The Silicon p-n-p-n Switch and Controlled Rectifier (Thyristor)

Nick Holonyak, Jr., *Life Fellow, IEEE*

Invited Paper

***Abstract*—Based on the two-transistor model of Jim Ebers (a p-n-p transistor driving an n-p-n, and the n-p-n driving the p-n-p), the two-terminal and three-terminal Si p-n-p-n switch (low power) originated at Bell Telephone Laboratories (BTL) in 1954–1955. The two-terminal version, with its various limitations (along with the Si technology supplied by BTL, Moll's group), went with Shockley to the West Coast. The two-terminal device and the Shockley enterprise failed, except as an unplanned, unpredicted transfer of technology that accidentally launched Silicon Valley. The three-terminal p-n-p-n device introduced by GE (1957) as the Si controlled rectifier (SCR, later thyristor) succeeded from the beginning, however, and became the dominant control device in the power industry. The early history of this work (1954–1960), including the shorted-emitter and symmetrical switch (TRIAC), is described. The early work proved the need to employ, besides the basic vertical p-n-p-n layering, lateral p-n patterning and the use of the lateral geometry for three-terminal operation, shorted emitters, symmetrical switches (TRIACs), regenerative gate operation, and ultimately gate-turn-off switches. Indeed, the two-terminal device could not match the performance of the three-terminal p-n-p-n switch, which became the premier megawatt control device of the power industry.**

***Index Terms*—Al evaporation, alloy junction, Al metallization, alpha sum, alpha sum unity, Au evaporation, Au metallization, avalanche breakdown, base current, breakover current, collector current, current continuity, diffused junction, diffused transistor, dV/dt problem, emitter current, Ga diffusion, Ga+P diffusion, hook collector, impurity diffusion, inversion layer, lateral current, local liquid-phase epitaxy (LPE), local LPE, n-p-n-p-n incomplete switch, n-p-n transistor, oxide masking, P diffusion, p-n-p-n switch, p-n-p transistor, point-contact transistor, short emitter, Si controlled rectifier (SCR), silicon (Si), Si oxide, Si technology, Si technology transfer, switching condition, symmetrical switch, temperature stability, three-terminal switch, thyristor, traps (defects), transistor alpha, TRIAC, two-terminal switch.**

I. Introduction

THE three terminal Si p-n-p-n switch, the Si controlled rectifier (SCR or now thyristor), is over 45 years old (1955) [1] and is the work-horse of the power industry. It is the premium power device. It is more, however, than the premium active element of the world of power. It is a special form of transistor that, besides its practical importance, has played a unique role in how Si emerged as the primary substance of electronics—the key material of electronics for which there has been no substitute (and maybe never will be!). In this article, we want to describe how the p-n-p-n switch, and thus SCR (thyristor), originated, starting with Bardeen and Brattain's hook-collector point contact transistor [2], [12], the Ebers' two-transistor circuit model [3], the Bell Telephone Laboratories (BTL) Si p-n-p-n switch (the generic device) [1], the General Electric (GE) Si controlled rectifier (SCR) [4], and the shorted-emitter and symmetrical switch (TRIAC) [5].The author's concern here is the early period of the existence of the p-n-p-n switch and SCR that was witnessed.

The world of power electronics and the power industry are, of course, well aware of the importance of the SCR (thyristor), but probably it is much less appreciated that the Si p-n-p-n switch lies at the beginning (1954–55) of modern Si electron device technology [1]. For example, Si p-n-p-n switch substrate wafers were some of the first, including some that were part of the work of [1], on which Carl Frosch grew (1955) the Si protective (masking) oxide [6], the oxide so critical to the Si integrated circuit and the so-called "chip." In addition, the Au and Al metallization procedures now standard in the "chip" industry were first developed in Moll's group at BTL to make contacts and shallow p-n junctions on the p-n-p-n switches of [1], [7] as well as on the diffused-base Si transistors that were also part of this work. This is the Si technology supplied by Moll's group that, indeed, was carried to the West Coast [8], and as a failed two-terminal crosspoint technology (but a source of Si "know-how") became the source, the point of origin, for the technology and people that generated "Silicon Valley." In other words, the Si p-n-p-n switch did more than serve as the source of the SCR (thyristor). It carried Si device work across the U.S. to a place that became known as "Silicon Valley."

II. Hook Collector

The culmination of the Bardeen and Brattain series of studies on semiconductor surface effects and the search for a field-effect amplifier device (not conceptually a transistor, then an unknown idea) was the major discovery (1947) of a new form of amplifying device, a device that could be named "the transistor" [2] (see also [9]). Not a voltage driven device like a vacuum tube, the new "transistor" operated with an input current (emitter current, I_e) at low impedance and with almost the same current

Manuscript received October 12, 2000; revised December 1, 2000. Recommended by Associate Editor A. Kelley.

The author is with the Electrical Engineering Research Laboratory and Center for Compound Semiconductor Microelectronics, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA.

Publisher Item Identifier S 0885-8993(01)01148-6.

(collector current, I_c) extracted at high impedance, hence gain and a "transfer resistor" or "transistor." This device, because of the history of how it occurred (and as last described by Bardeen in his NHK TV interview, Urbana, IL, June 1990) [9], quickly became two metal point contacts, an emitter and a collector, at close spacing (thousandths of an inch) on a smooth (etched) surface of Ge. The third electrode, the base (hence, I_b), was attached to the n-type Ge [2]. Generally in the common base configuration of input and output circuit, the ratio alpha (α) of output to input current approached unity ($\alpha \equiv I_c/I_e < 1$) [2].

To match the device collector impedance to the load, frequently the collector point contact was "formed," which was simply a more or less crude method of modifying the Schottky-barrier collector point contact by a heating pulse (e.g., discharge of a capacitor through the collector). Often this empirical collector modification procedure yielded an "alpha" exceeding unity ($I_c/I_e > 1$), which further yielded negative resistance effects and the designation "hook collector." It was apparently Shockley who proposed that the "hook collector" was owing to an inversion layer formed at the collector contact. Since the original transistor, the point contact transistor [2], is fully equivalent to a junction p-n-p device, the inversion-layer proposal made sense, giving thus the equivalent of a p-n-p-n device—hence Shockley's claim to paternity of the p-n-p-n switch. It remained yet to see a proper p-n-p-n switch, not to mention one made in Si and not Ge.

III. Ebers' p–n–p–n Switch Circuit Model

In the meantime, however, the p-n-p-n switch could be simulated by a circuit model, the Ebers' model [3] which is the center part of Fig. 1 between $A(+)$ at the bottom and $B(-)$ at the top. The idea of a p-n-p-n switch could be verified by a p-n-p transistor (bottom) driving an n-p-n (top) and, in turn, the n-p-n driving the p-n-p. The collector of one, either one, drives the base of the other. This is guaranteed to yield instability. When the voltage from A to B reaches avalanche breakdown of the "n-p" diode (center of Fig. 1) and sufficient current flows in emitter shunt resistors R_1 and R_2 to bias on the emitters, the sum of α_{pnp} and α_{npn} approaches unity, and to maintain current continuity switching occurs to low voltage. The two collectors switch from reverse to forward voltage, and to the "on" state of the A-B switch, which, of course, is still not a p-n-p-n switch in a single "slab" of Si. Could such a switch be built, and would it, indeed, work—circuit model notwithstanding?

IV. Silicon p–n–p–n Switch

In the early fall of 1954, J. M. Goldey (MIT) and the author (Bardeen's laboratory, Urbana) joined John Moll's Bell Telephone Laboratories (BTL) group with the specific task of constructing a Si p–n–p–n switch [1]. Moll quickly convinced us that we should try to build a Si p-n-p-n device, an undemonstrated new device, that potentially could compete with a two-terminal gas tube designed to be used, perhaps in large numbers, as a telephone crosspoint switch. The BTL switching group that wanted the crosspoint device (a Si device, Ge was obviously too leaky, $E_g(\mathrm{Ge}) = 0.67 < E_g(\mathrm{Si}) = 1.12$ eV) even wanted, because of the small negative resistance of the two-terminal gas

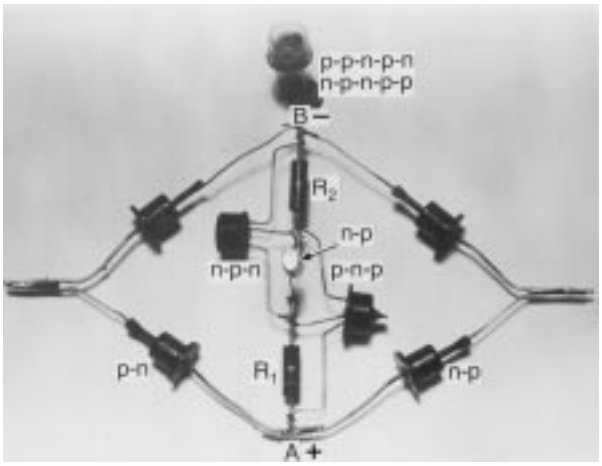

Fig. 1. Circuit model of a p-n-p-n switch. Between points A and B, the basic switch consists of a p-n-p transistor and an emitter shunt R_1, and an n-p-n transistor and an emitter shunt R_2, and an n-p avalanche diode (arrow) fixing the common collector breakdown voltage to a lower value than that of the collector junction of either transistor. The diamond-shaped diode (p-n) rectifier bridge makes the basic A-B p-n-p-n switch symmetrical and into a circuit equivalent of the single "chip" Si $\begin{matrix}\text{p-p-n-p-n}\\ \text{n-p-n-p-p}\end{matrix}$ shorted-emitter symmetrical switch at the top. (Circuit model made by T. P. Sylvan at GE in 1958.)

tube and hence possibility of circuit gain, a small on-state negative resistance in the proposed Si crosspoint. Moll, the principal "architect" in the beginning of the Si p-n-p-n switch (Shockley was nowhere near us or involved in this), argued that we must attack the problem of making a Si n-p-n switch by learning how to make diffused junctions and shallow evaporated (alloyed) contacts and junctions. He was not interested in the author's proposal to take an existing bar-geometry Si p-n-p-n transistor and modify it rapidly into a switch with either a point contact or small Al alloyed junction on the n-type collector body near the collector junction.[1] The author wanted to see if a Si p-n-p-n device really switched, and Moll wanted a switch rendered by a more rational, by a more robust and viable technology. It did not matter that this technology did not yet exist even for transistors. We would have to discover and develop it, as indeed happened—an entire new technology (oxidation, diffusion, metallization) to process Si substrate wafers into transistor devices. For example, later (Spring, 1955) came Frosch's oxide [6], which involved also our Si p-n-p-n switch work. The oxide gave us special capability to effect impurity diffusion. Needless to say, the Frosch oxide is now fundamental to the integrated circuit industry.

Not knowing in the beginning (1954) of the role of traps (defects) in governing the injection efficiency of Si p-n junctions [1], we pretty much followed Jim Ebers' two-transistor model for the proposed p-n-p-n switch [3]. For test reasons, we planned to make multiple terminal structures with resistive paths, including if necessary external resistors, to provide shunt leakage and variable bias at the p-n-p-n emitters, just as suggested in Fig. 1 by the five components stacked vertically from A to B ($+$ to $-$) within the diamond-shaped rectifier bridge. At avalanche

[1]The author later used this idea at GE (~ 1960) with Tom Mapother (father of Tom Cruise) to convert a bar-geometry n-p-n transistor to an n-p-n-p Si switch, and thus managed to keep an obsolete transistor in production longer as a latching device

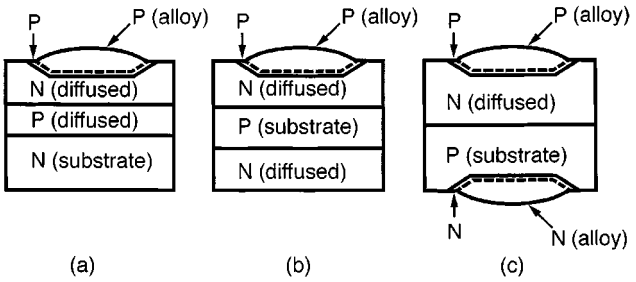

Fig. 2. Schematic cross sections (modified from [1, Fig. 4]) of three forms of Si p-n-p-n switches made at Bell Telephone Laboratories (BTL) in 1955. The n-p-n-p switch complementary to the p-n-p-n of (b) became the Si controlled rectifier (SCR) at GE in 1957.

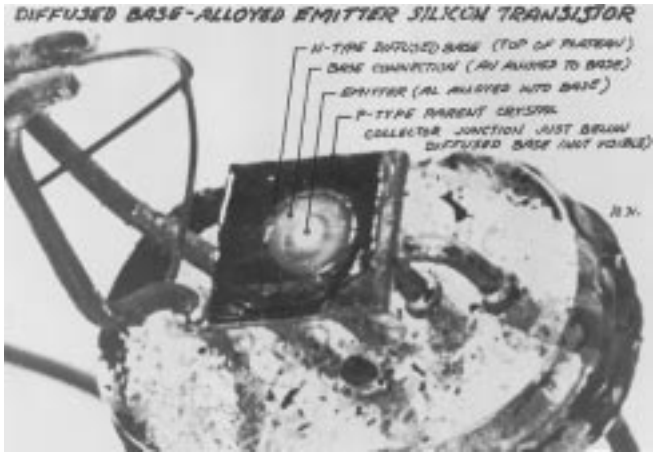

Fig. 3. Diffused-base Si p-n-p transistor with evaporated metal contacts made at BTL in 1955. The hand lettering is simply for the convenience of discussion (cf., Fig. 4). Note that thermo-compression bonding did not yet exist and the leads are spring contacts bent out of place for device viewing.

Fig. 4. M. Kikuchi and N. Holonyak, Jr. examining the diffused-base Si transistor of Fig. 3 at Denki Shikenjo (Tokyo, Japan) in 1956. (The note originally was on the back of the picture.)

breakdown of the center n–p collector junction (in the circuit model the diode at the center of Fig. 1) and as a result shunt current providing emitter bias, which would yield in turn current-variable (increasing) carrier injection and base transport, the common collector of the p-n-p-n device would be forced to switch from reverse $(+-)$ to forward $(-+)$ bias to preserve continuity of current. Specifically, the common electron-hole collector of the p-n-p-n switch, the center n-p junction, would be forced to switch from high reverse bias (voltage) to low forward bias to maintain continuity of current, all a consequence of $\alpha_{pnp} + \alpha_{npn} \rightarrow 1$ [1]. This is the key to the switch operation, the alpha sum approaching unity (and thus the need to switch).

Based on these ideas, we planned, generated enough of the new Si diffused-impurity transistor technology, and built relatively quickly (1954–1955) three and even four terminal Si p-n-p-n switches, but were pleasantly surprised to find we didn't need shunt leakage or resistors around the emitter junctions (R_1 and R_2 in Fig. 1) because of the saturable traps inherent in the junctions [1]. Simultaneously (1954–55) our BTL colleague Mort Prince was observing and independently confirming the effect of defects (traps) on the injection behavior (I-V characteristics) of diffused junction Si rectifiers [10]. Because of the traps in the junction transition region, at lower current levels the diffused Si rectifiers, as well as the two emitters of the p-n-p-n switch, behaved essentially as p-i-n [11] not p–n junctions, i.e., as $J \propto \exp(qV/2kT)$ instead of $J \propto \exp(qV/kT)$ until saturation of the traps and then onset of injection.

One of the p–n–p–n switches [1] [Fig. 2(a)] was automatically a three-terminal device because it was intended to be an n-p-n transistor but the shallow evaporated Al metallization [7] that was meant to form the base contact, by accident, did not reach through the top n-type layer into the middle p-type base layer and form the desired contact. It gave a p-n junction, an alloyed junction, on the top n-type layer [Fig. 2(a)]. Instead of yielding an n-p-n transistor it turned-out to give, because of the accidental Al alloy p-n junction and an evaporated Au-Sb ohmic contact (not shown) on the top n-type region (now a base, not an emitter), a three-terminal p-n-p-n switch—a small solid-state "thyratron" (thyristor). In any case, our first Si p-n-p-n switches were, in fact, multi-terminal devices, not two-terminal devices. These devices, of course, could be operated as two-terminal switches.

One of the switches [Fig. 2(a)] was double diffused from the top-side of the crystal and employed, as stated, an alloyed (evaporated) third junction on top (Tanenbaum). Another [Fig. 2(b)] was diffused on (from) both sides of the Si wafer and employed an evaporated and alloyed third junction on top, or bottom (Holonyak). The third [Fig. 2(c)] employed one central diffused junction and two outside alloyed (evaporated) junctions (Goldey). These are shown schematically in Fig. 2, which is a modified version of Fig. 4 of [1] showing, in addition, that a regrown (epitaxial) Si layer exists under each of the alloy regions.[2] Note that in the case of Fig. 2(b), if we remove the bottom diffused n-type layer (by polishing and etching), we are left with a diffused-base alloyed emitter (evaporated Al) [7] p-n-p transistor. One of these transistors that we made in 1955 is shown in Fig. 3. The hand lettering (N.H.) is simply for convenience in showing and discussing the device, which occurred later with M. Kikuchi at the Electrotechnical Laboratory (Denki Shikenjo, Tokyo, Japan, 1956). Note that thermo-compression

[2]The alloy process for making p-n junctions was introduced by R. N. Hall as part of his identification and study of the p-i-n diode and rectifier [11]. Hall's alloy process, which is really liquid phase epitaxy (local LPE!), was used to make billions of transistors and carried transistor development and study for ten or more years between ~1950 and ~1960 until diffused-impurity Si transistors began to take-over and dominate

bonding did not yet exist and the connecting leads were spring contacts that were bent out of place for convenience in viewing the device.

Employing a new technology at BTL in 1955, we made diffused-impurity Si transistors, and three-terminal as well as two-terminal p-n-p-n switches. Although it made more sense to build and study the three-terminal p-n-p-n switch than the two-terminal switch, and to change the system crosspoint logic to a three-terminal device configuration, which Moll advocated, this idea was not accepted. It did not fit the original gas-tube crosspoint logic notions, a BTL constraint. Hence, the first account of this work emphasized the two-terminal behavior of Si p-n-p-n switches, but for good reason did not overlook the fact that the three-terminal version operated as a thyratron [1].

V. From p–n–p–n Switch to Silicon Controlled Rectifier (SCR)

Not only had we introduced a new device in 1955, the Si p-n-p-n switch, we also introduced a new Si technology [1], [6], a technology employing Si wafer processing to make p–n junction devices by impurity diffusion, metallization (evaporation), and oxidation. This became an invariant technology that developed and developed, and after the appearance of [1] and [6] was bound to proliferate.

By the time [1] was published, the author was in the U.S. Army serving in Japan and, via an introduction from John Bardeen, became acquainted (1956) with George (Mitio) Hatoyama and Makoto Kikuchi, both then at the Electrotechnical Laboratory (ETL, Denki Shikenjo). Hatoyama was later the founding director of Sony's research laboratory, and somewhat later Kikuchi became its director (1974–1989). Fig. 4 shows the author describing the transistor of Fig. 3 to Kikuchi (at ETL). If the bottom-side n-type diffused layer is not removed, the transistor of Fig. 3 becomes the p-n-p-n switch of Fig. 2(b), which obviously is a three-terminal device (switch or transistor). When the author entered the Army, the BTL attorneys had restricted him from talking about only one matter, Frosch's oxide and oxide masking [6]. Thus soon after [1] became known in Japan, the author was invited and able to give a seminar in Tokyo on Si p-n-p-n switches and more broadly on diffused-impurity Si transistors. The note Kikuchi wrote inviting him to give the talk is shown in Fig. 5. A picture of the group to whom the author gave the seminar at the Electrotechnical Laboratory (Denki Shikenjo) in early Feb, 1957 is shown in Fig. 6 (Kikuchi on the author's right).

When the transfer of the new Si p-n-p-n switch and transistor technology, now a legendary story, was taking place from BTL (from Moll's group) [8] to Shockley, and hence planting the seeds of Silicon Valley, it was taking place also more broadly (e.g., to Japan and elsewhere). In fact, before the author entered the U.S. Army, preliminary planning was already underway at BTL to hold another transistor technology symposium (the second after the original Ge transistor symposium) to inform licensees and others of the new Si diffused-junction transistor technology. Licensees were beginning to catch-on that something new was afoot in Si transistor technology at BTL, and another symposium (January 1956) was needed to stem the constant interruption of visitors. The new transistor and switch technology was made broadly available to the world, but not as early or in as much detail as to Shockley [8].

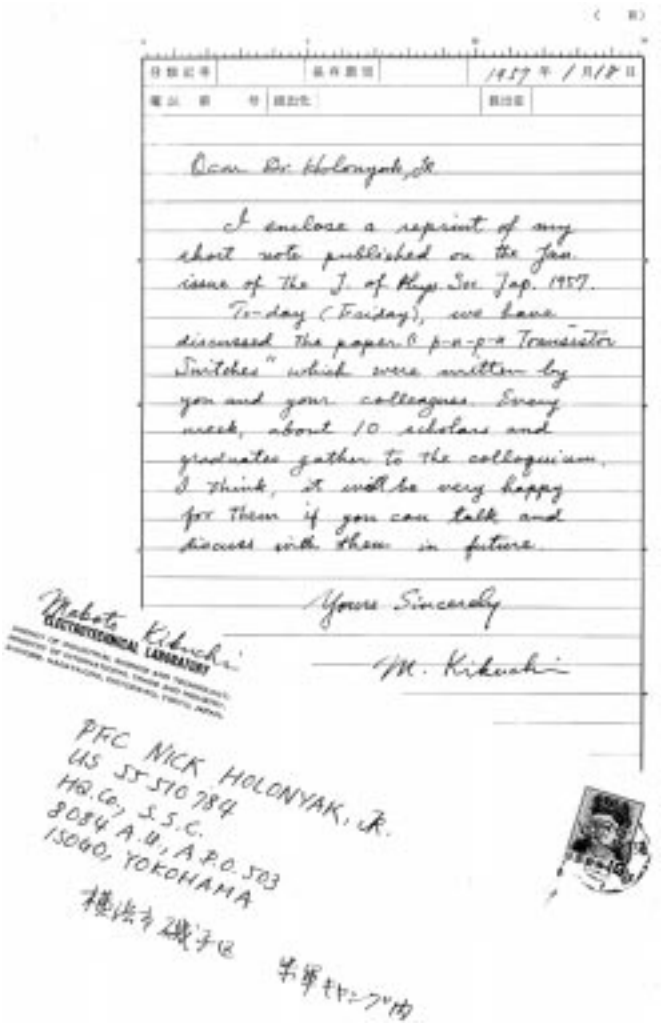

Fig. 5. Letter that Kikuchi (Tokyo) sent in January 1957, after the publication of [1], inviting Holonyak (Yokohama, Japan) to give a seminar at the Electrotechnical Laboratory (ETL, Denki Shikenjo, Japan) on Si transistors and p-n-p-n switches.

Fig. 6. Holonyak seminar, arranged by M. Kikuchi, on diffused-base Si transistor and p-n-p-n switch devices at Denki Shikenjo (Tokyo, Japan, February 1957). Kikuchi on N.H.'s right and Shibuya in corner on Kikuchi's right.

Of course, technology transfer occurs also in the more classical way of reports in the journal literature. The best example the author knows of this became apparent to him in the Fall of 1957 when he left the Army and decided to join General Electric (GE, Syracuse). The disclosure of the three-terminal operation of the p–n–p–n switch in [1] did not go unnoticed. The three-terminal Si p-n-p-n switch [1] turned out to be of considerable moment, but not the two-terminal device that enamored Shockley. General Electric rectifier department, specifically R. A. York, acting on advice from R. N. Hall alerting York to consult [1], started a program (1957) to build the silicon controlled rectifier (SCR) [4], the three-terminal p-n-p-n "thyratron" [1]. York was happy with the Si rectifier, and had asked why could there not be a Si thyratron? The answer was in [1].

York's staff had in the beginning a poor understanding of p-n-p-n switches, but had enough of the requisite Si technology to construct a p-n-p-n switch of complementary form, an n-p-n-p, to the one shown in Fig. 2(b) [1]. They followed the published "recipes" [1] but on a more or less cruder scale. The GE engineers built their device too thick (higher voltage) and too large in area (higher current), at least from the perspective of a BTL Si "crosspoint," and got for their efforts a device operating at hundreds of volts and 10's of amperes, i.e., the most welcome surprise of a many kilowatt "thyratron." What York wanted happened. The three-terminal p-n-p-n switch, the SCR or later thyristor, was thus launched and proved to be successful, successful from the beginning (1957). It was clear from the start, first in Moll's group (1955) and later at GE (1957), that there was need and purpose for the three-terminal p-n-p-n switch—and not particularly the two-terminal device. (In the case of the latter, the voltage and current parameters would have to be controlled too tightly, and input and output could not be adequately separated, the usual weakness of two-terminal devices.)

When the author joined GE (Syracuse, NY, November 1957), the author immediately went back to work building and studying the p-n-p-n switch and helping York's engineers to understand the SCR. A cross section of the type of Si p-n-p-n switch GE introduced as the SCR and that then went through many development cycles and device types is shown in Fig. 7. A $\sim$250 μm thick n-type Si wafer was diffused with Ga to a depth of $\sim$67 μm on both sides, giving overall a thick p-n-p. Then a rather thick Au (+Sb) layer, backed-up with a tungsten plate, was alloyed on the top-side of the wafer. The Au dissolved some of the p-type Ga-diffused Si ($\sim$40 μm) and in the cooling portion of the alloy cycle regrew $\sim$15 μm of n-type Si, thus giving a relatively thick ($\sim$210 μm) n-p-n-p switch capable of high voltage and high current operation. A gate connection to the top p-type layer could easily be made beyond the n-type regrown region. Note that the alloy cycle yielding the top n-type layer was simply an instance of liquid phase epitaxial (LPE) crystal growth [11], [15], i.e., "local LPE." There were obviously many versions of this basic form of SCR, a junction structure complementary to that of Fig. 2(b), before it was superseded by an all-diffused structure.

VI. Shorted Emitter and Symmetrical Switch (TRIAC)

Although the form of SCR shown in Fig. 7 was clearly important and was part of several generations of commercial devices,

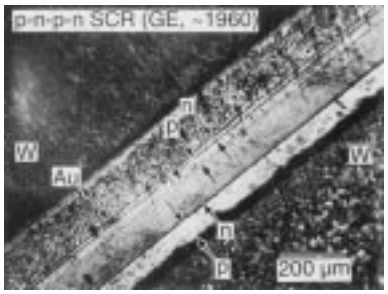

Fig. 7. Cross section of an early (1957–60) GE Si p-n-p-n SCR, with Ga diffused p-type regions) into both sides of a $\sim$250 μm n-type Si wafer and with an n-type emitter regrown on top from a Au (+Sb) alloy.

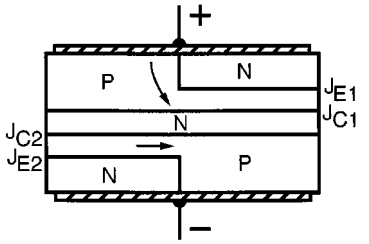

Fig. 8. Schematic cross section of a shorted-emitter symmetrical switch that in either polarity operates as a p-n-p-n switch. In the polarity shown the top emitter junction J_{E1} is inoperative and the bottom emitter junction J_{E2} is biased into operation by transverse (lateral) current.

it was not particularly convenient for us to build in the Advanced Semiconductor Laboratory (ASL) in Syracuse. In ASL we made p-n-p-n structures by a relatively convenient "one-shot" diffusion process employing as an impurity source Ga alloyed onto a small slab of Si and then saturated with phosphorus (P) in a closed-tube high temperature anneal cycle. This impurity source could be used over and over and, sealed into an ampoule with an n-type Si wafer, gave (simultaneously) a deeper Ga diffusion followed on top by a shallower P diffusion at higher impurity concentration. This occurred because Ga has a lower solid solubility in Si than P and, conveniently, a larger diffusion constant, leading to a deeper depth of diffusion. On an n-type wafer the simple Ga+P diffusion procedure yielded simultaneously an n and p layer ($n_{\rm P}$ shallower, $p_{\rm Ga}$ deeper) on both sides of the wafer, thus giving a symmetrical n-p-n-p-n unless the P diffusion was blocked with Frosch's oxide [6]. Ga diffusion is not blocked by the oxide, and the P diffusion (n-type regions) could be arranged in any desired form by patterning the oxide. Obviously we knew about a "symmetrical switch," say, an n-p-n-p-n in which the p-n-p-n portion switched but one (either one) of the shallower junctions ($\rm n_P$-$\rm p_{Ga}$) operated in reverse bias, say, in avalanche breakdown at low voltage, but not nearly as low as a forward-biased junction.

In a small meeting in Syracuse, NY, with the author and R. W. Aldrich in the Spring of 1958 York's engineers (F. Gentry [4], Gordon Hall, who made the first SCR, and others) asked why we could not devise for them a symmetrical Si switch. Afterall, we were not working with a thyratron, a gas, but a p-n-p-n switch, a solid (Si). The solid should be capable of much more, e.g., a much more intricate device geometry. We explained to them the problem of incomplete switching in an n–p-n-p-n, and our visitors left but Aldrich and the author stayed in the meeting room

and continued our discussion. Before we left the room, we devised the answer to a true symmetrical switch, one that switched down all the way to low voltage, to all the operative junctions in forward bias. By the next day we made the symmetrical switch, which is shown schematically in Fig. 8 [5], [13].

In general, when we made and studied n-p-n-p switches (SCRs) by the "one-shot" Ga+P diffusion process, we masked one side of an n-type Si wafer with oxide, giving only a p-type layer (Ga), and masked the other side partially with oxide stripes, giving side-by-side n-type (P) and p-type regions (Ga). On top we could attach an electrode to the $\mathrm{n_P}$ layer and, side-by-side, another electrode (the gate) to the $\mathrm{p_{Ga}}$ layer, which reached the surface where the Si was masked with oxide. Sometimes we did this without the aid of a microscope since, in losing an old contract, we lost some microscopes. We knew our metal electrodes, at the crystal surface, must have shorted across some of the n_{P}-$\mathrm{p_{Ga}}$ diffused junctions. Nevertheless, our n-p-n-p devices switched. We knew this meant transverse currents were creating the appropriate biases internally to make our $\mathrm{n_P}$-$\mathrm{p_{Ga}}$ emitters operative. This then led immediately to the shorted-emitter symmetrical switch of Fig. 8, after, of course, we gave a bit more thought to the problem.

When Aldrich and the author emerged from the ASL meeting room, we knew immediately how to proceed to make a symmetrical switch. We first oxidized an n-type Si wafer (both sides). We pulled some Apiezon black wax into threads, placed them on a diagonal at even spacing on one side, carefully heated and attached them to the wafer as "half-rounds," and etched off the oxide between the masking threads. Then the procedure was repeated on the other side with the threads placed also on a diagonal but at right angles to the threads on the first side. Next came the Ga+P diffusion. After removing the masking oxide stripes, we attached one side of a small piece of the wafer to a metal plate on a standard transistor header and attached another electrode on top. We could simply and easily make a nonpenetrating contact with Pb+Ti, and easily short across the n_{P}-p_{Ga} diffused junctions at the crystal surface (see Fig. 8). The device was completed with a deep etch into the wafer, but not so as to damage the electrodes. The way the masking threads were crossed on the two sides of the n-type wafer insured that we had, sight unseen, a shorted-emitter on top (electrode straddling the n_{P}-p_{Ga} junctions) aligned with a shorted-emitter on bottom, just as sketched in Fig. 8. The first device worked, and the shorted-emitter concept was proved immediately.

It is clear how the symmetrical switch of Fig. 8 operates. With the voltage polarity shown, the top shorted junction J_{E1} is inoperative, and positive current (curved arrow) flows on the left as shown. Hole injection at J_{C1} (acting as an emitter) occurs on the left and is collected on the left at J_{C2}, where, as majority carrier current, it causes a lateral voltage drop along the thin p layer below J_{C2}. Internally the bottom p-type layer is much more positive on the left than on the right where it is at the same potential "shorted" to the bottom n-type layer. At avalanche breakdown of J_{C2} and large enough current, the internal lateral positive-current biasing reaches a high enough level to cause the left side of the bottom n-type layer to reach sufficient forward bias to inject electrons. The bottom n-type layer, let us say, at the J_{E2} label, becomes an effective emitter, and we have a top-down operating p–n–p–n switch. All that is needed for switching is enough avalanche current at J_{C2} for lateral internal biasing to turn-on the emitter near the J_{E2} label. Obviously, because of the device symmetry, we could have employed the opposite voltage polarity. The two emitters, J_{E1} and J_{E2}, appear shorted, but internally they function very effectively with sufficient lateral current bias.

One of the first shorted-emitter symmetrical switches that we made is shown in operation in Fig. 9. It is now 42 years old and still operates. It is the progenitor of all shorted-emitter and symmetrical switches. Because the device was hand made and the top shorted-emitter not totally similar and symmetrical to the one on bottom, the break-over current, i.e., the avalanche current at switching $(\alpha_{npn} + \alpha_{pnp} \rightarrow 1)$, in the first quadrant (see arrow) is not identical to the one in the third quadrant (opposite arrow). It is obvious from the "on" current in either direction that there is no problem with the "on" voltage. All of the operative junctions are in forward bias. Furthermore, it did not take much thought to see how third electrode triggering could be accommodated in a symmetrical switch, making the switch into a true AC control device.

Besides being the progenitor of all symmetrical p-n-p-n switching devices, TRIACs, etc., the device of Fig. 9 is proof that internal lateral currents are important and can be turned to good purpose in p-n-p-n switch design. More generally it was clear from our first experiments that the p-n-p-n family of devices would become, unlike the device of Fig. 7, all-diffused structures, and that oxide and diffusion patterning would become an important part of device design and development.

We should mention that when we made the symmetrical switches of Figs. 8 and 9 and showed our device to Pete Sylvan (GE, Syracuse), a clever device applications and electronics circuit engineer, he immediately generalized the Ebers' p-n-p-n switch model of a one polarity switch (A+ to $B-$ in Fig. 1) to the symmetrical switch of Fig. 1. This required, besides the usual five A-to-B switch elements, four more bridge rectifier components. The point was to show that it took nine components to do with a circuit what a single piece of Si could do, which is the device at the top of Fig. 1 and the one of Fig. 9. Incidentally, we considered this a true integrated circuit, one in which all the device functioning was merged into the crystal.

When the author later showed Jim Early (BTL) one of our symmetrical switches on the occasion of a Syracuse IRE seminar, he was surprised and reported his observations to Ian Ross (BTL), who then came to see what we had done. The BTL competing device was the n-p-n-p-n that switched incompletely. The author must have hinted vaguely what we did (the shorted-emitter idea), because Ross said someone, somewhere else, someone not revealed, had beat us. The author later understood his reaction, because of a subsequent patent interference. Ross had underestimated, however, both what we had done, how much we had done, and the very early timing on our basic work (Spring 1958). The explanation for Ross' comments came later when our GE attorney filed very late for a patent [13] and we ran into an interference with Bob Noyce, then of Fairchild and later Intel, who had, in fact, discovered (after us) a piece of the basic shorted-emitter idea. Noyce and the author had a friendly debate on this matter (and the

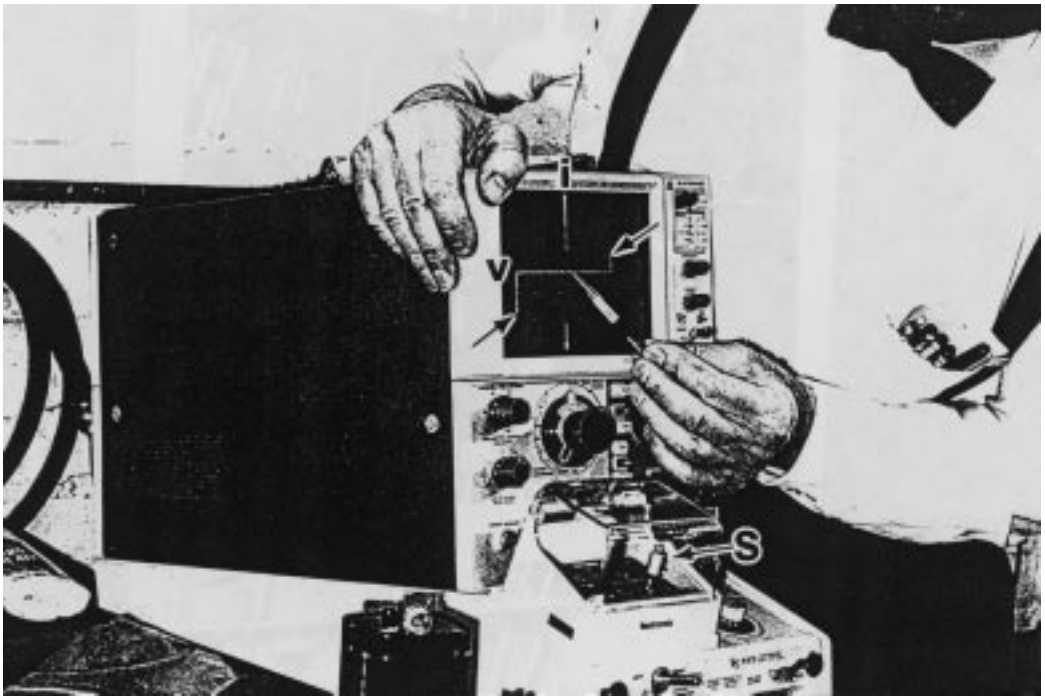

Fig. 9. First (1958) Si shorted-emitter symmetrical switch (the arrow S) showing (Urbana, IL, ~1995) its switching operation in the forward and reverse direction. Because this 1958 prototype of all symmetrical switches (TRIACs, etc.) was hand made, the forward and reverse switching currents (arrows) are different.

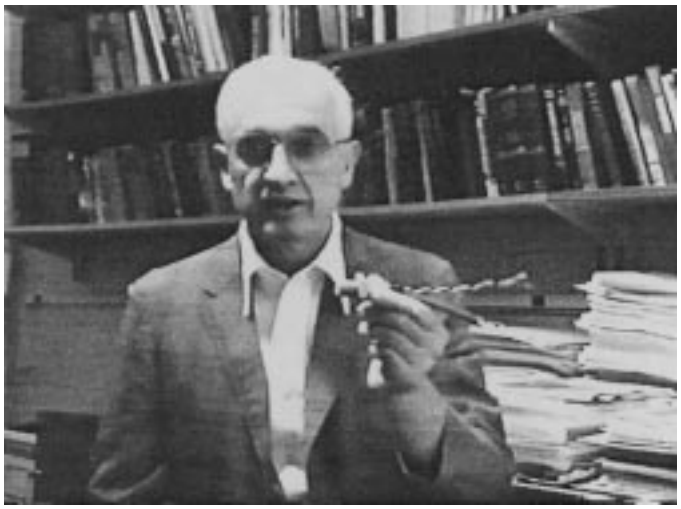

Fig. 10. General Electric SCR, 1960 gift to Bardeen, handed to N.H. by Bardeen near the end of June 1990 NHK Bardeen interview. (From the June 1990 NHK recording.)

Fig. 11. Bardeen and Holonyak talking about transistors and the NHK interview (June 1990, see [9]) just after he handed N.H. the 1960 GE SCR of Fig. 10. (From the June 1990 NHK recording.)

issue of research on Si versus III-V's) at the 1962 IRE Device Research Conference. Noyce told the author he would win the interference, but in spite of his smiling demeanor, he hesitated when the author told him, "Not unless your Fairchild work was done when you were still working for Shockley." Noyce lost the interference [13].

The shorted-emitter did more than make possible the symmetrical (ac) switch. It made it possible to set, by design, a certain current level before a p-n-p-n device switched, i.e., before the emitters became functional and thus $\alpha_{pnp} + \alpha_{npn} \rightarrow 1$. This gave the switch and the SCR more stability with temperature [5], [13]. Also, since capacitive current could fill Si traps and affect emitter operation and the alpha sum, thus giving false switching (the so-called "dV/dt problem"), the shorted-emitter and its built-in currents could reduce this effect. Once we introduced the shorted-emitter, and realized its advantages [5], it was certain to become part of all SCRs (thyristors). In fact, it is hard to understand why our patent attorney dawdled and filed so late (over a year) for a patent [5], [13].

VII. Comments and Conclusions

The author's intent here has been to describe the early days of the Si p-n-p-n switch and controlled rectifier (SCR), work prior to 1960. It happened that the author was the only one who participated in the first BTL work and then the early GE work.

In fact, at one point the GE corporate attorneys questioned the author to see if a fundamental device patent could be obtained on the SCR. The author considered this unlikely because the three-terminal p-n-p-n switch was covered, although briefly, in [1], and the BTL notebooks obviously contained further material but not, that the author recalled, on circuit applications. This made the shorted-emitter and symmetrical switch work of [5], [13] even more important in strengthening GE's position on SCRs, and on three-terminal p-n-p-n power devices in general, since a fundamental device patent was not forthcoming.

Shockley's mistake in believing in the two-terminal device, and not seeing or predicting the future correctly, was not GE's mistake. The three-terminal p-n-p-n device (SCR or thyristor) was the "right" choice and succeeded. The limitations of the two-terminal device were simply too great. The importance of the demonstration of the Si p-n-p-n switch [1] was not the two-terminal device, BTL's wish for a crosspoint. It was, in fact, the demonstration of the switching phenomenon itself and the three-terminal operation, something else (something, as it turned out, that York needed). The two-terminal device failed—and Shockley's enterprise—but, as we know, the transfer of Si technology from the East Coast to the West Coast was successful. It was not planned or predicted, nor even imagined. It occurred! Because of the Si p-n-p-n switch [1], Shockley's recruits (Noyce *et al.*) had access to a new Si technology supplied to Shockley [8] that could be taken in a different direction, back in the direction of transistors (three-terminal devices) and to a certain market.

We know what Shockley thought about the p-n-p-n switch but not necessarily what others thought, for example, John Bardeen. John never laid claim to the work of others, but since the p-n-p-n switch started from the hook collector of Bardeen and Brattain's point-contact transistor [2], [9], [14] he must have had some thoughts on the matter. Did he?

In 1960, on a Ph.D. recruiting trip to Urbana, IL, from GE, the author gave Bardeen a big SCR, one that operated at hundreds of amperes and a 1000 V or more (then a large device). The author wanted him to see a big transistor device, something that started with the point-contact transistor and its hook collector. The author wanted him to see this since, in 1952 when the author joined his laboratory, vacuum tube research people in Urbana with whom the author worked laughed at the minuscule power of a transistor, not to mention the rudimentary form of the point-contact transistor which reminded them of crystal-set radios. In their minds, and most minds, the thought was: How primitive, and how feeble compared to vacuum tubes!

The last time Bardeen told the transistor story was to NHK (Japanese television, June, 1990) and he wanted the author to be present [9]. Near the end of the interview, the NHK interviewer asked John to "introduce the author to the camera," and then asked him to say something about his work, which John already mentioned dealt with p–n–p–n switches, SCRs, tunnel diodes, LEDs, and lasers. In the author's response, the author added to what John said, and when the author mentioned (besides LEDs and lasers) the Si p-n-p-n switch and SCRs, Bardeen reached into his desk and handed the author the SCR the author is shown holding in Fig. 10. This was the same SCR the author gave him 30 years earlier, which, incidentally, is the same type of SCR that is shown in cross section in Fig. 7. Fig. 11 shows the author and Bardeen talking about transistors and the NHK interview a few minutes later, just after he handed the author the 1960 SCR.

The author did not know that Bardeen had the GE SCR in his desk for 30 years. Why did he still have it? He did not normally keep transistor artefacts and usually came to our laboratory for demonstration devices when he was to give a talk or seminar. He could have gotten an SCR from the author anytime he wanted one. The author knew John well enough to know that this high power Si p-n-p-n switch meant something to him, someone else's work that he could talk about without drawing attention to himself—yet something that started in his time. The "joke," the amusing irony, was that transistor devices were not doomed to operate only at low power—not as was commonly believed in the early days, including as voiced to to the author by his Urbana tube-lab friends (1952) when the author moved to Bardeen's laboratory. Great oaks indeed grow from little acorns, and megawatt SCRs (thyristors) came from micro- and milliwatt transistors, not to mention from the all-but-forgotten point contact transistor. Incidentally, Bardeen never claimed he could see megawatt Si switching devices coming from feeble point contact transistors. That required the work of many others.

Finally the author wants to mention that GE's SCR and the shorted-emitter, as well as the symmetrical switch (TRIACs, etc.), showed unambiguously the need and the advantages in employing device design and patterning in the lateral dimensions of a p-n-p-n device. The basic p-n-p-n switch, the two-terminal device, used the area of the device in the on-state very effectively to carry large currents, but made no use of lateral patterning in the doping and device geometry. Besides the basic vertical p-n-p-n layering, however, it was the lateral p-n patterning and lateral geometry that was vital for three-terminal operation, shorted emitters, symmetrical switches (TRIACs), regenerative gate operation, and ultimately for gate-turn-off switches. The Si p-n-p-n switch as a consequence became an amazing power device, the megawatt thyristor—a unique slab of Si much, much larger than a so-called "chip."

ACKNOWLEDGMENT

The author would like to thank D. A. Kellogg for help with Figs. 2, 7 and 8, M. Feng and S. W. Lee for rendering Figs. 10 and 11, B. L. Payne for assistance in manuscript preparation, the Sony Corporation for the support of the Bardeen Chair, his Urbana colleagues in power electronics for their interest in the subject of this paper and for encouraging him to write it, his BTL (1954–55) and GE colleagues for all they could do and learn about Si p-n-p-n switches and SCRs, M. Kikuchi and his colleages in Japan (1956–1957) who were interested in Si p-n-p-n switches and transistors and helped keep alive his interest in semiconductor materials and devices when he was stationed in Yokohama (1957 to 1963), and J. Moll who introduced him to the p-n-p-n switch problem and who argued, with great foresight, the need to explore, find, and use advanced Si technology in their work. Perhaps Silicon Valley owes more to J. Moll (and to C. Frosch) than to anyone else. Also, the author owes a special thanks to R. York (GE), for his wise support of the early SCR work, including much of the author's Si switch work, not

to mention later LED and laser work, the Air Force Cambridge Research Laboratory for a long period of support in exploratory work (at GE) on Si p–n–p–n switches, as well as later on III-V epitaxy and LEDs and lasers. The author owes more to John Bardeen and his memory than can be expressed. After John Bardeen, electronics changed, as well as, indeed, the entire field of the quantum theory of the conductivity of solids.

REFERENCES

[1] J. L. Moll, M. Tanenbaum, J. M. Goldey, and N. Holonyak, "P-N-P-N transistor switches," *Proc. IRE*, vol. 44, pp. 1174–1182, Sept. 1956.

[2] J. Bardeen and W. H. Brattain, "The transistor, semi-conductor triode," *Phys. Rev.*, vol. 74, pp. 230–231, June 1948.

[3] J. J. Ebers, "Four terminal P-N-P-N transistors," *Proc. IRE*, vol. 40, pp. 1361–1364, Nov. 1952.

[4] F. E. Gentry, F. W. Gutzwiller, N. Holonyak, Jr., and E. E. Von Zastrow, *Semiconductor Controlled Rectifiers*. Englewood Cliffs, NJ: Prentice-Hall, 1964, pp. 1–383.

[5] R. W. Aldrich and N. Holonyak, Jr., "Two-terminal asymmetrical and symmetrical silicon negative resistance switches," *J. Appl. Phys.*, vol. 30, pp. 1819–1824, Nov. 1959.

[6] C. J. Frosch and L. Derick, "Surface protection and selective masking during diffusion in silicon," *J. Electrochem. Soc.*, vol. 104, pp. 547–552, Sept. 1957.

[7] J. M. Goldey, "Evaporation and alloying to silicon," in *Transistor Technology*, F. J. Biondi, Ed. Princeton, NJ: D. Van Nostrand, 1958, vol. 3, pp. 231–244.

[8] J. L. Moll, "Fifty years of the transistor: The beginning of silicon technology," in *Proc. 1997 Symp. VLSI Circuits Dig. Tech. Papers*, Kyoto, Japan, June 1997, pp. 5–8.

[9] N. Holonyak, Jr., "John Bardeen and the point-contact transistor," *Phys. Today*, vol. 45, pp. 36–43, Apr. 1992.

[10] M. B. Prince, "Diffused *p-n* junction silicon rectifiers," *Bell Sys. Tech. J.*, vol. 35, pp. 661–684, May 1956.

[11] R. N. Hall, "Power rectifiers and transistors," *Proc. IRE*, vol. 40, pp. 1512–1518, Nov. 1952.

[12] J. Bardeen and W. H. Brattain, "Three-electrode element utilizing semiconductive materials," U.S. Patent 2 524 035 (Filed June 17, 1948), Oct. 3, 1950.

[13] R. W. Aldrich and N. Holonyak, Jr., "Five layer and junction bridging terminal switching device," U.S. Patent 3 476 993, Filed Sept. 8, 1959, Nov. 4, 1969.

[14] J. Bardeen, "Pre-history of the semiconductor laser," *Optoelectron.-Devices Technol.*, vol. 2, pp. 124–126, June 1987.

[15] R. N. Hall and W. C. Dunlap, "*P-N* junctions prepared by impurity diffusion," *Phys. Rev.*, vol. 80, pp. 467–468, Nov. 1, 1950.

Nick Holonyak, Jr. (LF'67) was born in Zeigler, IL, on November 3, 1928. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Illinois, Urbana, in 1950, 1951, and 1954, respectively.

He later was a Member of Technical Staff, Bell Telephone Laboratories, Murray Hill, NJ (1954 to 1955), and helped demonstrate the feasibility of diffused-impurity silicon devices, including transistors, oxide-masked transistors, p-n-p-n switches, and SCRs. He served with the U.S. Army Signal Corps (1955 to 1957), Fort Monmouth, NJ, and at Isogo-ku, Yokohama, Japan. In 1957, he joined the Advanced Semiconductor Laboratory, General Electric Company, Syracuse, NY, where he made contributions in the areas of power and signal p-n-p-n devices (including invention of the shorted-emitter and symmetrical SCR and thyristor switches—TRIACs, etc.), tunnel diodes, phonon-assisted tunneling (the initial observation of inelastic tunneling and the beginning of tunneling spectroscopy), halide transport and epitaxial growth of III-V compounds and compound mixtures (1960 to 1963), double injection and deep-impurity-level effects, junction luminescence (GaAsP LEDs), and III-V alloy semiconductor lasers (visible spectrum, GaAsP, 1962). His work, from 1960 to 1962, on GaAsP and the initial construction in 1960 of a p-n junction in this crystal system, and a visible-spectrum laser in 1962, led to the commercial introduction of GaAsP LEDs. He is the inventor of the first practical light emitting diode (the GaAsP LED), which also marks the beginning in the use of III-V alloys in semiconductor devices. Since 1963, he has been a Professor in the Department of Electrical and Computer Engineering, University of Illinois, and a member of the University of Illinois Center for Advanced Study. He has worked primarily on III-V semiconductors, III-V alloy crystal growth and the demonstration of red-orange-yellow-green stimulated emission in $In_{1-x}Ga_xP$, $In_{1-x}Ga_xP_{1-z}As_z$ and $Al_xGa_{1-x}As_{1-y}P_y$, stimulated emission on nitrogen trap transitions in the alloys $GaAs_{1-x}P_x$ and $In_{1-x}Ga_xP$, and heterojunctions in various ternary III-V's and in the quaternaries $Al_xGa_{1-x}As_{1-y}P_y$ and $In_{1-x}Ga_xP_{1-z}As_z$. He was the first to make quaternary III-V semiconductor devices (LEDs and lasers). His research, since 1976, has been concerned with quantum-well light emitters and lasers, and with impurity-induced layer disordering, which shifts lower gap quantum well layers to higher gap bulk crystal and serves as a basis for integrated optoelectronic devices. In 1990, he introduced higher temperature (400 °C) stable native oxides on Al-bearing III-V compounds and demonstrated their use in optoelectronic devices (LEDs and lasers). He was the first (1977) to construct p-n diode quantum well lasers (InP-InGaAsP, LPE) and first to achieve (1978) continuous (cw) room temperature (300 K) laser operation of quantum well heterostructures and superlattices, and later (1982) strained layer quantum well heterostructures. He and his students are the source of the name "quantum well laser." He is coauthor of the book *Semiconductor Controlled Rectifiers* (Englewood Cliffs, NJ: Prentice-Hall, 1964) and *Physical Properties of Semiconductors* (Englewood Cliffs, NJ: Prentice-Hall, 1989) and Editor of the Prentice-Hall series *Solid State Physical Electronics*. He has served on the boards of *Solid-State Electronics* (1970–1991) and the *Journal of Applied Physics and Applied Physics Letters* (1978–1980).

Dr. Holonyak received the General Electric Cordiner Award, and for his contributions to the field of visible-spectrum light emitting diodes and diode lasers, the IEEE Morris N. Liebmann Award, the John Scott Medal, the GaAs Symposium Award with Welker Medal, the IEEE Jack A. Morton Award, the Electrochemical Society Solid State Science and Technology Award, the Sigma Xi Monie A. Ferst Award, the IEEE Edison Medal, the Charles Hard Townes Award of the Optical Society of America, the National Academy of Sciences Award for the Industrial Application of Science, American Electronics Association 50th Anniversary Award, American Society for Engineering Education Centennial Medallion, Vladimir Karapetoff Eminent Members' Award of Etas Kappa Nu, 1995 TMS John Bardeen Award (The Minerals, Metals, and Materials Society), the 2000 IEEE Third Millennium Medal, the U.S. National Medal of Science, from Northwestern University an honorary doctor of science degree in 1992, was elected an honorary member of the Ioffe Physical-Technical Institute, St. Petersburg, Russia, an honorary doctor of engineering degree from Notre Dame University, in 1994, and the Japan Prize in 1995. In 1993, he was appointed (University of Illinois) the John Bardeen Chair Professor of Electrical and Computer Engineering and of Physics. He is a Member of the National Academy of Engineering, the National Academy of Sciences, the Russian Academy of Sciences, and Eta Kappa Nu. He is a Fellow of the American Academy of Arts and Sciences, the American Physical Society, and the Optical Society of America. He has served on the Editorial Board of the PROCEEDINGS OF THE IEEE (1966–1974).