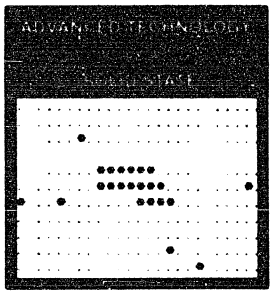
The trials of wafer-scale integration

Although major technical problems have been overcome since WSI was first tried in the 1960s, commercial companies can't yet make it fly



The superchips are coming! Units under development in industrial and academic laboratories both in the United States and Japan will integrate millions of logic gates into a single package the size of a man's hand. Ultimately, the circuits will encompass an entire wafer. In this size the superchips may be termed wafer-scale integrated (WSI) systems.

The concept is not new. Nearly two decades ago Texas Instruments used full wafers to implement the first large-scale integrated circuits [see "The early WSI effort at Texas Instruments," p. 34]. Gradually this early method of fabrication was supplanted by batch processing, in which 10 to 100 chip-sized systems are fabricated on one wafer and then diced, tested, and packaged individually. In 1967 the time was not ripe for WSI to emerge as the most economical fabrication and packaging concept. Neither the required fabrication technology nor the demand for such a level of integration was in place.

Today new design, fabrication, and testing techniques have evolved. More important, the market demand for more complex systems has climbed to the point where researchers are bumping into some hard lithography and yield barriers to the continued shrinkage of device size, long the hallmark of growth in the semiconductor industry.

At present a wafer—a thin slice of semiconductor material on which the devices used in microcircuits are fabricated—may contain 100 integrated circuits, or chips. A single circuit (or system) that fills the entire wafer represents the upper bound on the size of monolithic integration. Ranging in size from 2 to 8 inches in diameter, a single wafer can hold the equivalent of 25 to 100 microprocessors the size of an Intel 8086, or from 4 to 20 megabytes of dynamic memory if an 0.8-micrometer complementary-MOS process is used.

Though products of such density may be unfamiliar to many engineers, several commercial WSI memory products are already appearing in Japan, most notably a 4-megabit WSI ROM made by the NTT Musashino Electrical Communication Laboratory in Tokyo that holds the entire Kanji alphabet. In the United States several companies are busily pursuing WSI technology, including WaferScale Integration Inc., located in Santa Clara, Calif., and Mosaic Systems Inc. of Troy, Mich., among others.

Integrating circuits over an entire wafer offers several advantages over integrating circuits over much smaller chips and then interconnecting them. First, fewer pins protrude from a packaged system to connect the internal integrated circuitry to other packaged chips or wafers. Thus systems reliability is improved. Second, replacing a number of chips on a printed-circuit board with one integrated wafer eliminates interchip connections. These waste space, introduce erroneous signals (noise) into the circuit signal, slow the signal, and consume extra power to push the signal along the lengthy interconnections.

But wafer-scale integrated circuits have problems, too. One primary concern is heat removal; wafers may generate as much as 1000 watts of heat. While the total power consumed by a WSI system may be less than a similar IC system, because the interconnections are reduced, the power density on a single integrated wafer is much higher. Thus there is a need for packaging that can withstand high-power density and allow for efficient heat removal. In addition, WSI packaging must provide several thousand contact points to the circuit for signal and power distribution.

A second dominant problem is yield. Because wafers are so much larger than ICs, some means for repairing faulty gates is needed. Unlike a chip, a wafer is too costly to simply discard if bad. Circuit redundancy and discretionary wiring are solutions.

These hurdles are by no means slight. In August, Trilogy Systems Corp., a Santa Clara, Calif., company formed by Sperry Corp., Digital Equipment Corp., and CII-Honeywell Bull to commercialize a WSI process, ended its four-year effort to build a mainframe computer central processor implemented in WSI, laying off half of its 460 employees. The company will apply some of the technology to new packaging for commercial ICs, according to chairman Gene Amdahl, and will continue to experiment with related techniques, but it is "not clear" that WSI will rise again at Trilogy, according to Douglas L. Peltzer, Trilogy's vice president of technical operations.

Trilogy did develop several WSI products [Fig. 1], but could not envision making products economically, Mr. Pelzer said [see, "Why Trilogy dropped WSI," p. 37]. Nonetheless, other commercial companies, universities, and laboratories are still making progress in overcoming the hurdles to commercially competitive WSI products.

Reliability is high

The reliability of any packaged integrated device is improved if the number of pin connections protruding from the system is reduced. This is because input/output pads on the integrated circuit are especially vulnerable to electrostatic discharge, especially on NMOS circuits, the dominant very large-scale integrated-circuit technology. Also, the area consumed by I/O pads in small-scale ICs is substantial. Considerable chip area is also taken up by drivers, the circuitry next to the bonding pads that push signals off the chip onto package pins. And the drivers demand large amounts of current and power. On top of that, connection pins are vulnerable to mechanical failure.

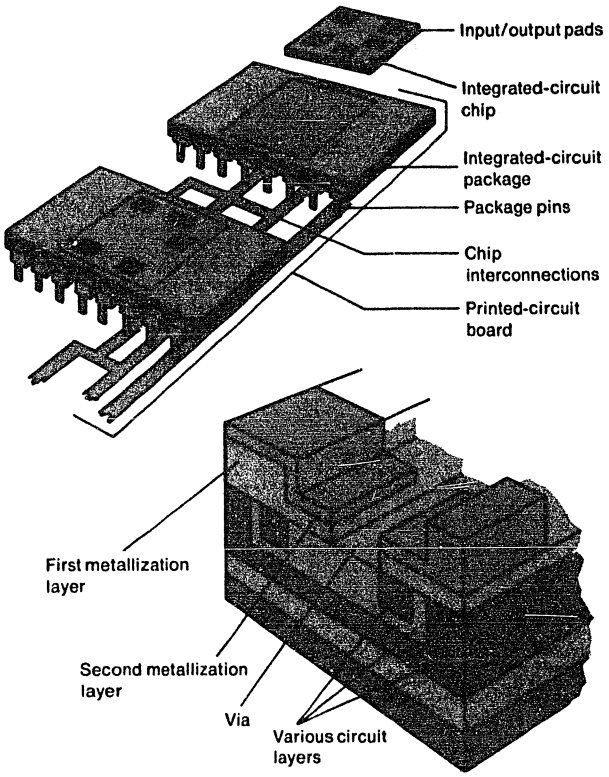
An estimate of the number of pins required for each package is provided by what is known as Rent's Rule. This is a statistical observation constructed from a broad collection of IC subsystems taken from many manufacturers. It states that the number of pins connecting a partitioned submodule to the rest of

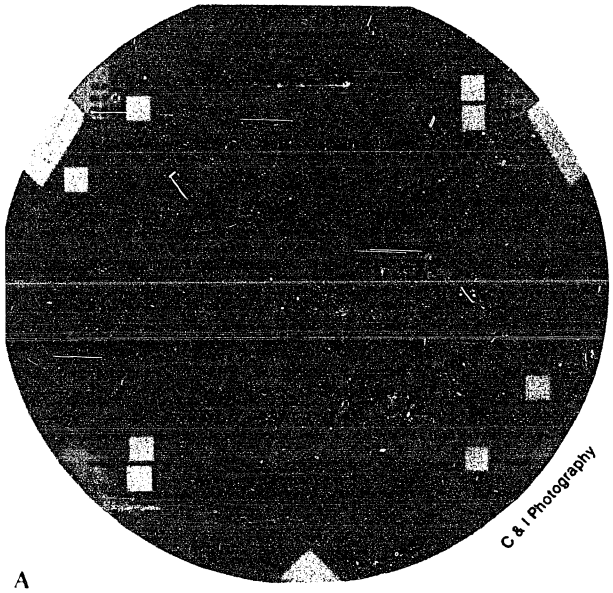
Jack F. McDonald, Edwin H. Rogers, Kenneth Rose, Andrew J. Steckl Rensselaer Polytechnic Institute

0018-9235/84/1000-0032$1.00©1984 IEEE

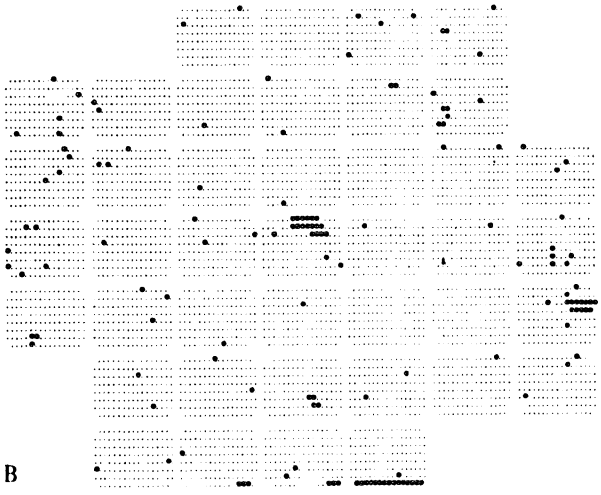
Defining terms

Dielectric insulator layer: a layer of nonconducting material laid down between layers of metallization to prevent electrical shorting between them.
Discretionary wiring: a network of metallization on a wafer that can be connected in various configurations to link working blocks of logic on the wafer.
Input/output pads: blocks of conducting material located on the periphery of an integrated circuit that send or receive electrical signals between chip circuitry and the package pins.
Integrated wafer: a wafer on which all the circuitry and interconnections are made in one monolithic structure on a continuous substrate.
Metallization: the physical network of metal lines that connect circuit elements on a chip.
Pins: metal leads extending from the chip package that transmit signals between the integrated circuit and the printed-circuit board into which it is inserted.
Vias: holes through the dielectric insulating layers that allow signals to transfer from one layer of metallization to another.
Yield: the ratio of good chips to bad, or area of good circuitry to bad, on a semiconductor wafer.





A



B

[1] A wafer-scale integrated wafer 100 millimeters in diameter (A) functions as one unit. Made by Trilogy Systems Corp., the active devices are contained within the 6-by-6-centimeter-square region, with process-monitoring structures along the periphery. Contact is made to the active devices through more than 2000 pins distributed over the surface. The wafer is trimmed into a square shape before packaging. One tool used by Trilogy to test WSI fabrication (B) is an array of over 5000 independent ring oscillators (small dots). The spatial distribution of circuit defects is shown by ring oscillators (test circuits) that fail to meet preset performance criteria (larger stars). Defects are grouped into "random" and "clustered" defects. Random defects are avoided by routing logic across redundant circuitry incorporated into the wafer design. Clustered defects, caused by fabrication process nonuniformities, are reduced by careful attention to manufacturing procedures.

a system is roughly proportional to the number of gates in the submodule raised to the 0.5 power. Rent's Rule shows that the ratio of the required pins to gates becomes more favorable in large systems in a single package; fewer pins are needed for a given number of gates, since more gates are integrated. This is certainly one of the driving forces for WSI.

Fortunately, Rent's Rule applies primarily to small submodules inside a much larger system. A wafer containing a complete system of 500 000 gates might not require 10^4 pins, as Rent's Rule would have it. When the package contains a major subsystem or the entire system, Rent's Rule can drastically overestimate the number of pins. Consequently WSI is particularly attractive when a nearly complete processor can be implemented on a single wafer.

A good example of what could happen if this were not quite achieved would occur in a system that required three wafers having 1800 pins per wafer package for interpackage communication. A factor of only three in improved density per wafer would provide a much lower pin count by permitting the entire system to be contained on one wafer; far fewer pins would be required.

For these reasons, memory systems may be the most attractive candidates for WSI. Only a few address, data, control, and timing pins would be required. Indeed, at the 1984 International Solid-State Circuits Conference, Eli Harari, president of WaferScale Integration Inc., speculated that WSI might displace hard disks for memory in some computer applications, opening an enormous market.

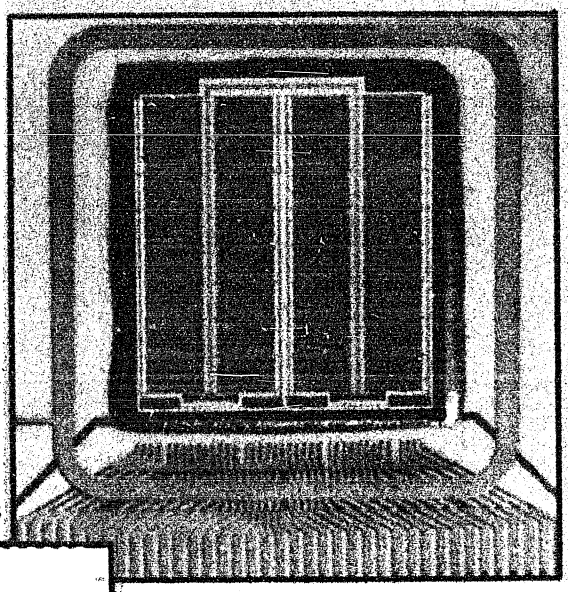
Interconnections minimized

Emphasis on minimizing interconnections in digital systems has been developing over the years. In the early days of digital systems the logic-switching elements in an integrated system were

The early WSI effort at Texas Instruments

The earliest attempt at wafer-scale integration (WSI) was the limited production in 1964 of the first large-scale integrated circuits by Texas Instruments in Dallas. The yield, or ratio of good chips to bad on a wafer, precluded the integration of more than a few tens of gates on a single chip. From 1965 to 1970 Jack Kilby, who patented the integrated circuit, and his co-workers developed the concept of discretionary wiring for the interconnection of systems directly on a wafer from small clusters of gates that had been tested and were known to work. One attempted application was undertaken in conjunction with Burroughs Corp. to make an Illiac IV computer implemented in WSI.

A wafer contained perhaps 50 cells with 10 or so gates apiece. Once the good cells were located, a computer program mapped the desired logic function onto these working sites and created the routing to interconnect them. Figure A shows one of these WSI systems in a 156-pin package. The regular pattern of discretionary wiring is clearly visible as the overlying grey lines. In developing the wafer, researchers found that successful WSI implementation would hinge on three (now classic) technologies: wafer-probe testing of circuitry, discretionary wiring, and wafer cooling and packaging.



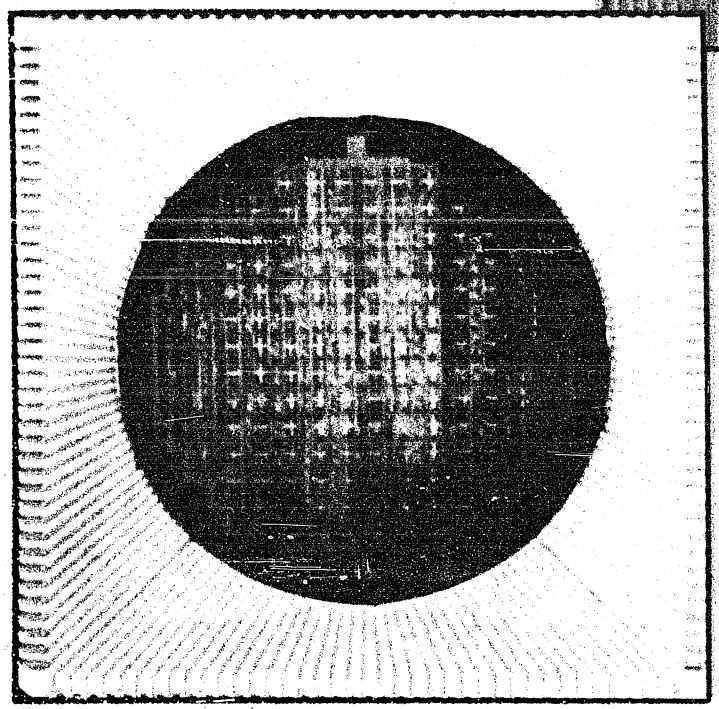
Texas Instruments Inc.

B

In the autumn of 1970 a Texas Instruments team led by Harvey Cragon designed and produced a 32-kilobit memory implemented in wafer-scale integration (Fig. B). Discretionary repair was used to attach good unit cells to the address and data buses. More than 50 percent of the wafers had a sufficient number of good unit cells.

Despite such successes, however, WSI never caught on. The required fabrication technology was then too expensive and not refined. Also, there was simply not a great demand for the high level of integration attainable with this technique.

—*J.F.M., E.H.R., K.R., A.J.S.*



A

A 2-inch-wide WSI logic array built by Texas Instruments in 1967 (A) used discretionary wiring. A 2-inch-wide, 32-kilobit wafer-scale memory developed by Texas Instruments in 1970 (B) used discretionary wiring to implement redundant circuits on a ceramic substrate.

larger and more expensive than the interconnections between them. Considerable ingenuity was devoted to devising design tools to minimize the size and power consumption of the switching elements for a given logic function.

This situation is now reversed in VLSI technology, because the active devices have shrunk in size faster than the interconnections. This is now widely recognized, and while reducing the numbers of active devices is still important, because of their power requirements, considerable effort is being expended devising methods to minimize interconnections. Aside from the obvious implications of signal delays induced along long wires, there is the problem of the area they take up.

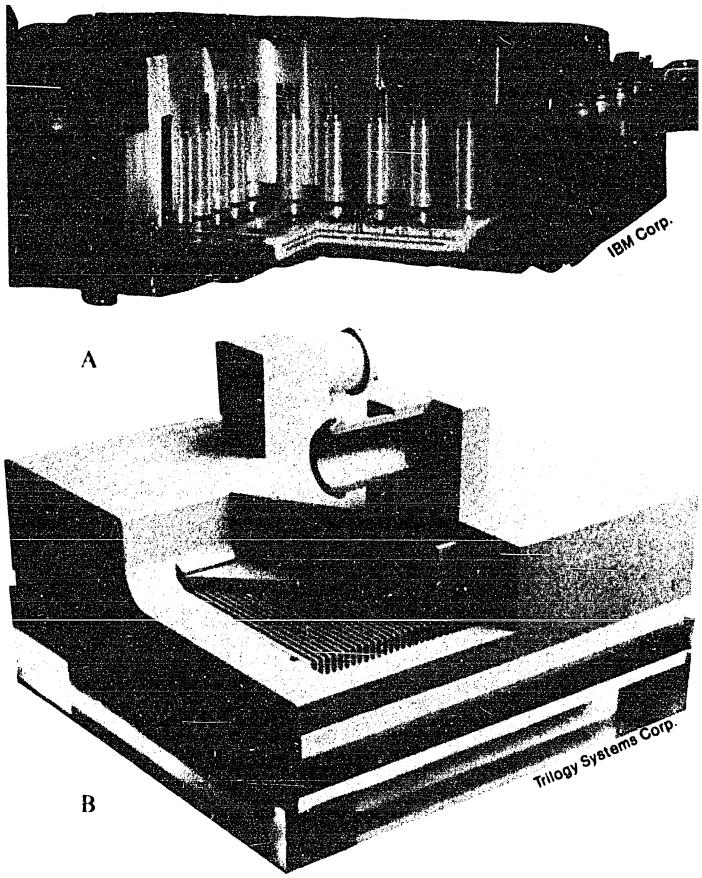
For example, it is not unusual to find VLSI circuits in which 60 to 80 percent of the surface area contains no active devices, only interconnections. Furthermore the more gates in a system, the longer the average interconnection length tends to be in typical digital architectures. This problem is likely to be worse in WSI unless the circuit's interconnection requirements are atypical.

A major concern in the interconnection layout for WSI is the requirement for wiring crossovers, so-called "vias." These take up a substantial area if the number of wiring layers is limited. Typically only three to four monolithic wiring layers are available in WSI, although additional wiring layers inside the WSI package are possible.

The design of the IBM 3081 provides a vivid picture of the interconnections required by a modern mainframe computer. The 3081 is implemented in bipolar (TTL) gate-array chips. Each chip uses three layers of metal to interconnect up to 704 logic gates, and the chip has 121 solder pads (IBM's version of bonding pads). A hundred or more logic chips are mounted on a ceramic substrate containing up to 33 layers of interconnections. This substrate is packaged in a thermal conduction module (TCM), a replaceable unit [Fig. 2A]. A 3081 processor unit with 16 channels contains 26 TCMs with additional interconnections.

Hence the interconnection area requirements for a WSI package (which can be regarded as a kind of monolithic TCM) will be high. Special computer-aided-design tools are required to manage this problem, even in architectures well suited to WSI implementation.

The architecture selected for WSI implementation is crucial. Non-Von Neumann, highly parallel architectures such as systolic arrays may be better matched to the technology now available, particularly if fast logic is employed. Here WSI technology points toward arrays of tightly coupled, fast parallel processors with relatively simple interprocessor interconnections. In spite of this, Trilogy's initial WSI effort is based on a conventional Von Neumann architecture. To accommodate this the Trilogy wafer consists of 20 processing layers. The reliable fabrication of these has proved to be one of the greatest technological obstacles.



[2] Two companies have built packages to remove large amounts of heat from dense, large-scale integrated systems. The IBM Thermal Conduction Module (A) used to hold and cool chips on the company's 3081 mainframe computer, mounts large chips on a 22-layer ceramic substrate. On the opposite side of the chips are spring-mounted metal plungers that act as individual heat sinks. Helium gas is circulated in the structure to extract heat from the module. In the Trilogy Systems package (B) an intact wafer is mounted in a molybdenum heat plate. The plate is water-cooled, and the module is hermetically sealed and filled with helium. Because the heat plate is eutectically bonded to the wafer its heat-transfer efficiency is high, but the number of wafer layers that can be interconnected is severely limited.

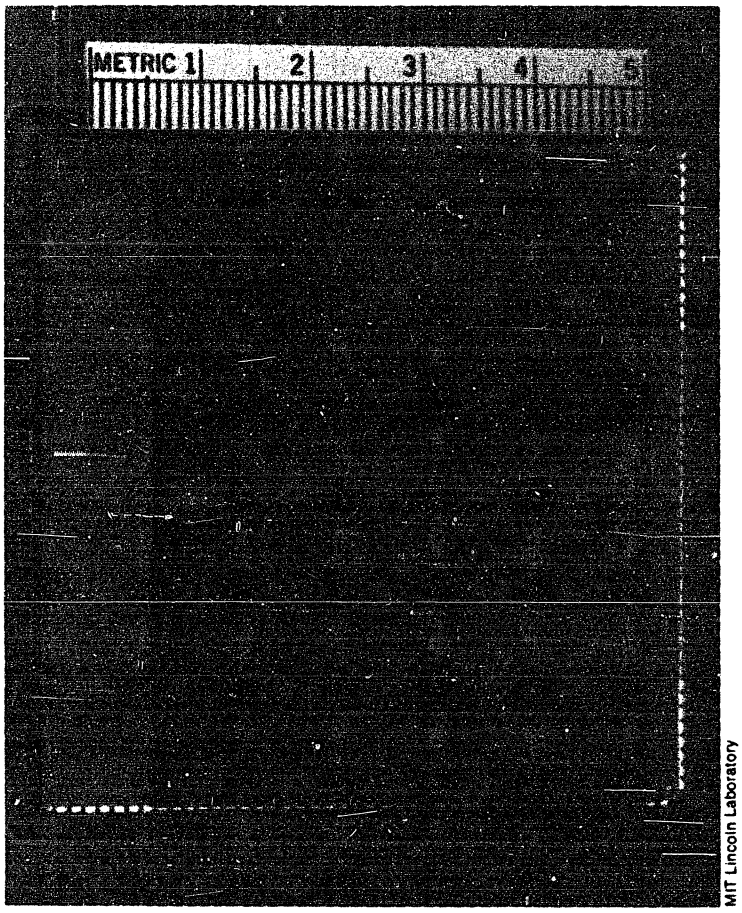
Signal delay decreased

The 100 or so chips that are typically diced, packaged, and mounted on a printed-circuit board or ceramic carrier are constrained to use interchip connections that are long, noisy, power-hungry, and slow. Signal transmission is aggravated by space between the chips and by the packages themselves. For example, the ratio between the area of a typical package and the area of the chip it contains can range from 2:1 to 10:1. If we include in the footprint of the package the average area of the printed-circuit board or supporting ceramic carrier, the ratio might become 40:1 or higher. The package pins introduce large parasitics that play on the electric characteristics of the package, which effectively can make the ratio look worse.

Integrating these chips on one wafer wastes much less space. Many of the connections through or between packages are eliminated. Also, with shorter and narrower interconnection lines, there is less self-capacitance in the circuitry, and thus less charge is needed to change the state of a logic gate from on to off and vice versa. This reduces the size and power consumption of line drivers. Since connections are such a dominant part of most circuit layouts, WSI can substantially speed systems and reduce their total internal power requirements while improving density and decreasing costs.

Nevertheless caution is advised, for though the number of pins per package is improving in WSI, the interconnections have not disappeared altogether. In essence, they are now inside the package—on the wafer. The average- and worst-case length of these internal interconnections increases, and so does the internal signal delay. These internal connections generally involve less capacitance than external connections, but signal delay on these fine paths can still be substantial, especially if the placement of components internal to the package is not optimal.

To keep the longest lines from producing unacceptable delay, it is possible to insert repeater amplifiers along these paths, a technique developed this year by H.B. Bakoglu and J.D. Meindl



[3] The "Restructurable VLSI" digital integrator circuit developed at MIT's Lincoln Laboratory is rewired by the use of an argon laser to cut and form connections around defective circuitry after wafer fabrication. The circuit has 64 working cells (130 000 transistors) out of a total of 192. Each cell consists of four 10-bit counters and access shift registers designed with a 5-micrometer CMOS process. This kind of structure lends itself nicely to implementation of cellular arrays of logic in which only the nearest neighbors communicate. The buses are organized with crossbar switches at the intersection of horizontal and vertical discretionary wiring channels.

at Stanford University in Palo Alto, Calif. By keeping the line resistance between repeaters less than the amplifier source resistance, this approach can make the delay linear. In effect, the interconnection lines appear to be without resistance. One can then decrease the source resistance while increasing the number of repeaters to increase speed further. However, increasing the number of repeaters usually uses more space and requires more power. The inherent delay caused by repeater internal parasitics usually can be ignored for the longer WSI lines.

Another way to reduce signal delay in interconnections is to use relatively thick metal-conductor and dielectric-insulator layers. Metal lines on the order of a few micrometers thick on dielectric layers roughly tens of micrometers thick will produce connections of transmission-line quality up to several gigahertz. Multiple layers like this, although much thicker than those normally found on today's ICs, can be made. This scheme has already been tried in certain packages employing fine pitch metal lines. Hence this approach amounts to fabricating part of the package directly on the surface of the wafer.

An obvious extension of fabricating thick metal transmission lines on thick dielectric layers would be the use of microminiaturized optical waveguides to carry light signals between the blocks of logic on the wafer rather than electric signals. These can be fabricated by deposition and etching of the glass guides on the wafer or by ion implantation of glass layers. One advantage of this approach would be the reduction of the power lost in conventional transmission-line terminators. However, the mixture of logic and optical circuits would have to await the introduction of gallium arsenide WSI. Also many structures, such as optional couplers and the optical equivalent of vias, would have to be developed.

Removing heat a key challenge

Although power dissipation can be kept fairly uniform over the wafer surface in a WSI system, thereby eliminating nonuniform, thermally induced mechanical stress, heat stress is a problem for the package. The kind of packaging required is already being used to mount large single devices for power semiconductor circuit elements, such as silicon-controlled rectifiers. Hence there exists a technology base for managing WSI heat removal, but further improvements are required.

The power-semiconductor approach appears capable of removing heat at the density rate of 10 to 20 watts per centimeter squared. For a wafer measuring 60 cm^2, this technology could dissipate 600 W of heat, but problems exist in scaling the technology up to this total power level.

One problem with scaling up concerns differences in the thermal expansion coefficients for the materials used in the substrate, interconnection layers, and heat sinks. Stresses built up by these differences can become large enough to damage the fragile metallization or dielectric layers. The effect of many large on-off temperature cycles can cause fatigue in the bonding surface between the wafer and the primary heat-removal surface of the package. The larger the wafer size, the larger the strain that can result from the stress.

Some materials that have shown promise as heat-removal substrates are silicon carbide and aluminum nitride, which match the thermal expansion coefficient of silicon very well. Both have high thermal conductivities. However, the best solution seems to be to first minimize the heat to be removed.

For comparative purposes, consider an Intel 8080A microprocessor. Made in NMOS technology, each chip dissipates about 1 W at a surface flux of 3 W/cm^2 while operating at a 1-megahertz clock rate. A hundred such chips operating on a single 4-inch wafer would dissipate 100 W. If a CMOS version of the same circuit operating at the same clock rate had been made, it would dissipate two orders of magnitude less power, or roughly only a few watts. This is because CMOS has almost no static power dissipation (the power required by a gate when it is off), while NMOS requires more power when off.

Of course, dynamic power dissipation (when gates are on) increases proportionally to clock rate, and CMOS can consume more power than TTL at higher frequencies. However, at those frequencies all technologies dissipate tremendous amounts of power. It would therefore appear that CMOS technology is most suitable for WSI. The MIT Lincoln Laboratory of the Massachusetts Institute of Technology in Cambridge has fabricated CMOS wafer-scale integrated chips containing 300 000 circuits by use of very conservative design rules.

Undaunted by the formidable heat dissipation of bipolar circuits, Trilogy has developed a packaging technology capable of removing 50 W/cm^2 [Fig. 2B]. Hence a Trilogy wafer could

dissipate over 1 kilowatt, allowing the use of emitter-coupled logic circuitry with substantially higher clock rates.

An interesting packaging alternative for WSI is to immerse the wafer directly into liquid nitrogen (4 degrees Kelvin). In addition to providing excellent heat transfer this would have other benefits. For example, the various mechanisms that cause the circuit to deteriorate with age tend to slow at lower temperatures; the mean time to failure will increase exponentially with lower temperature.

Higher carrier mobilities for the semiconductor switching elements that result at extremely low temperatures would increase speeds and actually lower the heat dissipation. Threefold increase in circuit speed or higher is possible in silicon-based systems. Shorter MOS channel lengths are possible at lower temperatures because of an improvement in the soft turn-on characteristics of the devices. Leakage currents decrease exponentially with temperature. As a result, dynamic latches behave like static latches when operated in liquid nitrogen temperatures. This would therefore permit a significant reduction in circuit size.

But the use of liquid nitrogen cooling would require smaller, more efficient, less expensive heat exchangers, nitrogen liquifiers, and pumps than those available today. In addition, the WSI system must be protected from thermal shock, which can crack the wafer.

Signal quality also questioned

Any designer of board-level systems is aware of signal quality, noise, and power-distribution problems. These arise particularly when high-frequency signals are transmitted in board systems, but in the dense layout found in VLSI or WSI they become acute at lower frequencies. Besides the signal delay accumulated across interconnections, an even more serious difficulty arises from the increase in interconnection density: interline coupling, caused by mutual capacitance.

In this mechanism, signals running along one line are conducted by a close parallel line. Interline coupling can arise between adjacent lines on a given layer or between lines on overlapping layers. This coupling is most dangerous on long runs of parallel lines. The decreased separation between lines in WSI, compared with that in other packaging arrangements, is a significant source of this problem. To help contain it, long parallel runs must be avoided in layout.

As for the power-supply distribution, pin and wiring inductance make it difficult to bypass switching transients on the power lines. Because of large precharge currents, this is a problem during refresh cycles in systems containing large dynamic RAMs. However, it may be even more severe in WSI logic systems because of the likely use of increased parallelism. In traditional architectures perhaps only 20 percent of the logic elements change state in any given clock cycle. In a highly parallel system this fraction of active changes can rise to 100 percent, demanding much larger surges of current.

One strategy that has been helpful in densely integrated dynamic random-access memories has been the creation of numerous skewed clock edges to spread out the current surges. This approach may help in wafer-scale memory, but the use of many skewed clocks in random logic may prove awkward and tend to place constraints on the design.

An additional consideration related to the distribution of power in WSI is electromigration. This phenomenon can cause the physical deterioration of the atomic structure of the metallization if the density of this current is too high. If the WSI system contains 100 processors drawing 100 milliamperes each, the total supply current would be 10 amperes. If 1-micrometer-thick metallization is used to carry this current, a total conductor width of at least 1 cm would be required to maintain the current density below the electromigration limit of 10^5 A/cm^2, above which the conductor begins to disintegrate.

Discretionary wiring needed

WSI systems also must be designed to tolerate circuit defects on the wafer. The origins of these faults are numerous, but not so abundant that patient research cannot identify and reduce them. Improved clean rooms, automated wafer handling, defect inspection, ultrapure chemicals, refined processing tools, and process-tolerant circuit design have already improved wafer yield. However, for wafer-scale integration some residual defects are inevitable. These can be descriptively classified as "point" or "cluster" defects. In this model a single point defect is assumed to render the circuit of area A inoperative. The larger A is, the lower the wafer yield becomes. Hence, one tries to fabricate WSI wafers from cells of small area.

To fabricate a system the size of a full wafer with reasonably high use of the available devices requires either a repair capability, a tolerance for failure, or a combination of both. Tolerance for failure implies redundancy in some form, while repair capability implies intervention in the fabrication process. The wafer-yield enhancement resulting from this redundancy or intervention remains one of the most potent benefits of WSI.

Why Trilogy dropped WSI

Although many popular press accounts of Trilogy Systems Corp.'s decision in August to drop its wafer-scale-integration effort cited a variety of technical "reasons why," Trilogy's wafers did, in fact, work. Unfortunately, they would have cost too much and taken too long to develop.

"We were caught in a tradeoff between circuit speed and yield," explained Douglas L. Peltzer, Trilogy's vice president for technical operations. Altering the wafers to make them affordable to potential customers would have made the wafer circuitry too slow.

Trilogy was attempting to make a large, IBM-compatible mainframe computer implemented in WSI. It planned to produce a compact, highly reliable mainframe in 1985 with speeds up to four times faster than IBM's largest uniprocessor, the 3083-J, according to Trilogy engineers. The company had overcome the problem of dissipating the large amount of heat generated by dense WSI circuitry with unique packaging schemes. In addition, although the wafers had four layers of metallization (most ICs have one or two), shorting between the layers was also overcome by using redundant metallization lines.

But the level of redundancy, a basic WSI technique for routing circuit and metal lines around defects and shorts, got too complex. "We used a lot of metallization to cover all the needed redundancy," Mr. Peltzer said, "but that lowered the circuit speed," reducing the attractiveness of the integrated wafers.

Trilogy's early wafers had very low yield due to defects and shorts. Good, fast wafers were made, but would have cost too much, their price having to offset the cost of making the wafers that did not work. The yield was raised by implementing many redundant lines to avoid defects, but the long lines took up area on the wafer, pushed active devices further apart from each other, and resulted in longer signal paths. Thus the circuits got too slow.

Trilogy thought of ways to rectify the speed versus cost tradeoff, Mr. Peltzer said, but it would have had to redevelop its computer-aided design equipment, redevelop its macrocell library (a set of circuit building blocks that are interconnected to make ICs or integrated wafers), and redesign its metallization scheme. Company leaders decided all the changes would cost too much and delay eventual products too long.

Mr. Peltzer suggested that it may be too great a leap in technology to economically implement WSI immediately. The alternative might be to "use pieces of the technology to enlarge existing ICs," until they themselves eventually grow to become integrated wafers.

—Mark A. Fischetti

From 1965 through 1970 Jack Kilby at Texas Instruments developed the concept of discretionary wiring to improve yield. With this technique the wafer was subdivided into small independent clusters of logic. A wafer had only 501 five-input gates and 95 six-input gates. These clusters of devices were of small enough area to enjoy reasonably good yield. Wafers were first fabricated and tested to identify which clusters worked, then the desired WSI system was configured from the working clusters.

Implicit in this approach was a degree of interchangeability between clusters that was exploited by a computer program matching the desired logic function components with the working cells. The program also produced the layout for a unique mask set for interconnecting the working cells. The mask set for this discretionary interconnection was then created with an electron-beam mask-masking machine.

In this scheme all tested wafers had to be personalized by their own mask sets. Rapid processing of the final metal interconnection layers was achieved by use of the mask set in conventional photolithography machines. However, at the time electron-beam mask-masking machines were slow and prohibitively expensive. Today, however, electron-beam machines have high throughput and are used widely in industry. What's more, direct-write-on-wafer electron-beam machines have emerged, which facilitate a full tailoring of the discretionary wiring between working cells directly on the wafer. The optical mask is unnecessary.

The IBM EL-2 direct-write electron-beam lithography machine, for example, is routinely used to personalize gate arrays automatically at the company's quick-turnaround fabrication facility in Fishkill, N.Y. The EL-2 can directly write complex and varied wafer-scale metallization patterns on different wafers and in different regions of a given wafer.

With discretionary wiring, wafer-scale yield then depends on the integrity of the discretionary interconnections, which themselves are vulnerable to defects. Pinholes in insulator layers, dirty contact openings, residual contaminants, and poor coverage of metallization over the dielectric insulator are among the sources of these defects.

The Texas Instrument's discretionary writing approach suffered primarily from the yield reduction inherent in the additional processing steps. It quickly succumbed as higher yields on larger ICs with fewer interconnection layers filled the most pressing market needs of the day.

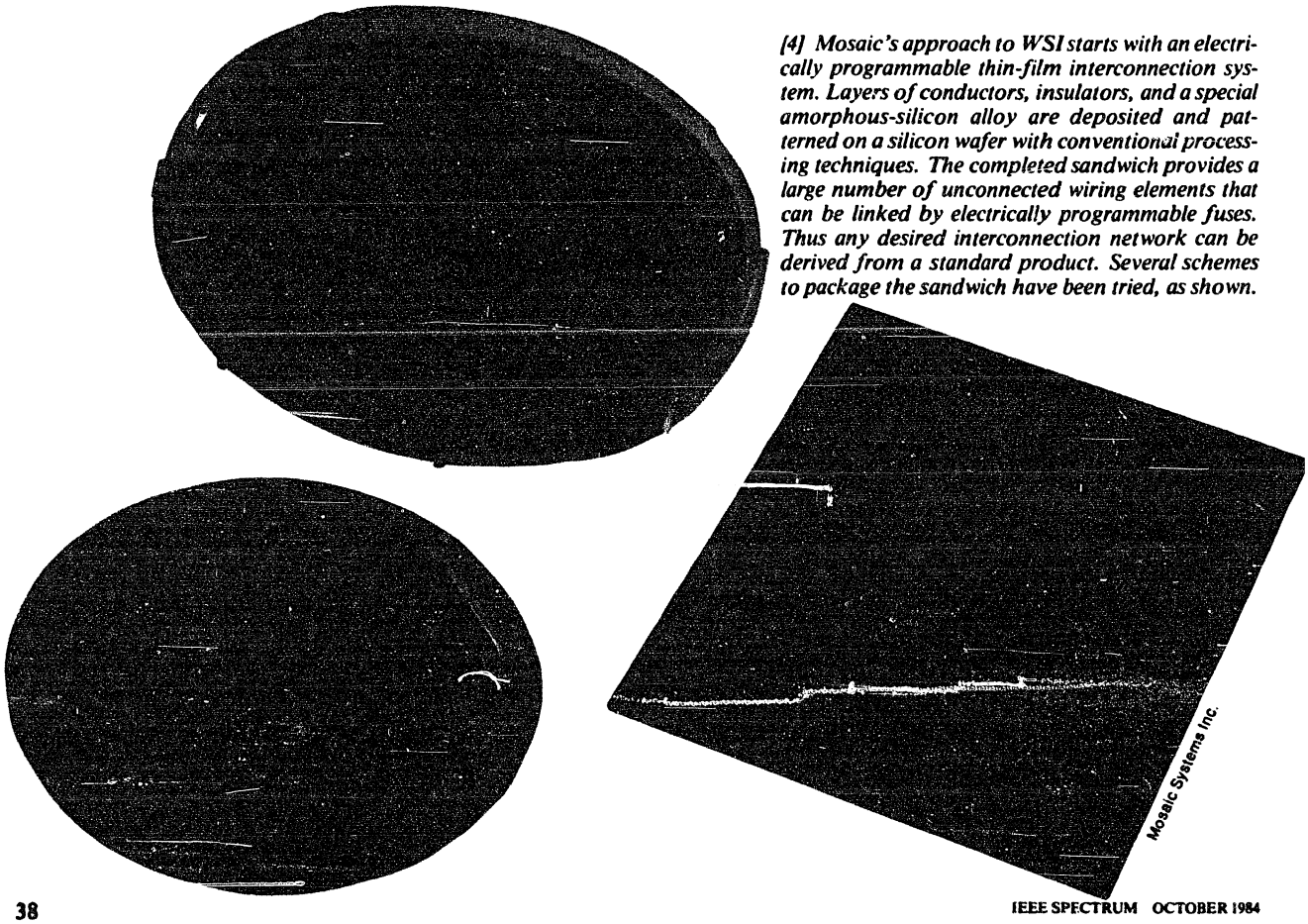
Recent discretionary approaches succeed

Through the 1970s WSI lay dormant as great strides were made in what became the conventional VLSI chip-packaging approach. Lithographic feature sizes were scaled down from mils to micrometers. Circuit density climbed by nearly three orders of magnitude, while yield improved. Microprocessors came into existence, spawning hundreds of new applications and millions of new jobs. One might be tempted to conclude that conventional IC systems could easily produce all the digital products of interest in the marketplace.

But by 1980 it was becoming clear that some market interest existed for even larger systems that could not easily be produced with good yield using conventional techniques. The industry began to reinvestigate discretionary techniques. In 1981 the NTT Musashino Electrical Communication Laboratory in Tokyo developed a 1-megabit RAM on a 3-inch wafer. In this case discretionary repair was accomplished by means of fusible links; defective memory cells were isolated when excess power was applied to blow specially placed links.

At the Lincoln Laboratory's Reconfigurable VLSI Project, researchers have been developing a generic fabrication methodology for WSI [Fig. 3]. In this scheme, arrays of small processors are fabricated along with internal, uncommitted interprocessor buses. Lasers are used both to cut interconnection lines and isolate cells from the bus. Moreover researchers have demonstrated the use of the laser to weld electrical contacts.

In another Lincoln Laboratory development researchers have



[4] Mosaic's approach to WSI starts with an electrically programmable thin-film interconnection system. Layers of conductors, insulators, and a special amorphous-silicon alloy are deposited and patterned on a silicon wafer with conventional processing techniques. The completed sandwich provides a large number of unconnected wiring elements that can be linked by electrically programmable fuses. Thus any desired interconnection network can be derived from a standard product. Several schemes to package the sandwich have been tried, as shown.

used an electron beam to program "floating" oxide-encapsulated polysilicon gates similar to those used in EPROMs. Floating refers to the complete isolation of the FET gate-electrode in the insulator layer. Both charging and discharging of the gates is feasible. A 128-K electron-beam programmable read-only memory has been demonstrated in a wafer-scale environment. The target gates were depletion-mode NMOS floating gates 6 micrometers on a side. The beam used had a spot diameter of 0.5 micrometer and was applied for 0.4 ms to set each gate. This approach could be employed in a variety of ways. For example, PROM microcode could be altered in logic systems. Connections to bus systems could be programmed or reprogrammed. Columns or rows in programmable logic arrays could be restructured.

At Mosaic Systems Inc., researchers are reducing the loss of wafer surface area because of defective cells. Mosaic was founded in 1981 by Robert R. Johnson to develop and apply an electrical thin-film interconnection system that appears in some respects to resemble a wafer-scale printed-circuit board. In this scheme, layers of conductors, insulators, and a special amorphous-silicon alloy are deposited and patterned in a silicon wafer by conventional processing methods. The completed sandwich provides many uncommitted wiring elements that can be linked by electrically programmable "antifuses" [Fig. 4].

WSI research is also being performed at several other companies. However, the work in well-established firms is not highly publicized. New ventures such as the highly visible Trilogy are perhaps more clearly committed. In Trilogy's case extensive new technology for wafer packaging, testing, design, routing, and discretionary repair have been developed. The Trilogy effort is centered on fast, hot-running bipolar logic, requiring a maximum effort at heat removal.

Laser and ion beams aid fabrication

Cautious enthusiasm marks a project at Columbia University in New York City. Directed by Richard Osgood, researchers are using lasers as a direct-write tool in circuit fabrication. Parallel to work at the Lincoln Laboratory and Lawrence Livermore National Laboratory in Livermore, Calif., the technique uses a focused laser beam to excite chemical reactions on the surface of the wafer. Deposition and etching operations that can be performed in succession may be possible through variations in the gaseous environment on the wafer surface. This may be useful in producing high-yield discretionary routing and repair.

At the Center for Integrated Electronics of Rensselaer Polytechnic Institute in Troy, N.Y., a research effort in focused electron and ion beams also has the potential to provide new discretionary routing or repair techniques. Work with ion-cluster beam-deposition systems may lead to improved multilayer films for both insulator and interconnection layers by providing a deposition method at low wafer temperature. In addition to fabrication research, the Rensselaer WSI effort includes architectural and computer-aided-design research. Three architectures are under investigation. These include wafer family sets for real-time video image processing, data-base machines, and design automation.

The future for WSI

It is difficult to predict the fate of any technology in the rapidly moving semiconductor industry. Nevertheless, the first commercial domestic WSI products in the United States are likely to be memory systems, just as they have been in Japan. Of seven U.S. and Japanese experimental integrated wafers displayed during a panel session at the International Solid-State Circuits Conference in February, five were memories. However, WSI random logic systems are likely to follow in certain specialized areas. Ultimately WSI may appear in a broad spectrum of products for military and commercial electronics systems.

One future possibility is suggested by the newly emerging GaAs technology. The performance levels of GaAs WSI might justify the cost of GaAs to the user, whereas GaAs ICs may not. Another possibility exists for three-dimensional stacking of the WSI systems. One advantage of the 3-D organization is the possibility of implementing extremely short signal paths through the wafer to nearby logic sites on other layers or other wafers.

The real question is whether WSI can produce today the complex systems of tomorrow at a slightly higher cost.

To probe further

The following papers all give specific information on current WSI research and that conducted in the late 1960s:

- Bakoglu, B.H., and Meindl, J.D., "Optimal Interconnect Circuits for VLSI," *Proceedings of the 1984 IEEE International Solid-State Circuits Conference*, Feb. 22-24, 1984, pp. 164-65.
- Deutch, A., and Ho, C.W., "Thin Film Interconnection Lines for VLSI Packaging," *Proceedings of the IEEE International Conference on Computer Design*, Rye, N.Y., 1983, pp. 222-26.
- Landman, B.S., and Russo, R.L., "On a Pin-Versus-Block Relationship for Partition of Logic Graphs," *IEEE Transactions on Computers*, Vol. C-20, December 1971.
- Lathrop, J.W., Clark, R.S., Hull, J.E., and Jennings, R.M., "A Discretionary Wiring System as the Interface between Design Automation and Semiconductor Array Manufacture," *Proceedings of the IEEE*, Vol. 55, no. 11, November 1967, pp. 1988-97.
- Lyman, J., "Electronics Components Conference to Spotlight Advances in Materials," *Electronics*, May 3, 1984, pp. 134-35.
- Pletzer, D.L., "Wafer Scale Integration—The Limits of VLSI?" *VLSI Design*, September 1973, pp. 43-47.
- Petritz, R.L., "Technological Foundations and Future Directions of Large-Scale Integrated Electronics," *Proceedings of the FJCC*, 1966, pp. 65-87.
- Rode, A., Flegal, T., and LaRue, G., "A High Yield GaAs Gate Array Technology and Applications," *Proceedings of the IEEE Gallium Arsenide Integrated Circuit Symposium (GaAs IC)*, October 1983, pp. 178-81.
- Stapper, C.H., Armstrong, F.M., and Saji, K., "Integrated Circuit Yield Statistics," *Proceedings of the IEEE*, Vol. 71, no. 4, April 1983, pp. 453-70.

About the authors

John F. McDonald (M) has been a professor of electrical engineering at RPI since 1974 and is one of the founding faculty of RPI's Center for Integrated Electronics. His current research interests involve WSI. He holds four technical patents. Prior to joining RPI he was an assistant professor at Yale University, where he received the Ph.D. in electrical engineering.

Edwin H. Rogers is a professor of computer science and mathematics. His research includes yield analysis in WSI, architecture for data bases, and logic programming and high-level CAD tools. Previously he worked at the Ballistic Research Laboratories at Aberdeen Proving Ground. He received the Ph.D. in mathematics from Carnegie-Mellon University.

Kenneth Rose is a professor of electrical engineering and also a founding faculty member of the Center for Integrated Electronics. His current research includes WSI architectures and processing issues. He received the Ph.D. in electrical engineering from the University of Illinois.

Andrew J. Steckl (SM) is a professor of electrical engineering and director of the Center for Integrated Electronics. His present research is in VLSI processing and devices. He has chaired the University Advisory Committee for the Semiconductor Research Corp. Prior to joining RPI he did research on charge-coupled devices for Rockwell International. He holds a Ph.D. in electrical engineering from the University of Rochester. ◆

The authors would like to acknowledge the help of the following individuals in preparing this article: Harvey Cragon at Texas Instruments Inc., Jack Raffel at MIT's Lincoln Laboratory, Herb Stapper at Mosaic Systems, and Michael Current and Douglas Peltzer at Trilogy Systems Inc.