R. H. Collins
E. G. Grochowski
W. D. North

Silicon Process Technology for Monolithic Memory

Abstract: The successful design and production of high density, high speed monolithic memory circuits that can operate efficiently over wide current ranges can be directly attributed to the development of a new silicon process technology based on the ⟨100⟩ crystal orientation. To provide optimum capability to fabricate these circuits with high yields, process modifications to accommodate larger starting substrates (wafers) have also been adopted. Detailed electrical characterization of the resulting devices has confirmed improved circuit operation, demonstrating that improved gain at low current and good junction quality are consequences of the process modifications.

Introduction

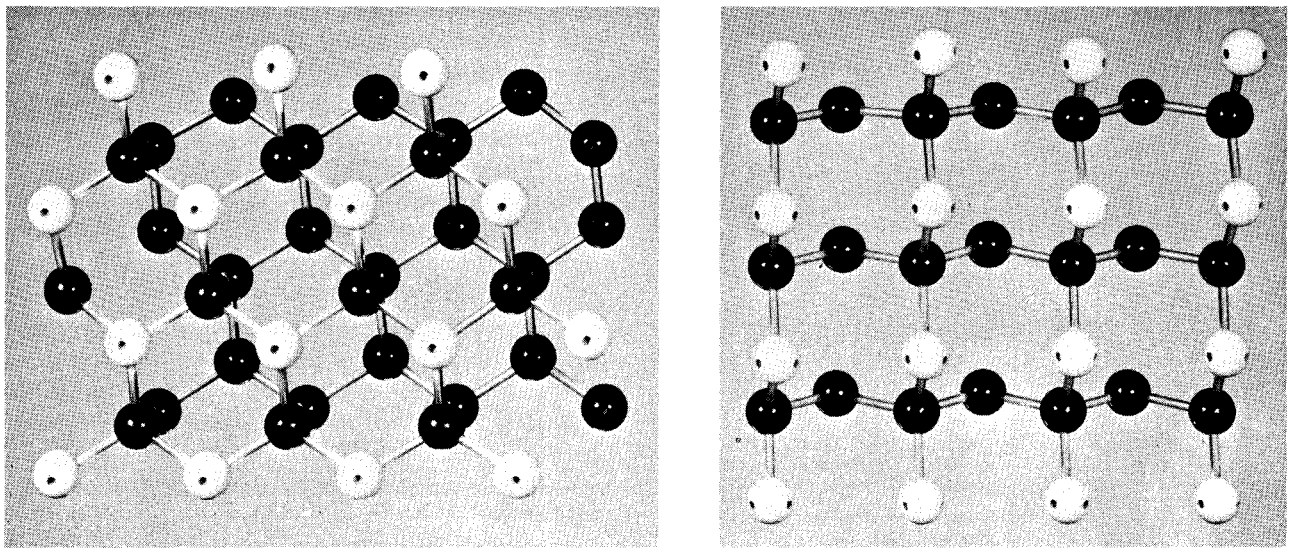
Ferrite core technology has predominated in the past in providing main-frame random access memories for computer applications. The new applications of bipolar integrated circuits for use in a high-speed buffer or scratchpad memory have clearly demonstrated superior performance, as well as optimum bit density, for those types of memory. The IBM announcement utilizing a semiconductor-based memory for main-frame applications in which large numbers of circuits are required was made feasible by developing a semiconductor process which yielded optimum performance at reduced cost.

Late in the 1960's, IBM undertook a major program to investigate the then current semiconductor technology, improve its capabilities and develop new processes and structures as required[1]. The increasing demand for improved circuit performance, particularly for monolithic memory applications, uncovered both material and processing limitations that were prevalent with ⟨111⟩ oriented silicon as the starting substrate material. Historically, the ⟨111⟩ orientation has been used for semiconductor processing since the development of alloy junction devices, and this technology had formed the basis for more recent structures. However, the achievement of high bit densities by microminiaturization required increased active-device pattern density and tighter tolerances, thereby requiring a larger percentage of the surface area on a given silicon substrate. This greater density increased the probability of including a crystallographic defect in a critical junction area and consequently demanded silicon wafers relatively free of imperfections.

Processes therefore had to be refined to produce silicon surfaces and interfaces with minimal contaminant levels, both on the semiconductor and in the passivating oxide, and to achieve adequate gettering in the bulk. Limitations in transistor gain at low currents (in the range of 1 μA) on ⟨111⟩-fabricated devices indicated that charges arising from surface states affect low current, low power operation. The ⟨100⟩ orientation in silicon exhibits a low surface state density[2], consequently offering a wider range of operating currents. This advantage, combined with improved crystallographic perfection, warranted the specific investigation of a semiconductor process based on the ⟨100⟩ orientation.

It is also important, because of the need for cost reduction, to minimize the number of individual handling operations per circuit by increasing the crystalline substrate (wafer) diameter. With advanced crystal pulling techniques, it has become possible to produce 2.25-inch diameter silicon wafers while maintaining perfection equivalent to that previously achieved with 1.25-inch crystals. However, subsequent processing, particularly high temperature heat cycles, had to be significantly modified to process wafers after what might be considered a minor substrate dimensional change.

The following is a description of the salient features of this new silicon process technology, emphasizing: 1) the ⟨100⟩ orientation, specifically the orientation effects on device structure that are pertinent to monolithic memory; 2) the processing of ⟨100⟩-oriented 2.25-inch wafers, with attention to diffusion and oxidation, gold diffusion,



⟨111⟩ ⟨100⟩

Figure 1 Models of ⟨111⟩ and ⟨100⟩ silicon crystallographic lattices.

and the use of metallurgical contacts; and 3) product characterization, specifically the identification of device parameters relating to ⟨100⟩ material properties and the characteristic yields encountered.

Materials

• *Orientation*

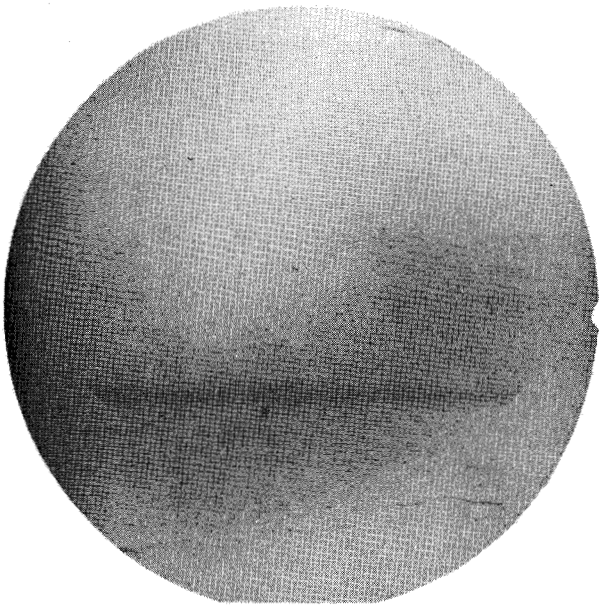
Silicon monocrystals for monolithic devices can be grown from the melt in many crystallographic directions. Due to the anisotropy of the silicon lattice, these crystals and the surfaces of substrates cut from them retain certain chemical, mechanical, and electrical characteristics that are orientation-dependent. Of the three major orientations with Miller indices ⟨111⟩, ⟨100⟩, and ⟨110⟩, semiconductor devices and circuits have historically been fabricated only from ⟨111⟩ material. This practice originated as a requirement for alloy-junction device manufacture, since it was observed that planar, controllable junction profiles in that technology could only be realized with ⟨111⟩ oriented substrates. A particular material technology was thus experimentally developed, later forming the basis for more complex device structures. With that technology well developed, investigation of processes based on other orientations was not thought justifiable until recently.

Figure 1 shows a comparison between models of the ⟨100⟩ and ⟨111⟩ orientations of the diamond lattice that characterizes Group IV-A elements, including silicon. The atomic density for a ⟨111⟩ surface plane is maximum for this type of lattice and greater than that for the ⟨100⟩ plane by more than a factor of two.

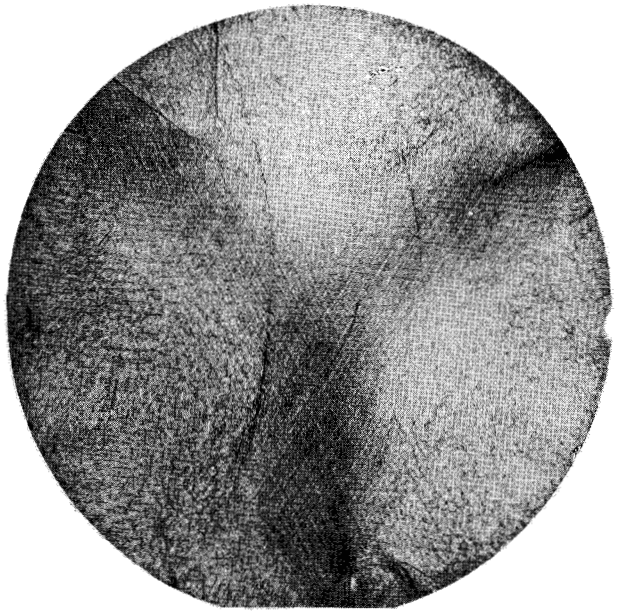
The most striking characteristic of silicon technology is that material processes and the resulting semiconductor devices can be drastically different from one another depending on the orientation of the starting crystal substrate. As a consequence, almost all semiconductor technology is governed by surface-controlled conditions; anisotropic behavior in the bulk is generally the direct result of varying atomic densities on the semiconductor surface.

The following description of orientation-controlled phenomena pertinent to ⟨100⟩ and ⟨111⟩ silicon is divided into three areas: 1) bulk silicon and substrate effects; 2) epitaxy, diffusion and oxidation; and 3) metallurgical contacts. Electrical characterization of ⟨100⟩ material structures is considered both in the orientation discussion and later in the section on device performance as applied to monolithic memory.

The minimization of both bulk-grown and process-induced crystallographic defects, which is vital for high-density monolithic circuits, is an inherent property of ⟨100⟩ silicon substrates. The observation of large numbers of crystalline substrates, sliced from crystals grown parallel to both the ⟨111⟩ and the ⟨100⟩ orientation by similar Czochralski techniques, has determined that crystallographic defect density as determined by both Sirtl etching and SOT (scanning oscillator technique) X-ray photography is much lower with the ⟨100⟩ orientation[3]. Figure 2 is a typical SOT display of this effect. The ⟨111⟩ crystals are characterized by linear arrays of diminishing density near the center of the wafer. Consequently, epitaxial films that are grown on these substrates and in which the circuits are fabricated exhibit a similar



⟨100⟩



⟨111⟩

Figure 2 SOT topographs of as-grown ⟨111⟩ and ⟨100⟩ crystals.

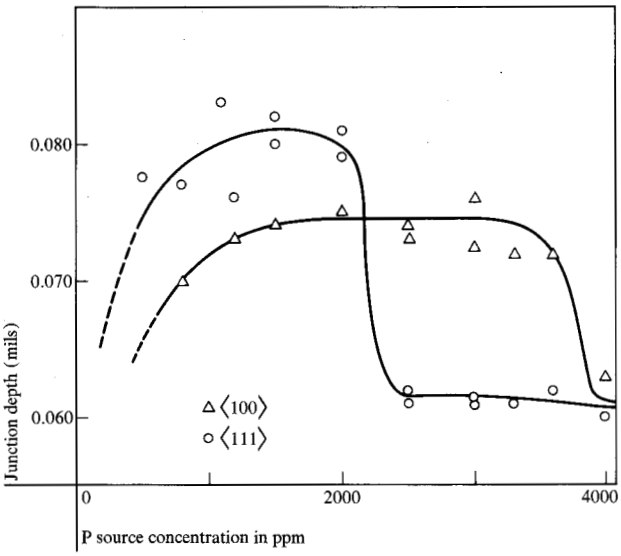
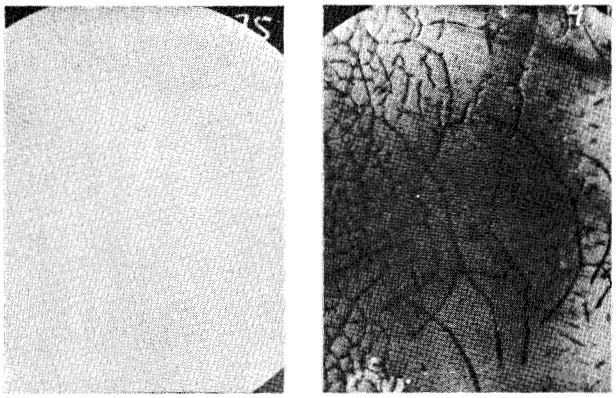


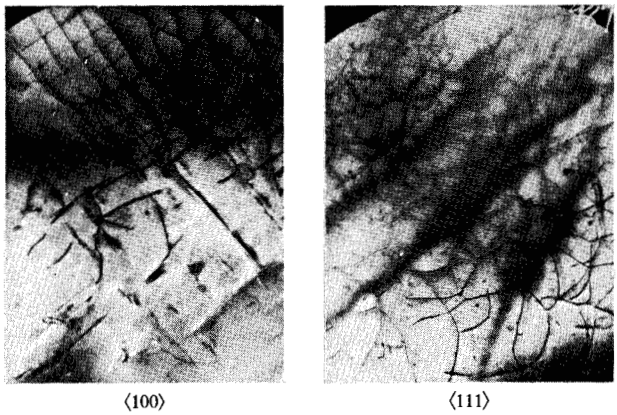
Figure 3 x_j vs P source concentration for ⟨111⟩ and ⟨100⟩ silicon.

Figure 4 Electron micrographs showing dislocation density for ⟨111⟩ and ⟨100⟩ materials at two different P source concentrations: (a) Source concentration 2500 ppm, (b) 4000 ppm.



⟨100⟩ ⟨111⟩

(a)



⟨100⟩ ⟨111⟩

(b)

defect density. This density is characteristically 10 to 1000 times larger for ⟨111⟩-oriented silicon epitaxial films than that measured for ⟨100⟩ material receiving equivalent processing.

Electrically active junctions are formed by diffusing Group III or Group V elements into the epitaxial layer; silicon does not exhibit anisotropic diffusivities for these. However, in diffusion processes in which surface characteristics control dopant concentration in the bulk, par-

ticularly by redistribution, differences between ⟨100⟩- and ⟨111⟩-oriented material are evident. For monolithic memory devices, the emitter diffusion should retain the highest dopant surface concentration for maximum injection efficiency. The emitter efficiency γ is conventionally defined as the ratio of injected-carrier current to the total emitter current[4]:

$$\gamma = I_{nE}/(I_{nE} + I_{pE}) = 1/[1 + (I_{pE}/I_{nE})]$$

for an npn device, where I_{nE} and I_{pE} are the current components due to electrons and holes respectively in the emitter region. Moreover, there is a proportionality between current and resistivity,

$$I_{pE}/I_{nE} \propto \rho_E/\rho_B.$$

The emitter diffusion is a two-step process comprising a deposition cycle, performed in an open-tube flow system using a controlled-concentration phosphorous source, followed by a reoxidation, drive-in cycle. In this process, phosphorous surface concentration depends on source concentration and is very close to the solubility limit for phosphorous in silicon; i.e., from 5×10^{20} to 10^{21} atoms/cm^3. These levels often promote diffusion effects characterized by precipitation formation, dislocations, localized junction pull-back and steep concentration profiles[5].

Determination and control of the diffusion depth into the silicon (junction depth) directly affects both base width and transistor gain. Figure 3 shows junction depth after drive-in as a function of source concentration for ⟨100⟩- and ⟨111⟩-oriented silicon. In each curve, the region of rapid fall-off indicates the point at which source concentration is high enough to cause drawback of phosphorus from the formation; this phase occurs at higher concentration in ⟨100⟩-oriented silicon (4000 ppm) than in the ⟨111⟩ (2500 ppm). Figure 4 is a series of transmission electron micrographs showing the two orientations at both 2500 and 4000 ppm source concentration. The photographs cover a range of depths into the silicon and show arrays of dislocations as regular squares for the ⟨100⟩ and triangles for the ⟨111⟩, as well as shallow precipitation sites as dark, irregular areas. The ⟨100⟩ has a greater tolerance for phosphorus at higher concentrations, allowing high injection efficiencies and good control of base width.

Surface concentration of active phosphorus, C_0, and diffused junction depth, x_j, are not related by a classical distribution function (a complementary error or an exponential function) at levels above 10^{20} atoms/cm^3. After using IR plasma resonance for non-destructive C_0 determinations, a comparison of diffused junction depths can be made from Figs. 5 and 6. Figure 5 relates surface concentration of electrically active phosphorous, which upon ionization yields a minimum in the reflection spectrum by absorption, to source concentration in the diffusion furnace. Figure 6 shows junction depth against the same C_0 values. Both ⟨111⟩ and ⟨100⟩ crystal orientations exhibit minimal deviations in depth up to near 1100 ppm. Above this region, the percentage of ionized donors decreases as indicated by the maximum at 4 ×

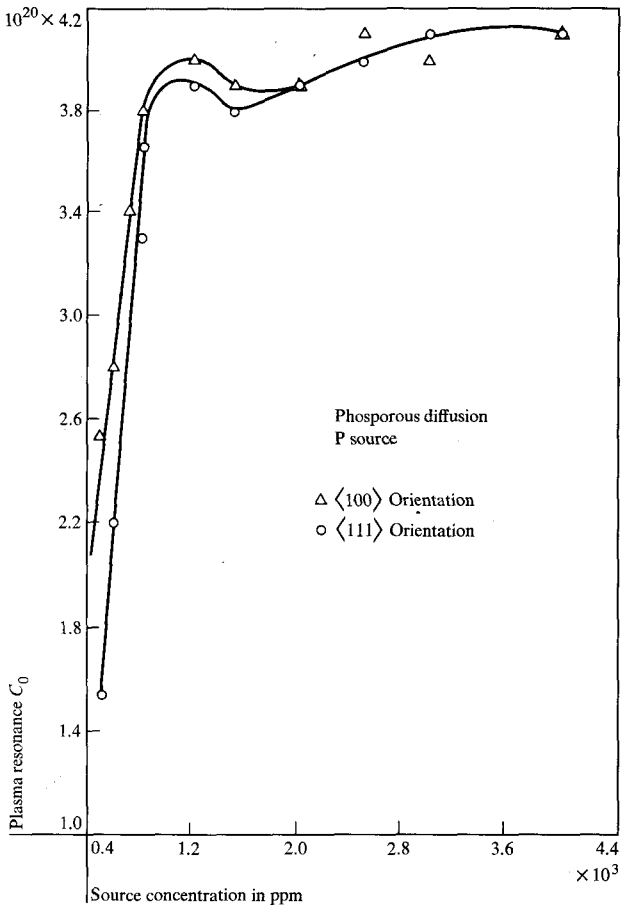
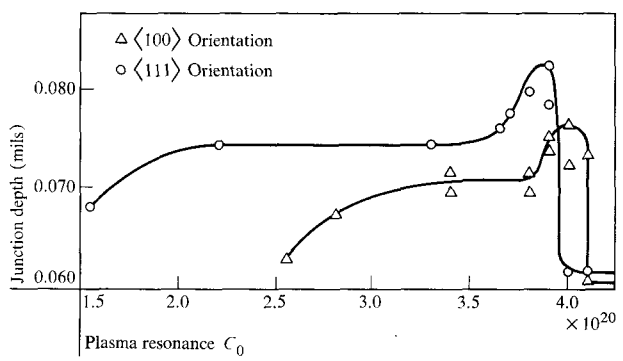


Figure 5 Surface concentration vs source concentration for both orientations.

Figure 6 Junction depth vs surface concentration for both orientations.



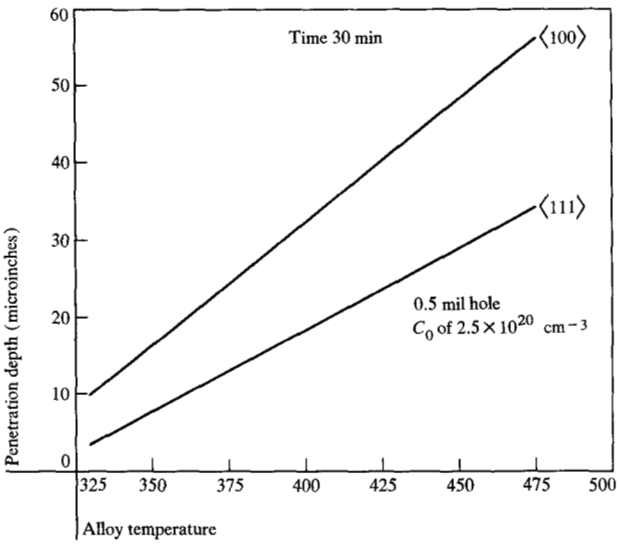


Figure 7 Alloy penetration depth vs temperature for aluminum contact metallization.

10^{20} phosphorous atoms/cm^3. The onset of precipitate formation with ⟨111⟩ material occurs here and reaches a maximum value between 1500 and 2000 ppm. The ⟨100⟩-oriented material, with a greater tolerance for surface phosphorus, reaches a maximum at 3000 ppm.

• *Metallurgical contact*

Low-resistance ohmic contacts are necessary to minimize voltage drops in monolithic arrays involving large numbers of devices. It has been a standard procedure to follow the evaporation of the contact metallurgy, usually aluminum, with a heating cycle at an elevated temperature to alloy a small portion of this metal into the body of the semiconductor. The mechanism for the motion of the alloy interface involves formation of an aluminum-silicon solid solution, dissolution of additional semiconductor, and replenishment of the solution with additional metal from the evaporated film[6].

At the alloying temperature, the solid phase dissolution is equivalent to a mild, preferential etching operation that removes all trace insulating oxides and organics between metal and silicon, effecting a low-resistance contact. Temperatures near 450°C have been found optimum for ⟨111⟩-oriented silicon. With identical processing the ⟨100⟩ orientation disclosed augmented alloying rates. Figure 7 compares maximum penetration depth as a function of alloy temperature for the two orientations. The morphology of ⟨100⟩ penetration suggests that the angles at which minor ⟨111⟩ planes intersect the surface plane control junction planarity and depth. This specific alloying behavior in the ⟨100⟩ orientation was, in the past, detrimental to controlling the base width of alloy junction devices. However, Fig. 7 indicates that an approximate 10% reduction in alloying temperature will yield depths for the ⟨100⟩ that are equivalent to those for the ⟨111⟩. This modification to lower alloy temperature has produced low-resistance aluminum contacts on ⟨100⟩ silicon while minimizing deep alloying.

Large wafer processing

• *Diffusion and oxidation*

The development of a large-wafer process for monolithic memory manufacture is economically justified by two factors: an increase in circuit capacity per wafer, and the possibility of improved yield due to a relatively lower edge loss. The capacity factor ratio of a 1.25-inch wafer compared to a 2.25-inch wafer is approximately one to four. Truncation losses arising from positioning rectangular circuits on circular wafers will be proportionately less with increasing wafer diameter since circumference increases linearly with radius while area increases as the radius squared. Each individual wafer-handling operation (etching, diffusion, oxidation, masking, evaporation) now simultaneously processes a larger number of circuits. Secondly, edge losses are also realized from non-geometric factors arising purely from processing. All process variations (i.e., junction depths, epitaxial thickness, oxide thickness and pattern fidelity) are proportionately larger in the edge region. Also, crystallographic damage originating from the original growing process, or from subsequent handling and temperature gradients within the wafer, is more prevalent near the peripheral region. If these factors are equivalent for 1.25- and 2.25- inch diameter wafers, optimum processing for maximum yield can be better realized with larger diameter wafers for circuits having areas of 10^4 mil^2.

Improved instrumentation and process equipment has made both production and processing of large-diameter wafers possible. Crystal pullers are now available that have capacity for 3-inch (and greater) diameter crystals of superior perfection, and large-muffle diffusion furnaces are commercially available which normally exhibit radial gradients of less than 0.5°C.

To increase the furnace loading density of wafers for simultaneous processing in diffusion and oxidation, the conventional arrangement is an array of rows of wafers separated by intervals of the order of the wafer thickness. Rapid warm-up and cool-down of the wafers by thermal radiation occurs during normal insertion into the furnace and withdrawal into the room ambient. Radial thermal gradients are produced in the center wafers of each row, and these generate tensile stress during warm up and compressive stress during cooling. Stress relief by means of wafer distortion and crystallographic slip is prevalent, particularly the latter for ⟨111⟩-oriented material. A

⟨111⟩ process using 2.25-inch substrates requires a slow cool after high temperature, open tube diffusion to minimize these thermal gradients and achieve maximum crystallographic perfection and yield near the peripheral region. The ⟨100⟩-oriented wafers also require slow cooling for optimum processing, but to a lesser degree. The mechanism of stress relief by dislocation formation is therefore anisotropic, and wafers sliced parallel to the ⟨100⟩ growth direction exhibit a greater elasticity. However, stress relief by another process (wafer deformation) is still prevalent. Slow cooling minimizes this loss of planarity to a significant degree for both orientations. Large-batch processing of 2.25-inch diameter ⟨100⟩ wafers has been shown to produce circuits having low dislocation density in which both high yield and high performance can be observed in the peripheral region.

• *Mechanical strength and gold diffusion*

To realize optimum processing with 2.25-inch wafers, the batch nature of current semiconductor technology requires adequate mechanical strength for minimum breakage. Flexure of 1.25- and 2.25-inch wafers has been analyzed theoretically. Both a membrane model and a bending-disc model have been applied, and each model has been exercised under both maximum symmetrical and cantilever loading. It can be concluded from these models that the maximum stress generated in a wafer can be approximated by

$$S_{\max} = \frac{3}{8}(3 + \mu)\frac{P_0 R^2}{t^2},$$

where μ is Poisson's ratio; P_0 is the pressure exerted (i.e., the loading); R is the radius of the wafer; and t is the wafer thickness. This expression relates radius and thickness linearly for equal stress and equal loading.

Membrane theory predicts small thicknesses, whereas bending theory requires considerably larger values. The value of 8 mils experimentally adopted for 1.25-inch wafers had previously yielded good strength characteristics with low breakage losses. The value of 15 mils for 2.25-inch wafers determined from this stress equation encompasses both theories and is a median value in each case.

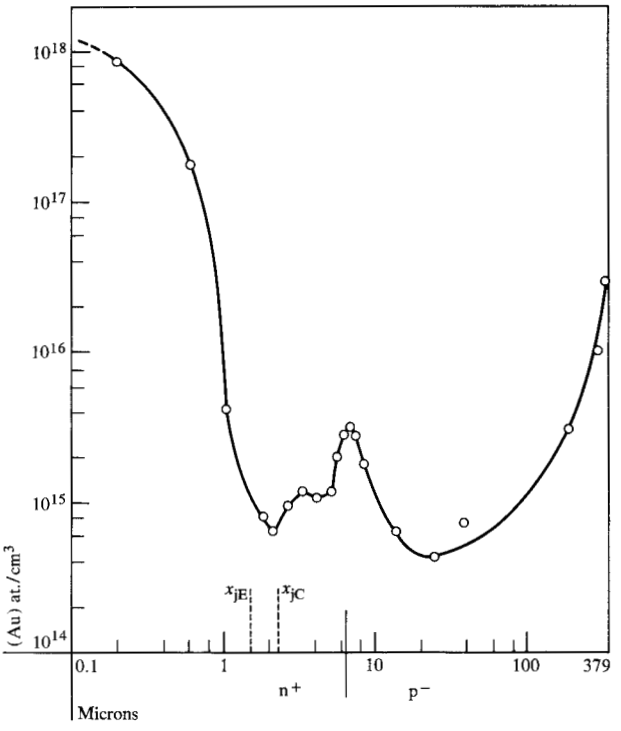
The required change in wafer thickness as well as the change in crystallographic orientation has influenced the diffusion of gold needed to control minority carrier lifetime in the collector region. Conventional small-wafer processing calls for evaporation of a thin gold film on the wafer backside, followed by a high-temperature diffusion cycle in which the gold is diffused through the entire wafer thickness. Wafer thickness thus determines lifetime characteristics if the process is source-limited[7].

Table 1 indicates that such variables as wafer thickness, gold film thickness, and the time-temperature cycle all have an influence on minority carrier lifetime. However, a maximum gold concentration within the transistor structure often limits gain and promotes precipitation at critical junction regions. Figure 8 shows an autoradiogram of a typical in-process gold profile. Regions of high n-type dopant concentrations also contain high gold levels due to a reactive affinity between gold and the dopant. Raising the phosphorous emitter concentration level for maximum efficiency in high speed memory circuits entails reduction of the active gold level within the collector region at a constant gold film thickness.

Table 1 Minority carrier lifetime as a function of wafer thickness and gold-diffusion process parameters.

Wafer thickness (mils)	*Process*	*Minority carrier lifetime (arbitrary units)*
8	30 min @1000°C; 400Å gold	2.43
15	30 min @1000°C; 400Å gold	5.0
15	60 min @1000°C; 400Å gold	2.71
15	120 min @1000°C; 800Å gold	1.0
15	30 min @1025°C; 800Å gold	2.57
15	20 min @1050°C 800Å gold	2.28

Figure 8 Gold concentration vs x_j for ⟨100⟩ devices.



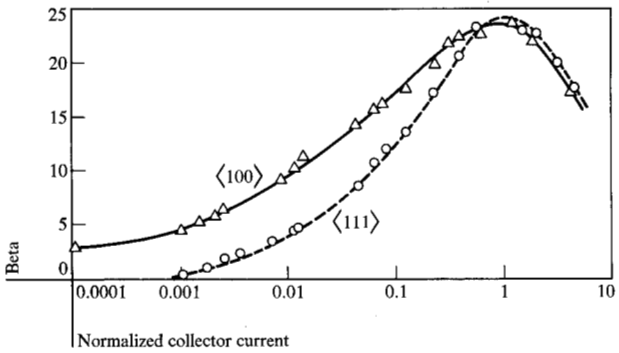


Figure 9 Beta vs normalized collector current for both orientations.

Table 2 Transistor test data.

Electrical parameter	⟨111⟩	⟨100⟩
BV_{EB} @ 10 μA	6.5 mV	7.2 mV
BV_{CB} @ 10 μA	35 mV	35 mV
BV_{CEO} @ 5 mA	12 mV	12 mV
FV_{CB} @ 1 mA	0.78 mV	0.78 mV
FV_{EB} @ 1 mA	0.79 mV	0.79 mV
$V_{CC} = 1$ V		
$\beta(I_C = 3$ mA)	25	25
$\beta(I_C = 100$ nA)	1	5

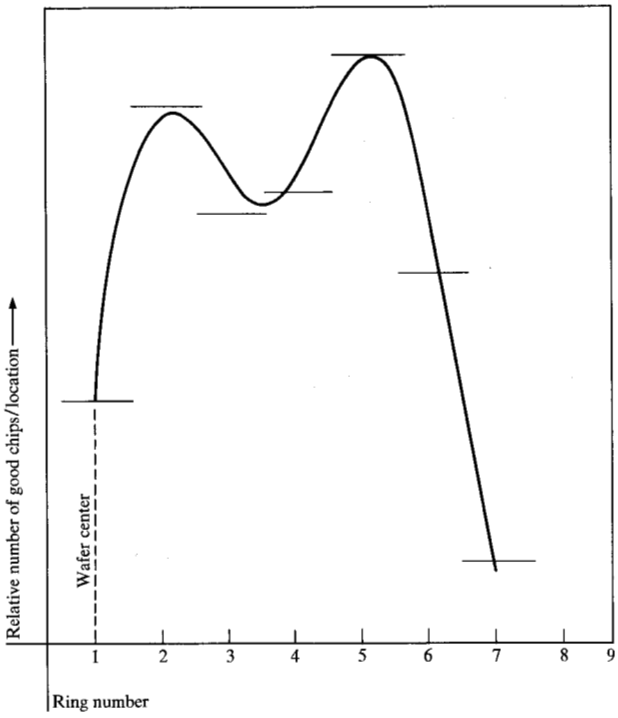


Figure 10 Yield vs distance from wafer center for 2.25-in. wafers of ⟨100⟩ material.

Characterization of processed devices

• *Device parameters*

To characterize and evaluate the processing results for individual lots of monolithic memory circuits, a preferred procedure is to include electrical test sites in the photolithographic mask design, which is also used to fabricate the product circuits on the wafer. The test sites are chip locations that have a special metal pattern that permits certain discrete memory components to be connected to test terminals. In the initial evaluation of an experimental process, large numbers of these sites are required to correlate device and circuit performance and also to optimize process parameters. However, since the test sites consume valuable wafer real estate and subtract from overall productivity, many of them can eventually be eliminated when optimization of the process produces maximum yield. Additional components and special "dumbell" resistor structures have been included in the kerf area (area removed during chip dicing). These test devices are measured on the wafer at various steps in the process and immediately after interconnection metallurgy by the use of automatic test equipment. Parameters such as substrate resistivity, epitaxial resistivity, and diffusion, deposition and drive-in resistivities, as well as many other electrical device parameters, can be measured on each run and displayed on computer-plotted quality control charts. The test data are used for yield analysis and diagnostics to improve control of the process and to optimize yield. The process has been practiced successfully under a high degree of control with these techniques.

Table 2 shows data taken from test devices located in the kerf area and represents typical device parameters for both ⟨111⟩- and ⟨100⟩-oriented material. It can be seen that device parameters are almost equivalent with the exception of low-current beta. At low currents, transistor gain becomes controlled by surface recombination current, and available base current drive to a device is reduced. It has been observed that ⟨100⟩-oriented material has a lower surface state charge than ⟨111⟩ orientation, and base recombination current is strongly related to the surface state charge. Figure 9 is a graphical illustration of transistor gain vs normalized collector current (I_C/I_{Cpeak}).

It is clear from this figure that beta diminishes much faster for the ⟨111⟩-oriented material. For currents of 0.001 times the peak value of collector current, the beta is 5 times higher for ⟨100⟩ than ⟨111⟩. This improvement

in low-current beta inherent in the process has permitted the design of a memory storage cell that can function under pulsed (bi-level) power. The cell is designed to remain stable at reduced current levels during periods of standby when the circuit is not addressed.

This pulsed power mode of operation enables maximum information storage at minimum power dissipation due to the low current levels. The ⟨100⟩ monolithic memories with increased packing densities that were heretofore impossible with conventional ⟨111⟩ processing can be fabricated without the need for liquid cooling systems. For example, a 128-bit, random access memory chip containing 1434 components can be efficiently packaged in a group of four chips on a 1/2-inch module. Furthermore, the packing density of the monolithic memory requires high values of low-current beta and these are inherent in ⟨100⟩ devices.

• *Yield*

Overall memory chip yield is determined by the product of process yield and final electrical test yields. The electrical tests are performed by computer-operated high-speed memory exercisers capable of operating the devices in a simulated computer environment in real time. Over 12,000 functional tests are performed in milliseconds, resulting in simultaneous ac and dc testing of the memory.

Figure 10 is a curve showing the relative yield of good chips as a function of distance from the center of the wafer. It should be noted that the distribution of good chips across the wafer is about constant out to a diameter of 6 or 7 rings. The outside ring, although it shows much lower yield values, still exhibits a finite chip yield.

Test results over a considerable period of time have demonstrated that the process, as developed, is capable of consistently producing good yields. The differences observed between ⟨100⟩ and ⟨111⟩ device characteristics, including better overall junction quality, combine to reduce the technological limitations encountered in the design and fabrication of large-scale circuits.

Summary

The formulation, experimental development, optimization, and application to production of a ⟨100⟩-oriented silicon process for high density, high-speed, low-power monolithic memory circuits has been discussed. Both the fundamental nature of the monolithic circuit for main memory application and the process to support it represent a significant departure from conventional semiconductor technology. The economical fabrication of these large area chips from increased diameter, thicker silicon wafers while increasing injection efficiency through higher phosphorous surface concentration has also been demonstrated. The attainment of high-speed switching capability with gold diffusion without diminishing junction quality has been the result of a judicious selection of both gold and phosphorous thermal cycles.

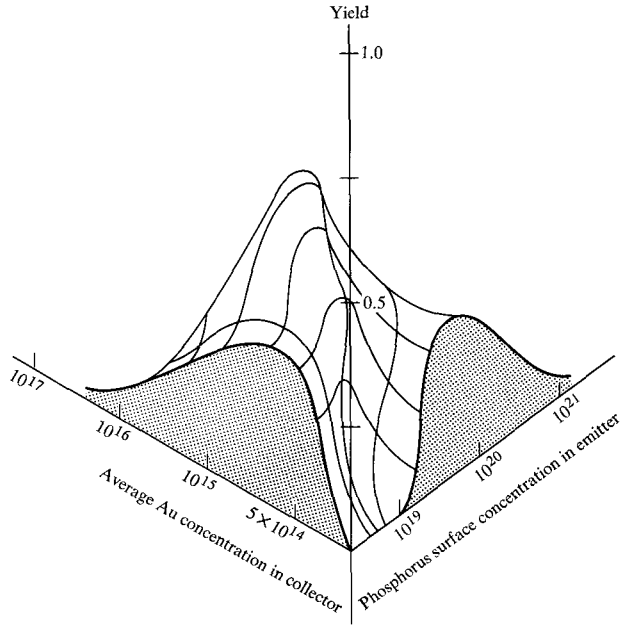


Figure 11 Yield vs gold and phosphorus concentrations.

Figure 11 is a three-dimensional display of the concentration of gold and phosphorus vs yield in arbitrary units. This figure conveys the concept that a maximum functional yield exists in a gold doped npn device circuit.

The development of a process using the ⟨100⟩ material, originally adopted for maximized low-current gain due to minimized surface charge density, has been shown to yield other processing advantages not attainable with ⟨111⟩. These have been observed as:

1) Higher surface concentration (near solid solubility) in the phosphorous doped emitter;
2) higher junction quality;
3) increased crystallographic perfection, both as grown and during processing; and
4) optimum epitaxial growth morphology, particularly over buried subcollector structures.

The initial advantage has also been responsible for increased emitter efficiency and low contact resistance, and the higher source concentrations are an important factor in the production of high-density, high-speed devices.

The final result of these process innovations is the fabrication of a high-density monolithic memory circuit functioning at nanosecond speeds without complex liquid

cooling. These advances also have value for equivalent high-speed circuits in logic applications where junction quality, cycle times, and similar parameters have equally fundamental importance. The improved charge density of oxide-passivated ⟨100⟩ surfaces, and their corresponding stability, demonstrate the possibility of direct application of the ⟨100⟩ orientation for future high-density devices.

Acknowledgments

The authors wish to express gratitude to their many colleagues for contributions to this study: to P. P. Castrucci for his technical encouragement and constructive criticism during the investigation; M. M. Roy, G. D. Watts and W. A. Westdorp for the wafer processing, planarity and SOT analyses; R. J. Herzdig and J. W. Mason for the aluminum penetration study; and many others unfortunately too numerous to acknowledge here.

References

1. E. G. Grochowski, G. Maheras, W. D. North and M. M. Roy, "The Effect of ⟨100⟩ Orientation in Fabrication of High Density, High Speed Monolithic Circuits," presented at the IEEE International Electron Devices Conference, Washington, D.C., October 1967.
2. P. V. Gray and D. M. Brown, "Density of SiO_2-Si Interface States," *Applied Physics Letters*, **8**, 2 (1966).
3. P. P. Castrucci, E. G. Grochowski, M. S. Hess, G. Maheras and W. D. North, U.S. Patent 3,585,464, June 15, 1971.
4. R. M. Warner, Jr., *Integrated Circuits*, McGraw-Hill Publishing Co. Inc., New York 1965
5. M. C. Duffy et al., *J. Electrochem. Soc.* **115**, 84 (1968).
6. P. A. Totta and R. P. Sopher, *IBM J. Res. Dev.* **13**, 226 (1969).
7. M. Zelm and C. E. Gaier, "Gold Concentration Profiles and Minority Lifetime in NPN Structures," presented at the Electrochemical Society Conference, Atlantic City, 1970.

Received December 8, 1970; revised September 1, 1971

The authors are located at the IBM Components Division Laboratory, East Fishkill (Hopewell Junction), New York 12533.