by Federico Faggin

The Making of the First Microprocessor

The Intel 4004 CPU-on-a-chip was developed under pressure on an extremely tight schedule—and it worked.

Although I didn't know it at the time, my early work experience turned out to be absolutely invaluable, setting the stage for my future career. Born, raised, and educated in northern Italy, I graduated in radio technology from the A. Rossi Technical Institute in Vicenza in 1960. My first job was assistant engineer at the Olivetti Electronic R&D Laboratory near Milan, where Olivetti was developing its early electronic computers. By a series of fortunate coincidences, in 1961 I ended up codesigning and building a small experimental electronic computer with 4K words of magnetic core memory. I was only 19 years old, and I had four technicians working for me, helping with the construction of that computer. The computer used approximately 1,000 logic gates, made with germanium transistors (fabricated in Italy by SGS-Fairchild), housed in a couple of hundred small printed circuit boards. Silicon transistors would have been faster, but they were too expensive, and integrated circuits (ICs) had

Digital Object Identifier 10.1109/MSSC.2008.930938

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.
1943-0582/09/$25©2009IEEE

just been invented and were not yet commercially available.

At the end of that project I decided to go back to school and study physics at Padua University, where I received a doctorate in physics, summa cum laude, with an experimental thesis in flying-spot scanners. In 1967 I joined SGS-Fairchild in Agrate Brianza, where I became MOS group leader, developed SGS's first manufacturing process for high-threshold-voltage MOS ICs, and designed the company's first two MOS integrated circuits. SGS-Fairchild was then the only Italian semiconductor company, 30% owned by Fairchild Semiconductor, and a licensee of Fairchild bipolar technology. In February 1968, SGS sent me to the R&D Laboratory of Fairchild Semiconductor in Palo Alto, California, as part of an engineer exchange program between the two companies. I was supposed to stay in the United States for six months and then return to Italy. I never went back except as a visitor!

The Early Years

In 1968, the vast majority of integrated circuits sold in the world used bipolar technology. They were all made with the revolutionary planar process, pioneered by Fairchild in the late 1950s to batch-fabricate silicon transistors. Competing with bipolar technology was another emergent technology, called MOS (metal oxide semiconductor) technology, considered by some, including myself, to be the future of ICs. The working principles of MOS transistors were quite different from those of bipolar transistors, relying on surface phenomena at the interface between silicon and silicon dioxide rather than the bulk semiconductor properties exploited in the bipolar devices. This fundamental difference made MOS integrated circuits physically smaller and simpler to fabricate than bipolar ICs, although their operating speed was far slower than bipolar.

For the same cost and the same power dissipation, MOS technology promised digital ICs with about ten times more logic gates than bipolar technology, although operating at a much slower speed. MOS technology was still controversial, however, with many people still skeptical about its viability, given its major speed limitations and its poor reliability record. Nonetheless, a few start-up companies had already sprouted in Silicon Valley to take advantage of MOS technology to make either serial memory, using dynamic shift registers, or to make emergent applications where high complexity at modest speed would be adequate—for example, desktop calculators, which at that point were still built with electromechanical technology.

In 1968, the only MOS technology in production was high-threshold-voltage p-channel MOS, and R&D work was being carried out around the world to develop a low–threshold-voltage technology that could more easily be made TTL compatible. (TTL, or transistor-transistor logic, had become the standard logic family of the industry and required 5-V operation. Standard MOS required 24 V, and low-threshold MOS was expected to use a supply voltage of 12 to 17 V.)

The Holy Grail

The holy grail of MOS technology was already recognized by many MOS experts. It was called MOS

© PHOTO FIX2 & IMAGESTATE

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 13:49:53 UTC from IEEE Xplore. Restrictions apply.

self-aligned gate, and it promised to greatly improve the speed of MOS ICs by eliminating a major parasitic effect plaguing the technology: overlap capacitance. In conventional MOS transistors, the source and drain regions of the transistor were made first, and formation of the thin-oxide gate region followed. The gate region overlapped the source and drain junctions by an amount sufficient to compensate for the misalignment introduced by the lithographic equipment and still guarantee a minimum overlap under worst-case conditions.

This requirement meant that much more overlap was necessary than the minimum required for the proper operation of the transistor. A particularly adverse effect occurred when the gate misalignment was in the direction to increase the overlap with the drain junction. In this case the overlap capacitance, multiplied by the gain of the stage (due to the so-called Miller effect), would appear as an additional gate capacitance. This substantial increase in gate capacitance not only reduced the circuit speed but also induced large speed variations from wafer to wafer.

The cure for this problem was to form the gate first, rather than last, and use the gate to define both the source and drain regions, thus creating perfect alignment every time. Unfortunately, the aluminum used for the gate electrode was not suitable for such a purpose because it could not withstand the high temperature required to create the source and drain junctions; a different, more refractory material, was required.

Silicon Gate Technology

My first assignment at Fairchild was to develop a low-threshold-voltage, self-aligned gate MOS technology using a gate electrode made of amorphous silicon, following the work of J.C. Sarace and collaborators who, at Bell Labs, had succeeded in building self-aligned gate MOS transistors using amorphous silicon. The structures built by Sarace, however, were adequate only to prove the working principle; they were not suitable for fabrication of integrated circuits. Much more work and innovation was required to have a technology suitable for mass production of silicon gate ICs.

During the first few weeks on the project, I invented the process architecture, followed by the design of the detailed process flow, succeeding in fabricating self-aligned gate p-channel, low-threshold voltage MOS devices a few months later. These low-threshold-voltage devices were built with amorphous silicon gates, using <111> silicon wafers, instead of the <100> silicon required for low-threshold-voltage transistors with metal gates. The transistors achieved low threshold voltage by taking advantage of the reduced work function (by about 1.1 V) between a properly doped silicon gate and the silicon substrate—an observation made by Tom Klein at Fairchild. In the following months I also designed an integrated circuit to prove that the new process technology was indeed manufacturable. This chip became the world's first commercial IC to use self-aligned gates. It was an 8-b analog multiplexer with decoding logic, called the Fairchild 3708 (Figure 1), replacing the Fairchild 3705, a functionally equivalent chip built with metal gates and difficult to consistently manufacture within specifications.

During the development of the 3708, it became apparent that vacuum-deposited amorphous silicon was unreliable, tending to break or crack at oxide steps. Soon Tom Klein and I found a way to replace it with polycrystalline silicon produced by vapor deposition. By the end of 1968, the 3708 could be reliably manufactured and became commercially available. Compared with the Fairchild 3705, the 3708 was about four times faster, the on resistance of its large multiplexing transistors was two and one-half times smaller, and the junction leakage was at least ten times smaller.

With silicon gate came also the ability to do phosphorus gettering (a method to reduce metal contaminants) after the completion of the

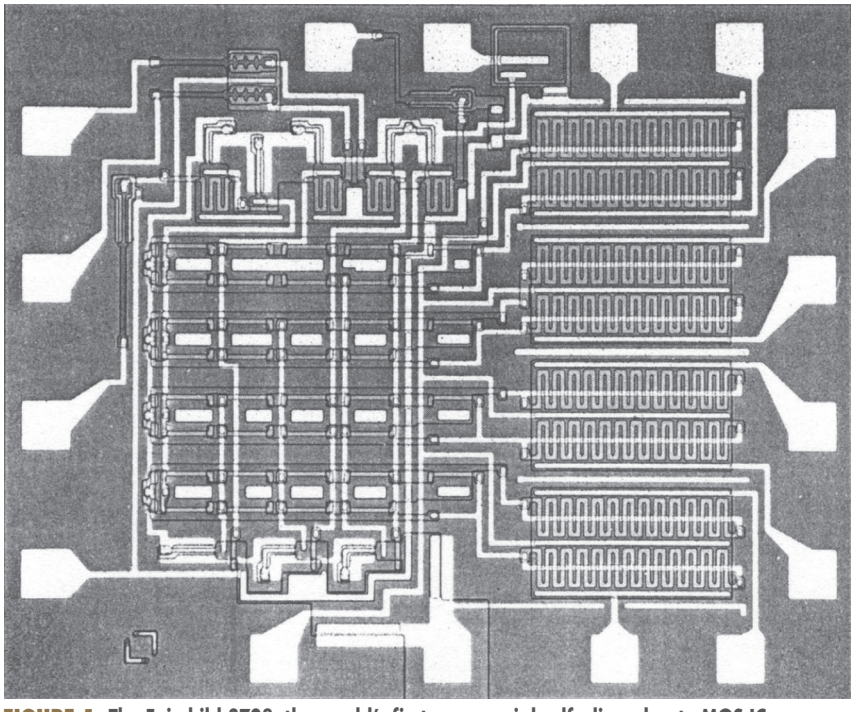

FIGURE 1: The Fairchild 3708, the world's first commercial self-aligned gate MOS IC, employing silicon gate technology and available in the market at the end of 1968.

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.

MOS structure, leading to a major improvement in both junction leakage and device reliability compared to metal gate, and opening up the path to the fabrication of dynamic random-access memories (DRAMs). This method was also first implemented in 1968 at Fairchild by the author. The unique ability to encase the silicon gate in thermal oxide—one of the best electrical insulators known—allowed also the creation of the first commercial nonvolatile memories (by Dov Frohman at Intel in 1971) and the first commercial image sensors made with charge-coupled devices (at Fairchild in 1973). Metal gate was unsuitable for such applications.

Essential Inventions

There are two other inventions I made while at Fairchild that proved essential to the microprocessor realization: the buried contact and the bootstrap load. The buried contact was a method to make a direct contact between polysilicon and junctions that did not involve the use of aluminum—therefore it was buried under a layer of silicon dioxide—and required only an additional masking step. This innovation made possible smaller contacts, but more important, it provided two layers of interconnections, one with polysilicon and one with aluminum, significantly increasing the circuit density, particularly for random-logic designs.

The bootstrap load was a very popular circuit design trick used in just about all MOS dynamic circuits of that time. It made possible an output signal swing that was not only equal to the power supply voltage, but was also faster than possible with normal MOS loads for the same power dissipation. In normal loads, the output swing was equal to the supply voltage minus the threshold voltage of the load transistor, which was significantly augmented by the "body effect." To make bootstrap loads, however, it was necessary to fabricate isolated capacitors, trivial to make with metal gate technology, but impossible with silicon gate technology without the use of an additional masking step, which would add significant cost to the process. In those days the p-channel MOS process required just five masking steps, compared with the 20–40 masking steps of contemporary ICs.

There are two other inventions I made while at Fairchild that proved essential to the microprocessor realization: the buried contact and the bootstrap load.

This limitation was considered very serious by the Fairchild chip designers and was delaying the adoption of silicon gate technology by the MOS division—a source of major frustration for me. For some time I also believed that this limitation was insurmountable as I struggled to find a solution. Eventually I realized that, under the normal operating conditions of a bootstrap load, there would always be a *virtual* junction under the polysilicon that could be used for one of the two electrodes of the capacitor, thus eliminating the need for a real junction and an extra masking step. The virtual junction was created by the inversion layer induced by the specific biasing conditions of the bootstrap capacitor. I then successfully designed and fabricated a number of bootstrap load structures to verify and optimize their operation, just months before joining Intel, where Intel engineers were still convinced that bootstrap loads could not be made with silicon gate.

With the addition of the buried contact and the bootstrap load, the silicon gate technology was now in all respects better than the incumbent metal gate technology. It allowed a designer to integrate in the same chip size about twice the number of random-logic transistors and achieve five to ten times the speed of the incumbent technology—for the same power dissipation and two-phase clock design (then called quasi-static design).

Silicon gate technology with buried contacts and bootstrap loads became one of the essential ingredients required to make the microprocessor feasible in 1970. The only other viable method to make complex random-logic circuits with high-threshold-voltage metal gate technology was to use fully dynamic circuits with four-phase clocks. This was a relatively complex technique, requiring computer-assisted design, successfully used by Rockwell Semiconductor and Four-Phase Systems to produce calculator and computer chips. Silicon gate technology, however, using simpler, quasi-static two-phase design, was superior in both speed and circuit density to the best four-phase designs. Low-threshold-voltage *metal* gate MOS technology was eventually developed with the help of ion implantation in 1970–1971, allowing metal gate MOS to narrow the performance gap and better compete with silicon gate technology for a short period of time before succumbing to it. By 1974–1975, the entire MOS industry had switched to silicon gate for all new MOS designs, and it is still in use today.

Intel Corporation

In the summer of 1968 I was shocked when I heard that Bob Noyce and Gordon Moore had left Fairchild Semiconductor to start another company. Soon Andy Grove and Les Vadasz, my boss, also left the Fairchild lab to join what later became known as Intel.

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.

Only weeks prior to Noyce's and Moore's departure I had decided to accept an offer to remain at Fairchild and not return to Italy, and now I was baffled by this start-up phenomenon that I had never witnessed in Italy. When my boss left, I immediately felt that Intel was going to use the silicon gate technology I had almost finished developing. In little more than a year, my hunch proved true.

In time the Intel mission became clear as well: it was to profit from the emergent market for semiconductor memories, particularly random-access memories (RAMs), by replacing the incumbent magnetic core memory with semiconductor memory. Intel intended also to sell semiconductor components to build small memory systems for which magnetic core memories were not well suited, because their large fixed overhead cost was independent from the number of bits. For such small memory systems, the only practical solution was to use serial memory made with either magnetostrictive materials or with MOS shift registers.

Toward the end of 1969, Fairchild had become a slow-moving company, crippled by the defection of many key people to Intel and other start-up companies and by its own success. I was frustrated by the slow adoption of silicon gate technology by the Fairchild MOS division, and soon after Intel announced its first silicon gate MOS product—a 256-b static RAM (the Intel 1101)—I decided to start looking for another job. My desire was to become a large-scale integration (LSI) chip designer using the very technology that was empowering this new trend; silicon gate technology. I felt this was the new frontier, and I wanted to return to my first love, building systems, but this time in a chip instead of in a rack of printed circuit boards as I had done at Olivetti.

In April, 1970 I joined Intel, working for my very first boss in the United States, Les Vadasz, who now was heading the Intel MOS Design department. During my interview process, Les had been purposefully very vague in describing the project I was supposed to lead, but he assured me that it would satisfy my hunger for a challenging chip-design project.

The Busicom Project

On my first day of work, I met Stan Mazor, an engineer working for Ted Hoff, the manager of the Application Research department, who described the "Busicom Project." He told me the story of how it evolved from a Busicom proposal of seven custom LSI chips, three of which were dedicated to make a special-purpose CPU, to an Intel proposal (spearheaded by Ted Hoff) of a set of four chips where the CPU was general-purpose and entirely integrated in one chip. Stan also gave me the basic specifications of the four chips, developed over a period of a few months between Intel's Hoff and Mazor and Busicom engineers, Masatoshi Shima being the lead engineer. Stan also told me that Shima was arriving in a few days to check on the progress, expecting to find the logic design of the CPU completed and the other chips in an advanced state of design. The problem was that since late 1969 no work had been done on the project, and Busicom was not told about it.

In the case of the 4000 family, this first step was led by Hoff with the assistance of Mazor and the Busicom team, Shima in particular.

When I saw the project schedules that were promised to Busicom, my jaw dropped: I had less than six months to design four chips, one of which, the CPU, was at the boundary of what was possible; a chip of that complexity had never been done before. I had nobody working for me to share the workload; Intel had never done random-logic custom chips before, and, unlike other companies in that business, had no methodology and no design tools for speedy and error-free design. Furthermore, my boss was consumed with the key project going on at that time, the 1103, and made it clear to me that he had little time for me. The Intel 1103 was the first dynamic 1024-b RAM, the product that was intended to make Intel successful in the semiconductor memory market, after a lukewarm market response to the 1101 and to the 3101 (64-b static bipolar RAM). Both Vadasz and Grove, my boss's boss, considered my project a diversion dreamed up by the marketing guys to make some money while waiting for the memory business—the real mission of Intel—to mature.

The Busicom project schedule had been clearly put together without much thought, since it had a CPU layout time of seven weeks, only two weeks more than a simple memory chip. A memory chip is a repetitive design whose layout is substantially faster to plan and draw than random logic, where almost every circuit is unique and has to be custom fitted. Therefore, not only was the project starting about five months later than promised to the customer, but also the duration of each project phase had been underestimated, particularly for the CPU.

Fortunately I was young and eager to prove myself in my newly chosen field. I understood computers, I could design both logic and circuits, and I had a lot of experience in developing MOS processes and MOS ICs—a very rare combination indeed, even in those days—therefore I felt that if I couldn't do

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.

it, nobody could. In particular, my intimate knowledge of silicon gate technology gave me an opportunity to develop a new methodology for random logic design that could take advantage of the strengths of that new technology. The methodology was indeed very successful and was used for all the early microprocessors from Intel and Zilog, the microprocessor company of which I later was cofounder and CEO.

Shima Arrives

Within a few days of joining Intel, Stan Mazor and I met Shima at the San Francisco airport on his arrival from Tokyo. Shima was eager to check the progress made since his last visit in December 1969. In particular, he wanted to check the logic design of the CPU and make sure that it would perform according to the agreed-on specification. When we arrived at the company, I gave Shima the material I was given by Stan a couple of days earlier. Shima was furious when he found out that no work had been done in the last five months and became very angry at me, the project leader, literally calling me names. I could not convince him that, having joined Intel only a few days before, I could not have done the work he expected. He said, "I came here to check, and there is nothing to check! This is just idea!"

He said that his project was irreparably compromised and that he had to call his management to find out what to do. It took almost one week for Shima to calm down and accept what happened. During that time I resolved the remaining architectural issues, I started working on the design methodology, and I prepared a new schedule that would give Busicom "first silicon" (the first chip to actually be fabricated from a design) of all four chips by the end of December, assuming I could get one engineer and a couple of draftsmen on time to help me.

This new schedule was extremely ambitious and would require me to work 70–80 hours per week to make up for the delay. I also told Shima that if he would help me there was a chance to meet the new schedule, since it would take time to hire the people I needed. Finally the difficulties were resolved; Busicom accepted the new schedule; Shima got permission to stay for six months to help me; and I could concentrate on designing what I now called the 4000 family.

The 4000 Family Takes Shape

Designing a production integrated circuit required many steps, starting with the definition of the chip architecture and its basic specifications. In the case of the 4000 family, this first step was led by Hoff with the assistance of Mazor and the Busicom team, Shima in particular, who verified the suitability of the specifications for Busicom's several applications. The task of the Application group was finished with the completion of the specifications.

The actual design and development of the chips was done in another department, the MOS Design department, and was entirely led by me without any further contribution by Hoff and Mazor. The design and development steps followed the sequence:

1) logic design
2) circuit design
3) composite layout design
4) ruby cutting (see "The Ruby-Cutting Procedure")
5) mask generation
6) wafer processing
7) first silicon
8) chip verification
9) debugging and characterization
10) production test-pattern development
11) transfer to manufacturing.

Steps 5 and 6 were usually done by groups outside the Design group. Generally the steps from specifications completed to first silicon, would take at least six months for a simple chip, longer for a complex chip. From specifications to transfer to production—at which point the responsibility for the product would move from the MOS Design department to Manufacturing—would normally take from ten to 18 months.

Since Intel had designed only memories up to that point, it had not yet developed a methodology for random-logic design as it existed

THE RUBY-CUTTING PROCEDURE

A few words about composite drawing and ruby cutting are in order, since the terms may sound foreign to today's chip designers, who do everything sitting in front of a workstation screen. The chip composite layout was drawn by hand with a straightedge and colored lead pencils at 400 to 500 times the actual scale, in a large, reclining drafting table over a Mylar quadrille sheet—Mylar was used for dimensional stability. The composite layout included all the masking layers of the chip superimposed with their proper registration.

Since the composite could not be used directly to generate the masks necessary for the manufacturing process, it was necessary to prepare a separate layer for each mask, to be photoreduced into a "reticle," a ten-times-larger version of one of the masks for the chip. The artwork used for generating a reticle was called rubylith, or ruby for short, and was obtained by first laying a sheet of Mylar covered by a thin red film over the composite drawing—which served as a guide—on a precision cutting table. The red film on the Mylar was cut and peeled off with tweezers in correspondence with the areas to be etched on the chip, thus producing a rubylith of the same size as the composite drawing, but showing only one of its layers.

The large ruby would then be photographed in a gigantic camera and reduced to a black-and-white reticle at ten times magnification. The reticle would then be mounted on a special "step-and-repeat" camera, which reduced the image to actual size and repeatedly exposed it onto a photographic glass plate until the plate's entire surface was covered with an array of patterns. This process produced the "master," out of which "submasters" and then "working plates" would be produced by contact photography. The working plates were then mounted in the lithographic equipment that transferred the pattern to the silicon wafers.

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 13:49:53 UTC from IEEE Xplore. Restrictions apply.

in companies like Fairchild, Texas Instruments, AMI, and others in the business of designing custom chips. Such companies had extensive libraries of circuits and circuit blocks; working layouts for various structures; computer simulation tools for logic, circuit design, and test program generation; characterization tools; random-logic testers; and, most important, random logic designers expert in the entire process.

Although I didn't know it at the time, my early work experience turned out to be absolutely invaluable, setting the stage for my future career.

Furthermore, silicon gate technology was new and required quite a different layout style than the one used with metal gate, particularly with buried contacts. In fact, several Fairchild MOS chip design engineers had complained to me that using silicon gate they always ended up with larger circuits than with metal gate, rather than the smaller ones I had promised. When I checked their layouts, I found out that they were trying to copy exactly the layout they had done with metal gate, instead of figuring out the natural way silicon gate needed to be laid out. Sure enough, when I showed them how to do the layout properly, the resulting silicon area was quite a bit smaller.

I was hired to design and lead the development of the four Busicom chips, and take them all the way through to the transfer to manufacturing. However, since Intel was new to random-logic custom circuits, I needed to carry out many more tasks than a typical project engineer working for a company already in the custom chip business had to do. For example, the logic design was normally done and verified by the customer. In fact, Busicom originally had come to Intel with the complete and verified logic design of their seven-chip architecture, but since the Busicom proposal had been rejected, I had to do the logic design and verification of the Busicom chips as well. Most important, I had to figure out and create a random-logic design methodology for silicon gate technology that didn't yet exist. I also had to design and build a characterization tester, and finally design and build a production wafer-sort tester for the 4004 to supplement the expensive final-test equipment that was purchased only toward the end of the project.

Since I had promised the customer, under duress, to deliver samples of all four chips by December, 1970—less than nine months from start—and since the CPU alone would take almost eight months, I had to work practically on all four chips in parallel, staggering them a bit so that the critical layout resources would be kept continually busy. I decided to design the 4001 first, followed by the 4003, 4002, and 4004—the last being the CPU. This sequence allowed me to incrementally develop the methodology and the building blocks I needed to use for the most complex chip, the 4004, and also to regain Busicom's confidence by showing early success with chips working first time.

The 4001 was a state-of-the-art 2048-b metal-mask-programmable read-only memory (ROM) with four metal-mask-programmable I/O lines. I did the logic and circuit design of the 4001 in a couple of weeks and gave it to Shima to check. Shima was an excellent logic designer and was also the engineer slated to develop the firmware of the Busicom desktop calculator—the first intended application of the 4000 family. Just like Hoff and Mazor, Shima was not a chip designer and didn't know much about MOS technology, but he was eager to learn and was very detail oriented—a highly valuable quality given the lack of verification tools at Intel.

One of the early challenges encountered during the 4001 design was to invent a flip-flop that was guaranteed to come up in a known state after turning the power supply on, since there were no extra pins in the 4001 to dedicate to a reset signal (each chip was packaged in a 16-pin dual in-line package, or DIP!). This flip-flop was to be used in the critical control of the tristate external bus that connected all the chips, to avoid contention after the power supply was turned on. I came up with a circuit that I later patented for Intel.

The 4001 layout started the day my first layout drafter, Rod Sayre, showed up for work. He was hired from Lockheed, where he was a mechanical drafter, and he had never seen a chip, never mind laid one out—in those days it was hard to find layout draftsmen. All the experienced Intel draftsmen were busy with memory projects, and I could not use any of them. I trained Sayre, and in time he became a very good drafter, but at the beginning, and for the duration of the 4001 and 4003 layouts, for which he was the only drafter, I had to draw myself all the building blocks freehand and Sayre would copy them properly in the composite layout.

After the 4001 layout was completed (Figure 2), Rod Sayre laid out the 4003, which was the only really simple chip of the 4000 family and only took two to three weeks to lay out. The 4003 was a 10-b static shift register with serial input, serial output, and gated parallel outputs. For its design I used a novel flip-flop that I had coinvented and patented in Italy while working for SGS. I also used the same circuit for many of the counters in the 4002 and 4004. The

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.

next chip to start was the 4002, the data RAM of the family. The 4002 was organized as four registers of 16 + 4 nibbles each, for a total of 320 b, and in addition it had a 4-b output port. Again, I did the logic and circuit design in a couple of weeks, and Shima checked my work. It was good to have somebody else check my work, given that there was no time to do any logic or circuit simulation, and Shima was very thorough.

In the 4002 design I used a three-transistor dynamic RAM cell, similar to the one that was being designed in the 1103. The chip included also a fair amount of logic for the memory refresh; the decoding of some instruction, timing, and control circuitry; and a 4-b output register. The layout of the 4002 was done by a new drafter just hired from Intel, Julie Hendricks, the same person who had laid out the Fairchild 3708 a few years before, when she was a trainee. Fortunately Hendricks was experienced, though mostly in bipolar layouts, helping me considerably. When Sayre was finished with the 4003 layout, he joined Hendricks to help speed up the 4002 layout (Figure 3).

Finally I could start the logic design of the 4004, though I was slowed down considerably by having to keep the other three chips moving, all at different phases of the design process. I also needed a debugging and characterization tester in a couple of months when I expected to receive the first silicon of the 4001. Fortunately Hal Feeney, a design engineer, and Paul Metrovich, an electronic technician, were assigned to me to help with the design and construction of such a tester. We started with a discarded memory system, and we built a programmable pattern generator by adding electronics to it. We also designed adjustable-pin electronics and added a paper tape reader. The entire contraption was ready only days before I received the first 4001 wafers.

After I had designed a good portion of the logic of the 4004, Shima offered to complete the logic design, particularly the control section of the CPU. At this point I felt very comfortable that he could do that task after the learning he acquired by assisting me with the design of the prior three chips. By now I had perfected the methodology, particularly the method of combining logic and circuit design in a single document that also contained the notion of how the chip would be naturally laid out. This method avoided the potential translation errors in going from the logic diagram to the circuit diagram; it allowed the designer to focus

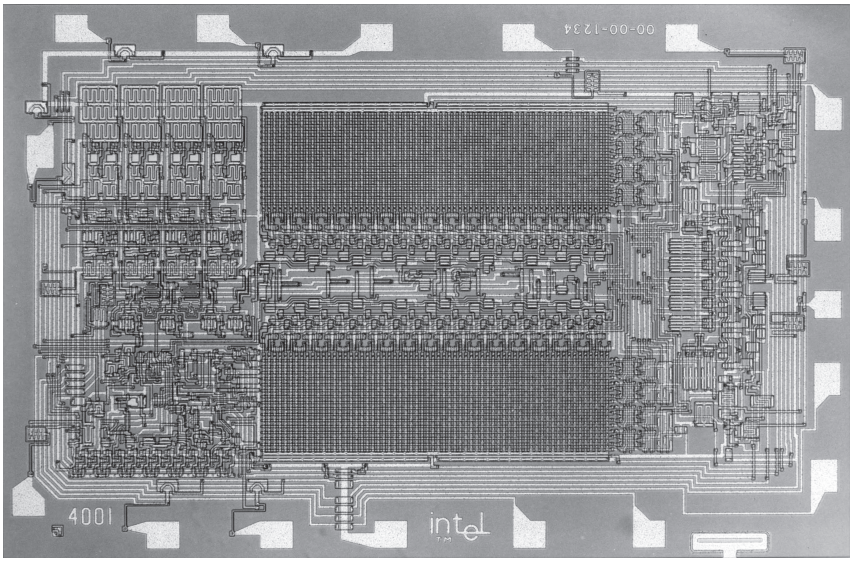

FIGURE 2: The Intel 4001. This chip was a 2048-b, metal-mask-programmable ROM, used to store the computer program. The chip also contained a section of metal-mask-programmable logic for its four input/output (I/O) lines.

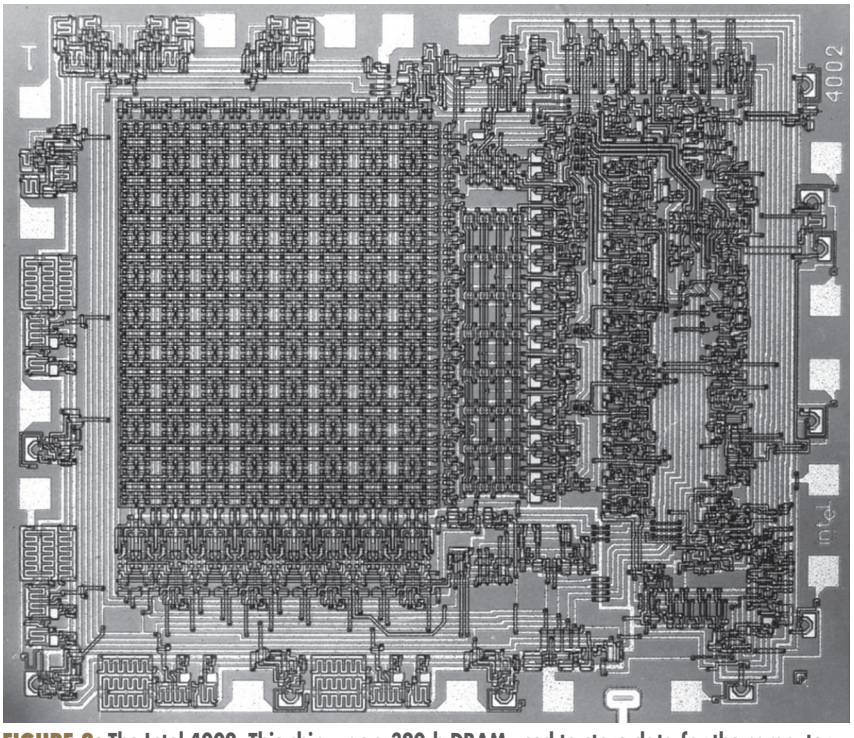

FIGURE 3: The Intel 4002. This chip was a 320-b DRAM used to store data for the computer. It contained its own memory refresh circuitry and four output lines with relative control logic.

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.

on the critical circuits, estimating the layout capacitances, and thus the transistor sizing, from the same document; and speeded up the layout by streamlining the translation from circuit design to layout, again reducing the potential for mistakes. Figure 4 shows the resulting chip.

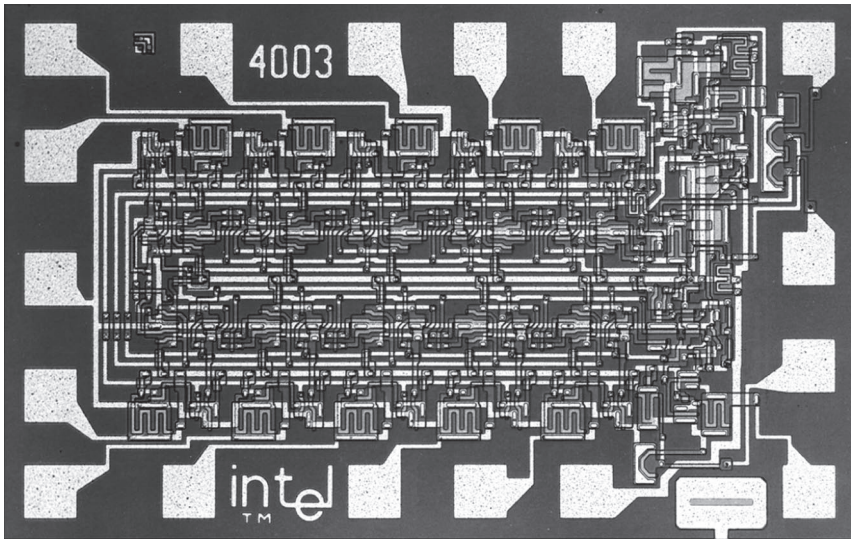

FIGURE 4: The Intel 4003. This chip was the only medium-scale integration (MSI) component of the 4000 family, containing a 10-b static shift register with parallel outputs to be used for I/O line expansion. The 4003 was quite helpful because each 4000 family chip was packaged in a 16-pin DIP package, severely limiting the number of I/O lines available in the LSI components.

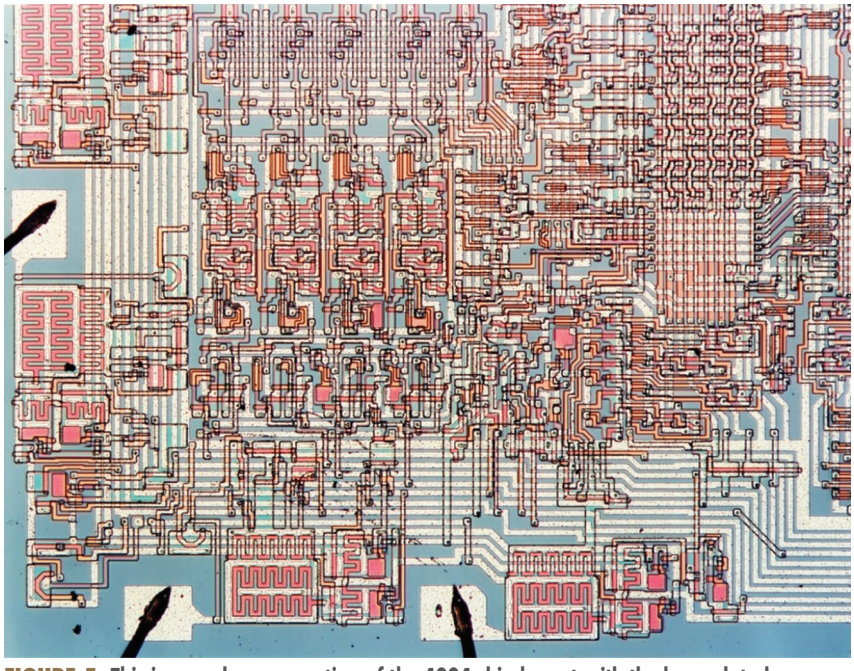

FIGURE 5: This image shows a portion of the 4004 chip layout with the large data bus drivers (the three large MOS transistors with the orange wavy lines, two on the bottom and one on the left of the photo). The external 4-b data bus was the main highway connecting all the chips together. For a system with many 4001s and 4002s, the capacitance of each data line could be several hundred picofarads, requiring powerful drivers. Moving from right to left on the image, one can see a portion of the control logic of the arithmetic unit, followed by a portion of the 4-b arithmetic unit. Notice the higher random logic layout density of the 4004 compared to the other three chips of the family. The 4004 was the only one of the four chips to use buried contacts.

Pressed for time, I had to start and closely supervise the 4004 layout before the design was completed, therefore I was coordinating with Shima so that I could keep the three drafters busy all the time while maintaining an excellent layout density, despite the fact that the design was not yet completed. For a chip at the limit of what could be economically produced, I could not afford to waste any precious silicon real estate. Joining Julie Hendricks and Rod Sayre in the 4004 layout team was Barbara Manness, an experienced memory layout drafter who had been at Intel nearly from the beginning. However, no Intel drafter had ever laid out complex random-logic circuits before, so the 4004 layout required close supervision on my part. The 4004 layout lasted about 14 weeks (42 worker-weeks) compared with the five worker-weeks of the 4001 layout. The original 4004 schedule prepared by Vadasz had predicted seven weeks with two drafters, for a total of 14 worker-weeks. Since each drafter had to work on a separate sheet of Mylar (using colored lead pencils), it was particularly challenging to maintain a good sense of the total chip.

When the 4004 layout was completed, I followed my impulse to sign my initials, F.F., on the metal mask, as artists autograph their creations. I felt it was a true work of art, where each stroke was not only aesthetic but also function-specific and meaningful. Figure 5 shows a portion of the 4004 chip (Figure 6, later in this article, shows the complete chip with my artist's initials).

Cutting the rubylith was a tedious and error-prone operation that required careful and time-consuming checking before the rubies could be sent to the mask making service. Since each rubylith represented only one layer of the chip, it was necessary to check its integrity and alignment by superimposing the rubies of the other layers, in various combinations.

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.

When the ruby cutting was finished, many people were recruited to help spot potential errors, and it could take several weeks to complete the checking, an operation that could be done only at the end of the cutting process. Checking the composite layout was far less stressful because it could at least be partially done during its drawing and not just at the end.

The ruby cutting of the 4004 was a challenge because the entire composite was larger than the cutting table and had to be done in two pieces (to the best of my recollection). Shima and I carried the brunt of the 4004 ruby checking—then the most complex chip ever done at Intel—but Hal Feeney and others also cheerfully joined forces to help speed up the process.

After the 4004 ruby checking, Shima returned to Japan with a brief detour to Egypt for a well-deserved vacation. I continued working 70–80 hours a week for many more months before I could take a brief break. During the peak of my work, my wife, Elvia, turned to her family for help with our three-month-old daughter, since I was so busy with the project. She went to Italy where she stayed for several months, allowing me to work very hard without feeling too guilty about being missing in action.

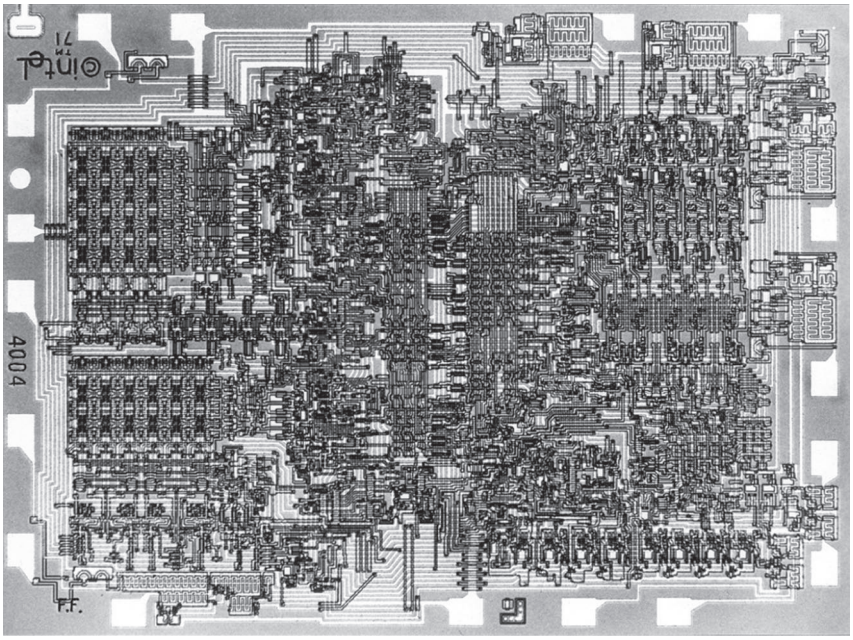

FIGURE 6: The Intel 4004, the world's first CPU-on-a-chip. This 4-b microprocessor contained approximately 2300 random-logic transistors. The 4004 basic instruction cycle was 10.7 μs and used eight clock periods of a two-phase clock running at 750 kHz. This time was longer than strictly necessary (by approximately a factor of 2.5) because of the heavy use of multiplexing to send the 12-b address, the 8-b instructions and the 4-b data on the same four data bus lines. The typical power dissipation of the 4004 was 750 mW. Notice the author's initials (F.F.) in the lower left corner.

The Moment of Truth

Shortly after Shima returned to Japan, I received the first silicon of the 4001. This was my first LSI chip design, and I was very nervous because it was the real test of my methodology: If the 4001 didn't work, all the other chips would have been hopeless because its design style was replicated in all of them. The characterization tester had just been completed enough to verify the 4001 operation, and I was delighted when the oscilloscope displayed the familiar waveforms I had drawn many times on paper and now were replayed live! I was stunned by the fact that the chip was doing exactly what it was supposed to do, after so much work and so many error-prone steps. The miracle of technology!

After a few days of checking and verification, everything was found to work as expected, not only functionally but also the clock speed at high temperature, and all the critical signal and supply margins were exceeding the design targets. It was a great relief! I passed the litmus test, and now I couldn't see any showstopper for the rest of the family. Busicom was also relieved that their first chip worked as expected.

A few weeks after receiving the first 4001 wafers, I also got the first silicon of the 4003. That chip also worked the first time, adding to my confidence level. In late November I received the first silicon of the 4002, which also was fully functional but for one minor mistake that was quickly identified and fixed.

Finally came the big day when I was given the first wafers of the 4004. The moment of truth had arrived. It was the end of the workday, a few days before New Year's Eve would usher out 1970, and most people had left the lab. It was fortunate, because nobody was around to see how nervous I was. My trembling hands placed the first wafer in the wafer prober. I lowered the probes into the first chip expecting to see the now familiar activity in the data bus, but instead nothing happened. "Oh, well," I said to myself, "that must be a bad chip." I lowered the probe on another chip with the same results, and then probed several more chips, always with the same symptoms. "Maybe this is a bad wafer," I thought. I tested another wafer, and got exactly the same behavior. By this time I was sweating profusely thinking, "Nothing works! How could I have screwed up so badly?" I decided to look at the chips under the microscope, and sure enough, the problem was obvious: during the manufacturing process the buried contact layer was left out by a technician's mistake, therefore most of the transistor gates were not connected, hence

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.

no life. Now my chance to meet the schedule had been blown away by a trivial mistake in manufacturing that was going to cost me about three weeks of delay.

About three weeks later I received a new run of 4004 chips. This time nothing was left out; I made sure of that by checking the wafers under the microscope *before* loading the first one on the probe station. As before, I received the wafers at the end of the workday when most people had already left the lab, and I set out to spend most of the night probing the 4004. I breathed much easier after the familiar signals in the data bus appeared in the scope. Now I was in business! I probed until 3 or 4 a.m., finding that everything was working as expected until, exhausted, I left for home.

Elvia had been waiting to hear the news. She woke up from a light sleep as soon as she heard my steps and immediately asked, "How did it go?" Still in a state of excitement I said, "It works!" And we shared feelings of exhilaration and happiness, knowing that something very significant had happened. That was the night the first microprocessor was born, almost nine months after I had started the project. I had just turned 29 the month before, and I realized that nine years earlier, at about this time of the year, I had just completed another computer, made with germanium transistors, that had about the same capabilities as this one, except the new one could all fit into a single printed circuit board, instead of a few hundred boards; had about ten times higher speed; and consumed almost 1,000 times less power. What a difference nine years make!

The story of the 1201 started with a visit to Intel by Computer Terminal Corporation toward the end of 1969.

FIGURE 7: Federico Faggin next to the engineering prototype of the Busicom calculator—the very first application of the microprocessor. This original artifact was a personal gift of Yoshio Kojima, president of Busicom, to the author in 1971. This working prototype was gifted by the author and his family in 1996, when this picture was taken, to the Computer History Museum, Mountain View, California.

In the following days I continued to check the 4004 and found a couple of minor problems that were relatively easy to diagnose and fix. In the meantime, after Shima's return to Japan, Busicom had finished building a 4000 system simulator with a full 4004 simulator and RAM replacing the 4001s so that the calculator firmware could be easily loaded, verified, and changed. This was necessary since the 4001 was a metal-mask-programmable ROM, taking several weeks to be fabricated and therefore appropriate only when the firmware was fully debugged. Shima, the developer of the calculator firmware, sent me the four verified ROM codes in January 1971, soon after he heard the news that the 4004 was working. Intel could then fabricate the 4001s in parallel with the new corrected version of the 4004. Therefore, by mid-March, when the revised silicon of the 4004 was expected, Busicom could build and verify the entire calculator with the final 4000 family components (Figure 7).

In January 1971, Young Feng, a new engineer, joined my team to help me with the extensive characterization and the transfer to production of the 4000 family. That same month, the success of the 4000 family brought another surprise: I was assigned the job of designing the 1201, supervising Hal Feeney who had been helping me with the 4000 family testing since about August 1970.

The Intel 1201

The story of the 1201 started with a visit to Intel by Computer Terminal Corporation (CTC, later renamed Datapoint Corp.), toward the end of 1969. CTC was a customer of Intel that purchased shift register chips for the memory of its computer terminals, a typical use of shift registers at that time. CTC had plans to build a new intelligent terminal, called Datapoint 2200, at the heart of which was a simple CPU of its own design implemented with TTL components. CTC wanted Intel to design a special custom *bipolar* RAM chip to be used as the stack register for its CPU.

When Stan Mazor found out the purpose of that custom chip—fresh from his participation in specifying the 4004—he ventured to tell CTC that Intel had the technology to put CTC's entire CPU on a chip, not just the stack memory. This was a pretty bold statement since the 4004 had not yet been designed, and he was not a chip designer! Eventually CTC was convinced that Intel could integrate the CTC 8-b CPU into a single MOS chip and signed a contract for the development of a custom product, called the 1201. Hal Feeney was hired to lead the design of the 1201, and joined Intel in March, 1970, just weeks before I did. Before Intel, Feeney had worked for General Instruments, where he designed a number of custom MOS random-logic chips.

I found out about the 1201 soon after I joined Intel, and I was disappointed to find out that there

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.

was another microprocessor in development at Intel. Clearly the 1201 would be finished before the 4004, I reasoned, since Feeney had to design only one chip while I had four, and the 4004 was going to be my last one. I was so busy with my own challenges, however, that I soon forgot about it. The 1201 project, however, dragged along for several months but never got into high gear and eventually was mothballed; Feeney was reassigned to a memory project and then to me. Fundamentally, he was overwhelmed by the magnitude of the project—not only by the complexity of the job, but also because he had never designed a chip with silicon gate technology before and the lack of design methodology and support made the task daunting.

I inherited the 1201 project after the 4004 was essentially completed and my experience, combined with the now-proven methodology, allowed me to lead the project to its successful conclusion with Feeney doing the detailed design. The 1201 took the entire year of 1971 to be designed, with first silicon out in December, and became commercially available in April 1972 with the name 8008. The 8008 is at the origin of the spectacularly successful x86 family of Intel microprocessors that are powering most of the personal computers in use today.

Another interesting development occurred in April 1971, when Texas Instruments (TI) announced having successfully designed a CPU-on-a-chip, as TI called it, only one month after the 4004 was fully functional. In other words, TI had also designed a microprocessor. We later found out that such development started as a custom project for CTC, which wanted a second source for its CPU. The specification of this chip was of course identical to the 8008, except TI used low-threshold-voltage metal gate technology for its design. The reported TI chip size was twice that of the 8008, a size Intel would have considered prohibitive to produce, clearly showing the advantages of the silicon gate technology with buried contacts. I later surmised that the highly competitive nature of the semiconductor business, stirred up by the customer's self-interest, convinced TI that, "If Intel can do a CPU-on-a-chip, so can we!"

I personally thought that there were many control applications where the 4000 family would do well, and I set out to find out for myself.

Many years later I was told by Vic Poor—CTC's vice president of engineering in 1971—that the TI chip never functioned, and of course it was never used. TI also never made that chip available in the market, even after Intel's announcement of the 4004 and the 8008. It was only used for public relations purposes. This simple fact serves to prove that the *implementation* of the microprocessor was far from a routine design job. TI was then a leader in MOS custom chip development with many powerful design tools and much experience in random logic design, yet it could not make its first microprocessor work, though TI claimed it did. If the TI chip did not work, my point is immediately evident, but even if the chip did work, it happened a few months after the 4004 was completed, thus proving again that "inventing" the microprocessor was still an issue of implementation.

Had I not worked so hard, TI would have beaten Intel to the punch, and they would now be properly hailed as the inventors of the microprocessor, rather than Intel. After all, the essence of the microprocessor was the successful design of a CPU into a *single* chip. That was the crucial step that had not been done before. Many people knew how to architect simple CPUs and many people knew how to do logic design. A CPU in a few MOS chips had already been done before the 4004 by Four-Phase Systems, for example. Furthermore, given the clear semiconductor industry trends, a CPU-on-a-chip was inevitable. It really was only a question of who would do it first, and unquestionably it happened at Intel.

Announcing the Microprocessor to the World

During the design of the 4000 family, I found out that Intel had entered into a contractual arrangement with Busicom giving them exclusive rights to their use. I was upset because, seeing the great potential of the microprocessor, I wanted my work to have a bigger impact than just being a custom job for Busicom. With the project nearing completion, I started lobbying with Intel's management to sell the 4000 family in the open market. Hoff and Mazor believed that Busicom would at most give up their rights for noncalculator applications, and they felt that the 4000 family was only good for calculator-like applications, therefore they were not initially convinced that selling the 4000 family in the open market for noncalculator applications was a good idea. They felt, however, that the more general-purpose architecture of the 1201 made it more suitable for general-purpose use than the 4000 family. Of course, introducing the 1201 in the market was also problematic since the 1201 was originally bound by an exclusivity arrangement with CTC similar to that with Busicom for the 4004.

I personally thought that there were many control applications

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:45:53 UTC from IEEE Xplore. Restrictions apply.

where the 4000 family would do well, and I set out to find out for myself. The opportunity came with a new project I needed to start: a wafer-sort tester for the 4004. I decided to use the 4004 to perform the control logic for the tester, instead of using random logic for it. I figured that in that way I would find out firsthand whether or not the 4004 was appropriate to the task. I would also gain insights into what a customer would have to do to apply the 4000 family to solve problems. Finally, I was very interested in programming the 4004, a task I had never done before.

Since there were no programming tools for the 4004—that was considered the job of Busicom—and I was pressed for time, I wrote the tester control program by using the instruction mnemonics, and then I had to literally translate it by hand into machine language—the ones and the zeros that needed to be stored in ROM, the 4001. However, since the 4001 was metal-mask-programmable, I decided to use instead a product that had just been developed by Dov Frohman at Intel: the 1702, the world's first electrically programmable, ultraviolet-erasable ROM. The 1702 was intended to aid the development, debugging, and prototyping of ROM codes that, at the end of the market testing, would then be translated into a conventional mask-programmable and pin-compatible ROM, the 1302, replacing the 1702 for production in exactly the same printed circuit boards. All I had to do, then, was to design an appropriate interface between the 4004 and the 1702 to make the 1702 behave like a 4001.

Finally came the big day when I was given the first wafers of the 4004.

The tester project was quite successful and convinced me that the 4004 could be effectively applied for control applications. I used that experience to argue to management that the 4000 family had market value, building a more convincing case for it, particularly with Ed Gelbach, Intel's new vice president of marketing. Finally, during a phone conversation with Shima, around the middle of 1971, I found out confidentially that Busicom was not doing well in the market and could not compete effectively because the price they were paying for the 4000 family chips was too high. Shima also told me that Intel CEO Bob Noyce and Ed Gelbach were going to visit Busicom shortly. That information gave me the break I needed: I told Bob Noyce about my conversation, suggesting that he might be able to get a release from exclusivity from Busicom, for noncalculator applications, in exchange for a lower price. Of course I also pushed once more the case that the 4004 was very good for control applications, as I had learned in my experience with the tester project.

Shortly after Noyce and Gelbach's visit to Busicom, I learned that Intel had been successful in negotiating a release from exclusivity and had decided to introduce the 4000 family in the market. I was delighted. Soon after that decision, Intel appointed Hank Smith to lead the marketing effort for the new microprocessor products, and Feeney and I from MOS R&D, together with Hoff and Mazor from Applications, helped the new marketing group prepare for the 4000 family market launch, with a new coined name: MCS-4, standing for microcomputer system 4-bit. The MCS-4 was soon to be followed in early 1972 by the MCS-8 introduction, with the 8008 at the core of the new family and the rest of the MCS-8 chip family being mostly standard Intel memories.

In November, 1971, the official birth announcement of the microprocessor to the world finally happened. A two-page spread in the well-read *Electronic News* magazine read: "Announcing a new era of electronics," and briefly described the microprocessor and its availability (Figure 8). This turned out to be a prophetic statement, a

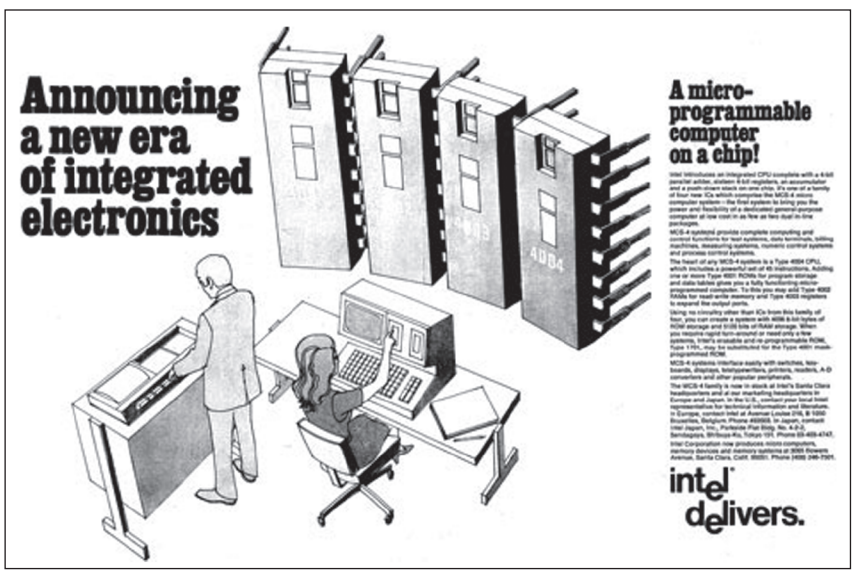

FIGURE 8: This is the first microprocessor advertisement, a two-page spread in *Electronic News* in November 1971.

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.

rare occurrence in advertising; the impact of the microprocessor on our lives has been truly extraordinary, as only a handful of other inventions in the last 100 years have been.

The success of the MCS-4 and the 8008, combined with my passion to see the microprocessor take root in the world, propelled my career for the next ten years. In late 1971 I conceived and architected the 8080, proposing the idea to Intel's management. It took about nine months before I got permission to start developing it. That delay reduced Intel's market lead to only six months, with strong competition from the Motorola 6800 microprocessor. I also conceived and architected the 4040, an improved version of the 4004 that could also use standard memories. I gradually took over more responsibility and more projects, and in early 1974 I was promoted to R&D department manager in charge of all MOS chip designs, except for dynamic memories. My major contribution in memories was the redesign of the Intel 2102 (5-V, 1024-b static RAM), proposing a new n-channel process technology with depletion loads that my boss, Vadasz, opposed. I eventually succeeded in convincing him to use depletion load, and the new product, called the 2102A, was spectacularly successful, with an access time four times less than the older version. That same process technology was then used for several future generations of memories and microprocessors.

Becoming an Entrepreneur

Intel was a memory company making microprocessors in order to sell more memories. I wanted to be in a company whose core business was microprocessors, not memories. Since such a company did not exist, I had to start my own. I invited Ralph Ungermann, one of my managers, to join me, and by the end of 1974 Zilog was born. At Zilog I initially conceived a microcontroller (which later became the Z8), but then I opted for a new generation 8-b microprocessor, the Z80. Introduced in mid-1976, the Z80 became wildly successful, powering many of the early personal computers and hundreds of other applications. The Z80 is still in high-volume production in 2008, 32 years after its market debut.

The Z80 was the last engineering project I directed, marking the end of my technical career and the beginning of my entrepreneurial career, which is still going on to this day. I was the first CEO of Zilog and the CEO of two other companies I cofounded. Five years ago I also became CEO of Foveon, a company with a highly innovative image sensor technology. I have been involved in many other start-up companies as an angel investor as well. I love to bring new ideas into the world, and I find that the most effective way to do so is through a start-up company where focus, passion, and energy are at their highest.

References

[1] F. Faggin and F. Capocaccia, "A new integrated MOS shift register," in *Proc. 25th Int. Scientific Congr. Electronics*, Rome, 1968, pp. 143–152.

[2] F. Faggin, T. Klein, and L. Vadasz, "Insulated gate field effect transistor integrated circuits with silicon gates," in *Int. Electron Devices Meeting Tech. Dig.*, Washington, DC, 1968, p. 22.

[3] F. Faggin and T. Klein, "A faster generation of MOS devices with low threshold is riding the crest of the new wave, silicon gate IC's," *Electronics*, Sept. 29, 1969.

[4] F. Faggin and T. Klein, "Silicon gate technology," *Solid-State Electronics*, vol. 13, pp. 1125–1144, 1970.

[5] F. Faggin and M.E. Hoff Jr., "Standard parts and custom design merge in four-chip processor kit," *Electronics*, pp. 112–116, Apr. 24, 1972.

[6] F. Faggin, et al., "The MCS-4-An LSI micro computer system," *Proc. IEEE Region 6 Conf.*, 1972.

[7] F. Faggin, "Power supply settable bi-stable circuit," U.S. Patent 3753011, Aug. 14, 1973.

[8] M. Hoff, S. Mazor, and F. Faggin, "Memory system for multi-chip digital computer," U.S. Patent 3821715, June 28, 1974.

About the Author

Federico Faggin (Faggin@foveon.com) is president, chief executive officer, and a director of Foveon Inc. He is also chair of the board of Synaptics Inc., and of Zilog, Inc. He was the cofounder of both companies as he was of Cygnet Technologies Inc. He worked for Intel Corp. from 1970 to 1974. In 1970–1971 he led the design and development of the 4004, the world's first microprocessor, and the other three chips of the MCS-4 product family. In 1971–1972, he led the development of the 8008, the world's first 8-b microprocessor. He architected and led the development of the 8080 microprocessor and the 4040 microprocessor. He also designed or supervised the design of more than 25 other commercial integrated circuits. He joined Fairchild Semiconductor in 1968, where he led the development of MOS silicon gate technology and designed the world's first commercial integrated circuit to use that technology, the Fairchild 3708. Born, raised, and educated in Italy, Federico Faggin became a naturalized U.S. citizen in 1978. He is the recipient of many honors and awards including the 1988 Marconi Prize, the 1994 IEEE W. Wallace McDowell Award, and the 1997 Kyoto Prize. In 1996, he was inducted in the National Inventor's Hall of Fame for the coinvention of the microprocessor. He received the Lifetime Achievement Award of the European Patent Organization in 2006.

He holds a doctorate in physics, summa cum laude, from the University of Padua (1965). He also holds honorary doctor degrees in computer science from the University of Milan (1994), in electronic engineering from the University of Rome Tor Vergata (2002), and in electronic engineering from the University of Pavia, Italy (2007). SSC

Authorized licensed use limited to: IEEE Xplore. Downloaded on July 22,2024 at 15:49:53 UTC from IEEE Xplore. Restrictions apply.