NOVEMBER 1977

HEWLETT-PACKARD JOURNAL

An NMOS Process for High-Performance LSI Circuits

Fast 16-bit microprocessors, 16K read-only memories, and a variety of special-purpose random-logic chips are the result of an NMOS process that produces high-performance large-scale integrated circuits.

by Joseph E. DeWeese and Thomas R. Ligon

THE MOST DRAMATIC IMPACT of LSI (large-scale integration) in recent years has been on computers and calculators. While the development of hand-held calculators has been highly visible to all, what has happened and continues to happen with their larger and more sophisticated cousins has been little less than profound. While the further integration of TTL and other small and medium-scale integrated circuits has in some cases resulted in much smaller machines consuming less power, more significantly it has led to much greater functional density allowing dramatic increases in performance for a given volume. Thus it is that the newer desktop calculators use more complex and powerful languages, contain more memory, and are able to solve more complex problems in a shorter time than their predecessors. Their new capabilities class them more properly as desktop computers.

Obtaining these advanced capabilities in HP desktop computers was not simply a matter of adapting available off-the-shelf LSI devices to meet objectives. This would have stretched development time since new devices must be made available before product design can be undertaken. With circuit design, LSI process development, and product design proceeding simultaneously on parallel paths, the result can be products based on LSI devices that are years ahead of anything available on the market. The products thus achieve exceptional performance/cost ratios. This is the approach used by Hewlett-Packard.

Developing an LSI process is a major undertaking (a process as used here refers to the procedure by which an integrated circuit is fabricated beginning with the bare silicon wafer). The technologies employed for forming or modifying particular parts of the structure, such as oxidation, diffusion, ion implantation, photolithography (pattern formation), and various types of thin-film deposition, are all complex subjects in themselves. What makes process development particularly challenging is the high degree of interaction that can and almost always does occur between the various steps in forming the device.

Undertaking NMOS

Previously described in these pages was a central processing unit (CPU) built on a single chip for mainframe computer systems by an HP-developed CMOS/SOS process.[1] We would now like to describe an NMOS process that was developed to meet the needs of desktop computers. This process provides the HP Model 9825A Desktop Computer with a CPU that has the speed and power of the HP Model 2100A 16-bit Computers on a hybrid circuit (Fig. 1) small enough to fit in a vest pocket.[2] Two of these hybrids are used in the new Model 9845A Desktop Computer (Fig. 2).

Why MOS? LSI circuits use MOS (metal-oxide-semiconductor) field-effect transistors more often than not because thousands of these devices can be packed on a single silicon chip of reasonable size. While bipolar devices (NPN and PNP transistors) are used in some LSI circuits, their manufacture requires several diffusion steps and the electrical isolation from the common substrate required by each device

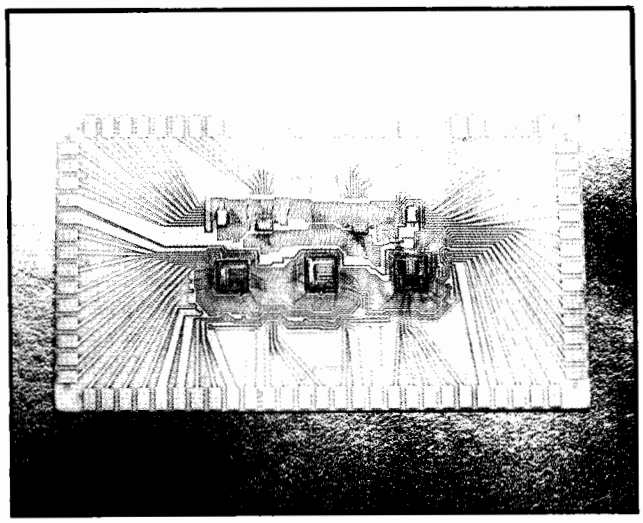


Fig. 1. *Hybrid microcircuit has a CPU chip, an input-output chip, an extended-math chip, and four bipolar interface buffers to form a 16-bit parallel processor that has the speed and power of a minicomputer.*

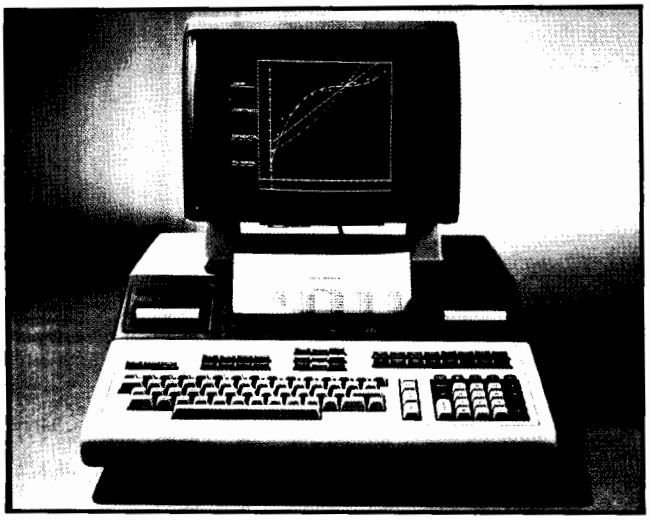


Fig. 2. *Recently-announced Model 9845A Desktop Computer uses two hybrid microprocessors similar to that shown in Fig. 1. These plus NMOS 16K ROMs mounted on 8-chip hybrid substrates give Model 9845A an extremely powerful central processor and an exceptionally large mass memory. Other NMOS LSI chips drive the thermal printhead and control the tape driver.*

has been a serious constraint on dense packing. MOS devices in most cases require only a single diffusion step and they are inherently self-isolating so they can be packed more closely.

There are powerful incentives for packing more devices on each chip, the primary incentive being an economic one. At today's prices the cost of a fully processed, 3-inch wafer might range from around $50 to $100. The wafer may have some hundreds of chips but only a fraction of that number will be sufficiently free of defects to be acceptable. The contents of each chip may range from a few hundred transistors, for simple logic or driving circuits, to well over 10,000 transistors for a microprocessor chip. The point is, the cost of chips may vary by a factor of only 10, depending on yield, process complexity, and so on, so processes that can put thousands of transistors on a single chip achieve substantial reductions in the cost per function.

Using a larger chip to allow more devices per chip may not be cost effective because the chances of a defect occurring on each chip then increases, with a consequent drop in yield. This is further compounded by the reduction in the number of chips per wafer. Given a particular defect environment, the yield Y can be related to active chip area A and defect density D (defects per unit area) by the expression:

$$Y = \left(\frac{1-e^{-AD}}{AD}\right)^2$$

The steep rise in the curves of Fig. 3, which relates device cost to yield Y, explains why the sizes of chips at the present time seldom exceed 25 to 30 mm² in even the most carefully controlled processes. Hence, major efforts are being made to make devices smaller so more can be placed on chips of limited size.

There are additional advantages to reducing device size. Scaled-down devices have less of the area-related capacitance that slows switching times. With more devices on a chip, more of the interconnections occur on the chip rather than between chips, resulting in better system speed and less power loss.

Why NMOS?

MOS devices exist in two basic types. In one, NMOS, the source and drain are doped n-type and the channel between them is p-type. A positive voltage applied to the gate causes the channel to turn n-type, allowing conduction between source and drain. In the other type, PMOS, the source and drain are p-type and the channel is n-type. A negative voltage applied to the gate then enables conduction between source and drain.

NMOS devices are faster because conduction is

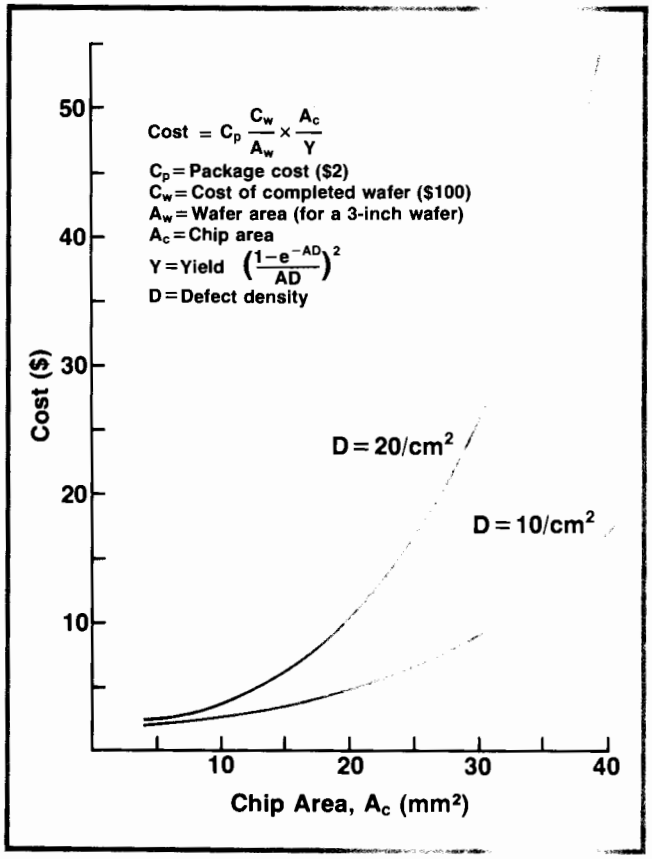


Fig. 3. *Curves plot typical device cost as a function of chip area A for two values of defect density D, showing why very large IC chips are uneconomical. These curves are for a 3-inch wafer with arbitrarily selected values of package and processed wafer costs.*

predominately by electrons whereas conduction in PMOS devices is by holes, which are less mobile. However, NMOS devices have been harder to make. This is largely because mobile ions from contaminants, principally sodium, migrate to the silicon-dioxide/silicon interfaces and tend to invert them (induce turn-on), giving rise to leakage currents.

In the late 1960's, a number of manufacturers were developing MOS memories to meet the anticipated demand that the substantial size and cost advantages of solid-state memories would generate, but none of the available devices met the particular needs of a new family of desktop computers being conceptualized at HP. Thus, the development of an NMOS process was undertaken at HP's Loveland, Colorado, facility for the purpose of manufacturing read-only memories (ROMs) for this family, the HP 9810A/20A/30A series. It was believed that the performance benefits of NMOS justified the additional process complexity.

The problem of contamination-caused leakage currents was largely reduced by the use of "back-gate" bias, a voltage applied to the substrate itself to counteract the potential caused by contaminating ions. As a result, HP was one of the first, if not the first, manufacturer to produce NMOS devices in quantity. The 4K ROMs developed in this facility enabled the 9810A/20A/30A family of desktop computers to have far greater capabilities than any similar products available at the time.

The Next Generation

The central processor in the 9810A/20A/30A machines included a number of off-the-shelf ICs occupying a good-sized printed-circuit board. Looking beyond the 9810A/20A/30A family, it was realized that if all this circuitry could be integrated on a large scale, the next generation machines could have a significant increase in processing speed and capability with a concurrent reduction in cost and size.

Using the experience gained with this first NMOS process, it was decided to develop a new NMOS LSI process that could enable integration of the central processor. The initial objectives were straightforward: twice the density of the current process with at least three times faster operating speeds. These objectives are complementary to a degree since small size means lower capacitance, the principal limit on speed.

The smallest feature size in the process at that time was 7μm. Therefore, to achieve a 50% reduction in device area, dimensions needed to be reduced to about 5 μm. This would also halve the capacitance area.

Obtaining a 30% reduction in linear dimensions was not a simple matter. As active portions of devices are brought closer together, undesired effects quickly arise. Two of these are the "short-channel" effect, in which gate-voltage control of the channel conductivity is degraded, and "punch-through", where conduction occurs between two depletion layers. Compounding these problems is the need to maintain more accurate dimensional control so the same allowable degree of dimensional error can be maintained. Tight tolerances must be obtained at each step in the process, otherwise there can be a surprisingly broad range of device parameters in finished chips, and since the designer must make allowances for the worst case, he sacrifices speed and efficiency if tolerances are too loose.

For the above reasons, and others, new and more controllable technologies had to be employed. The process that ultimately emerged, known within the company as NMOS II, makes use of the following

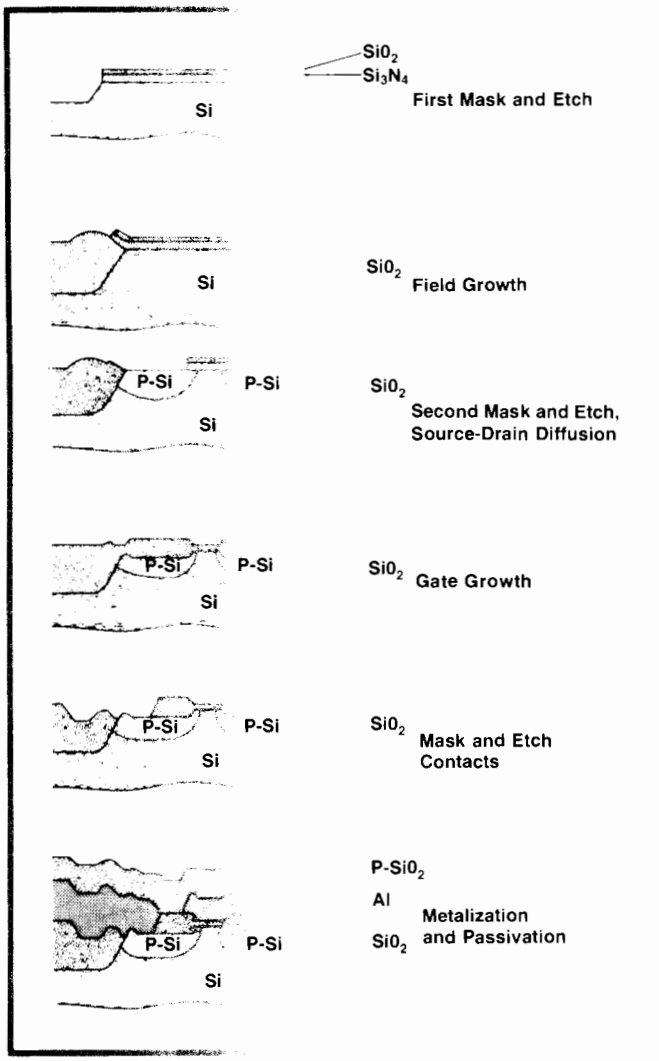


Fig. 4. *Drawings of magnified cross-sections of a silicon wafer show steps in the NMOS II process. The vertical scale is exaggerated for clarity; the depth of the phosphorous diffusion (P-Si) is actually less than 2 μm.*

technologies.

Local Oxidation of Silicon. Rather than oxidize the entire wafer and then etch away the oxide where the devices are to be formed, this technique first covers the wafer with a layer of silicon nitride and then removes it everywhere except where the devices are to be formed (Fig. 4). Silicon-dioxide is then grown on the exposed areas while the silicon nitride prevents the underlying areas from oxidizing. As shown by the cross-sectional drawing of Fig. 4, the growth of the oxide into the wafer forms partial sidewall containment for subsequent diffusions (Fig. 5), providing a margin against punch-through between adjacent diffusions, and it gives more gentle steps on the surface for metal passover. Furthermore, since the first etch is made through a thin layer of silicon nitride rather than through a relatively thick SiO2 layer, better definition of device features is achieved.

Ion implantation (I^2). This technique employs a particle accelerator to implant dopant ions into the silicon in a highly controlled manner (solid-state diffusions, such as that used for the source and drain, require temperatures around 1000°C, which can complicate a process by causing more than the desired effect). I^2 is used in NMOS II primarily for doping the gates of selected FETs so they are lightly depleted. These devices are then used as load resistors with the advantage that they are much smaller than resistors of equal value created by diffusion of a long, narrow area.

Self-aligned gates. This technique minimizes overlap of the gate structure onto the source or drain diffusion areas. Overlap results in a small gate-to-source or -drain capacitance that is magnified by the Miller effect, slowing switching speeds.

A number of techniques are used for self-alignment but a highly effective one was developed for the NMOS II process. It is based on the fact that under the proper conditions phosphorous-doped silicon oxidizes several times more rapidly than undoped silicon. After the source-drain diffusion, the protecting silicon nitride in the gate region is etched away, then the gate oxide is grown. The source-drain oxide forms simultaneously but at about five times the rate. The gate oxide is then centered perfectly between source and drain (see Fig. 4). Subsequent gate-metal placement is relatively uncritical. Furthermore, the feature size can be made smaller since less allowance is needed for aligning the gate to the source and drain.

Hard-surface masks. Until the time that the NMOS II process was being developed, the masks used to transfer patterns to the wafer were made of photoemulsion on glass. These were not well suited to the NMOS II process for two reasons: (1) the edge definition is not sharp enough for the 5 $\pm$0.5 μm line definition wanted, and (2) repeated clamping and aligning of the masks to the wafers soon damages the emulsion, leading to the generation of defects. Hard-surface masks are not so easily damaged. This was a brand-new technology at the time, so many difficulties were encountered in obtaining the required quality for the NMOS II process. Materials successfully used have been thin films of silicon and of iron oxide on glass.

Cu-Si-Al metalization. Aluminum is the standard material for making surface interconnections, one reason being that it alloys with silicon, providing good, low-resistance connections. On the other hand, with the shallow doping used in the NMOS II process, this alloying can extend through the doped silicon, shorting the contact to the substrate. Using silicon-saturated aluminum reduces this "alloy-through." The addition of copper to the alloy suppresses crystallization, smoothing the silicon-aluminum alloy. The smoothed alloy also covers steps more uniformly.

The Major Problems

The "alloy-through" was one of the most troublesome problems that had to be overcome. The boiling points of silicon and aluminum are quite different, making composition control difficult when using the evaporative technique of depositing thin-film conductors. Conventional sputtering techniques, while allowing easy composition control, are too slow and raise the temperature of the wafers to around 300°C, causing further problems with separation and crystallization of the alloy materials.

A new technique, commonly referred to as planar-magnetron sputtering, was developed. In P-M sputtering, a magnetic field in the plasma region causes electrons to follow spiral paths which results in more efficient ionization. Consequently, there is a large improvement in both deposition rate and a reduction in the substrate heating that was caused by secondary electron currents. P-M sputtering is now generally used throughout the industry.

Another troublesome area involved the problem of leakage. The introduction of phosphorous into silicon dioxide can be a good way to solve this problem because phosphorous is a rather effective "getter" for

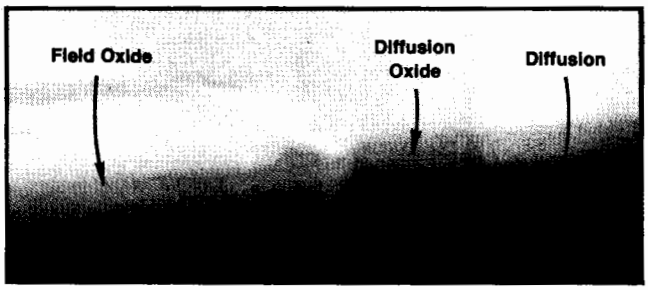
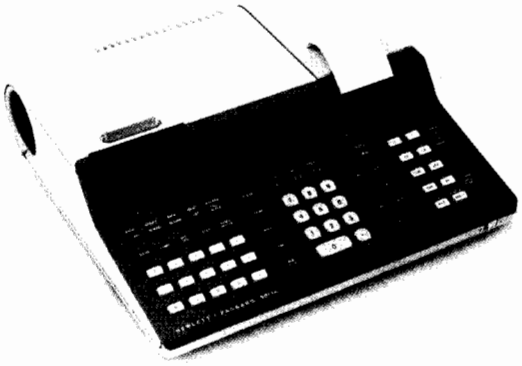


Fig. 5. *Scanning electron micrograph of NMOS II IC shows sidewall containment of diffusion (magnification: 8800×).*

Applications of the NMOS-II Process

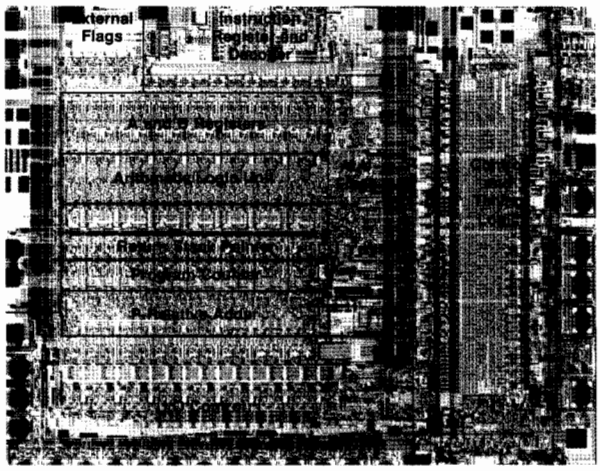
The first use of the NMOS II process described in the main text was for mask-programmable 16K ROMs used in the Model 9815A Desktop Computer[1] (DTC), introduced in September 1975. These ROMs are organized in a 2K × 8 structure for operation with a commercially available 8-bit microprocessor. In addition to providing four times as much firmware as the 4K ROMs used in earlier DTCs, the NMOS II ROMs decreased power consumption by turning themselves off when not being accessed by the microprocessor. This dynamic "power-pulse" operating mode decreased ROM power dissipation by a factor of greater than 10 without affecting access time significantly.



Model 9815A

The increased storage capability provided the Model 9815A by the NMOS II ROMs enabled the standard machine to have as part of its standard language many of the math functions and other options that were available only as add-on ROMs for the previous generation DTC's.

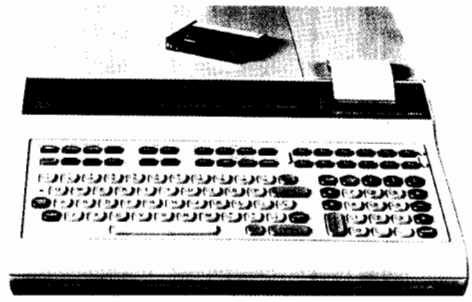
The first use of an NMOS II microprocessor was in the Model 9871A Impact Printer,[2] also introduced in 1975. Use of ROM-controlled processor logic gave the flexibility needed to improve performance at low cost. By enabling the printer to perform its internal functions itself, the processor also reduced the amount of communications needed between the DTC and printer.



Layout of BPC chip

This processor, known as the Binary Processor Chip (BPC), is a 16-bit parallel processor capable of executing 59 unique instructions. It has a memory address space of 32K 16-bit words. With NMOS II 16K ROMs organized in a 1K × 16-bit structure, the BPC forms a controlling processor that operates with a 6-MHz clock supplied by an external 2-phase, non-overlapping clock source. The processor controls the acceleration, speed, and deceleration of the motors that position the print wheel and it determines the hammer force according to the density of the character selected. Many features, such as self-test, automatic tabulation, plotting, and character substitution, were added at little cost simply by including them in the firmware.

The first totally NMOS II LSI-based desktop computer was the Model 9825A[3] introduced in January 1976. The heart of this DTC



Model 9825A

is the seven-chip hybrid microprocessor mentioned in the main text. Three of the chips are processors (a BPC, an IOC chip, and an extended math chip) that among them execute 86 instructions including control of all communications between the DTC and peripherals via the I/O port. The three processor chips are isolated from the electronics in the rest of the DTC by four bipolar interface buffers, minimizing the capacitance that the processor chips must drive. This allows the processor to operate at clock rates above 6 MHz.

Firmware is contained in 16K ROMs, organized 1K × 16. Besides saving energy with the power-pulsing (automatic self turn-off) feature, these ROMs have "three-state" (input, output, off) pins that allow multiplexing of addresses and data. The three-state arrangement reduces the overall capacitance on the parallel address-data bus by cycling non-addressed ROMs into a high impedance (low capacitance) state.

In addition to the processor and 16K ROMs, NMOS II is used in the Model 9825A in a random-logic chip that controls the operation of the keyboard, LED display, and thermal printer. This chip controls all operations of these "internal peripherals" and the data communications between them and the microprocessor. It generates key scans, allows for key debounce delay, generates repeat-key timing, and communicates key-code data to the microprocessor. It shifts and clocks data to the printer registers and controls paper advance and the burn timing of the thermal printhead. It shifts display dot character information to the LED display, controls display scan rate, and

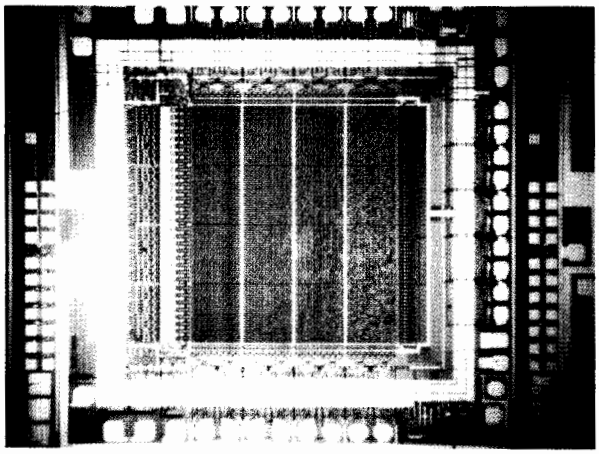
generates the display cursor. The character set used by the display and printer is contained in a mask-programmable ROM on this chip, enabling alternative character sets, such as Katakana, to be generated simply by changing photomasks during chip fabrication.

The most recent use of NMOS II LSI is in the recently announced 9845A Desktop Computer (to be described in a forthcoming issue of the HP Journal). The processor in this DTC was expanded to include two hybrid microprocessors of the type used in the Model 9825A. One processor executes the language processing while the other is dedicated to communications between the mainframe electronics and peripherals. A new BPC chip was designed for these microprocessors with four times the addressing space of the original BPC chip. The new BPC chip also makes use of NMOS power transistors for the processor's clock drivers, accepting a TTL-level clock input and generating a 12V, two-phase non-overlapping clock capable of charging 300 pF to above 10V in less than 30 ns.

The firmware memory of the 9845A has 16K ROMs on 8-chip hybrid substrates. Each of these ROMs has its own power-pulse transistor on the chip rather than using external bipolar transistors for this function as the earlier NMOS II ROMs did. Each of the memory hybrids is isolated from the mainframe electronics by bipolar interface buffers to reduce the capacitance the ROMs must drive, resulting in improved speed of operation.

Another random-logic NMOS II chip controls the operation of the minicartridge tape drive and data transfer between mainframe electronics and the tape, performing tape read-and-write control and detecting interrecord gaps. Another NMOS II chip, taking the place of a previously used bipolar chip, shifts and stores the data to the 80-character, page-width printer-plotter, directly driving the thin-film thermal printhead thin-film resistors with NMOS power transistors that are capable of sinking 150 mA with less than 6 ohms ON resistance.

Thus, it can be seen that the NMOS II process is not limited to memory and microprocessor applications but is branching out to encompass the whole array of desktop computer circuits, giving the designers of desktop computers more capability and greater flexibility in less space for enhanced performance.



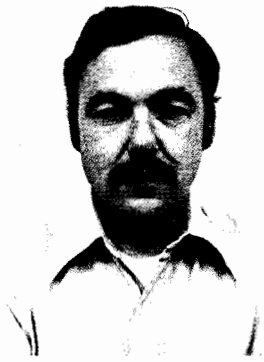
Microphotograph of NMOS II 16K ROM.

References

1. D.M. Clifford, F.T. Hickenlooper, and A.C. Mortensen, "Mid-Range Calculator Delivers More Power at Lower Cost," Hewlett-Packard Journal, June 1976.
2. R.B. Bump and G.R. Paulson, "Character Impact Printer Offers Maximum Printing Flexibility," Hewlett-Packard Journal, June 1976.
3. D.E. Morris, C.J. Christopher, G.W. Chance, and D.B. Barney, "Third Generation Programmable Calculator Has Computer-Like Capabilities," Hewlett-Packard Journal, June 1976.

sodium ions. However, the procedures used to achieve other desired features, such as the self-aligned gate, did not allow the easy introduction of sufficient phosphorous into the field oxide. The solution was to dope an oxide layer with phosphorous while the layer is deposited chemically at 400°C as the final passivation layer at the end of the process.

With the more sharply defined features, the shallower dopings, the self-alignment technique, and the use of depletion load transistors, plus refinements in circuit design, circuit speed improved by a factor of 10 with the factor-of-2 area reduction. During the two and a half years the NMOS II process has been on line, it has proved to be relatively high yielding. Besides providing the hardware basis for the 9815A/25A/45A desktop computer family and its many high-performance peripherals, it is also finding its way into other kinds of products, such as the new Model



Joseph E. DeWeese
Joe DeWeese has devoted five years to the NMOS-II process, beginning in 1972 with development of prototype process equipment and going on to development of products, production equipment and process monitors. The past two years, he was project manager for NMOS II LSI product support. He is now a project manager in the R and D Lab. A native of Clinton, Kentucky, Joe earned a BSEE degree at the University of Illinois and is working towards an MSEE degree at Colorado State University. In off hours, he does some woodworking and he also participates in the HP Golf League and intramural basketball. He and his wife have one small boy and a baby girl.

Thomas R. Ligon
After obtaining BS and MS degrees in physics from Oklahoma State University in 1966, Tom Ligon joined Hewlett-Packard, working initially on the development of precision high-frequency wirewound resistors, then the 17-layer printed-circuit ROM used in the Model 9100A Calculator, and finally integrated circuit development. He left HP in 1970 to pursue a private business venture but returned two years later to work on the NMOS-II process development, becoming project manager in 1973. A native of Wewoka, Oklahoma, Tom, his wife and three children live on a 1½ acre tract in Loveland, Colorado, raising fruits, vegetables, chickens, rabbits, and an occasional calf or pig.

8568A Spectrum Analyzer.

Acknowledgments

Major contributors to the NMOS II process were Larry Hall, Jim Mikkelson, Dana Seccombe, and Mark Lundstrom. In the LSI circuit development effort, Gene Zellmer and Howard Abraham were early notables while IC engineering manager Tom Haswell guided the overall effort. Many other people, too numerous to mention here, contributed significantly to the success of the process and products.

References

1. B.E. Forbes, "Silicon-on-Sapphire Technology Produces High-Speed Single-Chip Processor," Hewlett-Packard Journal, April 1977.
2. W.D. Eads and D.S. Maitland, "High-Performance NMOS LSI Processor," Hewlett-Packard Journal, June 1976.

(continued from page 25)

connector have not yet arrived.

Standardization is also a prerequisite of high-volume production. Because of low-volume developmental manufacture, the present costs of fiber, cable, solid-state lasers and LEDs are an order of magnitude higher than that considered economically feasible. It will take one or two years yet before the "gain" in the volume-cost-demand "loop" becomes great enough to create a steep increase in production and a drop in cost, as occurred with integrated circuits.

Then there is the usual reluctance of reliability-minded engineers to abandon highly developed, well proven systems for a totally new technology, with the active light source in the transmitter being particularly suspect. Material defects appear to be the major limiting factor on the life of LEDs and lasers. The specific failure mechanisms are being studied in a number of laboratories throughout the world with assurance of many solutions. Recent reports from Bell Laboratories cite an extrapolated life of up to 10^6 hours for a solid-state laser, corresponding to an uninterrupted service of about 100 years. Even a 10-year life span, more commonly achieved today, should be quite sufficient for reliable operation.

The Future

No doubt about it, wide-scale use of optical communications is coming, not only for telephone transmissions, cable TV, and distributed computer networks, but also for such short-haul applications as pc-board-to-pc-board signal transfer within an instrument. Most transmitters and receivers now are designed for specific bit rates and link lengths. Large-scale integration, and possibly trade-offs in receiver sensitivity and maximum bit rates, will result in compact circuits that will be able to accommodate a wide range of link lengths and bit rates without requiring any adjustments. This will facilitate the economical application of fiber-optics to a wide range of applications.

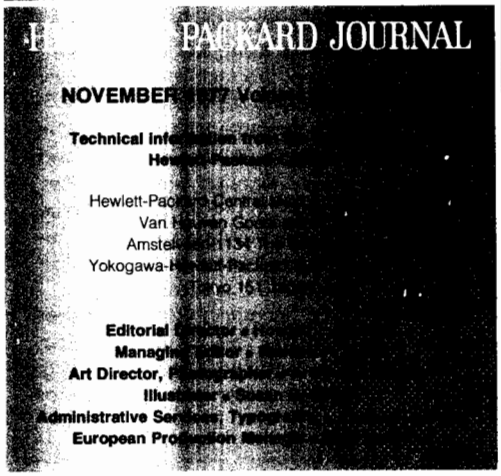
In the immediate future, we can see further developments in splices and connectors. To avoid undue loss of power, the faces of the adjoining fibers must be of equal size, in line, parallel, and as close to each other as possible. This leads to strict diameter control in fiber production and, given the small diameter of the fiber, requires a precision in the micrometre range for the connectors and splices. At present, the optical power loss in a typical connector is 0.5 to 2 dB, equivalent to a substantial length of low-loss fiber cable. The use of connectors or splices in a system therefore must be carefully planned, quite a different situation than that encountered with communications over metallic wire.

Farther in the future we anticipate developments in single-mode fibers and also in integrated optics, the optical equivalent of semiconductor ICs. These will consist of solid-state components, usually based on thin-film technology, in which optical rather than electrical signals are switched, amplified, modulated, and processed directly without conversion to electrical signals. The impact of this technique will be far-reaching.

Tom Hornak is Manager of the Applications and Engineering Department within the Solid-State Lab of HP's central research laboratories where he is responsible for R and D on LSI circuits and optoelectronic systems. He holds a Dipl. Ing. degree from the Slovak Technical University and a PhD degree from Czech Technical University and worked on radar, instruments and memories before joining Hewlett-Packard in 1968. He has published over two dozen papers on electronics and holds a like number of patents.

Hewlett-Packard Company, 1501 Page Mill Road, Palo Alto, California 94304

Bulk Rate
U.S. Postage
Paid
Hewlett-Packard
Company



C E ADDRESS: To change your address or delete your name from our mailing list please send us your old address label (it peels off). Send changes to Hewlett-Packard Journal, 1501 Page Mill Road, Palo Alto, California 94304 U.S.A. Allow 60 days.