Two inventions shaped the semiconductor industry and the modern information-driven world: the transistor and the integrated circuit. Transistors were made from semiconductor materials, and integrated circuits contained thousands and millions of transistors literally etched out of semiconductor materials, all on a single “chip.” By 2003, the semiconductor industry was shipping close to one billion billion (10^{18}) transistors each year. Virtually all of those transistors were contained within integrated circuits. This note describes the discovery of the transistor and the key technological steps leading to the development of integrated circuits. In addition, it describes the manufacturing process used to fabricate integrated circuits and the basic economics of the process.

**From Tubes to Transistors**

The 19th century saw the development of both the science and technology of electricity. In 1800, Count Allesandro Volta built the first electric battery. By 1825, electromagnets were in use and the first electric bell sounded. In 1834, Faraday invented the electric motor. By the end of the century, Morse had invented the telegraph, Bell the telephone, Edison the lightbulb, and Marconi the radio.

The 20th century saw the development of electronics. Whereas 19th century science used electricity as a source of physical power and warmth, the new technology of electronics used electricity to control and shape electricity itself. Electric “signals” could be amplified, and electric signals could act to switch other electric signals on and off. The devices performing these acts were termed “active” and were chiefly the vacuum tube and the transistor.

Invented in the early 1900s, vacuum tubes, similar in size and shape to small light bulbs (Figure 1), gave birth to the radio industry and, by 1939, television. In 1945, the first high-speed computer (ENIAC) was built, containing 17,000 vacuum tubes. Each vacuum tube consumed 5 to 100 watts of power and had a useful life of perhaps 10 years. In a TV set with 10 tubes, this meant that on average, one of the tubes had to be replaced every 12 months. However, in a device with 10,000 vacuum tubes, a failure could be expected every few hours. Despite accomplishments like ENIAC, progress in computing and telephone networks was dramatically limited by the unreliability and power needs of vacuum tubes.

In 1945, Bell Telephone Laboratories launched a research initiative to replace vacuum tube technology. William Shockley, who had joined Bell Labs in 1936 (and left during the War to develop radar for the Navy), led this effort. Shockley hired Bardeen and Brattain to aid in the project. The most intense development was done by Bardeen and Brattain, working as a team. After hundreds of trials, Bardeen, a theoretician, had a key insight: existing ideas about how electrons behaved in crystals were wrong—there were
unexpected effects at the surfaces. This discovery led to the creation of the pointcontact germanium transistor (Figure 2) in December 1947.\(^1\) It was named “transistor” for “transfer resistor”.

The point contact transistor demonstrated basic principles, but was not a practical device. Its performance depended upon the exact placement of contact whiskers on the germanium and was strongly influenced by surrounding conditions. Eager to show that he could do better, Shockley shut himself away in a hotel room in Chicago and, in a period of intense effort, developed the idea for the junction transistor.

AT&T licensed the transistor for a nominal fee to other manufacturers, and waived any fees for its use in hearing aids, a recognition of the interests of the company’s founder, Alexander Graham Bell. RCA, Raytheon, the newly formed Texas Instruments, and other firms began to manufacture transistors. Raytheon’s CK722 (Figure 3) was the first mass-produced transistor. Intended for hearing aids, the lower-quality rejects were sold to firms and amateurs at low prices, enabling thousands of engineers and enthusiasts to become acquainted with the properties of the device.

### The Planar Process and the First IC

Shockley, disappointed that he did not share in the royalties from patents he had helped win, left Bell Labs to return to his boyhood home in Santa Clara, on the peninsula south of San Francisco, where radar-oriented science labs had also taken root. There he started his own company, Shockley Semiconductor Laboratories. Hoping to capitalize on the discovery of the junction transistor, Shockley put together a team of renowned scientists.

One year later, Shockley’s company had not sold a single device. Many employees lost confidence in the venture; there were complaints about Shockley’s paranoia and managerial incompetence. In 1956, eight key people left Shockley Semiconductor Labs (“The Traitorous Eight”): Gordon E. Moore, C. Sheldon Roberts, Eugene Kleiner, Robert N. Noyce, Victor H. Grinich, Julius Blank, Jean A. Hoerni and Jay T. Last. Investing $3500 of their own, they started development of a double diffusion technique for manufacturing multiple transistors from a single wafer. Fairchild Camera invested $1.5 million in the team, obtaining an option to buy the company. The option was exercised in October of 1957, creating Fairchild Semiconductor. Within six months the company was turning a profit due to an order from IBM for 100 transistors at $150 each.

Fairchild and Texas Instruments were both developing ways to produce many transistors from a single piece of germanium or silicon. These transistors would then be separated, packaged, and inserted into circuit boards where they were connected to the other critical elements of electronics circuits: resistors and capacitors. A number of people saw that it would be a great simplification if the intermediate packaging steps could be elimi-

---

\(^1\) Shockley, Brattain, and Bardeen received the Nobel Prize in physics in 1956 for the invention of the transistor. Bardeen received a subsequent Nobel Prize in 1972 for superconductivity.
What if transistors and resistors and capacitors could all be built on a single wafer and then connected? Such a device would be an "integrated circuit."

Jack Kilby, an engineer at Texas Instruments, had a free summer and worked on a project of his own choosing—trying to build a whole electronic circuit out of semiconductor materials. Using mesa technology, he etched out the shapes of transistors, resistors, and capacitors in a sandwich of silicon, then connected the tops with tiny gold wires. By September 1958 he had a working device, and demonstrated it to senior management. A tiny sliver of germanium with protruding wires produced an undulating sine wave on an oscilloscope. Kilby received a Nobel Prize in 2000 for this innovation.

In the same year, Jean Hoerni, at Fairchild, created the planar process for fabricating transistors. Hoerni’s innovation was the methodical use of silicon dioxide as an intermediate masking material in building up the layers of doped silicon that made up a transistor. His process proved to be much more effective than the older “mesa” process with which Fairchild had begun. In the “mesa” process, layers of different types of silicon were built up on a wafer, resulting in mounds, or mesas. In the planar process, by contrast, the surface remained even and covered by a layer of silicon dioxide. Different “layers” were created by etching away portions of the silicon dioxide, treating the exposed material, then restoring the silicon dioxide surface (Figure 4).

At Fairchild, Noyce saw the connection between the planar process and the possibility of truly integrated circuits. Kilby’s approach required the intricate application of fine gold wires to connect the individual components etched on silicon. Noyce saw that the planar process could extend to the construction of metal layers, making much greater miniaturization possible. Patent disputes over the rights to the integrated circuit lasted ten years. In the end, Noyce and Kilby were recognized as joint inventors.

The power of integrated circuits was, at first, their miniaturization and lower cost per device. Use in the aerospace industry and computing industry became widespread. The first dramatic commercial success in integrated circuits was the series of TTL logic chips introduced in the mid-1960s by TI and Fairchild. Each chip contained 5-50 transistors arranged to perform specific logic functions (e.g., AND gates, 8-bit latches, etc.)

The first truly new device enabled by the IC was the small desktop calculator, developed by Kilby at TI in 1967. The ability to carry out high-precision arithmetic, the instant computation of square roots, logarithms, and powers, at one’s desk was revolutionary. By the early 1970s, the hand calculator had appeared.

Further developments in digital devices required advances in speed, memory, and display. Figure 5 shows Fairchild’s 4100, a 256 bit RAM memory “chip” introduced in 1970. The 16x16 matrix of memory cell transistors can be
clearly seen. The physical device measured about 100 mils (0.1") on each side, and was about 10 mils thick. The chip was an integrated circuit containing approximately 1000 transistors on a piece of silicon only as large as the letter "o" on this page. From about this year on, most transistors would be produced as elements of integrated circuits. The single packaged transistor would appear only in special applications.

Semiconductor Physics

Electrons are electricity. Atoms with one electron in their outer (valence) shells are good conductors—elements like copper, silver, and gold. These elements are good conductors because the single outer electron is not bound in a lattice and easily moves from place to place in the material.

Silicon and germanium have 4 electrons in their outer valence shell. When these elements form crystal lattices, each atom bonds with 4 nearby atoms; by sharing electrons, each atom is surrounded by 8 electrons (Figure 6). Physical law says that 8 valence electrons is a stable number. Such materials are very poor conductors because there are no electrons that are not tied down in the crystal lattice.

Phosphorus has 5 valence electrons. If a few atoms of phosphorus are mixed into a silicon crystal, they take up positions in the crystal. The 5th valence electron is not needed in the crystal structure and is free to roam (Figure 7). The “doped” silicon becomes a good conductor. Phosphorus doping creates an n-type semiconductor because the extra electrons are negatively charged. It is a “semiconductor” because its conductivity depends sharply on the amount of doping and other factors.

Boron has 4 valence electrons. If a few atoms of boron are scattered in a silicon crystal, there is a “missing” 4th electron at each place in the lattice where a boron atom lodges. The missing electron is called a “hole.” One of the critical discoveries in semiconductor science was that these holes conduct electricity just as well as electrons. Boron-doped silicon is p-type because the charged carriers are positively charged holes.

The basis of the transistor is the pn junction. Although p-type and n-type semiconductors are each conductors, when they are brought together the result is a device that only conducts in one direction. The conduction of electricity through the pn junction is a very subtle process and depends strongly on the chemistry of the materials and the electric fields that are present. The latter fact allows engineers to design devices in which the flow of electricity is itself controlled by other flows of electricity or electrical charges. When the flow of electricity is controlled by other flows of electricity, one has an active device that can be used for amplification or as a switch.

2 A mil was 1/1000 of an inch, so 100 mils = 0.1". Just as there were 39.37 inches per meter, there were 39.37 mils per millimeter.
3 A doped semiconductor has typically 5x10^{17} free electrons per cm^3, compared to 10^{10} in the intrinsic silicon.
Device Innovations

The first integrated circuits were built out of bipolar transistors. The mechanism in a bipolar transistor was that the current (electrons) flowing between two terminals was controlled, or switched, by the current flowing into a third terminal. A small signal applied to third terminal can cause a larger, or amplified, signal to appear between the first two. Bipolar transistors were fast, but as more and more were packed into a small space there was a problem. The control current used too much power and generated unwanted heat.

**MOSFETs**

Another, simpler, way of creating a transistor had been understood theoretically since 1925. Instead of controlling the transistor by passing a current into a third terminal, an electrical charge could be applied to a “terminal” which was separated from the material by a thin insulator (Figure 8). This terminal, called the *gate*, controlled the current flowing between the other two, called the *source* and *drain*. This structure was called a MOSFET: metal oxide silicon field-effect transistor. RCA produced the first commercial MOS device in 1963.

Intuitively, the operation of a MOSFET transistor is analogous a water faucet. Water would flow from the water (source) into the sink (drain). The water can flow, however, only if the faucet is turned on. In a MOSFET, electrons can flow from the source to the drain only when the gate is charged.

MOSFET transistors could also be arranged to form logic circuits. Logic gates and logic circuits performed simple logical (Boolean) operations by coding high and low voltages as 0 and 1, or as “true” and “false.” For example, the task of adding two binary numbers, say 0101 and 0110, obtaining their sum 1011, was essentially a task in “logic” and could be performed with and electronic circuit containing about 20 transistors. The circuit could be built from individual logic gates, commercially supplied as integrated circuits containing 5-10 transistors each. Or it could be created as a single integrated circuit. Or, it could simply be integrated into a much more complex device composed of thousands or millions of transistors on a single chip.
**CMOS**

In 1963 Frank Wanlass (Fairchild) patented the idea of complementary-MOS (CMOS) device. In a CMOS structure, two complementary MOS devices were created side-by-side. The devices were complements in that one transistor was turned "on" by a positive charge on its gate whereas the other was turned "off" by a positive charge and on by a zero voltage.

Although CMOS circuits tended to use more transistors, they reduced power consumption to new lows. Figure 9 shows a diagram of a CMOS inverter—a circuit that turned a high voltage ("1") into a low voltage ("0"). A high voltage at A turned the lower device on and the upper device off, connecting A to ground. A low voltage at A, reversed the process, connection A to V+.

By toggling the output between the high voltage and the low without running any current through a resistor, CMOS designs cut the power needs of integrated circuits by a factor of 1 million, making it possible to assemble tens of thousands of transistors without drawing undue power. By the 1990s, virtually all large-scale integrated circuits were CMOS.

**DRAM**

Another key innovation was the redesign of memory cells. The traditional way to store a "bit" was the two or four-transistor flip-flop. Robert Dennard, an IBM engineer, realized that a capacitor, storing charge, could represent the "bit" and a single MOSFET could charge or discharge the capacitor. This concept greatly simplified the design of RAM.

**Microprocessors**

In 1968 Robert Noyce and Gordon Moore quit Fairchild and started a new semiconductor company—Intel. They each invested $250 thousand and raised $2.5 million in venture capital. Intel's first products were memory chips.

When a Japanese company asked for a redesign of the chips used in its desktop calculators, Intel engineer Ted Hoff suggested that Intel build a small integrated microprocessor that would be programmable. That way, its functions could be changed by writing new software rather than redesigning the chips. Circuit designer Frederico Fagin led a team that, in 1971,
delivered the Intel 4004, the world’s first microprocessor. It measured 1/8” by 1/16” and contained 2300 transistors. By itself, it was as powerful as the world’s first digital computer, ENIAC, built in 1947, weighing tons.

A microprocessor was a “central processing unit” (CPU) built on a single chip. The microprocessor contained a few internal bytes of memory, called registers, and had connectors (pins) for input, output, and reading the contents of an external “memory” chip. It recognized some patterns of data as “instructions.” By coding these instructions properly, called writing software, the microprocessor could be made to perform arbitrarily complex logical functions.

The Intel 4004 (Figure 10) was not really a new concept, as it simply integrated onto one chip the functions that had been performed by a number of individual chips. Nevertheless, it represented a break with the past, more sharply delineating the line between hardware (the chip) and software (the specific code making the chip perform a particular task). Depending upon the program, a microprocessor could run a calculator, read a keyboard, control a room temperature, control an elevator, time an engine, control a printer, and so on.

**Manufacturing Overview**

Integrated circuits were constructed by applying processes to wafers of silicon. These processes created the thousands or millions of transistors which comprised an integrated circuit. The production process was more like printing or photography than mechanical manufacturing.

The basic process was to obtain a silicon wafer from an ingot of pure silicon. The wafer was then inscribed to demark a number (usually in the hundreds) of rectangular dice—each individual die being a potential chip, itself containing thousands or millions of transistors. Once fabricated and tested, the die became a chip by being mounted in a package, connected with tiny leads to external pins, and sealed.

The dice were not dealt with individually. Instead, the construction of integrated circuits involved the repeated application to the entire wafer of a relatively small number of processes. The process steps created and treated layers on the wafer. The build up of these layers was controlled by the cooperative action of photolithography, which created patterns of resist on the wafer surface, and chemical processes which added material to the wafer or removed material from its surface. Early integrated circuits used about 10 layer-building, or masking, steps. State-of-the-art wafers might use 25-30 layers. Each layer normally took 2 to 3 days of processing time.

**Growing Silicon**

Silicon ingots were grown (Figure 11) from a melt of pure silicon mixed with tiny amounts of active dopants (antimony, arsenic, boron, and phosphorus). A tiny “seed” of perfect crystalline silicon was lowered into a 1400 °C melt. As the liquid silicon was gradually cooled, new crystals formed around the seed. Slowly pulling the seed into a cylinder produced the ingot of perfect crystalline silicon. The ingot was then sliced, using a diamond saw, into wafers. In 2003, most wafers were 200mm (8 inches) in diameter, with

---

4 The kit for each calculator consisted of one 4004 cpu, two 4003 serial IO chips, two 4002 320 bit dynamic RAMs, and four 4001 2048 bit ROMs.
state-of-the-art wafers being 300mm (12 inches) in diameter. In 2003, an 8" wafer might be priced at $65. Adding an "epitaxial" layer of very high-quality silicon crystals might add another $25-35 to the price.

Photolithography Masks

The fabrication of integrated circuits was centered on a two-step process of masking. In the first step, photolithography was used to transfer a geometric pattern from a photomask to a light-sensitive photoresist that had been spread on the surface of the wafer. The resist was then developed, leaving a pattern of resist on the surface of the wafer that duplicated the original image on the photomask. This pattern of resist then acted itself as a mask—it protected the selected areas of the wafer from a process step such as etching, implantation, or oxidation.

Each layer required a separate mask defining the regions to be worked on that step. As feature sizes shrunk, it became increasingly important to control the fine-structure of the transistors. As more transistors were added to each circuit, increasing layers of metal interconnect were required. Both forces led to an increase in the number of masking steps. By 2003, advanced designs used 22-30 masking steps.

Masks were created by computer-aided design tools. These devices used data about the integrated circuit design to control an electron-beam which...
drew the pattern on a resist. Developing and plating led to a chrome mask on a glass or quartz base. In the early days of integrated circuit fabrication, the masks were applied directly to the wafer because any gap led to unfocused images. This contact process led to problems in mask wear and wafer contamination and was replaced by optical projection systems. Projection systems used a mask which contained the wafer pattern enlarged by a factor of 4 or 5. Laser light and complex lens systems reduced and focused the image onto the wafer. The mask itself was normally mounted inside a protective transparent frame.

Because the chips on a wafer were replicas of one another, it was not necessary for a mask to image the whole wafer. Masks imaged a number of chips in a block, called a reticle (Figure 13) and then used a step-and-repeat process or a scanner process to image the whole wafer.

The quality and accuracy of the mask was of critical importance for good results. Tiny imperfections in a mask would render all the chips produced by that image useless. Even if each mask was perfect, their overlaid images had to match for the build-up of 20 or more masking steps. Finally, the lithography equipment had to align each image accurately, with tolerances of a fraction of a wavelength of light.

The cost of a set of reticles depended upon the feature size being imaged. In 2003, a typically state-of-the-art mask set for 130nm technology might require 16 “critical tolerance” reticles, costing $32-38,000 each, and 10 “non-critical” reticles, costing $22-28,000 each. Each reticle had a life of about 4000 exposures (less than one week) before it had to be replaced. The cost per wafer processed of this whole set was, therefore, on the order of $100.

**Lithography Process:**

- **Photoresist** was an organic chemical that was light sensitive. Several drops were applied to the center of the wafer, and the wafer was then spun (3000 rpm) to spread it evenly over the surface. The thickness of the resist was approximately 1 micron. The resist was lightly baked in an oven to remove solvents.

- The patterns to be imposed on the wafer were projected onto the resist by passing light through a mask using a stepper or scanner. Where the mask (chrome on glass) was clear, light broke chemical bonds in the resist.

---

5 The five images used in this section were taken from Hutcheson & Hutcheson, "Technics in the Semiconductor Industry," *Scientific American*, 1997.
• The development of the resist, like the development of photographic film, required immersion in a chemical solution. This step removed the softened resist where the mask had been transparent. The remaining resist was baked in an oven to harden it. Then, the wafer was rinsed in deionized water and dried in extremely clean conditions.

• Processing step. The purpose of photolithography was to place a pattern of resist on the wafer which selectively protected the material below from whatever process was going to take place. The various processes that now could be performed on the exposed portions of the wafer included metal deposition, polysilicon deposition, plasma etching, wet etching, ion implantation, etc.

• The final step in the cycle was the removal of any remaining hardened photoresist. The normal method was exposure to a high temperature plasma. In many cases, the resist itself was not sufficiently robust to act as a mask for a process. For example, resists could not stand up to high temperature phosphorus doping, a process needed to create large n-type regions on the wafer. In such cases, a two-stage procedure was employed. The first step was to build a layer of silicon oxide on the wafer by exposing it to hot oxygen. Over time, the oxygen bonded with the silicon, “growing” a layer of SiO₂. After the oxide layer was built, photolithography built a pattern of resist that allowed the SiO₂ to be etched away in some places. With the removal of the resist, the oxide pattern left on the wafer acted like a super-resist. The wafer was exposed to a beam of phosphorus atoms, creating n-type doped areas wherever the oxide had been etched away.

**Wafer Processing Steps**

Photolithography defined the regions of the wafer on which treatments would occur. The actual treatments employed were these:

• **Etching** involved the removal of material not protected by a layer of resist. Wet etching used chemical acids (e.g., hydrofluoric acid). Dry etching was a process that used high energy atoms to chip away at the exposed surface.

• **Oxidation** was applied to either the whole wafer or to an exposed portion. Exposing the silicon to hot oxygen and hydrogen allowed SiO₂ to form (glass). In this process, the surface actually absorbed oxygen, swelling in size. The properties of the oxide layer could be adjusted by the temperature used to grow the oxide.

• **Implantation** was the process of doping silicon to be n-type or p-type. Diffusion implantation used hot gases and a furnace. An alternative method was ion implantation, wherein ions (charged atoms) were accelerated and driven into the surface. Ion implantation allowed
simultaneous control over the depth of the implantation and the concentration of dopants. But a consequence of ion implantation was damage to the crystal structure of the silicon. To remedy this problem, the silicon was annealed at high heat. This smoothed the surface, restoring the lattice structure of the atoms.

- **Deposition** involved adding a material layer to the complete wafer. Metal interconnects were created by sputtering a layer of aluminum, tungsten, or copper onto the wafer. (Subsequent etching steps cut away portions of the metal layer.) CMOS processes made use of polysilicon layers which were deposited using a chemical deposition process. Silicon nitride was often used as a temporary buffer during the formation of insulating oxides, and was applied using chemical vapor deposition methods.

- **Planarization** involved smoothing and flattening the surface, especially as multiple layers of metal conductors were added.

### Metal Interconnections

The metal interconnects among these devices were also constructed with these processes. As engineers integrated more transistors onto logic chips, more complex problems of interconnection arose. By 2003, complex logic chips contained 5 to 7 layers of metallic interconnects. Figure 14 shows a microphotograph of the interconnects on a DRAM chip where all the protective oxide has been etched away to reveal the metal conductors. Each of the four layers of metal required several process steps, as did the plugs which acted as vertical connectors among levels.

![Figure 14](Source: IBM)

### Yield

After a wafer was processed, it was probe-tested. The purpose of this step was to ascertain the quality of each die on the wafer before undertaking the costs of more han-
dling, shipping, and packaging. The probe-tester was a machine which held the wafer and stepped an electronic test probe over the dice. At each die, tiny needles were pressed onto the edge-connectors of the die and the integrated circuit was electronically tested. The percent of chips passing this test was called the *wafer-probe yield*. This yield was a key determinant in the cost structure of integrated circuit manufacturing.

Costs for DRAMs and microprocessors are estimated in Exhibit 1. A 1 Mbit DRAM chip contained about 1 million transistors. In 1988, the cost of processing a 150mm wafer (including the allocation of fixed overhead) was about $400. Each wafer contained, in this example, 546 dice (potential DRAM chips), giving a total processed cost per die of 400/546 = $0.73. However, the wafer-probe step, which tested the dice, had a yield of only 67%, so that the cost per good die rose to 400/366 = $1.09. Placing the chip into a package and connecting it to the external pins or balls (packaging) plus the final test steps added $0.31 in cost per die processed and resulted in the discard of another 44 dice per wafer. The final cost per chip was $1.59.

Exhibit 15 shows the relative importance of wafer processing, loss to probe-testing, and assembly, final test, and final test loss. Thirty-two percent of the final product cost was due to yield losses at the wafer-probe step.

On start-up, yield rates in the probe-step could be as low as 5-10%. Most fabs hoped to get yields up into the 90-98% range by the time their processes had "burned in." Nevertheless, the industry showed a great deal of variance in yield rates. The Berkeley benchmarking study\(^6\) showed DRAM yields varying from 35 to 90% across multiple fabs and production lines in 1997-98. Most of the data clustered within the 85-95% range.

Defects in dies had, historically, had three major sources: (a) 2% from defects in materials, including the wafer itself, (b) 8% from processing errors, including photolithography and mask defects, and (c) 90% from contamination.

Contamination meant unwanted particles adhering to the wafer or mask, causing defects to be etched or deposited onto the patterns on the wafer. If a defect was large enough, and fell in the wrong place, it would cause a critical failure. At the simplest level, two conductors might be improperly connected, shorting the circuit, or a connector might be interrupted, breaking a circuit. More subtly, device performance might be adversely affected by a defect, so that its operation was slow, temperature sensitive, or erratic.

In turn, contamination came from air particles, human presence, equipment, and processes. During the early years of the integrated circuits industry, a great deal of effort was expended on cleanroom technology—the methods for reducing particulate matter. Cleanroom practice started with HEPA air filters and added positive pressure control, so that uncleared air did not flow into the area. To this were added special clothing and handling systems. The old cleanroom standard classified cleanrooms according to the

---

number of particles 500nm and larger per cubic foot. Thus, a class 1 cleanroom had only 1 particle per ft³ whereas a class 100 cleanroom had 100 particles.

According to the new U.S. Federal Standards, if the ISO class of a clean room was the integer \( N \), then there had to be fewer than \( 10^N \) airborne particles per cubic meter having a size of 100nm or larger (Table 3). Both systems of measurement were in use.

Studies had shown that a person sitting still, properly garbed, emitted 100,000 particles per minute of size 0.3 mm or larger. Simple movements, like toe tapping, raised the rate of emission to 1,000,000 particles per minute! Thus, in a class 1 cleanroom, the limiting factor became the number of people per square-meter and the time the wafer was exposed to them.

The cost of cleanrooms rose dramatically with the stringency of its ratings. Class 100 cleanrooms cost about $1200 per ft² in 2002, whereas a class 1 cleanroom cost upwards of $10,000 per ft². A class 0.5 cleanrooms cost $25,000 per ft² and more.

Each wafer was divided into a number of dice. Each die was a potential chip (e.g., a DRAM or a microprocessor, a digital signal processor, a graphics controller, or a custom logic device.)

The gross yield was the number of dice on the wafer testing “good” divided by the total dice on the wafer. As can be seen in Figure 16, the larger a die, the more chance it had of encountering a particle of a killer size or larger. Much effort had been expended to quantify this association.

The simplest yield formula used in the industry was based on a model in which defects were distributed randomly on the wafer. Letting the expected number of killer defects per mm² be \( D \) (the defect density) and letting the active area of a die be \( A_d \) mm², then the expected number \( \gamma \) of killer defects on one particular die was

\[
\gamma = A_d D .
\]  

(1)

The probability of failure is the probability that \( \gamma \geq 1 \). The standard assumption was that (1) represented the mean of a gamma distribution. Integration gave this formula for yield (the probability that \( \gamma = 0 \)):

\[
Y = \left(1 + \frac{A_d D}{\alpha}\right)^{-\alpha} = \left(1 + \frac{\gamma}{\alpha}\right)^{-\alpha}.
\]  

(2)

Here, \( \alpha \) was the “clustering” parameter and reflected the degree to which particle positions were correlated. Normally, \( \alpha \) had a value in the neighborhood of 2, but ranged from 0.3 to 5.
Equation (2) implied that yield falls as the area of a die increases. Conversely, yield rises as the area of a die decreases (other factors being constant).

**Test and Packaging**

Once the wafer had been probe-tested, the individual dice were sawed apart using a diamond blade. Each die was mounted in a frame which became an integral part of the final package. At the edge of each die were bond pads which were its points of connection. Connections were made between these bond pads and spots on the frame that connected to the external pins.

As chips became more complex, more pins were required to attach to external circuits. In part, more power pins were required so that each part of the device was fed power, preventing internal lags. In addition, the integration of more functions and wider data buses (16 bit to 32 bit to 64 bit) meant more pins. In general, the number of pins on a logic chip tended to increase as the log of the number of transistors. However, sufficient modularity could break this rule of thumb.

In the mid 1990s, new styles of packaging arose to meet the needs of denser pin structures. Instead of having pins just on the edges of a chip, the whole undersurface of the package was covered with pins (pin grid array).

As densities increased, the creation of a connection between the integrated circuit and the outside world became more complex. To act as an intermediary between the scale of the chip and that of the circuit board, substrates became more complex. The substrate was like a miniature circuit board, handling the wiring from the chip bond pads to the pins or solder balls that are its external connections. By 2003, substrates were constructed with lithographic techniques to fine tolerances. The connections between the substrate and the outside circuits was made with either pins or a ball grid array (BGA).
BGA packages were directly connected to a dense circuit board or to another large substrate which integrated two to five BGA chips. These new structures provided a level of integration intermediate between the IC and the circuit board. Many observers noted that as feature sizes became smaller, more effort and cost went into packaging and connections.
## Exhibit 1
### Integrated Circuit Cost Structures

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Product</td>
<td>DRAM 1 Mbit</td>
<td>Pentium II 256 Mbit</td>
<td>DRAM 512 KB</td>
<td>Pentium 4 Cache</td>
</tr>
<tr>
<td>A Logic Transistors (millions)</td>
<td>0</td>
<td>7.5</td>
<td>0</td>
<td>41.9</td>
</tr>
<tr>
<td>B Memory Transistors (millions)</td>
<td>1.024</td>
<td>0</td>
<td>256</td>
<td>13.1</td>
</tr>
<tr>
<td>C Total Transistors (millions)</td>
<td>1.024</td>
<td>7.5</td>
<td>256</td>
<td>55</td>
</tr>
<tr>
<td>D Wafer Diameter (mm)</td>
<td>150</td>
<td>150</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>E Feature Size (nm)</td>
<td>1000</td>
<td>800</td>
<td>130</td>
<td>130</td>
</tr>
<tr>
<td>F Wafer Starts/Week</td>
<td>5600</td>
<td>4500</td>
<td>6900</td>
<td>6900</td>
</tr>
<tr>
<td>G Number of Masking Steps</td>
<td>10</td>
<td>17</td>
<td>25</td>
<td>23</td>
</tr>
</tbody>
</table>

### Cost of Processed Wafer ($/Wafer)

<table>
<thead>
<tr>
<th>H Raw Wafer</th>
<th>J Labor</th>
<th>K Supplies</th>
<th>L Equipment</th>
<th>M Facilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>132</td>
<td>116</td>
<td>132</td>
<td>12</td>
</tr>
<tr>
<td>60</td>
<td>163</td>
<td>70</td>
<td>497</td>
<td>38</td>
</tr>
<tr>
<td>60</td>
<td>102</td>
<td>261</td>
<td>962</td>
<td>93</td>
</tr>
<tr>
<td>85</td>
<td>174</td>
<td>522</td>
<td>1030</td>
<td>89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N Total</th>
<th>P Die Size (sq-mm)</th>
<th>Q Defect Density (per sq cm)</th>
<th>R Cluster Parameter</th>
<th>S Wafer Exclusion (mm)</th>
<th>T GrossDie()</th>
<th>U Gross Die</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>29</td>
<td>1.5</td>
<td>3</td>
<td>2</td>
<td>551</td>
<td>48</td>
</tr>
<tr>
<td>828</td>
<td>294</td>
<td>0.4</td>
<td>1</td>
<td>2</td>
<td>510</td>
<td>510</td>
</tr>
<tr>
<td>1478</td>
<td>55.1</td>
<td>0.2</td>
<td>2</td>
<td>3</td>
<td>185</td>
<td>185</td>
</tr>
<tr>
<td>1900</td>
<td>145.8</td>
<td>0.35</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V Test/Control Sites</th>
<th>W U-V Gross Available Die</th>
<th>X (1+P*Q/R)^R Probe Yield</th>
<th>Y int(p*n) Good Die</th>
<th>Z N/Y Cost/Good Die ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>546</td>
<td>67%</td>
<td>366</td>
<td>1.09</td>
</tr>
<tr>
<td>0</td>
<td>48</td>
<td>46%</td>
<td>22</td>
<td>37.64</td>
</tr>
<tr>
<td>5</td>
<td>505</td>
<td>90%</td>
<td>453</td>
<td>3.26</td>
</tr>
<tr>
<td>3</td>
<td>182</td>
<td>66%</td>
<td>120</td>
<td>15.83</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a Package Cost ($/die)</th>
<th>b Packaging Yield</th>
<th>c int(a*b) Good Packages</th>
<th>d Final Test Cost ($/die)</th>
<th>e Final Test Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>0.98</td>
<td>358</td>
<td>0.16</td>
<td>0.9</td>
</tr>
<tr>
<td>2</td>
<td>0.99</td>
<td>21</td>
<td>0.16</td>
<td>0.95</td>
</tr>
<tr>
<td>0.6</td>
<td>0.99</td>
<td>448</td>
<td>0.15</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>f int(c*e) Good Chips</th>
<th>g (N+Za+dc)/f Cost/Good Chip ($)</th>
<th>h g/B Cost/Mbit ($)</th>
<th>g/C Cost/M Transistors ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>322</td>
<td>1.59</td>
<td>1.59</td>
<td>1.59</td>
</tr>
<tr>
<td>19</td>
<td>42.31</td>
<td>0.016</td>
<td>5.64</td>
</tr>
<tr>
<td>443</td>
<td>4.09</td>
<td>0.016</td>
<td>0.470</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>h g/B Cost/Mbit ($)</th>
<th>g/C Cost/M Transistors ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.59</td>
<td>5.64</td>
</tr>
<tr>
<td>0.016</td>
<td>0.470</td>
</tr>
</tbody>
</table>
Notes to Exhibit 1

Costs have been estimated by combining technical information on devices with a number of industry sources, including Intel, Merrill Lynch, Goldman Sachs, IC Knowledge, SemaTech’s “Wafercost Calculator,” interviews with Mr. Howard Dicken, and the “Spreadsheet Model for Fab Economic Analysis,” created by Leachman, Ding, and Sato-Misaw’s UC Berkeley.

The yield equation used in Exhibit 1 was that given by (1) and (2) in the text.

The GrossDie() function calculates the number of die that fit on a wafer. The VBA function implemented is as follows:

```vbnet
Function DieCount(height As Double, width As Double, diameter As Double) As Integer
' (c) 2003 Richard P. Rumelt
' This function computes the number of die that can be fit onto a wafer. It takes the 'height and width of a single die and the diameter of the wafer as inputs and returns an 'integer count.
Dim h As Double
Dim w As Double
Dim row As Integer
Dim columns As Integer
Dim rowmax As Integer
DieCount = 0
h = height
w = width
If h > w Then
    h = width
    w = height
End If
If w < diameter Then
    rowmax = Int(Sqr(diameter ^ 2 - w ^ 2) / h)
    For row = 1 To rowmax
        columns = Int(Sqr(diameter ^ 2 - (row * h) ^ 2) / w)
        DieCount = DieCount + columns
    Next row
End If
End Function
```
### Exhibit 2

#### Die Area Computations

<table>
<thead>
<tr>
<th>Year</th>
<th>Product</th>
<th>2002</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>Pentium 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>256 Mbit</td>
<td>512KB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### DRAM Transistor Area

- **Cell area factor**: 7
- **Cell area (sq-microns)**: 0.1183
- **Cells (million)**: 256
- **Total DRAM Cell Area (sq-mm)**: 30.2848

#### Logic Transistor Area

- **Cell area factor**: 80
- **Cell area (sq-microns)**: 1.352
- **Cells (million)**: 41.944
- **Total Logic Cell Area (sq-mm)**: 56.71

#### SRAM Transistor Area

- **Cell area factor**: 16.8
- **Cell area (sq-microns)**: 1.352
- **Cells (million)**: 13.056
- **Total SRAM Cell Area (sq-mm)**: 17.65

#### Total Active Cell Area (Sq-mm)

- **Silicon Efficiency**: 0.55
- **Die Area (sq-mm)**: 55.1

<table>
<thead>
<tr>
<th>Year</th>
<th>Product</th>
<th>2002</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>Pentium 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>256 Mbit</td>
<td>512KB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Notes to Exhibit 2

Cell area factors were multipliers that related the square of feature size to the area taken up by a transistor. Thus, if the feature size was $F$ microns, and the cell area factor was 20, then the area taken by one transistor was $20F^2$ square microns. There were 1 million square microns in 1 square millimeter, so 1 million transistors would take up $20F^2$ square millimeters.

The active cell area on an integrated circuit were surrounded by a relatively empty margin which housed the metal leads running out to the terminals on the edge of the chip. Often, one-half of the area of the chip was devoted to this margin. The proportion of the chip used for actual active devices was termed the silicon efficiency.

DRAM memory normally used 1 transistor per bit. The actual data was stored as a charge on a capacitor. SRAM memory was much faster, but required 3 transistors per bit of storage. The fast cache memory included on microprocessors and logic chips was generally SRAM. A 512KB L2 cache required $0.512 \times 8 \times 3 = 12.3$ million transistors. The 13.056 transistor count for SRAM on the P4 (above) included smaller caches as well, but excluded the instruction pipeline storage on the chip.
Exhibit 3

US ISO 14644-1, *Cleanrooms and Associated Controlled Environments*

<table>
<thead>
<tr>
<th>CLASS</th>
<th>Number of Particles per Cubic Meter by Micrometer Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1 um</td>
</tr>
<tr>
<td>ISO 1</td>
<td>10</td>
</tr>
<tr>
<td>ISO 2</td>
<td>100</td>
</tr>
<tr>
<td>ISO 3</td>
<td>1,000</td>
</tr>
<tr>
<td>ISO 4</td>
<td>10,000</td>
</tr>
<tr>
<td>ISO 5</td>
<td>100,000</td>
</tr>
<tr>
<td>ISO 6</td>
<td>1,000,000</td>
</tr>
<tr>
<td>ISO 7</td>
<td></td>
</tr>
<tr>
<td>ISO 8</td>
<td></td>
</tr>
<tr>
<td>ISO 9</td>
<td></td>
</tr>
</tbody>
</table>

The older cleanroom nomenclature dealt in 0.5 µm particles per cubic-foot. There are 35.2 ft³ per cubic-meter. Under the old system, ISO Class 3 was “Class 1” in that it had 35.2 particles per cubic meter, or 1 particle per cubic foot. Similarly, ISO Class 4 was called “Class 10,” ISO Class 5 was called “Class 100,” and so on.