaturization, integration, and high density of components on chips apply to the substrates on which the chips are mounted, called modules or boards or cards, in multi-chip systems. The colloquial reference to area on chips and on chip substrates as "real estate" is essentially a sign of the value of surface area in semiconductor technology. The cost of processing depends most strongly on the area processed. In addition, placing chips close together reduces the time for signals to travel through a system and assists system speed. Further, space in buildings is valuable in itself, and the cost of the volume occupied by electronic equipment is an incentive for the pursuit of miniaturization and dense packing. Additional motivation towards high component density is provided by applications for mobile and portable systems.

At present a single small chip of silicon can contain several millions of transistors and tens of meters of wire. The design and assembly of so many individual components into a system that functions as a single entity is a very complex undertaking. The intricate relations between large numbers of parts that have proved to be necessary in practical systems become difficult for the human mind to contain. The difficulty is not a fundamental one, but stems from the very large number of facts that must be simultaneously brought together. It is not surprising that at an early date in the history of computers their builders resorted to automated machinery to store the vast assembly of facts and their interrelations. It was then a natural step to prepare programs to use the stored information to aid in the design and manufacture of information processing equipment. Automated design is a necessary part of the design of contemporary systems.

The design of a product requires that estimates of many of its physical characteristics be available at an early stage so that such things as feasibility, performance, and manufacturability can be assessed. Since interconnections play a major role in determining the physical character of a computing system their properties are intimately involved in any early estimates. The aim of this paper is to discuss the nature of connections and generalizations useful in estimating their characteristics.

Communication in computing

The might of modern computing systems has been made possible by technology that allows many components to be brought to bear on the performance of a single task. Besides the low cost that allows the large number of electronic devices to be provided, causing them to work together to achieve some desired ends requires a large amount of communication between them. As signals are represented electrically and the basic component, the transistor, is an electrical device, this communication is carried out by a dense network of electrical connections, wires.

The devices in computers are organized into logic circuits made of a few transistors that can perform an elementary logic function such as an AND or NOR. The circuits are interconnected to implement more complex functions. As all electrical components must be connected to other components to be part of a system, the number of connections is of the same order as the number of components. Thus, the number of connections in a system may be counted in the millions. The late Robert N. Noyce, one of the inventors of the integrated circuit, has said of the formidable task of physically providing these many connections "The integrated circuit is the component industry's solution to the interconnection problems" [2]. The conductors that carry electrical signals from one place to another on integrated circuits are manufactured by material deposition and removal processes on large wafers. Each fabrication step contributes to making a very large number of wires.

In addition to the assemblage of wiring and connectors used for internal communication, computers must provide for a large amount of communication with other devices, for example, tape and disk storage, workstations, printers and other human interfaces, and other computers. These external devices are physically at various distances from the central computer.

The interconnections have a major role in shaping the physical nature of computers. Provision of communication between devices on a chip is part of the chip manufacturing process, while communication between chips is part of the discipline known as packaging. [3] Wire on chips carries signals from transistor to transistor. The wire occupies an important fraction of the chip area. Chips are provided with input-output terminals that allow electrical connection to the wiring on the substrate to which the chips are attached. The connectors that join the two have proved difficult to miniaturize; connectors have not been affected in a significant way by the notable advances in electronic technology, such as fine lithography, well-controlled thin-film growth, and anisotropic etching. The size of the connectors is limited by the precision with which they can be located with respect to other parts during assembly of the system and by the mechanical forces to which they are subject. Their dimensions match the coarser of the parts being joined. One of the most serious limits on the speed at which systems of many chips can operate is the delays encountered by signals traversing the chip substrate, and minimizing these delays is a principal objective of packaging technology.

The internal operation and communication is controlled and synchronized by a clock in a cycle, a time period in which short sequences of logic and internal communication must be completed. Great effort is devoted to avoid unusually long sequences that would require a long cycle time during which much of the logic would be unused. The length of the cycle plays a major role in determining the performance of computers.

The communication with external devices is a form of telecommunication. In contrast to internal communication significant latency, long delays while signals pass from one place to another, are inherent in telecommunication links. Modern computers have cycle times on the order of 10 ns, during which time a signal travelling with the velocity of light moves only 3 meters. The latency involved in external communication may amount to many thousands of computer cycles. Since the transit delay is inevitable, means to deal with latency must be incorporated in communication with external devices. The processor can use the cycles for some other purpose, and there is little penalty associated with additional delays for signal processing. Thus telecommunication frequently involves multiplexing many signals onto one channel, coding and decoding for error detection and correction, data compression, and queuing.

The great success of modern telecommunication technology tempts one to propose extending it to the interior of the CPU, the central processing unit of a computer. However, the requirements of the two forms of communication are very different. In contrast to the latency in telecommunication, effort is devoted to saving picoseconds within the central computer. Telecommunication is oriented to handling concentrated point-to-point traffic on a single physical channel. High bandwidth is of primary importance, and investment in expensive components that can achieve it is justified. Traffic within the CPU is carried by many wires, each dedicated to connecting a specific source to specific destinations, and fan-out and fan-in are usually involved. Most wires will carry only a single bit in each CPU cycle; bandwidth is ordinarily not strained.

Value also accrues to carrying information through long distances in telecommunication. Indeed, the product of bit rate and distance of a link is regarded as a figure of merit. The lowest possible attenuation in the channel is sought. Values less than 1 dB/km are not uncommon. On the other hand, the wires within a computer are short, lengths are measured in microns rather than km, and attenuations exceeding 10^4 dB/km are acceptable.

The lack of opportunity for error detection and correction means that very low raw bit error rates must be maintained in a computer. A chip that sends a bit through each of 100 off-chip terminals per 10 ns cycle will send 10^{10} bits per second. Long periods of error-free operation will need bit error rates of 10^{-17} or 10^{-18}. A telecommunication channel with error correction can tolerate raw bit error rates of 10^{-9}.

The economic and the design considerations are also rather different in the two cases. The large number of components in the CPU calls for very low cost, low power, and high reliability. The maximization of bandwidth in telecommunication systems requires few but expensive components. Figure 2 summarizes these two regimes of information transmission technology.

Empirical rules

The interconnection schemes on chips and substrates are not simple, and their specification is one of the most complex tasks of design automation. It has not proved possible to perform the complicated processing of series of logic signals that is needed with regular arrays of connections. Although the components are usually placed in a well-ordered array, the embedding of function in the array requires patterns with little order in the arrangement of wires.

In fact, the lack of any pervasive regularity in the wiring pattern makes a probabilistic view of the wiring useful. Averages and distribution functions can provide useful descriptions of the wiring on a chip or other substrate. For example, the distribution of wire lengths might be given, or the average number of off chip connections. Estimates of these kinds of physical characteristics are needed at an early stage in the conception and design of a machine, before all of the detailed logic that will be implemented has been defined. Analysis of wirability depends on the lengths of wires and the area that will be available for them on chips and substrates and on the number of connections between them. Predictions of performance depend on the length of wires and the interfaces that will be encountered.

The desired early estimates of quantities useful in designing parts are offered by certain empirical relations that, while having only a fragile foundation in mathematical analysis, are well established by observation of the characteristics of chips and packages that have successfully provided the functions demanded by complex systems. The large number of wires on a part and the lack of knowledge of exactly what space they will occupy is reminiscent of problems of statistical mechanics. Thus generalizations about wiring deal with averages over large numbers of wires and ensembles of many parts.

The most used of the empirical relations concerns the number of connections that must be made to a part to incorporate it into a large system. Call the number of elementary logic circuits, collections of a few transistors that perform an elementary logic operation, in a partition of the system N. It is found that the number of connections that must be made to the partition to usefully employ most of the N gates is closely related to N. The number of signal connections, k_C, has the form [4, 5]

$$k_C = bN^s \tag{1}$$

Eq. (1) is known as Rent's rule. It is only approximate, there is scatter in the observed relation between k_C and N and a lack of unanimity among authors as to the best values for b and s. Representative numbers for chips that are part of a large multichip central processing unit are given by Blodgett [5], who sets $b = 2.5$ and $s = 0.61$. The values depend on the application in question, and are different, for example, for microprocessor chips and for memory chips.

Quantitative physical aspects of chips and packages depend on Rent's rule, Eq. (1), and one is led to ask how hard and fast such a “rule” is. In practice, various compromises that affect the value of k_C are made. The number of connections that are needed to effectively utilize the circuits on a chip varies from chip to chip. Too many connections will leave some unused on many chips. Too few connections means that a smaller fraction of the circuits on some chips will be usable and more chips may be needed to build a system. The optimum design depends on the relative costs of circuits and connections and the impact of the choice on performance.

Another compromise that could be made is the use of multiplexing. A chip could send several signals that are intended for the same destination chip sequentially through a connection, and they could be sorted out on the receiving chip. The additional circuitry and delays involved in doing this are not acceptable in the CPU of a high-performance system. The duplication of logic functions that generate multiple outputs is a more relevant illustration. For example, consider that both the result of a logic operation and its inverse are wanted in various places. Both could be generated by the logic circuit at the point of origin and transmitted, or a single output could be transmitted and its inverse formed at each of the receiving chips. The former uses more chip connections, the latter more circuit components. An engineering compromise must be made.

Rent's rule, based on a large body of empirical evidence is widely accepted as applying to physical partitions of a machine, that is, to a chip, or to a card or module or other substrate that contains some number of chips and that is to be connected to similar entities. Attempts have been also been made to apply it to conceptual partitions, defined in terms of imaginary lines drawn within a physical partition, but these have had no impact on design.

Another relation reflects a property of the wiring that may be called “nonlocality”. This means that if one examines some portion of a system, the greater the number of components in it, the greater the length of the average wire in it, measured in units of the component to component distance, known as the component “pitch” [6]. In other words, the number of components that can be reached by a wire of average length increases as the number of components in the system or part of it examined increases.

Quantitatively, in placing M circuits on a chip or M chips on a substrate let the average length of a wire, measured in terms of the pitch, the circuit-to-circuit or chip-to-chip distance, be m_1. Distances are measured along the perpendicular axes, and are often called “Manhattan” distances. For example, the distance between opposite corners of a square of side S is $2S$. The parameter m_1 increases with the number of chips to be packaged (the nonlocality), and the formula

$$m_1 = J(1 + 0.1\, ln\, M)M^{1/6} \qquad (2)$$

has been suggested [7]. This relation is well represented by the simpler $m_1 = J\,M^{1/4}$ for M over a large range. The value of the constant J is found to lie in the range 0.6 to 1.

Wire occupies a substantial area on chips and on packages and the need to devote area to wiring limits the packing density of components and increases the distance between them. The increase in average length of wires means that more space on a chip must be allotted to space for wiring. The increase can be quantified for masterslice chips by a probabilistic model. [8] Such chips (and parts of a system that contain a large number of chips, such as modules, and other entities that carry wire) are designed to take advantage of commonality among the parts, even though there are many different types of them. Much of the physical design is carried out without knowledge of the wiring patterns that they will eventually carry. Space for wiring is provided in the form of channels, in which wire segments can be placed as needed. Alternate layers of wiring on chips and substrates carry channels in mutually perpendicular directions. Provision for vias, places where connections between wire layers may be made, is necessary.

There are advantages to this approach. When the difference between parts lies only in their wiring patterns many of the same manufacturing tools can be used for different parts. Design is simpler, faster, and less expensive than that required to handle each part entirely independently. New, changed wire patterns can be designed and made available to the system builder rapidly. The masterimage is the method of choice for large machines with a variety of chip and module types, a small volume of each of which will be produced. There is, however, a penalty: custom design, a unique design throughout for each part, allows shorter average wire lengths and a smaller chip to be achieved. The added cost of custom design of a single part, such as a microprocessor chip, that is to be produced in very large volumes can be justified by gains in performance and productivity.

Many different kinds of chips personalized by wiring patterns may be used in the construction of large systems. Particular functions are implemented by choosing the occupation of the wire channels at a late stage. The chip is regarded as divided into N cells, each containing a logic gate, and the wiring space requirement is expressed as a number of channels traversing each cell that must be available for wires in the final stage. Statistical studies show that the number of channels or wire tracks needed to offer a high probability of success in placing interconnections depends mainly on the product $k_C m_1$, where m_1 is the average length of a wire and k_C is the average number of wires per cell. [7, 8] The product $k_C m_1$ is the length of wire per cell measured in cell pitches. A great many cases, including arrays of various sizes, were simulated in the work of reference 8 with results as summarized in Fig. 3.

A dashed line in Fig. 3 represents the relation *Wiring tracks* = $2k_C m_1$. The line thus represents points for which the total length of wiring tracks is twice the actual length of wire. The fraction of the wire channels that are actually occupied by wire is called the "utilization," and is 0.5 on the line. Utilization 0.5 is rarely achieved in this study.

The average wire length and the number of circuits play major parts in determining the physical configuration of chips. The distribution of wire lengths is also of interest, but it has received less attention. Although few examples are available, these suggest that the number of wires of length L decreases as $L^{-5/3}$, changing to a rapidly decreasing exponential for wires longer than some cutoff length. [9, 10] This view was adopted in reference 8. Cases can be found, however, that do not fit the simple rule.

The long wires represented by the exponential tail of the wire length distribution are often troublesome. Over 3% of the wires are longer than the cutoff point in the largest example in reference 9. The long wires have enough resistance to seriously degrade signals and demand special attention. Some might be eliminated by manual intervention. One approach to dealing with very long wires is to add layers with especially wide conductors. The use of repeaters to restore signals in long wires has been proposed.

The basis for empirical guidelines is a large number of designs that already exist. Quite a lot of work has usually gone into the designs, presumably satisfactorily. The rules are derived from designs for which the preliminary work has been done. This involves partitioning, the assignment of functions to particular chips or other levels of the package hierarchy, and placement, locating functions, chips, etc, that are highly interconnected in the neighborhood of one another. Another view of the empirical rules would be as a framework in which to compare the partitioning and placement with that of other good designs.

Wire model

These ideas concerning wire length can be used in a model of a chip or other wire substrate. It will be helpful to refer to a specific example to clarify the concepts involved in describing a model of wiring. The IBM Thermal Conduction Module (TCM) is chosen for this purpose because a readable and accessible description of it with excellent illustrations has been published by Blodgett [5], even though it represents an earlier stage of technology. The module is a 9 cm × 9 cm body of ceramic that can contain 100 chips in a 10 × 10 array. The chip pitch is 0.9 cm. There are 16 layers of signal wire in the ceramic, and the width of a wire channel is 0.05 cm. There is provision for 120 connections to each chip, 96 of which are for signals. The others deliver electrical power to the chip.

To construct a model of wiring, again consider M components, used here to mean devices or logic circuits or chips, or even multi-chip packages, that are interconnected by wiring on a substrate with area a per component. The component separation or lattice spacing, the pitch, is $a^{1/2}$. Let the average length of the wire channels that must be associated with each component be $ma^{1/2}$, that is, m is the length of the tracks needed to accommodate the wire. In the notation of the example above from reference 8 $m = k_C m_1/u$ As described above and in Fig. 3, because of the probabilistic and pseudo-random character of the wiring in computers only 40% to 60% of the available channel capacity is used (see illustration in Blodgett [5]). If the width of a wire channel is W, then the wire area per component is $m\,a^{1/2}W$.

The available area must equal the area needed to contain the wiring and the component [11, 12]. If there are K layers of wire then

$$Ka = ma^{1/2}W + a_C \tag{3}$$

Here a_C is the area needed for the component being wired, and is zero if the wires and the components are in different layers. In all cases a must be at least as large as a_C. It is often the case that a_C is small compared to a, when the quadratic equation (3) can be expanded to give

$$a = (mW/K)^2 + 2a_C/K \tag{4}$$

In many cases the approximations $a_C = 0$ and $a = (mW/K)^2$ are adequate. It should also be noted that as described above, Eq. (2), the average wire length m will depend on the number of components on the substrate.

To apply the model to the TCM example, assume, reasonably accurately, that there is one wire for every two chip terminals attached to the substrate. Then the the number of wires per chip, k_C, is half the number of terminals and, in Eq. (3), $m = m_1 k_C/u$. The wire channel utilization, u, will be taken to be 0.5. From Eq. (2) with $J = 0.9$, $m_1 = 2.85$ and

$m = 96 \times 2.85/2 \times 0.5 = 273$. The value of m is large because of the large number of terminals on a chip; there are many wires per chip. Wires can pass under chips, so $a_C = 0$ in Eq. (4) and $a = 0.73$ cm^2, to be compared with the actual 0.81 cm^2. The total length of wire channel needed is 100×273×0.9cm = 246 m, versus 260 m provided by the TCM. A typical TCM contained 130 m of physical wire, a utilization u of 0.5. The empirical rules give a reasonable first approximation to the package.

The wire channel density, the wire capacity per unit area, is a figure of merit for an interconnection technology. An area A filled with wire channels of pitch W contains a length of wire A/W. If there are K layers of such wire then the total length of wire is (KA/W) and the wire density is $D = K/W$. This figure is 16/0.05 cm = 320 cm/cm^2 for the TCM.

As levels of integration increased faster than the sizes of large systems in recent decades it became possible to package enough chips to constitute an entire multichip CPU on one module, instead of in the multilevel packaging hierarchies of the 1970's. The number of long module-to-module delays in a cycle was reduced. Fewer package components were needed, helping to contain costs. It even became possible to form a moderate sized computer on a single chip and microprocessors emerged.

The great advances in level of integration on chips in semiconductor technology were accomplished with a much more modest increase in the area of chips. However, the increased integration required an increased number of terminals per chip, The density of chip terminals increased and demanded continuing increases in the wire density of the substrates on which the chips are mounted. The dense wiring network must span a large area in a many-chip module. The packaging of highly integrated chips is limited by, among other things, the ability of technology to supply a dense network of signal wiring over a large area.

Reduction of W to reduce the area per logic circuit on a chip or module is limited by the difficulty of fabricating narrow lines, which is measured by yield, the fraction of substrates on which all of the wiring is good. The size of all of the powerful components of large machines, such as chips and modules, is limited by yield, the ability of technology to produce devices and wires completely free of defects over a large area.

Measuring connectivity

The complexity of wiring makes designing it a very hard task. Designers of logic circuitry must be able to choose one from among a great many possible interconnection patterns. [13] The number of patterns permitted within the constraints imposed by the wiring technology must be very large, and is a measure of its usefulness, its "connectivity." A tractable approach to a measure of connectivity is to ask: how many patterns of pairwise connection of all of the terminals to be wired are possible?

Given H ($H >> 1$) terminals on a substrate, the number of ways that they can be connected in pairs is given by

$$\Gamma_0 = H!/2^{H/2}(H/2)! \tag{5}$$

or, for large H

$$\log\ \Gamma_0 = (H/2)(\log H - 1) \tag{6}$$

Applying this, for example to the TCM, with 9600 chip terminals, it is found that $\Gamma_0 = \exp(3.9 \times 10^4)$.

In practice, quite a lot of wire may be required to provide this full flexibility. Let A be the area in which the H terminals are contained. The average "Manhattan" length of all possible pairwise connections in a square $H^{1/2}$ by $H^{1/2}$ array of terminals is $l_0 = (2/3)H^{1/2}$ terminal pitches. In the case of the TCM $l_0 = 65$. The average terminal pitch is $(A/H)^{1/2} = 0.092$ cm and the physical wire length would be $L_{av} = l_0(A/H)^{1/2} = 6.0$ cm. The wire density needed, with the utilization u as defined above, would be

$$D_0 = (H/2)L_{av}/uA \tag{7}$$

This works out to $D_0 = 720$ cm/cm^2 in the TCM case. As the actual density is 320 cm^{-1}, the full versatility suggested by Eqs. (6) is not available.

A package of lesser wiring density than Eq. (7) can be expected to admit fewer two terminal configurations. To estimate connectivity with limited wire density, observe that a smaller wire density can support only a proportionally shorter average connection length (Eq. (7), keeping the terminal density *(H/A)* constant). Then, if the wire density is $D < D_0$, the corresponding average wire length is $L'_{av} = (2uA/H)D$, as in Eq. (7). Let H' be is the number of terminals between which all possible connections can be supported by L'_{av} and note that, L'_{av} is also equal to $(2/3)H'^{1/2}(A/H)^{1/2}$. Equating these values for L'_{av} gives

$$H' = 9u^2AD^2/H \tag{8}$$

The number of possible connection patterns in the group of H' is, as in Eq.(6), $\log \Gamma' = (H'/2)(\log H' - 1)$. Now if the collection of H terminals is divided into (H/H') parts, the total number of connection patterns is $\log \Gamma = (H/H')\log \Gamma'$. Substituting the value of H' from Eq.(8) gives

$$\log \Gamma = (H/2)\log(9u^2AD^2/He) \tag{9}$$

Again using the example of the TCM, it is found that $(H/H') = 5$ and that the possible interconnection patterns are reduced to $\Gamma = \exp(3.1 \times 10^4)$. The reason, of course, that systems can be wired with average lengths much less than those associated with a random array is that automated design programs place things that are to be connected close to one another.

Figure 4 shows for chip substrates how the number of possible connection patterns depends on the logic content of a component. The figure plots $\log \Gamma$ against the number of logic gates packaged on a number of multichip substrates. Very large numbers are encountered. Those for chips are even larger, so large that one is forced to ask whether the power of the computers available to designers can keep up with the demands of the complicated tasks of Design Automation. Figure 5 addresses this question. The growth of $\log \Gamma$ with time for chips is shown. The numbers for chips are much less accurate than those for substrates because most connections involve fan-in and fan-out and questions may be raised as to what should be counted as a "terminal." These points in Fig. 5 can be interpreted as a guide to what can be achieved with purposeful design.

Measures of computer performance are also shown in Fig. 5 for comparison with the growth of $\log \Gamma$. The increase in the power of large mainframe computers lagged the growth in connectivity in recent years. The advent of various types of large multiprocessor computers enabled the design tools to keep up. There are reasons to believe that the pace of increase in the power of massively parallel systems will be even faster in the next decade and exceed a teraflop by the end of the decade, as illustrated by the dashed line. [14] Automated design will apparently be able to meet the needs.

Scaling Wire

Ho, et al, have advanced an interesting view of wire yield. [15] Their reasonable assumption is that the size of a defect that can cause a loss of function is proportional to the dimensions on the part, to the wire pitch, W, in the case of wire. If the concentration of defects per unit area in the size range, w, $w + dw$ is $C(w)$, and a defect with size greater than w_0, some fraction of W, is required to cause a failure, then the average number of fatal defects in an area A is

$$N_D = A \int_{w_0}^{\infty} C(w) dw \tag{10}$$

Now it is suggested that $C(w)$ is proportional to $1/w^3$. [15] Then $N_D \propto A/w_0{}^2$. The yield, the fraction of the parts produced that have no defect, is a function of N_D. Since A and $w_0{}^2$ are both proportional to W^2, it follows that changing all linear dimensions of a structure by the same factor does not change N_D and the yield. For example, it is implied that wire yields for chips may be scaled to the coarser wiring on substrates.

In this scaling view, multiplying all dimensions by some factor β>1 will decrease the amount of wire channel per unit of area, given as above by $K/W \equiv D$, by the factor β. The area to be wired would be increased by β^2 in the same scaling. Thus the scaling of wire capacity D, with the area that can be wired at that density, A, is defined by

$$AD^2 = \text{constant} \tag{11}$$

There is an appealing logic to Eq. (11) as a figure of merit. The difficulty of providing wire increases with both the density of wire and the area to be wired. Equation (11) describes the combination of these two technological limitations into a single measure. The implication is that technology develops in such a way that the product in Eq. (11) is independent of the size of substrate wired. In particular, it is the same for both packages and chips.

A different way to look at AD^2, Eq. (11), is as an aspect ratio of the wire channel. The length of channel in area A is AD and the length to width ratio is AD/W. For wire in a single layer $D=1/W$ and Eq. (11) is retrieved.

The framework for examining the state of the art of parts containing much wiring in the perspective of Eq. (11) is presented in Fig. 6. Here the logarithm of the density of wiring channels, measured as cm of channel per cm^2 of substrate, is the abscissa, and the logarithm of the area to be wired is the ordinate. Some examples of multichip modules built around 1990 are plotted. The same area-wire channel density space can be used to represent wiring on chips, and a few points derived from chips of the same era are also shown. The dashed line represents Eq. (11), and seems to lend credibility to the scaling notion.

The argument can be used at any particular point in time. In the course of time, however, improvements in cleanliness and processing methods reduce $C(w)$. This steady reduction of defect concentrations has made increasingly complex chips and packages possible, so that the "constant" in Eq. (11) increases with time. [13] Thus Eq. (11) should be written with a function of time on the right hand side:

$$AD^2 = G(t) \tag{12}$$

The implication of Eq. (12) is that a series of lines parallel to the one shown in Fig. 6 could be drawn, each representing the state of technology at a particular time.

Accepting that AD^2 is approximately constant at any particular time allows the function $G(t)$ to be determined. In this vein Fig. 7 plots the product AD^2 for the wiring on modules and chips as a function of year. For example, for the TCM, introduced in 1981, $D = 320$ cm/cm^2 and $A = 81$ cm^2, giving $AD^2 = 8.3 \times 10^6$. The line corresponds to the equation

$$G(t) = 240 \exp(0.432(Year - 1960)) \tag{13}$$

$G(t)$ roughly doubles every two years. Figure 7 and Eq. (20) characterize the trend of technology and specific parts may deviate from it by an order of magnitude. It is seen that in terms of the measure used wiring on chips and on substrates have progressed at comparable rates.

The future

Doubts are frequently expressed about the future of miniaturization and integration. Questions such as: How long can the technologies that have enabled rapid miniaturization persist? How small can a device be and still work properly? constantly arise. Confidence in continued progress at the historical rate through the next decade is supported by technological developments now in laboratories and has been expressed by the Semiconductor Industry Association through a quantitative "Roadmap" to the future. It is projected, for example, that some chips in year 2007 will have an area of 10 cm^2 and have 3600 connections with a substrate.

Regarding wiring, delays and attenuation caused by wire resistance as wires are miniaturized is of particular concern. The RC delay in wire varies as $t_{RC} = rcL_{av}^2$, where r and c are the resistance and capacitance of a wire per unit length. In the simplest view, when all dimensions are reduced by some factor β, the average length varies as β and r varies inversely as the wire cross section, as $1/\beta^2$. Capacitance/length c remains constant when dimensions are scaled. Then t_{RC} is also constant. Against this optimistic view it can be noted that the increase in chip size and content causes wires to be relatively longer, they will not simply scale as β, and that the other relevant time, the transit time of a signal along the wire decreases when the length decreases. The RC problem will be controlled for some time in the future, as it has been in the past, by making wires thicker and by using more layers to allow wider wires rather than simply scaling wire dimensions. Reduction of c through innovation in dielectric materials will also assist.

Conclusion

A number of generalizations about parameters useful in the design of chips and packages for computers can be drawn from the long history of successful designs that have been produced. The generalizations serve as a guide to the physical nature of a system in the early stages of design. They can be applied to aid in forecasting the nature of future systems.

REFERENCES

1. Moore, G. E., "Progress in digital electronics," 1975 IEDM Technical Digest, 11-13;

2. Noyce, R. N., "Large Scale Integration: What is Yet to Come?", Science, vol. 195, No. 4283, pp. 1102-1106 (18 March 1977).

3. Tummala, R.R., and Rymaszewski, E. J. Microelectronics Packaging Handbook (Van Nostrand-Reinhold, New York, 1989).

4. Landman, B. S., and Russo, R. L. "On a pin versus block relationship for partitions of logic graphs", IEEE Trans. Comput., vol. SC-13, pp. 1469-1479 (1971).

5. Blodgett, A. J. "Microelectronic Packaging," Scientific American, 121 (1), 86-96. 109-112 (July 1983)

6. Donath, W. E. "Placement and average interconnection lengths of computer logic", IEEE Trans. Ckt. Systems, vol. 26, pp. 272-277 (1979).

7. Heller, W.R., Hsi, C.G., and Mikhail, W.F. "Wirability - designing wiring space for chips and packages," IEEE Design Test (August), vol. 1, 43-51 (August 1984).

8. Heller, W.R., Mikhail, W.F., and Donath, W. E., "Prediction of wiring space requirements for LSI," Proceedings of the 14th Design Automation Conference, New Orleans, 32-43, 1977.

9. Donath, W. E. "Wire length distribution for placements of computer logic," IBM J. Res. Dev., vol. 25, pp. 152-155 (1981)

10. Nishi, Y. "Challenges in CMOS technology," Sol. State Technology, Vol. 31, No. 11, pp. 115-19 (Nov. 1988); Dansky, A. H., "Bipolar circuit design for VLSI gate arrays," ICCC80, pp. 674-7.

11. Keyes, R.W. "The wire-limited chip," IEEE J. Sol.-St. Ckts., **SC-17**, 1232-1233 (1982)

12. Keyes, R.W., "Device Limitations," Materials Research Society Symposium, Proceedings, Electronic Packaging Materials Science III, vol. 108, pp. 3-12 (1988)

13. Keyes, R. W., "The power of connections," IEEE Ckts. Dev. 7(3), 26, (May 1991); "Limits of Information Processing Systems," in Frontiers of Computing Systems Research, vol. 2, ed. S. K. Tewksbury, (Plenum, New York, 1991) pp. 1-59.

14. Bell, G., "Ulltracomputers: a teraflop before its time," Comm. ACM, vol. 35 (8), pp. 27-47 (August 1992)

15. Ho, C. W., et al. “The thin-film module as a high-performance semiconductor package”, IBM J. Res. Dev. vol. 26, pp. 286-296 (May 1982).

FIGURE CAPTIONS

Figure 1. The growth of the number of logic gates in large computers and on microprocessor chips.

Figure 2. The two regimes of information transmission technology.

Figure 3. The number of wiring tracks per terminal needed to successfully wire arrays of various sizes in the probabilistic simulations of reference 8. The line represents the relation *Wiring tracks* $= 2k_C m_1$.

Figure 4. Connectivity: The logarithm of Γ, Eq. (9), for chip substrates, as a funciton of the number of logic circuits packaged.

Figure 5. The growth of the logarithm of Γ, Eq. (9), for wiring on chips compared with the increase in the power of computing systems. The extrapolation shows the growth in power that may take place with massively parallel systems. [15]

Figure 6. A test of the scaling relation AD^2 = constant, Eq. (11), for chips and multichip modules manufactured around 1990.

Figure 7. The increase of the product AD^2 with time. The equation for the line is
$AD^2 = 240\exp(0.432(Year - 1960))$

OUTLINE

1. Introduction

   Early machines

   Miniaturization, integration, cost

   Many components, too much for mind

   Growth of numbers (components)

2. Communication in computers

   Communication as a determinant of physical form

   Regimes of communication

   Noyce quote

3. Empirical relations

   Rent's

   Wire length

   Master image, custom design

4. Wire model

5. Measuring connectivity

6. Scaling wire

7. The future

8. Conclusion

May 15, 1995