The Road to the Transistor

By Jed Margolin

The transistor was invented in 1947 at Bell Labs by the team of John Bardeen, Walter Brattain, and William Shockley, for which they later received the Nobel Prize. The first transistor was a germanium point-contact transistor consisting of two thin electrodes in point-contact with the surface of a piece of germanium and with a third wire attached to the base.

Many of the early experiments consisted of breaking open a point-contact germanium diode and, by trial and error, finding a place where the third wire could be attached so the device would have a gain greater than one. [IBM Reference] At the time, in 1947, germanium as well as silicon point-contact diodes had been available for several years, and therein lies a tale.

It started in 1906.

1906 was one of those years that would shape the world for years to come, although few people, if any, realized it at the time. In October 1906, Greenleaf Whittier Pickard (the grandnephew of the poet John Greenleaf Whittier) received a patent on a method for receiving radio signals that included a silicon point-contact diode. {U.S. Patent 836,531 was filed on August 30, 1906 and issued November 20, 1906} Shortly afterwards, Henry Dunwoody received a patent on a system using a point-contact detector made of carborundum (silicon carbide). {U.S. Patent 837,616 was filed on March 23, 1906 and issued December 4, 1906}.

These detectors were far superior to what was being used. Before 1906 a radio system (or wireless telegraph as it was called) had a transmitter consisting of an inductor that was charged with current and then allowed to discharge across a gap (the spark-gap). One side of the spark-gap was grounded and the other side was connected to the aerial which was as long and as high as possible. The receiver also used an aerial (which was as long and as high as possible) which was connected to a device called a coherer.

The coherer, developed by Edouard Branly in 1890, was a glass tube loosely filled with metal filings and with contacts on each end connected in series with a battery and some kind of signaling or recording instrument. Normally the metal filings were too loosely packed to conduct current, but when an electromagnetic signal reached the coherer the metal filings would coalesce (cohere) and current would flow through the circuit to actuate the signaling or recording instrument.

Unfortunately, when the signal stopped, the metal filings did not return to their non-conducting state so the tube had to be tapped to be made ready for the next signal. Advanced coherers used the recording device to actuate a small hammer to perform the tapping.

The coherer was not very sensitive; it is amazing that such a bizarre device worked at all, but it was this device that was used as the detector when Marconi bridged the Atlantic in 1901. (Of course, his antenna was 400 feet long at the end of a kite.) You can see this in Marconi's U.S. patent {U.S. Patent 586,193 which was filed on December 7, 1896 and issued July 13, 1897. This was the U.S. version of his famous British patent number 7,777.}

So, the silicon point-contact diode was invented in 1906. Why did it take so long to add another wire to make the transistor? There were several reasons.

One reason is because of something else that was invented in 1906 which was the result of something that happened in 1883.

In 1883 Thomas Edison noticed that if a wire was inserted in a light bulb and connected to a positive voltage current would flow (the Edison Effect). Edison did not follow it up, which was unusual for him. It was probably because he was busy setting up the world's first commercial generating station. It is ironic that Edison discovered the means for converting AC to DC because he was so strongly opposed to alternating current. All of his generating systems were DC, and some of these systems were still being used in parts of New York City as late as the 1940s.

In 1904 Sir John Ambrose Fleming patented the two-electrode vacuum tube; in other words, the vacuum tube diode. He had worked for Edison and knew about the Edison Effect. {Fleming's U.S. Patent 803,684 was filed April 19, 1905 and issued November 7, 1905.}

And it was in 1906 that Lee de Forest invented the triode vacuum tube. {U.S. Patent 879,532 filed on January 29, 1907 and issued February 18, 1908} The triode vacuum tube was the first device that could amplify the power of an analog signal. (Relays used in the telegraph industry to 'regenerate' the Morse signals could be be considered amplifiers for digital signals.)

According to his patent:

" The objects of my invention are to increase the sensitiveness of oscillation detectors comprising in their construction a gaseous medium by means of the structural features and circuit arrangements which are hereinafter more fully described."

Lee de Forest considered his triode simply as a more sensitive type of diode detector. In 1911 Edwin Armstrong was the first to recognize that the triode could operate as an amplifier.

The triode vacuum tube did not catch on right away because they were expensive, short lived, and did not amplify all that much. In 1914 AT&T bought the rights to the triode for use in radio receivers and began to improve it. The First World War came and spurred development even more.

So, when the point-contact diode was invented in 1906 amplification was a new concept and people were more interested in working on something that already worked better, the vacuum triode.

Another reason was that the semiconductors require a high level of purity along with a controlled amount of impurity to create p and n type materials. The techniques and equipment didn't exist; besides, people didn't know that they needed this. The reason they didn't know was because no one knew how point-contact diodes worked.

It wasn't until the late 1930's when Walter Schottky in Germany, Nevill F. Mott in England, and Alexander Davydov in the Soviet Union came up with the theory of metal-semiconductor junctions: the semiconductor material becomes depleted of current carriers at the junction, which creates an effective barrier to equilibrium electron flow across the junction. The application of an electric field that reduces the barrier (a higher potential on the semiconductor side) permits electron flow, while reversal of that field further depletes the semiconductor of carriers, thus heightening the barrier to electron flow. And we have a diode rectifier.

Although people didn't know how point-contact diodes worked, it didn't stop them from manufacturing and using them. Until the 1920s radio was in Morse Code and most radios were operated either by radio amateurs (hams), commercial stations in the business of sending messages, or by the Navy. Vacuum triodes were too expensive for most hams, so most of their radios used these crystal detectors. Some couldn't afford commercially available crystal detectors and made their own with a chunk of galena crystal and a cat's whisker. Galena is the name for lead sulfide which is the principal ore of lead. The cat's whisker was actually a very fine wire. The procedure was to made a good connection to the galena for one terminal and to probe around with the cat's whisker for a 'sweet spot' for the other terminal.

The first vacuum tubes were not capable of producing very much power. Once vacuum tubes were developed to the point where they could produce significant amounts of power, the transmission of speech and music became possible and Broadcasting was created. The manufacturers of vacuum tubes saw it as a way to create demand for their product and so, many of the early broadcast stations were started by manufacturers of vacuum tubes. Companies like the Radio Corporation of America.

Even though vacuum tubes held the center stage, some research into semiconductors did continue.

In 1926 Germanium was added to the list of semiconductors. It was a short list consisting of silicon, selenium, and tellurium.

The copper oxide rectifier was invented in 1927 by L.O. Grondahl and P.H. Geiger. Copper oxide rectifiers featured good conduction of forward current but poor reverse voltage rating, about 6 Volts. However, because they could be built on plates instead of with a point-contact, the larger surface area meant they could handle large currents (about 7 Amps) and were used for awhile in battery chargers. U.S. Patent 1,640,335 was issued August 23, 1927 to Grondahl. For some reason Geiger's name is not on the patent. (Thanks to Reader Dave for finding the patent and noting that Geiger's name is not on it.) That's a good question, why isn't Geiger's name on the patent since the reference for Geiger is: ***L.O. Grondahl, P.H. Geiger, "A New Electronic Rectifier", Proc. AIEE Winter Convention, p.357, NY 1927***. [The American Institute for Electrical Engineers (AIEE) merged with the Institute of Radio Engineers (IRE) in 1963 to form the Institute of Electrical and Electronic Engineers (IEEE).] The patent is assigned to the Union Switch & Signal Company. George Westinghouse founded Union Switch & Signal in 1881, consolidating the assets of companies that had pioneered closed track circuits and interlockings. It operated as an independent company until 1917 when it became a subsidiary of Westinghouse Air Brake. It's still around as a wholly-owned subsidiary of Ansaldo Signal. http://www.switch.com/uss_history.html.

The largest and longest use of copper oxide rectifiers was in multimeters so that a DC meter could be used to read AC voltages and currents. Simpson Electric Co. still uses copper oxide rectifiers in their non-electronic multimeters. [1] The ones they use have a small surface area to minimize the capacitance which results in a rectifier with high bandwidth extending to the video range.

The only company that still manufactures copper oxide rectifiers is Edal Industries in East Haven, Ct.. The best copper for the rectifiers has always come from Chile. During the Second World War German U Boats made it difficult to get; domestic copper was tried but did not produce very good rectifiers. [2] This suggests that the copper contains a still unidentified impurity that makes the device work.

The selenium rectifier was invented in 1933 by C.E. Fitts. Although selenium has a higher forward current resistance than copper oxide, it has higher reverse voltage rating, typically 20 to 30 Volts, so it was practical to stack several of them for higher voltages. The plate architecture readily lent itself to stacking. Selenium rectifiers probably reached their peak in TV sets of the 1950's. Edal Industries (the last existing source for copper oxide rectifiers) also still sells selenium rectifiers, mostly for the high voltage power supplies for industrial electrostatic air cleaners. When a selenium rectifier's ratings are exceeded and they burn up they give off a rather nasty smell. Nowadays their use would probably be restricted in consumer products because of the toxicity of selenium. There are several other companies that will sell you selenium rectifiers such as Cougar Electronics (USA), GD Rectifiers (UK) and even NTE . The U.S. Geological Survey continues to track the production of selenium. Their latest report is very interesting.

On January 28, 1930, U.S. Patent 1,745,175 was issued to J.E. Lilienfeld for a "Method and Apparatus For Controlling Electric Currents." The patent shows an insulating material, such as glass, coated with a metal film having "unidirectional conductivity." There is no evidence that the device actually worked but it was a brilliant hunch. The material to actually build a transistor just didn't exist at that point.

In 1939 William Shockley and Walter Brattain made an unsuccessful attempt to build a semiconductor amplifier by inserting a tiny controlling grid into a layer of copper oxide. The Second World War put an end to their experiments.

But the Second World War produced tremendous gains in technology.

One was in the development of radar. The first practical radars operated at the relatively low frequencies around 300 MHz. Higher frequencies were better because they gave higher target resolution. With the invention of the magnetron in 1940 microwave transmitters (3,000 MHz) were possible.

Receiving was another matter. Vacuum tubes could not be used at these frequencies because of the various parasitic capacitances between the electrodes. Guess what came to the rescue? Our old friend the point-contact diode. The reason is that since it is a point-contact device the 'junction' capacitance is extremely low. The way they were used was that the signal would come in and go directly to the diode which was operated as a mixer to heterodyne the signal down to a lower frequency where tubes could operate.

Even though it didn't have any gain, there wasn't any alternative. In fact it was not until the early 1980's when the Gallium Arsenide Transistor was perfected that there was an alternative to the straight diode mixer other than an expensive parametric amplifier that had to be cooled with liquid nitrogen.

At the beginning of the Second World War the point-contact diodes that were available were not very reliable and they were not very good. That is why the U.S. Government set up a crash program to improve them. By the end of the War great strides were made in producing silicon and germanium of very high purity, which made it possible to produce very good, reliable, low noise, point-contact diodes for the radar receiver.

Russel Ohl came up with methods for producing high purity silicon {U.S. Patent 2,402,661 was filed on March 1, 1941 and issued June 25, 1046; U.S. Patent 2,402,839 was filed on March 27, 1941 and issued June 25, 1946 } Note that both patents were filed before the United States entered the Second World War and issued after it was over.

Another method for purifying Silicon was developed by Gordon Teal and Keith Storks. {U.S. Patent 2,441,603 was filed on July 28, 1943 and issued May 18, 1948. }

It was because of people like Ohl, Teal, and Storks that, in 1947, high purity germanium and silicon were available.

The point-contact transistor was invented in 1947 by John Bardeen and Walter Brattain. William Shockley (the team's leader) wasn't there at the time and did not receive credit for the invention, which pissed him off greatly. It's just as well. The point-contact transistor was difficult to manufacture and not very reliable. It also wasn't the transistor that Shockley wanted, so he continued to work on his own idea which led to the junction transistor which was easier to manufacture and worked better. A good history of these events can be found in *The Invention That Changed The World.* [3] (The invention that changed the world was radar.)

The transistor that Shockley developed was the junction transistor which was made by diffusing various chemicals into the germanium. When the process was done correctly you ended up with some fairly decent transistors.

The earliest patent to come from this period (that I could find) is U.S. Patent 2,524,035 which was filed on June 17, 1948 and issued October 3, 1950. Note that the patent lists only Bardeen and Brattain as the inventors and leaves out Shockley. Also note that the application was a continuation-in-part of a previous application filed February 26, 1948 (and abandoned). In patents, a continuation-in-part means you have improved your basic invention in some material way while the original patent application was pending and you want to claim the specific improvements. It does not mean the original application has to be abandoned; I wonder why it was. It would be interesting to see the original application, but since it did not issue into a patent it is not available on the U.S. Patent and Trademark Office Web site.

Shockley did manage to get a number of his own patents, such as U.S. Patent 2,569,347 which was filed on June 26, 1948 and issued September 25, 1951, and U.S. Patent 2,502,488 which was filed on September 24, 1948 and issued April 4, 1950. U.S. Patent 2,623,105 which was filed on September 21, 1951 and issued December 23, 1952 appears to involve a junction transistor.

Therefore, the transistor has a long and documented paper trail, and the first transistors were made from germanium. And you can still buy germanium point-contact diodes, just go down to your local Radio Shack and ask for some 1N34As (stock number 276-1123).

Bonus #1: One of the items reprinted in *Semiconductor Devices: Pioneering Papers* [4] is a letter to the editor of a magazine called Electrical World, dated 1907. Its author,
H.J. Round, reports,

"On applying a potential of 10 volts between two points on a crystal of carborundum, the crystal gave out a yellowish light. Only one or two specimens could be found which gave a bright glow on such a low voltage, but with 110 volts a large number could be found to glow. In some crystals only edges gave the light and others gave instead of a yellow light green, orange or blue. In all cases tested the glow appears to come from the negative pole, a bright blue-green spark appearing at the positive pole. In a single crystal, if contact is made near the center with the negative pole, and the positive pole is put in contact at any other place, only one section of the crystal will glow and that the same section wherever the positive pole is placed."

Carborundum is Silicon Carbide, the same material used by Dunwoody in 1906 to make the first silicon point-contact diode.

Captain H. J. Round was an electronics pioneer who received 117 patents and was Marconi's personal assistant.

Round's life has been chronicled by his grandson on his Web site.

It appears that Round discovered the LED. He even had a blue LED, 80 years before Cree Research used Silicon Carbide to make theirs. Of course, their blue LED was easier to manufacture and more reliable.

The full text of Round's letter is here.

Bonus #2: I went to a local mall which had a store that sold Pretty Rocks and Things and bought an 11 oz. chunk of what is purported to be Carborundum (Silicon Carbide), which is what Dunwoody used in inventing the Silicon point-contact diode and Round used in discovering the LED. (Eleven ounces was the smallest piece the store had.)

The surface of the rock is mostly very rough; however, one end is almost smooth.

Across the face of the smooth part are two distinct areas. One is very light colored; the other is dark. See Picture 1.

I made a crude holder using #10 copper wire. On one side I made a connection to the dark side of the flat face.

I also made an arm to hold a section of #30 wirewrap wire to probe the surface. The wire is coiled to form a crude spring to hold the probe against the Silicon Carbide.

The first experiment was to make a diode.

I used a 24VAC transformer with a 50 Ohm 10 Watt series resistor to limit the current. An oscilloscope connected across the resistor displayed the current.

By probing the rough surface I eventually found a "sweet spot" that showed definite diode characteristics. See Picture 2. Zero current is where the kink is between the half-cycles. The ratio of forward to reverse current is approximately 1:3 . (I didn't say it was a great diode.) The reddish lights are from the camera flash.

The second experiment was to make an LED.

For this I used a 30VDC current-limited benchtop supply and no 50 Ohm resistor. (The reason for the change will become apparent shortly.)

Although the #30 wire probe produced blue and blue-green points of light, I got better results with a sewing needle, holding it in my hand to control the force against the rock.

The question is, " Was I looking at real LEDs, or was I just seeing very small sections where I was heating the rock to incandescence, aided and colored by the rock's impurities?"

When I reduced the current through the probe, the color did not change, as I believe it would have if it had been simple incandescence.

In any event, the light produced was very dim. In Picture 3 I'm not sure if we are seeing a tiny blue light at the end of the needle or if it is just another mineral facet reflecting the light.

In Picture 4, taken with the lights out and without flash, there is indeed a small point of light. Kudos to my Kodak DC25. I was unable to see the point of light even in the optical viewfinder. Finding it in the picture was a big surprise.

The experiment is surprisingly easy to perform. Try it yourself. All you need is a chunk of Silicon Carbide and a power supply. And don't electrocute yourself.

Bonus #3: According to the Kirk-Othmer Concise Encyclopedia of Chemical Technology (John Wiley & Sons, 1985, page 201-202):

" Silicon carbide (SiC) is a crystalline material, with a color that varies from nearly clear through pale yellow or green to black, depending upon the impurities. It occurs naturally only as the mineral moissanite in the meteoric iron of Canyon Diablo, Arizona.

The commercial product, which is made in an electric resistance furnace, is usually obtained as an aggregate of iridescent crystals. It takes 6-12 kWh to produce 1 kg of crude, depending on the grade and recovery from furnaces. The loose black or green grain of commerce is prepared from the manufactured product by crushing, purification treatments, and grading for size."

and a little later,

"The metallurgical, abrasive, and refractory industries are the largest users or silicon carbide. it is also used for heating elements in electric furnaces, in electronic devices, and in applications where its resistance to nuclear radiation damage is advantageous."

So, let's see what we have.

Silicon carbide, which was used to make one of the first silicon diodes, occurs naturally only in a single meteor impact site and is resistant to nuclear radiation.

Maybe what hit Canyon Diablo, Arizona, wasn't a meteor.

Either way, we didn't get the message. According to the Encylopedia Brittanica, Silicon Carbide was discovered by the American inventor Edward G. Acheson in 1891 while he was trying to make artifical diamonds, apparently before it was discovered at Canyon Diablo.

References:

[1] Telephone conversation with Al Arbitter of Simpson Electric Co., August 11, 1993.

[2] Telephone conversation with Dino Vagnini of Edal Industries, August 11, 1993.

[3] The Invention That Changed The Word by Robert Buderi, Simon & Schuster, 1996, pages 322 - 333.

[4] *Semiconductor Devices: Pioneering Papers*, Edited by S.M. Sze, World Scientific Publishing, 1991, page 879

Other interesting references:

Early Radio Wave Detectors; Vivian J. Phillips; Peter Pelegrinus Ltd; 1980

Thomas A. Edison, A Streak of Luck; Robert Conot; Da Capo Press, Inc., 1979

200 Meters & Down, The Story of Amateur Radio; Clinton DeSoto; American Radio Relay League; 1936, 1981

This article, *The Road to the Transistor*, originally appeared as a chapter in a book I wrote in 1993 called *The Fabric of Technology: How We Got To Where We Are.*

Unfortunately, none of the publishers I contacted were interested in publishing it because they felt it would appeal only to a small audience.

Now that I have a Web site, publishers are irrelevant.

This version of *The Road to the Transistor* has been expanded by adding more patent references, the patents themselves, as well as links to several of the companies mentioned.

(2/8/2004 - links updated, 9/10/2005 - reference to Grondahl patent added)

Please send comments here